Database Management, Graphing, and Statistical Analysis Using IBM-SPSS Statistics

Ryan W. Walters, MS
Research Analyst & Instructor
Division of Clinical Research and Evaluative Sciences
Department of Medicine
Creighton University
# TABLE OF CONTENTS

**Section I: Database Management**

Chapter 1 – Creating and naming variables ................................................................. 2

Chapter 2 – Entering Data............................................................................................. 4

Chapter 3 – Importing Data from Various Sources .................................................... 5
  Microsoft Excel ........................................................................................................ 5
  Microsoft Access ..................................................................................................... 8
  Delimited Files ....................................................................................................... 9

Chapter 4 – Exporting Your Data and Output ............................................................. 11
  Exporting Data to a New Database ........................................................................ 11
  Exporting Data to an Existing Database ................................................................ 12
  Exporting Output ................................................................................................... 14

Chapter 5 – Merging Files ......................................................................................... 16
  Add Cases ............................................................................................................... 16
  Add Variables ......................................................................................................... 17

Chapter 6 – Identifying Duplicate Cases ................................................................. 19

Chapter 7 – Restructuring the Dataset ...................................................................... 22
  Restructure Cases into Variables ......................................................................... 23

Chapter 8 – Recoding Variables ............................................................................. 25
  Recode Into Same Variables ............................................................................... 26
  Recode Into Different Variables ......................................................................... 29

Chapter 9 – Date and Time Wizard .......................................................................... 32
  Add or Subtract Time Units from a Date ............................................................... 32
  Calculate the Amount of Time between Two Dates ............................................ 33
Section II: Descriptive Statistics

Chapter 10 – Frequencies .................................................................37

Chapter 11 – Split File ....................................................................41

Chapter 12 – Select Cases ...............................................................43

Chapter 13 – Compute Variable ....................................................45

Chapter 14 – Descriptives ...............................................................47

Chapter 15 – Explore .....................................................................49

Output and Interpretation ...............................................................51

Section III: Graphing

Chapter 16 – Bar Charts .................................................................55

Simple Bar Chart, Summaries for Groups of Cases ......................55
Clustered Bar Chart, Summaries for Groups of Cases ..................57
Simple Bar Chart, Summaries of Separate Variables ...................59
Clustered Bar Chart, Summaries of Separate Variables ...............60

Chapter 17 – Line Charts .................................................................62

Simple Line Chart, Summaries for Groups of Cases .....................62
Multiple Line Chart, Summaries for Groups of Cases ..................64
Simple Line Chart, Summaries of Separate Variables ..................65
Multiple Line Chart, Summaries of Separate Variables ...............67

Chapter 18 – Boxplots .................................................................69

Simple Boxplot, Summaries for Groups of Cases .........................69
Clustered Boxplot, Summaries for Groups of Cases .....................71
Simple Boxplot, Summaries of Separate Variables .......................72
Clustered Boxplot, Summaries of Separate Variables .................73
Chapter 19 – Scatterplots ...................................................................................................75
  Simple Scatterplot ..................................................................................................75
  Matrix Scatterplot ..................................................................................................77
  Simple Dot Plot ......................................................................................................78
  Overlay Scatterplot .................................................................................................79
  3-D Scatterplot .......................................................................................................81

Section IV: Statistical Tests of Association

Chapter 20 – Pearson’s Product-Moment Correlation .......................................................84
  Assumptions ...........................................................................................................85
  Analysis ..................................................................................................................85
  Output ....................................................................................................................86
  Interpretation ..........................................................................................................87
  Example Results Section .......................................................................................87

Chapter 21 – Spearman’s Rank-Order Correlation Coefficient .........................................88
  Assumptions ...........................................................................................................88
  Evaluating the Linearity of Ranks Assumptions ...................................................89
  Analysis ..................................................................................................................89
  Output ....................................................................................................................90
  Interpretation ..........................................................................................................91
  Example Results Section .......................................................................................91

Chapter 22 – Simple Linear Regression ............................................................................92
  Assumptions ...........................................................................................................93
  Continuous Independent Variable ..........................................................................94
    Analysis ................................................................................................................94
    Output .................................................................................................................97
    Interpretation .......................................................................................................99
    Example Results Section ...............................................................................100
  Categorical Independent Variable ........................................................................100
    Analysis ................................................................................................................100
    Output .................................................................................................................103
    Interpretation .......................................................................................................105
    Example Results Section ...............................................................................106
Chapter 23 – Multiple Linear Regression........................................................................107
   Assumptions..............................................................................................................108
   Dummy Variables ..................................................................................................109
   Analysis ....................................................................................................................109
   Output ......................................................................................................................113
   Interpretation ..........................................................................................................115
   Example Results Section.......................................................................................116

Section V: Statistical Tests of Between-Subject Differences

Chapter 24 – One-Sample t-Test..................................................................................119
   Assumptions..............................................................................................................119
   Analysis ....................................................................................................................119
   Output ......................................................................................................................120
   Interpretation ..........................................................................................................120
   Example Results Section.......................................................................................120

Chapter 25 – Independent-Samples t-Test.................................................................121
   Assumptions..............................................................................................................122
   Analysis ....................................................................................................................123
   Output ......................................................................................................................124
   Interpretation ..........................................................................................................124
   Example Results Section.......................................................................................125

Chapter 26 – Mann-Whitney Test..............................................................................126
   Analysis ....................................................................................................................126
   Output ......................................................................................................................127
   Interpretation ..........................................................................................................128
   Example Results Section.......................................................................................128

Chapter 27 – Median Test.........................................................................................129
   Analysis ....................................................................................................................129
   Output ......................................................................................................................130
   Interpretation ..........................................................................................................131
   Example Results Section.......................................................................................131
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 28 – One-Way Between-Subjects ANOVA</td>
<td>132</td>
</tr>
<tr>
<td>Assumptions</td>
<td>132</td>
</tr>
<tr>
<td>Analysis</td>
<td>133</td>
</tr>
<tr>
<td>Output</td>
<td>137</td>
</tr>
<tr>
<td>Interpretation</td>
<td>141</td>
</tr>
<tr>
<td>Example Results Section</td>
<td>142</td>
</tr>
<tr>
<td>Chapter 29 – Kruskal-Wallis Test</td>
<td>143</td>
</tr>
<tr>
<td>Assumptions</td>
<td>144</td>
</tr>
<tr>
<td>Analysis</td>
<td>144</td>
</tr>
<tr>
<td>Output</td>
<td>145</td>
</tr>
<tr>
<td>Interpretation</td>
<td>145</td>
</tr>
<tr>
<td>Example Results Section</td>
<td>146</td>
</tr>
<tr>
<td>Chapter 30 – Factorial Between-Subjects ANOVA</td>
<td>147</td>
</tr>
<tr>
<td>Assumptions</td>
<td>147</td>
</tr>
<tr>
<td>Analysis</td>
<td>148</td>
</tr>
<tr>
<td>Output</td>
<td>153</td>
</tr>
<tr>
<td>Interpretation</td>
<td>156</td>
</tr>
<tr>
<td>Interpreting the Interaction Effect</td>
<td>157</td>
</tr>
<tr>
<td>Interpreting the Main Effects</td>
<td>157</td>
</tr>
<tr>
<td>Simple Main Effects Analysis</td>
<td>158</td>
</tr>
<tr>
<td>Interpretation of Simple Main Effects</td>
<td>159</td>
</tr>
<tr>
<td>Example Results Section</td>
<td>160</td>
</tr>
<tr>
<td>Chapter 31 – One-Way Between-Subjects ANCOVA</td>
<td>161</td>
</tr>
<tr>
<td>Assumptions</td>
<td>161</td>
</tr>
<tr>
<td>Evaluating the Homogeneity of Regression Assumption</td>
<td>163</td>
</tr>
<tr>
<td>Analysis</td>
<td>165</td>
</tr>
<tr>
<td>Output</td>
<td>168</td>
</tr>
<tr>
<td>Interpretation</td>
<td>171</td>
</tr>
<tr>
<td>Example Results Section</td>
<td>172</td>
</tr>
</tbody>
</table>
Section VI: Statistical Tests of Within-Subjects Change

Chapter 32 – Paired-Samples t-Test .................................................................174
  Assumptions ............................................................................................174
  Analysis ..................................................................................................175
  Output ....................................................................................................176
  Interpretation .........................................................................................177
  Example Results Section ........................................................................177

Chapter 33 – Wilcoxon Signed-Rank Test .....................................................178
  Assumptions ............................................................................................178
  Analysis ..................................................................................................178
  Output ....................................................................................................179
  Interpretation .........................................................................................180
  Example Results Section ........................................................................180

Chapter 34 – One-Way Repeated-Measures ANOVA ......................................181
  Assumptions ............................................................................................181
  Analysis ..................................................................................................182
  Output ....................................................................................................187
  Interpretation .........................................................................................190
  Example Results Section ........................................................................191

Chapter 35 – Friedman’s Test .......................................................................192
  Analysis ..................................................................................................192
  Output ....................................................................................................193
  Interpretation .........................................................................................194
  Example Results Section ........................................................................194

Chapter 36 – Mixed Between-Within ANOVA ..............................................195
  Assumptions ............................................................................................196
  Analysis ..................................................................................................197
  Obtaining Between-Person Residuals ......................................................203
  Output ....................................................................................................204
Interpretation........................................................................................................207
   Interpreting the Interaction Effect.................................................................208
   Interpreting the Main Effects ......................................................................209
Simple Main Effects Analysis of the Interaction Effect ..........................210
   Interpretation of Simple Main Effects ..........................................................211
Example Results Section......................................................................................212

Section VII: Statistical Tests for Categorical Variables

Chapter 37 – Pearson’s Chi-Square .................................................................214
   Assumptions...................................................................................................214
   Analysis...........................................................................................................215
   Output .............................................................................................................216
   Interpretation................................................................................................218
Example Results Section......................................................................................218

Chapter 38 – McNemar’s Test of Change .........................................................219
   Assumptions...................................................................................................219
   Analysis...........................................................................................................220
   Output and Interpretation..............................................................................220
Example Results Section......................................................................................221

Chapter 39 – Cochran’s Q .................................................................................222
   Assumptions...................................................................................................222
   Analysis...........................................................................................................222
   Output and Interpretation..............................................................................223
Example Results Section......................................................................................224

Chapter 40 – Simple Binary Logistic Regression ..............................................225
   Assumptions...................................................................................................226
   Categorical Independent Variable ..................................................................227
      Analysis......................................................................................................227
      Output .......................................................................................................230
      Interpretation ...........................................................................................232
Example Results Section........................................................................................................................................233
Continuous Independent Variable ....................................................................................................................233
Evaluating the Linearity in the Logit Assumption .........................................................................................233
Analysis..........................................................................................................................................................234
Output ..........................................................................................................................................................235
Interpretation..................................................................................................................................................237
Example Results Section................................................................................................................................238

Chapter 41 – Multiple Binary Logistic Regression ........................................................................................240
Assumptions..................................................................................................................................................241
Analysis..........................................................................................................................................................242
Output ..........................................................................................................................................................244
Interpretation..................................................................................................................................................247
Example Results Section................................................................................................................................248

Section VIII: Statistical Tests for Time-to-Event Data

Chapter 42 – The Kaplan-Meier Method .........................................................................................................250
Assumptions..................................................................................................................................................251
Analysis..........................................................................................................................................................252
Output ..........................................................................................................................................................255
Interpretation..................................................................................................................................................258
Example Results Section................................................................................................................................258

Chapter 43 – Simple Cox Regression without a Time-Varying Covariate .....................................................260
Assumptions..................................................................................................................................................262
Categorical Independent Variable..................................................................................................................264
Evaluating the Proportionality of Hazards Assumption .................................................................................264
Graphical Method..........................................................................................................................................264
Statistical Method..........................................................................................................................................266
Analysis..........................................................................................................................................................267
Output ..........................................................................................................................................................271
Interpretation..................................................................................................................................................274
Example Results Section................................................................................................................................275
Continuous Independent Variable..............................................................275
Evaluating the Form of the Independent Variable ................................275
Evaluating the Proportionality of Hazards Assumption ......................278
Analysis .................................................................................................279
Output .....................................................................................................281
Interpretation .........................................................................................282
Example Results Section ....................................................................284

Chapter 44 – Multiple Cox Regression without Time-Varying Covariates ....286
Assumptions ..............................................................................................287
Evaluating the Form of Continuous Independent Variables and Covariates ......289
Evaluating the Proportionality of Hazards Assumption .........................292
Analysis ..................................................................................................294
Output ....................................................................................................297
Interpretation .........................................................................................300
Example Results Section ....................................................................302

Chapter 45 – Simple Cox Regression with a Time-Varying Covariate ..........303
Assumptions ..............................................................................................304
Categorical Independent Variable ..........................................................304
   Evaluating the Proportionality of Hazards Assumption ......................304
   Analysis ...............................................................................................305
   Output ..................................................................................................308
   Interpretation ......................................................................................309
   Interpreting Interactions between a Categorical Independent Variable and Time ..............................................................................310
Example Results Section ....................................................................311

Continuous Independent Variable ..............................................................312
Evaluating the Form of the Independent Variable ................................312
Evaluating the Proportionality of Hazards Assumption ......................312
Analysis .................................................................................................313
Output .....................................................................................................315
Interpretation .........................................................................................316
Interpreting Interactions between a Continuous Independent Variable and Time .......................................................... 316
Example Results Section .......................................................................................................................... 317

Chapter 46 – Multiple Cox Regression with Time-Varying Covariates .................................................. 319
Assumptions ........................................................................................................................................... 320
Evaluating the Form of Continuous Independent Variables ............................................................... 321
Evaluating the Proportionality of Hazards Assumption ...................................................................... 321
Analysis ................................................................................................................................................. 322
Output .................................................................................................................................................. 325
Interpretation ........................................................................................................................................ 327
Example Results Section ....................................................................................................................... 329
Stratified Cox Model ............................................................................................................................... 330
   Evaluating the No Interaction Assumption ......................................................................................... 331
Analysis ............................................................................................................................................... 333
Output ............................................................................................................................................... 333
Interpretation ....................................................................................................................................... 335
Example Results Section ....................................................................................................................... 337
PREFACE

I want to welcome you to the world of SPSS. I chose to write this textbook using SPSS, as opposed to other statistical software packages, because SPSS is accessible and relatively straightforward to use. With that said, there is danger in accessibility and ease of use. Primarily, with SPSS you have the ability to indiscriminately point-and-click your way through some analysis to obtain some output. Garbage in, roses out? Definitely not!

I have written this text for an individual with at least one graduate level statistics course. However, those who have had a year of general linear model course (e.g., linear regression and ANOVA) will probably obtain the most use from it. Further, I have tried to keep technical terminology and mathematics to a minimum as well as use a conversational tone. This is an extremely applied text. My goal in writing it is to have you using SPSS quickly, while giving you the background knowledge of each statistical technique allowing you to maintain statistical rigor. With that said, I fully understand that some of you will find some holes in my descriptions and explanations. For this I apologize, but know that the answers to your questions are out there, you just have to dedicate the time and energy to find them. Asking for help is not an act of weakness; it is a means to an end in your quest for knowledge! I urge you to ask questions if you do not fully understand any topic. Your full understanding is the only way you will know if you conducted and interpreted an analysis correctly.

This text is slightly different from some of the other SPSS textbooks you may have used previously. I believe this is true for several reasons. First, the database management portion of this text should provide you with the tools required to guarantee that the data going into the analysis is set up correctly (roses in!) allowing you to be 100% certain that your test statistics are as accurate as possible. Second, I believe the discussion of each statistical test is thorough enough to allow you to complete each analysis with confidence. For every statistical technique, I provide you with:

1. A brief description of the technique
2. A small example study, with associated data allowing you to replicate my outputs
3. Thorough descriptions of all assumptions and how to evaluate them
4. Point-and-click instructions so you can conduct each technique immediately
5. Thorough descriptions of all dialog boxes you will encounter
6. Screenshots of all relevant output
7. Detailed explanations of how to properly interpret the results
8. An example results section providing you with relevant information minimally required to write up your results for presentation or publication

Two final points. First, I always appreciate feedback regarding this text—positive and negative. My goal is to make the best, most useful text possible. So, if (or should I say, when) you find any errors or if you do not agree with any points I make, please feel free to contact me directly to discuss. Second, this text was created using IBM-SPSS Statistics version 18.0.3. However, the SPSS interface has not changed much over the previous 10+ versions or so. With that said, you should find this text incredibly useful, regardless of version you are using!

Best of luck with your research pursuits,

Ryan W. Walters
9/7/11
Section I
Database Management

The ability to properly manage an SPSS database is an incredibly important skill to acquire primarily because you need to know your data are set up properly prior to conducting any statistical analysis. SPSS loves to produce outputs—it is what it was created to do—even if the numbers produced are completely meaningless and incorrect. That is, the program itself has no idea whether the numbers are right or wrong. Thus, it is critically important that you, the researcher, determine whether the data are correct prior to analysis. This ensures that all test statistics, and associated inference, are correct.

In this section, I cover data entry (both by hand and importing data from another source), merging SPSS datasets, restructuring the dataset as well as recoding variables and working with dates. Within each step, the **bolded** items are what I want you to click (or press) to use each procedure at an absolute minimal level. With that said, for each procedure, I provide in-depth detail of most of the dialog boxes allowing you to find and use all options so you can tailor each procedure to meet your specific needs. With that said, let’s get started…
Chapter 1
CREATING AND NAMING VARIABLES

In SPSS, data entry occurs in a very similar fashion as other table/spreadsheet based programs (e.g., Microsoft Excel). When you open SPSS a new SPSS dataset, you will see the Data View spreadsheet initially. All variables (both dependent and independent) will be designated their own column within this spreadsheet (note: a new SPSS spreadsheet initially has every column labeled var, which is short for variable). In this section, I will discuss the steps involved in creating and naming variables.

1. In the lower left hand corner, click the tab labeled Variable View. The spreadsheet setup with Variable View is different from Data View, and you can see the column names have changed. In Variable View, your variables will be listed under the Name column, and each row will signify a separate/individual variable. The columns signify the various attributes you apply to individual variables.

2. Click on the cell in row 1 under the Name column. Type the word ID and press Enter on your keyboard. A subject or ID variable should always be created as a unique identifier regardless of the analysis being conducted. This variable prevents you from losing track of individuals if you are required to copy data from another program or if you sort any one of your other variables. Notice that most of the columns for row 1 are now filled. Type Sex in row 2 of the Name column and press Enter. Your spreadsheet should look like Figure 1.1.

3. Here is a brief description of the contents within each column:
   a. Name – This column contains your variable names. Note that SPSS does not allow spaces or special characters (e.g., ! # $) within this column, except for an underscore (e.g., male_female) or a commercial-at sign (e.g., @male).
   b. Type – This column specifies the type of data you will be entering. The default value in this column is Numeric, stating that SPSS will only recognize numeric values for this variable in Data View. Within this column, you can click the blue ellipse box ( ). Clicking this will bring up the Variable Type dialog box. Here, you are able to specify what type of data will be entered for the Subject variable. You can customize characteristics based on the variable type.
c. **Width** – This column specifies the maximum number of digits you can enter in each cell while in *Data View*, which can range from 1 to 40 digits, letters, or symbols; depending on the type of data that is.

d. **Decimals** – This column specifies the number of decimal places for your *Numeric, Comma, Dot*, and *Scientific notation* data ranging from 0 to 16.

e. **Label** – This column allows you to type a description of each variable.

f. **Values** – This column is used if you need to code your categorical data. That is, tell SPSS what numeric values represent (e.g., 1 = Male; 0 = Female). Click the blue ellipsis ( ) for the *Sex* variable to bring up the *Value Labels* dialog box. The box to the right of *Value:* requires a numeric value, and the box to the right of *Label:* is where you provide a definition for each numeric value. Type 1 in *Value:* and type *Male* in *Label:*; then click *Add*. You have just signified that the numeric value of 1 within this variable signifies Male. Click *OK.*

g. **Missing** – This column signifies defined missing values. Clicking the blue ellipsis ( ) brings up the *Missing Values* dialog box. The default option is to have *No missing values*. However, you can indicate up to three individual missing values within the *Discrete missing values* box or select a *Range plus one option discrete missing value*.

h. **Columns** – This column signifies the column width for each variable presented in *Data View* individually.

i. **Align** – Specifies the alignment of your data within the cells in *Data View* for each variable. Clicking the down arrow ( ) within this column allows you to change the column alignment.

j. **Measure** – This column specifies your measurement level. Clicking the down arrow ( ) in this column allows you to change between *Nominal, Scale* (i.e., continuous or interval-level), or *Ordinal* measurement levels.

k. **Role** – This column allows you to identify variables with predefined roles (e.g., IVs, DVs, both IV and DV, no role, etc) so they can be easily identified for various analyses.

4. Continue to enter the names for all your variables, changing the necessary attributes and entering value labels along the way.

5. When you have finished entering all variable names, click on *Data View* in the lower left hand corner of the screen.

6. You should now see each of your variables listed as an individual column across the top of the *Data View* spreadsheet.

In conclusion, you are always allowed to add, delete, and move variables within *Variable View* and *Data View*. In addition, you do not need to label your variables if you so choose; however, I advise against this. If you choose not to label your variables, and just enter data directly into the spreadsheet in *Data View*, SPSS will provide a generic variable name (e.g., VAR00001, VAR00002, and so on).
Chapter 2
ENTERING DATA

When you have all of your variables created and named, you are ready to enter your data. As mentioned previously, the SPSS spreadsheet works in an identical fashion to other spreadsheet programs. To enter data:

1. In Variable View, label the first variable ID and the second variable Data.
2. Click the Data View button in the lower left hand corner.
3. Each column contains information pertaining to an individual variable. You see the first column is labeled Subject and the second column labeled Data. Note that in most cases each row signifies separate participants. Each individual cell in any column will contain one piece of data.
4. The cell you are currently going to enter data into will be highlighted in yellow with a thicker border around the cell.
5. Click and highlight the first cell under the ID column (that is, row 1, column 1).
   a. Type the number 1 and press Enter on your keyboard. The cell for row 2 should now be highlighted. You can now type 2, press Enter, type 3, press Enter, and so on. The procedure for entering data into the Data column, or any other column for that matter, is the same as the ID column.
6. Click the cell in the first row of the third (var) column. Enter any number into this cell and press Enter on your keyboard. Notice that SPSS produced a generic variable name as described in Chapter 1. Remember, while you are not required to name your variables, providing variable names are strongly recommended.

In conclusion, your spreadsheet should look like Figure 2.1. Of course, your numbers will be different. Overall, remember that each variable has its own column, and that in most cases individual rows signify separate individuals.

![Figure 2.1](image)
Chapter 3
IMPORTING DATA FROM VARIOUS SOURCES

There are a number of reasons why knowing how to import data files from several of the most widely used spreadsheet or database sources is an important skill. Often times you may be asked to analyze data that you have not entered yourself, which can be an ominous task especially if care was not taken at the data entry stage. To avoid having to re-enter the data you have been provided by hand, SPSS offers the option of importing the data directly from the source program, which inevitably makes the task a whole lot easier.

Although SPSS can import data from numerous sources, an exhaustive list is beyond the scope of this Chapter. Instead, I will walk you through the steps required to import data from three of the most common programs/formats: Microsoft Excel, Microsoft Access, and text files.

Microsoft Excel

As mentioned in previous Chapters, Data View in SPSS is extremely similar to Microsoft Excel. When importing from Excel you have two options. First, you can create and name each variable in Variable View and then highlight and copy the data in Excel and paste it directly into Data View in SPSS. This sounds reasonable, but I advise against this, as mistakes are very easily committed and there is an easier more reliable method by using SPSS to import the data. To import your data from Microsoft Excel:

1. First, you’ll want to check and make sure the variable names in your Excel file are in the first (top) row and that they do not contain any excessively long names, spaces, or special characters (see Figure 3.1). Similar to entering variables in Variable View, SPSS will not import variable names with spaces or special characters (e.g., “# of Pills” will not import but “NumPills” will). If you need to change variable names in Excel, you may find it easier to save the Excel file under a different name so you still have a copy of the original. When all variables are of acceptable length and style, you can save and close the Excel file.

2. Open a new SPSS database. Click on File, choose Open Database, and then click New Query.... This will bring up the Database Wizard dialog box.
3. Under **ODBC Data Sources**, click **Excel Files**, and then click **Next >**. This will bring up the **ODBC Driver Login** dialog box.

4. Click **Browse…** which brings up the **Open** dialog box. Here is where you will find the location of the Excel file you want to import. When you find the Excel file location, select it and then click **Open**. Finally, in the **ODBC Driver Login** dialog box click **OK**.

5. In the **Database Wizard** dialog box, you are asked to **Select Data**. In the **Available Tables**: box you will see a list of all sheets in your Excel file. Select the Excel sheet that contains your data. Typically the data you want to import is in the first sheet (**Sheet1$**), but if it is in a different sheet, change the sheet number appropriately.

   a. If you want to import all data in the Excel file, simply click and hold **Sheet1$** and drag it into the **Retrieve Fields in this Order**: box and unclick. All variables in the Excel sheet 1 should be highlighted and included as in Figure 3.2.

   b. If you do not want to import all variables, click the plus sign (⁺) next to the sheet your data is in. This will display all the variables in this sheet. Click the individual variable you want to import and then click the right arrow (→) to select the variable to be imported.

6. When all the variables you want to import are in the **Retrieve Fields in this Order**: box, click **Next >**.

7. If you are importing data from multiple Excel sheets containing some (or all) of the same variables you are next asked to **Specify Relationships**. For example, if you have the data for the intervention group in **Sheet1$** and data for the placebo group in **Sheet2$** you can connect identical variables (e.g., ID) so that they are imported as the same variable (i.e., one column) in SPSS. When you are satisfied, click **Next >**.

---

![Figure 3.2](image-url)
8. Next, you are asked to **Limit Retrieved Cases**. Here, you can import data for individuals who match a specific set of criteria or you can select a random sample of individuals. This is useful when your Excel database has more data than what you need. If you want to import all cases, simply click **Next >**.

9. Now you are asked to **Define Variables** and all variables to be imported are listed. Here, you have the option to rename any variable by double clicking the specific variable listed under the **Result Variable Name** column. You can also recode string (i.e., alphanumeric) variables to numeric by clicking the box within the **Recode to Numeric** column. You can also click **Minimize string widths based on observed values** to limit the column width in SPSS, which is desired if you are importing questionnaire responses to open-ended questions typically containing lengthy responses. When you have the variables named and in the format you want, click **Next >**.

10. Finally, you are presented with your **Results**. The highlighted portion in blue provides all the names of the variables to be imported. You have two options on how to proceed from here. You can use the default option and retrieve the data immediately or you can paste the instructions into the syntax editor. Because this is an introductory manual, you will choose **Retrieve the data I have selected** and then click **Finish**.

11. That’s it! The SPSS **Data View** should pop up containing the data you requested to be imported from Excel similar to Figure 3.3. You can see that the data in Figure 3.3 matches the data presented in Excel in Figure 3.1. Finally, you may want to switch into **Variable View** to format some of the variables (e.g., remove or add decimal places, choose a different date style, adjust column widths, etc).

![Figure 3.3](image_url)
Microsoft Access

Microsoft Access is able to handle large databases more easily and efficiently than Microsoft Excel; thus, if you work with large databases you may be required to import data from Access frequently. Access creates relational tables and is an easy way to have multiple users import data without overwriting. Further, it is used frequently for survey designs allowing easy storage of questionnaire responses. The importing process is identical to Excel. The primary difference will be in Step 6 below where you will frequently have data importing from multiple tables and so you will be asked to specify relationships. To import data from Microsoft Access:

1. Open a new SPSS database. Click on **File**, choose **Open Database**, and then click **New Query…**. This will bring up the **Database Wizard** dialog box.

2. Under **ODBC Data Sources**, click **MS Access Database**, and then click **Next >**. This will bring up the **ODBC Driver Login** dialog box.

3. Click **Browse…** which brings up the **Open** dialog box. Here is where you will find the location of the Access database containing the data you want to import. When you find the database location, click it, and then click **Open**. Finally, in the **ODBC Driver Login** dialog box click **OK**.

4. In the **Database Wizard** dialog box, you are asked to **Select Data**. In the **Available Tables** box you will see a list of all relational tables within the Access file. Select the table that contains your data
   a. If you want to import all data in the table, simply click and hold the specific table and drag it into the **Retrieve Fields in this Order:** box and unclick. All variables in this table should be highlighted and included and should look like Figure 3.2.
   b. If you do not want to import all variables, click the plus sign (+)

5. When all variables to import are in the **Retrieve Fields in this Order:** box, click **Next >**.

6. Next, you are asked to **Specify Relationships**. As stated above, this is used frequently with Access because most Access databases are relational in nature. That is, all data points for one specific individual may be found in multiple tables (e.g., one table for demographic information, another for medical history, another for comorbid disorders, etc). This step allows you to connect variables to unique individuals across the relational tables. If the variable names are consistent across tables, simply making sure **Auto Join Tables** is checked will connect all variables spelled identically. If the spelling of variable names are not consistent, you will have to connect the variables by selecting the variable in both tables, selecting the appropriate **Inner Join Type**, and then clicking **Join**. When all variables are connected, click **Next >**.

7. Next, you are asked to **Limit Retrieved Cases**. Here you can import data for all individuals, individuals who match a specific set of criteria, or you can select a random sample of individuals. This is especially useful when importing from Access because databases often have more data than what you need. When you are satisfied with your selections, click **Next >**.
8. Now you are asked to Define Variables and all variables to be imported are listed. Here, you have the option to rename any variable by double clicking the specific variable listed under the Result Variable Name column. You can also recode string (i.e., alphanumeric) variables to numeric by clicking the box under the Recode to Numeric column. You can also click Minimize string widths based on observed values to limit the column width in SPSS, which is desired if you are importing questionnaire responses to open-ended questions typically containing lengthy responses. When you have the variables named and in the format you want, click Next >.

9. Finally, you are presented with your Results. The highlighted portion in blue provides all the names of the variables to be imported. You have two options on how to proceed from here: you can use the default option and retrieve the data immediately or you can paste the instructions into the syntax editor. Because this is an introductory manual, you will choose Retrieve the data I have selected and click Finish. That’s it!

**Delimited Files**

Delimited files present your data as text-only with individual data points separated by some delimiter (e.g., comma, tab, space, pipe, etc). Whether created manually or exported from another program, delimited files are common and I guarantee that you will come across this type of file at some point during your research career. Common file extensions for these types of files include .txt, .csv, .dat, etc. The default program commonly used to open and edit delimited files is Notepad, but these files can be opened in Microsoft Word and Wordpad among others. SPSS can import any delimited file regardless of the delimiter; however, the process is different than Excel or Access.

An example of what a comma delimited file looks like is provided in Figure 3.4. Although it looks relatively unstructured, you can see the variable names are provided in the top row and data for each individual is provided in the subsequent rows. Also, note that each variable name and data point is separated (i.e., delimited) by a comma and that the data is identical to that presented in Figure 3.3.

![Figure 3.4](image-url)
To import data into SPSS from a delimited text file:

1. Open a new SPSS database window.
2. Click **File**, choose **Open**, and click **Data** to bring up the *Open Data* dialog box.
3. Click the **Files of type**: drop down menu and select **All Files (*.*)**.
4. Next, find the location of the delimited text file you want to import. Click it and then click **Open** to bring up the *Text Import Wizard* dialog box.
5. The first question you will see is *Does your text file match a predefined format?* Unless you import text files often, your answer will probably be **No**. Further, within this dialog box you will see what your text file looks like as you would have viewed it in Notepad. Feel free to scroll up and down, left and right if you want. When you are convinced the data you want to import is what you requested, click **Next >**.
6. Next, you are asked *How are your variables arranged?* Typically, the default response is **Delimited**, but occasionally it may be with fixed width columns. You are also asked if your variable names are listed at the top of your file. If they are, click **Yes**, if they aren’t click **No** (don’t worry if your answer is **No**, you can add in the variables names once the data are imported into SPSS). Click **Next >**.
7. Now you are asked *The first case of data begins on which line number?* The default value is **2**. This can be slightly confusing, but if the data in your original text file had variable names in row 1 and the data began in row 2, then the default value of **2** is correct (even though the data in the *Data preview* box says the data begins on line 1). In addition, you are asked *How are your cases represented?* Typically, the default option of **Each line represents a case** is the correct option. You are also asked *How many cases do you want to import?* You have the option to import all cases, a number of the first cases, or a random sample of cases. When you are satisfied with your choices, click **Next >**.
8. Now you are asked to identify the delimiter used in your text file (e.g., comma, tab, etc). SPSS does a pretty good job of identifying this for you and will usually have it check marked (from the example in Figure 3.4, **Comma** is checked). Further, you are asked if there is a text qualifier. Characters designated as text qualifiers are used to tell SPSS that you want all information in the quotes to be considered as one data point (e.g., variable 1 without qualifiers would be two data points, variable and 1; however, “variable 1”, with quotation marks as the text qualifier, would be one data point). If there are no text qualifiers in your file, select **None** (the default option). Check the *Data Preview* box to make sure your data appears correct. If you are satisfied, click **Next >**.
9. Next, you are able to rename variables and specify their data format (e.g., numeric, string, etc). To select a different variable to edit, within the *Data preview* box click on the column of the variable for which you want and make changes. When all variables are named and specified correctly, click **Next >**.
10. Finally, you can save this file format for future use. This allows you to choose **Yes** in the future to the first question in step 5 above. You can also paste your import instructions to SPSS syntax for further editing. When you are finished, click **Finish**. If you chose not to paste to syntax, your imported text file should appear in SPSS in *Data View*. 
Chapter 4
EXPORTING YOUR DATA & OUTPUT

Exporting your data and output into a more ubiquitous format/program is extremely common and an important skill to learn because not everyone has access to SPSS (or the same version of SPSS) and some individuals are more comfortable working in other formats. With that said, and as you probably guessed, SPSS is capable of exporting your data and output into numerous formats or programs.

Exporting Data to a New Database

Sometimes you may be asked to provide your raw data to someone who does not have access to SPSS, or you might have additional data in SPSS you want to append to data existing in another database program. There are three ways to export your data to another database program or format. The first involves simply highlighting and copying the data in SPSS and pasting it into a blank Excel or Access spreadsheet. This method should be considered only when you have a fairly small SPSS dataset. I believe you will find that the larger your dataset, the greater probability you will have of making errors during the copy and paste process. The second method involves saving the dataset in another format. This is the simplest method if you want to create a new database in another format. You are allowed to save the database in numerous alternate formats, with the most popular including: Excel (1997-2007), SAS (v. 6-9), STATA (v. 4-8), tab or comma delimited. To save your file in any of these formats:

1. Click File and then click Save as…
2. Type the name of your file in the File name: box.
3. Click the down arrow ( ) in the Save as type: box to open the drop down menu and select the file format you want to save your data.
4. If you do not want to save all variables click Variables... to bring up the Save Data As: Variables dialog box. Initially, all variables are chosen to be saved indicated by a checkmark in the Keep column. To remove variables you do not want to save simply uncheck the box in the Keep column. When you are satisfied, click Continue.
5. You’ll also notice that you are able to Write variable names to spreadsheet, Save value labels here defined instead of data labels, and Save value labels into a .sas file.
   a. Write variable names to spreadsheet provides you with the variable names in the first row of you newly saved spreadsheet. Note that this option is only available for programs using a spreadsheet format (e.g., Excel, delimited, etc). This box is checked by default.
   b. Save value labels here defined instead of data labels saves the value label instead of the data label. For example, say for your gender variable you coded males = 1 and females = 0. If this box is checked, the spreadsheet will contain “male” and “female” instead of 1s and 0s.
   c. Save value labels into a .sas file is as simple as the name implies and is only available for SAS or formats that SAS can read.
That’s it! You can either click **Paste** to paste instructions to SPSS syntax or click **Save** to save your data.

**Exporting Data to an Existing Database**

The third method is used to export your data into an already existing database. SPSS is very flexible in allowing you to control how you want to export your data. To export your data to an already existing database:

1. Open and save a new database. This will typically be a Microsoft Access database.
2. Close the newly created database.
3. In SPSS, click **File** and then click **Export to Database**…. This will bring up the **Export to Database Wizard**.
4. Under **ODBC Data Sources**, select the file format your already existing database is in and click **Next >**, which brings up the **ODBC Driver Login** dialog box.
5. Click **Browse**… to bring up the **Open** dialog box. Here, you need to find the location of the database to where you want to export your data. When you find the database, click it, and then click **Open**. Finally, in the **ODBC Driver Login** dialog box click **OK**.
6. Next, you are asked to **Choose how to export the data**. At times you have up to five options including:
   a. **Replace values in existing fields** – You will need to know which variables are primary, meaning that you need to know how to connect variables between the two databases.
   b. **Add new fields to an existing table** – You can only use this option if you have variables in SPSS that are different from those that already exist in your database. Again, you will need to know how to connect variables between databases.
   c. **Append new records to an existing table** – Here, you can add data for subjects that do not already exist in your existing database.
   d. **Drop an existing table and create a new table of the same name** – Here, you can overwrite the existing database.
   e. **Create a new table** – Here, you can simply create a new table to be added to your existing database.

For the purposes of this manual, I will walk you through the steps required to add data to a newly created blank database. At this point, SPSS has assessed what data and tables are already in your existing database. Because this database is blank, the only option available to you is to **Create a new table**.

7. With **Create a new table** already selected, you now have to **Name**: your new table. This name can be whatever you like and can contain special characters if required (as permitted by the program you’re exporting to). When you have named your table click **Next >**, which brings up the **Select variables to store in new table** dialog box in Figure 4.1.
In the PASW Statistics: box you will see all of the variables listed in your existing SPSS database. To include a variable in the new database’s table, click the variable name, and then click the right arrow ( ) next to the associated column within the Table: box. This variable should now be listed in the Variable to save column. When a variable is moved successfully into the Table: box additional information is presented, including the Field name, Type, and Width.

i. The Field name will be the name of the column heading in your new database. This can be changed to any name you wish with or without special characters depending on the program you are exporting to.

ii. The Type column describes whether the variable is numeric, alphanumeric, date, etc. If you highlight the cell and click the down arrow ( ) SPSS will provide you with a number of options, and although there are more options presented here they are very similar to the Type column in Variable View discussed in Chapter 1. In general, however, SPSS will correctly identify the variable Type.

iii. For alphanumeric variables, the Width column contains the number of characters the variable is allowed to have in the new database.

iv. Finally, you will notice the Primary Key ( ) column used to identify unique observations. That is, it is required that any variable selected as the primary key have only unique observations with unique values. If values are not unique an error message will be displayed.

b. Next, you are asked how you want to handle User-Missing Values. Here, you can either Export as valid values or Export numeric missing values as nulls and string missing values as blank spaces. The only difference between these two options is how missing values will appear in your existing database. When you are satisfied with your choices, click Next >.
8. Finally, you are provided with a *Summary* of your specifications. Check to make sure the *Summary* matches what you actually want SPSS to export. You are also asked whether you want to *Export the data* or *Paste the syntax* for further editing. When you have made your choice, click **Finish**.

Regardless of the method you use to export your data, make sure to open your new or existing database to determine whether the data exported correctly and in the format you requested. I will warn you that the programmers at SPSS have identified glitches in this process and some occur more frequently than others. However, if the data has not exported how you want, check to make sure you entered the steps correctly. When you have eliminated user error, then you know the problem came from the software, not you.

**Exporting Output**

When conducting statistical analysis in SPSS, your results are printed into a separate *Output* window. You need SPSS to open the output, and eventually you will be asked to provide your results in different software program because someone does not have SPSS. This process is relatively straightforward, and you have several options: (1) copy and paste the desired tables or results from the SPSS and paste them directly into the other software; (2) re-create the tables or results in another program; (3) export the output through SPSS. Overall, no choice is best; you need to decide which option works best for you.

If you do not have a large number of descriptive tables or statistical test results, the copy and paste method is straightforward and quick. Often times, though, the tables produced by SPSS do not meet publication requirements and you are left to re-create the tables in another program. This may be cumbersome and time-intensive, but is often time the only option that will afford you the flexibility needed to edit your table(s) to meet certain requirements. However, if you simply want to export the output so others can view it (because they either do not have SPSS or do not have the same version of SPSS), there is a quick and easy method of doing so.

In Figure 4.2, you can see an example of an output window. You can hide any executed task (e.g., *Frequencies* or *T-Test* in Figure 4.2) within your output that is currently visible by clicking the minus sign ( ), e.g., *T-Test* is hidden). You can make visible any executed task within your output that is hidden by clicking the plus sign ( ). Further, within each task, you can hide/make visible other aspects of your output (e.g., Title, Notes, Log, etc) by double-clicking their respective icons.

For example, you can see in Figure 4.2 that the *Log* for (i.e., above) both the *Frequencies* and *T-Test* headings are hidden, indicated by closed book icon ( ), while the *Title* and *Statistics* portions of the *Frequencies* are visible, indicated by open book icons ( and ). As you will see in the steps below, the portions you make visible or hidden are important to consider prior to exporting your output. To export your output to another program/format:

1. In your Output window, click **File**, then click **Export**... This will bring up the *Export Output* dialog box.

2. First, you are asked to identify the *Objects to Export*. You have the option to export *All* available output (regardless of what is visible or hidden), you can export *All visible* portions of your output, or only the *Selected* portions of your output (in Figure 4.2, *Statistics* and the *t-test*’s *Log* are selected, indicated by the yellow highlighting. So, you can select hidden portions as well).
3. Next, you are asked what Type: of Document you want to export your output to. Here you have a number of options. Click the down arrow (↓) to see the drop-down menu containing available options.

4. The Options: table contains information pertaining to your exported output document. You can click Change Options… to bring up the Export Output: document options dialog box. Note that the available options will vary depending on the type of document you are exporting to. When you are satisfied with your options, click Continue.

5. Next, you are asked to identify the File Name: and location for your exported output. Click Browse… to identify the location and name of your exported output. When you are satisfied, click Save.

6. Finally, you have the choice of selecting the file Type: you want to save your graphics as. You also have the ability to Change Options… for each file type. Note that this option is only available (1) if you have graphics (e.g., scatterplot, histogram, etc) and (2) if you chose your Document Type: to be either HTML, Text, or Graphics Only. In Figure 4.2, there are no graphics.

7. That’s it! When you are satisfied with your selections, click OK.
Chapter 5
MERGING FILES

There are a number of reasons why you need to know how to merge multiple files together to form one working dataset. Large databases often have their data partitioned into individual segments (e.g., demographics, clinical, laboratory, etc). Further, if you are collaborating with other person(s) or conducting a multi-site trial, you will have portions of study data that each contributor entered. I could list several more reasons, but you get the point: in both situations, datasets will need to be merged. There are two different reasons to merge files in SPSS. The first is to add additional cases or subjects to an existing dataset. The second is to add additional variables.

In either case, you can add variables and cases; however, it is relatively easy for you to end up with mismatched data. For example, say you want to add cases to Dataset 1. This dataset has cases and a set number of variables already entered. If the cases in Dataset 2 have only identical variables, then merging will simply append the cases in Dataset 2 to the end of Dataset 1. However, if the cases in Dataset 2 have different or additional variables, if you are not careful, all data will not be merged. This situation will create complete data only for variables that were identical in Datasets 1 and 2 (i.e., the cases in Dataset 1 will have missing responses for the additional variables added from Dataset 2). Bottom line, merging files correctly is an important step prior to analysis and this is why you must pay careful attention to detail when merging files.

Finally, all datasets to be merged must be in SPSS format. That is, you cannot merge an Excel and SPSS file together. Prior to merging datasets, you are required to import your data into SPSS (see Chapter 3 for how to import data into SPSS).

**Merge: Add Cases**

1. Make sure all data is imported into an SPSS dataset. Note that if you want to merge more than two files, you can only merge two files together at a time, so you will need to proceed in steps.

2. Open the dataset you want to merge the cases into (this is your active dataset) and click Data, then choose Merge Files, and finally click Add Cases…. This brings up the Add Cases to dialog box.

3. If you already have the dataset you want to merge opened, select it in the An open dataset box. If the dataset you want to merge in is not already opened click Browse... to bring up the Add Cases: Read File dialog box. Here, you will locate the SPSS dataset you want to merge in and click Open (this is your external dataset). Now, on the Add Cases to dialog box, click Continue.

4. This brings you to the Add Cases From dialog box.
   a. If any additional or misspelled variables are included in the external dataset they will be identified in the Unpaired Variables: box. If any variables are misspelled, you can click Rename... and enter a New Name: for this variable and the click Continue. If you do not wish to rename variables, you can pair two misspelled variables by selecting them (select one, press Ctrl on your keyboard, and select the other) in the Unpaired Variables: box and then click Pair.
b. If all variables in both datasets are identical the *Unpaired Variables* box will be empty and all variables will be included in the *Variables in New Active Dataset* box.

c. You always have to option of not merging a variable(s). If this is the case, simply select the variable(s) in the *Variables in New Active Dataset* box and either drag them into the *Unpaired Variables* box or click the left arrow ( ).

5. Finally, you can check *Indicate case source as variable* checkbox. Checking this box will create a new binary variable in your merged active dataset that indicates whether the case is from the original (active; 1) dataset or from the merged (external; 0) dataset. The default name SPSS provides for this new variable is *source01*; however, you can rename this variable (note that no special characters can be used in variable’s name).

6. When you are satisfied with your selections, click OK.

**Merge: Add Variables**

Prior to following the instructions below, you are going to want to make sure that your cases have a unique identifier variable(s). A unique identifier variable allows you to identify the data for an individual case across multiple datasets. The method you use to identify cases is arbitrary; although, you will want this identifier to be anonymous and private (so, no social security numbers). Finally, in order for the merge to occur, the unique identifier variable(s) you use must be sorted in ascending order. I want to reiterate this statement, the identifier variables must be in ascending order! I go into more detail in Step 5 below. To add variables:

1. Make sure all data is imported into SPSS datasets. Again, note that if you are merging more than two files, you can only merge two files together at a time.

2. Open the dataset you want to merge the variables into (this is your active dataset) and click *Data*, then choose *Merge Files*, and finally click *Add Variables*…. This brings up the *Add Variables to* dialog box.

3. If you already have the dataset you want to merge opened, select it in the *An open dataset* box. If the dataset you want to merge in is not already opened click *Browse*... to bring up the *Add Variables: Read File* dialog box. Here, you will locate the SPSS dataset you want to merge in and click *Open* (this is your external dataset). Now, on the *Add Variables to* dialog box, click *Continue*.

4. This brings you to the *Add Variables From* dialog box, which is more complex than the *Add Cases From* dialog box described above (see Figure 5.1).

   a. You can see in the *Excluded Variables* box that the *Subject* variable from the external dataset is excluded (note: you know the variable is from the external dataset by the + sign after the variable name). This variable is excluded because it is redundant (you can see the *Subject* variable from the active dataset in the *New Active Dataset* box).

   b. When merging two files, you will almost always need to have a unique identifier to allow you to match up individuals in one dataset to the other. Without this identifier, SPSS will merge the datasets together in no direct order; thus, you will
be unable to identify the case that the data belongs to in each row. To avoid this situation, click **Match cases on key variables in sorted files**. This will allow you to indicate whether both cases provide cases, or whether the active or external dataset provides the cases. In most situations, you will click **Both files provide cases**. Regardless of the choice you make here, you are going to need to identify **Key Variables**: Key variables are those that you will use as your unique identifier(s). There are numerous occasions when you will require multiple key variables (e.g., Name, DOB, Race, etc). In the situation in Figure 5.1, the Subject variable is the unique identifier, so you would click the Subject(+) variable in the **Excluded Variables** box, then click the left arrow ( ) to move this variable into the **Key Variables**: box. Notice that the unique identifier(s) were removed from the **New Active Dataset**: box as well.

c. Finally, you have the option of indicating the dataset source for each case, similar to above. Again, checking this box will create a new binary variable in your merged active dataset that indicates whether the case is from the original (active; 1) dataset or from the merged (external; 0) dataset. Further, you are allowed to rename this variable with no spaces or special characters.

5. When you are satisfied with your choices, click **OK**. A warning message will always pop up telling you that the **Keyed match will fail if data are not sorted in ascending order of Key Variables**. So, for the example in Figure 5.1, the Subject variable, in both the active and external datasets, must be in sorted ascending order. If they are not, the merge will fail. Do not worry if you click **OK** and the key variables are not in ascending order though, because although the merge will not take place, you will not lose any information.

6. That’s it! Check to make sure the datasets merged as you specified.
Chapter 6
IDENTIFYING DUPLICATE CASES

All between-subjects statistical tests assume cases are only measured on a variable one time. This is one key element that ensures the assumption of independence (discussed later). Further, you may be given a dataset in the stacked format but the analysis is only concerned with an endpoint analysis evaluating a case’s last measurement during the study period. Alternatively, there may also be instances when you are handed a dataset and asked to conduct an analysis with the assumption that each row only contains all the data for an individual case. If this dataset is large, you really have no way of knowing whether no duplicates exist without checking this for yourself. Thankfully, SPSS allows you easily to check whether duplicates cases exist in your dataset. To identify duplicate cases:

1. Click Data and then click Identify Duplicate Cases…. This will bring up the Identify Duplicate Cases dialog box similar to Figure 6.1.

2. First, you are asked to Define matching cases by:. Here, you will choose the variable(s) you want to use to identify any duplicate cases. Choose the variable you want to use and then click the right arrow (arrow) next to the Define matching cases by: box. Move all variables you want to use into this box.

3. Next, you are asked to Sort within matching groups by:. As the name implies, this box is where you will enter any variables that you wish to sort. You have several options when using this box:
   a. The most obvious example of this is when you have a time or date variable and you want to make sure the earliest observation is sorted first (or last). When you move a variable into this box, the Ascending and Descending radio buttons will become active. Ascending means the first observation for each case will be at the top of their file, whereas Descending means the last observation for each case will be at the top of their file.
b. In addition, you can sort by several variables based on the order in which they are entered. For example, if Variable A is entered first, followed by Variable B, SPSS will sort Variable B within Variable A for each case. Note that the method of sorting is specific to each individual variable. SPSS identified which method you use by either (A) or (B) next to the variable name. So, you can have Variable A sorted ascending and Variable B sorted descending.

c. Finally, you can change the sort order without removing all variables and starting over by clicking the up arrow (▲) or down arrow (▼).

4. In the Variables to Create box, you are provided with several options.

   a. The default is Indicator of primary cases (1=unique or primary, 0=duplicate). When this is checked, SPSS will create a new variable to identify the duplicate cases. You have the option to have SPSS designate the First case in each group is primary or the Last case in each group is considered primary. If you have chosen to indicate primary cases, you have to choose one of these options.

   b. Also, you will notice the Name: box contains PrimaryLast or PrimaryFirst depending on which observation you selected as primary. This is the name of the new variable SPSS will create in your dataset after you have finalized your selection(s). It can be changed to any variable name you wish, excluding spaces and special characters of course.

   c. You also have the option to Filter by indicator values. This allows you to select the primary observations within each case, while excluding all non-primary observations from any analyses. This does not mean non-primary observations are removed from the dataset, they are simply ignored. You can see non-primary cases that have been filtered in Figure 6.2 as they have a slash through their row number. That is, only the primary observations do not have slashes.

   d. Finally, you have the option to check the box informing SPSS to provide a Sequential count of matching case in each group (0=nonmatching case). Selecting this option will create a new count variable indicating the number of total observations within each case. Similar to option “4b” above, you can rename this count variable, with the default Name: being MatchSequence.

---

Figure 6.2
5. Next, you have the option to *Move matching cases to the top of the file*. This option moves cases with duplicate observations to the top of your dataset, whereas cases with only one observation are placed at the end of the dataset. No participants are deleted, they are simply reordered.

6. Finally, you have the option to *Display frequencies for created variables*. This option will produce frequency counts in the Output window for both the number of primary and duplicate cases as well as the number of within case observations if you chose option “4d” above.

   a. In Figure 6.3, notice that there were 6 *Primary Cases* with 7 *Duplicate Cases*. Thus, there are 6 unique individuals in the example dataset used here.

### Frequency Table

<table>
<thead>
<tr>
<th>Indicator of each first matching case as Primary</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid Duplicate Case</td>
<td>7</td>
<td>53.8</td>
<td>53.8</td>
<td>53.8</td>
</tr>
<tr>
<td>Valid Primary Case</td>
<td>6</td>
<td>46.2</td>
<td>46.2</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>13</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

### Sequential count of matching cases

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 Valid</td>
<td>1</td>
<td>7.7</td>
<td>7.7</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>38.5</td>
<td>46.2</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>30.5</td>
<td>94.6</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>15.4</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>13</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

*Figure 6.3*
Chapter 7
RESTRUCTURING THE DATASET

It may often be necessary to restructure your dataset to meet the assumptions of the statistical test you want to conduct. For example, consider a large dataset containing pharmacy refill or hospital claims data. In most instances, instead of each row containing all the data for an individual participant (i.e., multivariate format), a row will contain only the data for a single refill or hospital visit. The number of rows a participant has in the dataset is determined specifically by the number of refills or visits; thus, the number of rows a participant requires could vary considerably. This format is known as the stacked format, and is presented in Figure 7.1. Notice the number of refills varies across cases as indicated by the ID column.

For most statistical analyses (e.g., t-tests, ANOVA, linear regression), analyzing a data set the stacked format will violate the assumption of independence (you may remember this, but if you do not, no worries, we will get to this in a later Chapter); however, for other analyses and procedures in SPSS (e.g., multi-level modeling), the stacked format is required. In either case, you are required to have a solid understanding of how the data should be formatted for a particular statistical test.

There are two ways to restructure the dataset. First, you can restructure cases into variables. That is, transition from the stacked format to the multivariate format. Second, you can restructure variables into cases, which transitions from the multivariate format to the stacked format. Because we will not be discussing multi-level modeling in this workbook, and because the most common type of restructuring is based on the set up in Figure 6.1, I will only provide you with instructions on how to restructure cases into variables.

Figure 7.1
Restructure Cases into Variables

1. Click **Data** and then click **Restructure**…. This brings up the *Restructure Data Wizard* dialog box.

2. Here, you are asked *What do you want to do?* You have three options.
   a. *Restructure selected variables into cases* – Select this if you want to restructure a multivariate format into a stacked format.
   b. *Restructure selected cases into variables* – Select this if you want to restructure a stacked format into a multivariate format.
   c. *Transpose all data* – Select this if you want your cases to become variables or your variables to become cases. The most typical example of this is when data is outputted from survey software, where each row is a variable.

   Choose **Restructure selected cases into variables**, then click **Next >**.

3. Now, you are presented with the *Cases to Variables: Select Variables* dialog box. Within this dialog box, you are presented with two boxes to place variables.
   a. In the **Identifier Variable(s):** box, you will place the variable(s) that is the unique identifier for each individual participant. From Figure 7.1, this variable is labeled *ID*. To move a variable into this box, click the variable name in the *Variables in the Current File:* box and click the left arrow (←) next to the **Identifier Variable(s):** box.
   b. In the **Index Variable(s):** box, you will place variables that the program will use to label the new columns. The index variable is usually a string (i.e., alphanumeric) variable. No variables serve as index variables in Figure 7.1. To move a variable into this box, click the variable name in the *Variables in the Current File:* box and click the left arrow (←) next to the **Index Variable(s):** box.

   When you are satisfied with your variable choices, click **Next >**.

4. Next, you will be presented with the *Cases to Variables: Sorting Data* dialog box where you are asked whether you want to *Sort the current data?* For the data to be restructured, SPSS requires the data for each individual case be in adjacent rows. Reconsider Figure 7.1. Notice all the data is for each case are in adjacent rows. That is, all the data for participant 1 (i.e., *ID = 1*) precedes all the data for participant 2, and so on. SPSS would fail to restructure the dataset if some data from participant 1 was presented, then some data for participant 2, and then the rest of the data from participant 1. Now, this does not mean that you cannot have all participant 2 data before all the data for participant 1. If you are at all in doubt about whether your data is sorted correctly, select the default *Yes - data will be sorted by the Identifier and Index variables*. If you are certain your data is sorted correctly, you may select *No - use the data as correctly sorted*. When you are satisfied with your selection, click **Next >**.

5. Next, you will be presented with the *Cases to Variables: Options* dialog box. Here, you have several options:
a. In the Order of New Variable Groups box you can either Group by original variable or Group by index. The former groups the new variables together. So, from Figure 7.1, the restructured dataset would present consecutive Medication variables, followed by consecutive QTY variables, and so on. The latter will group the variables according to their order in the original dataset. You are required to choose one or the other, but both options will produce the same result, with the only difference being the order in which the new variables are presented in the restructured dataset.

b. The optional Case Count Variable box allows you to create a count variable in the restructured dataset containing the number of rows in the original dataset that were used to create a row in the restructured dataset. If you check the Count the number of cases in the current data used to create a new case box, you will be required to Name: the count variable, and optionally provide a variable Label:

c. Finally, the optional Indicator Variables box allows you to create one new variable for each unique value of the index variable you selected in Step 4. The indicator variable is not a count, but a binary variable depending on whether the case has a unique value for the index variable (1 = Yes; 0 = No). For example, from Figure 7.1, if the Medication variable were your index variable, participant 1 would have a value of 1 for medication D and 0 for medication C, whereas participant 5 would have values of 1 for both.

When you are satisfied with your selection(s), click Next >.

6. Finally, you are presented with the Finish dialog box. Here, you can either Restructure the data now or Paste the syntax generated by the wizard into the syntax window. The former will immediately produce the restructured dataset, the latter will leave the current dataset as is, but open a syntax window and paste the syntax required to restructure the dataset so you can run it in the future. When you have made your choice, click Finish.

7. Clicking Finish should mean you are finished, right? Well, SPSS will pop up a warning message stating that your dataset will be available after restructuring. Simply click OK.

   a. The restructured dataset should appear similar to Figure 7.2. Notice all individual case data is on the same row, but that cases do not necessarily have the same number of occurrences for any of the variables in the original dataset.
Chapter 8
RECODING VARIABLES

Understanding the recoding variables procedure is incredibly important, as most statistical procedures conducted by SPSS require numeric coding for categorical variables. Further, this procedure is useful if someone other than you has entered data into the dataset. For example, gender data was entered as Male and Female as opposed to 0 and 1 in Figure 8.1. This procedure will also come in handy when you need to collapse multiple groups or response categories. For example, consider a 4-point Likert-type response scale ranging from Strongly Disagree to Strongly Agree. Here, you could collapse Strongly Disagree and Disagree into simply one global Disagree category, and do the same for a global Agree category. Additionally, this procedure is required to dummy code variables for a linear regression analysis (see Chapter 38). That is, when you have a categorical variable with $k$ categories, you will need to create $k - 1$ dummy variables to represent category membership.

One primary item to note is that SPSS is case-sensitive. This is especially important to consider when you want to recode a string variable into a numeric variable. Recalling the gender example above, you would need to do a recode statement for every spelling of the word male. That is, if the person entering the data used Male, male, and MALE, you need to recode each spelling individually. SPSS will not recognize Male as male. This is why I recommend running descriptive statistics (see Chapter 11) on variables of interest prior to recoding variables.

SPSS has two methods of recoding variables. The first allows you to recode variables into the same variable. Choosing this option allows you to replace the current values with recoded values. For example, the value of 0 will replace all male entries for a gender variable. The original variable is replaced or copied over. The second option allows you to recode variables into a different variable. This option will create a new variable containing the recoded values, while keeping the original variable intact.

In general, I recommend you always recode into a different variable. This allows you to verify SPSS ran the procedure correctly while maintaining the original variable’s values. However, the choice is yours and I will walk you through both procedures. For both procedures, consider the small dataset presented in Figure 8.1.

![Figure 8.1](image-url)
Recode Into Same Variables

1. Click **Transform** and then click **Recode into Same Variables**.... This brings up the **Recode into Same Variables** dialog box.

2. Here, you will notice all the variables in your dataset are listed in the box on the left. You will also notice that you cannot click **Old and New Values**... until you move the variable to be recoded into the **Variables** box by selecting the variable and then clicking the right arrow ( ). It is noted that if you have several variables that you want to recode simultaneously, SPSS allows you to do so. However, there are several important aspects to consider:

   a. First, all variables to be recoded will be recoded using the exact same **Old and New Values**... That is, if you have two variables that require different coding schemes (e.g., male = 1 for variable 1 and male = 0 for variable 2), you will have to run the procedure separately for both variables.

   b. Second, the variable type must be the same for all variables recoded simultaneously. That is, all variables must be either numeric or string, you cannot have a combination of the two.

   c. Third, when recoding into the same variable, you are allowed to recode a string variable to numeric, but not a numeric variable to a string.

When you have selected the variable(s) you want to recode, click **Old and New Values**.... This brings up the **Recode into Same Variables: Old and New Values** dialog box, shown in Figure 8.2 for recoding a string variable to numeric.

---

**Figure 8.2**

---

26
3. On the left side of this menu, you will notice several options for the *Old Value* you want to recode. You can recode single values, ranges of values, or missing values. I will describe each option for you; however, the number of options available varies by the old variable type being recoded.

   a. The *Value:* box is available for both numeric and string variables. This option allows you to specify one specific value you want to recode into the box just below the radio button.

   b. The *System-missing* option is available only if you are recoding a numeric variable. This is because string variables allow any character, whereas numeric only allows numbers. System-missing values are defined as values in your data that are undefined according to the type of variable you specified. This primarily occurs when values of a string variable are considered numeric. SPSS would then replace any string value with a system-missing value, signified by a period.

   c. The *System- or user-missing* option is available for both numeric and string variables. A user-missing value is a value that is unknown to you. For example, if a participant did not respond to any demographic questions their gender would be unknown. This type of missing data is ubiquitous in studies using retrospective datasets.

   d. The *Range:* option is available only for numeric variables. This option allows you to specify a range of values to be recoded. For example, recoding an Age variable into 10-year segments. Say you want to recode all participants who are 40-49 years old. The lowest value (i.e., 40) would be placed in the box underneath *Range:* and the highest value (i.e., 49) would be placed in the box underneath *through*.

   e. The *Range, LOWEST through value:* option is available only for numeric variables. Here, you will specify the lowest value you want to recode and SPSS will collapse this value and all values greater than the value specified. For example, say you want to recode everyone less than 40 into the same variable. Here, you would type 40 into the box below *Range, LOWEST through value:*

   f. The *Range, value through HIGHEST:* option is available only for numeric variables. Here, you will specify the highest value you want to recode and SPSS will collapse this value and all values lower than the value specified. For example, say you want to recode everyone older than 18 into the same variable. Here, you would type 18 into the box below *Range, value through HIGHEST:*

   g. Finally, the *All other values* option is available for both numeric and string variables. This option is often used when you want to recode any other values not previously specified for recoding. Or, it can be used to create a constant. That is, if you select *All other values* and do not specify any other values to recode.

4. When you have chosen the old value, or range of values, you want to recode, you will next move on to the *New Value* box. This is where you will specify the new recoded *Value:* or tell SPSS you want the old value to be considered *System-missing.* The *Value:* can be either string or numeric.
5. After you have specified both the *Old Value* and *New Value*, you will click **Add**. This will move your selection into the *Old --> New: box*. Here, you will be able to review the recode information you selected. Clicking **Add** will also clear any values you had in the *Old Value* and *New Value* boxes.

   a. If you are unhappy with any choices you made, you can always click your choice in the *Old --> New: box*. This will place your values back into the *Old Value* and *New Value* boxes. These values can be changed as you wish. When you are satisfied with your choices, click **Change** and your new choices will replace your previous choices.

   b. Alternatively, if you recoded a variable incorrectly, you can click it in the *Old --> New: box* and then click **Remove**, which will delete the recode specification.

6. Repeat steps 3-5 until all values you want to recode are entered into the *Old --> New: box*. When you are satisfied with your recoding scheme, click **Continue**.

7. Additionally, you have the option of clicking **If...**. This brings up the *Recode into Same Variables: If Cases* dialog box. This option allows you to recode values for a selected subset of cases using conditional expressions. The default selection is *Include all cases*. However, say you only want to recode a variable for males (coded 1), while leaving females coded as is. In this situation, you would click the *Include if case satisfies condition: radio button*, which opens up all options to you. You would move your *Gender* variable into the *Include if case satisfies condition: box* by clicking the right arrow ( ). Because this is an If statement, you will use the equal sign (=) to designate what you want gender to equal, in this case males (i.e., *Gender = 1*); thus, any recoding will apply only to those coded 1 on the *Gender* variable. If statements follow logic similar to the *Compute* statement discussed in Chapter 13. When you have identified the subset of participants you want to recode, click **Continue**.

8. When you have made all of the necessary choices, you can either click **OK** to run the procedure immediately, or click **Paste** to paste the syntax for further editing.

9. When you run this procedure, you will notice that SPSS has replaced the old values with the new values you specified. As seen in Figure 8.3, you can see that I forgot to include one of the alternate spellings of *MALE*. This is simply a demonstration that all variations of string variables need to be specified individually. As stated above, SPSS will not recognize *male* as *MALE* or *Male* as *MALE*.

---

**Figure 8.3**
10. Finally, when recoding a string variable to numeric, you will have to go into Variable View and change the Type from String to Numeric. SPSS will not automatically do this for you.

**Recode Into Different Variables**

1. Click **Transform** and then click **Recode into Different Variables**… This brings up the *Recode into Different Variables* dialog box, shown in Figure 8.4.

![Figure 8.4](image.png)

2. Initially, you will notice all the variables in your dataset are provided on the left hand side of the dialog box. You will also notice that you cannot click **Old and New Values**… until you move the variable to be recoded into the **Input Variable -> Output Variable:** box by selecting the variable you want to recode and clicking the right arrow ( ). It is noted that if you have several variables that you want to recode simultaneously, SPSS allows you to do so. However, there are two important aspects to consider:

   a. First, all variables to be recoded will be recoded using the exact same **Old and New Values**… That is, if you have two variables that require different coding schemes, you will have to run the procedure separately for both variables.

   b. Second, the variable type must be the same for all variables recoded simultaneously. That is, all variables must be either numeric or string, you cannot have a combination of the two.

3. When you move a variable(s) into the **Input Variable -> Output Variable:** box, you must provide a name for each variable. This is because you are creating a new recoded variable and keeping the original variable intact. In the **Output Variable** box, you will provide the name of the new variable in the **Name:** box. As with all SPSS variable names, no spaces or special characters are allowed. Also, you can optionally provide a variable label in the **Label:** box. This label will be placed in the **Label** column in the **Variable View**. When you are satisfied with your **Name:** and optional **Label:**; click **Change**. You have to provide a new variable name for all variables to be recoded.
4. When you have renamed all variables to be recoded, click **Old and New Values...** to bring up the **Recode into Different Variables: Old and New Values** dialog box. Because much of the description of this dialog box has already been explained above in the discussion of Figure 8.2, I will only discuss a few subtle additions.

   a. First, you will notice in the **New Value** box a new option is available to **Copy old value(s)**. Choosing this option will not recode the value, or range of values, chosen in the **Old Value** box. Instead, the old value is retained.

   b. Second, in the lower left hand corner of the dialog box you will see **Output variables are strings**. This option is specifically chosen when you want to recode a numeric variable to a string variable (this was not available if you wanted to recode into the same variables, as described above). If you choose this option, you can choose the new variable’s **Width**: The default is 8 characters, but can be set up to 255 characters. After recoding, this value will be placed in the **Width** column in the **Variable View**.

   c. Third, in the lower left hand corner, you will also see **Convert numeric strings to numbers** (*5* to 5). This option converts string values containing numbers to numeric values. Note that strings containing numbers and letters will be assigned a system-missing value after recoding. This option is useful if you imported a numeric variable that SPSS read as string.

Follow steps 3-5 from the **Recode Into Same Variables** section above for all variable categories. When you are satisfied with your recoding scheme, click **Continue**.

5. Finally, you have the option of clicking **If...**. This brings up the **Recode into Different Variables: If Cases** dialog box. This option allows you to recode values for a selected subset of cases using conditional expressions. The default selection is **Include all cases**. However, say you only want to recode a variable for males, while leaving females coded as is. In this situation, you would click the **Include if case satisfies condition**: radio button, which opens up all options to you. You would move your **Gender** variable into the **Include if case satisfies condition**: box by clicking the right arrow ( ). Because this is an If statement, you will use the equal sign (=) to designate what you want gender to equal, in this case males (i.e., Gender = 1); thus, any recoding will apply only to those coded 1 on the Gender variable. If statements follow logic similar to the Compute statement discussed in Chapter 13. When you have identified the subset of participants you want to recode, click **Continue**.

6. When you have made all of the necessary choices, you can either click **OK** to run the procedure immediately, or click **Paste** to paste the syntax for further editing. An example of how your recoded values will appear is provided in Figure 8.5. Notice, the original variable is intact and that the new recoded variable has a system missing value for MALE. This is because I used the same recode scheme as shown in Figure 8.2. That is, I forgot to include a recode statement in the **Old --> New**: box within the **Recode into Different Variables: Old and New Values** dialog box for MALE. Again, SPSS will not recognize male as MALE or Male as MALE.
Figure 8.5
Chapter 9
DATE AND TIME WIZARD

The Date and Time Wizard provides an easy method for you to work with date variables. This procedure is important in situations where you need to determine how much time as passed between measurements. Useful, for example, if you are conducting a repeated-measures ANOVA where an assumption is equally spaced measurement points. Further, the Date and Time Wizard can be used to calculate a participant’s age at the beginning of a study. That is, it allows you to use the participant’s date of birth to calculate their age at any specific time point during the study period. In addition, you can also create date variables from string variables, which is useful after you have used SPSS to import data with unformatted date variables.

The primary advantage of this procedure is that it is the only application in SPSS allowing you to calculate with date variables. The most common aspects of this procedure are to add or subtract a specific duration from a date variable or to calculate the number of specific time units between two dates. In addition, you can also combine and extract various portions of a date variable. For the purposes of this Chapter, I will describe how to complete the former two tasks.

Add or Subtract Time Units from a Date

1. Click Transform and then click Date and Time Wizard…. This brings up the Date and Time Wizard dialog box.

2. The first question you are presented with asks What would you like to do? Click Calculate with dates and times. Click Next >. This brings up the Date and Time Wizard – Step 1 of 3 dialog box.

3. Click Add or subtract a duration from a date (e.g., add a month to an age or add a time variable to a date/time variable). Click Next >. This brings up the Date and Time Wizard – Step 2 of 3 dialog box shown in Figure 9.1.

Figure 9.1
4. You will notice all of the variables in your dataset are presented in the Variables: box. In addition, SPSS has created an optional Current date and time [$TIME$] variable that you can use in your calculation. This will always be the first variable listed. If you do not use this variable, SPSS will delete it immediately after you finish the procedure. The other aspects of this dialog box are described below.

   a. The Date: box is where you will place the date variable you are working with.

   b. You have two options when adding or subtracting from the date variable you selected above. First, you can select a Duration Variable:. This is a previously created non-date variable in your dataset, and can be found in the Variables: box. If you want to use this option, select the specific non-date variable you want to use from the Variables: box and click the right arrow (→). As an alternative, you can specify a Duration Constant: which will be added or subtracted from each participant’s date variable.

   i. The primary difference between these two options is that the Duration Variable: can vary in the amount added or subtracted, whereas the Duration Constant: is, as the name implies, a constant.

   c. Next, you must select the Units: of your duration variable or constant. That is, SPSS needs to know whether the variable or constant is in Years, Months, Days, Minutes, etc units. Click the down arrow (↓) to view and select the appropriate units.

   d. Finally, you are required to select the Operation. That is, whether you are using Addition or Subtraction.

When you are satisfied with your selections, click Next > to bring up the Date and Time Wizard – Step 3 of 3 dialog box.

5. You should notice that your Calculation: is presented mathematically near the top of the dialog box. If this calculation is incorrect, simply click < Back to make the necessary changes. Finally, upon completing this procedure, SPSS will create a new variable in your dataset. Thus, you are required to provide a name for the Result Variable: and optionally provide a Variable Label:. Remember, no special characters or spaces are permitted in the new variable’s name. When you are satisfied with your calculation and new variable name, you can either Create the variable now or Paste the syntax into the syntax window for further edition. After you have made your selection, click Finish.

6. In your Data View, you should notice your new date variable has been created in the last column of your dataset, with the appropriate duration either added or subtracted.

**Calculate the Amount of Time between Two Dates**

1. Click Transform and then click Date and Time Wizard…. This brings up the Date and Time Wizard dialog box.

2. The first question you are presented with asks What would you like to do? Click Calculate with dates and times. Click Next >. This brings up the Date and Time Wizard – Step 1 of 3 dialog box.
3. Click **Calculate the number of time units between two dates (e.g., calculate an age in years from a birthdate and another date)**. Click **Next >**. This brings up the **Date and Time Wizard – Step 2 of 3** dialog box shown in Figure 9.2.

![Date and Time Wizard – Step 2 of 3](image)

**Figure 9.2**

4. The elements of the **Date and Time Wizard – Step 2 of 3** dialog box are described as follows:
   
a. You will notice all of the variables in your dataset are presented in the **Variables:** box. In addition, SPSS has created an optional **Current date and time [$TIME$]** variable that you can use in your calculation. This will always be the first variable listed, and similar to above, if you do not use this variable, SPSS will delete it immediately after you finish the procedure.
   
b. In the **Date1:** box, you will select the most recent date variable. That is, with this procedure, you are always subtracting dates, thus, it is in your best interest to have the largest (i.e., most recent) date provided first, so that you avoid negative numbers after subtraction. Select the date variable you want to use from the **Variables:** box and click the right arrow ( ).
   
c. In the **minus Date2:** box, you will place the date variable you want to subtract from the **Date1:** variable. For example, if you want to calculate the participant’s age at the beginning of the study, you would place their date of birth in this box. Select the date variable you want to use from the **Variables:** box and click the right arrow ( ).
   
d. Next, you must select the **Units:** of your calculated difference. That is, SPSS needs to know whether this difference is in **Years, Months, Days, Minutes,** etc. Click the down arrow ( ) to select the appropriate units.
   
e. Finally, you must select the **Result Treatment**. Here, you have three options.
   
i. First, you can **Truncate to integer**. This option ignores any fractional portion of the calculation. For example, subtracting 10/28/2006 from 10/21/2007 returns a result of 0 for **Years** or 11 for **Months**.
ii. Second, you can Round to integer. Here, the result will be rounded to the closest integer. For example, subtracting 10/28/2006 from 10/21/2007 returns a result of 1 for Years or 12 for Months.

iii. Finally, you can Retain fractional part, where the complete value is retained. That is, subtracting 10/28/2006 from 10/21/2007 returns a result of 0.98 for Years or 11.76 for Months.

When you are satisfied with your selections, click Next to bring up the Date and Time Wizard – Step 3 of 3 dialog box.

5. You should notice that your Calculation: is presented in mathematical format near the top of the dialog box. If this calculation is incorrect, simply click < Back to make the necessary changes. Finally, upon completing this procedure, SPSS will create a new variable in the last column of your dataset. Thus, you are required to provide a name for the Result Variable: and optionally provide a Variable Label:. Remember, no special characters or spaces are permitted in the new variable’s name. When you are satisfied with your calculation and new variable name, you can either Create the variable now or Paste the syntax into the syntax window for further edition. After you have made your selection, click Finish.

6. In your Data View, you should notice your new duration variable has been created in the last column of your dataset.
In this section, I discuss various procedures for you to obtain descriptive statistics. I begin with the Frequencies procedure, which you will use often to obtain measures of central tendency, dispersion, and distribution. In the subsequent Chapters, I move away from descriptive statistics and into selecting subsets of participants in your dataset as well as creating new variables. I close this section with descriptive statistics that provide a nice transition into more inferential statistics of later sections.
Chapter 10
FREQUENCIES

In SPSS, the Frequencies procedure is synonymous with descriptive statistics. That is, this procedure is the one you will use to calculate measures of central tendency (e.g., means), distributional statistics (e.g., skewness), and dispersion measures (e.g., standard deviation). Further, you have the option to produce several graphs, which assist in interpreting the three calculations above. To run the Frequencies procedure:

1. Click **Analyze**, then choose **Descriptive Statistics**, then click **Frequencies**…. This brings up the *Frequencies* dialog box shown in Figure 10.1. (Note that the *Bootstrap*... option may be unavailable to you if you do not have the Complex Samples add-on).

![Figure 10.1](image)

2. You will notice that all of the variables in your dataset are presented on the left side of this dialog box. Move any variable(s) for which you want descriptive statistics into the *Variable(s):* box by selecting the variable and clicking the right arrow ( ).
   a. Note that you can move one or all of your variables into the *Variable(s):* box; however, the options you select below will be applied to all variables in this box. That is, if you request the *Mean* and you have both nominal and interval level variables, a mean will be provided for all variables, but the mean for any nominal variables is essentially useless.

3. Next, I will describe each of the options on the right side of the *Frequencies* dialog box in Figure 10.1 (excluding the *Bootstrap*... option).
   a. Clicking the **Statistics**… button will bring up the *Frequencies: Statistics* dialog box shown in Figure 10.2.
      i. Under *Percentile Values*, you have the option to group within-variable values into quartiles, cut points, or percentiles. The *Quartiles* option will divide your variable(s) into four groups of equal size based on the 25th, 50th (i.e., median), and 75th percentiles. The *Cut points for: n equal groups* option will divide the variable into *n* equal groups based on
percentiles. For example, typing 5 into this box will create equal groups based on the 20th, 40th, 60th, and 80th percentile. Finally, the Percentile(s): option allows you to create as many groups as you want based on any specific percentile(s).

ii. Under Central Tendency, you can request the Mean, Median, and Mode. In addition, you can select Sum to add all non-missing values within the variable.

iii. Under Dispersion, you can select the standard (Std.) deviation, Variance, and Range. You can also request the lowest and highest values within a variable by selecting Minimum and Maximum, respectively. Finally, you can request the standard error of the mean (S.E. mean), which will provide you with the standard deviation of the sampling distribution.

iv. Under Distribution, you can request the Skewness and Kurtosis. Requested values of skewness and kurtosis will range from $-\infty$ to $+\infty$. For skewness, positive numbers indicate positive (or right) skewness, whereas negative numbers indicate negative (or left) skewness. For kurtosis, positive numbers indicate a leptokurtic (i.e., thinner tailed) distribution, whereas negative numbers indicate a platykurtic (i.e., thicker tailed) distribution.

v. Finally, you have the option to request Values are group midpoints. You would use this option if your values are midpoints of groups. For example, if the ages of all people in their thirties are coded as 35. This option estimates the median and percentiles for the original, ungrouped data.

When you are satisfied with your selections, click Continue to return to the Frequencies dialog box.
b. Clicking the **Charts** button will bring up the *Frequencies: Charts* dialog box. Here, you have the option to choose one *Chart Type*. The default is *None*; however, you may choose either *Bar charts*, *Pie charts*, or *Histograms*. If you choose a histogram, you will also have the option to *Show normal curve on histogram*, which superimposes the normal curve in your distribution. Finally, you have the choice to have *Chart Values* reported as *Frequencies* or *Percentages*. When you are satisfied with your selections, click **Continue** to return to the *Frequencies* dialog box.

c. Clicking the **Format** button will bring up the *Frequencies: Format* dialog box. Here, you tell SPSS how you want to order values or handle multiple variables, and whether you want to suppress categories.

i. In the *Order by* section, you have the option to order the values in your variables as *Ascending values*, *Descending values*, *Ascending counts*, or *Descending counts*. These options pertain specifically to the frequency table you can request in Step 4 below.

ii. In the *Multiple Variables* section, you can select either *Compare variables* to display all variables in a single table or *Organize output by variables* to display separate statistics tables for each variable.

iii. Finally, you can select to *Suppress tables with many categories* by defining the *Maximum number of categories*: This option prevents the display of tables in your Output window with more than the specified number of values. This does not mean the descriptive statistics are not calculated, they simply show up as hidden and must be double-clicked to be viewed.

When you are satisfied with your selections, click **Continue** to return to the *Frequencies* dialog box.

4. Finally, you have the option to *Display frequency tables*. This option will display in your Output window a table containing four columns, including the count (frequency), percent, valid percent (i.e., no considering missing values), and cumulative percent.

a. Unless you are specifically interested in the frequency counts, I recommend that you do not select this option if you are only considering interval level variables. That is, if selected, the table will contain every value of your interval level variable, creating the possibility for an extremely long, unnecessary table. This, however, cannot be avoided if you are also calculating descriptive statistics for nominal- or ordinal-level variables in addition to interval-level variables.

5. When you are satisfied with your choices, you can click **OK** to produce your selected results. Click the *Output* window to view your results (if it does not popup automatically). An example of what a typical Output might look like is presented in Figure 10.3. Note that some normally visible items have been hidden as described in the introduction to the *Exporting Output* section of Chapter 4.

a. In this output, I requested the mean, median, standard deviation, skewness, kurtosis as well as the minimum and maximum values. I also requested a histogram with superimposed normal curve and a frequency table.
Frequencies

Statistics

<table>
<thead>
<tr>
<th></th>
<th>Group</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>N Valid</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mean</td>
<td>2.42</td>
<td>71.68</td>
</tr>
<tr>
<td>Median</td>
<td>2.00</td>
<td>71.62</td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>1.332</td>
<td>3.529</td>
</tr>
<tr>
<td>Skewness</td>
<td>0.114</td>
<td>0.029</td>
</tr>
<tr>
<td>Std. Error of Skewness</td>
<td>0.172</td>
<td>0.172</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-1.377</td>
<td>-1.054</td>
</tr>
<tr>
<td>Std. Error of Kurtosis</td>
<td>0.342</td>
<td>0.342</td>
</tr>
<tr>
<td>Minimum</td>
<td>1</td>
<td>65</td>
</tr>
<tr>
<td>Maximum</td>
<td>4</td>
<td>78</td>
</tr>
</tbody>
</table>

Frequency Table

<table>
<thead>
<tr>
<th>Group</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid Treatment 1</td>
<td>55</td>
<td>27.5</td>
<td>27.5</td>
<td>27.5</td>
</tr>
<tr>
<td>Treatment 2</td>
<td>53</td>
<td>26.5</td>
<td>26.5</td>
<td>54.0</td>
</tr>
<tr>
<td>Treatments 1 &amp; 2</td>
<td>44</td>
<td>22.0</td>
<td>22.0</td>
<td>76.0</td>
</tr>
<tr>
<td>Placebo</td>
<td>48</td>
<td>24.0</td>
<td>24.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Histogram

Figure 10.3
Chapter 11
SPLIT FILE

The Split File procedure is incredibly useful. This procedure allows you to obtain descriptive statistics or conduct statistical analyses on subsets of participants in your dataset. For example, say you want to compare the variances of your outcome variable across four treatment groups. That is, you want to inspect whether the homogeneity of variance assumption is tenable (discussed in later Chapters). When you request variances in the Frequencies procedure (Chapter 10), the Split File procedure will provide the variances for each individual group separately in your output. To use the Split File procedure:

1. Click Data and then click Split File…. This brings up the Split File dialog box, shown in Figure 11.1.

![Figure 11.1](image)

2. You will notice that all the variables in your dataset are presented in the box on the left hand side of this dialog box. Underneath the box containing your variable names, you are informed of your current Split File status. In the Figure, Current Status: Analysis by groups is off. On the right hand side, you have three options described below:

   a. The first is to Analyze all cases, do not create groups. This option is essentially the default for all SPSS datasets. That is, no groups are created, or, to put it another way, the Split File procedure is not used. Selecting this option will analyze all cases without grouping.

   b. The second option is to Compare groups. Selecting this option presents results grouped together in one table for comparisons purposes, as in Figure 11.2.

   c. The third option is to Organize output by groups. Here, the results for each group are presented in separate tables.

3. After you have selected either Compare groups or Organize output by groups, you are required to provide your grouping variable into the Groups Based on box. Click your grouping variable and then click the right arrow (>>) to move the variable into the box.
4. Next, you are asked whether you want to *Sort the file by grouping variable* or whether the *File is already sorted*. If you are at all uncertain about whether your dataset is sorted correctly, select the default *Sort the file by grouping variable*. If you are certain your data is sorted by the grouping variable, you may select *File is already sorted*. However, if the file was not already sorted, when you request any descriptive statistics or statistical analysis the following warning message will be provided in your output:

   **Warnings**

   The file is not sorted in a consistent manner on the split file variables. It is likely that any procedure output will be incorrect.

   Indeed, the results will be incorrect and you must re-run the Split File procedure choosing to *Sort the file by grouping variable*.

5. When you are satisfied with all of your selections, click **OK**.

6. After running this procedure, in *Data View*, you will notice the dataset is now sorted by your grouping variable in ascending order. You will also see in the lower right hand corner the phrase *Split by *, where the asterisk represents your grouping variable. This indicates SPSS has successfully split the file. Any procedure(s) you run will produce results for the individual groups, as long as the Split File procedure remains in effect. However, if you sort any column other than your grouping variable, the Split File procedure will be cancelled evidenced by the blank space where the *Split by* phrase used to be located.
Chapter 12
SELECT CASES

Selecting cases is useful when you want to conduct an analysis involving only a subset of participants in your dataset. For example, say you want to analyze cardiovascular risk only within a subset of participants 65 years or older. Further, you can select a random sample of participants. To select a subset of participants:

1. Click **Data** and then click **Select Cases**… This brings up the Select Cases dialog box shown in Figure 12.1.
   a. You will notice all the variables in your dataset are listed in the box on the left side of the dialog box, and underneath this box SPSS informs you of your **Current Status:** related to selecting cases.

   ![Select Cases Dialog Box](image)

   **Figure 12.1**

2. The **Select** section contains five options regarding how you want to select participants and for what specific purpose. They are each described below.
   a. The first option is to select **All cases**. This is the default option and is equivalent to using all cases. That is, choosing this option will select all participants in your dataset for analysis.
   b. The second option is the most popular and selects cases **If condition is satisfied**. After clicking this radio button, click **If...** to bring up the **Select Cases: If** dialog box. This dialog box operates similarly to the Compute procedure of Chapter 13. You select participants by entering a mathematical expression in the uppermost box. For example, to select only individuals aged 65 or older and your age variable is named **Age**, you would type \( Age \geq 65 \) in this box.
c. Third, you can select a Random sample of cases. After clicking this option, click Sample… to bring up the Select Cases: Random Sample dialog box. Here, you have two choices, you can select a percentage of participants, or you can select an exact number of participants from a specific number of cases.

d. Fourth, you can select participants Based on time or case range. After selecting this option, click Range… to bring up the Select Cases: Range dialog box. Here, you can select a range of cases based on specific row numbers in your dataset. Note that participant selection is based on row numbers, not their unique identifier variable.

e. Finally, you can Use a filter variable to select participants. For this option, any participants with any value other than 0 or missing for the filter variable would be selected. To use a filter variable, select your filter variable from the list of variables on the left hand side of the dialog box and click the right arrow ( ). Any variable in your dataset can serve as a filter variable.

3. After you have chosen how you want participants to be selected, you now must decide how they will be identified in your dataset. You have three options.

a. First, you can Filter out unselected cases. This option simply creates a new binary (i.e., 0/1) filter variable at the end of your dataset, named filter_$ by default. For this variable, 1 = selected and 0 = filtered. You can easily identify filtered cases by sorting the filter variable or by looking for row numbers with slashes through them ( ). Note that only selected cases will be used in analysis. To remove the filter, either delete the filter_$ variable or perform Step 2a above.

b. Second, you can Copy selected cases to a new dataset. Here, cases with a value of 1 for the filter_$ variable will be copied to a new dataset. You must provide a Dataset name:. Note that the names for this dataset follow the same rules as variables. That is, no spaces or special characters are allowed. Further, selected cases are not removed from the original dataset. That is, this option will create a new and separate dataset containing only the cases you selected. Thus, you can save this new dataset and perform necessary statistical procedures without making any changes to your original dataset (other than creating the filter_$ variable).

c. Finally, you can Delete unselected cases. Just as the name implies, any cases that are not selected will be deleted permanently from your dataset. This option should only be considered when you are 100% certain of what you are doing.

4. When you have satisfied with your choices, click OK.

a. Prior to conducting any analysis, make sure the filter_$ variable was created and that SPSS filtered your participants correctly.
Chapter 13

COMPUTE VARIABLE

The compute variable procedure in SPSS is extremely useful. This procedure allows you to create new variables that are mathematical transformations of original variables. These mathematical transformations could involve summing across variables, calculating the difference between two variables (e.g., pretest vs. posttest change), or calculating the square root or natural log. Further, this operation can be used to identify missing data, select random data, or convert variable types.

To use the Compute function:

1. Click **Transform** and then click **Compute Variable**…. This brings up the **Compute Variable** dialog box shown in Figure 13.1.

2. In the top left corner of this dialog box, you will see the **Target Variable**: box. This is where you will type the name of the new variable you are computing. Remember, no spaces or special characters are allowed in this new variable’s name.

3. Underneath **Target Variable**: box, you can click the **Type & Label**… button to bring up the **Compute Variable: Type and Label** dialog box.
   
   a. Here, you can create a variable **Label**: or you can use expression as label.
   
   b. Second, you can select the **Type** of variable you are creating. That is, whether the variable will be **Numeric** or **String**. If the variable is a **String**, you can indicate the number of characters the variable can have in the **Width**: box.
   
   c. When you have made your selections, click **Continue**.

4. The large box across the top is where you will type your **Numeric Expression**. The numeric expression is where you will provide your mathematic transformation. You have two options for creating your numeric expression:

   a. First, you can type the text yourself. I advise this option only if you know exactly what SPSS needs to produce the desired result (e.g., square root = sqrt).
b. Second, you can use the Function group: box. Here, you will select your function group and then scroll through the list presented in the Functions and Special Variables: box. A description of the function and the required format to run the expression is then provided under the calculator buttons. When you have found the function you need, click the up arrow (↑) to move this function into the Numeric Expression: box.

5. When you selected the necessary function, you will need to identify the variables you want to use. SPSS will require the variables to be within the function’s parentheses. Select the variable(s) from the left hand side of the Compute Variable dialog box and then click the right arrow (→) next to the Numeric Expression: box.

6. Finally, you can click the If... box to bring up the Compute Variable: If Cases dialog box. This is an optional step if you only want to compute a variable for a select number of participants. Upon opening this dialog box, you will notice immediate similarities between it and Figure 13.1. When you have identified the participants you want to include, click Continue.

7. When your Numeric Expression: is complete, click OK.

   a. Note that if any portion of your Numeric Expression: is incorrect, SPSS will pop up a warning message.
      i. If your mathematical function is incorrect, a warning message shown at the top of Figure 13.2 will pop up.
      ii. If you have incorrectly typed the variable names, the warning message at the bottom of Figure 13.3 will pop up.

b. Bottom line, if you see any warning message, you will need to correct your Numeric Expression: statement. When you have done so, click OK.

8. Verify your new variable was created in Data View and check to make sure it was computed correctly.

---

**Figure 13.2**
Chapter 14
DESCRIPTIVES

The Descriptives procedure allows you to calculate many of the same descriptive statistics available in the Frequencies procedure. However, two primary differences exist. First, in Descriptives you cannot chart your responses. That is, you do not have the option for histograms or pie charts. Second, the Descriptives procedure allows for easy calculation of standardized scores (i.e., \(z\)-scores). \(z\)-scores are useful for identifying outlying values or placing multiple variables on the same scale. They can be created for any interval- or ratio-level variable and are saved directly as a new variable in your dataset.

1. Click **Analyze**, then choose **Descriptive Statistics**, and then click **Descriptives**… to bring up the **Descriptives** dialog box.

2. All of the variables in your dataset are listed on the left hand side of this dialog box. To calculate descriptive statistics for any variable, click the variable name and then click the right arrow (\(\rightarrow\)) next to the **Variable(s):** box.

3. Next, you can click the **Options**… button to bring up the **Descriptives: Options** dialog box. Here, you can select various descriptive statistics including:
   a. The **Mean** as the primary measure of central tendency, or **Sum** to add all non-missing values within the variable.
   b. Under **Dispersion**, you can select the standard deviation (**Std. deviation**), **Variance**, and **Range**. You can also request the lowest and highest values within a variable by selecting **Minimum** and **Maximum**, respectively. Finally, you can request the standard error of the mean (**S.E. mean**), which will provide you with the standard deviation of the sampling distribution.
   c. Under **Distribution**, you can request the **Skewness** and **Kurtosis**. Requested values of skewness and kurtosis will range from \(-\infty\) to \(+\infty\). For skewness, positive numbers indicate positive (or right) skewness, whereas negative numbers indicate negative (or left) skewness. For kurtosis, positive numbers indicate a leptokurtic (i.e., thinner tailed) distribution, whereas negative numbers indicate a platykurtic (i.e., thicker tailed) distribution.
   d. Finally, you can select the **Display Order**. By default, variables are displayed in the order in which they were placed in the **Variable(s):** box. Optionally, you can display variables alphabetically, by ascending means, or by descending means.

When you are satisfied with your choices, click **Continue**.

4. Finally, you have the option to click **Save standardized values as variables**. Standardized values are a synonym for \(z\)-scores. To save these standardized values as variables means a new \(z\)-score variable will be created for every variable included in the **Variable(s):** box.

5. When you are satisfied with your selections, click **OK**. An output will be produced similar to Figure 14.1. Click your Output window (if it does not pop up automatically). You should notice some similarity between this output and the output produced by the Frequencies procedure.
### Descriptive Statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Skewness</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTH</td>
<td>200</td>
<td>15</td>
<td>100</td>
<td>37.31</td>
<td>14.609</td>
<td>1.444</td>
<td>.172</td>
</tr>
<tr>
<td>Osteocalcin</td>
<td>200</td>
<td>39</td>
<td>7.70</td>
<td>3.6884</td>
<td>1.18188</td>
<td>.236</td>
<td>.172</td>
</tr>
<tr>
<td>NTxCr</td>
<td>200</td>
<td>14.10</td>
<td>182.30</td>
<td>51.1874</td>
<td>27.74191</td>
<td>1.978</td>
<td>.172</td>
</tr>
<tr>
<td>CaAbsorb</td>
<td>197</td>
<td>.99</td>
<td>4.99</td>
<td>2.5978</td>
<td>.71048</td>
<td>.614</td>
<td>.173</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td>197</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 14.1**
Chapter 15
EXPLORE

All parametric statistical tests are based on distributional assumptions. The normality assumption states that the sampling distribution or residuals are normally distributed. The probability of incorrect inference increases as these distributions become increasingly non-normal. One approximation of the tenability of this assumption is to assure the within-group distributions of raw scores are normally distributed. That is, if you are evaluating group differences, the distributions for each category (i.e., level) of your independent variable(s) are normally distributed. There are several methods for testing the normality of a distribution, and the Explore procedure produces histograms, probability plots, and statistical tests. In general, it is best to use consider all three methods before drawing a conclusion.

To use the Explore procedure:

1. Click **Analyze** and then choose **Descriptive Statistics**, and the click **Explore**… to bring up the **Explore** dialog box shown in Figure 15.1.

![Figure 15.1](image)

2. All the variables in your dataset are listed on the left side of this dialog box. To use this procedure, you are only required to provide a variable in the **Dependent List:** box. This variable can be on any measurement scale. To move a variable into this box, click the variable name, and then click the right arrow ( ) next to the **Dependent List:** box. You can place as many variables here as you choose.

3. The **Factor List:** box is where you will place a categorical independent variable, if you have one. That is, separate descriptive statistics will be produced for each category of this variable. You can use as many categorical variables as you want.

4. You also have the option to **Label Cases by:**. If selected, the value labels (or values, if no labels are provided in **Variable View**) of this variable can be used to label outliers or extreme cases. The most common variable used here is the ID variable (e.g., unique identifier). Again, this is optional and if you do not select this option, the outliers and extreme cases will be identified by row number.
5. In addition, you must indicate whether you want to Display Statistics, Plots, or Both. You will notice that if you select Statistics, the Plots… button cannot be clicked. Likewise, if you select Plots, the Statistics… button cannot be clicked. In general, it is a good idea to simply select Both.

6. When you click the Statistics… button you are presented with four options.
   a. The Descriptives option will print measures of central tendency and dispersion including the mean, median, 5% trimmed mean, standard error, variance, standard deviation, minimum, maximum, range, interquartile range, skewness, and kurtosis. Further, the Confidence Interval for Mean: is also produced based on the value you input.
   b. The M-estimators option provides robust alternatives to the sample mean and median for estimating the location. The estimators calculated differ in the weights they apply to participants and include Huber's M-estimator, Andrews' wave estimator, Hampel's redescending M-estimator, and Tukey's biweight estimator.
   c. The Outliers option displays the five largest and five smallest values. Note that this does not necessarily indicate that these values are disconnected from the distribution (i.e., definition of an outlier).
   d. Finally, the Percentiles option displays the values for the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles.

When you are satisfied, click Continue.

7. When you click the Plots… button you are presented with several plots options.
   a. Under the Boxplots section, you can choose to Factor levels together which prints a separate display for each variable in the Dependent List: box. This option is similar to the Compare groups option in the Split File command described in Chapter 11. Or, you can choose the Dependent together option which prints a separate display for each group defined by the Factor List: variable. This option is similar to the Organize output by groups option in the Split File command. Finally, you can simply choose None.
   b. Under the Descriptive section, you can choose to display Stem-and-leaf plots or a Histogram.
   c. You also have the option to print Normality plots with tests. This option will provide you with normal probability and detrended probability plots. Further, the Kolmogorov-Smirnov and Shapiro-Wilk statistics are calculated.
   d. Finally, under the Spread vs Level with Levene Test section, you can choose various data transformations. The default option is None.
      i. The Power estimation option produces a plot of the natural logs of the interquartile ranges against the natural logs of the medians for all cells, as well as an estimate of the power transformation for achieving equal variances in the cells. The spread-versus-level plot helps to determine
the power for a transformation to stabilize variances across groups to be used in the \textit{Transformed} option described below.

ii. The \textit{Transformed} option allows you to select from a list of power alternatives including Natural Log, 1/square root, Reciprocal, Square Root, Square, or Cube. This option will produce plots of transformed data.

iii. Finally, the \textit{Untransformed} option produces plots of the raw data.

When you are satisfied with your selections, click \textbf{Continue}.

8. Finally, you can click the \textbf{Options…} button to indicate how you want SPSS to handle missing data. You have the option to \textbf{Exclude cases listwise}, \textbf{Exclude cases pairwise}, or \textbf{Report values}. When you are satisfied, click \textbf{Continue}.

9. After you have made all of you selections, click \textbf{OK}.

\textbf{Output and Interpretation}

When you click \textbf{OK}, SPSS will produce an \textit{Output} screen displaying your results. Click your \textit{Output} window if it does not popup automatically. I will describe the typical output you will see if you selected all of the bolded options in the Steps above.

The first table you will see is titled \textit{Case Processing Summary}. Here, you are provided the number of \textit{Valid}, \textit{Missing}, and \textit{Total} cases. Either this will be for the entire sample or within each group individually depending on whether you selected a \textit{Factor List:} variable in Step 3 above. The second table provides your \textit{Descriptives} statistics. This table is similar to the output produced by the Frequencies procedure described in Chapter 10.

Next, you are presented with the \textit{Tests of Normality} table, shown at in Figure 15.2. Here, the \textit{Kolmogorov-Smirnov} and \textit{Shapiro-Wilk} statistics are presented. If you used a categorical independent variable in the \textit{Factor List:} box in Step 3 above, separate tests are produced for each category. These tests evaluate whether the distribution follows a normal distribution; thus, a statistically significant result (i.e., \textit{Sig.} < .05) means that the distribution is significantly different from a normal distribution. Thus, you want to have non-statistically significant results.

\begin{table}[h]
\centering
\begin{tabular}{|l|c|c|c|c|c|c|}
\hline
\textbf{Group} & \multicolumn{3}{|c|}{\textit{Kolmogorov-Smirnov}} & \multicolumn{3}{|c|}{\textit{Shapiro-Wilk}} \\
& \textbf{Statistic} & \textbf{df} & \textbf{Sig.} & \textbf{Statistic} & \textbf{df} & \textbf{Sig.} \\
\hline
PTH & \text{.179} & 55 & \text{.000} & \text{.844} & 55 & \text{.000} \\
Treatment 1 & \text{.110} & 53 & \text{.160} & \text{.921} & 53 & \text{.002} \\
Treatment 2 & \text{.130} & 44 & \text{.058} & \text{.949} & 44 & \text{.049} \\
Treatments 1 & \text{.130} & 48 & \text{.022} & \text{.867} & 48 & \text{.000} \\
Placebo & \text{.130} & 48 & \text{.022} & \text{.867} & 48 & \text{.000} \\
\hline
\multicolumn{7}{|l|}{a Lilliefors Significance Correction} \\
\end{tabular}
\caption{Tests of Normality}
\end{table}
Next, you are presented with individual *Histograms*, similar to what obtained from the Frequencies procedure in Chapter 10. Although the example data used in this Chapter had four groups, only the histogram for Treatment 1 is shown in Figure 15.3. Histograms are incredibly useful for both visually evaluating the overall distribution of your data as well as identifying participants whose values are disconnected from the rest of the distribution (i.e., outliers).

Then, you are presented with *Normal Q-Q Plots*. The Q-Q plot for Treatment 1 is shown in Figure 15.4. A Q-Q plot, or quantile-quantile plot, is used to see how well a theoretical (or comparison) distribution matches your observed data. In this case, the theoretical distribution is the normal distribution. If the two distributions match, the points on the plot will form a linear pattern along the solid diagonal reference line. Thus, if the points fall along the reference line the observed distribution is close to normal; however, if the points do not fall along this line the distribution departs from normal.
In addition, SPSS will print *Detrended Normal Q-Q Plots*, which depict the actual deviations of data points from the theoretical distribution. However, no specific pattern in this plot indicates normality; thus, these plots are usually ignored.

Finally, a boxplot is produced, shown in Figure 15.5. The appearance of the boxplot will depend on whether you placed multiple variables in the *Dependent List:* box in Step 2 and/or categorical variables in the *Factor List:* box in Step 3. A detailed description of how to interpret a boxplot is provided at the beginning of Chapter 18.
The SPSS Frequencies procedure, described in Chapter 10, allowed you to produce several graph/chart options; however, the stand-alone graphing procedures in SPSS offer many more graph/chart options that can be customized to meet your needs. Within these procedures, you have the ability to produce bar charts, line charts, boxplots, and scatterplots among others. I will describe how to produce the four listed options in the next few Chapters.

It should be noted that I use the Legacy Dialogs procedure in the descriptions below. You also have the option to use the Chart Builder procedure. Some people find Chart Builder easier to work with, whereas others will find the procedures below easier. With that said, you will get the same results regardless of the procedure you choose.
Bar charts are appropriate when the variable you want use on the x-axis is categorical (i.e., nominal- or ordinal-level) to summarize either frequency counts within one variable (think, histograms) or a summary measure of central tendency across variables or categories.

1. Click **Graphs**, then choose **Legacy Dialogs**, then click **Bar**…. This brings up the **Bar Charts** dialog box.

2. You will notice initially you have three chart options **Simple**, **Clustered**, or **Stacked**.
   a. A **Simple** bar chart is used when you want to describe the categories within one variable. Note that each bar will present a single piece of information (described below).
   b. A **Clustered** bar chart is used when you have multiple variables—typically multiple independent variables. Here, the categories of one variable can be represented within the categories of another variable. For example, gender within treatment group.
   c. A **Stacked** bar chart is also used when you multiple variables, but results are stacked on top of one another in the same bar. I consider this type of bar chart unnecessarily confusing, and can honestly say I have never seen it in published literature. So, I am safe in assuming that it is not used very often. Thus, it will not be discussed here.

3. At the bottom of **Bar Charts** dialog box, you must tell SPSS what the **Data in Chart Are**.
   a. The default option, **Summaries for groups of cases**, summarizes groups of participants within an individual variable.
   b. The **Summaries of separate variables** option summarizes multiple variables. A bar is presented for each variable. That is, each bar represents one of the variables.
   c. Finally, the **Values of individual cases** option summarizes responses across participants. That is, the subject number serves as the individual categories along the X-axis. Similar to the stacked bar chart, this option is used sparingly and will not be discussed here.

In the sections that follow, I will provide you with descriptions for simple and clustered bar charts, both for summaries of groups of cases and for separate variables.

### Simple Bar Chart, Summaries for Groups of Cases

1. Click **Simple**, then click **Summaries for groups of cases**, and then click **Define**. This brings up the **Define Simple Bar: Summaries for Groups of Cases** dialog box, presented in Figure 16.1.
2. First, you are asked what the **Bars Represent**. Here you have five choices:
   a. \(N\) of cases: Each bar represents the number of response within each category (i.e., frequency count).
   b. Cum. \(N\): Each bar represents the cumulative total frequency count across categories.
   c. % of cases: Each bar represents the percentage of participants within each category.
   d. Cum. %: Each bar represents the cumulative percentages of participants across categories.
   e. Other statistic (e.g., mean): When selecting this option, you must provide a variable: used to represent each category. To select a variable, click the variable name on the left hand side of the dialog box, and then click the right arrow (\(\rightarrow\)) next to the **Variable:** box. Further, you can click Change Statistic... and select one of a number of descriptive, rank, or percentile options.

3. Next, you must identify the **Category Axis:**. This is where you will identify the primary nominal- or ordinal-level variable to be used on the x-axis. To select this variable, click the variable name on the left hand side of the dialog box, and then click the right arrow (\(\rightarrow\)) next to the **Category Axis:** box.

4. In addition, you have the option to panel your bar chart. A paneled chart is a grid of subcharts that share the same axes, but for a different group of one or more categorical variables. For example, you can present a separate bar chart for males and females on top of one another (\(\text{Rows}::\)) or next to one another (\(\text{Columns}::\)).
5. Clicking the *Titles*... button in the upper right hand corner brings up the *Titles* dialog box. Here, you can optionally provide the chart with a *Title*, *Subtitle:* or *Footnote.*

6. Clicking the *Options*... button in the upper right hand corner brings up the *Options* dialog box, where you can optionally indicate how you want to handle *Missing Values* or if you want to *Display error bars.*

7. Finally, you have the option to use a *Template* from a previous chart you have created. If you choose this option, only Steps 2 and 3 above are required.

8. When you are satisfied with your choices, click **OK.** SPSS will produce your simple bar chart in the output window, similar to Figure 16.2.

---

**Clipped Bar Chart, Summaries for Groups of Cases**

1. Click **Clustered**, then click **Summaries for groups of cases**, and then click **Define**. This brings up the *Define Clustered Bar: Summaries for Groups of Cases* dialog box.

   a. The only difference between this dialog box and the one presented in Figure 16.1 is that you now have to *Define Clusters by:* box where you will place the second categorical variable to group your participants. To do so, select the variable on the left hand side of the dialog box and click the right arrow (→) next to the *Define Clusters by:* box.

2. Similar to the Simple Bar Chart, you are asked what the *Bars Represent.* Here you have five choices:

   a. **N of cases**: Each bar represents the number of response within each category (i.e., frequency count).

   b. **Cum. N**: Each bar represents the cumulative total frequency count across categories.

   c. **% of cases**: Each bar represents the percentage of participants within each category.
d. *Cum, %*: Each bar represents the cumulative percentages of participants across categories.

e. *Other statistic (e.g., mean)*: When selecting this option, you must provide a *Variable* used to represent each category. To select a variable, click the variable name on the left hand side of the dialog box, and then click the right arrow (→) next to the *Variable*: box. Further, you can click *Change Statistic...* and select one of a number of descriptive, rank, or percentile options.

3. Next, you must identify the *Category Axis*: This is where you will identify the primary nominal- or ordinal-level variable to be used on the *x*-axis. To select this variable, click the variable name on the left hand side of the dialog box, and then click the right arrow (→) next to the *Category Axis*: box.

4. In addition, you have the option to panel your bar chart. A paneled chart is a grid of subcharts that share the same axes, but for a different group of one or more categorical variables. For example, you can present a separate bar chart for males and females on top of one another (*Rows:* or next to one another (*Columns:*)).

5. Clicking the *Titles...* button in the upper right hand corner brings up the *Titles* dialog box. Here, you can optionally provide the chart with a *Title*, *Subtitle:*, or *Footnote*.

6. Clicking the *Options...* button in the upper right hand corner brings up the *Options* dialog box, where you can optionally indicate how you want to handle *Missing Values* or if you want to *Display error bars*.

7. Finally, you have the option to use a *Template* from a previous chart you have created. If you choose this option, only Steps 2 and 3 above are required.

8. When you are satisfied with your choices, click *OK*. SPSS will produce your simple bar chart in the output window, similar to Figure 16.3.

---

**Figure 16.3**

![Bar chart showing count by group and gender with Treatment 1, Treatment 2, Treatments 1 & 2, and Placebo groups distinguished by gender: Male and Female](chart.png)
Simple Bar Chart, Summaries of Separate Variables

1. Click Simple, then click summaries of separate variables, and then click Define. This brings up the Define Simple Bar: Summaries for Separate Variables dialog box presented in Figure 16.4.

![Figure 16.4](image-url)

2. Because the bars will now represent individual variables, you must indicate what the Bars Represent: Here, you will place two or more variables you want to summarize. Typically, these variables will be interval-level. Note, however, that the y-axis will remain to scale, thus, if you have two variables on drastically different scales, separate charts will be appropriate.
   a. Click the variable(s) you want to use on the left hand side of the dialog box, and then click the right arrow (→) next to the Bars Represent: box.
   b. Optionally, you can click the Change Statistic... box to bring up the Statistic dialog box. Here, you will select the summary statistic for each of your variables. The default value is the Mean of values where the mean variable response will be presented. You have a number of choices to choose from including descriptive, rank, or percentile options. When you are satisfied, click Continue.

3. In addition, you have the option to panel your bar chart. A paneled chart is a grid of subcharts that share the same axes, but for a different group within one or more categorical variables. For example, you can present a separate bar chart for males and females on top of one another (Rows:) or next to one another (Columns:).
4. Clicking the *Titles*... button in the upper right hand corner brings up the *Titles* dialog box. Here, you can optionally add a *Title, Subtitle: or Footnote* to the chart.

5. Clicking the *Options*... button in the upper right hand corner brings up the *Options* dialog box, where you can optionally indicate how you want to handle *Missing Values* or if you want to *Display error bars*.

6. Finally, you have the option to use a *Template* from a previous chart you have created. If you choose this option, only Step 2 above is required.

7. When you are satisfied with your choices, click **OK**. SPSS will produce your simple bar chart in the output window similar to Figure 16.5.

---

**Figure 16.5**  
*Clustered Bar Chart, Summaries of Separate Variables*

1. Click **Clustered**, then click **Summaries of separate variables**, and then click **Define**. This brings up the *Define Clustered Bar: Summaries for Separate Variables* dialog box.

   a. The only difference between this dialog box and the one presented in Figure 16.4 is that you now have the *Category Axis: box where you will place a categorical variable to group your participants. To select this variable, click the variable name on the left hand side of the dialog box, and then click the right arrow (→) next to the *Category Axis: box.*

2. Because the bars will now represent individual variables, you must indicate what the **Bars Represent:**. Here, you will place two or more variables you want to summarize. Typically, these variables will be interval-level. Note, however, that the y-axis will remain to scale, thus, if you have two variables on drastically different scales, separate charts will be appropriate.

   a. Click the variable(s) you want to use on the left hand side of the dialog box, and then click the right arrow (→) next to the *Bars Represent: box.*
b. Optionally, you can click the **Change Statistic**... box to bring up the **Statistic** dialog box. Here, you will select the summary statistic for each of your variables. The default value is the *Mean of values* where the mean variable response will be presented. You have a number of choices to choose from including descriptive, rank, or percentile options. When you are satisfied, click **Continue**.

3. In addition, you have the option to panel your bar chart. A paneled chart is a grid of subcharts that share the same axes, but for a different group within one or more categorical variables. For example, you can present a separate bar chart for males and females on top of one another (*Rows:* or next to one another (*Columns:*).

4. Clicking the **Titles**... button in the upper right hand corner brings up the **Titles** dialog box. Here, you can optionally add a **Title**, **Subtitle** or **Footnote** to the chart.

5. Clicking the **Options**... button in the upper right hand corner brings up the **Options** dialog box, where you can optionally indicate how you want to handle *Missing Values* or if you want to display error bars.

6. Finally, you have the option to use a **Template** from a previous chart you have created. If you choose this option, only Step 2 above is required.

7. When you are satisfied with your choices, click **OK**. SPSS will produce your simple bar chart in the output window similar to Figure 16.6.
Chapter 17
LINE CHARTS

Line charts are appropriate for variables measured on any scale. They are useful for showing trends across variables or individuals.

1. Click **Graphs**, then choose **Legacy Dialogs**, then click **Line**…. This brings up the **Line Charts** dialog box.

2. You will notice initially you have three chart options **Simple**, **Multiple**, or **Drop-line**.
   a. A **Simple** line chart is used when you want to present data for one variable.
   b. A **Multiple** line chart is used when you have one continuous variable, but want to present a line across each level of a separate categorical variable.
   c. A **Drop-line** line chart is stacks results on top of each other. This type is not used often and will not be discussed here.

3. At the bottom of this dialog box, you must tell SPSS what the **Data in Chart Are**.
   a. The default option, **Summaries for groups of cases**, summarizes groups of participants within an individual variable.
   b. The **Summaries of separate variables** option summarizes multiple variables. A line is presented across the variable summary statistic.
   c. Finally, the **Values of individual cases** option summarizes responses across participants. That is, the subject number serves as the individual categories along the X-axis. This option is not used often and will not be discussed here.

Next, I will provide you with descriptions for each simple and clustered line charts, both for summaries of groups of cases and for separate variables.

**Simple Line Chart, Summaries for Groups of Cases**

1. Click **Simple**, then click **Summaries for groups of cases**, and then click **Define**. This brings up the **Define Simple Line: Summaries for Groups of Cases** dialog box. Aside from the heading, this dialog box is identical to Figure 16.1 from last Chapter.

2. First, you are asked what the **Lines Represent**. Here you have five choices:
   a. **N of cases**: Each data point represents the number of response within each category (i.e., frequency count).
   b. **Cum. N**: Each data point represents the cumulative total frequency count across categories.
   c. **% of cases**: Each data point represents the percentage of participants within each category.
   d. **Cum. %**: Each data point represents the cumulative percentages of participants across categories.
e. *Other statistic (e.g., mean):* When selecting this option, you must provide a

Variable: used to represent each category. To select a variable, click the variable

name on the left hand side of the dialog box, and then click the right arrow ( )

next to the Variable: box. Further, you can click Change Statistic... and select one

of a number of descriptive, rank, or percentile options.

3. Next, you must identify the Category Axis:. This is where you will identify the variable

used on the x-axis. To select this variable, click the variable name on the left hand side of

the dialog box, and then click the right arrow ( ) next to the Category Axis: box.

4. In addition, you have the option to panel your line chart. A paneled chart is a grid of

subcharts that share the same axes, but for a different group of one or more categorical

variables. For example, you can present a separate line charts for males and females

stacked on top of one another (Rows:) or next to one another (Columns:).

5. Clicking the Titles... button in the upper right hand corner brings up the Titles dialog box. Here, you can optionally provide the chart with a Title, Subtitle: or Footnote.

6. Clicking the Options... button in the upper right hand corner brings up the Options dialog box, where you can optionally indicate how you want to handle Missing Values or if you want to Display error bars.

7. Finally, you have the option to use a Template from a previous chart you have created. If you choose this option, only Steps 2 and 3 above are required.

8. When you are satisfied with your choices, click OK. SPSS will produce your simple bar chart in the output window, similar to Figure 17.1. Notice the similarity between this chart and Figure 16.2—same information simply presented differently.

---

**Figure 17.1**
**Multiple Line Chart, Summaries for Groups of Cases**

1. Click **Multiple**, then click **Summaries for groups of cases**, and then click **Define**. This brings up the *Define Clustered Line: Summaries for Groups of Cases* dialog box.
   
   a. The only difference between this dialog box and the one presented for the *Simple Line Chart, Summaries for Groups of Cases* section above is that you now have to **Define Lines by**: box. Here, you will select the categorical variable on the left hand side of the dialog box and click the right arrow (→) next to the **Define Lines by**: box.

2. Similar to the Simple Line Chart, you are asked what the **Lines Represent**. Here you have five choices:
   
   a. **N of cases**: Each data point represents the number of response within each category (i.e., frequency count).
   
   b. **Cum. N**: Each data point represents the cumulative total frequency count across categories.
   
   c. **% of cases**: Each data point represents the percentage of participants within each category.
   
   d. **Cum. %**: Each data point represents the cumulative percentages of participants across categories.
   
   e. **Other statistic (e.g., mean)**: When selecting this option, you must provide a **Variable** used to represent each category. To select a variable, click the variable name on the left hand side of the dialog box, and then click the right arrow (→) next to the **Variable**: box. Further, you can click **Change Statistic** and select one of a number of descriptive, rank, or percentile options.

3. Next, you must identify the **Category Axis**. This is where you will identify the variable used on the x-axis. To select this variable, click the variable name on the left hand side of the dialog box, and then click the right arrow (→) next to the **Category Axis**: box.

4. In addition, you have the option to panel your line chart. A paneled chart is a grid of subcharts that share the same axes, but for a different group of one or more categorical variables. For example, you can present a separate line charts for males and females stacked on top of one another (*Rows:* or next to one another (*Columns:*).

5. Clicking the **Titles**... button in the upper right hand corner brings up the **Titles** dialog box. Here, you can optionally provide the chart with a **Title**, **Subtitle**: or **Footnote**.

6. Clicking the **Options**... button in the upper right hand corner brings up the **Options** dialog box, where you can optionally indicate how you want to handle **Missing Values** or if you want to **Display error bars**.

7. Finally, you have the option to use a **Template** from a previous chart you have created. If you choose this option, only Steps 2 and 3 above are required.

8. When you are satisfied with your choices, click **OK**. SPSS will produce your simple bar chart in the output window, similar to Figure 17.2. Notice the similarity between this chart and Figure 16.3—same information simply presented differently.
Simple Line Chart, Summaries of Separate Variables

1. Click **Simple**, then click **Summaries of separate variables**, and then click **Define**. This brings up the **Define Simple Line: Summaries for Separate Variables** dialog box presented in Figure 17.3.
2. Because the plotted data will now represent individual variables, you must indicate what the Line Represents:. Here, you will place the variables you want to summarize. Typically, these variables will be interval-level. Note, however, that the y-axis will remain to scale, thus, if you have two variables on drastically different scales, separate charts may be appropriate.
   
   a. Click the variable(s) you want to use on the left hand side of the dialog box, and then click the right arrow (▼) next to the Line Represents: box.
   
   b. Click the Change Statistic... box to bring up the Statistic dialog box. Here, you will select the summary statistic for each of your variables. The default value is the Mean of values where the mean variable response will be presented. You have a number of choices to choose from including descriptive, rank, or percentile options. When you are satisfied, click Continue.

3. In addition, you have the option to panel your line chart. A paneled chart is a grid of subcharts that share the same axes, but for a different group of one or more categorical variables. For example, you can present a separate line chart for males and females stacked on top of one another (Rows:) or next to one another (Columns:).

4. Clicking the Titles... button in the upper right hand corner brings up the Titles dialog box. Here, you can provide the chart with a Title, Subtitle: or Footnote.

5. Clicking the Options... button in the upper right hand corner brings up the Options dialog box, where you can indicate how you want to handle Missing Values or if you want to Display error bars.

6. Finally, you can use a Template from a previous chart you have created. If you choose this option, only Step 2 is required.

7. When you are satisfied with your choices, click OK. SPSS will produce your simple line chart in the output window similar to Figure 17.4.

---

Figure 17.4
**Multiple Line Chart, Summaries of Separate Variables**

1. Click **Multiple**, then click **Summaries of separate variables**, and then click **Define**. This brings up the *Define Multiple Line: Summaries for Separate Variables* dialog box.
   a. The only difference between this dialog box and the one presented in Figure 17.3 is that now you have the *Category Axis:* box where you will place a categorical variable to group your participants. To select this variable, click the variable name on the left hand side of the dialog box, and then click the right arrow (→) next to the *Category Axis:* box.

2. Because the plotted data will now represent individual variables, you must indicate what the *Lines Represent:* Here, you will place the variables you want to summarize. Typically, these variables will be interval-level. Note, however, that the y-axis will remain to scale, thus, if you have two variables on drastically different scales, separate charts may be appropriate.
   a. Click the variable(s) you want to use on the left hand side of the dialog box, and then click the right arrow (→) next to the *Lines Represent:* box.
   b. Click the *Change Statistic:* box to bring up the *Statistic* dialog box. Here, you will select the summary statistic for each of your variables. The default value is the *Mean of values* where the mean variable response will be presented. You have a number of choices to choose from including descriptive, rank, or percentile options. When you are satisfied, click *Continue.*

3. In addition, you have the option to panel your line chart. A paneled chart is a grid of subcharts that share the same axes, but for a different group of one or more categorical variables. For example, you can present a separate line chart for males and females stacked on top of one another (*Rows:* ) or next to one another (*Columns:*).

4. Clicking the *Titles:* button in the upper right hand corner brings up the *Titles* dialog box. Here, you can provide the chart with a *Title, Subtitle:* or *Footnote.*

5. Clicking the *Options:* button in the upper right hand corner brings up the *Options* dialog box, where you can indicate how you want to handle *Missing Values* or if you want to *Display error bars.*

6. Finally, you can use a *Template* from a previous chart you have created. If you choose this option, only Step 2 is required.

7. When you are satisfied with your choices, click **OK.** SPSS will produce your simple line chart in the output window similar to Figure 17.5.
Figure 17.5
Chapter 18
BOXPLOTS

Boxplots, also known as box-and-whisker plots, are appropriate for variables measured on ordinal, interval, or ratio scales. The “box” of a boxplot contains the interquartile range (IQR), which is actually data ranging from the 25th to 75th percentile. The thick line in the middle of the box is the 50th percentile (i.e., median). The whiskers extend from the box at both ends, with one-quarter of the data lying in the top whisker and one-quarter in the bottom whisker. The whiskers will extend to a maximum of 1.5 times of the IQR. Data points lying 1.5 to 3.0 IQRs outside the box are termed outliers and are labeled with circles (o). Data points falling 3.0 IQRs outside the box are termed extreme cases and are labeled with asterisks (*). This will make more sense one you actually see the boxplot.

To create a boxplot:

1. Click **Graphs**, then choose **Legacy Dialogs**, then click **Boxplot**.... This brings up the **Boxplot** dialog box.
2. You will notice initially you have two boxplot options **Simple** or **Clustered**.
   a. A **Simple** boxplot is used when you want to present data for one variable.
   b. A **Clustered** boxplot is used when you want to present multiple variables.
3. At the bottom of this dialog box, you must tell SPSS what the **Data in Chart Are**.
   a. The default option, **Summaries for groups of cases**, summarizes categories of an individual variable.
   b. The **Summaries of separate variables** option summarizes multiple variables

Next, I will provide you with descriptions for each simple and clustered boxplots, both for summaries of groups of cases and for separate variables.

**Simple Boxplot, Summaries for Groups of Cases**

Selecting this option will create a boxplot summarizing a single ordinal-level or higher variable within levels of a categorical variable.

1. Click **Simple**, then click **Summaries for groups of cases**, and then click **Define**. This brings up the **Define Simple Boxplot: Summaries for Groups of Cases** dialog box shown in Figure 18.1.
2. First, you have to select your **Variable**: of interest. Click the variable name on the left hand side of the dialog box, and then click the right arrow (→) next to the **Variable**:
3. Next, you need to indicate the **Category Axis**. This variable is a nominal or categorical variable that is used to group your participants. That is, a separate boxplot will be produced for each group of participants. Click the variable name on the left hand side of the dialog box, and then click the right arrow (→) next to the **Category Axis**.
4. You have the option to Label Cases by:. If selected, the value labels (or values, if no labels are provided in Variable View) of this variable can be used to label outliers or extreme cases on the plot. The most common variable used here is the ID variable (e.g., subject number). Again, this is optional and if you do not select this option, the outliers and extreme cases will be identified by row number.

5. Also, you have the option to panel your boxplot by a categorical variable. For example, you can present a separate boxplots for males and females stacked on top of one another (Rows:) or next to one another (Columns:).

6. Finally, you can click the Options... button in the upper right hand corner to bring up the Options dialog box. Here, you can indicate how you want to handle Missing Values or if you want to Display error bars.

7. When you are satisfied with your choices, click OK. SPSS will produce your boxplot in the output window similar to Figure 18.2. Notice the outliers (o) and extreme cases (*)!

---

**Figure 18.1**

**Figure 18.2**
Clustered Boxplot, Summaries for Groups of Cases

1. Click Clustered, then click Summaries for groups of cases, and then click Define. This brings up the Define Simple Boxplot: Summaries for Groups of Cases dialog box.
   a. The only difference between this dialog box and Figure 18.1 is the addition of the Define Clusters by: box where you will place the second categorical variable to group your participants. To do so, click the secondary categorical variable you want to use and click the right arrow (→) next to the Define Clusters by: box.

2. First, you have to select your Variable: of interest. Click the variable name on the left hand side of the dialog box, and then click the right arrow (→) next to the Variable: box.

3. Next, you need to indicate the Category Axis:. This variable is a nominal or categorical variable that is used to group your participants. That is, a separate boxplot will be produced for each group of participants. Click the variable name on the left hand side of the dialog box, and then click the right arrow (→) next to the Category Axis: box.

4. You have the option to Label Cases by:. If selected, the value labels (or values, if no labels are provided in Variable View) of this variable can be used to label outliers or extreme cases on the plot. The most common variable used here is the ID variable (e.g., subject number). Again, this is optional and if you do not select this option, the outliers and extreme cases will be identified by row number.

5. Also, you have the option to panel your boxplot by a categorical variable. For example, you can present a separate boxplots for males and females stacked on top of one another (Rows:) or next to one another (Columns:).

6. Finally, you can click the Options... button in the upper right hand corner to bring up the Options dialog box. Here, you can indicate how you want to handle Missing Values or if you want to Display error bars.

7. When you are satisfied with your choices, click OK. SPSS will produce your boxplot in the output window similar to Figure 18.3. Again, notice the outliers and extreme cases.

Figure 18.3
Simple Boxplot, Summaries of Separate Variables

1. Click Simple, then click **Summaries of separate variables**, and the click **Define**. This brings up the Define Simple Boxplot: Summaries for Separate Variables dialog box shown in Figure 18.4.

![Figure 18.4](image)

2. Because the plotted data will now represent individual variables, you must indicate what the **Boxes Represent**. Here, you will place the variables you want to summarize. Note, however, that the y-axis will remain to scale, thus, if you have two variables on drastically different scales, separate charts may be appropriate.

   a. Click the variables you want to use on the left hand side of the dialog box, and then click the right arrow (↵) next to the **Boxes Represent** box.

3. You have the option to **Label Cases by**. If selected, the value labels (or values, if no labels are provided in Variable View) of this variable can be used to label outliers or extreme cases on the plot. The most common variable used here is the ID variable (e.g., subject number). Again, this is optional and if you do not select this option, the outliers and extreme cases will be identified by row number.

4. Also, you have the option to panel your boxplot by a categorical variable. For example, you can present a separate boxplots for males and females stacked on top of one another (**Rows:**) or next to one another (**Columns:**).

5. Finally, you can click the **Options...** button in the upper right hand corner to bring up the **Options** dialog box. Here, you can indicate how you want to handle **Missing Values** or if you want to **Display error bars**.

6. When you are satisfied with your choices, click **OK**. SPSS will produce your boxplot in the output window similar to Figure 18.5. Notice the severe difference in scales between the two variables!
Clustered Boxplot, Summaries of Separate Variables

1. Click **Clustered**, then click **Summaries of separate variables**, and then click **Define**. This brings up the **Define Simple Boxplot: Summaries for Separate Variables** dialog box.
   
a. The only difference between this dialog box and the one presented in Figure 18.4 is that you now have the **Category Axis:** box where you will place a categorical variable to group your participants. To select this variable, click the variable name on the left hand side of the dialog box, and then click the right arrow (→) next to the **Category Axis:** box.

   b. Note, however, that the y-axis will remain to scale, thus, if you have two variables on drastically different scales, separate boxplots may be appropriate.

2. Because the plotted data will now represent individual variables, you must indicate what the **Boxes Represent:**. Here, you will place the variables you want to summarize. Note, however, that the y-axis will remain to scale, thus, if you have two variables on drastically different scales, separate charts may be appropriate.
   
a. Click the variables you want to use on the left hand side of the dialog box, and then click the right arrow (→) next to the **Boxes Represent:** box.

3. You have the option to **Label Cases by:**. If selected, the value labels (or values, if no labels are provided in Variable View) of this variable can be used to label outliers or extreme cases on the plot. The most common variable used here is the ID variable (e.g., subject number). Again, this is optional and if you do not select this option, the outliers and extreme cases will be identified by row number.

4. Also, you have the option to panel your boxplot by a categorical variable. For example, you can present a separate boxplots for males and females stacked on top of one another (**Rows:**) or next to one another (**Columns:**).
5. Finally, you can click the *Options*... button in the upper right hand corner to bring up the *Options* dialog box. Here, you can indicate how you want to handle *Missing Values* or if you want to *Display error bars*.

6. When you are satisfied with your choices, click **OK**. SPSS will produce your boxplot in the output window similar to Figure 18.6. Notice the severe difference in scales between the two variables!

![Figure 18.6](image-url)
Chapter 19
SCATTERPLOTS

Scatterplots are used to compare values on two or more continuous (i.e., interval- or ratio-level) variables. They are useful in diagnosing linearity issues. SPSS provides offers five different types of scatterplots. With the exception of Simple Dot and 3-D Scatter, all variables are paired and plotted on two dimensional x-y axes.

1. Click Graphs, then choose Legacy Dialogs, then click Scatter/Dot…. This brings up the Scatter/Dot dialog box.
2. Here, you have five options.
   a. Simple Scatter: The most basic scatterplot option. Here, you are allowed to compare two continuous variable pairs.
   b. Matrix Scatter: This option allows you to compare multiple continuous variables at the same time printed as in matrix of two-way scatterplots.
   c. Simple Dot: This option plots individual observations for one continuous variable. Note that there is only an x-axis. There are no values on the y-axis; therefore, the vertical position of a dot does not indicate a particular value.
   d. Overlay Scatter: Here, two or more continuous variable pairs are plotted. Each pair is indicated by a different marker, or marker color. It is similar to the matrix scatter only now the plots are on top of one another.
   e. 3-D Scatter: The option allows you to compare three continuous variables in 3-D space.

Next, I will provide you with descriptions for each of the scatterplots described above.

Simple Scatterplot

1. Click Simple Scatter and then click Define to bring up the Simple Scatterplot dialog box shown in Figure 19.1.
2. The top two boxes of this dialog box require you to select the variables that scale the Y Axis: and X Axis:. The choice of variable for each axis is arbitrary and only the appearance of the scatterplot will change by switching axes.
   a. Click the continuous variable you want to place on the y-axis and then click the right arrow ( ) next to the Y Axis: box.
   b. Click the continuous variable you want to place on the x-axis and then click the right arrow ( ) next to the X Axis: box.
3. The Set Markers by: box is optional and allows you to use a categorical variable to identify different participant groups within the same scatterplot. Each group will be marked by a different symbol in the scatterplot. For example, men may have green circles and women may have blue circles in the same scatterplot.
4. Next, you have the option to Label Cases by:. If selected, the value labels (or values, if no labels are provided in Variable View) of this variable will be used to identify individual data points in the scatterplot. The most common variable used here is the ID variable (e.g., subject number); however, if your sample size is large this option may clutter the scatterplot.

5. Next, you have the option to panel your scatterplot. A paneled scatterplot is a grid of scatterplots that share the same axes, but for a different group of one or more categorical variables. For example, you can present a separate boxplots for males and females stacked on top of one another (Rows:) or next to one another (Columns:).

6. Clicking the Titles... button in the upper right hand corner brings up the Titles dialog box. Here, you can provide the scatterplot with a Title, Subtitle: or Footnote.

7. Clicking the Options... button in the upper right hand corner brings up the Options dialog box, where you can indicate how you want to handle Missing Values or if you want to Display error bars.

8. Finally, you can use a Template from a previous scatterplot you have created. If you choose this option, only Step 2 is required.

9. When you are satisfied with your choices, click OK. SPSS will produce your scatterplot in the output window similar to Figure 19.2.
1. Click **Matrix Scatter** and then click **Define** to bring up the **Scatterplot Matrix** dialog box.
   a. The only difference between this dialog box and the one shown in Figure 19.1 is that the **Y Axis:** and **X Axis:** boxes have been replaced by the **Matrix Variables:** box. Here, you can enter as many continuous variables as you want. To do so, select the variable you want to include in the matrix scatterplot and then click the right arrow ( ) next to the **Matrix Variables:** box.

2. The **Set Markers by:** box is optional and allows you to use a categorical variable to identify different participant groups within the same scatterplot. Each group will be marked by a different symbol in the scatterplot. For example, men may have green circles and women may have blue circles in the same scatterplot.

3. Next, you have the option to **Label Cases by:**. If selected, the value labels (or values, if no labels are provided in **Variable View**) of this variable will be used to identify individual data points in the scatterplot. The most common variable used here is the ID variable (e.g., subject number); however, if your sample size is large this option may clutter the scatterplot.

4. Next, you have the option to panel your scatterplot. A paneled scatterplot is a grid of scatterplots that share the same axes, but for a different group of one or more categorical variables. For example, you can present a separate boxplots for males and females stacked on top of one another (**Rows:**) or next to one another (**Columns:**).

5. Clicking the **Titles...** button in the upper right hand corner brings up the **Titles** dialog box. Here, you can provide the scatterplot with a **Title, Subtitle:** or **Footnote.**

6. Clicking the **Options...** button in the upper right hand corner brings up the **Options** dialog box, where you can indicate how you want to handle **Missing Values** or if you want to **Display error bars.**
7. Finally, you can use a Template from a previous scatterplot you have created. If you choose this option, only Step 2 is required.

8. When you are satisfied with your choices, click OK. SPSS will produce your scatterplot matrix in the output window similar to Figure 19.3.

---

![Figure 19.3](image)

**Simple Dot Plot**

1. Click Simple Dot and then click Define to bring up the Define Simple Dot Plot dialog box.
   
   a. The only difference between this dialog box and the one shown in Figure 19.1 is that the Y Axis: and X Axis: boxes have been replaced by the X-Axis Variable: box. Here, you select any one variable by clicking the right arrow (→) next to the X-Axis Variable: box.

2. Next, you have the option to panel your scatterplot. A paneled scatterplot is a grid of scatterplots that share the same axes, but for a different group of one or more categorical variables. For example, you can present a separate boxplots for males and females on top of one another (Rows:) or next to one another (Columns:).

3. Clicking the Titles... button in the upper right hand corner brings up the Titles dialog box. Here, you can provide the scatterplot with a Title, Subtitle: or Footnote.

4. Clicking the Options... button in the upper right hand corner brings up the Options dialog box, where you can indicate how you want to handle Missing Values or if you want to Display error bars.

5. Finally, you can use a Template from a previous scatterplot you have created. If you choose this option, only Step 2 is required.

6. When you are satisfied with your choices, click OK. SPSS will produce your dot chart in the output window similar to Figure 19.4.
1. Click **Overlay Scatter** and then click **Define** to bring up the *Overlay Scatterplot* dialog box shown in Figure 19.5.

2. Here, you are required to have at least one continuous variable pair. You can see in the *Y-X Pairs:* box that *Pair 1* requires both a *Y Variable* and an *X Variable.*
a. Click the continuous variable you want to place on the y-axis and then click the right arrow (\(\rightarrow\)) next to the Y-X Pairs: box.

b. Click the continuous variable you want to place on the x-axis and then click the right arrow (\(\rightarrow\)) next to the Y-X Pairs: box.

c. Repeat for as many variable pairs as you want.

d. You can switch variables between the x- and y-axis by click the horizontal, double-headed arrow (\(\leftrightarrow\)). Further, you can move variable pairs up or down in priority by clicking the row under the Pair column and then clicking either the up (\(\uparrow\)) or down arrow (\(\downarrow\)).

3. Next, you have the option to Label Cases by:. If selected, the value labels (or values, if no labels are provided in Variable View) of this variable will be used to identify individual data points in the scatterplot. The most common variable used here is the ID variable (e.g., subject number); however, if your sample size is large this option may clutter the scatterplot.

4. Next, you have the option to panel your scatterplot. A paneled scatterplot is a grid of scatterplots that share the same axes, but for a different group of one or more categorical variables. For example, you can present a separate boxplots for males and females stacked on top of one another (Rows:) or next to one another (Columns:).

5. Clicking the Titles... button in the upper right hand corner brings up the Titles dialog box. Here, you can provide the scatterplot with a Title, Subtitle: or Footnote.

6. Clicking the Options... button in the upper right hand corner brings up the Options dialog box, where you can indicate how you want to handle Missing Values or if you want to Display error bars.

7. Finally, you can use a Template from a previous scatterplot you have created. If you choose this option, only Step 2 is required.

8. When you are satisfied with your choices, click OK. SPSS will produce your overlay scatterplot in the output window similar to Figure 19.6.

![Figure 19.6](image)
3-D Scatterplot

1. Click 3-D Scatter and then click Define to bring up the 3-D Scatterplot dialog box shown in Figure 19.7.

![Figure 19.7](image)

2. The top three boxes require you to select the variables that scale the \( Y \) Axis, \( X \) Axis, and \( Z \) Axis. The choice variable for each axis is arbitrary and only the appearance of the scatterplot will change by switching axes.
   a. Click the continuous variable you want to place on the \( y \)-axis and then click the right arrow (\( \rightarrow \)) next to the \( Y \) Axis: box.
   b. Click the continuous variable you want to place on the \( x \)-axis and then click the right arrow (\( \rightarrow \)) next to the \( X \) Axis: box.
   c. Click the continuous variable you want to place on the \( x \)-axis and then click the right arrow (\( \rightarrow \)) next to the \( Z \) Axis: box.

3. The Set Markers by: box is optional and allows you to use a categorical variable to identify different participant groups within the same scatterplot. Each group will be marked by a different symbol in the scatterplot. For example, men may have green circles and women may have blue circles in the same scatterplot.

4. Next, you have the option to Label Cases by: If selected, the value labels (or values, if no labels are provided in Variable View) of this variable will be used to identify individual data points in the scatterplot. The most common variable used here is the ID variable (e.g., subject number); however, if your sample size is large this option may clutter the scatterplot.
5. Next, you have the option to panel your scatterplot. A paneled scatterplot is a grid of scatterplots that share the same axes, but for a different group of one or more categorical variables. For example, you can present a separate boxplots for males and females stacked on top of one another (Rows:) or next to one another (Columns:).

6. Clicking the Titles... button in the upper right hand corner brings up the Titles dialog box. Here, you can provide the scatterplot with a Title, Subtitle: or Footnote.

7. Clicking the Options... button in the upper right hand corner brings up the Options dialog box, where you can indicate how you want to handle Missing Values or if you want to Display error bars.

8. Finally, you can use a Template from a previous scatterplot you have created. If you choose this option, only Step 2 is required.

9. When you are satisfied with your choices, click OK. SPSS will produce your 3-D scatterplot in the output window similar to Figure 19.8.

10. In your output window, if you double click the 3-D scatterplot you will bring up the Chart Editor dialog box.
   a. Click Edit and then click 3-D Rotation. This will bring up the 3-D Rotation dialog box.
   b. Place your mouse over the 3-D scatterplot, the mouse pointer will become a hand with a pointed index finger. Click and hold the left mouse button and move the mouse around to rotate the scatterplot until the image is to your liking.
   c. Close the 3-D Rotation and Chart Editor dialog boxes. In your output window, the 3-D scatterplot should appear as you set it in Step 2.
Section IV

Statistical Tests of Association

I begin the statistical analysis portion of the text with correlation and linear regression analyses as well as their nonparametric alternatives when applicable. Linear regression is an extension of Pearson’s product-moment correlation (Chapter 20) and is the most direct application of the general linear model. Developing an understanding of linear regression prior to discussing $t$-tests, ANOVA, or ANCOVA (i.e., special cases of the general linear model) should make the special cases and their associated assumptions clearer and more easily understandable.

The remaining Chapters of this book each include a small dataset for you to practice your data entry and coding skills. I chose to include a dataset for you to use rather than refer to abstract examples so that you will be able to perform the analyses and replicate all output presented. This will allow you to verify you completed the analysis correctly. Similar to above, while all portions of the menus for each analysis are described in detail, the **bolded** instructions are minimally required to replicate the output provided.

Finally, at the end of each Chapter you will be presented with an example results section in APA format. This should give you a pretty good idea of what will be minimally required when reporting results for your future posters or manuscripts.
Chapter 20
PEARSON'S PRODUCT-MOMENT CORRELATION

Pearson’s product-moment correlation (or simply, Pearson’s correlation or Pearson’s $r$) is one of the most commonly used correlation measures. It measures the direction and strength of a linear relationship (or association) between two variables measured on continuous scales. It is important to note that Pearson’s $r$ is dimensionless (think, $z$-scores), so the actual scale of the two variables is irrelevant.

Pearson’s $r$ ranges from -1 to +1, with $r$ of 0 indicating no relationship. That is, as the correlation coefficient approaches -1 or 1, the relationship becomes stronger. A positive correlation indicates that as the values of one variable increase so do the values of the other variable, whereas a negative correlation indicates that as the values of one variable increase, the values of the other variable decrease. The absolute size of the correlation coefficient is directly related to how closely the data points fall along a best-fit line (also known as the regression line, discussed in later Chapters). You should note that if the best-fit line is perfectly horizontal, or perfectly vertical, the correlation coefficient cannot be calculated because one of your variables is a constant. That is, constants do not have variances, so during computation, you will essentially divide by zero and $r$ will be undefined. The statistical test of Pearson’s $r$ indicates whether the correlation of two variables is different from zero; thus, a statistically significant $r$ indicates the slope of the linear relationship is not horizontal.

Finally, it is important to point out that the relationship between two variables can be assessed visually by plotting the data on a scatterplot (Chapter 19). In fact, this practice is highly recommended, as you will be able to identify any outliers or curvilinearity, which violate two assumptions discussed below.

As an example, consider a study designed to test the association between minutes of sleep per night and body weight in adults. Here, you measure the number of minutes each participant sleeps per night and obtain their weight in pounds. The collected data is provided below.

<table>
<thead>
<tr>
<th>ID</th>
<th>Minutes</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>480</td>
<td>154</td>
</tr>
<tr>
<td>2</td>
<td>453</td>
<td>169</td>
</tr>
<tr>
<td>3</td>
<td>349</td>
<td>225</td>
</tr>
<tr>
<td>4</td>
<td>297</td>
<td>212</td>
</tr>
<tr>
<td>5</td>
<td>551</td>
<td>169</td>
</tr>
<tr>
<td>6</td>
<td>467</td>
<td>191</td>
</tr>
<tr>
<td>7</td>
<td>521</td>
<td>167</td>
</tr>
<tr>
<td>8</td>
<td>402</td>
<td>202</td>
</tr>
<tr>
<td>9</td>
<td>351</td>
<td>236</td>
</tr>
<tr>
<td>10</td>
<td>315</td>
<td>254</td>
</tr>
</tbody>
</table>

Following the data entry procedures described in Chapters 1 and 2, label the first variable ID, the second Minutes, and the third Weight. Enter your data appropriately.
Assumptions

Because Pearson’s $r$ is a parametric statistical test, it has numerous assumptions. First, both variables must be measured on at least an interval-level scale. One additional design assumption is independence. That is, each participant can contribute one, and only one, measurement value to each variable.

Additional assumptions include absence of bivariate outliers, linearity, and homoscedasticity. Each of these assumptions pertain directly to the best-fit line and are evaluated by a simple scatterplot (Chapter 19). Bivariate (i.e., two variable) outliers are defined as data points that are disconnected from the rest of the data points in the scatterplot. Outlying values can have undue influence on the best-fit line and can lead to incorrect inference. If outliers are identified you have several options. You can remove them from analysis through deletion (do not forget to describe them in your results section), transform the dependent variable (not advised), or use a nonparametric alternative that uses ranked data instead of the actual, raw scores (Chapter 21). Linearity requires that the relationship between the variables be best approximated by a straight line. That is, a straight best-fit line describes the data better than a curved best-fit line. Finally, homoscedasticity requires that the variability of data points around the best-fit line be constant. Your scatterplot should appear rectangular or (at minimum) oval shaped.

Analysis

Assuming independence, absence of bivariate outliers, linearity, and homoscedasticity, to conduct Pearson’s $r$:

1. Click Analyze, then choose Correlate, and finally click Bivariate…. This brings up the Bivariate Correlations dialog box, shown in Figure 20.1.
2. You will notice immediately all the variables in your dataset are listed on the left side of this dialog box. To include variables in the analysis all you need to do is move the variables you want into the Variables: box. You can include as many variables as you want; however, including a large number of variables creates an unruly correlation matrix that can be difficult to read.
   a. Click to highlight the Minutes variable and then click the right arrow ( ) next to the Variables: box.
   b. Click to highlight the Weight variable and then click the right arrow ( ) next to the Variables: box.
3. Click the Options... button to bring up the Bivariate Correlations: Options dialog box. Under the Statistics section, you have to option to display the Means and standard deviations for each variable as well as the Cross-product deviations and covariances. Further, you can tell SPSS how you want to handle Missing Values. When you are satisfied, click Continue.
4. Next, you must identify which of the three Correlation Coefficients you want to use. The default is Pearson. Note that the Kendall’s tau-b and Spearman options produce nonparametric correlations. Spearman’s rho will be discussed in Chapter 21.
5. In the Tests of Significance section, you can also indicate whether you want to conduct One-tailed or Two-Tailed statistical tests.
6. Finally, you can choose to Flag significant correlations. This box is checked by default and I recommend you keep it checked. SPSS flags significant correlations by providing asterisks next to the r value in the correlation matrix.
7. That’s it! Click OK to conduct Pearson’s r.

Output

When you click OK, SPSS will produce an Output screen displaying your results, which should be identical to those presented in Figure 20.2. Click the Output window to view your results (if it does not pop up automatically).

<table>
<thead>
<tr>
<th>Correlations</th>
<th>Minutes</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minutes</td>
<td>Pearson Correlation</td>
<td>-.772**</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>1</td>
<td>.000</td>
</tr>
<tr>
<td>N</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Weight</td>
<td>Pearson Correlation</td>
<td>1</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>-.772**</td>
<td>.000</td>
</tr>
<tr>
<td>N</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 20.2
The only table you will see in your output is the correlation matrix, provided in a table titled *Correlations*. This matrix displays the correlations between all possible pairs of variables included in your analysis. Each cell of the matrix contains information about the relationship between the two variables included at the intersection of the specific column and row. Note that the correlation values on the diagonal are all 1 because a variable will correlate perfectly with itself, and the values above and below the diagonal are identical. The top number in each cell is the *Pearson Correlation coefficient*, the middle number is the *p*-value (*Sig. 2-tailed*), and the final value is your sample size (*N*) for that specific correlation. Because you chose to flag significant correlation in Step 6 above, all statistically significant correlations are flagged with asterisks. Note that one asterisk (*) indicates *p* < .05 and two asterisks (**) indicate *p* < .01.

**Interpretation**

Interpretation of your results is completed by simply evaluating the correlation matrix. For this example, the correlation (-.772) was statistically significant. For reporting purposes, note that the degrees of freedom for Pearson’s *r* is the number of subjects in the analysis (*N*) minus 2. For this example, degrees of freedom was 18 (i.e., 20-2).

**Example Results Section**

Prior to analysis, no violation of independence was indicated. Further, a scatterplot indicated no bivariate outliers, curvilinearity, or heteroscedasticity.

The result of Pearson’s product-moment correlation analysis indicated a statistically significant negative association between minutes of sleep per night and body weight in pounds, *r*(18) = -.772, *p* < .05, where minutes of sleep per night increased as body weight decreased.
Chapter 21

SPEARMAN’S RANK-ORDER CORRELATION COEFFICIENT

Spearman’s rank-order correlation coefficient, also known as Spearman’s rho ($\rho$), is the nonparametric alternative to Pearson’s $r$. This correlation is used when the relationship between two continuous variables is monotonic (i.e., increasing or decreasing nonlinear) or when the variables are measured on an ordinal scale.

Spearman’s $\rho$ is based on rank ordered data as opposed to the actual values. That is, the values within each variable are ranked, with the highest value receiving the highest rank and the lowest value receiving the lowest rank. Because of this fact, Spearman’s correlation can handle data containing outliers. That is, disconnected values are a nonissue because the highest value (no matter how high) simply receives the highest rank and the lowest value (no matter how low) receives the lowest rank.

Finally, the values of Spearman’s $\rho$ range from -1 to 1, with 0 indicating no association. Thus, a statistically significant $\rho$ indicates that the association is significantly different from 0, and as $\rho$ approaches -1 or 1, the association between the two variables becomes stronger.

To show the similarity between Spearman’s $\rho$ and Pearson’s $r$, we will reconsider the example data presented in Chapter 20. As a reminder, you are conducting a study designed to test the association between minutes of sleep per night and body weight in adults. You have measured the number of minutes each participant sleeps per night and obtained their body weight in pounds. The collected data is provided below.

<table>
<thead>
<tr>
<th>ID</th>
<th>Minutes</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>480</td>
<td>154</td>
</tr>
<tr>
<td>2</td>
<td>453</td>
<td>169</td>
</tr>
<tr>
<td>3</td>
<td>349</td>
<td>225</td>
</tr>
<tr>
<td>4</td>
<td>297</td>
<td>212</td>
</tr>
<tr>
<td>5</td>
<td>551</td>
<td>169</td>
</tr>
<tr>
<td>6</td>
<td>467</td>
<td>191</td>
</tr>
<tr>
<td>7</td>
<td>521</td>
<td>167</td>
</tr>
<tr>
<td>8</td>
<td>402</td>
<td>202</td>
</tr>
<tr>
<td>9</td>
<td>351</td>
<td>236</td>
</tr>
<tr>
<td>10</td>
<td>315</td>
<td>254</td>
</tr>
<tr>
<td>11</td>
<td>500</td>
<td>146</td>
</tr>
<tr>
<td>12</td>
<td>461</td>
<td>200</td>
</tr>
<tr>
<td>13</td>
<td>373</td>
<td>181</td>
</tr>
<tr>
<td>14</td>
<td>399</td>
<td>192</td>
</tr>
<tr>
<td>15</td>
<td>426</td>
<td>119</td>
</tr>
<tr>
<td>16</td>
<td>342</td>
<td>232</td>
</tr>
<tr>
<td>17</td>
<td>464</td>
<td>159</td>
</tr>
<tr>
<td>18</td>
<td>535</td>
<td>134</td>
</tr>
<tr>
<td>19</td>
<td>403</td>
<td>184</td>
</tr>
<tr>
<td>20</td>
<td>515</td>
<td>129</td>
</tr>
</tbody>
</table>

Following the data entry procedures described in Chapters 1 and 2, label the first variable $ID$, the second $Minutes$, and the third $Weight$. Enter your data appropriately.

**Assumptions**

Spearman’s $\rho$ has two key assumptions— independence and linearity of ranks. Independence requires participants be measured once, and only once, on each variable. Linearity of ranks is trickier to evaluate because we are not talking about linearity of the actual raw values on each variable, but instead the ranked values. Linearity, however, still requires that the relationship between the ranked variables be best approximated by a straight line. Evaluating this assumption is discussed in detail below.
Evaluating the Linearity of Ranks Assumption

1. Click **Transform**… and then click **Rank Cases**… to bring up the **Rank Cases** dialog box.

2. You will notice that all of the variables in your dataset are listed on the left hand side of this dialog box.
   a. Click the **Minutes** variable and then click the right arrow ( ) next to the **Variable(s):** box.
   b. Click the **Weight** variable and then click the right arrow ( ) next to the **Variable(s):** box.

3. Under the **Assign Rank 1 to** section, make sure **Smallest value** remains checked.

4. Click **OK**.

You will notice that two new variables were created in your dataset (i.e., **RMinutes** and **RWeight**). Now, simply create a scatterplot of these two variables to evaluate the linearity assumptions (Chapter 19).

It is important to note that after this assumption has been assured, you will **use the original variables in the analysis**. The Spearman’s ρ procedure in SPSS routinely ranks the actual, raw values during analysis.

Analysis

Assuming independence and linearity of ranked values, to calculate Spearman’s ρ:

1. Click **Analyze**, then choose **Correlate**, and finally click **Bivariate**…. This brings up the **Bivariate Correlations** dialog box.

2. You will notice immediately all the variables in your dataset are listed on the left side of this dialog box. To include variables in the analysis all you need to do is move the variables you want into the **Variables:** box. You can include as many variables as you want; however, including a large number of variables creates an unruly correlation matrix that is potentially difficult to read.
   a. Click to highlight the **Minutes** variable and then click the right arrow ( ) next to the **Variables: box**.
   b. Click to highlight the **Weight** variable and then click the right arrow ( ) next to the **Variables:** box.

3. Click the **Options**… button to bring up the **Bivariate Correlations: Options** dialog box. Under the **Statistics** section, you have to option to display the **Means and standard deviations** as well as the **Cross-product deviations and covariances**. Further, you can tell SPSS how you want to handle **Missing Values**. When you are satisfied, click **Continue**.
4. Next, you must identify which of the three Correlation Coefficients you want to use. Click the Spearman checkbox, and then click to uncheck the default Pearson checkbox.

5. In the Tests of Significance section, you can also indicate whether you want to conduct One-tailed or Two-Tailed statistical tests.

6. Finally, you can choose to Flag significant correlations. This box is checked by default and I recommend you keep it checked. SPSS flags significant correlations by providing asterisks next to the \( \rho \) value in the correlation matrix.

7. That’s it! Click OK to conduct Spearman’s \( \rho \).

**Output**

When you click OK, SPSS will produce an Output screen displaying your results, which should be identical to those presented in Figure 21.1. Click the Output window to view your results (if it does not pop up automatically).

The only table you will see in your output is the correlation matrix, provided in a table titled Correlations. This matrix displays the correlations between all possible pairs of variables included in your analysis. Each cell of the matrix contains information about the relationship between the two variables included at the intersection of the specific column and row. Note that the correlation values on the diagonal are all 1 because a variable will correlate perfectly with itself, and the values above and below the diagonal are identical. The top number in each cell is the value of Spearman’s \( \rho \) (Correlation Correlation), the middle number is the \( p \)-value (Sig. 2-tailed), and the final value is your sample size (\( N \)) for that specific correlation. Because you chose to flag significant correlation in Step 6 above, all statistically significant correlations are flagged with asterisks. Note one asterisk (*) indicates \( p < .05 \), two asterisks (**) indicate \( p < .01 \).

<table>
<thead>
<tr>
<th>Correlations</th>
<th>Minutes</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spearman's rho</td>
<td>1.000</td>
<td>-.780**</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>.</td>
<td>.000</td>
</tr>
<tr>
<td>N</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Weight</td>
<td>-.780**</td>
<td>1.000</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>.</td>
</tr>
<tr>
<td>N</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 21.1
**Interpretation**

Interpretation of your results is completed by simply evaluating the correlation matrix. For this example, the Spearman’s $\rho$ (-.780) was statistically significant. Similar to Pearson’s $r$, for reporting purposes degrees of freedom is the number of subjects in the analysis ($N$) minus 2. For this example, degrees of freedom was 18 (i.e., 20-2).

**Example Results Section**

Prior to analysis, no violation of independence or linearity of ranks was indicated.

The result of a Spearman’s rho analysis indicated a statistically significant negative association between minutes of sleep per night and body weight in pounds, $\rho(18) = -.780, p < .05$, where minutes of sleep per night increased as body weight decreased.
Chapter 22
SIMPLE LINEAR REGRESSION

As stated in Chapter 20, the magnitude of the correlation coefficient is related directly to the strength of the linear relationship. This linear relationship is described by the slope of the best-fit line (from here on termed the regression line). That is, how changes in your independent variable affect your dependent variable.

Similar to Pearson’s $r$, the statistical test in simple linear regression is whether the slope of the regression line is statistically different from zero. However, simple linear regression takes Pearson’s $r$ a step further by actually predicting values of the dependent variable. The algebraic linear regression equation, $Y = a + BX + e$, is based on the regression line’s intercept with the $y$-axis ($a$) and slope ($B$), with prediction of the dependent variable ($Y$) based on the value of the independent variable ($X$). Note that the value of the correlation between the dependent and independent variables is incorporated into the equation for slope. This equation is the simplest form of the general linear model.

In addition, predicted values are rarely identical to the actual observed values. The difference between the predicted and observed value is referred to as the residual value (i.e., $e$ in the equation above). Residual values are calculated for all participants included in the analysis and most of the primary assumptions of linear regression are related directly to these residual values. That is, the residuals are assumed normally distributed, with constant variance, and independent of one another.

The primary statistical test used in simple linear regression is an $F$ test indicating whether the independent variable explains a statistically significant amount of variance in the dependent variable. If the $F$ test is statistically significant, so is the slope of the independent variable. The amount of variance explained is quantified by the coefficient of determination. This coefficient is calculated by squaring the correlation between the independent and dependent variables, and is referred to as $r^2$, or identically as $R^2$. The coefficient will be presented as a proportion ranging from 0 to 1, or as a percentage. In either case, higher values indicate more reliable prediction.

Interpretation of regression slopes differ if your independent variable is continuous or categorical. As an example of a continuous independent variable, say you want to know whether body weight in pounds significantly predicts minutes of sleep per night. As an example of a categorical independent variable, say you want to determine whether participants randomized to take your new sleep treatment sleep more minutes per night compared to participants who take the leading over-the-counter treatment. This analysis is identical to ANOVA described in Chapter 28! Each question will be analyzed separately. The collected data is presented below:

<table>
<thead>
<tr>
<th>ID</th>
<th>Tx</th>
<th>Wt</th>
<th>Minutes</th>
<th>ID</th>
<th>Tx</th>
<th>Wt</th>
<th>Minutes</th>
<th>ID</th>
<th>Tx</th>
<th>Wt</th>
<th>Minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>181</td>
<td>270</td>
<td>11</td>
<td>0</td>
<td>169</td>
<td>325</td>
<td>21</td>
<td>1</td>
<td>179</td>
<td>490</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>200</td>
<td>299</td>
<td>12</td>
<td>0</td>
<td>170</td>
<td>327</td>
<td>22</td>
<td>1</td>
<td>164</td>
<td>479</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>205</td>
<td>326</td>
<td>13</td>
<td>0</td>
<td>206</td>
<td>333</td>
<td>23</td>
<td>1</td>
<td>159</td>
<td>479</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>191</td>
<td>359</td>
<td>14</td>
<td>0</td>
<td>148</td>
<td>367</td>
<td>24</td>
<td>1</td>
<td>140</td>
<td>453</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>226</td>
<td>297</td>
<td>15</td>
<td>0</td>
<td>192</td>
<td>387</td>
<td>25</td>
<td>1</td>
<td>173</td>
<td>389</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>205</td>
<td>315</td>
<td>16</td>
<td>1</td>
<td>162</td>
<td>435</td>
<td>26</td>
<td>1</td>
<td>126</td>
<td>530</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>151</td>
<td>402</td>
<td>17</td>
<td>1</td>
<td>178</td>
<td>502</td>
<td>27</td>
<td>1</td>
<td>149</td>
<td>521</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>167</td>
<td>378</td>
<td>18</td>
<td>1</td>
<td>138</td>
<td>515</td>
<td>28</td>
<td>1</td>
<td>186</td>
<td>469</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>192</td>
<td>388</td>
<td>19</td>
<td>1</td>
<td>187</td>
<td>405</td>
<td>29</td>
<td>1</td>
<td>120</td>
<td>500</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>199</td>
<td>430</td>
<td>20</td>
<td>1</td>
<td>156</td>
<td>472</td>
<td>30</td>
<td>1</td>
<td>135</td>
<td>435</td>
</tr>
</tbody>
</table>
Following the data entry procedures described in Chapters 1 and 2, label the first variable ID, the second Tx (0 = OTC; 1 = treatment), the third Wt for body weight in pounds, and the fourth variable Minutes. Enter your data appropriately.

** Assumptions  

The assumptions of simple linear regression vary slightly depending on whether you are using a continuous or categorical independent variable. The first two assumptions—Independence and normality of residuals—are applicable to all independent variables.

Independence of residuals requires the residuals to be uncorrelated across participants. Remember, a residual is the difference between predicted and actual value of the dependent variable, so participants will have their own residual value. Independence of residuals is technically a design issue that requires no repeated measurement or clustering. It can be assured by using appropriate random sampling and random assignment techniques.

Normality of residuals is evaluated during analysis by requesting a histogram and P-P plot of standardized residuals. Note that a P-P plot, or probability-probability plot, is interpreted identically to the Q-Q plot discussed above in Chapter 15. If the histogram is normally distributed and the P-P plot follows the diagonal reference line, normality is assured.

Three additional assumptions for continuous independent variables include absence of bivariate outliers, linearity, and homoscedasticity of residuals. Bivariate (i.e., two variable) outliers are identified by data points that are disconnected from the rest of the data points in the scatterplot. Outlying values can severely influence the regression line and can lead to incorrect inference. If outliers are identified you have two options—remove them from analysis through deletion (do not forget to describe them in your results section) or transform the dependent variable (not advised). Notice that no assumption is made about the distribution or outliers within the independent variable specifically. That is, the outlying value is always in relation to the dependent variable, which is why it is identified via scatterplot.

Linearity requires the relationship between the independent variable and dependent variable to be best approximated by a straight regression line. This assumption is evaluated prior to analysis using a scatterplot (Chapter 19).

Homoscedasticity of residuals requires the residuals to have a constant variance across all predicted values of the dependent variable. This assumption is evaluated during the analysis by requesting a scatterplot of the standardized residuals and the standardized predicted values. If this scatterplot has a roughly rectangular shape the assumption is satisfied.

The only additional assumption for a categorical independent variable is homogeneity of variance. Homogeneity of variance requires that the dependent variable have roughly the same variance within all groups. This assumption is tested prior to analysis using Levene’s test (Chapter 28), where a non-statistically significant result (i.e., $p > .05$) indicates the variances are not statistically different, thereby satisfying the assumption. If a violation is indicated, however, the probability of committing a type I error increases, so a more conservative alpha level should be used (e.g., .01 instead of .05). You should note that with a categorical independent variable, homogeneity of variance and homoscedasticity are closely related. In fact, a violation of homogeneity of variance goes hand-in-hand with a violation of homoscedasticity. Thus, if homogeneity of variance is satisfied, homoscedasticity of residuals usually is too.
**Continuous Independent Variable**

**Analysis**

Assuming independence of residuals and linearity, to conduct a simple linear regression on the data above using *Minutes* as the dependent variable and body weight (*Wt*) as the continuous independent variable:

1. Click **Analyze**, then choose **Regression**, and finally, click **Linear**… to bring up the *Linear Regression* dialog box shown in Figure 22.1.

2. You will immediately see all of the variables in your dataset listed on the left hand side of the dialog box. For a linear regression analysis, you will be primarily concerned with the **Dependent:** and **Independent(s):** dialog boxes. The variable you place in either of these boxes is relatively self-explanatory.
   
a. Because *Minutes* is your dependent variable, click to highlight *Minutes* and then click the right arrow (→) next to the **Dependent:** box.

b. Body weight in pounds is your independent variable, so click to highlight *Wt* and then click the right arrow (→) next to the **Independent(s):** box.

3. You will also notice four buttons on the right hand side of this dialog box. Note that the **Bootstrap**… button is only available if you have the Exact Statistics add-on. Click the **Statistics**… button to bring up the *Linear Regression: Statistics* dialog box. In this dialog box, you have several descriptive and correlation options as well as options for your **Regression Coefficients** and **Residuals**.
   
a. On the right hand side of this box are five checkboxes. **Model fit** is selected by default. This option prints the several fit statistics including multiple $R$ (just the absolute value of Pearson’s $r$ in simple linear regression), $R^2$, and adjusted $R^2$ as well as the standard error of estimate and the $F$-test for your overall model. The $R^2$ square change checkbox is only useful when you have more than one independent
variable. The **Descriptives** option provides the number of participants with valid values, mean, and standard deviation for each variable in the analysis. The *Part and partial correlations* and *Collinearity diagnostics* option is useful only when you have more than one independent variable.

b. Under the **Regression Coefficients** section, the **Estimates** option is checked by default. You will want to leave this option checked because it requests the majority of the relevant output regarding your independent variable, which includes the unstandardized and standardized regression coefficients, standard error as well as the *t*-value and significance test. The **Confidence intervals** option provides the confidence interval around your unstandardized regression coefficient at the specific Level(%): you choose. Finally, you can choose to print the **Covariance matrix**, but this option does not provide additional useful information when you only have one independent variable (that is, it only provides the variance of your independent variable and its correlation with itself).

c. Under the **Residuals** section, you can choose to conduct the **Durbin-Watson** test assessing the independence of residuals assumption against a 1st order autoregressive pattern. Note that if the residuals are correlated, but not 1st order autoregressive, you will have violated the assumption, but passed this statistical check. You can also request **Casewise diagnostics** printing the standardized residual value, value of the dependent variable, predicted value, and unstandardized residual. You can print this information only for **Outliers outside:** of a specified number of standard deviations or for **All cases**.

When you are satisfied with your selections, click **Continue**.

4. Click the **Plots...** button. This option is required to check the normality and homoscedasticity of residuals assumptions. On the left hand side of the dialog box you will see a list of variable names which include the raw dependent variable (**DEPENDNT**), standardized predicted value (**ZPRED**), adjusted predicted value (**ADJPRED**), and four residual values including the standardized residual (**ZRESID**), deleted residuals (**DRESID**), Studentized residuals (**SRESID**), and Studentized deleted residuals (**SDRESID**). To evaluate homoscedasticity you need to plot the standardized predicted against the standardized residual values. Note that you can include as many combinations you want by clicking the **Next** button under the **Scatter** section.

a. Click to select the **ZRESID** variable, and then click the right arrow (→) next to the **Y:** box in the **Scatter** section.

b. Click to select the **ZPRED** variable, and then click the right arrow (→) next to the **X:** box in the **Scatter** section.

Under the **Standardized Residual Plots** section, you have the option to print both a **Histogram** and a **Normal probability plot** of your standardized residuals. Both of these options are used to evaluate the normality of residuals assumption. When you are satisfied with your choices, click **Continue**.

5. Click the **Save...** button. Here, you have the option to save many different values to your dataset. That is, any option you select in this dialog box will not print to your output, but instead a new variable will be created in your dataset.
a. Under the *Predicted Values* section, you have the option to save the *Unstandardized* or raw predicted value, the *Standardized* predicted value (think, *z*-scores), the *Adjusted* predicted value which is the predicted value when the participant is excluded from analysis, or the *S.E. of mean predictions* which is the standard deviation of the average value of the dependent variable for participants that have the same value on the independent variable.

b. Under the *Residuals* section, you can save up to five different residual values. The *Unstandardized* residual is the actual difference in raw units, whereas the *Standardized* residual are *z*-scored residuals. A *Studentized* residual is the residual value divided by a varying standard deviations based on the distance of the individual’s independent variable value from the mean of the independent variable. A *Deleted* residual is the residual when the participant has been excluded from the analysis and is the difference between the dependent variable and the adjusted predicted value. Finally, a *Studentized deleted* residual is the deleted residual value divided by its standard error. It is used to indicate how much of a difference eliminating a participant makes on its own prediction.

c. The options under the *Distances* section are used to identify participants who have unusual independent variable combinations. Combinations imply more than one independent variable, so these options are discussed in Chapter 23.

d. The options under the *Influence Statistics* section provide indicators of what would happen a participant was excluded from analysis. *DfBeta(s)* are the difference in standardized regression coefficient (i.e., *beta*) if the participant was removed from analysis with *Standardized DfBeta(s)* being simply *z*-scored DfBetas. *DfFit* is the change in the predicted value if the participant is excluded from analysis, and *Standardized DfFit* are simply *z*-scored DfFit values. Finally, the *Covariance ratio* is the ratio of the determinant of the covariance matrix with a particular participant excluded from analysis to the determinant of the covariance matrix with the participant included; thus, ratios closer to 1 indicate little change.

e. In the *Prediction Intervals* section, you can calculate the *Confidence Interval*: around your predicted values. The *Mean* option produces confidence intervals around the mean predicted response, whereas the *Individual* produces intervals around the value of the dependent variable.

f. Selecting *Create coefficient statistics* will save the regression coefficients to a new dataset (you must provide a *Dataset name:* ) or a completely new data file (click *File...* to locate the already existing file).

g. Finally, you can *Export model information to XML file*. Click *Browse...* to locate the already existing file. This option will save parameter estimates, and optionally the covariance matrix (clicking the checkbox is required) in XML format.

When you are satisfied with your selections, click *Continue*.

6. Click the *Options* button. The *Stepping Method Criteria* option is only useful if you are using stepwise regression (oh, and never use stepwise regression). In addition, you can *Include constant in equation*, where unchecking this option forces the regression
equation through the origin, which is rare; thus, it is wise to leave this option checked. Finally, you can tell SPSS how you want to handle Missing Values. When you are satisfied with your selections, click **Continue**.

7. That’s it! Click **OK** to conduct the analysis.

**Output**

When you click **OK**, SPSS will produce an **Output** screen displaying your results. Click the **Output** window to view your results (if it does not pop up automatically).

The first table is titled **Descriptive Statistics**, which contains the Mean, standard deviation (**Std. Deviation**), and sample size (**N**) for your dependent and independent variables. Next, is a table titled **Correlations**, which contains the correlation between your dependent and independent variables. Both tables are presented in Figure 22.2.

![Descriptive Statistics table](image)

<table>
<thead>
<tr>
<th>Variables Entered/Removed</th>
<th>Minutes</th>
<th>Volt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Correlation</td>
<td>Minutes</td>
<td>Volt</td>
</tr>
<tr>
<td>Std. (1-tailed)</td>
<td>Minutes</td>
<td>Volt</td>
</tr>
<tr>
<td>N</td>
<td>Minutes</td>
<td>Volt</td>
</tr>
</tbody>
</table>

**Figure 22.2**

The **Variables Entered/Removed** table is irrelevant for simple linear regression. It is used for hierarchical regression or stepwise regression. Both of these procedures enter variables in blocks or sets, and this table would indicate which variables were included in a particular analysis step.

The next table, titled **Model Summary**, is shown in Figure 22.3. This table contains the correlation between your two variables (**R**). Note that because we only have one IV, this value is the absolute value of Pearson’s **r** from Figure 22.2. It also contains your coefficient of determination, labeled **R Square**, which estimates the proportion of variance explained in your dependent variable by your independent variable. **Adjusted R Square** is the R Square value adjusted for the degrees of freedom lost during the regression, and is a more accurate estimate of your coefficient of determination. Finally, the standard error of estimate (**Std. Error of the Estimate**) measures the overall accuracy of your regression model. It is the square root of the mean square residual found the ANOVA table, discussed next. Remember, mean square is just another term for variance.

![Model Summary table](image)

**Figure 22.3**
The *ANOVA* table, shown in Figure 22.4, contains the *F*-test for your overall model. That is, does body weight explain a statistically significant proportion of variance in minutes of sleep per night? Or, stated another way, does body weight provide statistically significant prediction of minutes of sleep per night? The *Regression* row contains your hypothesis test and the *Residual* row contains the error information. If the omnibus *F* (*F*) is statistically significant (*Sig.*), then you know that the independent variable significantly predicts the dependent variable, assuming the assumptions were satisfied.

![ANOVA Table](image)

The next table, titled *Coefficients*, is shown in Figure 22.5. This table contains the information required to produce your regression equation. The *B* column contains the regression coefficient for the independent variable (i.e., the slope of your regression line) and the *(Constant)* is the *y*-intercept. The *Beta* column contains the standardized coefficient. Notice this value is identical to the Pearson’s *r* found in Figure 22.2 above. Note that as you add additional independent variables (see Chapter 23) the Beta values will diverge from Pearson’s *r* due to adjustment from the other variables included in the analysis (i.e., shared variance). The *t* and *Sig.* columns contain the results of the *t*-test for your independent variable. Note that the *t*-test for the *(Constant)* only indicates that the *y*-intercept is significantly different from zero. Finally, the last two columns provide the lower and upper limits of your 95.0% *Confidence Interval for B*. The confidence interval will give you an idea of the accuracy of your prediction, with a narrow interval indicating more precise predication. Further, if the interval contains zero, the slope (*B*) is not statistically significant using alpha of .05.

![Coefficients Table](image)
The next table is titled *Residuals Statistics*. This table contains the unstandardized and standardized (*Std.*) *Predicted Value* and *Residual* value at the *Mean*, *Minimum*, and *Maximum* values of the independent variable.

Finally, the *Charts* section provides the histogram, P-P plot, and scatterplot you requested in Step 4 above. The histogram is presented first and is shown on the left of Figure 22.6. Here, the distribution of the standardized residuals can be evaluated. Second, the *Normal P-P Plot* of the standardized residuals are presented, shown in the middle of Figure 22.6. This plot is interpreted identically to those in Chapter 15. Both the histogram and P-P plot are considered when evaluating the assumption of normality of residuals. Third, a scatterplot of the standardized residuals and standardized predicted values is presented, shown on the right side of Figure 22.6. This plot is used to evaluate the assumption of homoscedasticity of residuals.

**Figure 22.6**

*Interpretation*

Interpretation begins by to evaluating the assumption of normality of residuals assumption by examining the histogram and P-P plot under the *Charts* section of your output. From this example, the histogram appears normally distributed and the P-P plot does not depart much from the diagonal reference line. Thus, there is strong evidence for normality of residuals and the assumption is considered satisfied.

Next, you evaluate the homoscedasticity of residuals assumption by evaluating the scatterplot between the standardized residuals and standardized predicted values. From the example, this scatterplot is relatively rectangular (more oval) and variance appears relatively constant; thus, homoscedasticity of residuals is considered satisfied.

With normality and homoscedasticity of residuals satisfied, you can evaluate the results provided in the *ANOVA* table to determine whether your independent variable significantly predicts your dependent variable. In this example, the *F*-test was statistically significant, so you move on to the *Coefficients* table to determine the slope (*B*) of your regression line. With only one independent variable and a statistically significant *F*-test, you already knew that the *t*-test for the *Wt* variable provided in the *Coefficients* table would be statistically significant. In fact, the inference from the *t*-test and *F*-test are identical. That is, with one degree of freedom in the numerator of the *F* ratio, \( t^2 = F \) (which is within rounding error for this example).

Interpreting the slope (*B*) for a continuous independent variable is straightforward. If the slope is positive, a one-unit increase in the independent variable results in an *increase* in the predicted dependent variable. Alternatively, if the slope is negative, a one-unit increase in the
independent variable resulted in a decrease in the predicted value of the dependent variable. From the example, the negative slope for \( Wi \) of -1.896 indicates that a one-pound increase in body weight decreased predicted minutes of sleep per night by 1.896 minutes. It is important to note that this interpretation is additive; thus, a 10-pound increase in body weight results in an 18.96 (i.e., 1.896*10) decrease in predicted minutes of sleep per night.

The interpretation of the (Constant) states it is the predicted value of the dependent variable when the independent variable is zero. That is, it is the \( y \)-intercept. So, from the example, an individual who weighs zero pounds is predicted to sleep approximately 734.970 minutes per night. Make sense? Nope. Thus, it is important to remember that the constant is only interpretable if a value of zero makes theoretical sense.

You should also note that the Coefficients table is used to produce the regression equation based the \( B \) and (Constant) values. See the algebraic equation for the general linear model in the introduction for an explanation. Remember, you can only use this equation to predict values of the dependent variable for the range of independent variable values in your sample.

After you have interpreted your independent variable, you will evaluate the Adjusted \( R^2 \) from the Model Summary table to determine the proportion of variance explained by your independent variable. Adjusted \( R^2 \) ranges from 0 to 1, with higher values indicating more variance explained and better prediction. From this example, adjusted \( R^2 \) was .406 indicating that 40.6% of the variance in minutes of sleep per night can be explained by body weight. Or, stated another way, 59.4% of the reason why participants slept as long (or as short) as they did was not explained by body weight.

**Example Results Section**

No violation of independence, homoscedasticity, or normality of residuals was indicated. Further, no violation of linearity was indicated via scatterplot.

The results of a simple linear regression analysis indicated that body weight significantly predicted minutes of sleep per night, \( F(1,28) = 20.856, \ p < .05, \) adjusted \( R^2 = .406, \) with a one-pound increase in body weight producing a 1.896 minute decrease in minutes of sleep per night (95% CI = 1.046 to 2.746 minute decrease). The regression equation used to predict minutes of sleep per night for individuals weighting between 120 and 226 is \( Y_{\text{minutes}} = 734.970 - 1.896*(\text{body weight}). \)

**Categorical Independent Variable**

**Analysis**

Assuming independence of residuals and homogeneity of variance, to conduct a simple linear regression on the data above using Minutes as the dependent variable and Tx as the categorical independent variable:

1. Click Analyze, then choose Regression, and finally, click Linear... to bring up the Linear Regression dialog box that is identical to Figure 22.1. Click Reset to clear any previous analyses.

2. All of the variables in your dataset are listed on the left hand side. Again, you will be primarily concerned with the Dependent: and Independent(s): dialog boxes.
a. Because minutes of sleep per night is your dependent variable, click to highlight Minutes and then click the right arrow ( ) next to the Dependent: box.

b. Because treatment group is your independent variable, click to highlight Tx and then click the right arrow ( ) next to the Independent(s): box.

3. You will also notice four buttons on the right hand side of this dialog box. Click the Statistics… button to bring up the Linear Regression: Statistics dialog box. In this dialog box you have several descriptive and correlational options as well as options for your Regression Coefficients and Residuals.

   a. On the right hand side of this box are five checkboxes. Model fit is selected by default. This option prints the several fit statistics including multiple $R$ (just the absolute value of Pearson’s $r$ in simple linear regression), $R^2$, and adjusted $R^2$ as well as the standard error of estimate and the omnibus ANOVA table for your overall model. The $R$ square change checkbox is only useful when you have more than one independent variable. The Descriptives option provides the number of participants with valid values, mean, and standard deviation for each variable in the analysis. The Part and partial correlations and Collinearity diagnostics option is useful only when you have more than one independent variable and are described in detail in Chapter 23.

   b. Under the Regression Coefficients section, the Estimates option is checked by default. You want to leave this option checked because it requests the majority of the relevant output regarding your independent variable, which includes the unstandardized and standardized regression coefficients, standard error as well as the $t$-value and significance test. The Confidence intervals option provides the confidence interval around your unstandardized regression coefficient at the specific Level(%): you choose. Finally, you can choose to print the Covariance matrix, but this option does not provide additional useful information when you only have one independent variable (that is, it only provides the variance of your independent variable and its correlation with itself).

   c. Under the Residuals section, you can choose to conduct the Durbin-Watson test assessing the independence of residuals assumption against a 1st order autoregressive pattern. Note that if the residuals are correlated, but not 1st order autoregressive, you will have violated the assumption, but passed this statistical check. You can also request Casewise diagnostics printing the standardized residual value, value of the dependent variable, predicted value, and unstandardized residual. You can print this information only for Outliers outside: of a specified number of standard deviations or for All cases.

When you are satisfied with your selections, click Continue.

4. Click the Plots… button. This option is required to check the normality assumption. On the left hand side of the dialog box you will see a list of variable names which include the raw dependent variable (DEPENDNT), standardized predicted value (*ZPRED), adjusted predicted value (*ADJPRED), and four residual values including the standardized residual (*ZRESID), deleted residuals (*DRESID), Studentized residuals (*SRESID), and Studentized deleted residuals (*SDRESID). Because you have a categorical independent
variable, the Scatter section is not useful. Under the Standardized Residual Plots section, you have the option to print both the Histogram and Normal probability plot of your standardized residuals used to evaluate the normality of residuals assumption. When you are satisfied with your choices, click Continue.

5. Click the Save… button. Here, you have the option to save many different values to your dataset. That is, any option you select in this dialog box will not print to your output, but instead a new variable will be created in your dataset.

a. Under the Predicted Values section, you have the option to save the Unstandardized or raw predicted value, the Standardized predicted value (think, z-scores), the Adjusted predicted value which is the predicted value when the participant is excluded from analysis, or the S.E. of mean predictions which is the standard deviation of the average value of the dependent variable for participants that have the same value on the independent variable.

b. Under the Residuals section, you can save up to five different residual values. The Unstandardized residual is the actual difference in raw units, whereas the Standardized residual are z-scored residuals. A Studentized residual is the residual value divided by a varying standard deviations based on the distance of the individual’s independent variable value from the mean of the independent variable. A Deleted residual is the residual when the participant has been excluded from the analysis and is the difference between the dependent variable and the adjusted predicted value. Finally, a Studentized deleted residual is the deleted residual value divided by its standard error. It is used to indicate how much of a difference eliminating a participant makes on its own prediction.

c. The options under the Distances section are used to identify participants who have unusual independent variable combinations. Combinations imply more than one independent variable, so these options are discussed in Chapter 23.

d. The options under the Influence Statistics section provide indicators of what would happen a participant was excluded from analysis. DfBeta(s) are the difference in standardized regression coefficient (i.e., beta) if the participant was removed from analysis with Standardized DfBeta(s) being simply z-scored DfBetas. DfFit is the change in the predicted value if the participant is excluded from analysis, and Standardized DfFit are simply z-scored DfFit values. Finally, the Covariance ratio is the ratio of the determinant of the covariance matrix with a particular participant excluded from analysis to the determinant of the covariance matrix with the participant included; thus, ratios closer to 1 indicate little change.

e. In the Prediction Intervals section, you can calculate the Confidence Interval: around your predicted values. The Mean option produces confidence intervals around the mean predicted response, whereas the Individual produces intervals around the value of the dependent variable.

f. Selecting Create coefficient statistics will save the regression coefficients to a new dataset (you must provide a Dataset name:) or a completely new data file (click File… to locate the already existing file).
g. Finally, you can Export model information to XML file. Click Browse... to locate the already existing file. This option will save parameter estimates, and optionally the covariance matrix (clicking the checkbox is required) in XML format.

When you are satisfied with your selections, click Continue.

6. Click the Options... button. The Stepping Method Criteria option is only useful if you are using stepwise regression (oh, and never use stepwise regression). In addition, you can Include constant in equation, where unchecking this option forces the regression equation through the origin, which is rare; thus, it is wise to leave this option checked. Finally, you can tell SPSS how you want to handle Missing Values. When you are satisfied with your selections, click Continue.

7. That’s it! Click OK to conduct the analysis.

Output

When you click OK, SPSS will produce an Output screen displaying your results. Click the Output window to view your results (if it does not pop up automatically).

The first table is titled Descriptive Statistics, which contains the Mean, standard deviation (Std. Deviation), and sample size (N) for your dependent and independent variables. Next, is a table titled Correlations, which contains the correlation between your dependent and independent variables. Both tables are presented in Figure 22.7.

<table>
<thead>
<tr>
<th>Descriptive Statistics</th>
<th>Minutes</th>
<th>Std. Deviation</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minutes</td>
<td>459.23</td>
<td>76.309</td>
<td>39</td>
</tr>
<tr>
<td>Tx</td>
<td>50</td>
<td>508</td>
<td>39</td>
</tr>
</tbody>
</table>

The next table, titled Model Summary is shown in Figure 22.8. This table contains the correlation between your two variables (R). Note that because we only have one IV, this value is identical to the correlation in Figure 22.7. It also contains your coefficient of determination, labeled R Square, which estimates the proportion of variance explained in your dependent variable by your independent variable. Adjusted R Square is the R Square value adjusted for the degrees of freedom lost during the regression, and is a more accurate estimate of your coefficient of determination. Finally, the standard error of estimate (Std. Error of the Estimate) measures the overall accuracy of your regression model. It is the square root of the mean square residual (i.e., error variance) found the ANOVA table, discussed next.

<table>
<thead>
<tr>
<th>Correlations</th>
<th>Minutes</th>
<th>Tx</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Correlation Minutes</td>
<td>1.000</td>
<td>0.31</td>
</tr>
<tr>
<td>Pearson Correlation Tx</td>
<td>0.31</td>
<td>1.000</td>
</tr>
<tr>
<td>Std. (1-tailed) Minutes</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Std. (1-tailed) Tx</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>N Minutes</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>N Tx</td>
<td>30</td>
<td>30</td>
</tr>
</tbody>
</table>

Figure 22.7
The ANOVA table, shown in Figure 22.9, contains the \( F \)-test for your overall model. That is, does treatment group explain a statistically significant proportion of variance in minutes of sleep per night? Or, stated another way, does the treatment group a participants was randomized to provide statistically significant prediction of minutes of sleep per night? The Regression row contains your hypothesis test and the Residual row contains the error information. If the omnibus \( F \) \((F)\) is statistically significant (\( \text{Sig.} \)), then you know that the independent variable significantly predicts the dependent variable, assuming the assumptions were satisfied.

The next table, titled Coefficients, is shown in Figure 22.10. This table contains the information required to produce your regression equation. The \( B \) column contains the regression coefficient for your independent variable and the \( (\text{Constant}) \) term is the \( y \)-intercept. The column titled Beta is a standardized coefficient. Notice this value is identical to the correlation and \( R \) values found in the Figures 22.7 and 22.8, respectively. This occurred because you have one independent variable. As you include additional independent variables (see Chapter 23) the Beta values will diverge from Pearson’s \( r \) due to adjustment from the other variables included in the analysis (i.e., shared variance). The \( t \) and \( \text{Sig.} \) columns contain the results of the \( t \)-test for your independent variable. Note that the \( t \)-test for the \( (\text{Constant}) \) is irrelevant, and only indicates that the \( y \)-intercept is significantly different from zero. That is, it does not compare anything between groups. Finally, the last two columns provide the lower and upper limits of your 95.0% Confidence Interval for \( B \). The confidence interval will give you an idea of the accuracy of your prediction, with a narrow interval indicating more precise prediction. Further, if the interval contains zero, the slope \( (B) \) is not statistically significant using alpha of .05.
The next table is titled *Residuals Statistics*. This table contains the unstandardized and standardized (Std.) *Predicted Value* and *Residual value* at the *Mean, Minimum, and Maximum* values of the independent variable.

Finally, the *Charts* section provides the histogram and P-P plot you requested in Step 4 above. The histogram is presented first and is shown on the left side of Figure 22.11. Here, the distribution of the standardized residuals can be evaluated. Second, the *Normal P-P Plot* of the standardized residuals are presented, shown on the right side of Figure 22.11. This plot is interpreted identically to the Q-Q plot presented in Chapter 15. Both the histogram and P-P plot are considered when evaluating the assumption of normality of residuals.

### Figure 22.10

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
<th>95.0% Confidence Interval for B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>346.987</td>
<td>11.145</td>
<td>31.119</td>
<td>0.000</td>
</tr>
<tr>
<td>Tx</td>
<td>124.733</td>
<td>15.763</td>
<td>89.1</td>
<td>7.913</td>
<td>0.000</td>
</tr>
</tbody>
</table>

*a. Dependent Variable: Minutes*

### Figure 22.11

**Interpretation**

Interpretation begins by to evaluating the assumption of normality of residuals assumption by examining the histogram and P-P plot under the *Charts* section of your output. From this example, the histogram appears normally distributed and the P-P plot does not depart much from the diagonal reference line. Thus, there is strong evidence for normality of residuals and the assumption is considered satisfied.

With normality of residuals satisfied, you can evaluate the results provided in the *ANOVA* table to determine whether your independent variable significantly predicts your dependent variable. In this example, the *F*-test was statistically significant, so you move on to the *Coefficients* table to determine the slope (*B*) of your regression line. With only one independent variable and a statistically significant *F*-test, you already knew that the *t*-test for the *Wt* variable provided in the *Coefficients* table would be statistically significant. In fact, the inference from the
Statistical Analysis using SPSS – Walters (2011)

$t$-test and $F$-test are identical. That is, with one degree of freedom in the numerator of the $F$ ratio, $t^2 = F$ (which is within rounding error for this example).

Interpreting the slope ($B$) for a categorical independent variable is different slightly from the interpretation discussed above for a continuous independent variable. Here, because the independent variable is dichotomous, a one-unit increase now indicates going from the OTC group (coded 0) to the new sleep treatment group (coded 1). Thus, this slope is a direct comparison between the two groups. It is important to note that SPSS calculates the regression slope for the group coded 1, whereas the group coded 0 is always the reference group. So, in this case, the slope was calculated for the group receiving your new sleep treatment. Thus, being in the group receiving the new sleep treatment resulted in a 124.733 increase in predicted minutes of sleep per night compared to the OTC group. Notice the direct comparison in italics!

Alternatively, being in the OTC group resulted in a 124.733 decrease in predicted minutes of sleep per night compared to the group receiving the new sleep treatment.

Remember that the interpretation of the (Constant) is the predicted value of the dependent variable when the independent variable is zero. Here, a value of zero for the independent variable is empirically relevant as it represents the group receiving the OTC treatment. Thus, the OTC group is predicted to sleep 346.867 minutes on average. And, because being in the group receiving the new sleep treatment resulted in a 124.733 increase in predicted minutes of sleep per night compared to the OTC group, the new sleep treatment group is predicted to sleep 471.600 minutes (i.e., 346.867+124.733) on average.

Identical to above, the $B$ and (Constant) values in the Coefficients table are used to calculate the regression equation. See the algebraic equation in the introduction for an explanation.

Finally, after you have interpreted your independent variable, you will evaluate the Adjusted $R$ Square from the Model Summary table to determine the proportion of variance explained by your independent variable. Adjusted $R$ Square ranges from 0 to 1, with higher values indicating more variance explained and better prediction. From this example, adjusted $R^2$ was .680 indicating that 68.0% of the variance in minutes of sleep per night can be explained by treatment group. Or, stated another way, 32.0% of the reason why participants slept as long (or as short) as they did was not explained by treatment group.

**Example Results Section**

No violation of independence or normality of residuals was indicated. Further, no heterogeneity of variance was indicated via Levene’s test.

The results of a simple linear regression analysis indicated that treatment group significantly predicted minutes of sleep per night, $F(1,28) = 62.614, p < .05$, adjusted $R^2 = .680$. Participants who received the new sleep treatment had a 124.733 increase in minutes of sleep per night compared to those receiving the OTC treatment (95% CI = 92.444 to 157.023 minutes). The regression equation used to predict minutes of sleep per night is $Y_{\text{minutes}} = 346.867 + 124.733*(\text{new sleep treatment})$. 

106
Chapter 23
MULTIPLE LINEAR REGRESSION

Multiple linear regression is an extension of simple linear regression to situations involving more than one independent variable. The extension is most easily seen through the regression equation: \( Y = a + B_1X_1 + B_2X_2 + \cdots + B_kX_k + e \). Again, \( Y \) is the predicted value of the dependent variable, \( a \) is the \( y \)-intercept, the \( B \)s are the slopes for each independent variable, the \( X \)s are the values of the independent variable, and \( e \) is the residual. You can have as many independent variables as statistical power affords, and they can be measured on any scale.

One of the most common reasons for using multiple linear regression is to control for nuisance variables called covariates. Covariates are variables that are correlated with the dependent variable, but are not of primary research interest. Thus, in many situations, you may have only one or two independent variables and several covariates. In reality, however, the difference between an independent variable and covariates is semantic. The analysis treats them all the same. It is up to you to determine which variables are interesting and which are nuisance.

Because multiple linear regression is a direct extension of simple linear regression, the analysis and interpretation are straightforward. This, of course, assumes you have a solid understanding of simple linear regression, so feel free to re-read Chapter 22 if necessary. To avoid redundancy in explanation, we will jump right into the example.

Consider a study evaluating the effect of your new sleep treatment on minutes of sleep per night compared to a popular over the counter treatment and placebo. Because the amount of sleep per night can be affected by numerous variables, you decide to control for caffeine consumption after 2pm (in milligrams), whether the participant smokes, body weight in pounds, and level of global anxiety as measured by the Hamilton Anxiety Rating Scale (HAM-A).

Following the data entry procedures described in Chapters 1 and 2, label the first variable \( ID \), the second variable \( Tx \) (1 = Placebo; 2 = OTC; 3 = Treatment), the third variable \( Caffeine \), the fourth variable \( Smoke \) (0 = No; 1 = Yes), the fifth variable \( Wt \) for body weight, the sixth variable \( Anx \) for anxiety, and the seventh variable \( Minutes \). Enter the data appropriately.
Assumptions

The assumptions of multiple linear regression include all the assumptions for simple linear regression—indeed, independence, normality, and homoscedasticity of residuals as well as absence of bivariate outliers and linearity. However, two additional assumptions, directly related to the fact that you now have more than one independent variable, include absence of multivariate outliers and absence of multicollinearity.

Independence of residuals requires the residuals to be uncorrelated across participants. Remember, a residual is the difference between predicted and actual value of the dependent variable, so participants will have their own residual value. Independence of residuals is technically a design issue that requires no repeated measurement or clustering. It can be assured by using appropriate random sampling and random assignment techniques.

Normality of residuals is evaluated during analysis by requesting a histogram and P-P plot of standardized residuals. Note that a P-P plot, or probability-probability plot, is interpreted identically to the Q-Q plot discussed in Chapter 15. If the histogram is normally distributed and the P-P plot follows the diagonal reference line, normality is assured.

Homoscedasticity of residuals requires the residuals to have a constant variance across all predicted values of the dependent variable. This assumption is evaluated during the analysis by requesting a scatterplot of the standardized residuals and the standardized predicted values. If this scatterplot is roughly rectangular shaped, the assumption is considered satisfied. If a violation is indicated, however, the probability of committing a type I error increases, so a more conservative alpha level should be used (e.g., .01 instead of .05) when evaluating the effects of individual predictors.

Bivariate (i.e., two variable) outliers can severely influence the regression line and can lead to incorrect inference. Bivariate outliers are identified by data points that are disconnected from the rest of the data points in the scatterplot. Note that you will assess for outliers using a scatterplot of each independent variable-dependent variable pair. That is, you will not evaluate scatterplots of two independent variables. Further, no assumption is made about the distribution or outliers within the independent variables specifically. That is, an outlying value is always defined in relation to the dependent variable. If bivariate outliers are identified you have two options—remove them from analysis through deletion (do not forget to describe them in your results section) or transform the dependent variable (not advised).

Linearity requires the relationship between all independent variables and the dependent variable to be best approximated by a straight regression line. This assumption is evaluated prior to analysis by examining scatterplots of the dependent variable against each independent variable separately (Chapter 19).

Multivariate outliers are participants with unusual combinations of independent variable values. They can be identified through several methods, with one of the most common (i.e., simplest) being Mahalanobis distance. This identification process will be described in detail below. Briefly, Mahalanobis distance follows a chi-square distribution; thus, any distance greater than the critical chi-square value—at alpha of .001 with degrees of freedom equal to the number of independent variables—is considered a multivariate outlier.

Multicollinearity is defined as redundancy among your independent variables. While your independent variables are expected to be highly correlated with the dependent variable they should not be highly correlated with each other. If they are too highly correlated with each other, the standard errors increase making the regression coefficients ($B$) unstable. There are several
ways to identify multicollinearity—a statistically significant omnibus $F$ with no statistically significant regression parameters, beta values greater than 1, tolerance values below .10, or a variance inflation factor (VIF) above 10. The tolerance and VIF values are just benchmarks, however, as some researchers consider tolerance values below .40 and VIF values above 2.5 indicative of a collinearity issue. Prior to interpreting results, we will request collinearity diagnostics to examine the tolerance and VIF values. Briefly, tolerance is the proportion of variance in an independent variable that is independent of the other independent variables. So, a tolerance of .10 indicates that the variables shares 90% of its variance with the other independent variables and is largely redundant. The VIF is the reciprocal of tolerance and indicates how much of the variance of each estimated regression coefficient increased due to multicollinearity. If multicollinearity is observed you have several options including increasing sample size, combining offending variables into one variable, removing one of the variables from analysis, or using a different technique such as structural equation modeling or ridge regression.

**Dummy Variables**

One final point I want to make involves dummy variables and indicator coding. They are essentially synonyms and hopefully you are familiar with these terms. The linear regression procedure in SPSS requires you to create dummy variables for every categorical variable with three or more categories. The number of dummy variables required is calculated by taking the number of categories minus 1 (i.e., $k-1$). For example, the treatment variable in the example above has three categories (i.e., placebo, OTC, new treatment); thus, two dummy variables are required. Dummy variables are binary, usually coded 0/1, with 1 indicating the individual is a member of that category and 0 indicating they are not a member of that category. Note that one of the categories will always receive a value of 0 for all dummy variables—this is your reference category. That is, each dummy variable will compare the category coded 1 to the reference category. So, from the example, if we want to compare our new sleep treatment to both the OTC and placebo group, then it makes the most sense to have the new treatment group serve as the reference category. This way, both dummy variables (one for placebo, one for OTC) will be compared directly to the new sleep treatment and not to each other.

To create dummy variables, you will need to use the Recode into Different Variables procedure described in Chapter 8. For the example above, two dummy variables were created—one labeled *Placebo* and the other labeled *OTC*. If you created these variables correctly, group 3 in the original *Tx* variable, should have values of 0 for both the *Placebo* and *OTC* dummy variables.

**Analysis**

Assuming the assumptions of independence of residuals, absence of bivariate outliers, and linearity are satisfied, to conduct a multiple linear regression analysis:

1. Click **Analyze**, then choose **Regression**, and finally, click **Linear...** to bring up the *Linear Regression* dialog box. This dialog box is identical to the one used for simple linear regression presented in Figure 22.1 of the previous Chapter.

2. You will immediately see all of the variables in your dataset listed on the left hand side of the dialog box. Similar to simple linear regression, you will be concerned primarily with
the **Dependent:** and **Independent(s):** dialog boxes. The variable you place in either of these boxes is relatively self-explanatory. Remember, covariates are treated exactly the same as independent variables.

a. Because minutes of sleep per night is your dependent variable, click to highlight **Minutes** and then click the right arrow (→) next to the **Dependent:** box.

b. Your covariates include **Caffeine**, **Smoke**, body weight in pounds (**Wt**), and anxiety level (**Anx**). One-by-one, click to highlight each variable and then click the right arrow (→) next to the **Independent(s):** box.

c. Finally, your independent variables are two dummy variables—one representing the placebo group (**Placebo**), and one representing the group receiving the over the counter treatment (**OTC**). Remember, the group receiving your new sleep treatment is the reference group and, thus, is not represented by a dummy variable. One-by-one, click to highlight each variable and then click the right arrow (→) next to the **Independent(s):** box.

3. You will also notice four buttons on the right hand side of this dialog box. Click the **Statistics...** button to bring up the **Linear Regression: Statistics** dialog box. In this dialog box you have several descriptive and correlational options as well as options for your **Regression Coefficients** and **Residuals**.

   a. On the right hand side of this box are five checkboxes. **Model fit** is selected by default. This option prints the several fit statistics including the multiple correlation coefficient (**R**), **R**², and adjusted **R**² as well as the standard error of estimate and the omnibus ANOVA table for your overall model. The **R square change** checkbox is useful if you are entering the independent variables in sets. That is, covariates in first set, independent variables of interest in second set. It answers the question of whether the variables in subsequent sets explain a statistically significant amount of variance over and above the variance explained by the sets already in the analysis. The **Descriptives** option provides the number of participants with non-missing values as well as the mean and standard deviation for each variable in the analysis. The **Part and partial correlations** option will provide the zero-order (i.e., Pearson’s **r**), semipartial (i.e., part), and partial correlations. The latter two correlations estimate the relationship between the dependent variable and specific independent variable after controlling for the effects of the other independent variables. The semipartial correlation is the unique correlation between the dependent variable and specific independent variable. This correlation is what multiple regression is all about! The partial correlation is the correlation between the dependent variable and specific independent variable holding the remaining independent variables constant. The **Collinearity diagnostics** option will provide you with tolerance and VIF values for each independent variable used to check for multicollinearity.

   b. Under the **Regression Coefficients** section, the **Estimates** option is checked by default. You will want to leave this option checked because it requests the majority of the relevant output regarding your independent variables, which includes the unstandardized and standardized regression coefficients, standard error as well as the **t**-value and significance test. The **Confidence intervals** option
provides the confidence interval around your unstandardized regression coefficient at the specific Level(%): you choose. Finally, you can choose to print the Covariance matrix, but in general, you’ll be more interested in the standardized covariance (i.e., correlation).

c. Under the Residuals section, you can choose to conduct the Durbin-Watson test assessing the independence of residuals assumption against a 1st order autoregressive pattern. Note that if the residuals are correlated, but not 1st order autoregressive, you will have violated the assumption, but passed this statistical check. You can also request Casewise diagnostics printing the standardized residual value, value of the dependent variable, predicted value, and unstandardized residual. You can print this information only for Outliers outside: of a specified number of standard deviations or for All cases.

When you are satisfied with your selections, click Continue.

4. Click the Plots… button. This option is required to check the normality and homoscedasticity of residuals assumptions. On the left hand side of the dialog box you will see a list of variable names which include the raw dependent variable (DEPENDNT), standardized predicted value (*ZPRED), adjusted predicted value (*ADJPRED), and four residual values including the standardized residual (*ZRESID), deleted residuals (*DRESID), Studentized residuals (*SRESID), and Studentized deleted residuals (*SDRESID). To evaluate homoscedasticity you need to plot the standardized predicted against the standardized residual values. Note that you can include as many combinations you want by clicking the Next button under the Scatter section.

a. Click to select the *ZRESID variable, and then click the right arrow ( ) next to the Y: box in the Scatter section.

b. Click to select the *ZPRED variable, and then click the right arrow ( ) next to the X: box in the Scatter section.

Under the Standardized Residual Plots section, you have the option to print both a Histogram and a Normal probability plot of your standardized residuals. Both of these options are used to evaluate the normality of residuals assumption. When you are satisfied with your choices, click Continue.

5. Click the Save… button. Here, you have the option to save many different values to your dataset. That is, any option you select in this dialog box will not print to your output, but instead a new variable will be created in your dataset.

a. Under the Predicted Values section, you have the option to save the Unstandardized or raw predicted value, the Standardized predicted value (think, z-scores), the Adjusted predicted value which is the predicted value when the participant is excluded from analysis, or the S.E. of mean predictions which is the standard deviation of the average value of the dependent variable for participants that have the same value on the independent variable.

b. Under the Residuals section, you can save up to five different residual values. The Unstandardized residual is the actual difference in raw units, whereas the Standardized residual are z-scored residuals. A Studentized residual is the residual value divided by a varying standard deviations based on the distance of the
individual’s independent variable value from the mean of the independent variable. A **Deleted** residual is the residual when the participant has been excluded from the analysis and is the difference between the dependent variable and the adjusted predicted value. Finally, a **Studentized deleted** residual is the deleted residual value divided by its standard error. It is used to indicate how much of a difference eliminating a participant makes on its own prediction.

c. The options under the **Distances** section are used to identify participants who have unusual independent variable combinations (i.e., multivariate outliers). The **Mahalanobis** distance option provides an indicator of how much a participant’s independent variable value differs from the group mean. **Cook’s** distance indicates how much the residuals of all participants would change if the participant was excluded from analysis. Finally, **Leverage values** measure the influence the independent variable value has on the fit of the regression.

d. The options under the **Influence Statistics** section provide indicators of what would happen if a participant were excluded from analysis. **DfBeta(s)** are the difference in standardized regression coefficient (i.e., beta) if the participant was removed from analysis with **Standardized DfBeta(s)** being simply z-scored DfBetas. **DfFit** is the change in the predicted value if the participant is excluded from analysis, and **Standardized DfFit** are simply z-scored DfFit values. Finally, the **Covariance ratio** is the ratio of the determinant of the covariance matrix with a particular participant excluded from analysis to the determinant of the covariance matrix with the participant included; thus, ratios closer to 1 indicate little change.

e. In the **Prediction Intervals** section, you can calculate the **Confidence Interval**: around your predicted values. The **Mean** option produces confidence intervals around the mean predicted response, whereas the **Individual** produces intervals around the value of the dependent variable.

f. Selecting **Create coefficient statistics** will save the regression coefficients to a new dataset (you must provide a **Dataset name:** ) or a completely new data file (click **File...** to locate the already existing file).

g. Finally, you can **Export model information to XML file**. Click **Browse...** to locate the already existing file. This option will save parameter estimates, and optionally the covariance matrix (clicking the checkbox is required) in XML format.

When you are satisfied with your selections, click **Continue**.

6. Click the **Options...** button. The **Stepping Method Criteria** option is only useful if you are using stepwise regression (oh, and never use stepwise regression). In addition, you can **Include constant in equation**, where unchecked this option forces the regression equation through the origin, which is rare; thus, it is wise to always leave this option checked. Finally, you can tell SPSS how you want to handle **Missing Values**. When you are satisfied with your selections, click **Continue**.

7. That’s it! Click **OK** to conduct the analysis.
Output

When you click OK, SPSS will produce an Output screen displaying your results. Click the Output window to view your results (if it does not pop up automatically). Much of this output is largely redundant with what was printed for a simple linear regression analysis described in Chapter 22. Thus, extensive descriptions are avoided here to limit redundancy. However, new information will be described in detail.

The first table is titled Descriptive Statistics, which contains the Mean, standard deviation (Std. Deviation), and sample size (N) for your dependent and independent variables. Next, is a table titled Correlations, which contains the correlation between your variables. You should notice that all independent variables, except for OTC, have large, statistically significant Pearson Correlations with the dependent variable and relatively low correlations amongst each other. This is the exact pattern you want to observe!

The Variables Entered/Removed table is only relevant for hierarchical regression or stepwise regression. Both of these procedures enter variables in sets, and this table would indicate which variables were included in a particular analysis step.

The next table, titled Model Summary is shown in Figure 23.1. This table contains the multiple correlation coefficient (R), which is a summary measure indicates the strength of relationship between the dependent and all independent variables. It also contains the coefficient of determination, labeled R Square, which estimates the proportion of variance explained in your dependent variable by your independent variables, calculated by dividing the Sum of Squares for Regression by the Total Sum of Squares found in the ANOVA table, discussed next. Adjusted R Square is the R Square value adjusted for the degrees of freedom lost during the regression, and is a more accurate estimate of your coefficient of determination. Finally, the standard error of estimate (Std. Error of the Estimate) measures the overall accuracy of your regression model. It is the square root of the mean square residual found the ANOVA table. Remember, mean square is just another term for variance.

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.909*</td>
<td>.827</td>
<td>.781</td>
<td>48.406</td>
</tr>
</tbody>
</table>

* a. Predictors: (Constant), OTC, Smoke, Wt, Anx, Caffeine, Placebo
  b. Dependent Variable: Minutes

Figure 23.1

The ANOVA table, shown in Figure 23.2, contains information pertaining to the hypothesis test for your overall model. That is, does treatment group predict minutes of sleep per night after controlling for caffeine intake, smoking status, body weight, and anxiety level? The Regression row contains your hypothesis test and the Residual row contains the error information. If the omnibus F (F) is statistically significant (Sig.), then you know that at least one of your independent variables significantly predicts your dependent variable, assuming all assumptions were satisfied.
The next table, titled *Coefficients*, is shown in Figure 23.3. The *Coefficients* table contains the information required to produce your regression equation, identifies which independent variables are most important, and provides the tolerance and VIF values required to assess multicollinearity. The *B* column contains the regression coefficients for your independent variables (i.e., the slope of your regression line) and the (Constant) term is the *y*-intercept. The column titled *Beta* is a standardized coefficient, interpreted similarly to Pearson’s *r*. The absolute value of the Betas can be used to determine which independent variables are most important to prediction. The *t* and *Sig.* columns contain the results of the *t*-test for your independent variables. Note that the *t*-test and interpretation of the (Constant) is still irrelevant. That is, the constant is interpreted as the value of the dependent variable when all independent variables are zero. Thus, the constant is now the value of the dependent variable for an individual who received your new sleep treatment (i.e., Placebo = 0; OTC = 0), who did not ingest caffeine after 2pm or smoke, has no measured anxiety, but has a body weight of 0 pounds. That is, this person is weightless and probably not living (at least on Earth). The next two columns provide the lower and upper limits of your 95.0% *Confidence Interval for B*. The confidence interval will give you an idea of the accuracy of your prediction, with a narrow interval indicating more precise prediction. Again, if the interval contains zero, the slope (*B*) is not statistically significant using alpha of .05. The next section of output provides the *Correlations* you requested in Step 3a above. The Zero-order column will be identical to the Pearson’s correlations provided in the *Correlations* table. The *Partial* and *Part* (i.e., semipartial) correlations are also provided. Squaring the semipartial correlation will give you the amount of unique variance in the dependent variable explained by the individual independent variable. Finally, you are presented with *Collinearity Statistics*. Remember, *Tolerance* values below .10 and *VIF* values above 10 suggest multicollinearity.

### ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th><em>F</em></th>
<th>Sig.*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>239159.244</td>
<td>6</td>
<td>39896.674</td>
<td>10.277</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>48530.822</td>
<td>23</td>
<td>2195.506</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>287689.067</td>
<td>29</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*a. Predictors: (Constant), OTC, Smoke, M/F, Age, Caffeine, Placebo
b. Dependent Variable: Minutes*

### Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th><em>B</em></th>
<th>Std. Error</th>
<th><em>Beta</em></th>
<th><em>t</em></th>
<th><em>Sig.</em></th>
<th>95.0% Confidence Interval for <em>B</em></th>
<th>Correlations</th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>-----</td>
<td>------------</td>
<td>--------</td>
<td>-----</td>
<td>--------</td>
<td>Lower Bound</td>
<td>Upper Bound</td>
<td>Zero-order</td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-9.664</td>
<td>8.684</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Caffeine</td>
<td>-.117</td>
<td>-.081</td>
<td>-.014</td>
<td>-.294</td>
<td>.209</td>
<td>.090</td>
<td>460.223</td>
<td>691.946</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Smoke</td>
<td>.920</td>
<td>.171</td>
<td>.047</td>
<td>.473</td>
<td>.641</td>
<td>.100</td>
<td>-39.460</td>
<td>50.100</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>M/F</td>
<td>-.430</td>
<td>.283</td>
<td>-.133</td>
<td>-.432</td>
<td>.106</td>
<td>.070</td>
<td>-238.120</td>
<td>227.821</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>-.222</td>
<td>.369</td>
<td>-.163</td>
<td>-.720</td>
<td>.090</td>
<td>.404</td>
<td>-5.330</td>
<td>.445</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Caffeine</td>
<td>-.169</td>
<td>.264</td>
<td>-.014</td>
<td>-.264</td>
<td>.290</td>
<td>.100</td>
<td>-227.821</td>
<td>-109.290</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Smoke</td>
<td>.436</td>
<td>.231</td>
<td>.047</td>
<td>.827</td>
<td>.001</td>
<td>.046</td>
<td>41.705</td>
<td>-624</td>
<td></td>
</tr>
</tbody>
</table>

*a. Dependent Variable: Minutes*
The next table is titled Collinearity Diagnostics. Here, you are provided with more indicators of possible multicollinearity beyond tolerance and VIF values. Occasionally, Eigenvalues and the condition indices (Condition Index) calculated from the eigenvalues are used to diagnose multicollinearity. Opinions vary regarding benchmarks for identifying multicollinearity using the condition indices, typically values above either 15 or 30 are cause for concern. With that said, eigenvalues can be calculated several different ways, and there is controversy over which calculation is more appropriate for certain situations. Thus, it is probably better to consider the tolerance and VIF values as indicators.

The next table is titled Residuals Statistics. This table contains the unstandardized and standardized (Std.) Predicted Value and Residual value at the Mean, Minimum, and Maximum values of the independent variables.

Finally, the Charts section provides the histogram, P-P plot, and scatterplot used to evaluate the normality and homoscedasticity of residuals assumptions. The histogram is presented first and is shown on the left side of Figure 23.4. Next, the Normal P-P Plot of the standardized residuals is presented, and is interpreted identically to those in Chapter 15. Third, a scatterplot of the standardized residuals and standardized predicted values is presented, shown on the right side of Figure 23.4.

Interpretation

Interpretation begins by examining the normality and homoscedasticity of residuals assumption. To begin, examine the histogram, P-P plot, and scatterplot in the Charts section of your output to evaluate the assumptions of normality and homoscedasticity of residuals. In this example, the histogram and P-P plot provide fairly convincing evidence of normality of residuals. However, slight negative skewness and truncation is visible in the histogram, but this was not considered prohibitive. Next, consider the scatterplot between the standardized residuals and standardized predicted values, which is nicely rectangular shaped with constant variance across the standardized predicted values; thus, the homoscedasticity of residuals assumption is considered satisfied.

Next, consult the Tolerance and VIF columns of the Coefficients table to evaluate for multicollinearity. The lowest tolerance value is .394 and the highest VIF is 2.541. Both of these values are within suggested benchmarks; thus, the absence of multicollinearity assumption is considered satisfied.
Finally, you need to go back into your dataset to evaluate for multivariate outliers using
the Mahalanobis distances you saved in Step 5c above (named $MAH_1$ in your dataset).
Remember, Mahalanobis distance follows a chi-square distribution. So, using an alpha of .001
and 6 degrees of freedom (because you have 6 independent variables in the analysis), the critical
Mahalanobis distance is 22.458. Thus, a Mahalanobis distance greater than 22.458 is considered
a multivariate outlier. Note that the critical value can be found in the chi-square table in the back
of any statistics textbook. From the $MAH_1$ column of your dataset, the largest Mahalanobis
distance is 16.393, and because this value is smaller than the critical value, the absence
of multivariate outliers assumption is satisfied.

With the absence of multivariate outliers and multicollinearity as well as the normality
and homoscedasticity of residuals assumptions satisfied, you can evaluate the results provided in
the ANOVA table to determine whether your independent variables, as a set, predict a statistically
significant proportion of variance in your dependent variable. In the example, the $F$-test was
statistically significant, so you can move on to the Coefficients table to evaluate the individual
independent variables for statistical significance, of which only the Placebo and OTC variables
are statistically significant after adjusting for the other independent variables.

Because the Placebo and OTC variables are binary dummy variables and compared to the
reference group (i.e., your new sleep treatment), interpretation is simply a direct comparison of
the two groups involved. See Chapter 22 for a full explanation of how to interpret continuous
and categorical independent variables. The negative slope ($B$) for the Placebo variable states that
the placebo group had a 168.557 decrease in minutes of sleep per night compared to the group
receiving your new sleep treatment after adjusting for caffeine intake, smoking, body weight, and anxiety. Similarly, the interpretation for the OTC variable states that the group receiving the over
the counter treatment had a 90.766 decrease in minutes of sleep per night compared to the group
receiving your new sleep treatment after adjusting for caffeine intake, smoking, body weight, and anxiety. Note that both of these interpretations can be changed to state that your new sleep
treatment produced an increase in minutes of sleep per night after adjusting for the covariates.

Next, you will evaluate the semipartial correlation found in the Part column of the
Coefficients table. If you square the semipartial correlations for the Placebo and OTC variables
reported as $r^2_{sp}$, you will calculate the unique variance in the dependent variable that is
explained by each group, respectively. That is, being in the placebo group explained
approximately 26.11% (i.e., -.511\(^2\)) of the variance in minutes of sleep per night, while being
receiving the over-the-counter treatment explained approximately 11.02% (i.e., -.332\(^2\)) of the
variance in minutes of sleep per night. A regression equation can also be constructed using the
slopes and (Constant) values found in the Coefficients table.

Finally, you will evaluate the Adjusted $R$ Square value from the Model Summary table to
determine the proportion of variance explained by your independent variable. This model
explained 78.1% of the variance in the dependent variable, a huge proportion to say the least!

Example Results Section

No violation of linearity, multicollinearity, or independence, normality, or
homoscedasticity of residuals was indicated. Further, no multivariate outliers
were identified.

The results of a multiple linear regression analysis indicated that the two
independent variables and four covariates, as a set, significantly predicted minutes
of sleep per night, $F(6, 23) = 18.277$, $p < .05$, adjusted $R^2 = .781$. The group receiving the new sleep treatment slept significantly longer than both the placebo group (168.557 minutes; 95% CI = 109.293 to 227.821 minutes; $r^2_{sp} = .261$) and the group receiving the over-the-counter treatment (90.766 minutes; 95% CI = 41.705 to 139.827 minutes; $r^2_{sp} = .110$) after adjusting for amount of caffeine after 2pm, smoking status, body weight, and anxiety level. No covariate remained statistically significant after adjustment. The regression equation used to predict minutes of sleep per night is $Y_{\text{minutes}} = 570.634 - .117*(\text{mg caffeine intake after 2pm}) + 9.320*(\text{smoking status}) - .420*(\text{body weight in pounds}) - 2.221*(\text{HAM-A anxiety}) - 168.557*(\text{placebo}) - 90.766*(\text{OTC})$. 
Section V

Statistical Tests of Between-Subject Differences

The parametric analyses described in this section are special cases of the general linear model. That is, most of the assumptions discussed here are based on the residual values from the linear regression equation. Please feel free to re-read the introduction to Chapter 22 prior to beginning this section. Further, I also provide detailed descriptions of available nonparametric alternatives.

Similar to Part IV, the Chapters in this section each include a small dataset for you to practice your data entry and coding skills. I chose to include a dataset for you to use rather than refer to abstract examples so that you will be able to perform the analyses and replicate all output presented. This will allow you to verify you completed the analysis correctly. Similar to above, while all portions of the menus for each analysis are described in detail, the bolded instructions are minimally required to replicate the output provided.

Finally, at the end of each Chapter you will be presented with an example results section in APA format. This should give you a pretty good idea of what will be minimally required when reporting results for your future posters or manuscripts.
Chapter 24
ONE-SAMPLE $t$-TEST

A one-sample $t$-test is used to test for a statistical difference between the mean of your sample a *known* population mean.

For example, say you wanted to compare the psychometric intelligence of the students in your class against the intelligence of the general population. Standardized intelligence tests (e.g., WAIS, Stanford-Binet) have a standardized population mean of 100. Say, 10 students in your class are given an intelligence test; their data is presented below.

<table>
<thead>
<tr>
<th>ID</th>
<th>IQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>115</td>
</tr>
<tr>
<td>2</td>
<td>112</td>
</tr>
<tr>
<td>3</td>
<td>106</td>
</tr>
<tr>
<td>4</td>
<td>114</td>
</tr>
<tr>
<td>5</td>
<td>108</td>
</tr>
<tr>
<td>6</td>
<td>117</td>
</tr>
<tr>
<td>7</td>
<td>125</td>
</tr>
<tr>
<td>8</td>
<td>119</td>
</tr>
<tr>
<td>9</td>
<td>99</td>
</tr>
<tr>
<td>10</td>
<td>109</td>
</tr>
</tbody>
</table>

Following the data entry procedures described in Chapters 1 & 2, label the first variable *ID* and the second variable *IQ*. Enter the data appropriately.

**Assumptions**

Assumptions of the one-sample $t$-test include normality and absence of univariate outliers. Normality requires that the sampling distribution follow a relatively normal distribution, which for can be approximated by having normally distributed raw scores. Univariate outliers are identified by evaluating $z$-scores and a histogram (see Chapters 14 and 10 or 15, respectively) of your dependent variable. While $z$-scores greater than 3.29 or less than -3.29 may suggest an outlier, you need to determine whether the value is disconnected from the rest of the distribution. If it is not, the assumption is satisfied.

**Analysis**

Assuming normality and absence of outliers, to conduct a one-sample $t$-test based on a population mean of 100:

1. Click **Analyze**, choose **Compare Means**, and finally, click **One-Sample T Test…** to bring up the *One-Sample T Test* dialog box shown in Figure 24.1.

![Figure 24.1](image-url)
2. Notice all the variables in your dataset are listed on the left hand side of this dialog box. Click the IQ variable and then click the right arrow ( ) next to the Test Variable(s): box. Note that you can test as many variables as you want against the population mean.

3. In the Test Value: box, enter the population mean. For the example, it is 100.

4. You can click the Options… button to bring up the One-Sample T Test: Options dialog box. Here, you can select the Confidence Interval Percentage: as well as indicate how you want to handle Missing Values. When you are satisfied, click Continue.

5. Click OK to run the analysis.

Output

When you click OK, SPSS will produce an Output screen displaying your results. Click the Output window to view your results (if it does not pop up automatically). Your results should be identical to Figure 24.2. The first table, labeled One-Sample Statistics, provides you with the descriptive statistics of your sample and contains your dependent variable (IQ), sample size (N), mean, standard deviation (Std. Deviation), and standard error of the mean (Std. Error Mean). The second table is labeled One-Sample Test. At the top of this table is your test mean (Test Value = 100). This table contains your t value (t), degrees of freedom (df), significance level or p-level (Sig. 2-tailed), mean difference, and confidence intervals.

![Table 1: One-Sample Statistics](image1)

![Table 2: One-Sample Test](image2)

Figure 24.2

Interpretation

Interpretation begins by examining the statistical significance of your one-sample t-test provided in the One-Sample Test table. If this result is statistically significant, your sample mean is statistically different from the population mean. Next, evaluate your sample mean, provided in the One-Sample Statistics table, to the population mean.

Example Results Section

Prior to analysis, no violation of normality was indicated and no outliers were identified.

Results of a one-sample t-test indicated a statistically significant difference between the mean intelligence of the class \((M = 112.40; SD = 7.336)\) and the population mean of 100, \(t(9) = 5.345, \ p < .05\).
Chapter 25
INDEPENDENT-SAMPLES t-TEST

An independent-samples t-test is used to assess for a mean difference between two mutually exclusive groups. Specifically, the participants within each group have been selected randomly and are assessed for differences on one continuous dependent variable.

For example, say you want to test whether your new sleep treatment is effective in increasing total minutes of sleep per night compared to placebo. The collected data is below.

<table>
<thead>
<tr>
<th>Placebo Group</th>
<th>Treatment Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Minutes</td>
</tr>
<tr>
<td>1</td>
<td>237</td>
</tr>
<tr>
<td>2</td>
<td>297</td>
</tr>
<tr>
<td>3</td>
<td>295</td>
</tr>
<tr>
<td>4</td>
<td>275</td>
</tr>
<tr>
<td>5</td>
<td>261</td>
</tr>
<tr>
<td>6</td>
<td>311</td>
</tr>
<tr>
<td>7</td>
<td>322</td>
</tr>
<tr>
<td>8</td>
<td>214</td>
</tr>
<tr>
<td>9</td>
<td>346</td>
</tr>
<tr>
<td>10</td>
<td>269</td>
</tr>
</tbody>
</table>

Following the data entry procedure described in Chapters 1 and 2, label the first variable ID, the second variable Group (0 = Placebo; 1 = Treatment), and the last variable Minutes. Notice the inclusion of the Group variable where you indicate which group each participant was randomized. Enter data appropriately, as shown partially in Figure 25.1.

Figure 25.1
Assumptions

The assumptions of an independent-samples t-test include absence of univariate outliers and homogeneity of variance as well as independence, normality, and homoscedasticity of residuals (see Chapter 22 for explanation). With that said, no t-test procedures in SPSS produce residual values; however, you can use the general linear model procedure used in Chapter 28.

Univariate (i.e., one variable) outliers are identified by evaluating z-scores and a histogram of dependent variable raw scores within each group individually (see Chapters 14 and 15, respectively). Although z-scores greater than 3.29 or less than -3.29 may suggest a possible outlier, you need to determine whether the value is actually disconnected from the rest of the distribution using within-group histograms. If outliers are identified you have several options. You can remove them from analysis through deletion (do not forget to describe them in your results section), transform the dependent variable (not advised), or use a nonparametric alternative that uses ranked data instead of the actual, raw scores (Chapter 26).

Homogeneity of variance requires that the dependent variable have roughly the same variance within both groups. If a violation is detected, your statistical test will become too liberal and the probability of type I error increases. This assumption is tested during analysis by Levene’s test, where a non-statistically significant result indicates the variances are not statistically different, thereby satisfying the assumption. If a violation is indicated, however, the analysis adjusts (i.e., reduces) the degrees of freedom to make the test more conservative.

Independence of residuals requires that the residual for one participant not be correlated or associated with a residual for any other participant. Briefly, a residual is the difference between predicted and actual value of the dependent variable. Because the independent-samples t-test is a special case of the general linear model (see Chapters 22 and 23 for slightly more detailed discussion), each participant will have their own residual value. Independence of residuals is technically a design issue that requires no repeated measurement or clustering. It can be assured by using appropriate random sampling and random assignment techniques.

The normality assumption requires the residuals to be normally distributed. This assumption is evaluated by a Q-Q plot, a histogram, and Kolmogorov-Smirnov test of residuals (Chapter 15). However, simulation studies have shown that the analysis is robust (i.e., still provide correct inference) to violation of normality with greater than 20 error degrees of freedom and relatively equal group sizes.

Homoscedasticity requires the residuals to have a constant variance across all values the dependent variable. With a categorical independent variable, violation of homoscedasticity goes hand-in-hand with a violation of the homogeneity of variance assumption described above. Thus, if homogeneity of variance is satisfied, homoscedasticity of residuals usually is too. Homoscedasticity is evaluated by plotting (Chapter 19) of the predicted dependent variable values on the x-axis and the standardized residuals on the y-axis. If this scatterplot appears rectangular, the assumption is considered satisfied.

Again, residual values cannot be requested using the procedure described in this Chapter. However, they can be requested using the general linear model procedure shown in Chapter 28. Because an independent-samples t-test and one-way between-subjects ANOVA are mathematically equivalent through the general linear model, any software that conducts one-way between-subjects ANOVA can also conduct an independent-samples t-test. Remember, $t^2 = F$ with one degree of freedom between groups (i.e., two groups). Thus, if you take the example data above and use the procedure described in Chapter 28, inference will be identical.
Analysis

Assuming independence, normality, and absence of outliers were satisfied prior to analysis, to conduct an independent-samples $t$-test on the example data above:

1. Click **Analyze**, then choose **Compare Means**, and finally click **Independent-Samples T Test**… to bring up the **Independent-Samples T Test** dialog box, shown in Figure 25.2.

![Figure 25.2](image)

2. Click the **Minutes** variable on the left hand side of the dialog box and then click the right arrow ( → ) next to the **Test Variable(s):** box.

3. Next, click to highlight **Group** and then click the right arrow ( → ) next to the **Grouping Variable:** box.

4. Click the **Define Groups**… button underneath the **Grouping Variable:** box.
   
   a. Here, you are asked to specify which groups you are testing (this requires you to remember how you coded the **Group** variable). You have two options. First, you can **Use specified values**. Here, you will tell SPSS specifically what the group codes are. Second, you can use a **Cut point:** where you can enter a number that splits the values of the grouping variable into two sets. That is, all participants with values less than the cut point form one group, and participants with values that are greater than or equal to the cut point form the other group. This option is useful for median splits.

   b. Because we already coded our grouping variable, we will **Use specified values**. Type **0** in the **Group 1:** box and **1** in the **Group 2:** box. Click **Continue**.

5. In addition, you can click the **Options**… button to bring up the **Independent-Samples T Test: Options** dialog box where you can select the **Confidence Interval Percentage:** or indicate how you want SPSS to handle **Missing Values**. When you are satisfied, click **Continue**.

6. That’s it! Click **OK** to conduct the analysis.
Output

When you click OK, SPSS will produce an Output screen displaying your results. Click the Output window to view your results (if it does not pop up automatically).

The first table you see is titled Group Statistics shown in Figure 25.3. This table provides you with descriptive statistics including the sample size (N), Mean, standard deviation (Std. Deviation), and standard error of the mean (Std. Error Mean) for each group individually.

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minutes</td>
<td>10</td>
<td>282.70</td>
<td>30.789</td>
<td>12.576</td>
</tr>
<tr>
<td>Treatment</td>
<td>10</td>
<td>421.10</td>
<td>56.384</td>
<td>17.930</td>
</tr>
</tbody>
</table>

Figure 25.3

The second table, titled Independent Samples Test, is shown in Figure 25.4. This table contains the results of your independent-samples t-test. Note there are two rows of results within this table. The row you evaluate is determined by whether the homogeneity of variance assumption is satisfied, provided by F-test and p-value (Sig.) in the Levene’s Test for Equality of Variances section. Finally, the t-test for Equality of Means section provides the t-value (t), degrees of freedom (df), significance level (Sig. 2-tailed), Mean Difference, the standard error for this difference (Std. Error Difference), and Confidence Interval of the Difference.

<table>
<thead>
<tr>
<th></th>
<th>Levene’s Test for Equality of Variances</th>
<th>Independent Samples Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>Sig</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minutes</td>
<td>7.36</td>
<td>.402</td>
</tr>
<tr>
<td></td>
<td>Equal variances assumed</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Equal variances not assumed</td>
<td></td>
</tr>
</tbody>
</table>

Figure 25.4

Interpretation

Interpreting the results requires the information provided by both tables above. First, you will evaluate the homogeneity of variance assumption, tested by Levene’s test in the Independent Samples Test table. If Levene’s test is not statistically significant (i.e., p > .05), you consult the Equal variances assumed row; however, if Levene’s test is statistically significant, you consult the Equal variances not assumed row.

Next, you evaluate the results of your independent-sample t-test using the Sig. 2-tailed column. If this result is statistically significant, the means of your two groups are statistically
different. The *Mean Difference* and *95% Confidence Interval of the Difference* columns are the used to indicate the size of the difference and precision, respectively.

Finally, to complete interpretation, you will evaluate the group means provided in the *Group Statistics* table to determine which group actually had a higher (or lower) mean.

**Example Results Section**

Prior to analysis, no violation of independence or normality was indicated, and no outliers were identified. Further, no violation of homogeneity of variance was indicated by Levene’s test, $F = .736, p > .05$.

The results of an independent-samples $t$-test indicated a statistically significant difference in minutes of sleep per night between the two groups, $t(18) = -6.343, p < .05$, with the group receiving the new sleep treatment sleeping significantly longer than the placebo group (mean difference = 138.400 minutes, $SE = 21.819$, 95% CI = 92.560 to 184.240 minutes).
Chapter 26  
MANN-WHITNEY TEST

The Mann-Whitney test is the nonparametric alternative to the independent-samples $t$-test and is one of the most powerful nonparametric tests. It is used to assess for differences in two mutually exclusive groups when the normality of residuals assumption is violated (see Chapter 25), you have numerous outliers, or the dependent variable is measured on an ordinal scale.

The Mann-Whitney test is based on ranked data. That is, instead of using the actual values of the dependent variable, each dependent variable value is ranked (not considering groups), with the highest value receiving the highest rank and the lowest value receiving the lowest rank. The ranks within each group individually are then summed, and the test assesses whether the difference in ranked sums between groups is statistically different.

As an example, consider a performance improvement study assessing for differences in patient satisfaction of hospital stay following total hip replacement surgery. Patients were randomized to surgery and recovery in a new facility or an older facility. Surgical outcomes were expected to be identical. The measurement instrument used a Likert-type scale with five possible responses anchored from Unacceptable to Outstanding. The collected data is below:

<table>
<thead>
<tr>
<th>ID</th>
<th>Rating</th>
<th>ID</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>17</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>18</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>19</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>20</td>
<td>2</td>
</tr>
</tbody>
</table>

Following the data entry procedures described in Chapters 1& 2, label the first variable ID, the second variable Group (0 = Old; 1 = New), and the last variable Rating. Notice the inclusion of the Group variable where you indicate which group each participant was randomized (see Figure 25.1).

Analysis

1. Click Analyze, then choose Nonparametric tests, choose Legacy Dialogs, and finally click 2 Independent Samples… to bring up the Two-Independent-Samples Tests dialog box.

2. Click the Rating variable on the left hand side of the dialog box and then click the right arrow (ве) next to the Test Variable List: box.

   a. Note that you can compare as many dependent variables as you want.
3. Next, click to highlight **Group** and then click the right arrow (→) next to the **Grouping Variable:** box.

4. Click the **Define Groups...** button underneath the **Grouping Variable:** box to bring up the **Two Independent Samples: Define Groups** dialog box.
   a. Here, you are asked to indicate which groups you are testing (this requires you to remember how you coded the **Group** variable). You must tell SPSS specifically what the group codes are. Type **0** in the **Group 1:** box and type **1** into the **Group 2:** box. Click **Continue**.

5. Next, you must indicate what the **Test Type** is. Here, you have four options, with the default being **Mann-Whitney U**.

6. Finally, you can click the **Options...** button, where you can select two descriptive **Statistics** or identify how you want SPSS to handle your missing data.
   a. The **Descriptive** option displays the mean, standard deviation, minimum, maximum, and the number of non-missing values. The **Quartiles** option displays values corresponding to the 25th, 50th, and 75th percentiles.
   b. Under the **Missing Values** option, choosing to **Exclude cases test-by-test** evaluates each test separately for missing values. Choosing **Exclude cases listwise** excludes participants with missing values for any variable from all analyses.

When you are satisfied with your selections, click **Continue**.

7. Your **Two-Independent-Samples Tests** dialog box should appear similar to Figure 26.1. Click **OK** to conduct the Mann-Whitney test.

---

**Output**

When you click **OK**, SPSS will produce an **Output** screen displaying your results. Click the **Output** window to view your results (if it does not pop up automatically).

You are presented with two tables containing the results of the Mann-Whitney test. The first table is titled **Ranks**, and is shown at the top of Figure 26.2. Here, you are provided with the sample size \((N)\), **Mean Rank**, and **Sum of Ranks** for both levels of your grouping variable. In the **Test Statistics** table, shown at the bottom of Figure 26.2, you are provided with several test
statistics. The first two rows contain the Mann-Whitney U and Wilcoxon W values. While I will not describe the calculation of these values, it is important to know that for group sizes above 10 these both statistics approximate a normal distribution. Thus, statistical significance can be determined from a z-test, which is provided in the Z row. Note, the negative Z indicates the rank sums are lower than their expected values. Finally, you are presented with your p values. You will usually be presented with two p values. The first is the two-tailed asymptotic significance (Asymp. Sig. (2-tailed)). This value is only applicable for larger samples (N > 10 per group) and is the significance value if the normal distribution was used appropriately to approximate the Mann-Whitney U or Wilcoxon W values. The second is the two-tailed exact significance (Exact Sig. [2*(10tailed Sig.)]). This value is appropriate for smaller samples. Note that as the sample size increases, the asymptotic and exact significance values will converge. Unless you have a huge sample size, it is usually better to report and interpret the exact significance value.

### Mann-Whitney Test

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean Rank</th>
<th>Sum of Ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating</td>
<td>10</td>
<td>7.26</td>
<td>72.60</td>
</tr>
<tr>
<td>Rating</td>
<td>10</td>
<td>13.98</td>
<td>139.60</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Statistics</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mann-Whitney U</td>
<td>17.600</td>
</tr>
<tr>
<td>Wilcoxon W</td>
<td>72.600</td>
</tr>
<tr>
<td>Z</td>
<td>-2.581</td>
</tr>
<tr>
<td>Asymp. Sig. (2-tailed)</td>
<td>.010</td>
</tr>
<tr>
<td>Exact Sig. [2*(10tailed Sig.)]</td>
<td>.011*</td>
</tr>
</tbody>
</table>

a. Not corrected for ties
b. Grouping Variable: Group

### Interpretation

Interpretation begins by examining the result of the Mann-Whitney test provided in the Test Statistics table. If the Z value is statistically significant (Asymp. Sig. (2-tailed) < .05), your group have statistically different mean ranks. Interpretation is completed by evaluating the Mean Rank for each group, provided to you in the Ranks table to see which group had higher (or lower) mean ranks.

### Example Results Section

The results of the Mann-Whitney test indicated a statistically significant difference in ranked patient satisfaction between the old and new surgical facilities, $Z = -2.581$, exact $p < .05$, with patients in the new facility indicating greater ranked satisfaction on average compared to patients in the old facility.
Chapter 27
MEDIAN TEST

The median test is used to assess whether two or more mutually exclusive groups have different medians. It can be used appropriately for skewed data or data containing outliers. It can also be used when the dependent variable is measured on an ordinal scale. It should be noted that there is no parametric alternative to the median test; however, the Mann-Whitney test (Chapter 26) can be used as an adequate alternative for interval level data and Pearson’s chi-square test (Chapter 37) can be used for ordinal level data with larger samples. The test calculates the medians within each group and then classifies the data within each group as either above or below the respective group median. To show similarities between the median and Mann-Whitney tests, we will use the data and example provided in Chapter 26.

As a reminder, you are considering a performance improvement study assessing differences in patient satisfaction of hospital stay following total hip replacement surgery. Patients were randomized to surgery and recovery in either a new facility or an older facility and surgical outcomes were expected to be identical. The measurement instrument used a Likert-type scale with five possible responses anchored from Unacceptable to Outstanding. The collected data is provided below:

<table>
<thead>
<tr>
<th>Old Facility</th>
<th>New Facility</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ID</strong></td>
<td><strong>Rating</strong></td>
</tr>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
</tr>
</tbody>
</table>

Following the data entry procedures described in Chapters 1 & 2, label the first variable ID, the second variable Group (0 = Old; 1 = New), and the last variable Rating. Notice the inclusion of the Group variable where you indicate which group each participant was randomized (see Figure 25.1).

Analysis

1. Click **Analyze**, then choose **Nonparametric tests**, choose **Legacy Dialogs**, and finally click **K Independent Samples…** to bring up the **Tests for Several Independent Samples** dialog box.

2. Click the **Rating** variable on the left hand side of the dialog box and then click the right arrow ( ) next to the **Test Variable List: box**.

   a. Note that you can compare as many dependent variables as you want.

3. Next, click to highlight **Group** and then click the right arrow ( ) next to the **Grouping Variable: box**.
4. Click the **Define Groups**… button underneath the **Grouping Variable** box.
   a. Here, you are asked to indicate the range of groups you are testing (this requires you to remember how you coded the **Group** variable). Type 0 in the **Minimum** box and type 1 into the **Maximum** box. Click **Continue**.

5. Next, you must indicate what the **Test Type** is. Here, you have three options. You will need to uncheck the **Kruskal-Wallis H** checkbox, and check the **Median** checkbox.

6. Finally, you can click the **Options**… button, where you can select two descriptive **Statistics** or identify how you want SPSS to handle your missing data.
   a. The **Descriptive** option displays the mean, standard deviation, minimum, maximum, and the number of non-missing values. The **Quartiles** option displays values corresponding to the 25th, 50th, and 75th percentiles.
   b. Under the **Missing Values** option, choosing to **Exclude cases test-by-test** evaluates each test separately for missing values. Choosing **Exclude cases listwise** excludes participants with missing values for any variable from all analyses.

When you are satisfied with your selections, click **Continue**.

7. That’s it! Click **OK** to conduct the median test.

**Output**

When you click **OK**, SPSS will produce an **Output** screen displaying your results. Click the **Output** window to view your results (if it does not pop up automatically) which should be identical to those presented in Figure 27.1.

The first table is titled **Frequencies**. Here, you are provided with the frequency or count of the number of participants within each group who scored above the median or at or below the median. The second table, titled **Test Statistics**, provides you with the results of your statistical test. The first row contains the total sample size (**N**). The second row indicates the **Median**. Remember, this median is provided for the sample as a whole, regardless of group. Finally, you are presented with your **p**-value (**Exact Sig.**).

---

**Median Test**

<table>
<thead>
<tr>
<th>Frequencies</th>
<th>Group</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Old</td>
<td>New</td>
<td></td>
</tr>
<tr>
<td>Rating</td>
<td>&gt; Median</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>&lt;= Median</td>
<td>6</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Statistics²</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>20</td>
</tr>
<tr>
<td>Median</td>
<td>3.00</td>
</tr>
<tr>
<td>Exact Sig.</td>
<td>.070</td>
</tr>
</tbody>
</table>

---

**Figure 27.1**

130
Interpretation

Interpretation begins by examining the results of the median test provided in the Test Statistics tables. If the Exact Sig. is statistically significant, you will use the crosstabulation in the Frequencies table to determine which group had higher number of participants above (or at or below) the median.

You will notice immediately that the result for the example data was not statistically significant, whereas the same data using the Mann-Whitney test was statistically significant. This is because the median test has extremely low power (efficiency) compared to the Mann-Whitney test. For this reason, the median test is not used very often and is becoming obsolete. However, you may occasionally come across it in recently published journal articles, which is why it is described in this workbook.

Example Results Section

The results of the median test indicated no statistically significant difference in patient satisfaction between surgeries conducted in the old and new facilities, $Mdn = 3.00$, exact $p > .05$. That is, patients in the old facility responded similarly to patients in the new facility.
Chapter 28
ONE-WAY BETWEEN-SUBJECTS ANOVA

One-way between-subjects analysis of variance (ANOVA) is used when you have one continuous dependent variable and three (or more) mutually exclusive groups (or levels) within one independent variable. It is an extension of the independent-samples t-test to situations where you have three or more groups. As the name suggests, ANOVA assesses differences in the variance between groups.

ANOVA uses an F-test to evaluate overall group differences. This F-test is an omnibus test meaning that a statistically significant result only indicates that two of the groups differ, but does not indicate which groups are different. Thus, if you did not specify planned comparisons, post hoc tests are required. Post hoc tests can be thought of as adjusted independent-samples t-tests comparing all possible two-group comparisons. The adjustment is used to control the increased probability of type I error from doing multiple comparisons. There are numerous post hoc tests available and a full description is beyond the scope of this Chapter. So, while the example below uses the Tukey post hoc test, alternatives could have been used appropriately.

For example, say you want to assess the value of your new sleep treatment compared to a placebo group and a group receiving a popular over-the-counter treatment. If the omnibus F is statistically significant, you are interested in all between-group comparisons. The collected data is listed below:

<table>
<thead>
<tr>
<th>ID</th>
<th>Minutes</th>
<th>ID</th>
<th>Minutes</th>
<th>ID</th>
<th>Minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>250</td>
<td>11</td>
<td>356</td>
<td>21</td>
<td>480</td>
</tr>
<tr>
<td>2</td>
<td>219</td>
<td>12</td>
<td>369</td>
<td>22</td>
<td>477</td>
</tr>
<tr>
<td>3</td>
<td>301</td>
<td>13</td>
<td>348</td>
<td>23</td>
<td>433</td>
</tr>
<tr>
<td>4</td>
<td>245</td>
<td>14</td>
<td>402</td>
<td>24</td>
<td>447</td>
</tr>
<tr>
<td>5</td>
<td>234</td>
<td>15</td>
<td>399</td>
<td>25</td>
<td>467</td>
</tr>
<tr>
<td>6</td>
<td>215</td>
<td>16</td>
<td>387</td>
<td>26</td>
<td>419</td>
</tr>
<tr>
<td>7</td>
<td>225</td>
<td>17</td>
<td>361</td>
<td>27</td>
<td>500</td>
</tr>
<tr>
<td>8</td>
<td>269</td>
<td>18</td>
<td>330</td>
<td>28</td>
<td>482</td>
</tr>
<tr>
<td>9</td>
<td>175</td>
<td>19</td>
<td>305</td>
<td>29</td>
<td>478</td>
</tr>
<tr>
<td>10</td>
<td>278</td>
<td>20</td>
<td>370</td>
<td>30</td>
<td>451</td>
</tr>
</tbody>
</table>

Following the data entry procedures described above in Chapters 1 and 2, label the first variable ID, the second variable Group (1 = Placebo; 2 = OTC; 3 = New), and the final variable Minutes. Enter data appropriately. Notice the inclusion of the Group variable where you indicate which group each participant was randomized (see Figure 25.1).

Assumptions

The assumptions of the one-way between-subjects ANOVA include absence of univariate outliers and homogeneity of variance as well as independence, normality, and homoscedasticity of residuals.

Univariate (i.e., one variable) outliers are identified by evaluating z-scores and a histogram of dependent variable raw scores within each group individually (see Chapters 14 and
Although z-scores greater than 3.29 or less than -3.29 may suggest a possible outlier, you need to determine whether the value is actually disconnected from the rest of the distribution using within-group histograms. If outliers are identified you have several options. You can remove them from analysis through deletion (do not forget to describe them in your results section), transform the dependent variable (not advised), or use a nonparametric alternative that uses ranked data instead of the actual, raw scores (Chapter 29).

Homogeneity of variance requires that the dependent variable have roughly the same variance within all groups. If a violation is detected, your statistical test will become too liberal and the probability of type I error increases. This assumption is tested during analysis by Levene’s test, where a non-statistically significant result indicates the variances are not statistically different, thereby satisfying the assumption. If a violation is indicated, however, you should make the omnibus test more conservative by reducing alpha (e.g., from .05 to .01).

Independence of residuals requires the residuals to be uncorrelated across participants. Briefly, a residual is the difference between predicted and actual value of the dependent variable. Because one-way between-subjects ANOVA is a special case of the general linear model (see Chapters 22 and 23 for slightly more detailed discussion), each participant will have their own residual value. Independence of residuals is technically a design issue that requires no repeated measurement or clustering. It can be assured by using appropriate random sampling and random assignment techniques.

The normality assumption requires the residuals to be normally distributed. This assumption is evaluated by a Q-Q plot, a histogram, and Kolmogorov-Smirnov test of residuals (Chapter 15). With that said, simulation studies have shown that the analysis is robust (i.e., still provide correct inference) to violation of normality with greater than 20 error degrees of freedom and relatively equal group sizes. We will request residuals below, however.

Homoscedasticity requires the residuals to have a constant variance across all values the dependent variable. With a categorical independent variable, violation of homoscedasticity goes hand-in-hand with a violation of the homogeneity of variance assumption described above. Thus, if homogeneity of variance is satisfied, homoscedasticity of residuals usually is too. Homoscedasticity is evaluated by requesting a scatterplot matrix where the predicted dependent variable values are plotted against the standardized residuals (see Step 3e below). If this scatterplot appears rectangular, the assumption is considered satisfied.

### Analysis

Assuming independence and absence of outliers are tenable prior to analysis, to conduct a one-way between-subjects ANOVA:

1. Click **Analyze**, then choose **General Linear Model**, finally click **Univariate**… to bring up the **Univariate** dialog box, shown in Figure 28.1.

2. For this type of design, you will only need to consider two of the five boxes—the **Dependent Variable**: box, where you will place your dependent variable, and the **Fixed Factor(s)**: box, where you will place your independent variable.
   a. Click to select the **Minutes** variable and then click the right arrow (→) next to the **Dependent Variable**: box.
   b. Click to select the **Group** variable and then click the right arrow (→) next to the **Fixed Factor(s)**: box.
3. You will also notice six buttons on the right side of this dialog box. For one-way between-subjects ANOVA, only five of the six buttons are applicable. Note that the Bootstrap... button is only available if you have the Exact Tests add-on.

   a. Clicking the Contrasts... button allows you to test for differences among levels of the independent variable you specified in Step 2b above. You can select from several different contrasts described below. Note the default is None.

      i. Deviation: Compares the mean of each level, other than a reference category, to the grand mean.

      ii. Simple: Compares the mean of each level to the mean of a specified level. This is useful when you have a control group.

      iii. Difference: Compares the mean of each level (except the first) to the mean of previous levels.

      iv. Helmert: Compares the mean of each level (except the last) to the mean of subsequent levels.

      v. Repeated: Compares the mean of each level (except the last) to the mean of the subsequent level.

      vi. Polynomial: Compares the linear effect, quadratic effect, cubic effect, and so on depending on how many levels you have. The first degree of freedom contains the linear effect across all categories; the second degree of freedom, the quadratic effect; and so on. These contrasts are used to perform a trend analysis.

If you choose to Change Contrast, click the down arrow (▼) in the Contrast: box, select the contrast you want from the drop-down menu, and click Change. For the Deviation and Simple contrasts, you will be asked what your Reference Category should be, choose either Last or First. Remember, this is based on your coding scheme, with Last indicating your highest coded value and First indicating your lowest coded value. When you are satisfied, click Continue.
b. Clicking the **Plots…** button allows you to create a line chart plotting the estimated marginal means for all levels within your independent variable.

i. For the example above, click **Group** and then click the right arrow (→) next to the *Horizontal Axis* box. This tells SPSS to place the treatment groups on the x-axis.

ii. Click **Add** to move this instruction into the **Plots** box.

iii. Click **Continue**.

c. Click the **Post Hoc…** button. Here, you will select the post hoc tests you want to use based on whether the homogeneity of variance assumption is satisfied. If the assumption is satisfied, they you will select a post hoc test under the *Equal Variances Assumed* section. The 14 post hoc tests listed vary along a continuum of conservative to liberal. In general, it is probably better to go with a slightly more conservative test, with the **Tukey** or **Scheffe** being widely used options. If homogeneity of variance is violated, you will select a post hoc test under the *Equal variances Not Assumed* section. The **Games-Howell** and **Dunnett’s C** are used widely.

i. Click to select the **Group** variable and then click the right arrow (→) next to the *Post Hoc Tests for:* box.

ii. We are going to assume homogeneity of variance is satisfied (although it will be tested later), so click the **Tukey** checkbox.

iii. Click **Continue**.

d. Click the **Save…** button. Here, you have the option to save new variables to your dataset that contain information about **Predicted Values, Diagnostics, Residuals,** or **Coefficient Statistics.**

i. The options listed under the **Predicted Values** section allow you to save the values that your specific model predicts for each participant. You can save these values as **Unstandardized, Weighted** (available only if a weighted least squares variable was selected; **WLS Weight**: in Figure 28.1), or **Standard error** where an estimate of the standard deviation of the average value of the dependent variable for participants that have the same values of the independent variables.

ii. The options listed under the **Diagnostics** section provide measures to identify participants with unusual combinations of values for the independent variables or those who have unusual influence on the outcome of the model. For both options, larger values equal larger change or influence.

1. **Cook’s distance** is a measure of how much the residuals of all participants would change if the participant were removed from analysis.

2. **Leverage values** indicate the relative influence of each observation on the model’s fit.
iii. The options listed under the Residuals section provide the difference between the observed value of the dependent variable and the model predicted value.

1. **Unstandardized**: Provides the difference between an observed value and the value predicted by the model.

2. **Weighted**: Provides the weighted unstandardized residuals and is only available if a weighted least squares variable was previously selected (WLS Weight: in Figure 30.1).

3. **Standardized**: Provides the residual divided by an estimate of its standard deviation. These are also known as Pearson residuals. Similar to z-scores, they have a mean of 0 and a standard deviation of 1. The distribution of these residuals is used to evaluate normality.

4. **Studentized**: Provides the residual divided by an estimate of its standard deviation. However, the standard deviation is allowed to vary across participants depending on the distance of each participant's values on the independent variables from the means of the independent variables.

5. **Deleted**: Provides the residual for a participant if that participant is excluded from analysis. It is the difference between the value of the dependent variable and the adjusted predicted value.

iv. Finally, the options listed under the Coefficient Statistics section writes the variance-covariance matrix of parameter estimates, t-statistics, significance values, and residual degrees of freedom. You have the option to Create a new dataset where you must specify a Dataset name: or you can Write a new data file which creates a complete new dataset. Selecting the latter requires you to click the File... button and specify the dataset you want to write into.

When you are satisfied with your selections, click **Continue**.

e. Finally, click the **Options**... button. Here, you can tell SPSS to print a table of Estimated Marginal Means as well a number of descriptive statistics and diagnostic tests.

i. Under the Estimated Marginal Means section, in the Factor(s) and Factor Interactions: box, you will always be offered the option to display (OVERALL) means. If you choose this option, SPSS will print the grand mean, which considers all the data in your sample together irrespective of the independent variable groups. More importantly, however, this section allows you to display the estimated marginal means for each level of your independent variable.

1. Click **Group**, and then click the right arrow (→) next to the Display Means for: box.
ii. You can also *Compare main effects*. However, clicking this checkbox is redundant with to the post hoc tests you selected in Step 3c above, so it is never checked for one-way between-subjects ANOVA.

iii. Under the *Display* section, you have ten options. *Descriptive statistics* prints weighted means, standard deviations, and frequency counts for all levels of your independent variable. *Estimates of effect size* prints partial eta-squared, which is the proportion of total variability attributable to your independent variable. The *Observed power* option prints the power of the test based on your specific sample. The *Parameter estimates* option prints parameter estimates, standard errors, t-tests, and confidence intervals (similar to regression; Chapters 22 and 23). Clicking *Contrast coefficient matrix* prints the L matrix, but this option is not useful without covariates. *Homogeneity tests* produces Levene’s test and the *Spread-versus-level plot* compares cell means, standard deviations, and variances across the level combinations of all groups. Choosing a *Residual plot* produces a scatterplot matrix of observed, predicted, and residual values. It is useful for checking the homoscedasticity assumption. Finally, the *Lack of fit* option is used to check if the relationship between the dependent variable and independent variables can be described adequately by the model and the *General estimable function* allows you to construct custom hypotheses. You can also specify the *Significance level:* for your confidence intervals.

When you are satisfied with your selections, click **Continue**.

4. That’s it! All you need to do now is click **OK** to conduct the one-way between-subjects ANOVA.

**Output**

When you click **OK**, SPSS will produce an **Output** screen displaying your results. Click the **Output** window to view your results (if it does not pop up automatically).

The first table you see is titled *Between-Subjects Factors* containing the sample sizes ($N$) of your groups within each independent variable. The second table presented is titled *Test of Homogeneity of Variances*, shown in Figure 28.2. Here, you are provided with Levene’s statistic ($F$), numerator degrees of freedom ($df1$), denominator degrees of freedom ($df2$), and the statistical significance level ($Sig.$).

![Figure 28.2](image)

Levene's Test of Equality of Error Variances

<table>
<thead>
<tr>
<th>Dependent Variable: Minutes</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F$</td>
<td>$df1$</td>
<td>$df2$</td>
<td>$Sig.$</td>
</tr>
<tr>
<td></td>
<td>3.83</td>
<td>2</td>
<td>27</td>
<td>0.885</td>
</tr>
</tbody>
</table>

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + Group
The third table, titled *Between-Subjects Effects*, is shown in Figure 28.3. Do not be overwhelmed by the information provided in this table, it is explained in the Interpretation section below. This table contains the $F$-test, $p$-value ($\text{Sig.}$), and effect size (*Partial Eta Squared*) for your independent variable (i.e., the *Group* row). You are also provided with *Sums of Squares*, degrees of freedom ($df$), and *Mean Squares*. Finally, at the bottom of this Table is the overall effect size ($R^2$), which describes the proportion of variance in the dependent variable that is accounted for by the independent variable. Note that the $R^2$ value is identical to the *Partial Eta Squared* value found in the table. Also, *Adjusted R Squared* is simply the $R^2$ value adjusted for sample size. Thus, it will always be slightly smaller than the $R^2$ value; however, large differences between these two indices suggest model over-fit (i.e., little to no generalizability).

![Test of Between-Subjects Effects Table]

The next section is titled *Estimated Marginal Means*. Here, you will find descriptive statistics for all levels of your independent variable in the table labeled, *Group*, shown in Figure 28.4. Note that with equal group sizes this table will contain *Mean* values identical to those you could have calculated yourself using the Frequencies or Explore procedures (Chapters 10 and 15, respectively). However, these values will diverge as group sizes become more unequal; thus, you should always report the means presented in the *Estimated Marginal Means* table (because they are unweighted). This is also why I did not have you request *Descriptive statistics* in Step 3e above.

![Estimated Marginal Means Table]
The results of the Tukey post hoc tests are provided next under the *Post Hoc Tests* section in the *Multiple Comparisons* table, shown in Figure 28.5. The information provided in this table assesses for differences between all two-group combinations. Results include the *Mean Difference*, standard error (*Std. Error*), statistical significance level (*Sig.*), and 95% *Confidence Intervals*. Within the *Mean Difference (I-J)* column, an asterisk (*) following any value indicates a significant difference between groups. Note that half of this table is redundant. That is, the comparison between *Placebo* and *OTC* is also presented further down the table as *OTC* and *Placebo*.

### Multiple Comparisons

<table>
<thead>
<tr>
<th>(I) Group</th>
<th>(J) Group</th>
<th>Mean Difference (I-J)</th>
<th>Std. Error</th>
<th>Sig.</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placebo</td>
<td>OTC</td>
<td>-121.60*</td>
<td>13.762</td>
<td>.000</td>
<td>-155.72</td>
<td>-87.48</td>
</tr>
<tr>
<td></td>
<td>New</td>
<td>-222.30*</td>
<td>13.762</td>
<td>.000</td>
<td>-258.42</td>
<td>-188.18</td>
</tr>
<tr>
<td>OTC</td>
<td>Placebo</td>
<td>121.60*</td>
<td>13.762</td>
<td>.000</td>
<td>87.48</td>
<td>155.72</td>
</tr>
<tr>
<td></td>
<td>New</td>
<td>-100.70*</td>
<td>13.762</td>
<td>.000</td>
<td>-134.82</td>
<td>-66.58</td>
</tr>
<tr>
<td>New</td>
<td>Placebo</td>
<td>222.30*</td>
<td>13.762</td>
<td>.000</td>
<td>188.18</td>
<td>256.42</td>
</tr>
<tr>
<td></td>
<td>OTC</td>
<td>100.70*</td>
<td>13.762</td>
<td>.000</td>
<td>55.58</td>
<td>134.82</td>
</tr>
</tbody>
</table>

Based on observed means.
The error term is Mean Square(Error) = 947.015.

* The mean difference is significant at the .05 level.

*Figure 28.5*

The next table, under the *Homogenous Subtests* section, is titled *Minutes*. This table contains mean for each group, and is organized in an obscure way to present the exact same information as the *Multiple Comparisons* table above. Here, the three columns indicate that each mean differed significantly from one another. Another way to interpret this table is that because each group has their own column, they are their own subset and all three groups are statistically different from one another. Anytime you see group means in the same column this indicates that they are not statistically different from one another, thus forming a subset.

Next, you are presented with the residual scatterplot matrix you requested in Step 3e above, shown in Figure 28.6. This matrix presents all bivariate scatterplots between the *Observed* value of the dependent variable, the *Predicted* value of the dependent variable, and the standardized residual value (*Std. Residual*). The scatterplot in the middle row of the last column (i.e., *Predicted* by *Std. Residual*) is the one used to evaluate the homoscedasticity assumptions. This scatterplot may appear different from those you have seen previously. The reason there are three distinct rows of data points within the scatterplot is that predicted values are identical within each group. Or, stated another way, predicted values are only different across groups. Thus, each row of data points can be considered an individual group—three groups, three rows.
Finally, you will find the line chart you requested in Step 3c. This chart is shown in Figure 28.7. This chart displays the estimated marginal means for the *Placebo*, *OTC*, and *Treatment* groups individually; thus, it is the exact same data presented in Figure 28.4, just presented graphically.
**Interpretation**

Interpretation begins by evaluating the tenability of the normality of residuals assumption. You will notice that in your dataset, you have saved (or created) a new variable labeled \( ZRE_1 \). Further, if you click on Variable View, the Label for this variable states *Standardized Residual for Minutes*. Use the Explore procedure (Chapter 15) to evaluate the normality of the standardized residuals. Note that residuals are now considered for the entire sample simultaneously, irrespective of group. From the histogram provided in Figure 28.8, you can see the distribution is relatively normal. This evaluation was further supported by the Q-Q plot and non-statistically significant Kolmogorov-Smirnov test \((p > .05)\). Thus, the normality assumption is considered satisfied.

![Figure 28.8](image)

Next, you will evaluate the homogeneity of variance assumption provided in the *Levene’s Test of Equality of Error Variances* table. In this example, Levene’s test is not statistically significant, so the assumption is considered satisfied. In addition, you will evaluate the homoscedasticity of residuals assumption found in the scatterplot matrix. As stated above, when homogeneity of variance is satisfied, homoscedasticity of residuals is usually satisfied as well. As you can see in the *Predicted by Std. Residual* scatterplot, the width of the data points for each group are relatively equal. Thus, this assumption is considered satisfied as well.

Next, you will evaluate the omnibus \( F \)-test found in the *Tests of Between-Subjects Effects* table. Although there are six rows of data, you are only concerned with two of them—*Group* and *Error*. The *Group* row provides the result of the statistical for your independent variable, whereas the *Error* row is used for reporting degrees of freedom for error. You can see in this table that the *Group* variable is statistically significant with a massive effect size. Remember, Partial Eta Squared ranges from 0 to 1, with 1 indicating 100% explained variance. The statistically significant effect of *Group* indicated that at least two groups differ statistically; however, because this \( F \)-test is omnibus, it fails to identify which groups are statistically different from each other. Thus, you must consult the *Post Hoc Tests* section of your output, specifically, the *Multiple Comparisons* table. Remember, you are only allowed to consult and interpret the post hoc comparisons if your omnibus \( F \) is statistically significant.
In the *Multiple Comparisons* table you are provided with six individual statistical tests; however, as stated above, half the table is redundant. Under the *Mean Difference (I-J)* column, you can see that all comparisons have asterisks, which indicates that all group differences are statistically significant. This information is confirmed by evaluating the *Sig.* column where all *p*-values are less than .05. Interpretation of the post hoc tests can be completed two ways. First, you could use the *Mean Difference (I-J)* column to determine which group had the higher mean. Alternatively, and easier in my opinion, is to consult the *Group* table under the *Estimated Marginal Means* section. Here, you can evaluate each group mean to see which group had the higher (or lower) mean.

**Example Results Section**

Prior to analysis, no violation of independence was indicated and no outliers were identified. No violation of normality or homoscedasticity of residuals was indicated, and homogeneity of variance was assured by Levene’s test, \(F(2, 27) = .383, p > .05\).

The results of a one-way between-subjects ANOVA indicated a statistically significant treatment effect, \(F(2, 27) = 130.840, p < .05\), partial \(\eta^2 = .906\). Analysis of post hoc Tukey tests indicated the new sleep treatment produced a significant increase in minutes of sleep per night in comparison to both placebo and the group receiving the over the counter treatment (\(M = 463.40, 241.10,\) and 362.70 minutes per night, respectively; both \(p < .05\)). Results also indicated the over the counter treatment produced a significant increase in minutes of sleep per night when compared to the placebo group (\(p < .05\)).
Chapter 29  
KRUSKAL-WALLIS TEST

The Kruskal-Wallis test is the nonparametric alternative to the one-way between-subjects ANOVA. The test is also an extension of the Mann-Whitney test to assess group differences between three or more mutually exclusive groups. The Kruskal-Wallis test is typically used when the normality of residuals assumption is violated (see Chapter 28), when you have numerous outliers, or when the dependent variable is measured on an ordinal scale. Similar to the Mann-Whitney test, the Kruskal-Wallis test is based on rank sums, where the dependent variable scores are ranked from highest to lowest and then summed within each group individually. Finally, statistically significant group differences are evaluated based on these ranked sums.

Because the Kruskal-Wallis is used for three or more groups, it is considered an omnibus test. That is, a statistically significant Kruskal-Wallis test indicates there is a statistically significant difference between at least two groups, but does not identify which groups differ. Thus, you will need to conduct a series of post hoc tests to test for differences between each group. The post hoc test most commonly used is the Mann-Whitney test described in Chapter 26.

Any time you conduct multiple post hoc comparisons, the probability of committing a type I error increases (same as Chapter 28). Thus, you must statistically adjust the alpha-level (i.e., .05) to keep this value close to the nominal (aka, familywise or experimentwise) level. The most common approach employs a Bonferroni adjustment, which divides your nominal alpha by the number of post hoc tests you plan on conducting. For example, consider a study with four mutually exclusive groups. Testing for differences between all groups will require six post hoc tests. Thus, the Bonferroni adjustment reduces alpha from .05 to .0083 (i.e., .05/6), so any post hoc test you conduct requires $p < .0083$ to be considered statistically significant.

To show the similarity between the Kruskal-Wallis test and the one-way ANOVA, reconsider the example from Chapter 28. Say you want to assess the value of your new sleep treatment compared to a placebo group and a group receiving a popular over-the-counter treatment. The collected data is listed below:

<table>
<thead>
<tr>
<th>ID</th>
<th>Minutes</th>
<th>ID</th>
<th>Minutes</th>
<th>ID</th>
<th>Minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>250</td>
<td>11</td>
<td>356</td>
<td>21</td>
<td>480</td>
</tr>
<tr>
<td>2</td>
<td>219</td>
<td>12</td>
<td>369</td>
<td>22</td>
<td>477</td>
</tr>
<tr>
<td>3</td>
<td>301</td>
<td>13</td>
<td>348</td>
<td>23</td>
<td>433</td>
</tr>
<tr>
<td>4</td>
<td>245</td>
<td>14</td>
<td>402</td>
<td>24</td>
<td>447</td>
</tr>
<tr>
<td>5</td>
<td>234</td>
<td>15</td>
<td>399</td>
<td>25</td>
<td>467</td>
</tr>
<tr>
<td>6</td>
<td>215</td>
<td>16</td>
<td>387</td>
<td>26</td>
<td>419</td>
</tr>
<tr>
<td>7</td>
<td>225</td>
<td>17</td>
<td>361</td>
<td>27</td>
<td>500</td>
</tr>
<tr>
<td>8</td>
<td>269</td>
<td>18</td>
<td>330</td>
<td>28</td>
<td>482</td>
</tr>
<tr>
<td>9</td>
<td>175</td>
<td>19</td>
<td>305</td>
<td>29</td>
<td>478</td>
</tr>
<tr>
<td>10</td>
<td>278</td>
<td>20</td>
<td>370</td>
<td>30</td>
<td>451</td>
</tr>
</tbody>
</table>

Following the data entry procedures described above in Chapters 1 and 2, label the first variable ID, the second variable Group (1 = Placebo; 2 = OTC; 3 = New), and the final variable Minutes. Enter data appropriately. Notice the inclusion of the Group variable where you indicate which group each participant was randomized (see Figure 25.1).
Assumptions

Although the Kruskal-Wallis test does not require the data be distributed normally, the test does assume that the groups have the same population distribution. Thus, you should evaluate the distributions for each group and make sure they are similar, regardless of shape. That is, make sure they are all skewed left, all skewed right, or all normally distributed.

Analysis

1. Click Analyze, then choose Nonparametric tests, choose Legacy Dialogs, and finally click K Independent Samples… to bring up the Tests for Several Independent Samples dialog box.
2. Click the Minutes variable on the left hand side of the dialog box and then click the right arrow ( ) next to the Test Variable List: box.
   a. Note that you can compare as many dependent variables as you want.
3. Next, click to highlight Group and then click the right arrow ( ) next to the Grouping Variable: box.
4. Click the Define Groups… button underneath the Grouping Variable: box to bring up the Several Independent Samples: Define Groups dialog box.
   a. Here, you are asked to indicate the range of groups you are testing (this requires you to remember how you coded the Group variable). You must tell SPSS specifically what the group code range is. Type 1 in the Minimum: box and type 3 into the Maximum: box. Click Continue.
5. Next, you must indicate what the Test Type is. Here, you have three options. Conveniendy, the Kruskal-Wallis H checkbox is the default, so simply make sure this box remains checked.
6. Finally, you can click the Options… button, where you can select two descriptive Statistics or identify how you want SPSS to handle your missing data.
   a. The Descriptive option displays the mean, standard deviation, minimum, maximum, and the number of non-missing values. The Quartiles option displays values corresponding to the 25th, 50th, and 75th percentiles.
   b. Under the Missing Values option, choosing to Exclude cases test-by-test evaluates each test separately for missing values. Choosing Exclude cases listwise excludes participants with missing values for any variable from all analyses.

When you are satisfied with you selections, click Continue.
7. Click OK to conduct the Kruskal-Wallis test.
Output

When you click OK, SPSS will produce an Output screen displaying your results. Click the Output window to view your results (if it does not pop up automatically) which should be identical to those presented in Figure 29.1.

You are presented with two tables containing the results of the Kruskal-Wallis test. The first table is labeled Ranks. Here, you are provided with the sample size (N) and Mean Rank for all groups you identified in Step 4 above.

In the Test Statistics table, you are provided with your test results. The first row contains a Chi-square value because with more than five participants in each group, the Kruskal-Wallis test follows a chi-square distribution. Second, you are presented with your degrees of freedom, which is simply the number of groups minus one. Finally, you are presented with significance level (Asymp. Sig.), which is an asymptotic p-value.

### Kruskal-Wallis Test

<table>
<thead>
<tr>
<th>Ranks</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>N</td>
<td>Mean Rank</td>
</tr>
<tr>
<td>-------</td>
<td>----------</td>
<td>-----------</td>
</tr>
<tr>
<td>Minutes</td>
<td>10</td>
<td>5.50</td>
</tr>
<tr>
<td>Placebo</td>
<td>10</td>
<td>16.50</td>
</tr>
<tr>
<td>OTC</td>
<td>10</td>
<td>25.50</td>
</tr>
<tr>
<td>New</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Statistics a b</th>
<th>Minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square</td>
<td>25.806</td>
</tr>
<tr>
<td>of</td>
<td>2</td>
</tr>
<tr>
<td>Asymp. Sig.</td>
<td>.000</td>
</tr>
</tbody>
</table>

a. Kruskal/Wallis Test  
b. Grouping  
Variable Group  

Figure 29.1

Interpretation

Interpretation begins by evaluating the results provided in the Test Statistics table. Here, the result is statistically significant, indicating that at least two of the groups are statistically different from one another. Unlike the Mann-Whitney test, the Ranks table is not particularly useful to interpretation, even though the omnibus test indicated that the two ranks with the largest difference are statistically significant. That is, your new sleep treatment had a statistically higher rank than Placebo (25.50 and 5.50, respectively).

To determine which groups differ statistically, post hoc Mann-Whitney tests are required using Bonferroni-adjusted alpha. Because you have three groups, three post hoc tests are
required (i.e., *Placebo vs. OTC*, *Placebo vs. New*, and *OTC vs. New*). Thus, the adjusted alpha you will use for all post hoc tests is .017 (i.e., .05/3). See Chapter 26 regarding how to conduct these post hoc Mann-Whitney tests. While I do not show you how to conduct the post hoc Mann-Whitney tests, I have included their results below. You should be able to replicate these results.

**Example Results Section**

The results of the Kruskal-Wallis test indicated a statistically significant difference in ranked minutes of sleep per night between groups, $\chi^2(2) = 25.806, p < .05$. Bonferroni-adjusted post hoc Mann-Whitney tests indicated statistically significant differences between each group with participants receiving the new treatment ranking higher than participants receiving the OTC treatment ($p < .017$) or no treatment ($p < .017$). Further, those receiving the OTC treatment ranked higher compared to participants receiving no treatment ($p < .017$).
A factorial between-subjects ANOVA has one continuous dependent variable and two (or more) categorical independent variables. This Chapter will provide a brief description of analysis for a design with two independent variables; however, the same rationale/analysis can be completed for higher-level designs with more than two independent variables.

The primary advantage of a factorial design is that you can assess the interaction between your independent variables. That is, do the scores within groups of one of your independent variables differ across groups of another independent variable? For a design with two independent variables, you are provided with two main effects (one for each independent variable) and an interaction effect; thus, three separate statistical tests are provided. Note that statistical significance is determined separately for the main effects and the interaction effect.

Continuing with the sleep example, say you are looking to assess the value of your new sleep treatment, but now you want to know whether the treatment has a different effect on men and women. For this experiment, you have two treatment groups (a placebo group and a group receiving your new sleep treatment) and two gender groups (men and women). There are four possible groups for subjects (no treatment/men; no treatment/women; treatment/men; treatment/women). This is an example of a 2 X 2 factorial design because there are two levels within both independent variables. Note that if you added another independent variable with two levels to this design you would have a 2 X 2 X 2 design or if you added a third treatment level to the original design you would have a 3 X 2 design. The collected data is provided below:

<table>
<thead>
<tr>
<th>ID</th>
<th>Tx</th>
<th>Gender</th>
<th>Minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>243</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>239</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>225</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>210</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>185</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>254</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>268</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>215</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>180</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
<td>226</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>1</td>
<td>265</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>1</td>
<td>243</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>1</td>
<td>236</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>1</td>
<td>219</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>1</td>
<td>215</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>0</td>
<td>315</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>0</td>
<td>330</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>0</td>
<td>316</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>0</td>
<td>287</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>0</td>
<td>299</td>
</tr>
</tbody>
</table>

Following the data entry procedures described in Chapters 1 and 2, label the first variable ID, the second variable Tx (0 = Placebo; 1 = Treatment), the third variable Gender (0 = Women; 1 = Men), and the final variable Minutes. Enter the data appropriately.

Assumptions

The assumptions of the factorial between-subjects ANOVA are similar to one-way between-subjects ANOVA—absence of outliers and homogeneity of variance as well as independence, normality, and homoscedasticity of residuals.

Univariate (i.e., one variable) outliers are identified by evaluating z-scores and a histogram of dependent variable raw scores within each group individually (see Chapters 14 and
15, respectively). For the 2 X 2 design above, there are four separate groups. Although z-scores greater than 3.29 or less than -3.29 suggest possible outliers, determine whether the value is actually disconnected from the rest of the distribution using within-group histograms. If outliers are identified, you have several options. You can remove them from analysis through deletion (do not forget to describe them in your results section) or transform the dependent variable (not advised). Note that a nonparametric alternative is not widely available to evaluate an interaction effect.

Homogeneity of variance requires that the dependent variable have roughly the same variance within all groups. If a violation is detected, your statistical test will become too liberal and the probability of type I error increases. This assumption is tested during analysis by Levene’s test, where a non-statistically significant result indicates the variances are not statistically different, thereby satisfying the assumption. If a violation is indicated, however, you should make the omnibus test more conservative by reducing alpha (e.g., from .05 to .01).

Independence of residuals requires the residuals to be uncorrelated across participants. Briefly, a residual is the difference between predicted and actual value of the dependent variable. Because factorial between-subjects ANOVA is a special case of the general linear model (see Chapters 22 and 23 for slightly more detailed discussion), each participant will have their own residual value. Independence of residuals is technically a design issue that requires no repeated measurement or clustering. It can be assured by using appropriate random sampling and random assignment techniques.

The normality assumption requires the residuals to be normally distributed. This assumption is evaluated by a Q-Q plot, a histogram, and Kolmogorov-Smirnov test of residuals (Chapter 15). With that said, simulation studies have shown that the analysis is robust (i.e., still provide correct inference) to violation of normality with greater than 20 error degrees of freedom and relatively equal group sizes. We will request residuals below, however.

Homoscedasticity requires the residuals to have a constant variance across all values the dependent variable. With a categorical independent variable, violation of homoscedasticity goes hand-in-hand with a violation of the homogeneity of variance assumption described above. Thus, if homogeneity of variance is satisfied, homoscedasticity of residuals usually is too. Homoscedasticity is evaluated by requesting a scatterplot matrix where the predicted dependent variable values are plotted against the standardized residuals (see Step 3e below). If this scatterplot appears rectangular, the assumption is considered satisfied.

**Analysis**

Assuming independence, normality of the sampling distribution, and absence of outliers are tenable prior to analysis, to conduct a factorial between-subjects ANOVA:

1. Click **Analyze**, then choose **General Linear Model**, finally click **Univariate**… to bring up the **Univariate** dialog box, shown in Figure 30.1.

2. For this type of design, you will only need to consider two of the five boxes. The first is the **Dependent Variable**: box, where you will place your dependent variable. The second is the **Fixed Factor(s)**: box, where you will place your independent variables.

   a. Click to select the **Minutes** variable and then click the right arrow ( ) next to the **Dependent Variable**: box.
b. Click to select the **Tx** variable and then click the right arrow ( ) next to the **Fixed Factor(s):** box.

c. Click to select the **Gender** variable and then click the right arrow ( ) next to the **Fixed Factor(s):** box.

---

3. You will also notice six buttons on the right side of this dialog box. For factorial between-subjects ANOVA, all six buttons could possibly be applicable.

   a. Clicking the **Model…** button allows you to specify the specific analysis you want to conduct. That is, including or excluding any main effects or interaction terms. The default option is **Full factorial** and leaving this option selected will instruct SPSS to include all main effects and interactions. If you click the **Custom** option, the **Factors & Covariates:** and **Model:** boxes become available. Here, you can specify any main effects and interactions you want. Under the **Build Term(s)** section, clicking the down arrow ( ) on the **Type:** box will bring up all possible effects you could possibly include. So, to include a main effect, select **Main effects** from the **Type:** dropdown menu, select one of your independent variables in the **Factors & Covariates:** box, and then click the right arrow ( ) next to the **Model:** box. You can follow the same steps to include an interaction; you just need to select all variables included in the interaction by holding the Ctrl button on your keyboard as you are selecting variables (this will highlight two or more variables). Finally, you can also specify the type of **Sum of squares:** as well as indicate whether you want SPSS to **Include intercept in model.** When you are satisfied with your selections, click **Continue.**

   b. Clicking the **Contrasts…** button allows you to test for differences among levels of the independent variables you specified in Step 2 above. You can select from several different contrasts described below. Note the default is **None.**

      i. **Deviation:** Compares the mean of each level, other than a reference category, to the grand mean.
ii. **Simple:** Compares the mean of each level to the mean of a specified level. This is useful when you have a control group.

iii. **Difference:** Compares the mean of each level (except the first) to the mean of previous levels.

iv. **Helmert:** Compares the mean of each level (except the last) to the mean of subsequent levels.

v. **Repeated:** Compares the mean of each level (except the last) to the mean of the subsequent level.

vi. **Polynomial:** Compares the linear effect, quadratic effect, cubic effect, and so on depending on how many levels you have. The first degree of freedom contains the linear effect across all categories; the second degree of freedom, the quadratic effect; and so on. These contrasts are used to perform a trend analysis.

If you choose to *Change Contrast*, click the down arrow ( ) in the *Contrast:* box, select the contrast you want from the drop-down menu, and click *Change*. For the *Deviation* and *Simple* contrasts, you will be asked what your *Reference Category* should be, choose either *Last* or *First*. Remember, this is based on your coding scheme, with *Last* indicating your highest coded value and *First* indicating your lowest coded value. When you are satisfied, click *Continue*.

c. Clicking the *Plots*… button allows you to create a line chart plotting the estimated marginal means for all levels within your grouping variable. For factorial between-subjects ANOVA, you will have at least two grouping variables, so specifying specific variables in the *Horizontal Axis:*, *Separate Lines:*, or *Separate Plots:* boxes will significantly change the appearance of the plot. However, the variables you specify in each box are arbitrary and the interpretation will not change.

i. For the example above, click *Tx* and then click the right arrow ( ) next to the *Horizontal Axis:* box. This tells SPSS to place the treatment groups on the x-axis.

ii. Click *Gender* and click the right arrow ( ) next to the *Separate Lines:* box. This tells SPSS you want connect the group means for men and women across treatment group with separate lines.

iii. Click *Add* to move these instructions into the *Plots:* box.

iv. Click *Continue*.

d. Clicking the *Post Hoc*… button allows you to specify the post hoc tests you want to conduct for specific independent variables. Note that this option is only useful if you have three or more levels within an independent variable. If you only have two levels, then the *F*-test for your main effects would provide the exact same information as the post hoc test. That is, post hoc tests are unnecessary and SPSS will print a warning message when you have less than three levels for any independent variable. To specify a post hoc test, select any or all independent variables in the *Factor(s):* box, and then click the right arrow ( ) next to the *Post*
Hoc Tests for: box. This will allow you to select post hoc tests under the Equal Variances Assumed and/or Equal Variances Not Assumed sections. Click the checkbox for the specific post hoc test(s) you want to conduct. When you are satisfied with your selections, click Continue.

e. Clicking the Save… button brings up the Univariate: Save dialog box. Here, you have the option to save new variables to your dataset that contain information about Predicted Values, Diagnostics, Residuals, or Coefficient Statistics. Most of these options allow you to test the assumptions of your model.

i. The options listed under the Predicted Values section allow you to save the values that your specific model predicts for each participant. Remember, any ANOVA analysis is based on the general linear model—a mathematical equation based on deviations from the overall mean and group means. You can save these values as Unstandardized, Weighted (available only if a weighted least squares variable was selected; WLS Weight: in Figure 30.1), or Standard error where an estimate of the standard deviation of the average value of the dependent variable for participants that have the same values of the independent variables.

ii. The options listed under the Diagnostics section provide measures to identify participants with unusual combinations of values for the independent variables or those who have unusual influence on the outcome of the model. For both options, larger values equal larger change or influence.

1. Cook’s distance is a measure of how much the residuals of all participants would change if the participant were removed from analysis.

2. Leverage values indicate the relative influence of each observation on the model’s fit.

iii. The options listed under the Residuals section provide the difference between the observed value of the dependent variable and the model predicted value.

1. Unstandardized: Provides the difference between an observed value and the value predicted by the model.

2. Weighted: Provides the weighted unstandardized residuals and is only available if a weighted least squares variable was previously selected (WLS Weight: in Figure 30.1).

3. Standardized: Provides the residual divided by an estimate of its standard deviation. These are also known as Pearson residuals. Similar to z-scores, they have a mean of 0 and a standard deviation of 1. The distribution of these residuals is used to evaluate normality.

4. Studentized: Provides the residual divided by an estimate of its standard deviation. However, the standard deviation is allowed to
vary across participants depending on the distance of each participant's values on the independent variables from the means of the independent variables.

5. **Deleted**: Provides the residual for a participant if that participant is excluded from analysis. It is the difference between the value of the dependent variable and the adjusted predicted value.

iv. Finally, the options listed under the Coefficient Statistics section writes the variance-covariance matrix of parameter estimates, t-statistics, significance values, and residual degrees of freedom. You have the option to Create a new dataset where you must specify a Dataset name: or you can Write a new data file which creates a complete new dataset. Selecting the latter requires you to click the File… button and specify the dataset you want to write onto.

When you are satisfied with your selections, click **Continue**.

f. Finally, clicking the **Options…** button tells SPSS to print a table of Estimated Marginal Means as well a number of descriptive statistics and diagnostic tests.

i. Under the Estimated Marginal Means section, in the Factor(s) and Factor Interactions: box, you will always be offered the option to display (OVERALL) means. If you choose this option, SPSS will print the grand mean, which considers all the data in your sample together irrespective of the independent variable groups. More importantly, however, this section allows you to display the estimated marginal means for your independent variables and their interaction. For the example, click **Tx**, and then click the right arrow ( ) next to the **Display Means for:** box. Follow the same procedure for **Gender** and **Tx*Gender**.

ii. For any main effects and interactions you selected you can Compare main effects. Clicking this checkbox allows you to conduct post hoc tests (called pairwise comparisons) between all levels of your independent variables. Here, you can select three adjustments which include LSD(none), Bonferroni, or Sidak. Adjustment is required to control for type I error due to multiple pairwise comparisons. Note that with only two levels per independent variable, this option is redundant, as the pairwise comparisons will produce identical results to the F-tests.

iii. Under the Display section, you have ten options. Descriptive statistics prints means, standard deviations, and frequency counts for all levels of your independent variable(s). Estimates of effect size prints partial eta-squared, which is the proportion of total variability attributable to your independent variables and interaction(s). The Observed power option prints the power of the test based on your specific sample. The Parameter estimates option prints parameter estimates, standard errors, t-tests, and confidence intervals (similar to regression). Clicking Contrast coefficient matrix prints the L matrix, but this option is not useful without covariates. Homogeneity tests produces Levene’s test and the Spread-versus-level
plot compares cell means, standard deviations, and variances across the level combinations of all groups. Choosing a Residual plot is useful for checking the homoscedasticity of residuals assumption. Finally, the Lack of fit option is used to check if the relationship between the dependent variable and independent variables can be described adequately by the model and the General estimable function allows you to construct custom hypotheses. You can also specify the Significance level: for your confidence intervals.

When you are satisfied with your selections, click Continue.

4. That’s it! Click OK to conduct the factorial between-subjects ANOVA.

Output

When you click OK, SPSS will produce an Output screen displaying your results. Click your Output window to view your results (if it does not pop up automatically).

The first table you see is titled Between-Subjects Factors containing the sample sizes (N) of your groups within each independent variable. Conveniently, this table provides any Value Labels you provided in the Values column of Variable View to make group identification easy.

The second table is titled Levene’s Test of Equality of Error Variances, shown in Figure 30.2. This table is interpreted in the same way as it was for independent-samples t-test or one-way between-subjects ANOVA, where a significant F indicates a violation of this assumption.

<table>
<thead>
<tr>
<th>Levene’s Test of Equality of Error Variances*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent Variable: Minutes</td>
</tr>
<tr>
<td>F</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>0.947</td>
</tr>
</tbody>
</table>

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

* Design: Intercept + Tx + Gender + Tx * Gender

Figure 30.2

The third table, titled Between-Subjects Effects, is shown in Figure 30.3. This table contains the F-tests, p-values (Sig.), and effect sizes (Partial Eta Squared) for all main effects and interactions. You are also provided with Sums of Squares, degrees of freedom (df), and Mean Squares for the associated effects. Interpretation begins with the interaction effect. If the interaction (Tx * Gender) is statistically significant, it indicates that the levels of one independent variable are dependent on the levels of the other independent variable involved in the interaction. Thus, you cannot interpret your main effects because they only consider each independent variable without respect to the other independent variable(s). However, if your interaction effect is not statistically significant, you can interpret the main effect of each independent variable. You can see in the Figure 30.2 that the interaction effect is statistically significant; thus, the main effects will be ignored. Finally, at the bottom of this Table is the overall effect size (R Squared), which describes the proportion of variance in the dependent variable that is accounted for by all
main effects and interaction(s). Note that Adjusted $R^2$ is simply adjusted for sample size and will always be lower than the $R^2$ value. In addition, large differences between these two indices suggest model over-fit (i.e., little to no generalizability).

### Tests of Between-Subjects Effects

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>Partial $R^2$ Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Model</td>
<td>25304.203*</td>
<td>3</td>
<td>8434.733</td>
<td>13.856</td>
<td>.000</td>
<td>723</td>
</tr>
<tr>
<td>Intercept</td>
<td>1235045.000</td>
<td>1</td>
<td>1235045.000</td>
<td>2033.247</td>
<td>.000</td>
<td>992</td>
</tr>
<tr>
<td>Tx</td>
<td>9405.000</td>
<td>1</td>
<td>9405.000</td>
<td>13.837</td>
<td>.002</td>
<td>464</td>
</tr>
<tr>
<td>Gender</td>
<td>5379.200</td>
<td>1</td>
<td>5379.200</td>
<td>8.856</td>
<td>.009</td>
<td>355</td>
</tr>
<tr>
<td>Tx * Gender</td>
<td>11520.000</td>
<td>1</td>
<td>11520.000</td>
<td>19.955</td>
<td>.000</td>
<td>542</td>
</tr>
<tr>
<td>Error</td>
<td>9718.800</td>
<td>16</td>
<td>607.425</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1270088.000</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>35023.000</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 30.3

The next section is titled *Estimated Marginal Means*. Because the interaction was statistically significant, in Figure 30.4, I have only presented the estimated marginal means for the interaction effect. That is, for each gender individually within each treatment group. In this table, you will find descriptive statistics for all levels within each independent variable. Note that with equal group sizes this table will contain *Mean* values identical to those you could have calculated yourself using the Frequencies or Explore procedures (Chapters 10 and 15, respectively). However, these values will diverge as group sizes become more unequal; thus, you should always report the means presented in the *Estimated Marginal Means* table (because they are unweighted). This is also why I did not have you request *Descriptive statistics* in Step 3f above. Here, you are also provided with standard errors (*Std. Error*) and *Confidence Intervals*.

### 3. Tx * Gender

<table>
<thead>
<tr>
<th>Tx</th>
<th>Gender</th>
<th>Mean</th>
<th>Std. Error</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
<td>Upper Bound</td>
</tr>
<tr>
<td>Control</td>
<td>Women</td>
<td>220.400</td>
<td>11.022</td>
<td>197.034</td>
</tr>
<tr>
<td></td>
<td>Men</td>
<td>235.600</td>
<td>11.022</td>
<td>212.234</td>
</tr>
<tr>
<td>Treatment</td>
<td>Women</td>
<td>309.400</td>
<td>11.022</td>
<td>286.034</td>
</tr>
<tr>
<td></td>
<td>Men</td>
<td>228.600</td>
<td>11.022</td>
<td>205.234</td>
</tr>
</tbody>
</table>

Figure 30.4
Next, you are presented with the residual scatterplot matrix you requested in Step 3f above, shown in Figure 30.5. This matrix presents all bivariate scatterplots between the *Observed* value of the dependent variable, the *Predicted* value of the dependent variable, and the standardized residual value (*Std. Residual*). The scatterplot in the middle row of the last column (i.e., *Predicted* by *Std. Residual*) is the one used to evaluate the homoscedasticity assumptions. This scatterplot may appear different from those you have seen previously. The reason there are two distinct rows of data points within the scatterplot is that predicted values are identical within each unique *Tx by Gender* combination group.

**Figure 30.5**

Finally, you will find the line chart you requested in Step 3c. This chart is shown in Figure 30.6. This chart displays the dependent variable means for the *Placebo* and *Treatment* levels for *Men* and *Women* individually. This chart provides a convenient method for visually inspecting a statistically significant interaction effect. From Figure 30.6, it is clear that the treatment was more effective for *Women* than for *Men*.

**Figure 30.6**
Interpretation

Interpretation begins by evaluating the tenability of the normality of residuals assumption. You will notice that in your dataset, you have saved (or created) a new variable labeled ZRE_1. Further, if you click on Variable View, the Label for this variable states Standardized Residual for Minutes. Use the Explore procedure (Chapter 15) to evaluate the normality of the standardized residuals. Note that residuals are now considered for the entire sample simultaneously, irrespective of group. From the histogram provided in Figure 30.7, you can see the distribution is relatively normal. This evaluation was further supported by the Q-Q plot and non-statistically significant Kolmogorov-Smirnov test ($p > .05$). Thus, the normality assumption is considered satisfied.

![Figure 30.7](image)

Next, you will evaluate the homogeneity of variance assumption provided in the Levene’s Test of Equality of Error Variances table. In this example, Levene’s test is not statistically significant, so the assumption is considered satisfied. In addition, you will evaluate the homoscedasticity of residuals assumption found in the scatterplot matrix. As stated above, when homogeneity of variance is satisfied, homoscedasticity of residuals is usually satisfied as well. As you can see in the Predicted by Std. Residual scatterplot, the width of the data points for each group are relatively equal. Thus, this assumption is considered satisfied as well.

Next, you interpret the results provided in the Tests of Between-Subjects Effects table. First, you are going to evaluate the interaction effect provided in the Tx * Gender row, which for this example is statistically significant with a large effect size (i.e., Partial Eta Squared). Remember, if the interaction is statistically significant, you do not interpret the main effects. Alternatively, if the interaction is not statistically significant, the main effects can be interpreted. Thus, for this example, you will ignore the results for the main effects of Tx and Gender. Please note, however, that I do provide interpretations of the main effects below so you will know how to interpret them in the future if your results indicate a non-statistically significant interaction effect.
Interpreting the Interaction Effect

The interpretation of the interaction effect can be confusing. A statistically significant interaction effect indicates that the independent variables are dependent on one another. Or, stated another way, the values of the dependent variable for one independent variable depend on the levels of the other independent variable. In the example above, the effect of the treatment was dependent on the gender of the participant. That is, women responded to the treatment differently than men. Post hoc tests, called simple main effects, are required to determine where statistically significant differences occurred. These are discussed below.

Interpreting the Main Effects

Although the interaction effect was statistically significant in the example above, for didactic purposes I will provide you with an interpretation of the main effects as well. Interpretation of main effects is very similar to one-way between-subjects ANOVA. A main effect for one independent variable in factorial ANOVA collapses across the levels of the other independent variable. That is, it essentially conducts the analysis considering only one of the independent variables. For example, the main effect of treatment group collapses across gender. Thus, the analysis is compares the difference between treatment and placebo, irrespective of gender. Think, independent-samples t-test on all data using only the treatment variable. Alternatively, the main effect for gender collapses across treatment groups. Thus, the difference between men and women is evaluated irrespective of treatment group.

In the example above both main effects were statistically significant. Because both independent variables had only two levels, interpretation only requires that you examine the estimated marginal means to determine which group slept more (or less) minutes per night. From the example, the main effect of treatment group indicated participants receiving your new sleep treatment slept significantly longer than participants in the placebo group (269.000 minutes vs. 228.000 minutes, respectively). The main effect of gender indicated women slept significantly longer than men (264.900 minutes vs. 232.100 minutes, respectively).

If, however, you have an independent variable with three or more levels, the main effect is an omnibus test. That is, a statistically significant difference exists between at least two of the levels, but you do not know exactly which levels differ. Post hoc tests, called simple comparisons, are required to determine which levels were statistically different. Again, these tests only consider the independent variable of interest by collapsing across the other independent variable. The procedure for conducting these comparisons in SPSS is identical to the simple main effects discussed below. They cannot be conducted appropriately through the point-and-click method. That is, all you need to do is include the COMPARE(*) and ADJ(*) statement on the appropriate /EMMEANS line, where the asterisk represents the independent variable of interest and adjustment for type I error, respectively. For the example above, using a Bonferroni adjustment, you would include the following code: COMPARE(Gender) ADJ(BONFERRONI) to compare across gender or COMPARE(Tx) ADJ(BONFERRONI) to compare across treatment groups. This, of course, should make much more sense one you have read the simple main effects analysis section below.
Simple Main Effects Analysis

At this point you should be asking, how we know whether men and women differ in the treatment group, placebo group, or both? You can see in Figure 30.6 that there is a substantial difference between men and women in the treatment group. Because this is the biggest difference, and there is no violation of homogeneity of variance, this can immediately be considered statistically significant. However, for a factorial between-subjects ANOVA, you must conduct simple main effects to interpret (or tease apart) a statistically significant interaction. Simple main effects assess for differences between the levels of one independent variable while ignoring the other independent variable. For example, differences between men and women are assessed only in the treatment group and then only in the placebo group. Two analyses total.

Simple main effects cannot be performed in SPSS using point-and-click methods because the error terms and degrees of freedom will not be used, leading to incorrect inference. However, they can be performed using the Syntax Editor. Up to this point, we have not discussed syntax, but the procedure is straightforward. To perform simple main effects using the Syntax Editor:

1. Follow the point-and-click method described above in Steps 1 through 3. However, instead of clicking Continue in Step 4, click Paste. A Syntax Editor window will pop up containing the syntax required to conduct the factorial between-subjects ANOVA, shown in Figure 30.8.

2. Without describing the entire syntax, the primary line in the syntax you will be concerned with is the 8th line of code shown in Figure 30.8. This is presented below:

   /EMMEANS=TABLES(Tx*Gender)
This line of code was created when you requested estimated marginal means for the interaction effect in Step 3f above. All you need to do to tell SPSS you want to compare for differences in gender across treatment groups, using a Bonferroni adjustment to control type I error, is type `COMPARE(Gender) ADJ(BONFERRONI)` at the end of the line of code presented above. Note that all three adjustments for type I error described in Step 3f above are available for these comparisons (i.e., LSD and Sidak). When you do this, line 8 should now look like this:

```
/EMMEANS=TABLES(Tx*Gender) COMPARE(Gender) ADJ(BONFERRONI)
```

3. That’s it! To conduct the analysis, in the `Syntax Editor` window, click **Run** and then click **All**.

**Interpretation of Simple Main Effects**

When you click **All**, SPSS will produce an `Output` screen displaying your results. Click your `Output` window to view your results (if it does not pop up automatically). You should notice that for the most part the output is identical to the point-and-click output described previously. The only difference is that under the `Estimated Marginal Means` section you now have new tabled in the 3. `Tx * Gender` section, titled `Pairwise Comparisons` and `Univariate Tests`. These tables are shown in Figure 30.9.

<table>
<thead>
<tr>
<th>Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent Variable: Minutes</td>
</tr>
<tr>
<td>( \text{Tx} )</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>Control</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Treatment</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pairwise Comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent Variable: Minutes</td>
</tr>
<tr>
<td>( \text{Tx} )</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>Control</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Treatment</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

* The mean difference is significant at the .05 level.

Figure 30.9
The term “pairwise comparisons” should immediately indicate that these are post hoc tests. Further, you can see in footnote \textit{a} under the \textit{Pairwise Comparisons} table in Figure 30.9 that \textit{Adjustment for multiple comparisons: Bonferroni} indicates the Bonferroni adjustment was applied as requested by the syntax. You should also notice that while it appears as though there are four statistical tests provided in the \textit{Pairwise Comparisons} table, two of them are redundant. That is, the \textit{Placebo} group comparison of \textit{Women} to \textit{Men} in row 1 is the same as the comparison of \textit{Men} to \textit{Women} in row 2. This is why the Bonferroni adjustment was based on 2 post hoc tests—one test for the Placebo group; one test for the Treatment group.

Finally, if you compare the \textit{Pairwise Comparisons} and \textit{Univariate Tests} tables, you can see that they provide identical statistical results, just in different forms. This is specifically because our Gender independent variable has only two levels. If we had three or more levels of an independent variable, the \textit{F}-test in the \textit{Univariate Tests} table would be an omnibus test for the simple main effect, and the \textit{Pairwise Comparisons} table would then provide the post hoc tests. Because these two tables provide identical information, we only consider the \textit{Pairwise Comparisons} table for the statistical test and use the \textit{Univariate Tests} table for effect size.

Using the \textit{Pairwise Comparisons} table, evaluation begins by examining the \textit{Sig.} column, which contains your \textit{p}-values. Alternatively, you can quickly identify a statistically significant difference by looking for an asterisk next to the value in the \textit{Mean Difference (I-J)} column. If a difference is statistically significant, you then look at the \textit{Mean Difference (I-J)} column to evaluate which group had higher values as well as the standard error (\textit{Std. Error}) and confidence interval (\textit{95% Confidence Interval for Difference}). Next, you evaluate the effect size (\textit{Partial Eta Squared}) in the \textit{Univariate Tests} table.

\textbf{Example Results Section}

Prior to analysis, no violation of independence was indicated and no outliers were identified. No violation of normality or homoscedasticity of residuals was indicated, and homogeneity of variance was assured by Levene’s test, $F(3,16) = .947, p > .05$.

The results of a 2 (Group: Treatment vs. Placebo) X 2 (Gender: Men vs. Women) factorial between-subjects ANOVA indicated a statistically significant interaction effect, $F(1, 16) = 18.965, p < .05$, partial $\eta^2 = .542$. A simple main effects analysis, using the Bonferroni adjustment to reduce the probability of type I errors, indicated a statistically significant difference between men and women in the treatment group ($p < .05$; partial $\eta^2 = .627$), with women sleeping more minutes per night than men (mean difference = 80.800; 95% CI = 47.756 to 113.844). No statistically significant gender differences were indicated in the Placebo group.
Chapter 31
ONE-WAY BETWEEN-SUBJECTS ANCOVA

One-way between-subjects analysis of covariance (aka, ANCOVA) is a combination of one-way between-subjects ANOVA and linear regression. That is, you have one categorical independent variable and one (or more) continuous covariate that is related to the continuous dependent variable, but not of research interest. ANCOVA statistically adjusts group means using the association (i.e., slope) between the dependent variable and covariate. Thus, effective adjustment requires the covariate to be highly related to the dependent variable.

From the definition and description of ANCOVA above, you probably figured out that any design applicable to ANCOVA is also applicable to multiple linear regression (see Chapter 23). In fact, ANOVA, ANCOVA, and linear regression are mathematically equivalent by the general linear model (think, regression equation). Thus, a thorough understanding of both linear regression and ANOVA is required prior to the use and proper interpretation of ANCOVA, so feel free to review Chapters 22, 23, and 28.

Finally, you should be aware that ANCOVA can be employed for both factorial and repeated-measures designs, and while these analyses are extensions of the discussion here, they are beyond the scope of this Chapter.

As an example, say you want to know whether participants who take your new sleep treatment sleep more minutes per night compared to participants who take the leading over the counter treatment after statistically adjusting for milligrams of caffeine ingested after 2 pm. The collected data is presented below:

<table>
<thead>
<tr>
<th>ID</th>
<th>Tx</th>
<th>Caffeine</th>
<th>Minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>267</td>
<td>270</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>159</td>
<td>299</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>125</td>
<td>326</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>75</td>
<td>359</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>115</td>
<td>297</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>127</td>
<td>315</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>90</td>
<td>402</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>85</td>
<td>378</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>61</td>
<td>388</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>430</td>
</tr>
</tbody>
</table>

Following the data entry procedures described in Chapters 1 and 2, label the first variable ID, the second Tx (0 = OTC; 1 = Treatment), the third Caffeine, and the fourth variable Minutes. Enter the data appropriately.

Assumptions

The assumptions for one-way between-subjects ANCOVA are a combination of one-way between-subjects ANOVA and linear regression. They include independence, normality, and homoscedasticity of residuals as well as absence of univariate and bivariate outliers, linearity between the dependent variable and covariate, and homogeneity of variance. Additional assumptions include homogeneity of regression, assured reliability of covariates, and (possibly) absence of multivariate and multicollinearity.
Independence of residuals requires the residuals to be uncorrelated across participants. Briefly, a residual is the difference between predicted and actual value of the dependent variable. Because one-way between-subjects ANCOVA is essentially multiple linear regression (Chapter 23), and thus, uses the general linear model, each participant will have their own residual value. Independence of residuals is technically a design issue that requires no repeated measurement or clustering. It can be assured by using appropriate random sampling and random assignment techniques.

The normality assumption requires the residuals to be normally distributed. This assumption is evaluated by a Q-Q plot, a histogram, and Kolmogorov-Smirnov test of residuals (Chapter 15). With that said, simulation studies have shown that the analysis is robust (i.e., still provide correct inference) to violation of normality with greater than 20 error degrees of freedom and relatively equal group sizes. We will request residuals below, however.

Homoscedasticity requires the residuals to have a constant variance across all values the dependent variable. With a categorical independent variable, violation of homoscedasticity goes hand-in-hand with a violation of the homogeneity of variance assumption described below. Thus, if homogeneity of variance is satisfied, homoscedasticity of residuals usually is too. Homoscedasticity is evaluated by requesting a scatterplot matrix where the predicted dependent variable values are plotted against the standardized residuals (see Step 3e below). If this scatterplot appears rectangular, the assumption is considered satisfied.

Univariate (i.e., one variable) outliers are identified by evaluating $z$-scores and a histogram of dependent variable raw scores within each group individually (see Chapters 14 and 15, respectively). Although $z$-scores greater than 3.29 or less than -3.29 may suggest a possible outlier, you need to determine whether the value is actually disconnected from the rest of the distribution using within-group histograms. Bivariate (i.e., two variable) outliers can severely influence the regression line and can lead to incorrect inference. Bivariate outliers are identified by data points that are disconnected from the rest of the data points in the scatterplot. Note that you will assess for outliers using a scatterplot of the covariate and dependent variable. Similar to linear regression, no assumption is made about the distribution or outliers within the covariate specifically; outliers are always defined in relation to the dependent variable. If outliers are identified you have several options. You can remove them from analysis through deletion (do not forget to describe them in your results section) or transform the dependent variable (not advised). Note that there is no readily available nonparametric alternative that uses ranked data and covariates.

Linearity requires the relationship between the dependent variable and covariate to be best approximated by a straight (or linear) regression line. This assumption is evaluated prior to analysis using a scatterplot (Chapter 19).

Homogeneity of variance requires that the dependent variable have roughly the same variance within all groups. If a violation is detected, your statistical test will become too liberal and the probability of type I error increases. This assumption is tested during analysis by Levene’s test, where a non-statistically significant result indicates the variances are not statistically different, thereby satisfying the assumption. If a violation is indicated, however, you should make the omnibus test more conservative by reducing alpha (e.g., from .05 to .01).

The two new assumptions specific to ANCOVA include reliability of covariates and homogeneity of regression. Reliability of covariates requires that the covariates be measured without error. This can be a tough assumption to satisfy with behavioral or self-report measures. Homogeneity of regression states that the slopes of the regression coefficient between the
dependent variable and covariate are the same within all groups. Or, stated another way, there is no interaction between the categorical independent variable and the covariate. This assumption is tested prior to analysis and the procedure described below. A violation of this assumption renders ANCOVA inappropriate because ANCOVA cannot include independent variable-covariate interactions. However, if this assumption is violated, all you would need to do is use multiple linear regression instead (Chapter 23). Just make sure to include the interaction in the regression analysis.

Finally, if you have more than one covariate, the assumptions of absence of multivariate outliers and multicollinearity must also be evaluated. Multivariate outliers are identified using Mahalanobis distance and multicollinearity is tested using tolerance and VIF values. Both assumptions are evaluated using the linear regression procedure in SPSS, and are described in detail in Chapter 23.

**Evaluating the Homogeneity of Regression Assumption**

1. Click **Analyze**, then choose **General Linear Model**, and then click **Univariate**… to bring up the **Univariate** dialog box, shown in Figure 31.1.

![Figure 31.1](image)

2. You will notice immediately that all of your variables are listed on the left hand side of this dialog box. For ANCOVA, you will be using three boxes—**Dependent Variable:**, **Fixed Factor(s):**, and **Covariate(s):**. Remember, fixed factor is another term for categorical independent variable.

   a. Click to select the **Minutes** variable and then click the right arrow ( ) next to the **Dependent Variable:** box.

   b. Click to select the **Tx** variable and then click the right arrow ( ) next to the **Fixed Factor(s):** box.
c. Click to select the Caffeine variable and then click the right arrow (►) next to the Covariate(s): box.

3. Next, you need to tell SPSS that you want to include the interaction between the independent variable and covariate in the analysis. This is the explicit test of homogeneity of regression. Click the Model... button to bring up the Univariate: Model dialog box shown in Figure 31.2.
   a. Under the Specify Model section, click the Custom radio button.
   b. Under the Factors & Covariates section, click to highlight the Tx variable, and then click the right arrow (►) next to the Model: box.
   c. Click the Caffeine variable, and then click the right arrow (►) next to the Model: box.
   d. Next, under the Build Term(s) section, under Type:, make sure that Interaction is chosen. If it is not, click the down arrow (▼) to select it from the drop-down menu. Click to highlight the Tx variable, then press and hold the Ctrl button on your keyboard, and then click the Caffeine variable. With both variables selected, click the (►) next to the Model: box. Your dialog box should be identical to Figure 31.2.
   e. Click Continue.

4. That’s it! Click OK.

In your output window, scroll down to the Test of Between-Subjects Effects box. It should be the last table in your output. Look only at the statistical significance (Sig.) of the Tx * Caffeine variable. Here, the result of the F-test is not statistically significant, F(1,26) = 1.901, p > .05, which indicates that there is no interaction between treatment group and caffeine intake. Thus, the homogeneity of variance assumption is satisfied.
Analysis

Assuming independence of residuals, absence of outliers, linearity, homogeneity of regression, and reliability of the covariate, to conduct a one-way between-subjects ANCOVA on the example data above:

1. Click Analyze, then choose General Linear Model, and then click Univariate… to bring up the Univariate dialog box identical to Figure 31.1. Click Reset.

2. Again, you will notice immediately that all of your variables are listed on the left hand side of this dialog box.
   a. Click to select the Minutes variable and then click the right arrow ( ) next to the Dependent Variable: box.
   b. Click to select the Tx variable and then click the right arrow ( ) next to the Fixed Factor(s): box.
   c. Click to select the Caffeine variable and then click the right arrow ( ) next to the Covariate(s): box.

3. You will notice six buttons on the right side of this dialog box, and for one-way between-subjects ANCOVA, all these buttons could possibly be applicable.
   a. Clicking the Model… button allows you to specify the specific analysis you want to conduct. That is, including or excluding any main effects or interaction terms. The default option is Full factorial and leaving this option selected will instruct SPSS to include only main effects for your independent variable and covariate. Note that interactions will only be included if you have more than one independent variable. That is, no independent variable-covariate interactions will ever be included, because this type of interaction renders ANCOVA inappropriate! Finally, you can also specify the type of Sum of squares: as well as indicate whether you want SPSS to Include intercept in model. When you are satisfied with your selections, click Continue.
   b. Clicking the Contrasts… button allows you to test for differences among levels of the independent variable(s) you specified in Step 2 above. You can select from several different contrasts described below. Note the default is None.
      i. Deviation: Compares the mean of each level, other than a reference category, to the grand mean.
      ii. Simple: Compares the mean of each level to the mean of a specified level. This is useful when you have a control group.
      iii. Difference: Compares the mean of each level (except the first) to the mean of previous levels.
      iv. Helmert: Compares the mean of each level (except the last) to the mean of subsequent levels.
      v. Repeated: Compares the mean of each level (except the last) to the mean of the subsequent level.
vi. **Polynomial**: Compares the linear effect, quadratic effect, cubic effect, and so on depending on how many levels you have. The first degree of freedom contains the linear effect across all categories; the second degree of freedom, the quadratic effect; and so on. These contrasts are used to perform a trend analysis.

If you choose to *Change Contrast*, click the down arrow ( ) in the *Contrast:* box, select the contrast you want from the drop-down menu, and then click *Change*. For the *Deviation* and *Simple* contrasts, you will be asked what your *Reference Category* should be, choose either *Last* or *First*. Remember, this is based on your coding scheme, with *Last* indicating your highest coded value and *First* indicating your lowest coded value. When you are satisfied, click *Continue*.

c. Clicking the **Plots** button allows you to create a line chart plotting the *adjusted* estimated marginal means for all levels within your grouping variable. For one-way between-subjects ANCOVA, you will have only have one grouping variable, so you need to specify this variables in the *Horizontal Axis:* box.

i. Click *Tx* and then click the right arrow ( ) next to the *Horizontal Axis* box. This tells SPSS to place the treatment groups on the x-axis.

ii. Click *Add* to move these instructions into the *Plots:* box.

When you are satisfied with your selections, click *Continue*.

d. For this example, the **Post Hoc** button is not available because your independent variable has only two categories. That is, a statistically significant result can be interpreted simply by evaluating the adjusted estimated marginal means. However, if you have three or more categories, this option allows you to specify the post hoc tests you want to conduct for specific independent variables. To specify a post hoc test, select any or all independent variables in the *Factor(s)* box, and then click the right arrow ( ) next to the *Post Hoc Tests for:* box. This will allow you to select post hoc tests under the *Equal Variances Assumed* and/or *Equal Variances Not Assumed* sections. Click the checkbox for the specific post hoc test(s) you want to conduct. When you are satisfied with your selections, click *Continue*.

e. Clicking the **Save** button brings up the *Univariate: Save* dialog box. Here, you have the option to save new variables to your dataset that contain information about *Predicted Values, Diagnostics, Residuals*, or *Coefficient Statistics*. Most of these options allow you to test the assumptions of your model.

i. The options listed under the *Predicted Values* section allow you to save the values that your specific model predicts for each participant. Remember, ANCOVA is based on the general linear model—a mathematical equation based on deviations from the overall mean and group means. You can save these values as *Unstandardized, Weighted* (available only if a weighted least squares variable was selected; *WLS Weight:* in Figure 31.1), or *Standard error* where an estimate of the standard deviation of the average value of the dependent variable for participants that have the same values of the independent variables.
ii. The options listed under the Diagnostics section provide measures to identify participants with unusual combinations of values for the independent variables or those who have unusual influence on the outcome of the model. For both options, larger values equal larger change or influence.

1. *Cook’s distance* is a measure of how much the residuals of all participants would change if the participant were removed from analysis.

2. *Leverage values* indicate the relative influence of each observation on the model’s fit.

iii. The options listed under the Residuals section provide the difference between the observed value of the dependent variable and the model predicted value. Note that the Standardized: option will be used to evaluate the normality of residuals assumption.

1. *Unstandardized:* Provides the difference between an observed value and the value predicted by the model.

2. *Weighted:* Provides the weighted unstandardized residuals and is only available if a weighted least squares variable was previously selected (*WLS Weight:* in Figure 30.1).

3. *Standardized:* Provides the residual divided by an estimate of its standard deviation. These are also known as Pearson residuals. Similar to z-scores, they have a mean of 0 and a standard deviation of 1.

4. *Studentized:* Provides the residual divided by an estimate of its standard deviation. However, the standard deviation is allowed to vary across participants depending on the distance of each participant's values on the independent variables from the means of the independent variables.

5. *Deleted:* Provides the residual for a participant if that participant is excluded from analysis. It is the difference between the value of the dependent variable and the adjusted predicted value.

iv. Finally, the options listed under the Coefficient Statistics section writes the variance-covariance matrix of parameter estimates, $t$-statistics, significance values, and residual degrees of freedom. You have the option to Create a new dataset where you must specify a Dataset name: or you can Write a new data file which creates a complete new dataset. Selecting the latter requires you to click the File... button and specify the dataset you want to write onto.

When you are satisfied with your selections, click **Continue**.

f. Finally, clicking the Options... button tells SPSS to print a table of adjusted Estimated Marginal Means as well a number of descriptive statistics and diagnostic tests.
i. Under the *Estimated Marginal Means* section, in the *Factor(s) and Factor Interactions:* box, you will always be offered the option to display the *(OVERALL)* mean. If you choose this option, SPSS will print the grand mean, which considers all the data in your sample together irrespective of the independent variable groups. More importantly, this section allows you to display the adjusted marginal means for your independent variable(s), and, if need be, their interaction. For the example, click *Tx,* and then click the right arrow ( ) next to the *Display Means for:* box.

ii. For any main effects you included in the previous step, you can *Compare main effects.* Clicking this checkbox allows you to conduct post hoc tests (called pairwise comparisons) between all levels of your independent variables. Here, you can select three adjustments which include *LSD*(none), *Bonferroni,* or *Sidak.* Adjustment is required to control for type I error due to multiple pairwise comparisons. Note that your independent variable has only two levels, so this option is redundant.

iii. Under the *Display* section, you have ten options. The *Descriptive statistics* option prints means, standard deviations, and frequency counts for all levels of your independent variable(s). *Estimates of effect size* prints partial eta-squared, which is the proportion of total variability attributable to your independent variables and interaction(s). The *Observed power* option prints the power of the test based on your specific sample. The *Parameter estimates* option prints parameter estimates, standard errors, t-tests, and confidence intervals. This is used to further evaluate the covariate. Clicking *Contrast coefficient matrix* prints the L matrix. *Homogeneity tests* produces Levene’s test and the *Spread-versus-level plot* compares cell means, standard deviations, and variances across the level combinations of all groups. Choosing a *Residual plot* provides a scatterplot matrix of standardized and unstandardized predicted values and residuals. Finally, the *Lack of fit* option is used to check if the relationship between the dependent variable and independent variables can be described adequately by the model and the *General estimable function* allows you to construct custom hypotheses. You can also specify the *Significance level:* for your confidence intervals.

When you are satisfied with your selections, click *Continue.*

4. That’s it! All you need to do now is click *OK.*

**Output**

When you click *OK,* SPSS will produce an *Output* screen displaying your results. Click your *Output* window to view your results (if it does not pop up automatically). The first table you see is titled *Between-Subjects Factors* containing the sample sizes (*N*) of your groups within each independent variable. Conveniently, this table provides any *Value Labels* you provided in the *Values* column of *Variable View* to make group identification easy. The second table, titled *Descriptive Statistics,* contains the (weighted) *Mean,* standard deviation (*Std. Deviation*), and sample sizes (*N*) for your groups within each independent variable.
The next table is titled *Levene’s Test of Equality of Error Variances*. This table is interpreted in the same way as it was for independent-samples t-test or one-way between-subjects ANOVA, where a significant $F$ indicates a violation of the homogeneity of variance assumption.

The fourth table, titled *Between-Subjects Effects*, is shown in Figure 31.3. This table contains the $F$-tests, $p$-values (Sig.), and effect sizes (*Partial Eta Squared*) for both your independent variable and covariate. You are also provided with *Sums of Squares*, degrees of freedom ($df$), and *Mean Squares* for the associated effects. Finally, at the bottom of this table is the overall effect size (*$R$ Squared*), which describes the proportion of variance in the dependent variable that is accounted for by the independent variable and covariate. Remember, *Adjusted $R$ Squared* is adjusted for sample size and will always be smaller than the *$R$ Squared* value.

![Figure 31.3](image)

The next table, titled *Parameter Estimates*, is shown in Figure 31.4. Here, you can evaluate the slope ($B$) of your regression coefficient that was used to adjust the group means. You will also notice that SPSS has dummy coded your categorical independent variable. Notice that the OTC group (coded 0 in your dataset) is not the reference group. I know this is confusing, but critical to interpretation.

![Figure 31.4](image)
The next section is titled *Estimated Marginal Means*, and is presented in Figure 31.5. Here, you will find the adjusted means (*Mean*), standard error (*Std. Error*), and *95% Confidence Intervals* for each level of your independent variable. I want to reiterate that this table contains adjusted means, and the footnote at the bottom of the table indicates the value of the covariate where these means were calculated. Note that these adjusted means will be different from the weighted means presented in the *Descriptive Statistics* table, even if sample sizes are equal.

<table>
<thead>
<tr>
<th>Tx</th>
<th>Mean</th>
<th>Std. Error</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower Bound</td>
<td>Upper Bound</td>
<td></td>
</tr>
<tr>
<td>CTC</td>
<td>356.813 a</td>
<td>10.593</td>
<td>335.078</td>
</tr>
<tr>
<td>Treatment</td>
<td>461.654 a</td>
<td>10.593</td>
<td>439.919</td>
</tr>
</tbody>
</table>

*a. Covariates appearing in the model are evaluated at the following values: Caffeine = 94.67.*

**Figure 31.5**

Finally, you will find the line chart you requested in Step 3c above shown in Figure 31.6. This chart displays the adjusted dependent variable means for the *OTC* and *Treatment* groups individually. This chart provides the exact same information as the adjusted estimated marginal means in Figure 31.5.

**Figure 31.6**
Interpretation

Interpretation begins by evaluating the tenability of the normality of residuals assumption. You will notice that in your dataset, you have saved (or created) a new variable labeled \( ZRE_1 \). Further, if you click on Variable View, the Label for this variable states Standardized Residual for Minutes. Use the Explore procedure (Chapter 15) to evaluate the normality of the standardized residuals. Note that residuals are now considered for the entire sample simultaneously, irrespective of group. From the histogram provided in Figure 31.7, you can see the distribution is relatively normal. This evaluation was further supported by the Q-Q plot and non-statistically significant Kolmogorov-Smirnov test (\( p > .05 \)). Thus, the normality assumption is considered satisfied.

![Figure 31.7](image)

Next, you will evaluate the homogeneity of variance assumption, tested by Levene’s test provided in the Levene’s Test of Equality of Error Variances table. The test is not statistically significant (i.e., \( p > .05 \)); thus, homogeneity of variance is satisfied and you can move on to the Tests of Between-Subjects Effects table. Here, you will evaluate the significance test of your independent variable, which for this example is statistically significant with a substantial effect size (Partial Eta Squared) of .620. Although you are primarily concerned with the \( Tx \) variable, notice that the covariate, Caffeine, is also statistically significant. This was expected, as your covariate should always be significantly related to the dependent variable. If it is not, you have chosen a poor covariate and have sacrificed statistical power. Interpretation of your independent variable is completed by examining the adjusted marginal means. You can see that after adjusting for caffeine intake, the treatment group slept significantly longer than the OTC group (461.654 vs. 356.813 minutes, respectively). Also, notice the Adjusted R Squared value of .743 in the footnote underneath this table. This indicates the total proportion of variance in minutes of sleep per night explained by both treatment group and caffeine intake after 2pm.

Next, you will evaluate the contribution of the covariate in the Parameter Estimates table. Note that much of this information is redundant with the results provided in the Tests of Between-Subjects Effects table. The slope for this continuous covariate is interpreted identically to the results of a linear regression analysis (see Chapter 22). That is, a one-milligram increase in caffeine after 2pm resulted in a .349 minute decrease in minutes of sleep per night. Again, you
do not need to report the significance test for this parameter estimate because it is identical to the test reported in the *Tests of Between-Subjects Effects* table. In addition, pay attention to the *Tx* dummy variables. Here, the slope of -104.840 indicates the mean difference between the OTC and Treatment groups. Because the Treatment group was the reference group (indicated by the slope value of 0), and the slope parameter is negative, you know the OTC group slept 104.840 minutes less than the Treatment group. Finally, the *Intercept* value of 494.731 is meaningful and interpreted as the average minutes of sleep per night for a participant in the Treatment group who consumed no caffeine after 2pm.

**Example Results Section**

No violation of independence or normality of residuals, homogeneity of variance or regression, or linearity was observed, and no outliers were identified.

The results of a one-way between-subjects ANCOVA indicated a statistically significant difference in minutes of sleep per night across treatment groups after adjusting for caffeine intake after 2pm, $F(1,27) = 44.058$, $p < .05$, partial $\eta^2 = .620$. Adjusted means indicated the treatment group slept significantly longer than the group receiving the over-the-counter treatment (461.651 minutes vs. 356.813 minutes, respectively). Finally, a one-milligram increase in caffeine intake after 2pm resulted in a .349 minute decrease in minutes of sleep per night.
Section VI
Statistical Analysis of Within-Subjects Change

Again, the parametric analyses described in this section are special cases of the general linear model. That is, most of the assumptions discussed here are based on the residual values from the linear regression equation. Except now, however, the residual term has been partitioned into between-subjects variance and within-subject variance. In repeated measures analysis, all between-subject variance is removed from analysis, which in turn reduces the error term thereby increasing statistical power. Similar to above, nonparametric alternatives are also presented when applicable.

Similar to Parts IV and V, the Chapters in this section each include a small dataset for you to practice your data entry and coding skills. I chose to include a dataset for you to use rather than refer to abstract examples so that you will be able to perform the analyses and replicate all output presented. This will allow you to verify you completed the analysis correctly. Similar to above, while all portions of the menus for each analysis are described in detail, the **bolded** instructions are minimally required to replicate the output provided.

Finally, at the end of each Chapter you will be presented with an example results section in APA format. This should give you a pretty good idea of what will be minimally required when reporting results for your future posters or manuscripts.
The paired-samples t-test (also known as the matched t-test, nested t-test, or dependent t-test) is used when one group of participants is measured twice on a continuous dependent variable, or two groups of participants are matched on specific characteristics and measured once on a continuous dependent variable. In both cases, the assumption of independence, or mutually exclusive groups, is violated. When one group of participants is measured twice, it is known as a repeated measures design. Alternatively, when two groups of participants are matched on specific characteristics, it is called a matched design. Repeatedly measuring participants or matching participants is a valid method for reducing error and increasing statistical power.

The more common of the two designs discussed above is the repeated measures design, which will be used in the example for this Chapter. Note, however, that the analysis of a matched design follows the exact same steps.

As an example of a repeated measures design, say you want to test the effectiveness of your new sleep treatment using each participant as his or her own control. Before beginning the treatment, you select your sample and measure the minutes of sleep each night for each participant. Then, you implement your treatment. One month after implementing your treatment, you measure the amount of sleep each night for those same participants. You now have pre-treatment and post-treatment data on the same dependent variable for the same subjects. The collected data is below.

<table>
<thead>
<tr>
<th>ID</th>
<th>Pre</th>
<th>Post</th>
<th>ID</th>
<th>Pre</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>250</td>
<td>396</td>
<td>6</td>
<td>267</td>
<td>428</td>
</tr>
<tr>
<td>2</td>
<td>302</td>
<td>417</td>
<td>7</td>
<td>330</td>
<td>500</td>
</tr>
<tr>
<td>3</td>
<td>210</td>
<td>285</td>
<td>8</td>
<td>235</td>
<td>285</td>
</tr>
<tr>
<td>4</td>
<td>285</td>
<td>402</td>
<td>9</td>
<td>265</td>
<td>480</td>
</tr>
<tr>
<td>5</td>
<td>280</td>
<td>460</td>
<td>10</td>
<td>246</td>
<td>379</td>
</tr>
</tbody>
</table>

Following the data entry procedures described in Chapters 1 and 2, label your first variable ID, your second variable Pre, and your final variable Post. Enter the data appropriately.

Assumptions

The assumptions of the paired-samples t-test include absence of outliers as well as normality and homoscedasticity of residuals.

Because the paired-samples t-test is based directly on one difference score, the absence of outliers assumption is applicable directly to the difference score. That is, from the example, the difference between Pre and Post. You can evaluate this assumption by first calculating the difference between your two measures using the Compute procedure (Chapter 13) and then evaluating the z-scores and histogram using the Descriptives and Frequencies procedures (Chapters 14 and 10, respectively).

The normality assumption requires the residuals to be normally distributed. However, for repeated measures designs, residual values are not interpreted as they have been previously. While a full discussion is beyond the scope of the Chapter, briefly, repeated-measures designs increase statistical power by removing between-person differences. The residuals (i.e., e)
produced by a repeated-measures analysis are what remain after between-person differences have been removed. That is, the residuals represent only within-person variance. Because residuals are based on the repeated measurements, a residual value is produced for each repeated measurement. Thus, for the example above, two residual values would be produced, and the normality assumption on each residual distribution separately. Similar to one-way between-subjects ANOVA, this assumption is evaluated by a Q-Q plot, a histogram, and Kolmogorov-Smirnov test of residuals (Chapter 15). With that said, simulation studies have shown that the analysis is robust (i.e., still provide correct inference) to violation of normality with greater than 20 error degrees of freedom.

Homoscedasticity requires the residuals to have a constant variance across all values the predicted repeated measurement. Homoscedasticity is evaluated by requesting a scatterplot matrix where the predicted repeated measurement values are plotted against the standardized residuals. If these scatterplots appear rectangular, the assumption is considered satisfied.

Similar to the situation described in Chapter 25, you cannot request residuals using the dedicated paired-samples t-test procedure in SPSS. However, they can be requested using the general linear model procedure shown in Chapter 34. Because a paired-samples t-test and one-way repeated-measures ANOVA are mathematically equivalent through the general linear model, any software that conducts one-way repeated-measures ANOVA can also conduct a paired-samples t-test. Remember, \( t^2 = F \) with one degree of freedom between groups (i.e., two groups). Thus, if you take the example data above and use the procedure described in Chapter 34, inference will be identical.

**Analysis**

Assuming normality and absence of outliers within the difference scores, to conduct a paired-samples t-test on the data above:

1. Click **Analyze**, then choose **Compare Means**, and finally choose **Paired-Samples T Test**… to bring up the **Paired-Samples T Test** dialog box shown in Figure 32.1.

![Figure 32.1](image)

2. For this analysis, you are required to designate the variables in pairs. These variable pairs will be indicated in the **Paired Variables**: box.
   a. Click the **Pre** variable and then click the right arrow (\( \rightarrow \)) next to the **Paired Variables**: box. This variable should now be listed under the **Variable1** column.
b. Click the **Post** variable and then click the right arrow ( ) next to the **Paired Variables** box. This variable should now be listed under the **Variable2** column.

c. Notice that after you complete Step 2b, another row becomes available for you to insert a variable pair. You can test as many variable pairs as you want.

3. Clicking the **Options...** button brings up the **Paired-Samples T Test: Options** dialog box. Here, you can indicate your **Confidence Interval Percentage** as well as how you want SPSS to handle missing data. When you are satisfied with your choices, click **Continue**.

4. That it! Click **OK** to conduct the analysis.

**Output**

When you click **OK**, SPSS will produce an **Output** screen displaying your results. Click the **Output** window to view your results (if it does not popup automatically).

The first table, titled **Paired Samples Statistics**, is shown in Figure 32.2. This table contains the descriptive statistics for each variable individually. Here, you have the **Mean**, sample size within each measurement (N), standard deviation (**Std. Deviation**), and the standard error of the mean (**Std. Error of the Mean**).

<table>
<thead>
<tr>
<th>Paired Samples Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Pre</td>
</tr>
<tr>
<td>Post</td>
</tr>
</tbody>
</table>

**Figure 32.2**

The second table, titled **Paired Samples Correlations**, contains the number of participants with data on both variables in the analysis (N), the bivariate correlation between your pre-treatment data (**Pre**) and your post-treatment data (**Post**), as well as the significance level (**Sig.**) for this correlation. In general, this correlation should be statistically significant because the data is from the same person across variable pairs.

The final table is titled **Paired Samples Test** and is shown in Figure 32.3. This table contains the results of the paired-samples t-test. This table provides you with the mean difference between variables (**Mean**), the standard deviation (**Std. Deviation**) of this difference, the standard error of the mean (**Std. Error of the Mean**), confidence intervals based on the percentage you selected in Step 3 above, and finally your **t** value (**t**), degrees of freedom (**df**), and **p-value** (**Sig. 2-tailed**).

<table>
<thead>
<tr>
<th>Paired Samples Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Pair 1: Pre - Post</td>
</tr>
</tbody>
</table>

**Figure 32.3**
Interpretation

Interpretation begins by examining the results provided in the *Paired Samples Test* table. Here, you will evaluate the *Sig. (2-tailed)* column to determine if your paired-samples *t*-test is statistically significant, which, for this example, it is. A statistically significant result for a repeated-measures design indicates that there was a statistically significant change from the first measurement to the second.

Next, you need to interpret the direction of the change (i.e., increase or decrease). To do so, you can either evaluate the *Mean* column in the *Paired Samples Test* table to evaluate the difference between the measurements, or you can simply evaluate the *Mean* in the *Paired Samples Statistics* table to compare means directly. They both provide identical results. For this example, there was an increase from *Pre* to *Post*.

Finally, you should evaluate the 95% confidence interval around the mean difference to evaluate the precision of your estimate.

**Example Results Section**

Prior to analysis, within the difference scores, no violation of normality was indicated and outliers were identified.

The results of a paired-samples *t*-test indicated a statistically significant treatment effect, *t*(9) = -8.714, *p* < .05, with an increase in minutes of sleep per night from pre-treatment to post-treatment (mean difference = 136.200, 95% CI = 100.843 to 171.557).
Chapter 33
WILCOXON SIGNED-RANK TEST

The Wilcoxon signed-rank test (or simply, the signed-rank test) is the nonparametric alternative to a paired-samples \( t \)-test. The test is used typically when the difference scores are not normally distributed, when the numerous outliers are indicated, or when the dependent variable is measured on an ordinal scale. The signed-rank test assesses for differences between two repeated measurements (or two matched groups). The test is based on ranked difference scores (e.g., difference between pretest and posttest), with the highest difference score receiving the highest rank and the lowest difference score receiving the lowest rank.

To show the similarities between the signed-rank test and the paired-samples \( t \)-tests, we will use the data and example provided in Chapter 32. Say you want to test the effectiveness of your new sleep treatment using each participant as his or her own control. Before beginning the treatment, you select your sample and measure the minutes of sleep each night for each participant. Then, you implement your treatment. One month after implementing your treatment, you measure the amount of sleep each night for those same participants. You now have pre-treatment and post-treatment data on the same dependent variable for the same subjects. The collected data is below.

<table>
<thead>
<tr>
<th>ID</th>
<th>Pre</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>250</td>
<td>396</td>
</tr>
<tr>
<td>2</td>
<td>302</td>
<td>417</td>
</tr>
<tr>
<td>3</td>
<td>210</td>
<td>285</td>
</tr>
<tr>
<td>4</td>
<td>285</td>
<td>402</td>
</tr>
<tr>
<td>5</td>
<td>280</td>
<td>460</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Pre</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>267</td>
<td>428</td>
</tr>
<tr>
<td>7</td>
<td>330</td>
<td>500</td>
</tr>
<tr>
<td>8</td>
<td>235</td>
<td>285</td>
</tr>
<tr>
<td>9</td>
<td>265</td>
<td>480</td>
</tr>
<tr>
<td>10</td>
<td>246</td>
<td>379</td>
</tr>
</tbody>
</table>

Following the data entry procedure described in Chapters 1 and 2, label your first variable \( ID \), your second variable \( Pre \), and your final variable \( Post \). Enter the data appropriately.

Assumptions

The only assumption of the signed-ranks test is that you do not have a great number of tied values. That is, participants who have the same value on both variables (i.e. no change). This is most common when the dependent variable is ordinal with a small number of categories. Regardless, tied values essentially indicate that your treatment or intervention was not effective.

Analysis

Assuming few or no tied values, to conduct the signed-rank test on the example data above:

1. Click Analyze, then choose Nonparametric tests, choose Legacy Dialogs, and finally click 2 Related Samples… to bring up the Two-Related Samples Tests dialog box.
2. Exactly like the paired-samples \( t \)-test, you are required to designate variable pairs. These variable pairs will be indicated in the Test Pairs: box.
a. Click the **Pre** variable on the left hand side of the dialog box and then click the right arrow (→) next to the **Test Pairs:** box. This variable should now be listed under the **Variable1** column.

b. Click the **Post** variable on the left hand side of the dialog box and then click the right arrow (→) next to the **Test Pairs:** box. This variable should now be listed under the **Variable2** column.

3. Next, you must indicate what the **Test Type** is. Here, you have four options. Conveniently, the **Wilcoxon** checkbox is the default, so simply make sure this box remains checked.

4. Finally, you can click the **Options…** button, where you can select two descriptive statistics or identify how you want SPSS to handle your missing data.

   a. The **Descriptive** option displays the mean, standard deviation, minimum, maximum, and the number of non-missing values. The **Quartiles** option displays values corresponding to the 25th, 50th, and 75th percentiles.

   b. Under the **Missing Values** option, choosing to **Exclude cases test-by-test** evaluates each test separately for missing values. Choosing **Exclude cases listwise** excludes participants with missing values for any variable from all analyses.

   When you are satisfied with your selections, click **Continue**.

5. That’s it! Click **OK** to conduct the signed-rank test.

**Output**

When you click **OK**, SPSS will produce an **Output** screen displaying your results. Click the **Output** window to view your results (if it does not pop up automatically).

The first table is titled **Ranks** and is provided in Figure 33.1. Here, you are provided with frequency counts (*N*) of negative and positive ranks as well as ties. Note that a tie indicates no change from pre-treatment to post-treatment, or, stated another way the pre-treatment and post-treatment scores were identical. You are also provided with the **Mean Rank** and **Sum of Ranks** for both **Negative Ranks** and **Positive Ranks**.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean Rank</th>
<th>Sum of Ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post - Pre</td>
<td>0a</td>
<td>.00</td>
<td>00</td>
</tr>
<tr>
<td></td>
<td>10b</td>
<td>5.50</td>
<td>55.00</td>
</tr>
<tr>
<td>Positive Ranks</td>
<td>0c</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ties</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 33.1
The second table, titled *Test Statistics*, is shown in Figure 33.2. Here, you are provided with your test statistic. While I will not describe the calculation of this value, it is important to know that for group sizes above 10 an approximation of the normal distribution is used. Thus, statistical significance can be determined from a $z$-test, which is provided in the $Z$ row. Note, the negative $Z$ indicates the rank sums are lower than their expected values, but because the normal distribution is symmetric, you can remove the negative if you choose. Finally, you are presented with your $p$-value (*Asymp. Sig. (2-tailed)*). The asymptotic significance is only applicable for larger samples ($N > 10$) and is the significance value if the normal distribution was used.

<table>
<thead>
<tr>
<th>Test Statistics$^b$</th>
<th>Post - Pre</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z$</td>
<td>-2.803$^a$</td>
</tr>
<tr>
<td>Asymp. Sig. (2-tailed)</td>
<td>.005</td>
</tr>
</tbody>
</table>

*a. Based on negative ranks.*

*b. Wilcoxon Signed Ranks Test*

Figure 33.2

**Interpretation**

Interpretation begins by evaluating the statistical significance of the $z$-test found in the *Test Statistics* table. If this result is statistically significant, you will consider both the number of *Negative Ranks* and *Positive Ranks* as well as their *Mean Rank* found in the *Ranks* table. A higher *Mean Rank* for *Positive Ranks* indicates an increase in score from pre-treatment to post-treatment. Conversely, a higher *Mean Rank* for *Negative Ranks* indicates an decrease in score from pre-treatment to post-treatment.

From the example data, the $z$-test was statistically significant with no negative ranks. This indicates the treatment increased ranked minutes of sleep per night from *Pre* to *Post*.

**Example Results Section**

The results of a Wilcoxon signed-rank test indicated a statistically significant increase in minutes of sleep per night from pre-treatment to post-treatment, $Z = -2.803$, $p < .05$. 


Chapter 34
ONE-WAY REPEATED-MEASURES ANOVA

A one-way repeated-measures ANOVA is an extension of a paired-samples t-test to situations where you are measuring the same group of individuals on three or more separate occasions—think, pre-test, post-test, and one-year follow-up. Alternatively, you can use this analysis for three or more matched groups, but again, the more common use involves repeated measurements. If necessary, please re-read or consult Chapter 32 prior to working through this Chapter.

As an example, reconsider the study of your new sleep treatment extended to include a one-year follow-up. Based on the paired-samples t-test data analyzed in Chapter 32, we know the new sleep treatment is effective from pre-treatment to post-treatment; however, now you want to assess the persistence of your sleep treatment over time on a new sample of participants to assess its long-term success. For this experiment, you assessed minutes of sleep per night before implementing your treatment. Then, six months after implementing your treatment, you assess minutes of sleep again. Finally, to assess long-term success you assess minutes of sleep again at a one-year follow-up. At the completion of your study, you will have three measurements for each participant. The data collected is provided below.

<table>
<thead>
<tr>
<th>ID</th>
<th>Pre</th>
<th>Six</th>
<th>Year</th>
<th>ID</th>
<th>Pre</th>
<th>Post</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>250</td>
<td>396</td>
<td>340</td>
<td>6</td>
<td>267</td>
<td>428</td>
<td>375</td>
</tr>
<tr>
<td>2</td>
<td>302</td>
<td>417</td>
<td>346</td>
<td>7</td>
<td>330</td>
<td>500</td>
<td>382</td>
</tr>
<tr>
<td>3</td>
<td>210</td>
<td>315</td>
<td>250</td>
<td>8</td>
<td>220</td>
<td>342</td>
<td>300</td>
</tr>
<tr>
<td>4</td>
<td>285</td>
<td>402</td>
<td>350</td>
<td>9</td>
<td>265</td>
<td>480</td>
<td>450</td>
</tr>
<tr>
<td>5</td>
<td>280</td>
<td>460</td>
<td>412</td>
<td>10</td>
<td>246</td>
<td>379</td>
<td>287</td>
</tr>
</tbody>
</table>

Following the data entry procedures described in Chapters 1 and 2, label the first variable ID, the second variable Pre, the third variable Post, and the last variable Year. Enter the data appropriately.

Assumptions

The assumptions of the one-way repeated measures ANOVA include absence of univariate outliers, normality and homoscedasticity of residuals as well as sphericity and balanced time periods.

Univariate (i.e., one variable) outliers are identified by evaluating z-scores and a histogram of dependent variable raw scores within each measurement individually (see Chapters 14 and 15, respectively). Although z-scores greater than 3.29 or less than -3.29 may suggest a possible outlier, you need to determine whether the value is actually disconnected from the rest of the distribution using within-measurement histograms. If outliers are identified you have several options. You can remove them from analysis through deletion (do not forget to describe them in your results section), transform the dependent variable (not advised), or use a nonparametric alternative that uses ranked data instead of the actual, raw scores (Chapter 35).

The normality assumption requires the residuals to be normally distributed. However, residual values are not interpreted as they have been previously. While a full discussion is beyond the scope of the Chapter, briefly, repeated-measures designs increase statistical power by
removing between-person differences. The residuals produced by a repeated-measures analysis are what remain after between-person differences have been removed. That is, the residuals represent only within-person variance. Because residuals are based on the repeated measurements, a residual value is produced for each repeated measurement. Thus, for the example above, three residual values will be produced, and the normality assumption on each residual distribution separately. Similar to one-way between-subjects ANOVA, this assumption is evaluated by Q-Q plots, histograms, and Kolmogorov-Smirnov tests of residuals (Chapter 15). With that said, simulation studies have shown that the analysis is robust (i.e., still provide correct inference) to violation of normality with greater than 20 error degrees of freedom. We will request residuals below, however.

Homoscedasticity requires the residuals to have a constant variance across all values of each predicted repeated measurement. Homoscedasticity is evaluated by requesting a scatterplot matrix where the predicted repeated measurement values are plotted against the standardized residuals. If these scatterplots appear rectangular, the assumption is considered satisfied.

The next assumption is termed sphericity. This assumption states that the variances of the difference scores between the repeated measurements are equal. This is the repeated measures version of homogeneity of variance. Note that a violation of homoscedasticity goes hand-in-hand with a violation of the sphericity. Thus, if sphericity is satisfied, homoscedasticity of residuals usually is too. Sphericity is a tough assumption to satisfy, however, especially as the number of repeated measurements increases. It is tested during the analysis, and, if violated, SPSS provides several options to adjust the test statistics. The options are discussed below; however, to note that the multivariate approach to repeated measures or a linear mixed model (neither discussed in this text) would be a more appropriate modeling strategy if sphericity were violated.

The final assumption is that the repeated measurements are balanced across time (i.e., exactly equal intervals of time between measurements for each person). This assumption, of course, is not applicable to cross-sectional matched-groups designs. For repeated-measures designs, if the duration of time between measurements is unbalanced, time becomes a confounding factor and the analysis will yield incorrect results.

**Analysis**

Assuming absence of univariate outliers and equal time intervals across the repeated measurements, to conduct a one-way repeated-measures ANOVA:

1. Click **Analyze**, choose **General Linear Model**, and finally click **Repeated Measures**… to bring up the **Repeated Measures Define Factor(s)** dialog box shown in Figure 34.1.
   a. The first box you see is labeled **Within-Subject Factor Name:** (default states *factor1*). Here, you have the option to tell SPSS what your within-subject factor is. That is, for the study above, because the measurements were made at pre-treatment, 6-months post-treatment, and at a one-year follow-up, your within-subject factor is time. That is, the passage of time separates the measurements. Thus, type **Time** into this box.
   b. The next box is labeled **Number of Levels:**. This box is required, as you must tell SPSS how many repeated measures you have. Enter 3 (this is the number of times you measured your dependent variable—pre-treatment, post-treatment, and one year follow-up).
c. Click **Add** next to the first large box. This places the within-subjects factor name and number of levels into the first box from the top. If you need to make edits to this or if you want to remove the name and levels altogether, simply click *Time(3)* and then click either the **Change** or **Remove** buttons.

d. Finally, you have the option to name (or label) your dependent variable in the *Measure Name:* box. Again, this is optional, but you can enter *Minutes* into this box and click **Add.** Similarly to above, if you need to make edits to this or if you want to remove the name, simply click *Minutes* and then click either the **Change** or **Remove** buttons.

When you have completed these steps, click **Define.** This brings up the *Repeated Measures* dialog box shown in Figure 34.2.

---

**Figure 34.1**

**Repeated Measures Define Factor(s) dialog box**

- **Within-Subject Factor Name:**
  - *Time(3)*
- **Number of Levels:**
  - *
- **Measure Name:**
  - *Minutes*

**Figure 34.2**

**Repeated Measures dialog box**
2. For a one-way repeated-measures ANOVA, the only portion of this dialog box you are going to have to be concerned with is the **Within-Subjects variables:** box near the top. This is where you are going to place your measurement variables in chronological order. You will notice initially that this box has the same number of levels as the number you indicated in Step 1b above. If you followed the example correctly, you should have three slots available (\(_?_(1,\text{Minutes})\), \(_?_(2,\text{Minutes})\), and \(_?_(3,\text{Minutes})\)).

   a. Click to select the **Pre** variable and then click the right arrow (\(\Rightarrow\)) next to the **Within-Subjects variables:** box.

   b. Click to select the **Post** variable and then click the right arrow (\(\Rightarrow\)) next to the **Within-Subjects variables:** box.

   c. Click to select the **Year** variable and then click the right arrow (\(\Rightarrow\)) next to the **Within-Subjects variables:** box.

   If you have done this correctly, your dialog box should look like Figure 34.2.

3. You will also notice six buttons on the right side of this dialog box. For a one-way repeated-measures ANOVA, only the **Contrasts...**, **Plots...**, **Save...**, and **Options** buttons are applicable. They are described below.

   a. Clicking the **Contrasts...** button allows you to test for differences among levels of the factor(s) you specified in Steps 1a through 1c above. You can select from several different contrasts described below. Note the default is **Polynomial**.

      i. **Deviation**: Compares the mean of each level, other than a reference category, to the grand mean.

      ii. **Simple**: Compares the mean of each level to the mean of a specified level. This is useful when you have a control group.

      iii. **Difference**: Compares the mean of each level (except the first) to the mean of previous levels.

      iv. **Helmert**: Compares the mean of each level (except the last) to the mean of subsequent levels.

      v. **Repeated**: Compares the mean of each level (except the last) to the mean of the subsequent level.

      vi. **Polynomial**: Compares the linear effect, quadratic effect, cubic effect, and so on depending on how many repeated measurements you have. The first degree of freedom contains the linear effect across all categories; the second degree of freedom, the quadratic effect; and so on. These contrasts are used to perform a trend analysis.

      If you choose to **Change Contrast**, select the contrast you want from the drop-down menu and click **Change**. When you are satisfied, click **Continue**.

   b. Clicking the **Plots...** button allows you to create a line chart of the estimated marginal means for each repeated measurement.

      i. Click **Time** and then click the right arrow (\(\Rightarrow\)) next to the **Horizontal Axis** box. This tells SPSS to place the repeated measurements on the x-axis.
ii. Click **Add** to move this variable into the *Plots:* box. When you are satisfied with your selections, click **Continue.**

c. Clicking the **Save...** button brings up the *Repeated Measures: Save* dialog box. Here, you have the option to save new variables to your dataset that contain information about *Predicted Values, Diagnostics, Residuals,* or *Coefficient Statistics.* Most of these options allow you to test the assumptions of your model.

   i. The options listed under the *Predicted Values* section allow you to save the values that your specific model predicts for each participant. Remember, one-way repeated-measures ANOVA is based on the general linear model—a mathematical equation based on deviations from the overall mean and group means. You can save these values as *Unstandardized* or *Standard error* where an estimate of the standard deviation of the average value of the dependent variable for participants that have the same values of the independent variables.

   ii. The options listed under the *Diagnostics* section provide measures to identify participants with unusual combinations of values for the repeated measurements or those who have unusual influence on the outcome of the model. For both options, larger values equal larger change or influence.

      1. *Cook’s distance* is a measure of how much the residuals of all participants would change if the participant were removed from analysis.

      2. *Leverage values* indicate the relative influence of each observation on the model’s fit.

   iii. The options listed under the *Residuals* section provide the difference between the observed value of the repeated measurement and the model predicted value. Note that the *Standardized:* option will be used to evaluate the normality of residuals assumption.

      1. *Unstandardized:* Provides the difference between an observed value and the value predicted by the model.

      2. *Standardized:* Provides the residual divided by an estimate of its standard deviation. These are also known as Pearson residuals. Similar to z-scores, they have a mean of 0 and a standard deviation of 1.

      3. *Studentized:* Provides the residual divided by an estimate of its standard deviation. However, the standard deviation is allowed to vary across participants.

      4. *Deleted:* Provides the residual for a participant if that participant is excluded from analysis. It is the difference between the value of the repeated measurement and the adjusted predicted value.

iv. Finally, the options listed under the *Coefficient Statistics* section writes the variance-covariance matrix of parameter estimates, *t*-statistics,
significance values, and residual degrees of freedom. You have the option to Create a new dataset where you must specify a Dataset name: or you can Write a new data file which creates a complete new dataset. Selecting the latter requires you to click the File... button and specify the dataset you want to write onto.

When you are satisfied with your selections, click Continue.

d. Finally, by clicking the Options... button you can print a table of Estimated Marginal Means as well a number of descriptive statistics and diagnostic tests.

i. Under the Estimated Marginal Means section, in the Factor(s) and Factor Interactions: box, you will always be offered the option to display (OVERALL) means. If you choose this option, SPSS will print the grand mean, which considers all the data in your sample together irrespective of measurement point. More importantly, you will have the option to display the estimated marginal means for your individual measurement points (the variable you identified in Steps 1a through 1c). Click Time, and then click the right arrow ( ) next to the Display Means for: box.

ii. If you have three or more measurement points, the Compare main effects checkbox will be available. Clicking this checkbox will allow you to conduct post hoc tests (called pairwise comparisons) between all repeated measurements. Upon clicking the checkbox, the Confidence interval adjustment: box becomes available. Here, you can select three adjustments—LSD(none), Bonferroni, or Sidak. Adjustment is required to control for type I error due to multiple pairwise comparisons.

iii. Under the Display section, you have twelve options. The Descriptive statistics option prints means, standard deviations, and frequency counts for each repeated measurement. Estimates of effect size prints partial eta-squared, which is the proportion of total variability attributable to your repeated measurements. The Observed power option prints the power of the test based on your specific sample. The Parameter estimates option prints parameter estimates, standard errors, t-tests, and confidence intervals (similar to regression). Further, you can display SSCP matrices and the Residual SSCP matrix. Clicking Transformation matrix produces the M matrix. Homogeneity tests produces Levene’s test and the Spread-versus-level plot compares cell means, standard deviations, and variances across the level combinations of all factors (groups); however, because you have no between-subjects factor in a one-way repeated measures design these two options are not applicable. Choosing a Residual plot is useful for checking the homoscedasticity assumption. Finally, the Lack of fit and General estimable function options are not applicable in one-way repeated measures designs.

If you have selected all the options in bold type, your options dialog box should appear exactly like Figure 34.3. When you are satisfied with your selections, click Continue.
4. That’s it! Click **OK** to conduct the one-way repeated-measures ANOVA.

**Output**

When you click **OK**, SPSS will produce an **Output** screen displaying your results. Click the **Output** window to view your results (if it does not pop up automatically).

The first table is titled *Within-Subjects Factors* containing the names of your repeated measurements. The second table is titled *Descriptive Statistics*, which contains the *Mean*, standard deviation (*Std. Deviation*), and within-measurement sample size (*N*) for each repeated measurement. The next table is titled *Multivariate Tests*, which uses MANOVA (i.e., the multivariate approach to repeated measures, also known as profile analysis) for statistical analysis, but because we are using repeated-measures ANOVA, this table is ignored.

The next table you see is titled *Mauchly’s Test of Sphericity* shown in Figure 34.4. This table contains information pertaining to a fundamental assumption (sphericity) of repeated-measures ANOVA. Sphericity requires that the variances of the between measurement difference scores are relatively equal. *Mauchly’s W* is your test statistic and it follows a chi-square distribution, which is why you see the approximate (*Approx.*) Chi-Square column. The statistical test is based on the chi-square value, degrees of freedom (*df*) and the *p*-value provided in the *Sig.* column. Remember, *p* < .05 indicates a violation of the assumption, as the variances of the differences scores are not equal. Note that the probability of a type 1 error increase with a violation of this assumption. If sphericity is violated, you will use a correction provided in the next table. Finally, you are provided three *Epsilon* values. Epsilon values close to 1 indicate less violation of sphericity.
Figure 34.4

The results of your one-way repeated measures ANOVA are provided in the table titled *Tests of Within-Subjects Effects* shown in Figure 34.5. This table contains your selected *Sum of Squares*, degrees of freedom (*df*), *Mean Square*, omnibus *F*, significance level (*Sig.*) as well as your effect size (*Partial Eta Squared*). If Mauchly’s Test was not statistically significant, you will interpret the *Sphericity Assumed* row; if sphericity assumption was violated, consult the *Huynh-Feldt* row. Note that research has shown that the *Greenhouse-Geisser* and *Lower-bound* options are unnecessarily conservative—that is, their use has been shown to increase the probability of type II errors. Each correction simply multiplies the degrees of freedom by their associated epsilon. Because epsilon is less than 1, degrees of freedom are decreased, making the test more conservative; thus, controlling for type I errors. In this example, you can see that there is a statistically significant change across the repeated measurements based on the *F* and *Sig.* columns in the *Huynh-Feldt* row.

Figure 34.5

The next table is titled *Tests of Within-Subjects Contrasts* and is relevant if you decided to do a trend analysis (as mentioned in Step 3a above). Here, you are informed whether your measurement means followed a *Linear* or *Quadratic* trend over time, with a *Linear* trend being a straight line and a *Quadratic* trend having one bend (think, parabolic). For each trend, you are provided with your selected *Sum of Squares*, degrees of freedom (*df*), *Mean Square*, *F*, significance level (*Sig.*) and *Partial Eta Squared*. For interpretation, you only consider the highest statistically significant trend. So, the *Quadratic* trend is statistically significant which indicates there is a significant change in direction across the measurement points.
The next table, titled *Tests of Between-Subjects Effects* is not relevant to the current example and can be ignored because a one-way repeated-measures ANOVA has no between subjects independent variable. That is, there is only one group of participants.

With three or more repeated measurements, the $F$-test result provided in the *Tests of Within-Subjects Effects* table is an omnibus test. This means that a statistically significant result indicates that at least two of the measurement points are statistically different from one another. However, this test does not indicate which two measurements are different. Thus, the next section, titled *Estimated Marginal Means* is incredibly important to interpretation following a statistically significant omnibus $F$-test. The *Estimates*, shown in Figure 34.6, contains the *Mean*, standard error (*Std. Error*), and 95% *Confidence Interval* information for each measurement.

<table>
<thead>
<tr>
<th>Measure Minutes</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>Mean</td>
<td>Std Error</td>
<td>Lower Bound</td>
<td>Upper Bound</td>
</tr>
<tr>
<td>1</td>
<td>265.480</td>
<td>11.476</td>
<td>239.536</td>
<td>291.484</td>
</tr>
<tr>
<td>2</td>
<td>411.900</td>
<td>10.471</td>
<td>370.115</td>
<td>453.685</td>
</tr>
<tr>
<td>3</td>
<td>349.200</td>
<td>10.327</td>
<td>306.927</td>
<td>391.903</td>
</tr>
</tbody>
</table>

**Figure 34.6**

The subsequent table, titled *Pairwise Comparisons* and shown in Figure 34.7, provides adjusted post hoc tests (if you selected *Bonferroni* or *Sidak* in Step 3c above) comparing all repeated measurements to each other. Remember, this section is evaluated if, and only if, the omnibus $F$-test was statistically significant. A statistically significant result is indicted by an asterisk next to the *Mean Difference* value. Note the *Mean Difference* (I-J) column simply provides the difference in estimated marginal means between the two groups being compared. You are also provided with the standard error (*Std. Error*), *p*-value (*Sig.*), and 95% *Confidence Interval for Difference*.

<table>
<thead>
<tr>
<th>Measure Minutes</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) Time</td>
<td>(j) Time</td>
<td>Mean Difference (I-J)</td>
<td>Std Error</td>
<td>Sig*</td>
<td>95% Confidence Interval for Difference*</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>-148.400</td>
<td>11.027</td>
<td>.000</td>
<td>-178.747 to -118.055</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>-23.700</td>
<td>14.671</td>
<td>.001</td>
<td>-52.320 to 64.920</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>158.400</td>
<td>11.027</td>
<td>.000</td>
<td>114.065 to 178.747</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>62.700</td>
<td>8.136</td>
<td>.001</td>
<td>36.867 to 88.533</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>33.700</td>
<td>14.671</td>
<td>.000</td>
<td>40.060 to 127.320</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>-62.700</td>
<td>8.136</td>
<td>.000</td>
<td>-88.533 to -36.867</td>
<td></td>
</tr>
</tbody>
</table>

Based on estimated marginal means:

* The mean difference is significant at the .05 level.

a. Adjustment for multiple comparisons: Bonferroni.

**Figure 34.7**
Finally, SPSS produces a line chart (if you requested one in Step 3b above) titled *Estimated Marginal Means of (*) shown in Figure 34.7. Note that the asterisk represents the name you provided your dependent variable in Step 2d above (i.e., *Minutes* in the example). Note that if you did not name your dependent variable, the asterisk would be replaced with *MEASURE_1*. This chart plots the estimated marginal means of each variable allowing you to compare the repeated measures to each other in graphical format. For this graph, your dependent variable (*Minutes*) is on the *y*-axis, while your data collection points are placed on the *x*-axis (labeled 1, 2, and 3 for *Pre*, *Six*, and *Year*, respectively). This plot is incredibly useful to interpretation of the pairwise comparisons (and trend analysis) as you can see that there was a substantial drop off in *Minutes* from *Post* to *Year*.

![Figure 34.7](image)

**Figure 34.7**

**Interpretation**

Interpretation begins by evaluating the tenability of the normality of residuals assumption. You will notice that in your dataset you have saved (or created) three new variables labeled *ZRE_1*, *ZRE_2*, and *ZRE_3*. Use the Explore procedure (Chapter 15) to evaluate the normality of the standardized residuals for all three measurements. From the histograms provided in Figure 34.8, you can see the distributions are normal. This evaluation was further supported by the Q-Q plot and non-statistically significant Kolmogorov-Smirnov test (*p* > .05). Thus, the normality assumption is considered satisfied.

![Figure 34.8](image)
Next, you will consult the Mauchly’s Test table to determine whether the sphericity assumption has been violated. In this example, Mauchly’s test is not statistically significant, indicating that the sphericity assumption has not been violated. Further, we are going to assume homoscedasticity of residuals is satisfied as well. Thus, you can next evaluate the Tests of Within-Subjects Effects table to determine whether you have a statistically significant omnibus $F$. Remember, the specific row you consult based on whether the sphericity assumption has been violated. For this example, because sphericity was not violated, we will consult the Sphericity Assumed row.

Note, however, that if the sphericity assumption was violated, you would report results for one of the corrections provided, typically the Huynh-Feldt correction (although, again, the multivariate results or a linear mixed model would be more appropriate). Notice for the Huynh-Feldt correction, the degrees of freedom from the Sphericity Assumed row were simply multiplied by the Huynh-Feldt Epsilon value of .779 found in the Mauchly’s Test of Sphericity table. Thus, the degrees of freedom in the Tests of Within-Subjects Effects table were reduced to 1.559 and 14.029, for Time and Error(Time), respectively (i.e., $2 \times .779 = 1.559$ and $18 \times .779 = 14.029$). If you use the Huynh-Feldt correction, you would report these degrees of freedom in your results section.

Next, if the omnibus $F$ is statistically significant and you have three or more repeated measurements, you will consult the Pairwise Comparisons table to determine which measurements are statistically different from one another. From the example, the omnibus F-test is statistically significant, and the results presented in the Pairwise Comparisons table indicate that all three measurements are statistically different from one another using the Bonferroni adjustment (see footnote $a$ in the Pairwise Comparisons table to see which correction was used).

Next, you need to interpret the direction of the change (i.e., increase or decrease). To do so, you can either evaluate the Mean Difference (I-J) column in the Pairwise Comparisons table or simply evaluate the Means provided in the Estimates table, as they both provide identical results. For this example, there was a statistically significant increase from pretest to the six-month follow-up and a statistically significant decrease from the six-month follow-up to the one-year follow-up. However, participants still slept longer at the one-year follow-up compared to pretest.

Finally, you should evaluate the 95% confidence interval around the mean difference to evaluate the precision of your estimate.

**Example Results Section**

Prior to analysis, no outliers were identified and time was balanced across repeated measurements. Further, no violation of normality of residuals or sphericity was indicated.

The results of a one-way repeated-measures ANOVA indicated a statistically significant change across the repeated measurements, $F(2, 18) = 79.156$, $p < .05$, partial $\eta^2 = .898$. Bonferroni-adjusted pairwise comparisons indicated a statistically significant increase in minutes of sleep per night between pretest and six-month follow-up (mean difference = 146.400, 95% CI = 114.053 to 178.747; $p < .05$), but a statistically significant decrease from six-month follow-up to one-year follow-up (mean difference = 62.700, 95% CI = 38.833 to 86.567; $p < .05$). However, the statistically significant increase in minutes of sleep per night was maintained from pretest to one-year follow-up (mean difference = 83.700, 95% CI = 40.080 to 127.320; $p < .05$).
Chapter 35
FRIEDMAN’S TWO-WAY ANOVA BY RANKS

The Friedman two-way ANOVA by ranks test (or simply the Friedman’s test) is the nonparametric alternative to the one-way repeated measures ANOVA. That test can also be considered an extension of the signed-rank test to situations with three or more repeated measurements. Similar to the other nonparametric tests, it is most often used when the distributions of your repeated measurements are not normal or when the dependent variable is measured on an ordinal scale. Friedman’s test is initially based on within-participant ranked data, with higher scores receiving higher ranks. These ranks are then summed within each measurement individually, and these ranked sums are used to assess for statistically significant differences between repeated measurements.

Because Friedman’s test is used for three or more repeated measurements (or matched groups), it is considered an omnibus test. That is, a statistically significant Friedman’s test indicates there is a statistically significant difference between at least two measurements, but does not identify which measurements differ. Thus, you will need to conduct a series of adjusted post hoc tests to test for differences between each measurement. The post hoc test used is the signed-rank test described in Chapter 33.

To show the similarity between Friedman’s test and the one-way repeated measures ANOVA, we will use the example presented in Chapter 34, with different that violates the normality of residuals assumption. Reconsider the study of your new sleep treatment extended to include a one-year follow-up to assess long-term effectiveness. For this experiment, you assessed minutes of sleep per night before implementing your treatment. Then, after implementing your treatment, you assess minutes of sleep again at a six-month follow-up. Finally, to assess long-term success you assess minutes of sleep again at a one-year follow-up. At the completion of your study, you will have three measurements for each participant. The data collected is provided below.

<table>
<thead>
<tr>
<th>ID</th>
<th>Pre</th>
<th>Six</th>
<th>Year</th>
<th>ID</th>
<th>Pre</th>
<th>Post</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>250</td>
<td>396</td>
<td>340</td>
<td>6</td>
<td>267</td>
<td>428</td>
<td>400</td>
</tr>
<tr>
<td>2</td>
<td>302</td>
<td>417</td>
<td>346</td>
<td>7</td>
<td>330</td>
<td>500</td>
<td>498</td>
</tr>
<tr>
<td>3</td>
<td>210</td>
<td>285</td>
<td>250</td>
<td>8</td>
<td>235</td>
<td>285</td>
<td>189</td>
</tr>
<tr>
<td>4</td>
<td>285</td>
<td>402</td>
<td>350</td>
<td>9</td>
<td>265</td>
<td>480</td>
<td>428</td>
</tr>
<tr>
<td>5</td>
<td>280</td>
<td>460</td>
<td>401</td>
<td>10</td>
<td>246</td>
<td>379</td>
<td>301</td>
</tr>
</tbody>
</table>

Following the data entry procedures described in Chapters 1 and 2, label the first variable ID, the second variable Pre, the third Six, and the last Year. Enter the data appropriately.

Analysis

1. Click Analyze, then choose Nonparametric Tests, choose Legacy Dialogs, and finally click K Related Samples….
2. Initially, you will notice that all of your variables in your dataset are listed on the left hand side of this dialog box. The Test Variables: box is where you will identify the repeated measures on which you want to conduct the analysis.
a. Click to select the **Pre** variable and then click the right arrow (→) next to the **Test Variables**: box.

b. Click to select the **Post** variable and then click the right arrow (→) next to the **Test Variables**: box.

c. Click to select the **Year** variable and then click the right arrow (→) next to the **Test Variables**: box.

3. Next, you need to tell SPSS what the **Test Type** is. In this section, you have three options, with the default being **Friedman**. Make sure this checkbox remains checked.

4. Finally, you have the option to click the **Statistics**... button where you can request **Descriptive** statistics, which will print the sample size, mean, standard deviation, minimum, and maximum for each repeated measurement. You can also choose **Quartiles**, which will print the 25th, 50th, and 75th percentile values for each measurement. When you are satisfied with your selections, click **Continue**.

5. That’s it! Click **OK** to conduct the analysis.

**Output**

When you click **OK**, SPSS will produce an **Output** screen displaying your results, which should be identical to those presented in Figure 35.1. Click the **Output** window to view your results (if it does not pop up automatically).

Within the results, you are presented with two tables containing the results of Friedman’s test. The first table is labeled **Ranks**. Here, you are only provided with the **Mean Rank** of each repeated measurement. The second table is titled **Test Statistics**. Here, you are provided with your test statistics. The first row of this table contains the number of participants with data for all repeated measurements (**N**). The test statistics for Friedman’s test uses the chi-square distribution to determine statistical significance. Thus, your inference is based on the **Chi-square** value and degrees of freedom (**df**) which is the number of repeated measurements minus 1. Finally, you are presented with your **p-value** (**Asymp. Sig.**).

---

<table>
<thead>
<tr>
<th>Ranks</th>
<th>Mean Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre</td>
<td>1.10</td>
</tr>
<tr>
<td>Sck</td>
<td>3.00</td>
</tr>
<tr>
<td>Year</td>
<td>1.90</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Statistics&lt;sup&gt;a&lt;/sup&gt;</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>10</td>
</tr>
<tr>
<td>Chi-square</td>
<td>18.203</td>
</tr>
<tr>
<td>df</td>
<td>2</td>
</tr>
<tr>
<td>Asymp. Sig.</td>
<td>.003</td>
</tr>
</tbody>
</table>

<sup>a</sup> Friedman Test

**Figure 35.1**

193
Interpretation

Interpreting the results of your study is based on the information provided to you in Test Statistics table. For this example, the result is statistically significant, indicating that at least two of the repeated measurements are statistically different from one another. Because the Friedman’s test is an omnibus test, the Ranks table is not particularly useful to interpretation; however, the statistically significant result indicates that the two ranks with the largest difference are probably statistically significant. That is, the difference between Pre and Six (1.10 and 3.00, respectively).

To determine which measurements differ statistically you need to conduct a series of adjusted post hoc signed-rank tests. Because you have three repeated measurements, three post hoc tests are required (i.e., Pre vs. Six, Pre vs. Year, and Six vs. Year). Thus, the adjusted alpha you will use for all post hoc tests is .017 (i.e., .05/3). While I do not show you how to conduct the post hoc signed-rank tests, I have included their results below. You should be able to replicate these results. See Chapter 33 regarding how to conduct the signed-rank test.

Example Results Section

The result of Friedman’s test indicated a statistically significant change in ranked minutes of sleep per night across the repeated measurements, $\chi^2(2) = 18.200, p < .05$. Bonferroni-adjusted signed-rank tests indicated a statistically significant increase in ranked minutes of sleep per night from pretest to six-month follow-up ($Z = -2.803, p < .017$), but a statistically significant decrease from six-month follow-up to one-year follow-up ($Z = -2.805, p < .017$). However, the increase in ranked minutes of sleep per night remained statistically significant between pretest and one-year follow-up ($Z = -2.497, p < .017$).
Chapter 36
MIXED BETWEEN-WITHIN ANOVA

A mixed between-within ANOVA design is a factorial design that has one (or more) between-subjects independent variable(s) and one (or more) repeated measures independent variable(s). This analysis can be thought of as an extension of factorial between-subjects ANOVA (Chapter 30) to include an additional repeated-measures independent variable, with the primary difference being a mixed between-within ANOVA has a between-subjects-by-repeated measurement interaction effect.

Consistent with any factorial design, the primary advantage is that you can assess the interaction between your between-subjects independent variable(s) and your repeated measures independent variable(s). That is, you can assess whether two (or more) groups change differently over the course of your repeated measurements. This analysis is often used in longitudinal studies to pinpoint when statistically significant change occurred during the study period.

For a design with one between-subjects independent variable and one repeated-measures independent variable, you are provided with two main effects (i.e., one between-subjects, one repeated measures) and an interaction effect; thus, three separate statistical tests are provided. Statistical significance is determined separately for each main effect and interaction effect.

Continuing with the sleep example, say you want to know whether your new sleep treatment is more effective than an over-the-counter treatment during a one-year study period. For this experiment, say you have two randomly assigned groups. One group receives an effective over-the-counter treatment, whereas a second group receives your new sleep treatment. In addition, you have three repeated measurements. You assess minutes of sleep per night before implementing the treatments, six months after implementing the treatments, and one-year after implementing the treatments. This is an example of a 2 X 3 mixed between-within design because there are two levels within the between-subjects independent variable and three repeated measures. Note that if you added a third treatment level you would have a 3 X 3 factorial design or if you added another between-subjects independent variable with two levels to the original design you would then have a 2 X 2 X 3 design. The collected data is provided below:

<table>
<thead>
<tr>
<th>ID</th>
<th>Tx</th>
<th>Pre</th>
<th>Six</th>
<th>Year</th>
<th>ID</th>
<th>Tx</th>
<th>Pre</th>
<th>Six</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>280</td>
<td>326</td>
<td>352</td>
<td>11</td>
<td>1</td>
<td>269</td>
<td>370</td>
<td>426</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>239</td>
<td>315</td>
<td>360</td>
<td>12</td>
<td>1</td>
<td>250</td>
<td>350</td>
<td>435</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>248</td>
<td>370</td>
<td>385</td>
<td>13</td>
<td>1</td>
<td>258</td>
<td>313</td>
<td>399</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>261</td>
<td>301</td>
<td>326</td>
<td>14</td>
<td>1</td>
<td>262</td>
<td>335</td>
<td>456</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>227</td>
<td>300</td>
<td>302</td>
<td>15</td>
<td>1</td>
<td>219</td>
<td>341</td>
<td>486</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>301</td>
<td>376</td>
<td>359</td>
<td>16</td>
<td>1</td>
<td>284</td>
<td>375</td>
<td>500</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>282</td>
<td>347</td>
<td>335</td>
<td>17</td>
<td>1</td>
<td>285</td>
<td>359</td>
<td>476</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>310</td>
<td>358</td>
<td>350</td>
<td>18</td>
<td>1</td>
<td>272</td>
<td>400</td>
<td>512</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>316</td>
<td>400</td>
<td>375</td>
<td>19</td>
<td>1</td>
<td>289</td>
<td>367</td>
<td>429</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>215</td>
<td>278</td>
<td>300</td>
<td>20</td>
<td>1</td>
<td>241</td>
<td>383</td>
<td>455</td>
</tr>
</tbody>
</table>

Following the data entry procedures described in Chapters 1 and 2, label the first variable \(ID\), the second \(Tx\) (0 = OTC; 1 = Treatment), the third variable \(Pre\), the fourth variable \(Six\), and the final variable \(Year\). Enter your data appropriately.
**Assumptions**

The assumptions of the mixed between-within ANOVA consist of a combination of the assumptions for one-way between-subjects ANOVA and one-way repeated measures ANOVA. Several assumptions are applicable only to the repeated measures, several applicable only to the between-subjects variables(s). The assumptions include absence of univariate outliers and homogeneity of variance as well as independence, normality, and homoscedasticity of residuals in addition to multi-sample sphericity and equal time intervals between measurements.

The assumption applicable to all variables is absence of univariate outliers. Univariate (i.e., one variable) outliers are identified by evaluating $z$-scores and a histogram of dependent variable raw scores within each measurement individually and within each group individually (see Chapters 14 and 15, respectively). Although $z$-scores greater than 3.29 or less than -3.29 may suggest a possible outlier, you need to determine whether the value is actually *disconnected* from the rest of the distribution using within-group and within-measurement histograms. If outliers are identified you have several options. You can remove them from analysis through deletion (do not forget to describe them in your results section) or transform the dependent variable (not advised). Note that a nonparametric alternative is not widely available to evaluate a between-within interaction.

The assumption applicable to the between-subjects independent variable is homogeneity of variance. This assumption requires that the repeated measurements have roughly the same variance across groups, and is evaluated for each repeated measurement individually. If a violation is detected, your statistical test will become too liberal and the probability of type I error increases. This assumption is tested during analysis by Levene’s test, where a non-statistically significant result indicates the variances are not statistically different, thereby satisfying the assumption. If a violation is indicated, however, you should make the $F$-test for the between-subjects effect more conservative by reducing alpha (e.g., from .05 to .01).

Independence of residuals requires the residuals to be uncorrelated across groups, but (obviously) not across repeated measurements. This is a change from the independence assumption described previously. Briefly, a residual is the difference between predicted and actual value of the repeated measurement. Because mixed between-within ANOVA is a special case of the general linear model (see Chapters 22 and 23 for slightly more detailed discussion), each participant will have their own residual value for each repeated measurement. Again, independence of residuals is technically a design issue that requires groups to be (and stay) mutually exclusive across measurements.

The normality of residuals assumption requires the residuals to be normally distributed. For mixed between-within ANOVA, two types of residuals are calculated—residuals representing within-subject variance and residuals representing between-subject variance. While a full discussion is beyond the scope of the Chapter, briefly, any design that includes repeated-measures increase statistical power by removing between-subject differences. Thus, the residuals produced for repeated-measures effects (i.e., the repeated measures main effect and between-within interaction) are what remain after between-subject differences have been removed. That is, these residuals represent only *within-subject* variance. However, a between-subjects main effect considers only between-subject differences; thus, residuals for this effect represent *between-subject* variance (in this case, the residuals represent total variance). Residual values for the repeated measures are calculated for each repeated measurement individually, similar to one-way repeated-measures ANOVA. However, you cannot request between-subject residuals.
directly in SPSS. You can obtain between-subject residuals by calculating the mean of the repeated measurements for each participant (Chapter 13) and then using the procedure described in Chapter 28 to save the standardized residuals. This process is described in detail below. Although I know this seems like a lot of work, and it is, the steps involved to calculate the between-subjects residuals provides a good illustration of how main effects simply collapse (i.e., use the mean) across levels of the independent variable not involved in the analysis directly. Regardless of whether the residual represents between-subject or within-subject variance, the normality assumption is evaluated by Q-Q plots, histograms, and Kolmogorov-Smirnov tests on these residuals (Chapter 15). With that said, simulation studies have shown that the analysis is robust (i.e., still provide correct inference) to violation of normality with greater than 20 error degrees of freedom. We will request and evaluate the residuals for all effects below, however.

Heteroscedasticity of residuals requires the residuals to have a constant variance across all values of the repeated measurements. A violation of heteroscedasticity goes hand-in-hand with a violation of the homogeneity of variance (and multi-sample sphericity, discussed below). Thus, if homogeneity of variance is satisfied, heteroscedasticity of residuals usually is too.

Additional assumptions applicable only to the within-subjects or repeated-measures effect include multi-sample sphericity and equal time intervals between measurements. Multi-sample sphericity states that the variances of the difference scores between the repeated measurements are equal. If you have more than one repeated-measures independent variable, multi-sample sphericity is tested for each variable individually. Note that a violation of heteroscedasticity of residuals goes hand-in-hand with a violation of the sphericity. Thus, if sphericity is satisfied, heteroscedasticity of residuals usually is too. It should be clear that for a mixed between-within ANOVA you now have two proxy tests for homogeneity of variance (i.e., homogeneity of variance and sphericity). Sphericity is a tough assumption to satisfy, however, especially as the number of repeated measurements increases. It is tested during the analysis, and, if violated, SPSS provides several options to adjust the test statistics. The options are discussed below; however, to note that the multivariate approach to repeated measures or a linear mixed model (neither discussed in this text) would be a more appropriate modeling strategy if sphericity were violated.

The final assumption is that the repeated measurements are balanced across time (i.e., exactly equal intervals of time between measurements for each person). This assumption, of course, is not applicable to cross-sectional matched-groups designs. For repeated-measures designs, if the duration of time between measurements is unbalanced, time becomes a confounding factor and the analysis will yield incorrect results.

### Analysis

Assuming independence of residuals, absence of outliers, and equal duration between measurements are tenable prior to analysis, to conduct a mixed between-within ANOVA:

1. **Click Analyze**, choose **General Linear Model**, and finally click **Repeated Measures**… to bring up the **Repeated Measures Define Factor(s)** dialog box, identical to Figure 34.1.
   a. The first box you see is labeled *Within-Subject Factor Name:* (default states *factor1*). Here, you have the option to tell SPSS what your within-subject factor is. That is, for the study above, because the measurements were made at pre-treatment, 6-month post-treatment, and at a one-year follow-up, your within-
subject factor is time. Or, stated another way, the passage of time separates the measurements. Type **Time** into the **Within-Subject Factor Name**: box.

b. The next box is labeled **Number of Levels**. This box is required, as you must tell SPSS how many repeated measures you have. Enter 3 (this is the number of times you measured your dependent variable—Pre, Six, and Year).

c. Click **Add** next to the upper-most large box. This places the within-subjects factor name and number of levels into the large box near to top. If you need to make edits to this or if you want to remove the name and levels altogether, simply click **Time(3)** and then click either the **Change** or **Remove** buttons.

d. Finally, you have the option to label your dependent variable in the **Measure Name**: box. Again, this is optional, but you can enter **Minutes** into this box and click **Add**. Similar to above, if you need to make edits to this or if you want to remove the name, simply click **Minutes** and then click either the **Change** or **Remove** buttons.

When you have completed these steps, click **Define**. This brings up the **Repeated Measures** dialog box shown in Figure 36.1.

![Repeated Measures Dialog Box](image)

**Figure 36.1**

2. In a mixed between-within ANOVA, you have one between-subjects independent variable and one repeated measures independent variable. Thus, you are going to need to specify both types of variables in this dialog box. In the **Within-Subjects variables**: box near the top you are going to place your repeated measures variables in chronological order. You will notice initially that this box has the same number of levels as the number you indicated in Step 1b above. If you followed the example correctly, you should have three slots available (\(_\_\_\_1,Minutes\), \(_\_\_\_2,Minutes\), and \(_\_\_\_3,Minutes\)).

   a. Click to select the **Pre** variable and then click the right arrow (\(\rightarrow\)) next to the **Within-Subjects variables**: box.
b. Click to select the **Six** variable and then click the right arrow (\(
\rightarrow\)) next to the *Within-Subjects variables:* box.

c. Click to select the **Year** variable and then click the right arrow (\(
\rightarrow\)) next to the *Within-Subjects variables:* box.

You also need to specify all between-subjects independent variables by placing them in the *Between-Subjects Factor(s):* box.

a. Click to select the **Tx** variable and then click the right arrow (\(
\rightarrow\)) next to the *Between-Subjects Factor(s):* box.

If you have done this correctly, your dialog box should look like Figure 36.1.

3. You will also notice six buttons on the right side of this dialog box. When conducting a mixed between-within ANOVA, all of these buttons could possibly be applicable. They are described below.

a. Clicking the **Model…** button allows you to specify the specific analysis you want to conduct. That is, including or excluding any main effects or interaction terms. The default option is **Full factorial** and leaving this option selected will instruct SPSS to include all main effects and interactions, including all between-within interaction, in the analysis. If you click the **Custom** option, four boxes become available—two for the *Within-Subjects:* and two for the *Between-Subjects:* portions of the model. Here, you can specify any main effects and interactions you want. Under the *Build Term(s)* section, clicking the down arrow (\(
\downarrow\)) will bring up all possible effects you could possibly include. So, to include a main effect, select *Main effects* from the dropdown menu, select one of your within- or between-subjects independent variables, and then click the right arrow (\(
\rightarrow\)) next to the associated box. You can follow the same steps to include an interaction; you just need to select all variables included in the interaction by holding the *Ctrl* button on your keyboard as you are selecting variables (this will highlight two or more variables). Note that while you cannot specify between-within interactions directly, they are included automatically. Finally, you can also specify the type of *Sum of squares:* When you are satisfied with your selections, click **Continue.**

b. Clicking the **Contrasts…** button allows you to test for differences among levels of the between-subjects or repeated measures independent variables you specified in Step 2 above. You can select from several different contrasts described below. Note the default is **Polynomial** for repeated measures variables and **None** for between-subjects variables.

   i. **Deviation:** Compares the mean of each level, other than a reference category, to the grand mean.

   ii. **Simple:** Compares the mean of each level to the mean of a specified level. This is useful when you have a control group.

   iii. **Difference:** Compares the mean of each level (except the first) to the mean of previous levels.

   iv. **Helmert:** Compares the mean of each level (except the last) to the mean of subsequent levels.
v. **Repeated**: Compares the mean of each level (except the last) to the mean of the subsequent level.

vi. **Polynomial**: Compares the linear effect, quadratic effect, cubic effect, and so on depending on how many levels you have. The first degree of freedom contains the linear effect across all categories; the second degree of freedom, the quadratic effect; and so on. These contrasts are used to perform a trend analysis.

If you choose to *Change Contrast*, click the down arrow ( ) in the *Contrast:* box, select the contrast you want from the drop-down menu, and click *Change.* For the *Deviation* and *Simple* contrasts, you will be asked what your *Reference Category* should be, you can choose either *Last* or *First.* Remember, this is based on your coding scheme, with *Last* indicating your highest coded value and *First* indicating your lowest coded value. When you are satisfied, click *Continue.*

c. Clicking the **Plots…** button allows you to create a line chart plotting the estimated marginal means across the repeated measurements individually for all levels of your between-subjects independent variable. For mixed between-within ANOVA, you will have at least one grouping variable and at least one repeated measures variables. Thus, you’ll always want to specify your repeated measures variables as the *Horizontal Axis:* and your grouping variable as *Separate Lines:.* If you have a second grouping variable, the *Separate Plots:* box is used to assist in decomposing the three-way interaction.

i. Click *Time* and then click the right arrow ( ) next to the *Horizontal Axis* box. This tells SPSS to place the repeated measures along the *x*-axis.

ii. Click *Tx* and click the right arrow ( ) next to the *Separate Lines:* box. This tells SPSS you want separate lines for the treatment and OTC groups.

iii. Click *Add* to move these instructions into the *Plots:* box.

When you are satisfied with your selections, click *Continue.*

d. Clicking the **Post Hoc…** button allows you to specify the post hoc tests you want to conduct only for a specific between-subjects independent variable(s). Note that this option is only used if you have three or more levels within your between-subjects independent variable(s). If you only have two levels, then the *F*-test for your between-subjects main effect would provide the exact same information as the post hoc test. That is, post hoc tests are unnecessary and SPSS acknowledges this fact by printing a warning message when you have less than three levels for any independent variable. To specify a post hoc test, select any or all between-subjects independent variables in the *Factor(s)* box, and then click the right arrow ( ) next to the *Post Hoc Tests for:* box. This will allow you to select post hoc tests under the *Equal Variances Assumed* and/or *Equal Variances Not Assumed* sections. Click the checkbox for the specific post hoc test(s) you want to conduct. When you are satisfied with your selections, click *Continue.*

e. Clicking the **Save…** button brings up the *Repeated Measures: Save* dialog box. Here, you have the option to save new variables to your dataset that contain
information about Predicted Values, Diagnostics, Residuals, or Coefficient Statistics. Most of these options allow you to test the assumptions of your model.

i. The options listed under the Predicted Values section allow you to save the values that your specific model predicts for each participant. Remember, any ANOVA analysis is based on the general linear model—a mathematical equation based on deviations from the overall mean and group means. You can save these values as Unstandardized or as Standard error. The Standard error option provides an estimate of the standard deviation of the average value of the dependent variable for participants that have the same values of the independent variables.

ii. The options listed under the Diagnostics section provide measures to identify participants with unusual combinations of values for the independent variables or those who have unusual influence on the outcome of the model. For both options, large values equal larger change or influence.

1. Cook’s distance is a measure of how much the residuals of all participants would change if the participant were removed from analysis.

2. Leverage values indicate the relative influence of each observation on the model’s fit.

iii. The options listed under the Residuals section provide some form of the difference between the observed value of the repeated measurements and the model predicted value. Remember, these residuals only represent within-subject variance.

1. Unstandardized: Provides the difference between an observed value and the value predicted by the model.

2. Standardized: Provides the residual divided by an estimate of its standard deviation. These are also known as Pearson residuals. Similar to z-scores, they have a mean of 0 and a standard deviation of 1.

3. Studentized: Provides the residual divided by an estimate of its standard deviation. However, the standard deviation is allowed to vary across participants depending on the distance of each participant's values on the independent variables from the means of the independent variables.

4. Deleted: Provides the residual for a participant if that participant is excluded from analysis. It is the difference between the value of the dependent variable and the adjusted predicted value.

iv. Finally, the options listed under the Coefficient Statistics section writes the variance-covariance matrix of parameter estimates, t-statistics, significance values, and residual degrees of freedom. You have the option to Create a new dataset where you must specify a Dataset name: or you
can Write a new data file which creates a complete new dataset. Selecting the latter requires you to click the File... button and specify the dataset you want to write onto.

When you are satisfied with your selections, click Continue.

f. Finally, by clicking the Options... button you have the option to print a table of Estimated Marginal Means as well a number of descriptive statistics and diagnostic tests.

i. Under the Estimated Marginal Means section, in the Factor(s) and Factor Interactions: box, you will always be offered the option to display (OVERALL) means. If you choose this option, SPSS will print the grand mean, which considers all the data in your sample together, irrespective of groups or repeated measurements. More importantly, however, this section allows you to display the estimated marginal means for your independent variables and their interaction.

1. Click Tx, and then click the right arrow ( ) next to the Display Means for: box.
2. Click Time, and then click the right arrow ( ) next to the Display Means for: box.
3. Click Tx*Time, and then click the right arrow ( ) next to the Display Means for: box.

ii. For any main effects and interactions you selected you can Compare main effects. Clicking this checkbox allows you to conduct post hoc tests (called pairwise comparisons) between all levels of your independent variables. It is this option that allows you to conduct post hoc tests across your repeated-measures. Here, you can select three adjustments which include LSD(none), Bonferroni, or Sidak. Adjustment is required to control for type I error due to multiple pairwise comparisons. Note that with only two levels per independent variable, this option is redundant, as the pairwise comparisons will produce identical results to the $F$-tests.

iii. Under the Display section, you have twelve options. The Descriptive statistics option prints means, standard deviations, and frequency counts for all levels of your independent variable(s). Estimates of effect size prints partial eta-squared, which is the proportion of total variability attributable to your independent variables and interaction(s). The Observed power option prints the power of the statistical test for each effect based on your specific sample. The Parameter estimates option prints parameter estimates, standard errors, t-tests, and confidence intervals (similar to regression). Further, you can display SSCP matrices and the Residual SSCP matrix. Clicking Transformation matrix produces the M matrix, whereas Contrast coefficient matrix prints the L matrix. Again, the L matrix is not useful without covariates. Homogeneity tests produces Levene’s test and the Spread-versus-level plot compares cell means, standard deviations, and variances across the level combinations of
all factors (groups). Choosing a *Residual plot* is useful for checking the normality assumption. Finally, the *Lack of fit* option is used to check if the relationship between the dependent variable and independent variables can be described adequately by the model and the *General estimable function* allows you to construct custom hypotheses. You can also specify the *Significance level* for your confidence intervals.

When you are satisfied with your selections, click **Continue**.

4. That’s it! All you need to do now is click **OK** to conduct the mixed between-within ANOVA.

### Obtaining Between-Person Residuals

To obtain the between-person residuals, you first need to collapse the data across the repeated measurements by calculating the mean across repeated measurements for each participant using the Compute procedure (Chapter 13). Then, use the general linear model procedure described in Chapter 28 to save the standardized residual values to your dataset. Finally, you will use the Explore procedure (Chapter 15) to evaluate the normality of these residuals.

For the example data above, I have provided an abbreviated list of steps below to obtain the standardized residual values. At this point in this text, you should be fairly skilled at using the Explore procedure to evaluate the normality of these residuals; thus, those steps are not provided here. For full details about all procedures used, please consult Chapters 13, 15, and 28 as needed.

1. Click **Transform** and then click **Compute Variable**…
   a. In the **Target Variable:** box, type **M_minutes** to represent mean minutes of sleep per night across the repeated measurements.
   b. In the **Numeric Expression:** box, type *(Pre + Six + Year)/3*. This will calculate the mean minutes of sleep per night for each participant.
   c. Click **OK**.
2. Click **Analyze**, then choose **General Linear Model**, and finally click **Univariate**…
   a. Click **M_minutes** and then click the right arrow (**→**) next to the **Dependent Variable:** box.
   b. Click **T** and then click the right arrow (**→**) next to the **Fixed Factor(s):** box.
   c. Click the **Save…** button.
      i. Under the **Residuals** section, click **Standardized**.
      ii. Click **Continue**.
   d. Click **OK**.

203
Output

When you click OK, SPSS will produce an Output screen displaying your results. Click your Output window to view your results (if it does not pop up automatically).

The first table is titled **Within-Subjects Factors** containing the variable names of your repeated measurements. The second table is titled **Between-Subjects Factors** containing the sample sizes \((N)\) of your groups within each level of our between-subjects independent variable. Conveniently, this table provides any **Value Labels** you provided in the **Values** column of Variable View to make group identification easy.

The next two tables are titled **Box’s Test of Equality of Covariance Matrices** and **Multivariate Tests**. The former tests a primary assumption of the multivariate approach to repeated measures and the latter provides the multivariate results of your analysis (i.e., repeated measures MANOVA). However, we are using ANOVA in this Chapter, so these tables are ignored.

The next table you see is titled **Mauchly’s Test of Sphericity** shown in Figure 36.2. This table contains information pertaining to fundamental assumption of repeated-measures ANOVA (i.e., multi-sample sphericity). Sphericity requires that the variances of the difference scores between measurements, across groups are relatively equal. **Mauchly’s \(W\)** is your test statistic and it follows a chi-square distribution, which is why you see the approximate (Approx.) **Chi-Square** column. The statistical test is based on the chi-square value, degrees of freedom \((df)\) and the \(p\)-value provided in the **Sig.** column. Remember, \(p < .05\) indicates a violation of the assumption, as the variances are not equal. If sphericity is violated, you will address this violation in the next table. Finally, you are provided with three **Epsilon** values. Epsilon values closer to 1 indicate smaller violations of sphericity.

![Mauchly's Test of Sphericity](image)

<table>
<thead>
<tr>
<th>Time</th>
<th>Mauchly's W</th>
<th>Approx. Chi-Square</th>
<th>df</th>
<th>Sig.</th>
<th>Epsilon</th>
<th>Greenhouse-Geisser</th>
<th>Huynh-Feldt</th>
<th>Lower-bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>.770</td>
<td>4.445</td>
<td>2</td>
<td>.100</td>
<td>.813</td>
<td>.932</td>
<td>.500</td>
<td></td>
</tr>
</tbody>
</table>

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b. Design: Intercept + Tx
   Within Subjects Design: Time

**Figure 36.2**

The results of your repeated measures main effect and between-within interaction are provided in the table titled **Tests of Within-Subjects Effects**, shown in Figure 36.3. Note that the between-within interaction is included in this table because it uses the repeated measures error term \((Error(Time))\) when calculating the \(F\)-statistic. This table also contains **Sum of Squares**, degrees of freedom \((df)\), **Mean Square**, omnibus \(F\), significance level \((Sig.)\) as well as the effect size (**Partial Eta Squared**). If Mauchly’s Test was not statistically significant, you will interpret the **Sphericity Assumed** row; if the sphericity assumption was violated, consult the **Huynh-Feldt** row. Note that research has shown that the **Greenhouse-Geisser** and **Lower-bound** options are
unnecessarily conservative—that is, their use has been shown to increase the probability of type II errors. In this example, sphericity was assured, so you can see from the *Sphericity Assumed* row that the interaction effect is statistically significant. This indicates that the treatment and OTC groups changed differently over the three repeated measurements.

### Tests of Within-Subjects Effects

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>Partial Eta Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>188922.533</td>
<td>2</td>
<td>9311.267</td>
<td>207.752</td>
<td>.000</td>
<td>.520</td>
</tr>
<tr>
<td>Greenhouse-Geisser</td>
<td>186922.533</td>
<td>1.626</td>
<td>11479.755</td>
<td>207.752</td>
<td>.000</td>
<td>.420</td>
</tr>
<tr>
<td>Horn-Fisher</td>
<td>186922.533</td>
<td>1.964</td>
<td>100127.937</td>
<td>207.752</td>
<td>.000</td>
<td>.420</td>
</tr>
<tr>
<td>Lower-bound</td>
<td>186922.533</td>
<td>1.900</td>
<td>196822.533</td>
<td>207.752</td>
<td>.000</td>
<td>.420</td>
</tr>
<tr>
<td>Time * Tx Time</td>
<td>31160.600</td>
<td>2</td>
<td>15950.468</td>
<td>42.504</td>
<td>.000</td>
<td>.702</td>
</tr>
<tr>
<td>Greenhouse-Geisser</td>
<td>31160.600</td>
<td>1.626</td>
<td>25462.968</td>
<td>42.504</td>
<td>.000</td>
<td>.702</td>
</tr>
<tr>
<td>Horn-Fisher</td>
<td>31160.600</td>
<td>1.964</td>
<td>26059.012</td>
<td>42.504</td>
<td>.000</td>
<td>.702</td>
</tr>
<tr>
<td>Lower-bound</td>
<td>31160.600</td>
<td>1.900</td>
<td>38188.800</td>
<td>42.504</td>
<td>.000</td>
<td>.702</td>
</tr>
<tr>
<td>Error(Time)</td>
<td>18169.333</td>
<td>86</td>
<td>449.148</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Greenhouse-Geisser</td>
<td>18169.333</td>
<td>29.267</td>
<td>552.481</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Horn-Fisher</td>
<td>18169.333</td>
<td>33.548</td>
<td>481.968</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower-bound</td>
<td>18169.333</td>
<td>36.908</td>
<td>895.265</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 36.3**

The next table is titled *Tests of Within-Subjects Contrasts* and is relevant if you decided to do a trend analysis. Here, you are informed whether each effect followed a *Linear* or *Quadratic* trend over time. For each trend, you are provided with your selected *Sum of Squares*, degrees of freedom (df), *Mean Square*, *F*, significance level (Sig.) and *Partial Eta Squared*. We are not conducting a trend analysis; thus, this table can also be ignored.

The next table is titled *Levene's Test of Equality of Error Variances* shown in Figure 36.4. This table is interpreted in the same way as it was for one-way between-subjects ANOVA (see Chapter 28), where a statistically significant *F* indicates a violation of this assumption. Note that homogeneity of variance is tested for each repeated measurement individually. A statistically significant result here suggests inference may be too liberal; thus, a more conservative alpha should be used.

### Levene's Test of Equality of Error Variances

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>df1</th>
<th>df2</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre</td>
<td>3.959</td>
<td>1</td>
<td>18</td>
<td>.061</td>
</tr>
<tr>
<td>Sx</td>
<td>3.347</td>
<td>1</td>
<td>18</td>
<td>.004</td>
</tr>
<tr>
<td>Year</td>
<td>5.91</td>
<td>1</td>
<td>18</td>
<td>.452</td>
</tr>
</tbody>
</table>

Tests the null hypothesis is that the error variance of the dependent variable is equal across groups.

* a. Design: Intercept + Tx
  Within Subjects Design: Time

**Figure 36.4**
Next, you are provided with the *Between-Subjects Effects* table shown in Figure 36.5. This table contains the *F*-tests, *p*-values (*Sig.*), and effect sizes (*Partial Eta Squared*) for your between-subjects independent variable(s). You are also provided with *Sums of Squares*, degrees of freedom (*df*), and *Mean Squares* for the associated effect(s).

![Figure 36.5](image)

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>Partial Eta Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>6861401.667</td>
<td>1</td>
<td>6861401.667</td>
<td>3241.244</td>
<td>.000</td>
<td>.994</td>
</tr>
<tr>
<td>Tx</td>
<td>28253.400</td>
<td>1</td>
<td>29253.400</td>
<td>13.347</td>
<td>.002</td>
<td>.426</td>
</tr>
<tr>
<td>Error</td>
<td>38104.267</td>
<td>18</td>
<td>2116.904</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The next section, titled *Estimated Marginal Means*, provides the unweighted Mean, standard error (*Std. Error*), and 95% *Confidence Interval* information for your between-subjects independent variable, repeated measures independent variable, and between-within interaction. Because the interaction was statistically significant, only the estimated marginal means for each group across the repeated measures (i.e., *Tx * Time*) is shown in Figure 36.6. We are not interested in the means collapsed across the repeated measures (i.e., *Tx* table) or the means collapsed across group (i.e., *Time* table) because minutes of sleep per night across the repeated measurements was dependent on the group the participant was randomized to.

![Figure 36.6](image)
Finally, a line chart titled *Estimated Marginal Means of Minutes* is presented. This chart is shown in Figure 36.7. This chart plots the estimated marginal means of each group separately across the repeated measurements. Here, your dependent variable (minutes of sleep per night) is on the y-axis, while your repeated measures are placed on the x-axis (labeled 1, 2, and 3 for Pre, Six, and Year, respectively). Separate lines are printed for the OTC and Treatment groups.

**Figure 36.7**

**Interpretation**

Interpretation begins by evaluating the tenability of the normality of residuals assumption. You will notice that in your dataset you have saved (or created) four new variables, \(ZRE_1\), \(ZRE_2\), and \(ZRE_3\) to represent within-subject variance for each repeated measurement individually, and \(ZRE_4\) to represent between-subject variance. Use the Explore procedure (Chapter 15) to evaluate the normality of the standardized residuals for all four measurements. The histograms for the repeated measurements are provided in Figure 36.8, whereas the histogram for the between-subjects residuals is provided in Figure 36.9 on the next page. You can see all distributions are relatively normal. This evaluation was further supported by the Q-Q plot and non-statistically significant Kolmogorov-Smirnov tests \((p > .05)\). Thus, the normality assumption is considered satisfied.

**Figure 36.8**
Next, you consult the Mauchly’s Test table to determine whether the sphericity assumption has been violated. For the data above, Mauchly’s test was not statistically significant; thus, sphericity is satisfied.

Then, you consult the Levene’s Test of Equality of Error Variances table to determine if the homogeneity of variance assumption has been violated. Remember, Levene’s test is calculated for each repeated measure. In the example, no violation of homogeneity of variance was indicated within any of the repeated measures.

With all assumptions satisfied, you can move on to the Tests of Within-Subjects Effects table. First, you will determine whether you have a statistically significant omnibus $F$ for the interaction effect. Remember, the specific row you consult is based on whether sphericity was violated. For this example, sphericity was not violated, so you will consult the Sphericity Assumed row. If your interaction effect is statistically significant, you do not interpret either main effects. If, however, your interaction is not statistically significant, you will report and interpret the repeated measures main effect from the Tests of Within-Subjects Effects table and the between-subjects main effect from the Tests of Between-Subjects Effects table. Please note, in the example above that the interaction effect is statistically significant; however, I do provide interpretations of the main effects below so you will know how to interpret them in the future if your results indicate a non-statistically significant interaction effect.

**Interpreting the Interaction Effect**

For the example above, the interaction effect was statistically significant, so you will ignore the main effects because the change over time is dependent on the specific group to which the participant was randomized. That is, minutes of sleep per night changed differently over the repeated measures depending on whether the participant was randomized to the new sleep treatment group or the group receiving the over-the-counter treatment. Without statistical tests, you can see from both the estimated marginal means and the profile plot (which present the same information) that there is little difference in the change from pretest to the 6-month follow-up, but a large change from 6-month to one-year follow-up. To break down this interaction effect, you need to conduct a simple main effects analysis. This process is described in detail below. Briefly, you will evaluate for differences between treatment groups at each measurement point individually.
Interpreting Main Effects

Although the interaction effect was statistically significant in the example above, for didactic purposes I will provide you with an interpretation of both the repeated-measures and between-subjects main effects.

Interpretation of the statistically significant repeated-measures main effect is very similar to one-way repeated-measures ANOVA (Chapter 34). From the example above, the repeated-measures main effect evaluates whether there is a change across time, for the entire sample on average, irrespective of treatment group. This analysis can essentially be viewed by averaging across groups for each repeated measurement. For both groups, you can clearly see the increase across time in Figure 36.7. If you have a statistically significant repeated-measures main effect, and the repeated-measures independent variable has only two levels, interpretation is completed by evaluating the estimated marginal means. If, however, you have three or more repeated measurements, as the example does, the main effect is an omnibus test. That is, the main effect indicates that there is a statistically significant change across the repeated measures on average, but does not indicate between which measurements the change occurred. Thus, post hoc simple comparisons are required to determine which measures differ from one another. These simple comparisons cannot be conducted appropriately using the point-and-click approach, so you must use a similar process to the simple main effects discussed in the next section. All you need to do is include the `COMPARE(*)` statement on the appropriate `/EMMEANS` line, where the asterisk represents name of your repeated-measures independent variable you provided in Step 1 above. So, the line 7 in Figure 36.10 would read: `/EMMEANS=TABLES(Time) COMPARE(Time) ADJ(BONFERRONI)`. This should make much more sense once you have read the next section; however, if you complete this analysis, you will see that there were statistically significant increases from Pre to Six (mean difference = 82.800), Six to Year (mean difference = 52.700) as well as Pre to Year (mean difference = 135.500).

Interpretation of between-subjects main effect is identical to factorial between-subjects ANOVA (Chapter 30). This main effect evaluates between-group differences by collapsing across the repeated measures. That is, the analysis considers the average minutes of sleep per night across the repeated measures for each group individually. In the example above, the between-subjects main effect was statistically significant indicating there was an overall difference in minutes of sleep per night between treatment groups. You can see this fact by comparing the results in Figure 36.5 to the results produced when you calculated the between-subjects residual. The $F$-value, $p$-value (Sig.), and effect size (Partial Eta Squared and $R$ Squared) are identical. Because the between-subjects independent variable had only two levels, interpretation would be completed by evaluating the estimated marginal means found in the 1. Tx table, which for this example indicated that the group receiving the new sleep treatment slept longer than the OTC group (359.867 vs. 316.467 minutes, respectively). If this variable has three or more levels, the main effect is an omnibus test. Thus, post hoc simple comparisons are required to determine which measures differ from one another. Similar to above, these simple comparisons cannot be conducted appropriately using the point-and-click approach, so you will need to use the syntax editor as described in the simple main effects section below. For the between-subjects main effect, all you need to do is include the `COMPARE(*)` statement on the appropriate `/EMMEANS` line, where the asterisk represents name of your repeated-measures independent variable you provided in Step 1 above. So, the line 6 in Figure 30.10 would read: `/EMMEANS=TABLES(Tx) COMPARE(Tx) ADJ(BONFERRONI)`. Again, this will make much more sense one you have read the next section.
Simple Main Effects Analysis of the Interaction Effect

You probably know by now that when you have three or more levels of any independent variable, any $F$-test involving that variable is omnibus. Further, the interaction effect is always an omnibus test. For a statistically significant interaction effect, simple main effects are required to determine at which measurement the groups differed. Because the example study was interested in whether the two groups changed differently over time, simple main effects of the between-subjects independent variable are most appropriate. That is, we will test for statistically significant group differences at each repeated measurement individually.

Simple main effects cannot be performed in SPSS using point-and-click methods because the correct error term and degrees of freedom will not be used, leading to incorrect inference. Thus, the Syntax Editor must be used. I discussed the Syntax Editor when discussing simple main effects for factorial between-subjects ANOVA in Chapter 30, and the procedure is similar and straightforward. To perform simple main effects using the Syntax Editor:

1. Follow the point-and-click method described above in Steps 1-3 of the main Analysis section. However, do not click Continue in Step 4, instead click Paste. A Syntax Editor window will pop up containing the syntax required to conduct the mixed between-within ANOVA, shown in Figure 36.10.

![Syntax Editor Screenshot](image)

Figure 36.10

2. Without describing the entire syntax, the primary line in the syntax you should consider is the 8th line of code in Figure 36.10. This line is presented below:

```
/EMMEANS=TABLES(Tx*Time)
```

This line of code was created when you requested estimated marginal means for the between-within interaction effect in Step 3f above. To evaluate differences in gender across the repeated measurements using a Bonferroni adjustment, type `COMPARE(Tx) ADJ(BONFERRONI)` at the end of the line of code presented above. Note that all three
adjustments for type I error described in Step 3f above are available for these comparisons (i.e., LSD and Sidak). When you do this, line 8 should now look like this:

\[ /EMMEANS=TABLES(Tx*Time)\] 
\[ COMPARE(Tx)\] 
\[ ADJ(BONFERRONI)\]

3. That’s it! In the Syntax Editor window, click Run and then click All.

**Interpretation of Simple Main Effects**

When you click All, SPSS will produce an Output screen displaying your results. Click your Output window to view your results (if it does not pop up automatically). You should notice that, for the most part, the output is identical to the output described above. The only difference is that under the Estimated Marginal Means section, you now have the two new tables under the 3. Tx * Time heading, titled Pairwise Comparisons and Univariate Tests. Both are shown in Figure 36.11.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Minutes</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tx</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>OTC</td>
<td>257.900</td>
</tr>
<tr>
<td></td>
<td>Std. Error</td>
<td>9.312</td>
</tr>
<tr>
<td></td>
<td>Lower Bound</td>
<td>248.335</td>
</tr>
<tr>
<td></td>
<td>Upper Bound</td>
<td>267.466</td>
</tr>
<tr>
<td>2</td>
<td>OTC</td>
<td>337.100</td>
</tr>
<tr>
<td></td>
<td>Std. Error</td>
<td>10.452</td>
</tr>
<tr>
<td></td>
<td>Lower Bound</td>
<td>315.141</td>
</tr>
<tr>
<td></td>
<td>Upper Bound</td>
<td>359.059</td>
</tr>
<tr>
<td>3</td>
<td>Treatment</td>
<td>344.400</td>
</tr>
<tr>
<td></td>
<td>Std. Error</td>
<td>10.274</td>
</tr>
<tr>
<td></td>
<td>Lower Bound</td>
<td>322.816</td>
</tr>
<tr>
<td></td>
<td>Upper Bound</td>
<td>365.984</td>
</tr>
<tr>
<td>Treatment 1</td>
<td>OTC</td>
<td>262.900</td>
</tr>
<tr>
<td></td>
<td>Std. Error</td>
<td>9.312</td>
</tr>
<tr>
<td></td>
<td>Lower Bound</td>
<td>243.335</td>
</tr>
<tr>
<td></td>
<td>Upper Bound</td>
<td>282.466</td>
</tr>
<tr>
<td>2</td>
<td>Treatment</td>
<td>359.300</td>
</tr>
<tr>
<td></td>
<td>Std. Error</td>
<td>10.452</td>
</tr>
<tr>
<td></td>
<td>Lower Bound</td>
<td>337.341</td>
</tr>
<tr>
<td></td>
<td>UpperBound</td>
<td>361.259</td>
</tr>
<tr>
<td>3</td>
<td>Treatment</td>
<td>457.400</td>
</tr>
<tr>
<td></td>
<td>Std. Error</td>
<td>10.274</td>
</tr>
<tr>
<td></td>
<td>Lower Bound</td>
<td>435.816</td>
</tr>
<tr>
<td></td>
<td>Upper Bound</td>
<td>478.984</td>
</tr>
</tbody>
</table>

**Pairwise Comparisons**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Minutes</th>
<th>95% Confidence Interval for Difference²</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>(I) Tx</td>
<td>(J) Tx</td>
</tr>
<tr>
<td>1</td>
<td>OTC</td>
<td>Treatment</td>
</tr>
<tr>
<td></td>
<td>Treatment</td>
<td>OTC</td>
</tr>
<tr>
<td></td>
<td>Treatment</td>
<td>OTC</td>
</tr>
<tr>
<td>3</td>
<td>OTC</td>
<td>Treatment</td>
</tr>
<tr>
<td></td>
<td>Treatment</td>
<td>OTC</td>
</tr>
</tbody>
</table>

*Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

* The mean difference is significant at the .05 level.
The term “pairwise comparisons” should immediately indicate that these are post hoc tests. Further, you can see in footnote a under the Pairwise Comparisons table that Adjustment for multiple comparisons: Bonferroni indicates the Bonferroni adjustment was applied as requested by the syntax. You should also notice that while it appears as though there are six statistical tests provided in the Pairwise Comparisons table, three of them are redundant. That is, for the Time 1 comparing OTC to Treatment in row 1 is the same as comparing Treatment to OTC in row 2.

Finally, if you compare the Pairwise Comparisons and Univariate Tests tables, you can see that they provide identical statistical results, just in different forms. This is specifically because our Tx independent variable had only two levels. Remember, for one degree of freedom between groups, $t^2 = F$. If we had three or more levels of an independent variable, the $F$-test in the Univariate Tests table would be an omnibus test for the simple main effect, and the Pairwise Comparisons table would provide the post hoc tests. Because these two tables provide identical information, we only consider the Pairwise Comparisons table for the statistical test and use the Univariate Tests table for effect size.

Using the Pairwise Comparisons table, evaluation begins by examining the Sig. column, which contains your $p$-values. Alternatively, you can quickly identify a statistically significant difference by looking for an asterisk next to the value in the Mean Difference (I-J) column. If a difference is statistically significant, you then look at the Mean Difference (I-J) column to evaluate which group had higher values. You also evaluate the standard error (Std. Error) and confidence interval (95% Confidence Interval for Difference). For this example, the two groups were statistically different only at the one-year follow-up (i.e., Time 3). The estimated marginal means indicated that the group receiving the new treatment slept more minutes per night compared to the group receiving the over-the-counter treatment.

Finally, you evaluate the effect size (Partial Eta Squared) in the Univariate Tests table, which for the statistically significant simple main effect is large (i.e., .667).

**Example Results Section**

Prior to analysis, no violation of independence was indicated and no outliers were identified. Further, no violation of normality of residuals, multi-sample sphericity or homogeneity of variance were indicated. Finally, equal time elapsed between repeated measurements. The results of a 2 (Group: Treatment vs. OTC) X 3 (Measure: pre-treatment, 6-months, one-year) mixed between-within ANOVA indicated a statistically significant between-within interaction effect, $F(2, 36) = 42.504, p < .05$, partial $\eta^2 = .702$, indicating minutes of sleep per night changed differently between treatment groups across the three repeated measurements. A simple effect analysis, using the Bonferroni adjustment to reduce the probability of type I errors, compared the Treatment and OTC groups on each repeated measure individually.

Results indicated a statistically significant difference at the one-year follow-up ($p < .05$; partial $\eta^2 = .771$), with participants receiving the new sleep treatment sleeping significantly longer than participants taking the OTC treatment (mean difference = 113.00, 95% CI = 82.475 to 143.525). No statistically significant group differences were identified at pre-treatment or at the 6-month follow-up.
Section VII

Statistical Tests for Categorical Variables

Categorical variables are ubiquitous in the medical sciences. There analysis is slightly different from the analyses discussed above. Primarily, we are moving away from the general linear model and into analyses that are all nonparametric. This section begins with the chi-square test, which you will see often in the literature. Next, two repeated-measures options are presented. We finish up with binary logistic regression, which uses the generalized linear model, but is interpreted very similarly to linear regression, with a few slight differences.

Similar to above, the Chapters in this section each include a small dataset for you to practice your data entry and coding skills. I chose to include a dataset for you to use rather than refer to abstract examples so that you will be able to perform the analyses and replicate all output presented. This will allow you to verify you completed the analysis correctly. Similar to above, while all portions of the menus for each analysis are described in detail, the bolded instructions are minimally required to replicate the output provided.

Finally, at the end of each Chapter you will be presented with an example results section in APA format. This should give you a pretty good idea of what will be minimally required when reporting results for your future posters or manuscripts.
Chapter 37
PEARSON’S CHI-SQUARE TEST

Pearson’s chi-square test (or simply, the chi-square test) is used when you have two
categorical variables. It is used to assess for significant differences between two or more
mutually exclusive groups for two variables measured on nominal, dichotomous, or categorical
scales. Further, the data may also be measured on an ordinal scale if the number of ranks is
small; however, this test does not consider rank order. The chi-square test assesses for
differences between observed and expected frequencies, where expected frequencies are defined
as the number of observations that would be expected if there actually were no differences.

It is important to note that a chi-square test based on a contingency table larger than 2 X 2
is an omnibus test. That is, when one or both variables have three or more levels, the chi-square
test will indicate whether a statistically significant difference exists, but will not indicate which
levels differ. In these situations, Bonferroni-adjusted 2 X 2 post hoc chi-square tests (or Fisher’s
exact tests if expected frequencies are low) are used to determine where statistically significant
differences occurred. Use the Select Cases procedure (Chapter 12) to select relevant groups.

Continuing with the sleep example, say you now want to know whether your new sleep
treatment helps people fall asleep within one hour of ingestion at a greater rate than the leading
over-the-counter treatment. Here, you are not interested in specifically how long it takes people
to fall asleep, just whether they fell asleep within one hour of taking their respective treatment.
Note that this example shows how you lose information when you collapse data into categories.
That is, you know whether the participant fell asleep within one hour or not, but you do not know
exactly how many minutes it took them to fall asleep. The collected data is presented below:

<table>
<thead>
<tr>
<th>ID</th>
<th>Group</th>
<th>Asleep</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Following the data entry procedures described in Chapters 1 and 2, label the first variable
ID, the second Group (0 = OTC; 1 = Treatment), and the third variable Asleep (0 = No; 1 = Yes).
Enter your data appropriately.

**Assumptions**

Assumptions of the chi-square test include independence and adequacy of expected
frequencies.

Independence requires that the categories of your independent variables be mutually
exclusive. It is technically a design issue. For the example data above, this assumption requires
that no participant in the treatment group can also be in the placebo group.
Adequacy of expected frequencies requires that the sample size be sufficiently large so that no cells have an observed frequency count of 0, and no more than 20% of cells have expected frequencies less than 5. This assumption is evaluated during analysis.

**Analysis**

Assuming independence, to conduct the chi-square test on the data above:

1. Click **Analyze**, then choose **Descriptive Statistics**, and then click **Crosstabs**… to bring up the **Crosstabs** dialog box shown in Figure 37.1. Note that if you do not have the Complex Samples or Exact Statistics add-on, the **Exact**... and **Bootstrap**… buttons will not be available to you.

![Figure 37.1](image)

2. For this analysis, you will be primarily concerned with the **Row(s):** and **Column(s):** boxes. The variables in these boxes construct the contingency table for the analysis. It does not matter which variable you place in the rows or columns boxes, the outcome will remain the same; however, generally the most important grouping variable is placed in the **Column(s):** box. Finally, if you have multiple variables you want to compare you can enter multiple variables into either box and separate analyses will be conducted.
   a. Click the **Group** variable, and then click the right arrow ( ) next to the **Column(s):** box.
   b. Click the **Asleep** variable, and then click the right arrow ( ) next to the **Row(s):** box.
3. Click the **Statistics**… button to bring up the **Crosstabs: Statistics** dialog box.
   a. Here, you are presented with 15 possible test statistics. Because this Chapter is dedicated to the chi-square test, this is the only option considered. However, it is worthwhile to note that the majority of these other options are forms of correlation (i.e., association) with their use determined by the specific levels of measurement of the variables included in the analysis.
   b. Click the **Chi-square** checkbox and then click **Continue**.
4. Click the **Cells...** button to bring up the **Crosstabs: Cell Display** dialog box. The options in this dialog box are incredibly useful for interpretation.
   
a. In the **Counts** section, you have the option to print the **Observed** and/or **Expected** frequencies. I suggest you select both because you need to know both the actual and expected number of participants within each cell for interpretation.
   
b. In the **Percentages** section, you can select percentages, for observed frequencies, that sum across the **Rows** or down **Columns**. Further, you can select the percentages of the **Total** number of participants.
   
c. In the **Residuals** section, you can select three options that provide you with a summary of the difference between the observed and expected values. **Standardized** and **Adjusted standardized** residual options divide the **Unstandardized** residual by the standard deviation and standard error, respectively.
   
d. If you selected a weight variable (where only cell counts are included in your dataset), observed values can take on fractional values. In the **Noninteger Weights** section, you can tell SPSS how you want to present any fractional values.

When you are satisfied with your selections, click **Continue**.

5. Clicking the **Format...** button allows you to determine how you want SPSS to present the **Row Order**, which can be either **Ascending** or **Descending**. This option can be used with both numeric or alphanumeric data. When you are satisfied, click **Continue**.

6. You also have the option to use **Layer** variables. If you select one (or more) layer variables, a separate chi-square test is produced for each category of each layer variable. For example, if you wanted to layer the sleep example across gender, separate analyses would be conducted for men and women. You can include more layer variables by clicking the **Next** button.

7. The **Display clustered bar charts** option will present clustered bars charts (see Chapter 16) for the variable categories you specified in the **Row(s):** box, with the variable you placed in the **Column(s):** box serving as the primary grouping variable.

8. Finally, if you know the contingency table you are outputting will be large—that is, you have a large number of categories for one or both variables, or you have a large number of layer variables—you can select the **Suppress tables** option and a portion of the table will be shown in your output. Note that you can double click the table to view it fully.

9. That’s it! Click **OK** to conduct the analysis.

**Output**

When you click **OK**, SPSS will produce an **Output** screen displaying your results. Click your **Output** window to view your results (if it does not pop up automatically). Here, you are presented with three tables.

The first table titled, **Case Processing Summary**, provides you with the number (**N**) and **Percent of Valid, Missing, and Total** participants. Note that the **Valid** column indicates how many participants were included in the analysis.
The second table provides the results of your *Crosstabulation*, shown in Figure 37.2. Note that the observed (*Count*) and expected (*Expected Count*) frequencies as well as the column percentages (% within Group) are provided in each cell.

<table>
<thead>
<tr>
<th>Asleep</th>
<th>Group</th>
<th>OTC</th>
<th>Treatment</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td></td>
<td>11</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>Yes</td>
<td></td>
<td>7.5</td>
<td>7.5</td>
<td>15.0</td>
</tr>
<tr>
<td>% within Group</td>
<td>73.3%</td>
<td>29.7%</td>
<td>50.0%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>15</td>
<td>15</td>
<td>30</td>
</tr>
<tr>
<td>Expected Count</td>
<td>7.5</td>
<td>7.5</td>
<td>15.0</td>
<td></td>
</tr>
<tr>
<td>% within Group</td>
<td>26.7%</td>
<td>73.3%</td>
<td>50.0%</td>
<td></td>
</tr>
</tbody>
</table>

*Figure 37.2*

Finally, the *Chi-Square Tests* table, shown in Figure 37.3, provides you with several test statistics (*Value*), their degrees of freedom (*df*), and p-values (*Asymp. Sig. (2-sided)*, *Exact Sig. (2-sided)*, and *Exact Sig. (1-sided)*). Notice in footnote a that SPSS calculates the percentage of cells that have expected frequencies less than 5. This provides all the information you need regarding the adequacy of expected frequencies assumption. Remember, if more than 20% have expected frequencies less than 5, the assumption has been violated. In a special case that you have a 2 X 2 contingency table and adequacy of expected frequencies assumption has been violated, you can interpret the results of *Fisher’s Exact Test* (this is why this test is often used as a post hoc test when breaking down larger contingency tables). *Fisher’s Exact Test* provides both 1- and 2-sided p-values and interpretation is similar to Pearson’s chi-square. The other two options you see in this table include Yates’ *Continuity Correction* and *Linear-by-Linear Association*. Yates’ *Continuity Correction*, applicable only for 2 X 2 tables, reduces the difference between the observed and expected frequencies by 0.5. Thus, the chi-square is reduced and the p-value is increased (i.e., statistical power is reduced). The *Linear-by-Linear Association* is only interpreted if both variables included in analysis are at least ordinal level. This value is the squared Pearson’s product-moment correlation multiplied by the number of participants minus 1 (i.e., $r^2(N-1)$).

<table>
<thead>
<tr>
<th>Chi-Square Tests</th>
<th>Value</th>
<th>df</th>
<th>Asymp. Sig (2-sided)</th>
<th>Exact Sig. (2-sided)</th>
<th>Exact Sig. (1-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Chi-Square</td>
<td>6.503</td>
<td>1</td>
<td>.011</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continuity Correction</td>
<td>4.639</td>
<td>1</td>
<td>.032</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>6.784</td>
<td>1</td>
<td>.009</td>
<td>.027</td>
<td>.013</td>
</tr>
<tr>
<td>Fisher’s Exact Test</td>
<td>6.316</td>
<td>1</td>
<td>.012</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear-by-Linear Association</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 37.3*

---
a. 8 cells (28%) have expected count less than 5. The minimum expected count is 7.50.
b. Computed only for a 2x2 table
**Interpretation**

Interpretation begins by evaluating the adequacy of expected frequencies assumption in footnote a of the Chi-Square Tests table. If less than 20% of cells have expected frequencies less than 5 and no cells have observed frequencies of 0, report the Pearson Chi-Square value with associated degrees of freedom and significance. If these assumptions are not met, and you have a 2 X 2 table, report the significance of Fisher’s Exact Test. In the example above, none of the four cells had expected counts less than 5, so the assumption is satisfied.

Regardless of whether you report the chi-square or Fisher’s exact test results, if your results are statistically significant, consult the Crosstabulation table to compare the expected frequencies to the observed frequencies. Indicate which cells had higher (or lower) observed frequencies than expected and report the observed frequency percentages. From the example, the expected count for people who fell asleep within one hour after ingesting your new sleep treatment was 7.5; however, the actual number was 11. Thus, more participants than expected fell asleep within 1 hour after taking your sleep treatment.

Finally, expected frequencies are rarely, if ever, reported in publication. Instead, the observed frequencies (Count) and associated percentages are reported. For this example, you will use the column percentages (% within Group) comparing those who fell asleep within one hour across groups (i.e., 26.7% OTC and 73.3% new treatment).

**Example Results Section**

Prior to analysis, no violation of independence was observed, and no cells had expected frequencies less than five or observed frequencies of zero.

The results of Pearson’s chi-square indicated a statistically significant difference between groups, $\chi^2(1) = 6.533$, $p < .05$, with participants receiving the new sleep treatment falling asleep within one-hour of dosing more often than those receiving the over-the-counter treatment (73.3% vs. 26.7%, respectively).
**Chapter 38**

**McNEMAR'S TEST OF CHANGE**

McNemar’s test of change (or simply, McNemar’s test) is an extension of the Pearson’s chi-square and Fisher’s exact test (Chapter 37) to situations where participants are measured repeatedly on two separate occasions. The test assesses the statistical significance of observed changes between two repeated measurements or two matched groups. Again, repeated measurement is more common. McNemar’s test is a bit limited in that it is only applicable to a dependent variable measured on a dichotomous scale; however, continuous variables can be artificially dichotomized, with the understanding that information will be lost.

A critically important caveat is that McNemar’s test only considers participants who changed across the measurements. Participants who did not change are not considered in analysis. Thus, if the data indicate change was a rare occurrence, McNemar’s test will probably be unable to detect a statistically significant effect due to the substantial decrease in statistical power resulting from the removal of participants who did not change from analysis.

As you probably guessed, a one-shot pretest-posttest design is one commonly analyzed by McNemar’s test if the outcomes are dichotomous. Continuing with the example from Chapter 37, say you now want to know whether your sleep treatment helps people fall asleep within one hour of ingestion. Here, you measure whether participants fall asleep within one hour at pretest. The next night, you give all participants your new sleep treatment and re-measure whether they fall asleep within one hour. The collected data is presented below:

<table>
<thead>
<tr>
<th>ID</th>
<th>Pre</th>
<th>Post</th>
<th>ID</th>
<th>Pre</th>
<th>Post</th>
<th>ID</th>
<th>Pre</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>21</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>1</td>
<td>22</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>1</td>
<td>23</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>14</td>
<td>0</td>
<td>1</td>
<td>24</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>15</td>
<td>1</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td>1</td>
<td>26</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>17</td>
<td>0</td>
<td>1</td>
<td>27</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1</td>
<td>18</td>
<td>1</td>
<td>1</td>
<td>28</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>1</td>
<td>19</td>
<td>0</td>
<td>1</td>
<td>29</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>30</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Following the data entry procedures described in Chapters 1 and 2, label the first variable ID, the second variable Pre (0 = No; 1 = Yes), and the third variable Post (0 = No; 1 = Yes). Enter your data appropriately.

**Assumptions**

The only assumption for McNemar’s test is that the categories for both variables must be mutually exclusive and exhaustive. That is, if you are in one category, you cannot be in any other category.
Analysis

1. Click **Analyze**, then choose **Nonparametric tests**, choose **Legacy Dialogs**, and finally click **2 Related Samples**.

2. You will notice immediately that this dialog box is identical to the one used for the Wilcoxon signed-rank test described in Chapter 33. Thus, you are required to identify variable pairs, which are indicated in the **Test Pairs** box.

   a. Click the **Pre** variable on the left hand side of the dialog box and then click the right arrow ( ) next to the **Test Pairs** box. This variable should now be listed under the **Variable1** column.

   b. Click the **Post** variable on the left hand side of the dialog box and then click the right arrow ( ) next to the **Test Pairs** box. This variable should now be listed under the **Variable2** column.

4. Next, you must indicate what the **Test Type** is. Here, you have four options, but we are only concerned with the **McNemar** checkbox. Check this checkbox and uncheck the **Wilcoxon** checkbox.

5. Finally, you can click the **Options** button, where you can select two descriptive statistics or identify how you want SPSS to handle your missing data.

   a. The **Descriptive** option displays the mean, standard deviation, minimum, maximum, and the number of non-missing values. The **Quartiles** option displays values corresponding to the 25th, 50th, and 75th percentiles.

   b. Under the **Missing Values** option, choosing to **Exclude cases test-by-test** evaluates each test separately for missing values. Choosing **Exclude cases listwise** excludes participants with missing values for any variable from all analyses.

   When you are satisfied with your selections, click **Continue**.

6. That’s it! Click **OK** to conduct McNemar’s test.

**Output and Interpretation**

When you click **OK**, SPSS will produce an **Output** screen displaying your results. Click your **Output** window to view your results (if it does not pop up automatically), which should be identical to Figure 38.1.

In the output, you are presented with two tables. The title of the first table changes based on the names of two variables you selected in Step 2 above. Here, it is titled **Pre & Post**. This table is a 2 X 2 contingency table providing information about those who changed and those who did not. Remember, participants who did not change across measurements are not considered in analysis. You can find the frequency counts for these participants by looking at the diagonal, which in this example include 10 participants (8 – No→No; 2 – Yes→Yes). Alternatively, you know that 20 participants did change (16 – No→Yes; 4 – Yes→No).

The second table is titled **Test Statistics**. Here, you are provided with the size of your sample (**N**) and your **p-value** (Exact Sig. (2-tailed)). If your **p-value** (i.e., Exact Sig. 2-tailed) is statistically significant, you will interpret the **Pre & Post** table to determine the direction of the change. Here, more individuals went from No to Yes than Yes to No.
McNemar Test

Crosstabs

<table>
<thead>
<tr>
<th>Pre &amp; Post</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Pre</td>
<td>8</td>
</tr>
<tr>
<td>No</td>
<td>4</td>
</tr>
<tr>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>

Test Statistics

<table>
<thead>
<tr>
<th></th>
<th>Pre &amp; Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>30</td>
</tr>
<tr>
<td>Exact Sig. (2-tailed)</td>
<td>.012a</td>
</tr>
</tbody>
</table>

a. Binomial distribution used.
b. McNemar Test

Figure 38.1

Example Results Section

The results of McNemar’s test of change indicated a statistically significant change from pretest to posttest ($p < .05$; $N = 30$), with 16 participants (53.3%) who had not previously fallen asleep within 1 hour at baseline doing so after taking the new sleep treatment.
Chapter 39
COCHRAN’S Q TEST

The Cochran Q test (or simply Cochran’s Q) is an extension of McNemar’s test to situations where participants are measured on a dichotomous outcome across three or more separate occasions or when you have three or more matched groups.

Because Cochran’s Q is used to assess change over three or more repeated measures, it is an omnibus test. That is, the test will determine whether a statistically significant change exists between the repeated measurements, but will not indicate where the change occurred. Thus, adjusted post hoc tests are required. These involve separate Bonferroni-adjusted McNemar’s tests comparing across each repeated measurement.

Continuing with the example from Chapter 38, say you now want to know whether your sleep treatment helps people fall asleep within one hour of ingestion long-term. Thus, you measure whether participants fall asleep within one hour at baseline. Then, you give all participants your new sleep treatment and re-measure whether they fall asleep within one hour. Finally, you measure whether they fall asleep within one hour at a six-month follow-up. The collected data is presented below:

<table>
<thead>
<tr>
<th></th>
<th>ID</th>
<th>Pre</th>
<th>Post</th>
<th>Six</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Following the data entry procedures described in Chapters 1 and 2, label the first variable ID, the second variable Pre (0 = No; 1 = Yes), the third variable Post (0 = No; 1 = Yes), and the final variable Six (0 = No; 1 = Yes). Enter your data appropriately.

Assumptions

The only assumption for Cochran’s Q is that the categories for all variables must be mutually exclusive and exhaustive. That is, if you are in one category, you cannot be in any other category.

Analysis

1. Click Analyze, then choose Nonparametric Tests, choose Legacy Dialogs, and finally click K Related Samples....

2. You will notice immediately that this dialog box is identical to the one used for the Friedman’s test described in Chapter 35. You will also notice that all of your variables in
your dataset are listed on the left hand side of this dialog box. The **Test Variables:** box is where you will identify the repeated measures on which you want to conduct your analysis.

a. Click to select the **Pre** variable and then click the right arrow (→) next to the **Test Variables:** box.

b. Click to select the **Post** variable and then click the right arrow (→) next to the **Test Variables:** box.

c. Click to select the **Six** variable and then click the right arrow (→) next to the **Test Variables:** box.

3. Next, you need to tell SPSS what the **Test Type** is. In this section, you have three options.

   Click the **Cochran’s Q** checkbox and uncheck the **Friedman** checkbox.

4. Finally, you can click the **Options…** button, where you can select two descriptive statistics or identify how you want SPSS to handle your missing data.

   a. The **Descriptive** option displays the mean, standard deviation, minimum, maximum, and the number of non-missing values.

   b. The **Quartiles** option displays values corresponding to the 25th, 50th, and 75th percentiles.

When you are satisfied with your selections, click **Continue**.

5. That’s it! Click **OK** to conduct the analysis.

**Output and Interpretation**

When you click **OK**, SPSS will produce an **Output** screen displaying your results, which should be identical to those presented in Figure 39.1. Click the **Output** window to view your results (if it does not pop up automatically).

The first table you see is titled **Frequencies**. Here, the frequency counts for your outcome across the repeated measurements are presented. Notice that the value labels you applied in **Variable View** are not included, so you will have to remember the coding scheme you used. The second table contains your **Test Statistics**. Here, you are presented with the total sample size (**N**), test statistic (**Cochran’s Q**), degrees of freedom (**df**), and **p-value** (**Asymp. Sig.**). It is noted that the **Cochran’s Q** test statistic follows a chi-square distribution and that the degrees of freedom is equal to the number of measurements minus 1.

Interpreting the results of your study is based on the information provided to you in **Test Statistics** table. Remember, Cochran’s Q is an omnibus test, so a statistically significant **Cochran’s Q** value indicates that a change occurred between at least two of the measurements, but you do not know specifically between which measurements. Thus, you need to conduct a series of adjusted post hoc McNemar’s tests. Again, the most common alpha adjustment used for nonparametric tests is the Bonferroni-adjustment. Here, you divide your experimentwise alpha (i.e., .05) by the number of post hoc tests you plan on conducting (i.e., 3). While I do not show you how to conduct the post hoc tests (Chapter 38), I present their results as well.
The results of Cochran’s Q indicated a statistically significant change across the three repeated measurements, $Q = 16.357$, $p < .05$. Bonferroni-adjusted post hoc McNemar’s tests indicated a statistically significant increase in the number of participants falling asleep within 1 hour from pretest to posttest (53.3%; $p < .017$) and from pretest to 6-month follow-up (66.6%; $p < .017$). Further, the treatment effect from pretest to posttest was maintained through six-month follow-up as no statistically significant change occurred between posttest and the six-month follow-up.
Chapter 40
SIMPLE BINARY LOGISTIC REGRESSION

Binary logistic regression analysis is used when you have one dichotomous (i.e., binary) dependent variable. The term, binary, implies that the dependent variable is coded 0/1. The analysis and interpretation is very similar to linear regression, with one important difference—the use of log-odds. Log-odds are variously indicated by \( \ln(p/1-p) \) or \( \ln(\mu/1-\mu) \), where the log portion is the natural log and the odds portion is calculated as the probability divided by 1 minus the probability (i.e., \( p/1-p \)).

As hinted at by the equation for log-odds above, binary logistic regression predicts the probability of being in one outcome category compared to the other. Note that the logistic regression in SPSS predicts the probability of being in the group coded 1. This is important for interpretation purposes. Using a linear regression analysis for a binary dependent variable is inappropriate because linear regression will fit a model with a constant, infinite slope. When your dependent variable is binary, the probability of being in one group or the other is bounded between 0 and 1. Logistic regression shuts off this slope at 0 and 1, so the results based on probability are nonlinear. Do not fear this nonlinear model, however. Binary logistic regression uses the logit link function to transform nonlinear probability into log-odds (aka, logits), which are linear and symmetric. Thus, after transformation, what are essentially linear regression techniques are applied to these log-odds, and interpretation is similar to linear regression but in log-odds units.

You should also note that you can exponentiate (i.e., \( e \) or \( \exp \)) the log-odds to obtain the odds. Odds, which are often presented and interpreted in the biomedical sciences and epidemiology, are asymmetric and range from 0 to infinity. Odds equaling 1 indicate an equal probability of being in either category of the dependent variable (i.e., probability of .5). Because log-odds and odds are dependent on one another, given one, you can easily calculate the other. That is, exponentiated log-odds equal odds, and the natural log of the odds equal log-odds.

Just like linear regression, logistic regression can be used to calculate the predicted log-odds using a regression equation known as the linear predictor. The algebraic simple logistic regression equation is: \( \ln(p/1-p) = a + BX \), where the log-odds \( \ln(p/1-p) \) are predicted by the intercept with the y-axis \( a \) and slope \( B \) as well as the value of the independent variable \( X \).

Notice there is no residual term (i.e., \( e \)) in this equation because the residual variance is dependent on probability. Thus, no assumptions are made regarding the variance of the residuals as in linear regression. However, residual values do exist in binary logistic regression. They are calculated as the difference between predicted probability of being in the category coded 1 and the participant’s actual, raw value of the dependent variable (i.e., 0 or 1). Thus, residuals can only take on two possible values for a given participant. That is, 0 minus predicted probability or 1 minus predicted probability. Therefore, the residuals in logistic regression cannot have a mean of zero, they cannot be normally distributed, and they cannot have constant variance (i.e., homoscedastic).

One final point to consider is that binary logistic regression can be viewed as an extension of Pearson’s chi-square to situations where you want to predict some outcome. Thus, any analysis that can be conducted using chi-square can also be conducted using logistic regression, with only interpretation changing. Logistic regression is more flexible that Pearson’s chi-square, however, because you can also include continuous independent variables. When we
discuss assumptions, you should notice immediately that they combine the assumptions of linear regression and Pearson’s chi-square.

Interpretation of regression slopes and odds ratios differ if your independent variable is categorical or continuous. As an example of a categorical independent variable, say you want to know whether participants who take your new sleep treatment have greater probability of falling asleep within one hour compared to participants who take the leading over the counter treatment. As an example of a continuous independent variable, say you want to know whether body weight in pounds significantly predicts whether participants fell asleep within one hour of going to bed. Each question will be analyzed separately. The collected data is presented below:

<table>
<thead>
<tr>
<th>ID</th>
<th>Group</th>
<th>Wt</th>
<th>Asleep</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>181</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>200</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>205</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>191</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>256</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>205</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>151</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>167</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>192</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>199</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>169</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>170</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>206</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>148</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>192</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>162</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>178</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>138</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>187</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>156</td>
<td>1</td>
</tr>
</tbody>
</table>

Following the data entry procedures described in Chapters 1 and 2, label the first variable \(ID\), the second variable \(Tx\) (0 = OTC; 1 = Treatment), the third variable \(Wt\) for body weight, and the final variable \(Asleep\) (0 = No; 1 = Yes). Enter your data appropriately.

**Assumptions**

There are three assumptions for simple logistic regression, but only two of them will be applicable at a time. Binary logistic regression is a between-subjects analysis; thus, the first assumption is independence of residuals and is always required. It may seem confusing that this assumption is included, as residual values were not included explicitly in the equation above. However, remember residuals can be calculated as described above. Similar to the analyses discussed previously, independence is technically a design issue and can be satisfied by data that is not clustered and by not measuring participants repeatedly.

The applicability of the two remaining assumptions varies according to the measurement scale of your independent variable. If your independent variable is categorical, the primary assumption is adequacy of expected frequencies, evaluated following the exact same steps as Pearson’s chi-square in Chapter 37. This assumption requires that no more than 20% of cells have expected frequencies less than 5 and no cells have observed frequencies of 0.

If your independent variable is continuous, the primary assumption is that the continuous independent variable is linear in the logit. Remember, logistic regression transforms probability to log-odds or logits, whether these log-odds are indeed linear is an empirical question. To evaluate the assumption, you need to include both the independent variable and the interaction between the independent variable and the natural log of the independent variable—a technique called the Box-Tidwell transformation. For example, if your continuous independent variable is body weight \((Wt)\), the interaction term would be \(Wt*\ln(Wt)\). So, in your model you would
include both $W_t$ and the $W_t \times \ln(W_t)$ interaction in the analysis. If the interaction is statistically significant, the assumption has been violated and regression parameters and standard errors become biased. You will need to use the Compute procedure (Chapter 13) to calculate the natural log of your continuous IV. Testing this assumption is described in detail below.

Finally, because binary logistic regression uses maximum likelihood estimation to estimate regression parameters, large sample sizes are required—at least 50 participants with a minimum of 10 events (i.e., event = dependent variable category coded 1). Note that the example above uses a much smaller sample than recommended.

**Categorical Independent Variable**

**Analysis**

Assuming independence of residuals and adequacy of expected frequencies are satisfied, to conduct a simple logistic regression on the data above using Asleep as the binary dependent variable and Tx as the categorical independent variable:

1. Click **Analyze**, then choose **Regression**, and finally, click **Binary Logistic…** to bring up the **Logistic Regression** dialog box, shown in Figure 40.1.

![Figure 40.1](image)

2. You will immediately see all of the variables in your dataset listed on the left hand side of the dialog box. You will be concerned primarily with the **Dependent:** and **Covariates:** boxes. The variable you place in either of these boxes is relatively self-explanatory. Remember, the only difference between a “covariate” and an “independent variable” is semantic.

   a. Because whether the participants fell asleep within one hour is your binary dependent variable, click to highlight **Asleep** and then click the right arrow (\( \rightarrow \)) next to the **Dependent:** box.

   b. Your independent variable is treatment group. Click to highlight **Group** and then click the right arrow (\( \rightarrow \)) next to the **Covariates:** box.
3. You will also notice three buttons on the right hand side of the *Logistic Regression* dialog box. Click the **Categorical…** button. Here, only the variables you entered in the **Covariates** box above are shown on the left hand side. You will use this option when you have a categorical independent variable or covariate with more than two levels. That is, you will not have to create new dummy variables as you did for linear regression. For this example, however, your independent variable is dichotomous, so contrast coding is not required. To use contrast coding, select an independent variable from the **Covariates** box on the left hand side and click the right arrow (→) next to the **Categorical Covariates** box. Then, select the **Contrast**: you want from the drop down list by clicking the down arrow (▼), and then click **Change**. A description of each **Contrast** option is described below. Note that for **Deviation**, **Simple**, and **Indicator** coding you will need to choose your **Reference Category**:, with **Last** indicating the highest coded category and **First** indicating the lowest coded category.

   a. **Indicator**: Contrasts the presence (i.e., 1) or absence (i.e., 0) of the independent variable to a reference category.

   b. **Simple**: Each category of the independent variable is compared to the unweighted average of all categories.

   c. **Difference**: Each category of the independent variable, except lowest coded category, is compared to the average effect of the previous categories. Also known as reverse Helmert contrasts, because…

   d. **Helmert**: Each category of the independent variable, except highest coded category, is compared to the average effect of the previous categories.

   e. **Repeated**: Each category of the independent variable, except lowest coded category, is compared to previous category (e.g., 2 to 1; 3 to 2, etc).

   f. **Polynomial**: Available only for continuous independent variables; uses orthogonal polynomial contrasts when categories are equally spaced. Think, trend-type analysis.

   g. **Deviation**: Each category of the independent variable, except the reference category, is compared to the unweighted overall effect.

   When you are satisfied with your selection, click **Continue**.

4. Click the **Save…** button. Here, you have the option to save many different values to your dataset. That is, any option you select in this dialog box will not print to your output, but instead a new variable will be created in your dataset.

   a. Under the **Predicted Values** section, you have the option to save the predicted **Probabilities** or **Group membership**. Because logistic regression is based on the probability of being in one outcome group or the other, the **Probabilities** option will create a new variable containing the predicted probability of being in the dependent variable category coded 1. Further, the **Group membership** option uses the predicted probabilities to indicate which group each participant should belong to. This is useful when reporting classification accuracy.
b. Under the *Residuals* section, you can save up to five different residual values. The *Unstandardized* residual is the difference between the observed value and the predicted probability in raw units. Note that the unstandardized residual can only take on two values. The *Logit* residual is the unstandardized residual divided by the predicted probability times 1 minus the predicted probability. A *Studentized* residual provides the change in the model deviance (i.e., likelihood ratio) if the participant is excluded. A *Standardized* residual is the unstandardized residual divided by the participant’s standard deviation. Finally, a *Deviance* residual is based on the model deviance, and is most appropriate for evaluating the distributional properties of the model when plotted against predicted values.

c. The options under the *Influence* section are used to measure the influence individual participants have on predicted probabilities. The *Cook’s* option indicates how much the unstandardized residuals of all participants would change if the participant was excluded from analysis. The *Leverage values* option measures the influence each participant has on model fit. Finally, the *DfBeta(s)* option indicates the difference in the regression coefficient (i.e., log-odds) if the participant was removed from analysis.

d. Finally, you can *Export model information to XML file*. Click *Browse...* to locate the already existing file. This option will save parameter estimates, and optionally the covariance matrix (clicking the checkbox is required) in XML format.

When you are satisfied with your selections, click *Continue*.

5. Click the **Options...** button.

a. The overwhelming majority of the information provided by logistic regression is going to be requested in the *Statistics and Plots* section. The *Classification plots* option prints a plot used to identify correctly classified participants. That is, the plot provides frequency counts of individual participants who were correctly classified to their observed outcome based on the predicted probabilities. It is more useful with you have a continuous independent variable or more than one independent variable or covariate measured on any scale. The *Hosmer-Lemeshow goodness-of-fit* option assesses how well your model fits the data. This statistic is calculated by grouping participants into categories (usually 10, but may be less for smaller samples) based on predicted probabilities and then compares the observed and expected frequencies across categories of the outcome variable. Sound familiar? It should, it is essentially a 2 X 10 chi-square. A non-statistically significant result (i.e., \( p > .05 \)) indicates good fit. The *Casewise listing of residuals* option will present a table identifying individual participants who are *Outliers outside* of a specified number of standard deviations (i.e., standardized residual) or for *All cases*. You can also print the *Correlations of estimates*, which will provide a correlation matrix between all independent variables and/or covariates. You can also print an *Iteration history*. Because logistic regression uses maximum likelihood estimation, an iterative procedure, you can evaluate the change in -2 log-likelihood as the estimation procedure converges on a solution. Finally, you should request the *CI for exp(B)*, which will print the confidence interval around your odds and odds ratios.
b. Under the Display section, you can choose to display results At each step or At last step. Because the only step in simple logistic regression is to compare the model with the independent variable to the model without the independent variable, both options will produce identical results.

c. You can indicate the Entry: and Removal: in the Probability for Stepwise section, but this option is not used in simple logistic regression.

d. Finally, you can indicate the Classification cutoff:, where 0.5 indicates equal probability of being in either dependent variable category. If you know (or hypothesize) that the actual probability of being the one group or another is not equal, you can change this value accordingly. You can also identify the number of Maximum Iterations for the maximum likelihood estimation. Finally, you can indicate whether you want to Include constant in model. Notice that this option is selected by default.

When you are satisfied with your selections, click Continue.

6. That it! Click OK to conduct the analysis.

Output

When you click OK, SPSS will produce an Output screen displaying your results. Click the Output window to view your results (if it does not pop up automatically). The first table you see is titled Case Processing Summary. Here, you are presented with the number (N) of participants Included in Analysis as well as the number of Missing Cases. This informs you whether your variables were coded correctly. The second table, titled Dependent Variable Encoding, indicates the coding scheme of your binary dependent variable. Remember, SPSS predicts the probability of being in the category coded 1.

The first large section, titled Block 0: Beginning Block contains information regarding your constant-only model. This model does not include your independent variable, and is for comparison purposes only. That is, this model is compared directly to the model with your independent variable (aka, full or saturated model). Thus, your full model should predict probability of being in one category or the other significantly better than the model with no independent variables. The Classification Table is presented in Figure 40.2. Much of the output in this section is redundant with what is discussed next, but you should notice that your independent variable, Group, is included in the Variables not in the Equation table.

![Classification Table](image)

**Figure 40.2**
The primary output you will be concerned with is found in the *Block 1: Method = Enter* section. The first table in this section, titled *Omnibus Tests of Model Coefficients*, is shown in Figure 40.3. This table contains the likelihood ratio test indicating whether your saturated model predicts the probability of being in category 1 of the dependent variable significantly better than the constant-only model. You can think of this test as similar to the model $F$-test for linear regression.

![Figure 40.3](image)

<table>
<thead>
<tr>
<th>Omnibus Tests of Model Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>Step 1</td>
</tr>
<tr>
<td>Block</td>
</tr>
<tr>
<td>Model</td>
</tr>
</tbody>
</table>

Figure 40.3

The next table is titled *Model Summary* and is shown in Figure 40.4. This table provides the $-2 \text{ Log likelihood}$ as well as two pseudo-$R^2$ measures—Cox & Snell and Nagelkerke—termed “pseudo” because they are approximations (underestimations) of $R^2$ for linear regression.

![Figure 40.4](image)

<table>
<thead>
<tr>
<th>Model Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

* Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Figure 40.4

In the *Classification Table*, shown in Figure 40.5, you can evaluate how well the model predicted both categories of your outcome variable as well as the *Overall Percentage* of correctly classified participants. Note that the overall percentage in this table should be substantially greater than the overall percentage resulting from the constant-only model, here 73.3% vs. 50.0% from the constant-only model above.

![Figure 40.5](image)

<table>
<thead>
<tr>
<th>Classification Table*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
</tr>
<tr>
<td>Predicted</td>
</tr>
<tr>
<td>Observed</td>
</tr>
<tr>
<td>Predicted</td>
</tr>
<tr>
<td>Percentage Corrected</td>
</tr>
<tr>
<td>------------------------</td>
</tr>
<tr>
<td>Step 1: Asleep</td>
</tr>
<tr>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
</tr>
<tr>
<td>Overall Percentage</td>
</tr>
</tbody>
</table>

*a. The cut-value is 0.003*

Figure 40.5
Finally, the Variables in the Equation table, shown in Figure 40.6, contains the results of your independent variable as well as the Constant. The $B$ column presents your regression coefficient in log-odds. Squaring $B$ and dividing by the squared standard error ($S.E.$) produces the Wald statistic, which is a 1 degree of freedom ($df$) test using a chi-square distribution. The statistical significance ($Sig.$), odds ratio ($Exp(B)$) as well as 95% confidence interval for the odds ratio (95% C.I. for EXP(B)) are also provided. Remember, odds ratios greater than 1 indicate an increase, odds ratio less than 1 indicate a decrease, and odds ratios equal to 1 indicate no effect.

<table>
<thead>
<tr>
<th>Variables in the Equation</th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>Sig.</th>
<th>Exp(B)</th>
<th>95% C.I. for EXP(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1 # Tx</td>
<td>2.023</td>
<td>.026</td>
<td>6.004</td>
<td>1</td>
<td>.014</td>
<td>7.562</td>
<td>1.499</td>
</tr>
<tr>
<td>Constant</td>
<td>-1.012</td>
<td>.094</td>
<td>3.002</td>
<td>1</td>
<td>.363</td>
<td>364</td>
<td>1.152</td>
</tr>
</tbody>
</table>

a. Variable(s) entered on step 1: Tx.

Figure 40.6

Interpretation

Interpretation begins by examining the omnibus test provided in the Omnibus Tests of Model Coefficients table in the Block 1: Method = Enter section. Because you have not used sequential or stepwise logistic regression, the Step, Block, and Model rows in this table will present identical information. If this result is statistically significant, which for this example it is, you can evaluate the results of the independent variable in the Variables in the Equation table. Note that because you have only one independent variable in this model, the omnibus test already informed you that your independent variable is statistically significant (assuming no assumption violations). Evaluate the regression coefficient ($B$) and standard error ($S.E.$) as well as the odds ratio ($Exp(B)$) and confidence interval (95% C.I. for EXP(B)). Large standard errors and wide confidence intervals indicate poor measurement. In addition, a 95% confidence interval that contains 1 indicates the variable is not statistically significant.

It is also noted that for dichotomous independent variables, the regression coefficient and odds ratio are calculated for the group coded 1 in reference to the group coded 0. Thus, the log-odds of 2.023 and odds ratio of 7.562 in Figure 40.6 were calculated for the group receiving the new sleep treatment because they were coded 1 in our data. Because log-odds are linear and symmetric around 0, the log-odds for the OTC group are simply -2.023. Further, to calculate the odds ratio for the reference (OTC) group, simply take the reciprocal of the $Exp(B)$. Thus, $1/7.562 = .132$. Alternatively, you can exponentiate the log-odds for OTC as $e^{-2.023} = e^{2.023} = .132$.

The interpretation of log-odds are identical to linear regression—a one-unit increase in the independent variable results in an increase (+) or decrease (-) in the log-odds of experiencing the category coded 1 on the dependent variable. Here, because the independent variable is dichotomous, a one-unit increase indicates going from the OTC group to the new sleep treatment group. Thus, being in the group receiving the new sleep treatment resulted in a 2.023 increase in the log-odds of falling asleep within one hour. Alternatively, being in the OTC group resulted in a 2.023 decrease in the log-odds of falling asleep within one hour.
When interpreting odds ratios, it is important to focus on the terminology. Some researchers like to state that odds ratios above 1 indicate an increased likelihood of experiencing the outcome (or event) of interest. That is, the dependent variable category coded 1. This is often incorrect because being more or less likely to experience an outcome is associated with risk, and odds only approximate risk if the incidence of the event is low (i.e., ≤ .1). Thus, from Figure 40.6, the odds ratio of 7.562 for the new sleep treatment indicates that those who received the new sleep treatment had a 7.562 (or 756.2%) increase in the odds of falling asleep within one hour compared to those who received the OTC group. Alternatively, those who received the OTC treatment had an 86.8% (i.e., [1-.132]*100) decrease in the odds of falling asleep within one hour compared to those who received the new sleep treatment.

Two additional points. First, note that odds ratios always compare two groups; thus, interpretation always needs to identify the two groups that were compared. This is why the italics were used as emphasis above. Second, note that odds are not symmetric—a 756.2% increase versus a 86.8% decrease. Odds can be tough to get your head around because they are not additive, they are multiplicative.

After you have evaluated the results of your independent variable, you will evaluate one of the pseudo-$R^2$ measures in the Model Summary table. Typically, you will want to report the Nagelkerke $R^2$ because it will always be the largest pseudo-$R^2$ measure provided by SPSS. However, the Cox & Snell $R^2$ can be reported as well. Finally, you will evaluate the results of the Classification Table to determine evaluate how well the model correctly classified participants. Obviously, you want this to be as close to 100% as possible.

**Example Results Section**

Prior to analysis, no violation of independence or adequacy of expected frequencies was identified.

The results of a simple logistic regression analysis indicated that the treatment group participants were randomized to significantly predicted whether they fell asleep within one hour, $\chi^2(1) = 6.794$, $p < .05$, Nagelkerke $R^2 = .270$. Participants receiving the new sleep treatment had a 2.023 increase in the log-odds of falling asleep within one hour or a 7.562 increase in the odds of falling asleep within one hour (95% CI = 1.499 to 38.152) compared to the over-the-counter treatment. Overall, the model correctly classified 73.3% of participants, a 23.3% increase over the constant-only model.

**Continuous Independent Variable**

**Evaluating the Linearity in the Logit Assumption**

To evaluate the assumption of linearity in the logit using the Box-Tidwell transformation you need to include the interaction between the independent variable and the natural log of the independent variable in your analysis. Use the Compute procedure, described in Chapter 13, to create the natural log of $Wt$ (i.e., $\ln(Wt)$). Label this new variable $lnWt$. To test linearity in the logit:

1. Click **Analyze**, then choose **Regression**, and finally, click **Binary Logistic**… to bring up the **Logistic Regression** dialog box, identical to Figure 40.1.
2. You will immediately see all of the variables in your dataset listed on the left hand side of the dialog box. Again, you will be concerned primarily with the **Dependent:** and **Covariates:** boxes.
   a. Because whether the participants fell asleep within one hour is your binary dependent variable, click to highlight **Asleep** and then click the right arrow (>) next to the **Dependent:** box.
   b. Your continuous independent variable is body weight (Wt). Click to highlight **Wt** and then click the right arrow (>) next to the **Covariates:** box.
   c. Next, to include the interaction term, click to highlight **Wt**. Then, press and hold the Ctrl button on your keyboard, and then click to highlight the **lnWt** variable. With both variables highlighted, click the >a*b< button (>) next to the **Covariates:** box. This will place the interaction term in this box, which should appear as **Wt*lnWt**.

3. That’s it! Click **OK**.

   In your output window, scroll down to the **Variables in the Equation** box under the **Block 1: Method = Enter** section. It should be the last table in your output. Look only at the statistical significance (**Sig.**) of the **Wt by lnWt** variable. Here, the variable is not statistically significant (**Sig. = p = .911**), which indicates that the body weight variable (Wt) is linear in the logit and the assumption is assured.

**Analysis**

Assuming independence of residuals and linearity in the logit are satisfied, to conduct a simple logistic regression on the data above using **Asleep** as the binary dependent variable and **Wt** as the continuous independent variable:

1. Click **Analyze**, then choose **Regression**, and finally, click **Binary Logistic...** to bring up the **Logistic Regression** dialog box, shown in Figure 40.1. Click **Reset** to clear what you had used to assess linearity in the logit.

2. You will immediately see all variables are listed on the left hand side of the dialog box. Again, you will be concerned primarily with the **Dependent:** and **Covariates:** boxes.
   a. Because whether the participants fell asleep within one hour is your binary dependent variable, click to highlight **Asleep** and then click the right arrow (>) next to the **Dependent:** box.
   b. Your continuous independent variable is body weight. Click to highlight **Wt** and then click the right arrow (>) next to the **Covariates:** box.

3. You will also notice three buttons on the right hand side of the **Logistic Regression** dialog box. Because your only independent variable is continuous, you do not need to be concerned with the **Categorical...** button. Further, if you click the **Save...** button all the options described in Step 4 of the Categorical Independent Variable section above still apply. Re-read that section for full description.
4. Click the **Options…** button. A full description of the options is described in Step 5 of the Categorical Independent Variable section above; however, you will select different options when you have a continuous independent variable.

   a. Click the **Hosmer-Lemeshow goodness-of-fit** option to assess how well your model fits the data. Remember, this statistic is calculated by grouping participants into categories (usually 10) based on predicted probabilities and then compares the observed and expected frequencies across categories of the outcome variable. Remember, a non-statistically significant result (i.e., \( p > .05 \)) indicates good model fit.

   b. Click the **CI for exp(\(B\))** to print the confidence interval around your odds and odds ratios.

   c. Click **Continue**.

5. That it! Click **OK** to conduct the analysis.

**Output**

When you click **OK**, SPSS will produce an **Output** screen displaying your results. Click the **Output** window to view your results (if it does not pop up automatically). The first table you see is titled **Case Processing Summary**. Here, you are presented with the number (\( N \)) of participants **Included in Analysis** as well as the number of **Missing Cases**. The second table, titled **Dependent Variable Encoding**, indicates the coding scheme of your binary dependent variable. Remember, SPSS predicts the probability of being in the category coded 1.

The first large section, titled **Block 0: Beginning Block** contains information regarding your constant-only model. This model does not include your independent variable, and is for comparison purposes only. That is, this model is compared directly to the model with your independent variable (aka, full or saturated model). Thus, your full model should predict probability of being in one category or the other significantly better than the model with no independent variables. If it does not, game over. The **Classification Table** is identical to the one presented in Figure 40.2. Again, much of the output in this section is redundant with what is presented next, but you should notice that your independent variable, \( W_t \), is included in the **Variables not in the Equation** table.

The primary output you will be concerned with is found in the **Block 1: Method = Enter** section. The first table in this section, titled **Omnibus Tests of Model Coefficients**, is shown in Figure 40.7. This table contains the likelihood ratio test indicating whether your saturated model predicts the probability of being in category 1 of the dependent variable significantly better than the constant-only model. You can think of this test as similar to the model \( F \)-test for linear regression.

![Omnibus Tests of Model Coefficients](image)

**Figure 40.7**
The next table is titled *Model Summary* and is shown in Figure 40.8. This table provides the -2 Log likelihood as well as two pseudo-\( R^2 \) measures—Cox & Snell and Nagelkerke—termed “pseudo” because they are approximations (underestimations) of \( R^2 \) for linear regression.

### Model Summary

<table>
<thead>
<tr>
<th>Step</th>
<th>-2 Log likelihood</th>
<th>Cox &amp; Snell R Square</th>
<th>Nagelkerke R Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>33.823*a</td>
<td>0.226</td>
<td>0.304</td>
</tr>
</tbody>
</table>

*a.* Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Figure 40.8

The overall fit of your model is provided as a statistical test in the table titled *Hosmer and Lemeshow Test*, shown at the top of Figure 40.9. Remember, this test is a chi-square analysis where a non-statistically significant result indicates good model fit as the observed frequencies were not statistically different from the expected frequencies. The contingency table the Hosmer-Lemeshow test was based on is provided in the *Contingency Table for Hosmer and Lemeshow Test* table, shown at the bottom of Figure 40.9. The two columns you should consider are titled *Asleep = No* and *Asleep = Yes*. When results are not statistically significant, you can see that the *Observed* and *Expected* values are similar. Note that for this example, this test was based on a 2 X 9 table.

### Hosmer and Lemeshow Test

<table>
<thead>
<tr>
<th>Step</th>
<th>Chi-square</th>
<th>df</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.240</td>
<td>7</td>
<td>.862</td>
</tr>
</tbody>
</table>

### Contingency Table for Hosmer and Lemeshow Test

<table>
<thead>
<tr>
<th></th>
<th>Asleep = No</th>
<th>Asleep = Yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed</td>
<td>Expected</td>
<td>Observed</td>
</tr>
<tr>
<td>Step 1</td>
<td>1</td>
<td>4</td>
<td>3.423</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2.961</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2.007</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1.735</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1.446</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1.236</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>.958</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>.702</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>.533</td>
<td>4</td>
</tr>
</tbody>
</table>

Figure 40.9
Next, the Classification Table is presented, shown in Figure 40.10. Here, you can evaluate how well the model predicted both categories of your outcome variable as well as the Overall Percentage of correctly classified participants. Note that the overall percentage in this table should be substantially greater than the overall percentage resulting from the constant-only model, here 66.7% vs. 50.0% from the constant-only model above.

<table>
<thead>
<tr>
<th>Classification Table a</th>
<th>Predicted</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>Asleep</td>
<td>Percentage Correct</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Step 1: Asleep No</td>
<td>10</td>
<td>66.7</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>Overall Percentage</td>
<td></td>
<td>66.7</td>
</tr>
</tbody>
</table>

Figure 40.10

Finally, the Variables in the Equation table, shown in Figure 40.11, contains the results of your independent variable as well as the Constant. The B column presents your regression coefficient in log-odds. Squaring B and dividing by the squared standard error (S.E.) produces the Wald statistic, which is a 1 degree of freedom (df) test using a chi-square distribution. The statistical significance (Sig.), odds ratio (Exp(B)) as well as 95% confidence interval for the odds ratio (95% C.I. for EXP(B)) are also provided. Remember, odds ratios greater than 1 indicate an increase, odds ratio less than 1 indicate a decrease, and odds ratios equal to 1 indicate no effect.

<table>
<thead>
<tr>
<th>Variables in the Equation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
</tr>
<tr>
<td>Step 1: VM</td>
<td>-0.45</td>
</tr>
<tr>
<td>Constant</td>
<td>2.734</td>
</tr>
</tbody>
</table>

Figure 40.11

Interpretation

Interpretation begins by examining the omnibus test provided in the Omnibus Tests of Model Coefficients table in the Block 1: Method = Enter section. Again, because you have not used sequential or stepwise logistic regression, the Step, Block, and Model rows in this table will present identical information. If this result is statistically significant, you can evaluate the results of the independent variable in the Variables in the Equation table. Note that because you have only one independent variable, the omnibus test already informed you that your independent
variable is statistically significant. Evaluate the regression coefficient \((B)\) and standard error \((S.E.)\) as well as the odds \((Exp(B))\) and confidence interval \((95\% \text{ C.I. for } EXP(B))\). Large standard errors and wide confidence intervals indicate poor measurement. In addition, a 95% confidence interval that contains 1 indicates the variable is not statistically significant.

For continuous independent variables, the regression coefficient and odds are calculated for a one-unit increase in the independent variable. However, it is important to note that log-odds are additive and linear, whereas odds are multiplicative and nonlinear. What this means to you is that, based on the results of Figure 40.11, the log-odds of -.045 indicates that for every one-pound increase in body weight, the log-odds of falling asleep within one hour decreased by .045. Because log-odds are additive, you can calculate the effect of a 2 pound increase or 10 pound increase directly (.09 and .45 decrease in log-odds, respectively). Note the similarity between interpretation of log-odds in logistic regression and slopes from a continuous independent variable in linear regression. They are identical.

Odds, on the other hand, are a different animal. Although they represent the change in odds for a one-unit increase in the independent variable, they are not additive. Thus, a two-unit change will not have exactly twice the effect of a one-unit change. However, you can quickly calculate the odds for any unit change by multiplying the log-odds by your desired increase and then exponentiating this value. For example, as calculated above, the log-odds for a two-pound increase were -.09 (i.e., -.045*2). Exponentiating this value \(i.e., e^{-0.09}\) gives odds of .914. A 10-pound increase gives odds of .638. This example shows how odds are not additive. That is, if they were, a 10-pound increase in body weight should have reduced the odds to .55 \(i.e., 1-.45\) instead of .638. Interpreting the odds from Figure 40.11 above, a one-pound increase in body weight resulted in a 4.4% \(i.e., [1-.956]*100\) decrease in the odds of falling asleep within one hour. Note that there is no comparison group for a continuous independent variable. Similarly, a 10-pound increase in body weight resulted in a 36.2% decrease in the odds of falling asleep within one hour. One final point I want to reiterate is that when you are interpreting odds, it is important to focus on the terminology. Remember, odds can only be interpreted as risk when the incidence of your outcome of interest \(i.e.,\) category coded 1 on your dependent variable) is low.

After you have evaluated the results of your independent variable, you will evaluate how well your model fit the data. This information is provided by the Hosmer and Lemeshow Test table. Remember, you want a non-statistically significant result \(i.e., p > .05\), which indicates your model fit the data well. Next, you will evaluate one of the pseudo-\(R^2\) measures in the Model Summary table. Again, you will typically report the larger Nagelkerke \(R^2\), but the Cox & Snell \(R^2\) can also be reported. Finally, you will evaluate the results of the Classification Table to determine evaluate how well the model correctly classified participants. Obviously, you want this to be as close to 100% as possible.

Example Results Section

Prior to analysis, no violation of independence of errors was observed and no violation of linearity in the logit was identified using the Box-Tidwell transformation.

The results of a simple logistic regression analysis indicated that body weight significantly predicted whether participants fell asleep within one hour, \(\chi^2(1) = 7.755, p < .05\), Nagelkerke \(R^2 = .304\). Model goodness-of-fit was excellent according to the Hosmer-Lemeshow test, \(\chi^2(7) = 3.240, p > .05\). Overall, the
model correctly classified 66.7% of participants, a 16.7% increase over the constant-only model. Results indicated a one-pound increase in body weight resulted in a .045 decrease in the log-odds of falling asleep within one hour or a 4.4% decrease in the odds of falling asleep within one hour (95% CI = 7.9% to 0.7% decrease). Further, a ten-pound increase in body weight produced a .45 decrease in log-odds or 36.2% decrease in the odds of falling asleep within one hour.
Chapter 41
MULTIPLE BINARY LOGISTIC REGRESSION

Multiple binary logistic regression is an extension of simple binary logistic regression to situations involving more than one independent variable. The extension is most easily seen through the binary logistic regression equation: \( \ln(p/(1-p)) = a + B_1X_1 + B_2X_2 + \cdots + B_kX_k \). Again, \( \ln(p/(1-p)) \) are the predicted log-odds, \( a \) is the \( y \)-intercept, the \( B \)s are the slopes of the log-odds for each independent variable, and the \( X \)s are the values of the independent variable. You can have as many independent variables as statistical power affords, measured on any scale.

One of the most common reasons for using multiple binary logistic regression is to control for nuisance variables called covariates. Covariates are variables that are correlated with the dependent variable, but are not of primary research interest. Thus, in many situations, you may have only one or two independent variables and several covariates. In reality, however, the difference between an independent variable and covariates is semantic. The analysis treats them all the same. It is up to you to determine which variables are interesting and which are nuisance.

Because multiple binary logistic regression is a direct extension of simple binary logistic regression, the analysis and interpretation should be straightforward assuming you have a solid understanding of simple binary logistic regression. Feel free to re-read Chapter 40 if necessary. To avoid redundancy in explanation, we will jump right into the example.

Say you want to know whether participants who take your new sleep treatment have greater probability of falling asleep within one hour compared to participants who take the leading over-the-counter treatment or placebo. Because the amount of sleep per night can be affected by numerous variables, you decide to control for caffeine consumption after 2pm (in milligrams), whether the participant smokes, body weight in pounds, and level of global anxiety as measured by the Hamilton Anxiety Rating Scale (HAM-A). The collected data is presented below:

<table>
<thead>
<tr>
<th>ID</th>
<th>Tx</th>
<th>Caffeine</th>
<th>Smoke</th>
<th>Wt</th>
<th>Anx</th>
<th>Asleep</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>215</td>
<td>1</td>
<td>181</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>478</td>
<td>1</td>
<td>225</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>269</td>
<td>0</td>
<td>172</td>
<td>19</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>246</td>
<td>1</td>
<td>191</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>178</td>
<td>1</td>
<td>256</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>99</td>
<td>0</td>
<td>205</td>
<td>35</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>360</td>
<td>1</td>
<td>151</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>421</td>
<td>1</td>
<td>167</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>105</td>
<td>1</td>
<td>192</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>149</td>
<td>0</td>
<td>199</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>198</td>
<td>1</td>
<td>184</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>145</td>
<td>1</td>
<td>170</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>269</td>
<td>1</td>
<td>236</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>10</td>
<td>0</td>
<td>135</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>215</td>
<td>1</td>
<td>192</td>
<td>19</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Tx</th>
<th>Caffeine</th>
<th>Smoke</th>
<th>Wt</th>
<th>Anx</th>
<th>Asleep</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>2</td>
<td>141</td>
<td>1</td>
<td>154</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>296</td>
<td>1</td>
<td>162</td>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>304</td>
<td>0</td>
<td>178</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>2</td>
<td>369</td>
<td>0</td>
<td>172</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>137</td>
<td>0</td>
<td>191</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>3</td>
<td>200</td>
<td>1</td>
<td>187</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>3</td>
<td>149</td>
<td>0</td>
<td>156</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>3</td>
<td>56</td>
<td>0</td>
<td>179</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>3</td>
<td>15</td>
<td>0</td>
<td>185</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>25</td>
<td>3</td>
<td>145</td>
<td>0</td>
<td>159</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>26</td>
<td>3</td>
<td>98</td>
<td>0</td>
<td>140</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>27</td>
<td>3</td>
<td>125</td>
<td>1</td>
<td>173</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>28</td>
<td>3</td>
<td>10</td>
<td>1</td>
<td>155</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>29</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>132</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>30</td>
<td>3</td>
<td>89</td>
<td>0</td>
<td>186</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

Following the data entry procedures described in Chapters 1 and 2, label the first variable \( ID \), the second variable \( Tx \) (1 = Placebo; 2 = OTC; 3 = Treatment) for treatment group, the third
variable *Caffeine*, the fourth variable *Smoke* (1 = Yes; 0 = No), the fifth variable *Wt* for body weight, the sixth variable *Anx* for anxiety, and the seventh variable *Asleep* (1 = Yes; 0 = No) to indicate whether the participants fell asleep within the first hour. Enter the data appropriately.

### Assumptions

The assumptions for multiple binary logistic regression can be viewed as a combination of simple binary logistic regression and multiple linear regression.

Multiple binary logistic regression is a between-subjects analysis; thus, the first assumption is independence of residuals. Remember, although residuals are not represented explicitly in the binary logistic regression equation, they can be calculated as described in Chapter 40. Independence is technically a design issue and can be satisfied by data that is not clustered and by not measuring participants repeatedly.

For categorical independent variables, the primary assumption is adequacy of expected frequencies, evaluated following the exact same steps as Pearson’s chi-square in Chapter 37. This assumption requires that no more than 20% of cells have expected frequencies less than 5 and no cells have observed frequencies of 0.

For continuous independent variables, the primary assumption is linearity in the logit. Remember, logistic regression transforms probability to log-odds or logits, whether these log-odds are linear is an empirical question. The explicit steps to evaluate this assumption are described in the Evaluating the Linearity in the Logit Assumption in Chapter 40. However, now you will include all continuous independent variables and their interactions, simultaneously, in one analysis. Remember, statistically significant interactions indicate a violation of this assumption.

Because you have multiple independent variables, absence of multicollinearity and absence of multivariate outliers are two additional assumptions required—both are evaluated using the linear regression procedure in SPSS (see Chapters 22 and 23). Multicollinearity is tested following the procedures described in Chapter 23, Step 3a. Note that it is ok to use a binary dependent variable to test this assumption in SPSS’s linear regression procedure because the tolerance and VIF values used to evaluate multicollinearity are independent of the scale of the dependent variable. For this example, the lowest tolerance value was .396 and the highest VIF was 2.526, which are at or below benchmarks.

Absence of multivariate outliers is assessed by Mahalanobis distance requested and saved through the linear regression procedure described in Chapter 23, Step 5c. Note that you will need to create two dummy variables for the *Tx* variable when using the linear regression procedure to evaluate multicollinearity. Follow the steps described in Dummy Variables section in Chapter 23. For this example, the largest Mahalanobis distance is 14.80329, which is well below the critical chi-square value of 22.458 using 6 degrees of freedom at $\alpha = .001$.

Finally, because binary logistic regression uses maximum likelihood estimation to estimate regression parameters, large sample sizes are required. Sample size benchmarks require at least 50 participants per independent variable and at least 10 events (i.e., outcome of interest; dependent variable category coded 1) per 50 participants. So, for example, if your study has five independent variables, you should have at minimum 250 participants and 50 participants with the outcome of interest. Note that the example above uses an extremely small sample size!
Assuming that independence of residuals, adequacy of expected frequencies, linearity in
the logit as well as absence of multicollinearity and multivariate outliers have been evaluated and
are satisfied, to conduct a multiple binary logistic regression on the example data above:

1. Click **Analyze**, then choose **Regression**, and finally, click **Binary Logistic…** to bring up
the **Logistic Regression** dialog box. Again, this dialog box is identical to the one shown in
Figure 40.1.

2. You will immediately see all of the variables in your dataset listed on the left hand side,
and you will be concerned primarily with the **Dependent:** and **Covariates:** boxes. The
variable you place in either of these boxes is relatively self-explanatory.
   a. Because whether the participants fell asleep within one hour is your binary
dependent variable, click to highlight **Asleep** and then click the right arrow (→) next to the **Dependent:** box.
   b. Your covariates include **Caffeine, Smoke**, body weight in pounds (**Wt**), and
anxiety level (**Anx**). One-by-one, click to highlight each variable and then click
the right arrow (→) next to the **Covariates:** box.
   c. Your independent variable of interest is treatment group. Click to highlight **Tx**
variable and then click the right arrow (→) next to the **Covariates:** box.

3. Click the **Categorical…** button. Here, only the variables you entered in the **Covariates:**
box above are shown on the left hand side. You will use this option when you have a
categorical independent variable or covariate with more than two levels. That is, you will
not have to create new dummy variables as you did for linear regression. Because
treatment group has three categories (i.e., Placebo, OTC, and new treatment), you need to
identify a reference category for comparison. Because you want to compare your new
sleep treatment to both placebo and the OTC treatment, you should use your new sleep
treatment group as the reference category. Remember, your new sleep treatment was
coded as category 3 on the **Tx** variable.
   a. Click to highlight the **Tx** variable and then click the right arrow (→) next to the
**Categorical Covariates:** box.
   b. Notice the **Change Contrast** section is now available. There are seven **Contrast:**
available from the drop down list. These include:
      i. **Indicator**: Contrasts the presence (i.e., 1) or absence (i.e., 0) of the
independent variable to a reference category.
      ii. **Simple**: Each category of the independent variable is compared to the
unweighted average of all categories.
      iii. **Difference**: Each category of the independent variable, except lowest
coded category, is compared to the average effect of the previous
categories. Also known as reverse Helmert contrasts, because…
      iv. **Helmert**: Each category of the independent variable, except highest coded
category, is compared to the average effect of the previous categories.
v. **Repeated**: Each category of the independent variable, except the lowest coded category, is compared to previous category (e.g., 2 to 1; 3 to 2, etc).

vi. **Polynomial**: Available only for continuous independent variables; uses orthogonal polynomial contrasts when categories are equally spaced.
Think, trend-type analysis.

vii. **Deviation**: Each category of the independent variable, except the reference category, is compared to the unweighted overall effect.

Note the default **Indicator** option is another term for dummy coding, the option we want to use for this example. Finally, for several of the options you will have to choose your **Reference Category**. Because your new sleep treatment was coded 3, it is the **Last** category, which is also the default. Click **Continue**.

4. Click the **Save**... button. Here, you have the option to save many different values to your dataset. That is, any option you select in this dialog box will not print to your output, but instead a new variable will be created in your dataset.

   a. Under the **Predicted Values** section, you have the option to save the predicted **Probabilities** or **Group membership**. Because logistic regression is based on the probability of being in one outcome group or the other, the **Probabilities** option will create a new variable containing the predicted probability of being in the outcome group coded 1. Further, the **Group membership** option, uses the predicted probabilities to indicate which group each participant belongs to, useful when reporting classification accuracy.

   b. Under the **Residuals** section, you can save up to five different residual values. The **Unstandardized** residual is the difference between the observed value and the predicted probability in raw units. Note that the unstandardized residual can only take on four values. The **Logit** residual is the unstandardized residual divided by the predicted probability times 1 minus the predicted probability. A **Studentized** residual provides the change in the model deviance (i.e., likelihood ratio) if the participant is excluded. A **Standardized** residual is the unstandardized residual divided by the participant’s standard deviation. Finally, a **Deviance** residual is based on the model deviance, and is most appropriate for evaluating the distributional properties of the model when plotted against predicted values.

   c. The options under the **Influence** section are used to measure the influence individual participants have on predicted probabilities. The **Cook’s** option indicates how much the unstandardized residuals of all participants would change if the participant was excluded from analysis. The **Leverage values** option measures the influence each participant has on model fit. Finally, the **DfBeta(s)** option indicates the difference in the regression coefficient (i.e., log-odds) if the participant was removed from analysis.

   d. Finally, you can **Export model information to XML file**. Click **Browse**... to locate the already existing file. This option will save parameter estimates, and optionally the covariance matrix (clicking the checkbox is required) in XML format.

When you are satisfied with your selections, click **Continue**.
5. Click the **Options**... button.

   a. The overwhelming majority of the information provided by logistic regression is going to be requested in the *Statistics and Plots* section. The *Classification plots* option is used to identify correctly classified participants. That is, the plot provides frequency counts of individual participants who were correctly classified to their observed outcome based on the predicted probabilities. The *Hosmer-Lemeshow goodness-of-fit* option assesses how well your model fits the data. This statistic is calculated by grouping participants into categories (usually 10, but may be smaller for smaller samples) based on predicted probabilities and then compares the observed and expected frequencies across categories of the outcome variable. It is essentially a 2 X 10 chi-square, where a non-statistically significant result (i.e., \( p > .05 \)) indicates good fit. The *Casewise listing of residuals* option will present a table identifying individual participants who are *Outliers outside* of a specified number of standard deviations (i.e., standardized residual) or for *All cases*. You can also print the *Correlations of estimates*, which will provide a correlation matrix between all independent variables and/or covariates. You can also print an *Iteration history* evaluating the change in -2 log-likelihood as the estimation procedure converges on a solution. Finally, you should request the *CI for exp(B)*, which will print the confidence interval around your odds ratios.

   b. Under the *Display* section, you can choose to display results *At each step* or *At last step*. Because the only step in simple logistic regression is to compare the model with the independent variable to the model without the independent variable, both options will produce identical results.

   c. You can indicate the *Entry*: and *Removal*: in the *Probability for Stepwise* section, but this option is not used in simple logistic regression.

   d. Finally, you can indicate the *Classification cutoff:*, where 0.5 indicates equal probability of being in either dependent variable category. If you know (or hypothesize) that the actual probability of being the one group or another is not equal, you can change this value accordingly. You can also identify the number of *Maximum Iterations* for the maximum likelihood estimation. Finally, you can indicate whether you want to *Include constant in model*. Notice that this option is selected by default.

   When you are satisfied with your selections, click **Continue**.

6. That it! Click **OK** to conduct the analysis.

**Output**

When you click **OK**, SPSS will produce an *Output* screen displaying your results. Click the *Output* window to view your results (if it does not pop up automatically). The first table you see is titled *Case Processing Summary*. Here, you are presented with the number (\( N \)) of participants *Included in Analysis* as well as the number of *Missing Cases*. This informs you whether your variables were coded correctly. The second table, titled *Dependent Variable Encoding*, indicates the coding scheme of your binary dependent variable. Remember, SPSS predicts the probability of being in the category coded 1.
Next, you are presented with a table titled *Categorical Variables Codings*, shown in Figure 41.1. This table provides a matrix of the coding scheme you specified for your categorical covariate in Step 3 above. Note that if you had multiple categorical covariates, you will have multiple tables. Because you selected dummy (i.e., Indicator) coding, this coding scheme should look familiar to you. That is, you have one variable representing the Placebo group and no other group (i.e., 1.000 under the (1) column) and another variable representing the OTC group and no other group (i.e., 1.000 under the (2) column), while the Treatment group is your reference group (i.e., .000 under both columns). You will appreciate the value of this table when you have to determine the coding scheme when reviewing the semi-ambiguous results of the analysis (see Table 41.5 below) or refreshing your memory after a lengthy period of time has passed (e.g., manuscript submission with reviewer comments).

<table>
<thead>
<tr>
<th>Categorical Variables Codings</th>
<th>Parameter coding</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency (1) (2)</td>
</tr>
<tr>
<td>Tx Placebo</td>
<td>10 1.000 .000</td>
</tr>
<tr>
<td>OTC</td>
<td>10 .000 1.000</td>
</tr>
<tr>
<td>Treatment</td>
<td>10 .000 .000</td>
</tr>
</tbody>
</table>

*Figure 41.1*

The first large section, titled *Block 0: Beginning Block* contains information regarding your constant-only model. This model does not include your independent variables, and is for comparison purposes only. That is, this constant-only model is compared directly to the model that includes all of your independent variable (aka, full or saturated model). Thus, your full model should predict the probability of being in one category or the other significantly better than the constant-only model with no independent variables. The *Classification Table* in this section indicates the percentage of correctly classified participants without considering any independent variables. The *Variables in the Equation* table contains only regression parameters for your *Constant*. This information is not useful, and the significance test for the *Constant* is not informative. Finally, you should notice that all of your independent variables are included in the *Variables not in the Equation* table.

The primary output you will be concerned with is found in the *Block 1: Method = Enter* section. The first table in this section, titled *Omnibus Tests of Model Coefficients* is shown in Figure 41.2. This table presents the results of the likelihood ratio test indicating whether your saturated model predicts the probability of being in category 1 of the dependent variable significantly better than the constant-only model. You can think of this test as similar to the model *F*-test for linear regression.

<table>
<thead>
<tr>
<th>Omnibus Tests of Model Coefficients</th>
<th>Chi-square</th>
<th>df</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1</td>
<td>23.856</td>
<td>6</td>
<td>.001</td>
</tr>
<tr>
<td>Block</td>
<td>23.856</td>
<td>6</td>
<td>.001</td>
</tr>
<tr>
<td>Model</td>
<td>23.856</td>
<td>6</td>
<td>.001</td>
</tr>
</tbody>
</table>

*Figure 41.2*
The Model Summary table, shown in Figure 41.3, provides the -2 Log likelihood as well as two pseudo-$R^2$ measures—Cox & Snell and Nagelkerke—termed “pseudo” because they are approximations (underestimations) of $R^2$ for linear regression. Again, higher values on these pseudo-$R^2$ measures are always better, and the value for this example are huge.

<table>
<thead>
<tr>
<th>Step</th>
<th>-2 Log likelihood</th>
<th>Cox &amp; Snell $R^2$</th>
<th>Nagelkerke $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16.515*</td>
<td>.549</td>
<td>.742</td>
</tr>
</tbody>
</table>

*a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

**Figure 41.3**

Next, you are presented with the results of the Hosmer and Lemeshow test. The overall fit of your model is provided as a statistical test in the table titled Hosmer and Lemeshow Test. Remember, this test is a chi-square analysis where a non-statistically significant result indicates good model fit as the observed frequencies were not statistically different from what was expected. The contingency table the Hosmer-Lemeshow test was based on is provided in the Contingency Table for Hosmer and Lemeshow Test table. The two columns you should consider are titled Asleep = No and Asleep = Yes. When results are not statistically significant, you can see that the Observed and Expected values are similar. Both tables are presented in Figure 41.4.

<table>
<thead>
<tr>
<th>Step</th>
<th>Chi-square</th>
<th>df</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15.149</td>
<td>8</td>
<td>.056</td>
</tr>
</tbody>
</table>

**Figure 41.4**
In the Classification Table, you can evaluate how well the model with all of your independent variables and covariates predicted both categories of your outcome variable as well as the Overall Percentage of correctly classified participants. Note that the overall percentage in this table should be substantially greater than the overall percentage resulting from the constant-only model, for this example: 60.0% vs. 90.0%.

Finally, the Variables in the Equation table, shown in Figure 41.5, contains the individual results for your independent variable and covariates as well as the Constant. The B column presents your regression coefficient in log-odds. Squaring B and dividing by the squared standard error (S.E.) produces the Wald statistic, which is a 1 degree of freedom (df) test using a chi-square distribution. The statistical significance (Sig.), odds ratio (Exp(B)) as well as 95% confidence interval for the odds ratio (95% C.I. for EXP(B)) are also provided. Notice that your dummy variables are labeled Tx(1) and Tx(2), for the placebo and OTC groups, respectively. If you did not remember how the groups were coded, you can always revisit the Categorical Variable Codings table shown in Figure 41.1. Note that the (1) and (2) following Tx are the column labels from Figure 41.1.

<table>
<thead>
<tr>
<th>Variables in the Equation</th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>Sig.</th>
<th>Exp(B)</th>
<th>95% C.I. for EXP(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Step 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caffeine</td>
<td>.001</td>
<td>.003</td>
<td>.015</td>
<td>1</td>
<td>.904</td>
<td>1.001</td>
<td>.999</td>
</tr>
<tr>
<td>Smoke</td>
<td>-2.139</td>
<td>1.501</td>
<td>2.143</td>
<td>1</td>
<td>.143</td>
<td>.111</td>
<td>.008</td>
</tr>
<tr>
<td>WH</td>
<td>-.030</td>
<td>.038</td>
<td>.867</td>
<td>1</td>
<td>.346</td>
<td>.965</td>
<td>.955</td>
</tr>
<tr>
<td>Anx</td>
<td>-.158</td>
<td>.173</td>
<td>.945</td>
<td>1</td>
<td>.331</td>
<td>.845</td>
<td>.602</td>
</tr>
<tr>
<td>Tx</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tx(1)</td>
<td>-3.052</td>
<td>1.929</td>
<td>3.069</td>
<td>1</td>
<td>.046</td>
<td>.021</td>
<td>.000</td>
</tr>
<tr>
<td>Tx(2)</td>
<td>-3.170</td>
<td>2.036</td>
<td>2.309</td>
<td>1</td>
<td>.129</td>
<td>.042</td>
<td>.001</td>
</tr>
<tr>
<td>Constant</td>
<td>11.329</td>
<td>6.050</td>
<td>2.660</td>
<td>1</td>
<td>.104</td>
<td>63076.858</td>
<td></td>
</tr>
</tbody>
</table>

Figure 41.5

Interpretation

Interpretation begins by examining the omnibus test provided in the Omnibus Tests of Model Coefficients table in the Block 1: Method = Enter section. Again, because you have not used sequential or stepwise logistic regression, the Step, Block, and Model rows in this table will present identical information. Because this result is statistically significant, at least one of your predictors is statistically significant, but you do not know which one(s). Thus, you evaluate the results for each independent variable or covariate in the Variables in the Equation table. Here, you evaluate the log-odds (B) and standard error (S.E.) as well as the odds (Exp(B)) and confidence interval (95% C.I. for EXP(B)). Large standard errors and wide confidence intervals indicate poor measurement. Because the purpose of the study was to evaluate the effectiveness of your new sleep treatment, you are only concerned with the results of the Tx variable. Remember, the remaining variables were covariates, or nuisance variables, that were only used to help
explain why the participant may not fall asleep within one hour. They were not of specific research interest. It is noted, however, that although each individual covariate significantly predicted the probability of falling asleep within one hour, none of the covariates remained statistically significant after controlling for each other. That is, no covariate explained a statistically significant amount of unique variance in the probability of falling asleep within one hour after removing shared variance.

Because the independent variable (Tx) used a dummy coding scheme, two categorical variables are used for direct comparison—Tx(1) compares the placebo group to the new sleep treatment group and Tx(2) compares the OTC group to the new sleep treatment group. The Tx(1) variable is statistically significant, but the Tx(2) variable is not. These results indicate that the group receiving the new sleep treatment differs from the placebo group, but not from the OTC group. Thus, we interpret the log-odds and odds ratio only for the Tx(1) variable. The log-odds value of -3.852 indicates that the placebo group (coded 1 on the dummy variable) had a 3.852 decrease in the log-odds of falling asleep within one-hour compared to the group receiving your new sleep treatment after statistically adjusting for caffeine intake, smoking, body weight, and anxiety level. Further, the odds ratio of .021 indicate that the placebo group had a 97.9% (i.e., [1-.021]*100) decrease in the odds of falling asleep within one hour compared to the group receiving your new sleep treatment after statistically adjusting for caffeine intake, smoking, body weight, and anxiety level. Alternatively, you could state that the group receiving your new sleep treatment had a 3.852 increase in log-odds or that the odds increased 4708.7% compared to the placebo group after statistically adjusting for caffeine intake, smoking, body weight, and anxiety level. With that said, notice the unreasonably large standard error (S.E.) and extremely wide confidence interval. These values indicate unacceptable precision of the true population odds ratio and are a direct result of using such a small sample size.

After you have evaluated the results of your independent variable, you will evaluate how well your model fit the data. This information is provided by the Hosmer and Lemeshow Test table. For this example, the model fits the data well as the test is not statistically significant (i.e., \( p > .05 \)). Next, you will evaluate one of the pseudo-\( R^2 \) measures in the Model Summary table. Again, you will typically report the larger Nagelkerke \( R^2 \), but the Cox & Snell \( R^2 \) can also be reported. Finally, you will evaluate the results of the Classification Table to determine evaluate how well the model correctly classified participants. Obviously, you want this to be as close to 100% as possible.

**Example Results Section**

Prior to analysis no violation of independence of residuals, adequacy of expected frequencies, or linearity in the logit were observed. Further, no multicollinearity was evident and no multivariate outliers were identified.

The results of a multiple binary logistic regression analysis indicated that the predictors, as a set, significantly predicted whether participants fell asleep within one hour, \( \chi^2(6) = 23.866, p < .05 \). Nagelkerke \( R^2 = .742 \). Model goodness-of-fit was assured according to the Hosmer-Lemeshow test, \( \chi^2(8) = 15.149, p > .05 \). Overall, the model correctly classified 90.0% of participants, a 30.0% increase over the constant-only model. Results indicated the group receiving the new sleep treatment had a 3.852 increase in the log-odds, or a 4708.7% increase in the odds, of falling asleep within one hour compared to placebo after statistically adjusting for caffeine intake, smoking status, body weight, and anxiety level.
Section VIII

Statistical Tests for Time-to-Event Data

Survival analysis (or failure analysis) consists of two of the most commonly used statistical techniques in the biomedical sciences—the Kaplan-Meier method and Cox regression. In general, survival analysis is concerned with time-to-event data; that is, the time to experience an outcome of interest. For example, consider a study designed to examine whether your new sleep treatment helps participants fall asleep faster compared to an effective OTC medication within one hour of ingestion. Notice that both the time required to fall asleep and whether the participants fell asleep is of specific importance.

In survival analysis, participants who do not experience the outcome of interest (which for survival analysis is called the event) are considered to survive, while those who experience the event are considered to fail. Although this is fairly grim terminology, survival and failure do not necessarily imply living or dying. That is, in the example above, failure was defined as falling asleep during the 60-minute study period. In survival analysis, the outcome will typically be dichotomous, patients do not need to enter the study at the same time (i.e., enrollment can be continuous), and patients are followed until the study ends, which is known as end of follow-up.

Patients who do not experience the event by the end of follow-up, who are lost to follow-up, or who drop out of the study are termed censored. The key advantage survival analysis has over the analyses presented above (particularly logistic regression) is it can handle censored (i.e., incomplete) data. That is, censored data are considered incomplete because the researcher does not know when these participants experienced the event. For example, someone who does not fall asleep within 60-minutes did not experience the event. Briefly, survival analysis can handle censored data because the dependent variable is time, which is a continuous variable allowed to vary for each participant. Thus, as long as a participant has a recorded survival time, they are included in analysis. The key point here is that survival time for a censored participant is the time until they were censored. Thus, all participants who entered the study are included in analysis regardless of whether they experienced the event or are censored because the analysis is concerned specifically with time in the study.

This section will cover two techniques for assessing time-to-event data—the Kaplan-Meier method and the Cox proportional hazards model (also known as Cox regression).
Chapter 42

THE KAPLAN-MEIER METHOD

Prior to discussion of the Kaplan-Meier method, a quick discussion of life tables are necessary. Life tables present time-to-event data in table format. That is, a life table tabulates the time that has elapsed until the event is experienced. Life tables are used to indicate the proportion of individuals surviving (i.e., not experiencing the event) based on fixed time intervals. For example, reconsider the example the time required to fall asleep following ingestion of your new sleep treatment or an OTC medication. A life table allows the researcher to tabulate the cumulative proportion of participants who fall asleep at any interval (e.g., 5-minutes, 10-minutes, 30-minutes). Life tables based on fixed time intervals have a significant weakness, as they do not consider the exact time within the specified interval when the participant experienced the event. Thus, depending on the length of the time interval, a great amount of information is lost regarding the exact time the event was experienced. That is, the longer the time interval, the less precise a researcher can be regarding the exact moment the event occurred.

The Kaplan-Meier method is an extension of the life table using varying time intervals. Using this method, the cumulative proportion surviving is recalculated every time an event occurs. For example, instead of assessing the total number of participants who fell asleep at fixed intervals of 5-minutes, the Kaplan-Meier method recalculates the proportion of participants not falling asleep every time a participant falls asleep. In most studies, Kaplan-Meier data is presented as a graph of the cumulative survival (or hazard) over the study period known as a Kaplan-Meier curve. The Kaplan-Meier curve presents either a survival or hazard functions, where the survival function is the cumulative frequency of participants not experiencing the event, while the hazard function is the cumulative frequency of participants experiencing the event. Note that the hazard function will be provided most often as this function is particularly important to Cox regression discussed in Chapters 43 through 45. A Kaplan-Meier curve is provided in the vast majority of published literature using survival analysis, and this curve is the primary focus of this chapter.

The Kaplan-Meier curve can be presented for single or multiple groups. From the example above, two Kaplan-Meier curves would be plotted on the same graph—one for your new sleep treatment and one for the OTC medication. With two or more treatment groups, a statistical test can be conducted to assess for a statistically significant group difference in survival rate. The most common test in the biomedical literature is the Log-Rank Test (also known as the Mantel-Cox test). A statistically significant log-rank test indicates there is a significant difference in survival (or hazard) rate between the groups. However, it is important to note the log-rank test has no way of indicating why one group of participants fell asleep quicker beyond the possibility of one of the treatments being ineffective.

Finally, when three or more treatment groups are being compared (e.g., placebo, OTC, treatment) the log-rank test is an omnibus test. That is, with three or more groups, a statistically significant log-rank test will indicate that a statistically significant difference exists between groups, but will not indicate which groups differ specifically. Thus, Bonferroni corrected post hoc log-rank tests are required to determine where a significant difference in survival occurred. Using the three group example above, three Bonferroni corrected log-rank tests are required using an adjusted alpha of 0.017(0.05/3). These adjusted post hoc tests determine whether statistically significant differences in survival rate occurred between your new sleep treatment
and placebo, your new sleep treatment and the OTC medication, and between placebo and the OTC medication.

Say you want to determine whether your new sleep treatment is more effective in helping people fall asleep faster than an effective OTC medication. Here, your dependent variable the time required to fall asleep. Participants were censored at 60 minutes. The collected data is presented below:

<table>
<thead>
<tr>
<th>ID</th>
<th>Tx</th>
<th>Asleep</th>
<th>Minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>53</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>36</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1</td>
<td>54</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
<td>26</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>1</td>
<td>59</td>
</tr>
</tbody>
</table>

Following the data entry procedures described in Chapters 1 and 2, label the first variable ID, the second variable Tx (1 = Treatment; 0 = OTC) for treatment group, the third variable Asleep (1 = Yes; 0 = No), and the fourth variable Minutes to indicate the number of minutes the participant was in the study. Enter the data appropriately.

Assumptions

The assumptions the Kaplan-Meier method include the requirement of no differences between withdrawn and remaining participants, constant probability of survival over the entire study period, and that the event occurred at the time specified. The majority of these assumptions can be properly assured by maintaining experimental rigor throughout the study period.

No differences between withdrawn and remaining participants require that participants who are censored (or lost to follow-up) do not differ systematically from those who experience the event. For example, in the data above, a participant who was censored prior to study completion (e.g., ID = 1) may have been censored due to an adverse reaction to the medication. This is not an easily testable assumption because censoring can occur for a variety of reasons. However, binary logistic regression (Chapters 40 or 41) can be used to evaluate for differences between withdrawn and remaining participants. That is, those who withdrew from the study prior to completion are coded 1 with all remaining participants coded 0. This new binary variable serves as the dependent variable in the logistic regression analysis when assessing for statistically significant differences in the independent variable or covariates. Failure to achieve statistical significance satisfies the assumption. However, remember logistic regression requires large samples, so evaluation of this assumption may not be tenable in all situations.

Having constant probability of survival over the study period requires (1) that the same factors affecting survival are present throughout the entire study period and (2) participants entering the study at different time points have equal probability of survival. For example, a violation of this assumption would occur if the researcher forgot to turn off the lights after dosing
the participant or the OTC medication was discontinued requiring substitution with a close, but not identical, medication. This assumption is evaluated by both examining the study design and evaluating the protocol for violations. In addition, and more appropriate for longer-term studies, additional novel, effective, or ancillary treatments may become available during the study period that require ethical implementation. In these situations, the assumption is tested by examining the survival functions within subsets of participants (i.e., before vs. after differing treatment modalities).

Finally, assuring that the event occurred at the time specified can be difficult, especially for left- or interval-censored participants (e.g., participant fell asleep 10-15 minutes after dosing, but the researcher is not exactly sure when). A violation of this assumption tends to bias survival estimates upwards.

**Analysis**

Assuming no differences between withdrawn and remaining participants, constant probability of survival over the entire study period, and that the event occurred at the time specified, to conduct a Kaplan-Meier analysis using the log-rank test:

1. Click **Analyze**, then choose **Survival**, and finally, click **Kaplan-Meier**… to bring up the Kaplan-Meier dialog box, shown in Figure 42.1.

![Kaplan-Meier dialog box](image)

**Figure 42.1**

2. You will immediately see all of the variables in your dataset listed on the left hand side of the dialog box. You will be concerned primarily with three boxes—**Time**, **Status**, and **Factor** boxes. The **Time** box is fairly self-explanatory. Here, you will place the dependent variable that measures time-to-event. The **Status** box is essentially the censoring variable. That is, the variable that identifies who experienced the event or not. Finally, the **Factor** box is where the categorical independent variable will be placed. In addition, the optional **Strata** and **Label Cases by** boxes are available. The **Strata** box is used when you have an additional categorical variable for which you want to use to conduct Kaplan-Meier analyses across. For example, you could evaluate the treatment differences across men and women. Finally, the **Label Cases by** box will use the unique
identifier variable to indicate which patient experienced the event, or was censored, directly on the printed Kaplan-Meier curve.

a. Click to highlight the **Minutes** variable, and then click the right arrow (►) next to the **Time:** box.

b. Click to highlight the **Asleep** variable, and then click the right arrow (►) next to the **Status:** box.

c. Click to highlight the **Tx** variable, and then click the right arrow (►) next to the **Factor:** box.

3. You should notice immediately that there is a question mark in parentheses after the **Asleep** variable in the **Status:** box (see Figure 42.1). This is because you need to tell SPSS which numeric value the event was assigned. Click the **Define Event...** button to bring up the **Kaplan-Meier: Define Event for Status Variable** dialog box shown in Figure 42.2. In this dialog box you have three options.

a. First, you can specify a **Single Value**: Although you may have multiple values coded for various events, the use of this box allows you to specify one value as the event and all other values as non-events.

   i. For the example data above, type **1** in this box.

b. Next, you can specify a **Range of values**: This option is useful if you have multiple values coded for various events, and you are interested in two or more events. However, note that the events must be coded in numerical order. That is, you cannot have events of interest coded 1, 2, and 4. Using this option would include the event coded 3, which would not be of interest to you.

c. Finally, you can specify a **List of values**: This option can be used if you have any number of events, coded in any numerical order. Thus, if you are interested in events coded 1, 2, and 4, you can specify each event individually by clicking the radio button, typing the event number of interest and clicking **Add**.

d. When you have specified the event(s) of interest, click **Continue**.

---

![Figure 42.2](image-url)
4. Click the **Compare Factor**… button to bring up the *Kaplan-Meier: Compare Factor Levels* dialog box.

   a. In the *Test Statistics* section, you have three options. Each option is used to test statistically for differences in survival functions across levels of your independent variable (and within strata, if you so choose). The **Log rank** test weights all time points equally. The **Breslow** test gives slightly greater weight to earlier events as time points are weighted by the number of participants at risk (i.e., not censored) at each time point. Finally, the **Tarone-Ware** test gives greater weight to earlier events as time points are weighted by the square root of the number of participants at risk at each time point.

   b. Next, you must decide how you want to compare levels of your independent variable (or strata variable, if using).

      i. The **Linear trend for factor levels** option if self-explanatory and allows you to test for linear trend across levels of your *Factor*: variable. This option is only available for omnibus tests (i.e., *Pooled over strata* or *For each stratum*) and only makes sense if the *Factor*: variable has natural ordering (i.e., ordinal).

      ii. If you did not specify a stratification (i.e., *Strata:* ) variable, you have only two options—*Pooled over strata* or *Pairwise over strata*. If the *Factor:* variable has only two levels, selecting either option will produce identical results. However, with three or more levels, the *Pooled over strata* option is an omnibus test identifying that at least two groups have different survival functions but will not indicate which groups differ specifically. Thus, selecting *Pairwise over strata* will provide pairwise comparisons comparing all groups to each other. Note that the pairwise comparisons are unadjusted. Thus, the Bonferroni adjustment must be employed to reduce the probability of type I error.

      iii. The remaining two options are only available if you specified a stratification variable. Selecting *For each stratum* performs a separate statistical test at each level of the strata variable. Similar to above, this will be an omnibus test for any stratum having three or more *Factor:* levels. Thus, selecting *Pairwise for each stratum* allows you to conduct unadjusted post hoc tests within each stratum.

   c. When you have made your selections, click **Continue**.

5. Click the **Save** button to bring up the *Kaplan-Meier: Save New Variables* dialog box. Selecting any of these options will create a new variable in your dataset that can be used to test hypotheses or evaluate certain assumptions. All of the options except for *Hazard* are provided for you in the Survival Table printed in the output.

   a. The **Survival** option provides the cumulative survival probability estimate for each participant.

   b. The **Standard error of survival** option provides the standard error of the cumulative survival estimate.
c. The *Hazard* option provides the cumulative hazard estimate.

d. The *Cumulative events* option provides the cumulative frequency of events when participants are sorted by their survival time and status coding.

e. When you are satisfied with your selections, click *Continue*.

6. Finally, click the **Options** button to bring up the *Kaplan-Meier: Options* dialog box. Here, you can select the actuarial *Statistics* you want printed as well as any specific *Plots*.

a. Under the *Statistics* section, you can select **Survival table(s)**, which will print the time until each event occurred in ascending order, which might not be the order presented in your dataset. This table will contain much of the information you could have requested in using the *Save* options described in Step 5 above, but in a more concise format. Further, you can select **Mean and median survival**, which will print the mean and median survival time for each level of the *Factor* variable as well as standard errors and 95% confidence intervals. Finally, you can request **Quartiles** to print a table displaying average survival time at the 25th, 50th (i.e., median), and 75th percentile along with the associated standard errors.

b. Under the *Plots* section, you can select four separate plots to examine the survival function of your data for each level of your *Factor* variable individually. Selecting **Survival** prints the cumulative survival function on a linear scale (i.e., the actual value of your dependent time variable). The function will always be decreasing as fewer participants are expected to remain in the risk pool as time passes. Selecting the **One minus survival** option, will transpose the survival function above, so that as time passes the function increases. In general, this is a difficult function to interpret and is not used often. Selecting the **Hazard** option prints the cumulative hazard function on the original linear time scale. This function will always be increasing as the risk of experiencing the event will necessarily increase as time passes. Finally, you can choose to print the **Log Survival** function, which is similar to the survival function above, except the natural log of cumulative survival is used instead.

c. When you are satisfied with your selections, click *Continue*.

7. That’s it! Click **OK** to conduct the analysis.

**Output**

When you click **OK**, SPSS will produce an *Output* screen displaying your results. Click the *Output* window to view your results (if it does not pop up automatically). The first table you see is titled *Case Processing Summary*. Here, you are presented with the number of participants (*Total N*) within each level of the independent variable (i.e., the *Factor:* variable) as well as for your entire sample (*Overall*). Further, you are presented with the number of events within each group (*N of Events*) as well as the number of participants (*N*) and percentage (*Percent*) that were *Censored*.

Next, you are presented with the *Survival Table*, shown in Figure 42.3. You will notice immediately that the table is sectioned by each level of the *Factor:* variable, which in this case was treatment group, where *OTC* is in the upper half, and *Treatment* is in the lower half. The
Time column displays the time at which the event or censoring occurred, and the Status column indicates whether the case actually experienced the event or was censored. The Cumulative Proportion Surviving at the Time section includes the proportion of participants surviving from the start of the study until the specific Time. Note that if multiple participants experience the event at the same time, the Estimate and standard error (Std. Error) are only printed once during the time period. Finally, the number of participants that experienced the event from the start of the study is found in the N of Cumulative Events column, and the number of participants at the specific time point that have yet to experience the event are found in the N of Remaining Cases column. Notice that censored cases are not considered to have experienced the event. For example, participant 1 in the dataset coincides with the third row (3) in the OTC group. You can see that this participant was censored (i.e., Status = No; that is, they did not experience the event) at 35 minutes, and does not contribute to the cumulative number of events while being removed from the total number of remaining participants.

<table>
<thead>
<tr>
<th>Tx</th>
<th>Time</th>
<th>Status</th>
<th>Cumulative Proportion Surviving at the Time</th>
<th>N of Cumulative Events</th>
<th>N of Remaining Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>OTC</td>
<td>1</td>
<td>16:00h</td>
<td>Yes</td>
<td>0.933</td>
<td>0.084</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>28:00h</td>
<td>Yes</td>
<td>0.987</td>
<td>0.089</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>35:00h</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>38:00h</td>
<td>Yes</td>
<td>0.784</td>
<td>0.108</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>48:00h</td>
<td>Yes</td>
<td>0.722</td>
<td>0.119</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>51:00h</td>
<td>Yes</td>
<td>0.650</td>
<td>0.127</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>53:00h</td>
<td>Yes</td>
<td>0.578</td>
<td>0.132</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>54:00h</td>
<td>Yes</td>
<td>0.506</td>
<td>0.134</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>58:00h</td>
<td>Yes</td>
<td>0.433</td>
<td>0.133</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>60:00h</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>60:00h</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>60:00h</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>60:00h</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>60:00h</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>60:00h</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>1</td>
<td>10:00h</td>
<td>Yes</td>
<td>0.933</td>
<td>0.084</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>12:00h</td>
<td>Yes</td>
<td>0.987</td>
<td>0.089</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>15:00h</td>
<td>Yes</td>
<td>0.810</td>
<td>0.103</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>16:00h</td>
<td>Yes</td>
<td>0.733</td>
<td>0.114</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>17:00h</td>
<td>Yes</td>
<td>0.657</td>
<td>0.122</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>20:00h</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>22:00h</td>
<td>Yes</td>
<td>0.553</td>
<td>0.129</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>28:00h</td>
<td>Yes</td>
<td>0.410</td>
<td>0.133</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>35:00h</td>
<td>Yes</td>
<td>0.444</td>
<td>0.133</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>38:00h</td>
<td>Yes</td>
<td>0.370</td>
<td>0.129</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>41:00h</td>
<td>Yes</td>
<td>0.285</td>
<td>0.120</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>50:00h</td>
<td>Yes</td>
<td>0.222</td>
<td>0.112</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>60:00h</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>60:00h</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>60:00h</td>
<td>No</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 42.3
Next, you are provided with the Means and Medians for Survival Time table, shown in Figure 42.4. Here, you are provided with the Mean and Median survival time as well as associated standard errors and 95% confidence intervals for each group individually. Note that both the Mean and Median survival time include participants who were censored before the end of follow-up. The mean is an estimated quantity, but can be approximated by hand by summing all survival times and dividing by the total number of participants within the group. The median is the survival time where the probability of survival is 0.50 or 50%. This can be estimated visually using drawing a horizontal reference line at 0.50 on the survival function in Figure 42.6 below. Further, and as with any statistical test discussed in the above Chapters, large overlap in the confidence intervals indicates little difference in survival time between groups.

The Overall Comparisons table is presented next, as shown in Figure 42.5. Here, you are presented with the results of the log-rank test (Log Rank(Mantel-Cox)). This test follows a chi-square distribution (Chi-Square) with degrees of freedom (df) equal to the number of groups (or Factor: levels) minus 1. The statistical significance of this test is indicated in the Sig. column.

Finally, you are presented with the Survival and Hazard Functions, presented in Figure 42.6, you requested in Step 6b above. Both functions present individual lines for each level of your Factor: variable. Notice that the x-axis displays the time to event in the original time units. The Survival Function displays, at any time point, the probability that a participant on a specific treatment will not have experienced the event. The Hazard Function displays, at any time point, the probability that a participant will experience the event. There are two important characteristics to notice about both functions. First, they are both step functions. That is, the function steps down in survival or up in hazard each time a participant experiences the event. For example, consider participant 16 in your dataset (i.e., row 1 of the
**Treatment** group in Figure 42.3). This participant experienced the event at 10 minutes; thus, the green line steps down in survival or up in hazard at 10 minutes. The magnitude of the step down or step up is determined by the sample size, with smaller steps associated with larger samples and vice versa. Further, the horizontal distance between steps is determined by the length of time between events. Second, note that participants who were censored prior to end of follow-up are marked with a vertical dash through their group’s respective survival or hazard function and not a step in the functions. For example, participant 1 in your dataset (i.e., row 3 of OTC group in Figure 42.3) was censored at 35 minutes; thus, there is a vertical line through both the blue survival and hazard functions at 35 minutes, but no associated step down or up, respectively.

**Interpretation**

Interpretation begins by evaluating the *Overall Comparisons* table where you will find the results of the log-rank test. For this example, there is a statistically significant difference in survival (i.e., no falling sleep within 60 minutes) between the two treatment groups. Next, you will evaluate and report the *Median* estimates (and associated standard errors and confidence
intervals) from the *Means and Medians for Survival Table*. I recommend you report the median because the Kaplan-Meier method (and associated log-rank test) is a nonparametric method and the most appropriate indicator of central tendency for rank ordered data is the median. Note that if cumulative survival never reaches 50%, the median will not be calculated, so report the mean. Further, you should also report the number of participants that experienced the event as well as the number of participants that were censored. This information is provided to you in the *Case Processing Summary* table. Finally, you should always present either the survival or hazard function (aka, Kaplan-Meier curve) in your results. The choice between the two depends on your research question and the information you are attempting to portray.

**Example Results Section**

Prior to analysis, two participants were censored prior to end of follow-up for unknown reasons; however, no differences were identified between participants who were censored and those who fell asleep during follow-up. Further, no changes to the risk pool or protocol violations were indicated.

The results of the log-rank test indicated a statistically significant difference between the two groups, $\chi^2_1 = 4.029, p < .05$, with participants receiving the new sleep treatment falling asleep at in less time and at a higher rate ($Mdn = 35$ minutes, $SE = 11.628$, 95% CI = 12.210 to 57.790; 73.3% falling asleep) than participants receiving the OTC medication ($Mdn = 59$ minutes, $SE = 5.507$, 95% CI = 48.207 to 69.793; 53.3% falling asleep). Survival and hazard functions are presented in Figure 42.6.
Chapter 43
SIMPLE COX REGRESSION WITHOUT A TIME-VARYING COVARIATE

Although the Kaplan-Meier method is effective in assessing for overall differences in survival, it is unable to identify the association between covariates and survival. That is, the Kaplan-Meier method cannot identify whether an independent variable significantly predicts survival, which for many studies is a far more important consideration. Thus, a form of regression analysis is required.

The Cox proportional-hazards model (or simply, Cox regression) is a semiparametric technique (discussed below) used to predict the time to experience some categorical event (e.g., fell asleep vs. did not fall asleep; event coded 1). As stated in Chapter 42, Cox regression models the hazard function. It should also be noted that most published research progresses from Kaplan-Meier curves and log-rank tests to Cox regression analysis. That is, the Kaplan-Meier curve will first present the hazard (and/or survival) functions for the independent variable of interest as well as associated log-rank test(s), and then Cox regression will be conducted assessing for the relationship between the independent variable and the event of interest after adjusting for important covariates. Note that just like logistic regression you must identify the primary event of interest and the associated reference category prior to analysis.

The results of a Cox regression analysis are considered time dependent. That is, Cox regression is concerned with the time-dependent risk of experiencing an event. Thus, Cox regression can be thought of as an extension of relative risk (i.e., #exposed / #unexposed at a given time point) and incidence rates. Further, a Cox model has a particular advantage of using more information for each participant (i.e., survival time and censoring) than logistic regression. In reality, however, either Cox or logistic regression can be used to evaluate the same data, with the distinction lying in your particular research question. That is, if you are interested in time-to-event, use Cox regression; if you are interested in the overall occurrence of events, use logistic regression.

As stated above, a Cox model predicts the hazard function, which necessarily increases with the passage of time and, thus, ranges between 0 and infinity. Stated more simply, the hazard function is dependent on time. In addition, because simple Cox regression includes one independent variable, the hazard function is also conditional on the participant’s value of this variable. With these two components in mind, the algebraic simple Cox regression equation is: \( h(t|X) = h_0(t)(\exp(B_1X_1)) \). While this equation may appear intimidating, it contains similar information to what you already know from Chapters 40 and 41. Here, \( h(t|X) \), is the predicted hazard at a specific time point \( t \) for a given value of the covariate \( X \), \( h_0(t) \) is the baseline hazard function (i.e., hazard function at time \( t \) when the covariate is 0), and \( \exp(B_1X_1) \) is the regression component with the exponent required to keep the hazard function positive. It is important to note that there are two different forms of Cox regression—with or without time-varying covariates. The use of Cox regression with time-varying covariates (i.e., covariates that fluctuate over time) is based on a violation of the proportionality of hazards assumption, which will be discussed in detail below. Similar to logistic regression, there is no residual term in the Cox regression equation; thus, no assumptions are made about the residuals as in linear regression aside from independence. However, residuals do exist in Cox regression and they are used for diagnostic purposes as discussed below; specifically, Martingale and deviance residuals.

As you may have guessed from the use of exponentiation in the equation above, the hazard function is nonlinear. Cox regression uses the log link function to linearize the
relationship between the hazard function, regression parameter and covariate in the equation above. Thus, if you find the natural log of both sides of the equation: \( \text{ln}(h(t|X)) = h_0(t) + B_1X_1 \), where the left side of the equation is termed the log-hazard (i.e., risk) and the right side is a linear function of the baseline hazard and independent variable. In general, the log link function used here differs from the logit link in logistic regression in name alone, but note that the regression parameters produced from Cox regression are termed **log-hazards**. Exponentiating the log-hazards produces a **hazard ratio**, a task very similar to moving from log-odds to odds ratios in logistic regression. The hazard ratio produced by a Cox regression analysis is calculated and interpreted similarly, but not identically, to the relative risk. Briefly, hazard ratios range from 0 to infinity, with 1 indicating equal risk. Therefore, a hazard ratio above 1 indicates an increase in the risk of experiencing the event, while a hazard ratio below 1 indicates a decrease in the risk of experiencing the event.

It is important to note that while the interpretation of the baseline function is similar to other regression techniques, \( h_0(t) \) remains unspecified as it constantly changes depending on the specific time point. Thus, the distribution of \( h_0(t) \) is allowed to follow any functional form. This is the primary reason why Cox regression is semiparametric. It should be noted that truly parametric Cox regression models do exist if you know the form of the baseline hazard prior to analysis (e.g., Weibull), and while parametric models usually have more statistical power than their nonparametric (or, semiparametric) counterparts, methodological research has shown that the semiparametric Cox model described in this text closely approximate the unspecified baseline hazard function, regardless of its true functional form.

Finally, no true pseudo-\( R^2 \) exists for Cox regression, although several have been suggested, and one will be discussed in detail below. The pseudo-\( R^2 \) for Cox regression is not interpreted as the percentage of variance explained. That is, the value does not indicate how much of the overall reason why a participant experienced the event. Instead, the value represents the relative association between survival and the independent variable. It is calculated as: 
\[
1 - \exp\left(-\frac{G^2}{n}\right),
\]
where \( G^2 \) is the likelihood ratio statistic and \( n \) is the sample size.

Interpretation of regression slopes and hazard ratios differ if the independent variable is categorical or continuous. As an example of a categorical independent variable, say you want to know whether participants who fell asleep within 60 minutes fell asleep faster taking your new sleep treatment or the leading over the counter treatment. As an example of a continuous independent variable, say you want to know whether body weight in pounds significantly predicts whether participants who fell asleep within one hour fell asleep faster. Each question will be analyzed separately. The collected data is presented below:

<table>
<thead>
<tr>
<th>ID</th>
<th>Tx</th>
<th>Wt</th>
<th>Asleep</th>
<th>Minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>210</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>191</td>
<td>1</td>
<td>53</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>135</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>185</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>140</td>
<td>0</td>
<td>36</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>159</td>
<td>1</td>
<td>54</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>120</td>
<td>1</td>
<td>26</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>175</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>223</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>204</td>
<td>1</td>
<td>59</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Tx</th>
<th>Wt</th>
<th>Asleep</th>
<th>Minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>0</td>
<td>156</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>169</td>
<td>1</td>
<td>48</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>157</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>134</td>
<td>1</td>
<td>51</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>205</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>126</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>145</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>140</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>235</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>198</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>163</td>
<td>1</td>
<td>22</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>139</td>
<td>1</td>
<td>35</td>
</tr>
<tr>
<td>23</td>
<td>1</td>
<td>127</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>24</td>
<td>1</td>
<td>148</td>
<td>1</td>
<td>29</td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>170</td>
<td>1</td>
<td>39</td>
</tr>
<tr>
<td>26</td>
<td>1</td>
<td>191</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>27</td>
<td>1</td>
<td>185</td>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>28</td>
<td>1</td>
<td>200</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>29</td>
<td>1</td>
<td>193</td>
<td>1</td>
<td>41</td>
</tr>
<tr>
<td>30</td>
<td>1</td>
<td>131</td>
<td>1</td>
<td>17</td>
</tr>
</tbody>
</table>
Following the data entry procedures described in Chapters 1 and 2, label the first variable ID, the second variable Tx (1 = Treatment; 0 = OTC), the third variable Wt for body weight in pounds, the fourth variable Asleep (1 = Yes; 0 = No), and the fifth variable Minutes to indicate the number of minutes the participant was in the study. Enter the data appropriately.

Assumptions

The assumptions for a simple Cox proportional-hazards model include all three assumptions of the Kaplan-Meier method: no differences between withdrawn and remaining participants, constant probability of survival over the entire study period, and that the event occurred at the time specified. Four additional assumptions include independence of residuals, proportionality of hazards, that the independent variable assumes a linear form, and absence of outliers. Note that the last two assumptions apply only to continuous independent variables.

No differences between withdrawn and remaining participants require that participants who are censored (or lost to follow-up) do not differ systematically from those who experience the event. For example, in the data above, a participant who was censored prior to study completion (e.g., ID = 1) may have been censored due to an adverse reaction to the medication. This is not an easily testable assumption because censoring can occur for a variety of reasons. However, binary logistic regression (Chapters 40 or 41) can be used to evaluate for differences between withdrawn and remaining participants. That is, those who withdrew from the study prior to completion are coded 1, with all remaining participants coded 0. This new binary variable serves as the dependent variable in the logistic regression analysis when assessing for statistically significant differences in the independent variable or covariates. Failure to achieve statistical significance satisfies the assumption. However, remember logistic regression requires large samples, so evaluation of this assumption may not be tenable in all situations.

Having constant probability of survival over the study period requires (1) that the same factors affecting survival are present throughout the entire study period and (2) participants entering the study at different time points have equal probability of survival. For example, a violation of this assumption would occur if the researcher forgot to turn off the lights after dosing the participant or the OTC medication was discontinued requiring substitution with a close, but not identical, medication. This assumption is evaluated by both examining the study design and protocol for violations. In addition, and more appropriate for longer-term studies, additional novel, effective, or ancillary treatments may become available that require ethical use. In these situations, the assumption is tested by examining the survival functions within subsets of participants (i.e., before vs. after differing treatment modalities).

Assuring that the event occurred at the time specified can be difficult, especially for left-or interval-censored participants (e.g., participant fell asleep 10-15 minutes after dosing, but the researcher is not exactly sure when). A violation of this assumption tends to bias survival estimates upwards.

Simple Cox regression is a between-subjects analysis; thus, the first assumption is independence of residuals. Independence is technically a design issue and can be satisfied by data that is not clustered and by not measuring participants repeatedly.

These first four assumptions were mostly assured via proper experimental design and protocol implementation. The three remaining assumptions are strictly related to the regression component, with the most important of these being the proportionality of hazards assumption. However, nonlinearity can often result in an erroneous violation of the proportionality of hazards
assumption. Thus, evaluating the form of a continuous independent variable is incredibly important. In Cox regression, continuous variables are assumed to have a linear form. This assumption is similar to the linearity in the logit assumption described for logistic regression. There are several methods for testing this assumption, but we will focus primarily on the use of Martingale residuals. Martingale residuals are skewed, ranging from negative infinity to 1. Values near 1 indicate the participant experienced the event sooner, whereas large negative residuals indicate the participant experienced the event later or was censored. The most straightforward approach to testing this assumption using Martingale residuals is to estimate a Cox regression model without the continuous independent variable. Note that Martingale residuals cannot be requested directly from SPSS, so you will need to save the cumulative hazard function from this analysis and compute them using the syntax provided. Next, print a simple scatterplot (Chapter 19) with the Martingale residuals on the y-axis and the previously omitted continuous variable on the x-axis. Finally, request a smoothing line (e.g., loess) to determine the functional form of the independent variable. If the smoothing line appears linear, the assumption is satisfied. However, if nonlinearity is evident you can include appropriate polynomial terms, use a piecewise or spline regression model, or, finally, transform the variable to make it linear (not recommended).

The proportionality of hazards assumption requires that the hazard ratio remain constant over time. Or, stated another way, the hazard for one individual is proportional to the hazard for any other individual. A third way of viewing this assumption is easily seen from the Cox regression equation presented in the introduction to this Chapter. Notice in the right side of the equation that the independent variable (i.e., X) does not interact with time (t). Thus, the assumption requires that there is no covariate-by-time interaction. There are several methods for evaluating this assumption, which include both graphical and statistical techniques, and although many techniques exist, only two of the more common methods will be described here—one graphical and one statistical. In general, graphical techniques are most appropriate for categorical independent variables and employ the Kaplan-Meier method (Chapter 42), whereas statistical techniques can be used for either categorical or continuous variables.

The graphical approach employs ln-ln (pronounced, log minus log often written as log-log) survival curves, which is simply a transformation of the estimated probability of survival from a Kaplan-Meier analysis. Here, the survival probability is saved for each participant and then transformed by taking the natural log of these probabilities twice. Note that the probability of survival is only saved for participants who experienced the event, so do not be alarmed if you are missing probabilities for the censored individuals. The ln-ln values are then plotted against the time variable (e.g., Minutes from the example data above) for each independent variable category (e.g., Tx in the example data). If the two plotted lines are relatively parallel, then the assumption is assured. There is no statistical test to determine whether the lines are indeed parallel; thus, you may have guessed the primary drawback of this (or any other) graphical approach is subjectivity. In general, however, unless there is drastic non-parallelism (i.e., stark intersection) you can generally assume proportionality of hazards. In the event that the proportionality of hazards assumption is violated for a categorical variable, you should conduct a stratified Cox model, where the offending predictor is used as a stratification variable. Although this technique is not described in full in this text, it is straightforward to implement in SPSS and will be presented brief detail below.

The statistical technique involves for testing the proportionality of hazards assumption evaluates for a statistically significant interaction between the independent variable of interest
and time. To use this method, you first create a new time variable that is calculated by taking the natural log of the time variable (e.g., \( \ln(\text{Minutes}) \)). Next, the interaction between the independent variable and the natural log of time (e.g., \( \text{Wt} \times \ln(\text{Minutes}) \)) from the example data above) as well as the main effect of the independent variable (e.g., \( \text{Wt} \)) is included in a Cox regression analysis. If the interaction effect is statistically significant, the independent variable interacts with time and the proportionality of hazards assumption is violated. In this situation, you need to include the interaction in all subsequent analyses, which result in a Cox regression with time-varying covariates (Chapter 45). This may sound tedious and confusing, but SPSS has a built in procedure for testing this assumption.

Finally, the absence of outliers assumption is evaluated by examining the deviance residuals, which are transformed Martingale residuals that are symmetric around 0, with an approximate standard deviation of 1. Positive values indicate the participant experienced the event sooner and negative values indicate experiencing the event later. Thus, very small (i.e., < -3.29) or very large (i.e., > 3.29) disconnected values indicate outliers. The absence of outliers assumption is evaluated by creating a simple scatterplot (Chapter 19) with the deviance residuals on the y-axis and each participant (i.e., \( \text{ID} \)) on the x-axis. Deviance residuals cannot be requested directly in SPSS, but can be computed by requesting the cumulative hazard function from a Cox regression analysis that included the independent variable and using the syntax provided.

### Categorical Independent Variable

#### Evaluating the Proportionality of Hazards Assumption

As stated above, there are two methods for testing the proportionality of hazards assumption for a categorical covariate—one graphical, one statistical. Both methods will be described in detail here, with the graphical method presented first. Note that because the graphical method uses the Kaplan-Meier method described in detail in Chapter 42, only the necessary steps for evaluating the assumption will be provided. If needed, refer to Chapter 42 for a more thorough description of the options for using the Kaplan-Meier method.

### Graphical Method

1. Click **Analyze**, then choose **Survival**, and finally click **Kaplan-Meier**… to bring up the **Kaplan-Meier** dialog box.
2. All of the variables in your dataset are listed on the left hand side of the dialog box.
   a. Click to highlight the **Minutes** variable, and then click the right arrow (\( \rightarrow \)) next to the **Time**: box.
   b. Click to highlight the **Asleep** variable, and then click the right arrow (\( \rightarrow \)) next to the **Status**: box.
      i. Click the **Define Event**… button to bring up the **Kaplan-Meier: Define Event for Status Variable** dialog box.
         1. Click the **Single Value**: radio button and type **1** in this box
         2. Click **Continue**.
   c. Click to highlight the **Tx** variable, and then click the right arrow (\( \rightarrow \)) next to the **Factor**: box.
3. Click the **Save** button to bring up the *Kaplan-Meier: Save New Variables* dialog box.
   a. Click the **Survival** checkbox and click **Continue**.

4. Click **OK**.

   In your dataset, a new variable has been created. This variable will be labeled *SUR_1* if this is the first time you saved the survival probabilities to your dataset (if the variable is *SUR_2* or *SUR_3*, change the syntax below accordingly. Notice that values were only calculated for participants who experienced the event. The next steps in this testing the assumption involve transforming these survival probabilities into ln-ln survival probabilities. This process involves using the Compute procedure as described in Chapter 13.

5. Click **Transform** and then click **Compute Variable**… to bring up the *Compute Variable* dialog box.

6. In the **Target Variable:** box, type: *lnlnSUR_1*.

7. In the **Numeric Expression:** box type: *ln(-ln(SUR_1))*.

8. Click **OK**.

   Next, you need to produce a scatterplot of the time variable plotted against the ln-ln transformed survival probabilities for each group separately. Again, this process only describes the minimal steps involved in producing the scatterplot. For a full description of the scatterplot procedure, see Chapter 19.

9. Click **Graphs**, then choose **Legacy Dialogs**, then click **Scatter/Dot**… to bring up the *Scatter/Dot* dialog box.

10. Click **Simple Scatter** and then click **Define**.

11. Click the **lnlnSUR_1** variable and then click the right arrow ( ) next to the **Y Axis:** box.

12. Click the **Minutes** variable and then click the right arrow ( ) next to the **X Axis:** box.

13. Click the **Tx** variable and then click the right arrow ( ) next to the **Set Markers by:** box.

14. Click **OK**.

   When you click **OK**, SPSS will produce an *Output* screen displaying the scatterplot. Click the *Output* window to view your results (if it does not pop up automatically). Notice that the plot for the two groups is roughly parallel and do not intersect, indicating the assumption is assured.

![Figure 43.1](image)
Statistical Method

The statistical method described below is used to test for an interaction between the independent variable and (the natural log of) time. There are two methods for testing this assumption. First, you can manually create the interactions between the categorical and time variable using the Compute procedure as described in Chapter 13. That is, compute a new variable that represents the natural log of time (e.g., ln(\textit{Minutes})), and then compute a new variable representing the interaction between the natural log of time and the categorical independent variable (e.g., ln(\textit{Minutes})*\textit{Tx}). This term, as well as the main effect of the categorical independent variable would then be included in a Cox regression described for the full analysis below.

The alternative procedure for testing this assumption, described here, uses SPSS to calculate the time variable for you; thus, this procedure will begin with a Cox regression procedure that is slightly different from the procedure used in the full analysis below.

1. Click **Analyze**, then choose **Survival**, and then click **Cox w/ Time-Dep Cov…** to bring up the **Compute Time-Dependent Covariate** dialog box, shown in Figure 43.2.

![Figure 43.2](image)

2. The dialog box above is used specifically to transform the time variable. You will notice that all the variables in your dataset are provided on the left hand side of this dialog box. However, notice that the first variable is new and listed as \textit{Time [T]}. This is an internal time variable created by SPSS to be used for all time related transformations. To test the proportionality of hazards assumption, you need to use the natural log of time.

   a. In the _Expression for T\_COV_: box, type: \textit{ln(T)}, as shown in Figure 43.2.

   b. Click the **Model...** button on the right hand side of the dialog box to bring up the **Cox Regression** dialog box. Note that this dialog box is identical to the one used for the full analysis; thus, only the minimal required steps are detailed here. For a full description of the options available, reference the descriptions provided below.
3. Click the **Minutes** variable and then click the right arrow (→) next to the **Time:** box.

4. Click to highlight the **Asleep** variable, and then click the right arrow (→) next to the **Status:** box.
   a. Click the **Define Event...** button to bring up the *Cox Regression: Define Event for Status Variable* dialog box.
      i. Click the **Single Value:** radio button and type 1 in this box.
      ii. Click **Continue.**

5. Click to highlight the **Tx** variable, and then click the right arrow (→) next to the **Covariates:** box.

6. Click the **Tx** variable, press and hold the **Ctrl** button on your keyboard, and then click the **T_COV_[T_COV_]** variable. With both variables highlighted, click the >a*b< button (→) next to the **Covariates:** box. This will place the interaction term in this box, which should appear as **T_COV_ *Tx.**

7. Click **OK.**

When you click **OK,** SPSS will produce an *Output* screen displaying the results. Click the *Output* window to view your results (if it does not pop up automatically). Scroll down to the **Block 1: Method = Enter** section of the output and find the **Variables in the Equation** table. It should appear identical to Figure 43.3 below. To evaluate the assumption, you will only be concerned with the interaction term. If this interaction is not statistically significant, the assumption is assured. That is, a not statistically significant interaction indicates there is no interaction between the independent variable and time. For the results below, the *p*-value (**Sig.**) for the interaction term is .179, which is not statistically significant. Thus, the proportionally of hazards assumption is considered satisfied. If this assumption had not been satisfied, you will need to include the interaction in all subsequent Cox models. This process and interpretation is described in detail in the Chapter 45.

<table>
<thead>
<tr>
<th>Variables in the Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Tx</td>
</tr>
<tr>
<td>T_COV_ *Tx</td>
</tr>
</tbody>
</table>

**Figure 43.3**

**Analysis**

With the proportionality of hazards assumption satisfied and assuming no differences between withdrawn and remaining participants, constant probability of survival over the entire study period, and that the event occurred at the time specified, to conduct a simple Cox regression with a categorical independent variable:
1. Click **Analyze**, then choose **Survival**, and then click **Cox Regression**… to bring up the Cox Regression dialog box, shown in Figure 43.4.

![Figure 43.4](image)

2. You will immediately see all of the variables in your dataset listed on the left hand side of the dialog box. You will be concerned primarily with three boxes labeled **Time**, **Status**, and **Covariates**: The **Time**: box is fairly self-explanatory. Here, you will place the dependent variable that measures time-to-event. The **Status**: box is essentially the censoring variable. That is, the variable that identifies who experienced the event or not. Finally, the **Covariates**: box is where the categorical independent variable will be placed. In addition, the **Strata**: box is available, but is not used for simple Cox regression.

   a. Click to highlight the **Minutes** variable, and then click the right arrow ( ) next to the **Time**: box.

   b. Click to highlight the **Asleep** variable, and then click the right arrow ( ) next to the **Status**: box.

   i. Click the **Define Event**… button to bring up the Cox Regression: Define Event for Status Variable dialog box. In this dialog box, you have three options.

      1. First, you can specify a **Single Value**: Although you may have multiple values coded for various events, the use of this box allows you to specify one value as the event and all other values as non-events.

         a. For the example data above, type 1 in this box.

      2. Alternatively, you can specify a **Range of values**: This option is useful if you have multiple values coded for various events, and you are interested in two or more events. However, note that the events must be coded in numerical order. That is, say you are interested only in the events coded 1 and 3, not event 2. By specifying a range of values, you would necessarily include the event coded 2, which is not of interest to you.
3. Alternatively, you can specify a *List of values*. This option can be used if you have any number of events, coded in any numerical order. Thus, if you are interested in events coded 1 and 3, you can specify each event individually by clicking the radio button, typing the event number of interest and clicking *Add*.

When you have specified the event(s) of interest, click *Continue*.

c. Click to highlight the **Tx** variable, and then click the right arrow (→) next to the *Covariates*: box.

3. Click the **Categorical…** button. Here, only the variables you entered in the *Covariates*: box in Step 2c above are shown on the left hand side. You will use this option when you have *any* categorical independent variable or covariate. Not specifying the independent variable as categorical will not influence your results; however, you would not be able to print separate survival or hazard functions for each group (Step 4 below). In addition, the categorical option removes the need for you to create new dummy variables as you did for linear regression for variables consisting of three or more groups.

a. Click to highlight the **Tx** variable and then click the right arrow (→) next to the *Categorical Covariates*: box.

b. Notice the *Change Contrast* section is now available. There are seven *Contrast* available from the drop down list. These include:

i. **Indicator**: Compares the presence (i.e., 1) or absence (i.e., 0) of the independent variable to a reference category. Note that *Indicator* is a synonym for dummy coding, the option we want to use for this example and described in detail in Chapter 23.

ii. **Simple**: Each category of the independent variable is compared to the unweighted average of all categories.

iii. **Difference**: Each category of the independent variable, except lowest coded category, is compared to the average effect of the previous categories. Also known as reverse Helmert contrasts, because…

iv. **Helmert**: Each category of the independent variable, except highest coded category, is compared to the average effect of the previous categories.

v. **Repeated**: Each category of the independent variable, except lowest coded category, is compared to previous category (e.g., 2 to 1; 3 to 2, etc).

vi. **Polynomial**: Available only for continuous independent variables; uses orthogonal polynomial contrasts when categories are equally spaced. Think, trend-type analysis.

vii. **Deviation**: Each category of the independent variable, except the reference category, is compared to the unweighted (i.e., mean) overall effect.

c. For several of the options listed above you need to identify the *Reference Category*: You can select either the *First* or the *Last* category as reference. If you are more interested in a middle category, the variable will need to be recoded (see
Chapter 8). For this example, however, we only have two groups, so it does not matter.

i. Select the **First** category as reference and then click the **Change** button.

When you are satisfied with your selections, click **Continue**.

4. Click the **Plots...** button. Here, you can select the **Plot Type** you want to print. There are several options and each is listed below.

   a. **Survival**: Selecting this option will print the cumulative survival function on the original linear time scale.

   b. **Hazard**: Selecting this option prints the cumulative hazard function on the original linear time scale.

   c. **Log minus log**: Selecting this option prints the cumulative survival function after the ln-ln transformation.

   d. **One minus survival**: Selecting this option prints the one-minus survival function on the original linear time scale.

When you select a **Plot Type**, the **Covariate Values Plotted at:** box becomes available. This box will contain the independent variable you specified in Step 2c above. When you click on the independent variable in this box, the **Change Value** options becomes available. Each plot depends on the value of the independent variable, so you must use some constant value of the variable to plot against time. As you can see, the default is **Mean**; however, you can enter any value you would like by clicking the **Value** radio button, typing the value in the box, and clicking **Change**. In addition, the **Separate Lines for:** box is available, which allows you to produce separate lines for each category of your independent variable within each plot selected above.

   a. In the **Covariate Values Plotted at:** click the **Tx (Cat) (Mean)** variable, and then click the right arrow ( ) next to the **Separate Lines for:** box. Note that if you do not see the (Cat) in the name of this variable, you did not specify it as categorical in Step 3 above.

When you are satisfied with your selections, click **Continue**.

5. Click the **Save** button to bring up the **Cox Regression: Save** dialog box. Selecting any of these options will create a new variable in your dataset that can be used to test hypotheses or evaluate certain assumptions.

   a. The **Survival function** option saves the probability of survival for each participant given the duration of the study.

   b. The **Standard error of survival function** option saves the standard error of the survival function in the option above.

   c. The **Log minus log survival function** option saves the cumulative survival function for each participant after the ln-ln transformation has been applied.

   d. The **Hazard function** option saves the cumulative hazard estimate, also known as the Cox-Snell residuals. These residuals are used to calculate Martingale and Deviance residuals.
e. The Partial Residuals option saves the partial residuals, also known as Schoenfeld residuals, for each independent variable in the model.

f. The DfBeta(s) option saves the estimated change in the coefficient value if the participant was removed from analysis. A new variable is saved for each independent variable in the model.

g. The $X * Beta$ option saves the estimated linear predictor score. It is the sum of the product of the mean-centered covariate and the corresponding parameter estimates for each participant.

Finally, you can Export Model Information to XML File by identifying a previously created XML file via the Browse... button. When you are satisfied with your selections, click Continue.

6. Finally, click the Options... button to bring up the Cox Regression: Options dialog box. Here, you can select several Model Statistics you want printed as well as information regarding the required Probability for Stepwise entry or removal.

   a. Under the Model Statistics section, you can choose to print the confidence interval for the hazard ratios (CI for $\exp(B)$), as well as set the confidence interval level, with the default being 95%. You can also print the Correlation of estimates, but this option is only appropriate when you have more than one independent variable. Finally, you have the choice to Display Model Information At each step or At last step.

   b. Under the Probability for Stepwise section, you can indicate how you want SPSS to use $p$-values to enter (Entry:) or remove (Removal:) variables in a stepwise regression analysis. Because we are not conducing stepwise regression in the example, these options can be ignored.

   c. Because Cox regression uses partial likelihoods, it uses an iterative estimation process to maximize the likelihood that the estimates are correct. You can set the number of Maximum Iterations: which essentially controls how long the procedure will search for a solution.

   d. Finally, you can Display baseline function, which prints the baseline hazard function and cumulative survival at the mean of the covariates. Note that this option is not available if you have time-varying covariates.

When you are satisfied with your selections, click Continue.

7. That’s it! Click OK to conduct the analysis.

Output

When you click OK, SPSS will produce an Output screen displaying the results. Click the Output window to view your results (if it does not pop up automatically). The first table you see is titled Case Processing Summary, which contains important information about your sample including the frequency and percentage of participants experiencing the event (Event) and Censored, the Total number of participants, as well as participants with missing values, negative time, and those that were left-censored.
The next table, titled *Categorical Variable Codings*, is presented in Figure 43.5. This table contains information that is incredibly important to interpretation as it indicates how the categorical independent variable was coded as well as the number of participants in each category.

<table>
<thead>
<tr>
<th>Categorical Variable Codings²</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency</td>
<td>(i)</td>
</tr>
<tr>
<td>T² – 1=OTC</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>T² – 1=Treatment</td>
<td>15</td>
<td>1</td>
</tr>
</tbody>
</table>

*Fig. 43.5*

The next section of the output is titled *Block 0: Beginning Block*. This block can be thought of as similar to the Block 0 from a logistic regression analysis. That is, it contains information for the model without the independent variable and serves as the comparison or reference model. The table in this section is titled *Omnibus Tests of Model Coefficients* and contains the -2 Log Likelihood value.

Next, the *Block 1: Method=Enter* section is shown which contains the results of the Cox regression analysis. The first table, titled *Omnibus Tests of Model Coefficients*, shown in Figure 43.6, contains the overall test of the model based on a statistical comparison between Block 0 and Block 1. This table contains the -2 Log Likelihood value as well as omnibus tests for the *Overall (score)* model as well as *Change from Previous Step* and *Change from Previous Block*. Note that the latter two sections of this table provide identical results because we did not use hierarchical, sequential, or stepwise Cox regression. Further, note that the Chi-square value from the *Change From Previous Block* section is the likelihood ratio statistic (i.e., $G^2$) and is used when calculating effect size.

<table>
<thead>
<tr>
<th>Omnibus Tests of Model Coefficients³</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>-2 Log Likelihood</td>
<td>Overall (score)</td>
<td>Change From Previous Step</td>
<td>Change From Previous Block</td>
</tr>
<tr>
<td></td>
<td>Chi-square</td>
<td>df</td>
<td>Sig.</td>
</tr>
<tr>
<td>101.082</td>
<td>4.342</td>
<td>1</td>
<td>.026</td>
</tr>
</tbody>
</table>

*Fig. 43.6*

The next table is titled *Variables in the Equation* and is shown in Figure 43.7. This is the most important table in the output because it contains the results of the independent variable. You probably noticed that there is no constant (i.e., y-intercept), as provided in both linear and logistic regression. This is due to the fact that the baseline hazard function is unspecified. Or, stated another way, the baseline function is dependent on the specific time point. This table contains the regression slope ($B$), standard error ($SE$), *Wald* statistic, degrees of freedom ($df$), $p$-value ($Sig.$), hazard ratio ($Exp(B)$) as well as the 95% confidence interval around the hazard ratio ($95.0\% CI for Exp(B)$).
Next, you are presented a table titled *Covariate Means and Pattern Values* and a plot titled *Survival Function at mean of covariates*. The plot presents the survival function for a hypothetical participant who is at the mean of the covariates. For the example data above, the mean of the covariates is .50 because there were 15 participants in each group (i.e., \([(1*15)+(0*15)]/30\)). Because the groups were coded 0/1, this survival function is for a participant who is neither in the treatment nor OTC group. This same scenario is applied to the plot further down in the output titled, *Hazard Function at mean of covariates*. In general, survival and hazard functions at the mean of the covariate are irrelevant for categorical independent variables.

Finally, two plots that are extremely useful for publication purposes are titled *Survival Function for patterns 1-2* and *Hazard Function for patterns 1-2*, and are presented in Figure 43.8. There are two important characteristics to notice about both functions. First, they are both step functions. That is, the function steps down in survival, or up in hazard, each time a participant experiences the event. The magnitude of the step down or step up is determined by the sample size, with smaller steps associated with larger samples. Further, the horizontal distance between steps is determined by the length of time between events. The primary difference between these functions and those printed using the Kaplan-Meier method is that censored participants are not indicted on the plot by vertical slashes. Thus, if you want to show graphically when participants were censored, use the Kaplan-Meier method described in Chapter 42.
Interpretation

Because all assumptions for a categorical independent variable were evaluated prior to analysis, interpretation begins by evaluating the Omnibus Tests of Model Coefficients table in the Block 1: Method=Enter section shown in Figure 43.6. Consider only the omnibus test found in the Change From Previous Block columns. The Chi-square value of 4.700 is statistically significant as \( p = .030 \). Note that this value was calculated as the difference between the \(-2\ \text{Log Likelihood}\) from Block 0: Beginning Block (i.e., 105.763) and the \(-2\ \text{Log Likelihood}\) from Block 1: Method=Enter (i.e., 101.062) sections. This value, termed the likelihood ratio statistic (or \( G^2 \)) is used when calculating the effect size estimate. As stated above, the effect size estimate is calculated as: 1 – \( \exp(-G^2/n) \). Thus, the \( R^2 \) estimate equals .145 (i.e., 1 – \( \exp(-4.700/30) \)), which represents the relative association between survival time and the independent variable.

Next, you evaluate the effect of the independent variable, which is found in the Variables in the Equation table. Note that because you have only one independent variable in this model, the omnibus test already informed you that your independent variable is statistically significant (assuming no assumption violations). Evaluate the regression coefficient (\( B \)) and standard error (S.E.) as well as the hazard ratio (\( \exp(B) \)) and confidence interval (95.0% C.I. for \( \exp(B) \)). Large standard errors and wide confidence intervals indicate poor measurement. In addition, a 95% confidence interval that contains 1 indicates the variable is not statistically significant.

It is also noted that for categorical independent variables, the regression coefficient and hazard ratio are calculated for the group coded 1 in reference to the group coded 0. Thus, the log-hazards of 1.048 and hazard ratio of 2.851 in Figure 43.7 were calculated for the group receiving the new sleep treatment because they were coded 1 in the data. Because log-hazards are linear and symmetric around 0, the log-hazard for the OTC group are simply -1.048. Further, to calculate the hazard ratio for the reference (OTC) group, simply take the reciprocal of the \( \exp(B) \). Thus, 1/2.851 = .351. Alternatively, you can exponentiate the log-hazards for OTC as \( \exp(-1.048) = e^{1.048} = .351 \).

The interpretation of log-hazards are similar to log-odds in logistic regression and slopes in linear regression—a one-unit increase in the independent variable results in an increase (+) or decrease (-) in the log-hazards of experiencing the category coded 1 on the dependent variable. Here, because the independent variable is dichotomous, a one-unit increase indicates going from the OTC group (coded 0) to the new sleep treatment group (coded 1). Thus, being in the group receiving the new sleep treatment resulted in a 1.048 increase in the log-hazards of falling asleep within one hour. Alternatively, being the in OTC group resulted in a 1.048 decrease in the log-hazards of falling asleep within one hour.

When interpreting hazard ratios, you can now indicate which group is more likely to experience the event. This is primarily due to the fact that hazard ratios are based on risk, which are based on the number of participants not experiencing the event up to a given time point (i.e., a rate ratio). Thus, from Figure 43.7, the hazard ratio of 2.851 for the new sleep treatment indicates that participants who received the new sleep treatment were 2.851 times more likely to fall asleep within one hour compared to those who received the OTC treatment. Alternatively, those who received the OTC treatment were 64.9% (i.e., 1-.351) less likely to fall asleep within one hour compared to those who received the new sleep treatment.
Two additional points need deserve attention. First, note that hazard ratios always compare two groups; thus, interpretation always needs to identify the two groups that were compared. This is why the italics were used as emphasis above. Second, note that hazard ratios are not symmetric—a 285.1% increase versus a 64.9% decrease.

Finally, the survival and hazard functions presented in Figure 43.8 should be provided in the manuscript. Both plots are incredibly useful when depicting group differences visually. The choice of which plot you present to your audience is completely up to you. By quickly looking through the literature, you will see that each kind of plot is presented equally as often. The interpretation of the plots differs slightly for survival and hazard functions. The survival function ranges from 0 to 1. A value of 0 indicates everyone has experienced the event (or been censored) at a specific time point, whereas a value of 1 indicates no one has experienced the event at a specific time point. It is always important to consider the time aspect, as the survival function represents the probability of not experiencing the event in the next time interval, which for this example is the next minute. Alternatively, the hazard function ranges from 0 to infinity, with greater values indicating greater likelihood of experiencing the event in the next time unit.

Example Results Section

Prior to analysis, no differences between withdrawn and remaining participants were indicated and the probability of survival was assumed constant over the entire study period. Further, the proportionality of hazards assumption was satisfied as evidenced using both graphical and statistical methods. The log-minus-log survival plots indicated parallel survival functions and there was no statistically significant interaction between treatment group and time ($p = .179$).

The results of a simple Cox regression analysis indicated treatment group significantly predicted time to fall asleep during the 60-minute study period, $\chi^2_{1} = 4.942$, $p < .05$, $R^2 = .145$. Participants receiving the new sleep treatment had a 1.048 increase in the log-hazards and were 2.851 times more likely to fall asleep during the study period compared to participants receiving the OTC treatment. This effect is evidenced by both the survival and hazard plots.

Continuous Independent Variable

Evaluating the Form of the Independent Variable

Cox regression assumes the form of a continuous independent variable is linear. This, of course, is an empirical question. If a violation of this assumption is indicated, nonlinearity can result in a violation of the proportionality of hazards assumption. Thus, you should always check this assumption prior to evaluating proportionality of hazards. To test this assumption we need to calculate the Martingale residuals from a baseline model excluding the independent variable. That is, from a model where the independent variable of interest is not included in the analysis. This may seem awkward, and it is, because for a simple Cox regression, testing the assumption requires that no independent variable be modeled. Note the steps below are the minimal requirements to test the assumption.

1. Click Analyze, then choose Survival, and then click Cox Regression… to bring up the Cox Regression dialog box.
2. Click **Reset** to clear all previous selections.

3. Click to highlight the **Minutes** variable, and then click the right arrow (►) next to the **Time**: box.

4. Click to highlight the **Asleep** variable, and then click the right arrow (►) next to the **Status**: box.
   a. Click the **Define Event…** button and in the **Single Value**: box, type 1.

5. Click **Continue**.

6. Click the **Save…** button.
   a. Click the **Hazard function** checkbox to request the Cox-Snell residuals.
   b. Click **Continue**.

7. Click **OK**.

From Step 6 above, the Cox-Snell residuals were saved as a new variable in the dataset named `HAZ_1`. Note that if you conducted an analysis previously in which you requested Cox-Snell residuals, the variable name might be different (e.g., `HAZ_2`, `HAZ_3`, etc). We assume the name of the variable is `HAZ_1` in this example, but change the name accordingly if needed. To calculate the Martingale residuals, we need to transform the Cox-Snell residuals, a process that requires the syntax commands shown below.

8. Click **File**, then choose **New**, and then click **Syntax…**. A new syntax window will pop-up automatically.

9. Copy and paste (or type) the syntax commands below into the syntax window as shown in Figure 43.9.

   ```plaintext
   COMPUTE Mart_1 = (Asleep = 1) - HAZ_1.
   FORMATS Mart_1 (f8.5).
   EXECUTE.
   ```

10. Click **Run** and then click **All…**

---

**Figure 43.9**
This syntax will create a new variable in the dataset named Mart_1, which represents the Martingale residuals. To test the form of the independent variable:

11. Click **Graphs**, then choose **Legacy Dialogs**, and then click **Scatter/Dot**.
12. Click **Simple Scatter** and then click **Define**.
13. Click the **Mart_1** variable and then click the right arrow ( ) next to the **Y Axis**: box.
14. Click the **Wt** variable and then click the right arrow ( ) next to the **X Axis**: box.
15. Click **OK**.

When you click **OK**, SPSS will produce an **Output** screen displaying the results. Click the **Output** window to view your results (if it does not pop up automatically). The scatterplot should appear similar to Figure 43.10.

![Figure 43.10](image)

Although we can see subjectively the form is relatively linear, we need an objective indicator of linearity. Thus, we will employ a smooth curve known as a loess line. The loess line is a locally weighted regression line that is generally smooth with its location and direction determined by the percentage of data points in its vicinity. Previous literature has shown that a loess line fitting approximately 75% of data points is sufficient to test the assumption.

16. In the **Output** window, double click the scatterplot to bring up the **Chart Editor** window, shown in Figure 43.11.
17. Click **Elements** and then click **Fit Line at Total** to bring up the **Properties** dialog box.
18. Click the **Loess** radio button, type **75** into the **% of points to fit**: box, and then click the **Apply** button.
19. Click **Close**.
20. Close the **Chart Editor** window either by clicking **File** and then clicking **Close**, or by clicking ✗ in the upper right hand corner of the window.
Whew, that was a lot of steps! Based on the loess line shown in Figure 43.11, we can see that Martingale residuals are linear across all values of the independent variable. Thus, we can consider the independent variable to have a linear form.

**Evaluating the Proportionality of Hazards Assumption**

For continuous independent variables, only the statistical method can be used to evaluate the proportionality of hazards assumption. Note that the graphical method cannot be used because, with one continuous independent variable, there are no groups to differentiate. Testing the assumption uses the exact same steps as described above in the *Categorical Independent Variable* section.

1. Click **Analyze**, then choose **Survival**, and then click **Cox w/ Time-Dep Cov…** to bring up the *Compute Time-Dependent Covariate* dialog box, shown in Figure 43.2.

2. The dialog box is used specifically to transform the time variable. You will notice the first variable listed on the left hand side of the dialog box is $T_{[\ldots]}$. This is an internal time variable created by SPSS to be used for all time transformations. To test the proportionality of hazards assumption, you need to use the natural log of time.
   a. In the **Expression for $T_{COV}$**: box, type: $\ln(T_{[\ldots]})$, as shown in Figure 43.2.
   b. Click the **Model…** button on the upper right corner of the dialog box to bring up the *Cox Regression* dialog box.

3. Click the **Minutes** variable and then click the right arrow (→) next to the **Time**: box.

4. Click to highlight the **Asleep** variable and then click the right arrow (→) next to the **Status**: box.
   a. Click the **Define Event…** button to bring up the *Cox Regression: Define Event for Status Variable* dialog box.
      i. In the **Single Value**: box type 1 in this box and then click **Continue**.
5. Click to highlight the Wt variable, and then click the right arrow (►) next to the Covariates: box.

6. Click the Wt variable, press and hold the Ctrl button on your keyboard, and then click the T_COV_*[T_COV_] variable. With both variables highlighted, click the >a*b> button (▶) next to the Covariates: box. This will place the interaction term in this box, which should appear as T_COV_*Wt.

7. That’s it! Click OK to see the results.

When you click OK, SPSS will produce an Output screen displaying the results. Click the Output window to view your results (if it does not pop up automatically). Scroll down to the Block 1: Method = Enter section of the output and find the Variables in the Equation table. It should appear identical to Figure 43.12 below. To evaluate the assumption, you will only be concerned with the interaction term where a non-statistically significant interaction indicates that the assumption is assured. That is, a non-statistically significant interaction indicates there is no interaction between the independent variable and time. For the results below, the p-value (Sig.) for the interaction term is .115, which is not statistically significant; thus, the proportionally of hazards assumption is considered satisfied. If this assumption had not been satisfied, you will need to include this interaction in the primary Cox model, a process described in Chapter 45.

![Variables in the Equation Table]

<table>
<thead>
<tr>
<th>Variables in the Equation</th>
<th>B</th>
<th>SE</th>
<th>Wald</th>
<th>df</th>
<th>Sig</th>
<th>Exp(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wt</td>
<td>-.174</td>
<td>.088</td>
<td>3.851</td>
<td>1</td>
<td>.050</td>
<td>.841</td>
</tr>
<tr>
<td>T_COV_*Wt</td>
<td>.038</td>
<td>.024</td>
<td>2.483</td>
<td>1</td>
<td>.115</td>
<td>1.039</td>
</tr>
</tbody>
</table>

**Figure 43.12**

**Analysis**

With the proportionality of hazards assumption satisfied, the independent variable shown to have a linear form, and assuming no differences between withdrawn and remaining participants, constant probability of survival over the entire study period, and that the event occurred at the time specified, to conduct a simple Cox regression with a continuous covariate: (note that the description at each step has been reduced, see the Categorical Independent Variable section for a full description of each available option):

1. Click Analyze, choose Survival, and then click Cox Regression… to bring up the Cox Regression dialog box, shown in Figure 43.4.

2. You will immediately see all of the variables in your dataset listed on the left hand side of the dialog box.
   a. Click to highlight the Minutes variable, and then click the right arrow (►) next to the Time: box.
   b. Click to highlight the Asleep variable, and then click the right arrow (►) next to the Status: box.
i. Click the Define Event… button to bring up the Cox Regression: Define Event for Status Variable dialog box.

1. In the Single Value: box, type 1 and then click Continue.
   c. Click to highlight the Wt variable, and then click the right arrow (→) next to the Covariates: box.

3. The Categorical… button is available, but we are using a continuous independent variable, so we ignore it.

4. Click the Plots… button. Here, you can select the Plot Type you want to print. There are several options, these are listed below.
   a. Survival: Selecting this option will print the cumulative survival function on the original linear time scale.
   b. Hazard: Selecting this option prints the cumulative hazard function on the original linear time scale.

When you select a Plot Type, the Covariate Values Plotted at: box becomes available. This box will contain the independent variable you specified in Step 2c above. When you click on the independent variable in this box, the Change Value options becomes available. Each plot depends on the value of the independent variable, so you must use some constant value of the variable to plot against time. As you can see, the default is Mean; however, you can enter any value you would like by clicking the Value: radio button, typing the value in the box, and clicking Change. It is important to print meaningful survival and hazard functions. For this example, however, we will just have the functions estimated at mean body weight. In addition, the Separate Lines for: box is available, but is not an option for continuous independent variables as there are no groups to compare. When you are satisfied with your selections, click Continue.

5. Click the Save button to bring up the Cox Regression: Save dialog box. Selecting any of these options will create a new variable in your dataset that can be used to test hypotheses or evaluate certain assumptions.
   a. Click the Hazard function checkbox to save the Cox-Snell residuals to your dataset that will be used to calculate the Martingale and Deviance residuals.

6. Finally, click the Options… button to bring up the Cox Regression: Options dialog box. Here, you can select several Model Statistics you want printed as well as information regarding the required Probability for Stepwise entry or removal.
   a. Under the Model Statistics section, you can choose to print the confidence interval for the hazard ratios (CI for exp(B)), as well as set the confidence interval level, with the default being 95%. You can also print the Correlation of estimates, but this option is only appropriate when you have more than one independent variable. You also have the choice to Display Model Information At each step or At last step.

When you are satisfied with your selections, click Continue.

7. That’s it! Click OK to conduct the analysis.
Output

When you click OK, SPSS will produce an Output screen displaying the results. Click the Output window to view your results (if it does not pop up automatically). The first table you see is titled Case Processing Summary, which contains important information about your sample including the frequency and percentage of participants experiencing the event (Eventa) and Censored, the Total number of participants, as well as participants with missing values, negative time, and those that were left-censored.

The next section of the output is titled Block 0: Beginning Block. This block can be thought of as similar to that from logistic regression. That is, it contains information for the model without the independent variable and serves as the comparison model. The only table in this section is titled Omnibus Tests of Model Coefficients, which contains the -2 Log Likelihood information.

The Block 1: Method=Enter section contains the results of the Cox regression. The first table, titled Omnibus Tests of Model Coefficients and shown in Figure 43.13, contains the overall test of your model, and is a statistical comparison between Block 0 and Block 1. This table contains the -2 Log Likelihood value as well as omnibus tests for the Overall (score) model as well as Change From Previous Step and Change from Previous Block. Note that the latter two produce identical results because we did not use hierarchical, sequential, or stepwise Cox regression. Note that the Chi-square value from the Change From Previous Block section is used when calculating the effect size.

<table>
<thead>
<tr>
<th>Omnibus Tests of Model Coefficientsa</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2 Log Likelihood</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>Chi-square</td>
</tr>
<tr>
<td>89.020</td>
</tr>
</tbody>
</table>

Figure 43.13

The next table is titled Variables in the Equation and is shown in Figure 43.14. This is the most important table in the output because it contains the results of the independent variable. Notice there is no constant, as in both linear and logistic regression, because the baseline hazard function is unspecified. The table contains the regression slope (B), standard error (SE), Wald statistic, degrees of freedom (df), p-value (Sig.), hazard ratio (Exp(B)) as well as the 95% confidence interval around the hazard ratio (95.0% CI for Exp(B)).

<table>
<thead>
<tr>
<th>Variables in the Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
</tr>
<tr>
<td>Wt</td>
</tr>
</tbody>
</table>

Figure 43.14
Next, you are presented a table titled Covariate Means and two plots titled Survival Function at mean of covariates and Hazard Function at mean of covariates. These plots present the survival or hazard function for a hypothetical participant who is at the mean of the independent variable. For the example data above, the mean body weight across the sample of 30 participants is 168.467. Both plots are shown in Figure 43.15.

![Figure 43.15](image)

**Interpretation**

Interpretation begins by evaluating the absence of outliers assumption, which requires the calculation of deviance residuals. Because SPSS does not calculate deviance residuals directly, we need to calculate the residual values based on the Cox-Snell residuals we requested in Step 5a above. This process requires syntax commands. In the dataset, a new variable was saved named HAZ_2 (HAZ_1 was used when testing the form of body weight above). Note that if you conducted this analysis previously, the variable name might be different (e.g., HAZ_3, HAZ_4, etc). If this is the case, simply change the name in the syntax commands.

1. Click File, then choose New, and then click Syntax…. A new syntax window will pop-up automatically.
2. Type the syntax commands below into the syntax window as shown in Figure 43.16 on the next page.

```plaintext
COMPUTE Mart_2 = (Asleep = 1) - HAZ_2.
COMPUTE Dev_2 = SQRT(-2*(Mart_2+(Asleep=1)*LN((Asleep=1)-Mart_2))). IF Mart_2 < 0 Dev_2 = -Dev_2.
FORMATS Mart_2 Dev_2 (f8.5).
EXECUTE.
```
3. Click Run and then click All….

Two new variables are now created in the dataset named Mart_2 and Dev_2. I used the name Dev_2 to keep the names consistent and to remind you from which Martingale residuals the deviance residuals were calculated. To evaluate for outliers, we simply print a scatterplot of the deviance residuals against the participant’s ID variable. That is, plot the deviance residual for each participant.
1. Click **Graphs**, then choose **Legacy Dialogs**, and then click **Scatter/Dot**.
2. Click **Simple Scatter** and then click **Define**.
3. Click **Reset** to clear all previous selections.
4. Click the **Dev_2** variable and then click the right arrow (→) next to the **Y Axis:** box.
5. Click the **ID** variable and then click the right arrow (→) next to the **X Axis:** box.
6. Click **OK**.

When you click **OK**, SPSS will produce an **Output** screen displaying the results. Click the **Output** window to view your results (if it does not pop up automatically). The scatterplot should appear as shown in Figure 43.17. No deviance residuals are extremely large (e.g., > 3.29 or < -3.29) and no residual values are disconnected. Thus, we can consider the assumption satisfied.

With all assumptions satisfied, we can now be interpreting the results of the Cox regression analysis. Interpretation begins by evaluating the **Omnibus Tests of Model Coefficients** table in the **Block 1: Method=Enter** section shown in Figure 43.13. Consider only the **Change From Previous Block** columns. The Chi-square value of 16.742 is statistically significant as $p < .0005$. Note that this value was calculated as the difference between the -2 Log Likelihood from **Block 0: Beginning Block** (i.e., 105.763) and the -2 Log Likelihood from **Block 1: Method=Enter**
This value, termed the likelihood ratio statistic (or $G^2$) is used when calculating the effect size estimate. As stated above, the effect size estimate is calculated as: $1 - \exp(-G^2/n)$. Thus, the $R^2$ estimate equals 0.428 (i.e., $1 - \exp(-16.742/30)$), which represents the relative association between survival time and the independent variable. In addition, degrees of freedom are calculated as the difference in the number of parameters between models; that is, the model in Block 1 has one more parameter (i.e., body weight) than the Block 0.

Next, you evaluate the effect of the independent variable, which is found in the *Variables in the Equation* table. Note that because you have only one independent variable in this model, the omnibus test already informed you that the independent variable is statistically significant (assuming no assumption violations). Evaluate the regression coefficient ($B$) and standard error (S.E.) as well as the hazard ratio ($\exp(B)$) and confidence interval (95.0% C.I. for $\exp(B)$). Large standard errors and wide confidence intervals indicate poor measurement. In addition, a 95% confidence interval that contains 1 indicates the variable is not statistically significant.

The interpretation of log-hazards are similar to log-odds in logistic regression and slopes in linear regression—a one-unit increase in the independent variable results in an increase (+) or decrease (-) in the log-hazards of experiencing the event. Thus, every one-pound increase in body weight resulted in a 0.039 decrease in the log-hazards of falling asleep within one hour. It is important to note that log-hazards are symmetric and additive; thus, for every ten-pound increase in body weight, the log-hazard decreases by 0.390 (i.e., -0.039*10).

When interpreting hazard ratios, a one-unit increase results in how much more or less likely a participant was to experience the event. This is primarily due to the fact that hazard ratios are based on risk, which are based on the number of participants not experiencing the event up to a given time point (i.e., a rate ratio). Thus, from Figure 43.14, the hazard ratio of 0.962 indicates that for every one-pound increase in body weight, a participant was 3.8% (i.e., 1-0.962) less likely to fall asleep within one hour. It is important to note that hazard ratios are neither symmetric nor additive; thus, the decrease in hazard for a ten-pound increase in body weight is not simply 3.8% multiplied by 10. Instead, calculate a ten-pound increase using the log-hazard (i.e., -0.039*10) and then calculate the hazard ratio (i.e., $e^{-0.390} = 0.677$ or a 32.3% decrease).

Finally, the survival and hazard functions presented in Figure 43.15 should be provided in the manuscript, with the choice of which plot you present to your audience being completely up to you. Be sure to note what value of the independent variable was used when calculating the plots. In this example, the plots were calculated at a mean body weight of 168.467 pounds. The interpretation of the plots differs slightly for survival and hazard functions. The survival function ranges from 0 to 1. A value of 0 indicates everyone has experienced the event (or been censored) at a specific time point, whereas a value of 1 indicates no one has experienced the event at a specific time point. It is always important to consider the time aspect, as the survival function represents the probability of not experiencing the event in the next time interval, which for this example is the next minute. Alternatively, the hazard function ranges from 0 to infinity, with greater values indicating greater likelihood of experiencing the event in the next time unit.

**Example Results Section**

Prior to analysis, no differences between withdrawn and remaining participants were indicated and the probability of survival was assumed constant over the entire study period. The proportionality of hazards assumption was satisfied as evidenced as no statistically significant interaction between body weight and the outcome.
weight and time was indicated ($p = .115$). In addition, the linear form of body weight was assured by examining Martingale residuals and no outliers were indicated via deviance residuals.

The results of a simple Cox regression analysis indicated body weight in pounds significantly predicted time to fall asleep during the 60-minute study period, $\chi^2_{1} = 16.742, p < .05, R^2 = .428$. Every one-pound increase in body weight decreased the log-hazards of falling asleep within one hour by .039. Further, a one-pound increase in body weight resulted in participants being 3.8% less likely to fall asleep within one hour. This effect is evidenced by both the survival and hazard plots based on a mean body weight of 168.467 pounds.
Chapter 44
MULTIPLE COX REGRESSION WITHOUT TIME-VARYING COVARIATES

Multiple Cox regression is an extension of simple Cox regression to situations involving more than one independent variable. The extension is most easily seen through the Cox regression equation: \( \ln(h(t|X)) = h_0(t) + B_1X_1 + B_2X_2 + \cdots + B_kX_k \). For a more complete derivation of the equation, see Chapter 43. Briefly, \( \ln(h(t|X)) \) are the predicted log-hazards, \( h_0(t) \) is the unspecified baseline hazard function, the \( Bs \) are the slopes of the log-hazards for each independent variable or covariate, and the \( Xs \) are the values of the independent variable or covariate. You can have as many independent variables and covariates as statistical power affords, and they can be measured on any scale.

One of the most common reasons for using multiple Cox regression is to control for nuisance variables called covariates. Covariates are variables that are correlated with the dependent variable, but are not of primary research interest. Thus, in many situations, you may have only one or two independent variables and numerous covariates. In reality, however, the difference between an independent variable and covariates is semantic. The analysis treats them all the same. It is up to you to determine which variables are interesting and which are nuisance.

Because multiple Cox regression is a direct extension of simple Cox regression, the analysis and interpretation should be straightforward assuming you have a solid understanding of simple Cox regression. Feel free to re-read Chapter 43 if necessary. To avoid redundancy in explanation, we will jump right into the example.

Say you want to know whether participants who take your new sleep treatment fall asleep faster during a 60-minute study period compared to participants who take the leading over-the-counter treatment or placebo. Because the amount of sleep per night can be affected by numerous variables, you decide to control for caffeine consumption after 2pm (in milligrams), whether the participant smokes, body weight in pounds, and level of global anxiety as measured by the Hamilton Anxiety Rating Scale (HAM-A). The collected data is presented below:

<table>
<thead>
<tr>
<th>ID</th>
<th>Tx</th>
<th>Caff</th>
<th>Smk</th>
<th>Wt</th>
<th>Anc</th>
<th>Asleep</th>
<th>Mins</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>89</td>
<td>1</td>
<td>162</td>
<td>8</td>
<td>1</td>
<td>57</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>191</td>
<td>1</td>
<td>209</td>
<td>22</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>115</td>
<td>0</td>
<td>172</td>
<td>10</td>
<td>1</td>
<td>58</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>246</td>
<td>1</td>
<td>191</td>
<td>18</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>178</td>
<td>1</td>
<td>189</td>
<td>13</td>
<td>0</td>
<td>45</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>99</td>
<td>0</td>
<td>146</td>
<td>15</td>
<td>1</td>
<td>49</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>192</td>
<td>1</td>
<td>193</td>
<td>11</td>
<td>0</td>
<td>39</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>210</td>
<td>1</td>
<td>167</td>
<td>10</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>105</td>
<td>1</td>
<td>192</td>
<td>20</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>149</td>
<td>0</td>
<td>199</td>
<td>16</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>198</td>
<td>1</td>
<td>184</td>
<td>12</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>145</td>
<td>1</td>
<td>170</td>
<td>9</td>
<td>1</td>
<td>52</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>269</td>
<td>1</td>
<td>203</td>
<td>19</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>10</td>
<td>0</td>
<td>135</td>
<td>7</td>
<td>1</td>
<td>39</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>215</td>
<td>1</td>
<td>192</td>
<td>19</td>
<td>0</td>
<td>60</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Tx</th>
<th>Caff</th>
<th>Smoke</th>
<th>Wt</th>
<th>Anc</th>
<th>Asleep</th>
<th>Mins</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>2</td>
<td>141</td>
<td>1</td>
<td>154</td>
<td>11</td>
<td>1</td>
<td>48</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>296</td>
<td>1</td>
<td>162</td>
<td>23</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>150</td>
<td>0</td>
<td>178</td>
<td>18</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>19</td>
<td>2</td>
<td>137</td>
<td>0</td>
<td>144</td>
<td>10</td>
<td>1</td>
<td>27</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>137</td>
<td>0</td>
<td>191</td>
<td>8</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>21</td>
<td>3</td>
<td>200</td>
<td>1</td>
<td>187</td>
<td>11</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>22</td>
<td>3</td>
<td>149</td>
<td>0</td>
<td>156</td>
<td>11</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>23</td>
<td>3</td>
<td>56</td>
<td>0</td>
<td>161</td>
<td>5</td>
<td>1</td>
<td>22</td>
</tr>
<tr>
<td>24</td>
<td>3</td>
<td>175</td>
<td>0</td>
<td>185</td>
<td>16</td>
<td>0</td>
<td>41</td>
</tr>
<tr>
<td>25</td>
<td>3</td>
<td>145</td>
<td>0</td>
<td>186</td>
<td>15</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>26</td>
<td>3</td>
<td>98</td>
<td>0</td>
<td>140</td>
<td>12</td>
<td>1</td>
<td>33</td>
</tr>
<tr>
<td>27</td>
<td>3</td>
<td>125</td>
<td>1</td>
<td>173</td>
<td>9</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>28</td>
<td>3</td>
<td>10</td>
<td>1</td>
<td>155</td>
<td>6</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>29</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>132</td>
<td>7</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>30</td>
<td>3</td>
<td>223</td>
<td>0</td>
<td>186</td>
<td>12</td>
<td>0</td>
<td>50</td>
</tr>
</tbody>
</table>
Following the data entry procedures described in Chapters 1 and 2, label the first variable \textit{ID}, the second variable \textit{Tx} (1 = Placebo; 2 = OTC; 3 = Treatment) for treatment group, the third variable \textit{Caff} for caffeine intake, the fourth variable \textit{Smk} (1 = Yes; 0 = No) for smoking status, the fifth variable \textit{Wt} for body weight, the sixth variable \textit{Anx} for anxiety, the seventh variable \textit{Asleep} (1 = Yes; 0 = No) to indicate whether the participants fell asleep within the first hour, and the eighth variable \textit{Mins} for minutes required to fall asleep. Enter the data appropriately.

\section*{Assumptions}

The assumptions for a multiple Cox proportional-hazards model include all three assumptions of the Kaplan-Meier method, the four assumptions of simple Cox regression, and two additional assumptions related specifically to the inclusion of multiple independent variables or covariates. Assumptions include no differences between withdrawn and remaining participants, constant probability of survival over the entire study period, that the event occurred at the time specified, independence of residuals, that the continuous covariate assumes a linear form, proportionality of hazards, absence of outliers as well as absence of both multicollinearity and multivariate outliers.

No differences between withdrawn and remaining participants require that participants who are censored (or lost to follow-up) do not differ systematically from those who experience the event. For example, in the data above, a participant who was censored prior to study completion (e.g., ID = 18) may have been censored due to an adverse reaction to the medication. This is not an easily testable assumption because censoring can occur for a variety of reasons. However, a binary logistic regression (Chapters 40 or 41) can be used to evaluate for differences between withdrawn and remaining participants. That is, those who withdrew from the study prior to completion are coded 1, with all remaining participants coded 0. This new binary variable serves as the dependent variable in the logistic regression analysis when assessing for statistically significant differences in the independent variables or covariates. Failure to achieve statistical significance satisfies the assumption. However, remember logistic regression requires large samples, so evaluation of this assumption may not be tenable in all situations—most notably for the example data above.

Having constant probability of survival over the study period requires (1) that the same factors affecting survival are present throughout the entire study period and (2) participants entering the study at different time points have equal probability of survival. For example, a violation of this assumption would occur if the researcher forgot to turn off the lights after dosing the participant or the OTC medication was discontinued requiring substitution with a close, but not identical, medication. In addition, and more appropriate for longer-term studies, additional novel, effective, or ancillary treatments may become available during the study period that require ethical implementation. In these situations, the assumption is tested by examining the survival functions within subsets of participants (i.e., before vs. after differing treatment modalities). This assumption is evaluated by both examining the study design and evaluating the protocol for violations.

Assuring that the event occurred at the time specified can be difficult, especially for left- or interval-censored participants (e.g., participant fell asleep 10-15 minutes after dosing, but the researcher is not exactly sure when). This assumption is satisfied by ensuring proper data collection and data entry. A violation of this assumption tends to bias survival estimates upwards.
Cox regression is a between-subjects analysis; thus, the assumption of independence of residuals is relevant. Independence is technically a design issue and can usually be satisfied by data that is not clustered and by not measuring participants repeatedly.

The first four assumptions mentioned above were mostly assured via proper experimental design and protocol implementation. The remaining assumptions are strictly related to the regression component, with the most important of these being the proportionality of hazards assumption. However, nonlinearity can often result in an erroneous violation of the proportionality of hazards assumption. Thus, evaluating the form of a continuous independent variable or covariate is incredibly important and should be completed prior to evaluating proportionality of hazards. In Cox regression, continuous covariates are assumed to have a linear form, an assumption similar to the linearity in the logit assumption for logistic regression. There are several methods for testing this assumption, but we will focus primarily on the use of Martingale residuals. Martingale residuals are skewed, ranging from negative infinity to 1. Values near 1 indicate the participant experienced the event sooner, whereas large negative residuals indicate the participant experienced the event later or was censored. The most straightforward approach to testing this assumption using Martingale residuals is to estimate a Cox regression model without the continuous independent variable (i.e., a baseline model). Note that Martingale residuals cannot be requested directly from SPSS, so you will need to save the cumulative hazard function from this analysis and compute them using the syntax provided. Next, print a simple scatterplot (Chapter 19) with the Martingale residuals on the y-axis and the previously omitted continuous variable on the x-axis. Finally, request a smoothing line (e.g., loess) to determine the functional form of the independent variable. If the smoothing line appears linear, the assumption is satisfied. However, if nonlinearity is evident, you can include appropriate polynomial terms, use a piecewise or spline regression model, or, finally, transform the variable to make it linear (not recommended).

The proportionality of hazards assumption requires that the hazard ratios remain constant over time. Stated another way, the hazard for one individual is proportional to the hazard for any other individual. A third way of viewing this assumption is easily seen from the Cox regression equation presented in the introduction to this Chapter. Notice in the right side of the equation that the independent variable (i.e., X) does not interact with time (t). Thus, the assumption requires that there is no covariate-by-time interaction. There are several methods for evaluating this assumption, including both graphical and statistical techniques. However, for multiple Cox regression models, we use only the statistical technique. This is primarily due to the ease of implementation and because it lends itself nicely to extended Cox models if a violation of the assumption is indicated.

After the linear form of each continuous independent variable or covariate has been assured, evaluating the proportionality of hazards assumption involves testing for a statistically significant interaction between the independent variable(s) of interest or covariates and time. The assumption will be evaluated for each variable individually, but in one analysis. To use this method, you first create a new time variable that is calculated by taking the natural log of the time variable (e.g., ln(Minutes)). Next, the interaction between each independent variable or covariate and the natural log of time (e.g., Wt*ln(Minutes) from the example data above) as well as the main effect of each independent variable (e.g., Wt) or covariate is included in a Cox regression analysis. If any interaction effect is statistically significant, the variable interacts with time and the proportionality of hazards assumption is violated for this variable. In this situation, you need to include any statistically significant interaction(s) in all subsequent analyses, which
result in a *Cox regression with time-varying covariates* (Chapter 45). This may sound tedious and confusing, but SPSS has a built-in procedure for testing this assumption.

The absence of outliers assumption is evaluated by examining the deviance residuals, which are transformed Martingale residuals that are symmetric around 0, with an approximate standard deviation of 1. Positive values indicate the participant experienced the event sooner and negative values indicate experiencing the event later. Thus, very small (i.e., \(-3.29\)) or very large (i.e., \(3.29\)) *disconnected* values indicate outliers. The absence of outliers assumption is evaluated by creating a simple scatterplot (Chapter 19) with the deviance residuals on the y-axis and each participant (i.e., ID) on the x-axis. Deviance residuals cannot be requested directly in SPSS, but can be computed by requesting Cox-Snell residuals, calculated from the multiple Cox regression analysis, using the syntax provided below.

Because there are multiple independent variables, absence of multicollinearity and absence of multivariate outliers are two additional assumptions required. Both are evaluated using the linear regression procedure described in Chapter 23. Multicollinearity is tested following the procedures described in Chapter 23, Step 3a. Note that it is okay to use a binary dependent variable (e.g., *Asleep* in the example data above) to test this assumption in the linear regression procedure because the tolerance and VIF values used to evaluate multicollinearity are independent of the scale of the dependent variable. For the example data above, the lowest tolerance value was .477 and the highest VIF value was 2.094; thus, the assumption can be considered satisfied.

Finally, the absence of multivariate outliers assumption is assessed by Mahalanobis distance requested and saved through the linear regression procedure described in Chapter 23, Step 5c. Note that you will need to create two dummy variables for the Tx variable when using the linear regression procedure to evaluate multicollinearity. Follow the steps described in *Dummy Variables* section in Chapter 23. Mahalanobis distance approximates a chi-square distribution. Thus, to identify multivariate outliers, find the critical chi-square value with degrees of freedom equal to the number of independent variables and covariates in the analysis using an alpha of .001. For the example data above, there are six independent variables (note that the time variable, Mins, is never included when testing the assumption); thus, with six degrees of freedom and an alpha of .001, the critical chi-square is 22.458. Therefore, any Mahalanobis distance greater than 22.458 would be considered a multivariate outlier. For the example data above, the largest Mahalanobis distance was 10.091, which is far less than 22.458; thus, the assumption can be considered satisfied.

**Evaluating the Form of Continuous Independent Variables and Covariates**

Cox regression assumes the form all continuous independent variables and covariates be linear. This, of course, is a testable empirical question. If a violation of this assumption is indicated, nonlinearity can result in an erroneous violation of the proportionality of hazards assumption. Thus, you should always check this assumption prior to evaluating the proportionality of hazards. To test this assumption we need to calculate the Martingale residuals from a baseline model *excluding* all independent variables and covariates (i.e., a baseline model). After calculating the baseline model, the assumption is evaluated for each independent variable or covariate separately (e.g., Caff, Wt, and Anx) by plotting the Martingale residuals on the y-axis and the continuous independent variables and covariates, separately, on the x-axis.
1. Click **Analyze**, then choose **Survival**, and then click **Cox Regression…** to bring up the *Cox Regression* dialog box.

2. Click to highlight the **Mins** variable, and then click the right arrow ( ) next to the **Time:** box.

3. Click to highlight the **Asleep** variable, and then click the right arrow ( ) next to the **Status:** box.
   
a. Click the **Define Event**… and in the **Single Value:** box, type 1.

4. Click **Continue**.

5. Click the **Save**… button.
   
a. Click the **Hazard function** checkbox to request the Cox-Snell residuals.
   
b. Click **Continue**.

6. Click **OK**.

From Step 6 above, the Cox-Snell residuals were saved as a new variable in the dataset named *HAZ_1*. Note that if you conducted an analysis previously in which you requested Cox-Snell residuals, the variable name might be different (e.g., *HAZ_2*, *HAZ_3*, etc). We assume the name of the variable is *HAZ_1* in this example, but change the name accordingly if needed. To calculate the Martingale residuals, we need to transform the Cox-Snell residuals, a process that requires the syntax commands shown below.

7. Click **File**, then choose **New**, and then click **Syntax**…. A new syntax window will pop-up automatically.

8. Copy and paste (or type) the syntax commands below into the syntax window as shown in Figure 44.1.

   ```
   COMPUTE Mart_1 = (Asleep = 1) - HAZ_1.
   FORMATS Mart_1 (f8.5).
   EXECUTE.
   ```

9. Click **Run** and then click **All**…. 

---

**Figure 44.1**
This syntax will create a new variable in the dataset named *Mart_1*, which represents the Martingale residuals. To test the form of all continuous independent variable(s) and covariates:

10. Click **Graphs**, then choose **Legacy Dialogs**, and then click **Scatter/Dot**.
11. Click **Matrix Scatter** and then click **Define**.
12. Click the **Mart_1** variable and then click the right arrow (→) next to the **Matrix Variables**: box.
13. Click the **Caff** variable, hold the **Ctrl** button on your keyboard, and then click the **Wt** and **Anx** variables. Then, click the right arrow (→) next to the **Matrix Variables**: box.
14. Click **OK**.

When you click **OK**, SPSS will produce an **Output** screen displaying the results. Click the **Output** window to view your results (if it does not pop up automatically). The scatterplot matrix should appear similar to Figure 44.2.

![Figure 44.2](image)

When evaluating this assumption, we are only concerned with the top row that has **Mart_1** on the *y*-axis and each covariate on the *x*-axis of their respective scatterplot. Although we can see subjectively the form is relatively linear for each variable, we need an objective indicator of linearity. Thus, we will employ a smooth curve known as a **loess** line. The loess line is a locally weighted regression line that is generally smooth with its location and direction determined by the percentage of data points in its *vicinity*. For this example, the loess line will fit approximately 50% of data points.

15. In the **Output** window, double click the scatterplot to bring up the **Chart Editor** window, shown in Figure 44.3.
16. Click **Elements** and then click **Fit Line at Total** to bring up the **Properties** dialog box.
17. Click the **Loess** radio button, type 75 into the % of points to fit: box, and then click the **Apply** button.
18. Click **Close**.
19. Close the **Chart Editor** window either by clicking **File** and then clicking **Close**, or by clicking ✗ in the upper right hand corner of the window.
Whew, that was a lot of steps! Based on the loess lines shown in the top row of Figure 44.3, we can see that Martingale residuals are linear across all values of the covariates. Thus, we can consider them to each have a linear form.

**Evaluating the Proportionality of Hazards Assumption**

For multiple Cox regression, only the statistical method will be used to evaluate the proportionality of hazards assumption. The statistical method tests directly whether any independent variable(s) or covariates interact with time.

1. Click **Analyze**, then choose **Survival**, and then click **Cox w/ Time-Dep Cov…** to bring up the *Compute Time-Dependent Covariate* dialog box, shown in Figure 43.2 found in the previous Chapter.

2. The dialog box is used specifically to transform the time variable. You will notice the first variable listed on the left hand side of the dialog box is \( T_\). This is an internal time variable created by SPSS to be used for all time transformations. To test the proportionality of hazards assumption, you need to use the natural log of time.

   a. In the *Expression for T_COV_* box, type: \( \ln(T)_\).

   b. Click the **Model…** button on the upper right corner of the dialog box to bring up the *Cox Regression* dialog box.

3. Click the **Mins** variable and then click the right arrow (\( \rightarrow \)) next to the *Time*: box.

4. Click to highlight the **Asleep** variable and then click the right arrow (\( \rightarrow \)) next to the *Status*: box.

   a. Click the **Define Event…** button to bring up the *Cox Regression: Define Event for Status Variable* dialog box.

      i. In the **Single Value**: box type 1 in this box and then click **Continue**.
5. Click the Tx variable, hold the Ctrl button on your keyboard, and then click the Caff, Smk, Wt, Anx variables. With all five variables highlighted, click the right arrow ( ) next to the Covariates: box.

6. Click the Tx variable, press and hold the Ctrl button on your keyboard, and then click the T_COV_[T_COV_] variable. With both variables highlighted, click the >a*b< button ( ) next to the Covariates: box. This will place the interaction term in this box, which should appear as T_COV_ *Wt. Follow these same steps for the Caff, Smk, Wt, and Anx variables. If completed correctly, the dialog box should contain five interaction terms.

7. Click the Categorical button.
   a. Click the Tx variable and then click the right arrow ( ) next to the Categorical Covariates: box.
   b. Note that while it does not matter what the reference group is when testing this assumption, we will set the reference category as we would in the primary analysis. In the Change Contrast box, click the First radio button, and then click the Change button.
   c. Click Continue.

8. That’s it! Click OK to see the results.

When you click OK, SPSS will produce an Output screen displaying the results. Click the Output window to view your results (if it does not pop up automatically). Scroll down to find the Variables in the Equation table within the Block 1: Method = Enter section. This table is shown in Figure 44.4. To evaluate the assumption you need to verify that no variable-by-time interaction is statistically significant. For the example data above, none of these interactions are statistically significant indicating the assumption is satisfied. However, had any interaction been statistically significant, you would need to include the interaction(s) in all subsequent analyses, with interpretation described in Chapter 45.

<table>
<thead>
<tr>
<th>Variables in the Equation</th>
<th>B</th>
<th>SE</th>
<th>Wald</th>
<th>df</th>
<th>Sig</th>
<th>Exp(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tx</td>
<td></td>
<td></td>
<td>1.882</td>
<td>2</td>
<td>.390</td>
<td></td>
</tr>
<tr>
<td>T_COV_Tx</td>
<td>-13.374</td>
<td>12.237</td>
<td>1.185</td>
<td>1</td>
<td>.276</td>
<td>.900</td>
</tr>
<tr>
<td>T_COV_Tx(1)</td>
<td>23.823</td>
<td>34.274</td>
<td>.483</td>
<td>1</td>
<td>.487</td>
<td>2.220E10</td>
</tr>
<tr>
<td>Caff</td>
<td>.056</td>
<td>.073</td>
<td>.848</td>
<td>1</td>
<td>.391</td>
<td>.391</td>
</tr>
<tr>
<td>Smk</td>
<td>-4.669</td>
<td>4.341</td>
<td>1.131</td>
<td>1</td>
<td>.287</td>
<td>.090</td>
</tr>
<tr>
<td>Wt</td>
<td>.121</td>
<td>.190</td>
<td>.405</td>
<td>1</td>
<td>.524</td>
<td>.908</td>
</tr>
<tr>
<td>Anx</td>
<td>.513</td>
<td>.731</td>
<td>.692</td>
<td>1</td>
<td>.408</td>
<td>1.870</td>
</tr>
<tr>
<td>T_COV_Tx</td>
<td></td>
<td></td>
<td>1.512</td>
<td>2</td>
<td>.470</td>
<td></td>
</tr>
<tr>
<td>T_COV_Tx(1)</td>
<td>3.248</td>
<td>3.329</td>
<td>.992</td>
<td>1</td>
<td>.327</td>
<td>25.816</td>
</tr>
<tr>
<td>T_COV_Tx(2)</td>
<td>-9.618</td>
<td>10.655</td>
<td>.386</td>
<td>1</td>
<td>.534</td>
<td>.001</td>
</tr>
<tr>
<td>Caff*T_COV</td>
<td>-0.019</td>
<td>.021</td>
<td>.797</td>
<td>1</td>
<td>.369</td>
<td>.492</td>
</tr>
<tr>
<td>Smk*T_COV</td>
<td>1.113</td>
<td>1.295</td>
<td>.739</td>
<td>1</td>
<td>.390</td>
<td>3.044</td>
</tr>
<tr>
<td>T_COV_Wt</td>
<td>.012</td>
<td>.054</td>
<td>.052</td>
<td>1</td>
<td>.820</td>
<td>1.012</td>
</tr>
<tr>
<td>Anx*T_COV</td>
<td>-.329</td>
<td>.266</td>
<td>1.532</td>
<td>1</td>
<td>.216</td>
<td>.720</td>
</tr>
</tbody>
</table>

Figure 44.4
Analysis

With the proportionality of hazards assumption satisfied, the continuous covariates shown to have a linear form, no multicollinearity or multivariate outliers evident, and assuming no differences between withdrawn and remaining participants, constant probability of survival over the entire study period, and that the event occurred at the time specified, to conduct a multiple Cox regression:

1. Click **Analyze**, choose **Survival**, and then click **Cox Regression**… to bring up the **Cox Regression** dialog box, shown in Figure 44.5.

![Figure 44.5](image)

2. Click **Reset** to clear any previous selections.

3. You will immediately see all of the variables in your dataset listed on the left hand side of the dialog box. You will be concerned primarily with three boxes labeled **Time:**, **Status:**, and **Covariates:**. The **Time:** box is fairly self-explanatory. Here, you will place the variable representing time-to-event. The **Status:** box is essentially the censoring variable. That is, the variable that identifies who experienced the event or not. Finally, the **Covariates:** box is where the categorical independent variable will be placed. In addition, the **Strata:** box is available, but is not used for simple Cox regression.
   
a. Click to highlight the **Mins** variable, and then click the right arrow (→) next to the **Time:** box.

   b. Click to highlight the **Asleep** variable, and then click the right arrow (→) next to the **Status:** box.

   i. Click the **Define Event**… button to bring up the **Cox Regression: Define Event for Status Variable** dialog box. In this dialog box, you have three options.

      1. First, you can specify a **Single Value**. Although you may have multiple values coded for various events, the use of this box allows you to specify one value as the event and all other values as non-events.

         a. For the example data above, type 1 in this box.
2. Alternatively, you can specify a *Range of values*:. This option is useful if you have multiple values coded for various events, and you are interested in two or more events. However, note that the events must be coded in numerical order. That is, say you are interested only in the events coded 1 and 3, but not event 2. By specifying a range of values, you would necessarily include the event coded 2, which is not of interest to you.

3. Alternatively, you can specify a *List of values*:. This option can be used if you have any number of events, coded in any numerical order. Thus, if you are interested in events coded 1 and 3, you can specify each event individually by clicking the radio button, typing the event number of interest and clicking *Add*.

When you have specified the event(s) of interest, click **Continue**.

   c. Click to highlight the **Caff** variable, hold the **Ctrl** button on your keyboard, and then click the **Smk, Wt, Anx** variables. With all four variables highlighted, click the right arrow ( ) next to the **Covariates:** box.

   d. Click to highlight the **Tx** variable and then click the right arrow ( ) next to the **Covariates:** box.

4. Click the **Categorical…** button. Here, all variables you entered in the **Covariates:** box in Steps 3c and 3d above are shown on the left hand side. You will use this option when you have *any* categorical independent variable or covariate. Not specifying a categorical variable as categorical will only influence your results if you have three or more categories. However, failing to specify variables as categorical will prevent you from print separate survival or hazard functions for each group (Step 5 below). In addition, the categorical option removes the need for you to create new dummy variables for variables consisting of three or more groups as you did when conducting linear regression.

   a. Click to highlight the **Tx** variable, hold the **Ctrl** button on your keyboard and then click the **Smk** variable. Click the right arrow ( ) next to the **Categorical Covariates:** box to move both variables into the box.

   b. Notice the *Change Contrast* section is now available. There are seven *Contrast:* available from the drop down list. These include:

      i. **Indicator**: Compares the presence (i.e., 1) or absence (i.e., 0) of the independent variable to a reference category. Note that **Indicator** is a synonym for dummy coding (see Chapter 23) and is the option we want to use for this example.

      ii. **Simple**: Each category of the independent variable is compared to the unweighted average of all categories.

      iii. **Difference**: Each category of the independent variable, except lowest coded category, is compared to the average effect of the previous categories. Also known as reverse Helmert contrasts, because…

      iv. **Helmert**: Each category of the independent variable, except highest coded category, is compared to the average effect of the previous categories.
v. **Repeated**: Each category of the independent variable, except the lowest coded category, is compared to previous category (e.g., 2 to 1; 3 to 2, etc).

vi. **Polynomial**: Available only for continuous independent variables; uses orthogonal polynomial contrasts when categories are equally spaced. Think, trend-type analysis.

vii. **Deviation**: Each category of the independent variable, except the reference category, is compared to the unweighted (i.e., mean) overall effect.

c. For several of the options listed above you need to identify the **Reference Category**. You can select either the **First** or the **Last** category as reference. If you are more interested in a middle category, the variable will need to be recoded (see Chapter 8).

When you are satisfied with your selections, click **Continue**.

5. Click the **Plots** button. Here, you can select the **Plot Type** you want to print. There are several options and each is listed below.

   a. **Survival**: Selecting this option will print the cumulative survival function on the original linear time scale.

   b. **Hazard**: Selecting this option prints the cumulative hazard function on the original linear time scale.

   c. **Log minus log**: Selecting this option prints the cumulative survival function after the ln-ln transformation.

   d. **One minus survival**: Selecting this option prints the one-minus survival function on the original linear time scale.

In the **Covariate Values Plotted at:** section, click to highlight the **Tx (Cat) (Mean)** variable and then click the right arrow ( ) next to the **Separate Lines for:** box. This will provide the survival and hazard functions for each of the three treatment groups individually on the same plot for comparison. When you are satisfied with your selections, click **Continue**.

6. Click the **Save** button to bring up the **Cox Regression: Save** dialog box. Selecting any of these options will create a new variable in your dataset that can be used to test hypotheses or evaluate certain assumptions.

   a. The **Survival function** option saves the probability of survival for each participant given the duration of the study.

   b. The **Standard error of survival function** option saves the standard error of the survival function in the option above.

   c. The **Log minus log survival function** option saves the cumulative survival function for each participant after the ln-ln transformation has been applied.

   d. The **Hazard function** option saves the cumulative hazard estimate, also known as the Cox-Snell residuals. These residuals are used to calculate Martingale and Deviance residuals.
e. The **Partial Residuals** option saves the partial residuals, also known as Schoenfeld residuals, for each independent variable in the model.

f. The **DfBeta(s)** option saves the estimated change in the coefficient value if the participant was removed from analysis. A new variable is saved for each independent variable in the model.

g. The **X * Beta** option saves the estimated linear predictor score. It is the sum of the product of the mean-centered covariate and the corresponding parameter estimates for each participant.

Finally, you can **Export Model Information to XML File** by identifying a previously created XML file via the **Browse**... button. When you are satisfied with your selections, click **Continue**.

7. Click the **Options**... button to bring up the **Cox Regression: Options** dialog box. Here, you can select several **Model Statistics** you want printed as well as information regarding the required **Probability for Stepwise** entry or removal.

e. Under the **Model Statistics** section, you can choose to print the confidence interval for the hazard ratios (**CI for exp(B)**), as well as set the confidence interval level, with the default being **95%**. You can also print the **Correlation of estimates**.

Finally, you have the choice to **Display Model Information At each step** or **At last step**.

f. Under the **Probability for Stepwise** section, you can indicate how you want SPSS to use **p-values** to enter (**Entry:**) or remove (**Removal:**) variables in a stepwise regression analysis. Because we are not conducting stepwise regression in the example, these options can be ignored.

g. Because Cox regression uses partial likelihoods, it uses an iterative estimation process to maximize the likelihood that the estimates are correct. You can set the number of **Maximum Iterations:** which essentially controls how long the procedure will search for a solution.

h. Finally, you can **Display baseline function**, which prints the baseline hazard function and cumulative survival at the mean of the covariates. Note that this option is not available if you have time-varying covariates (Chapter 45).

When you are satisfied with your selections, click **Continue**.

8. That’s it! Click **OK** to conduct the analysis.

**Output**

When you click **OK**, SPSS will produce an **Output** screen displaying the results. Click the **Output** window to view your results (if it does not pop up automatically). The first table you see is titled **Case Processing Summary**, which contains important information about your sample including the frequency and percentage of participants experiencing the event (**Event**') and **Censored**, the **Total** number of participants, as well as participants with missing values, negative time, and those that were left-censored.
Next, you are presented with a table titled *Categorical Variable Codings*, presented in Figure 44.6. This table is incredibly important as it informs you how your categorical variables were coded and which group is serving as the reference. Remember, the group coded 0 across the dummy variables is your reference group (i.e., $3=\text{Treatment}$ for $Tx$ and $1=\text{Yes}$ for $Smk$).

<table>
<thead>
<tr>
<th>Categorical Variable Codings&lt;sup&gt;a-d&lt;/sup&gt;</th>
<th>Frequency</th>
<th>(1)&lt;sup&gt;b&lt;/sup&gt;</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Tx^a$</td>
<td>10</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1=Placebo</td>
<td>10</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2=QTC</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3=Treatment</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$Smk^a$</td>
<td>14</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>0=No</td>
<td>16</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1=Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Indicator Parameter Coding  
b. The (0,1) variable has been recoded, so its coefficients will not be the same as for indicator (0,1) coding  
c. Category variable: $Tx$  
d. Category variable: $Smk$

**Figure 44.6**

The next section of the output is titled *Block 0: Beginning Block*. This block can be thought of as similar to that from logistic regression. That is, it contains information for the model without any independent variable(s) or covariates, and serves as the comparison model. The only table in this section is titled *Omnibus Tests of Model Coefficients*, which contains the $-2 \log \text{Likelihood}$ information which is used to calculate the $G^2$ statistic (i.e., likelihood-ratio test).

The *Block 1: Method=Enter* section contains the results of the multiple Cox regression analysis. The first table, titled *Omnibus Tests of Model Coefficients* and shown in Figure 44.7, contains the overall test of your model, and is a statistical comparison between Block 0 and Block 1. This table contains the $-2 \log \text{Likelihood}$ value as well as omnibus tests for the Overall (score) model as well as Change from Previous Step and Change from Previous Block. Note that the latter two produce identical results because we did not use hierarchical, sequential, or stepwise Cox regression, and that the Chi-square value from the Change From Previous Block section is used when calculating the effect size.

**Figure 44.7**
The next table is titled *Variables in the Equation* and is shown in Figure 44.8. This is the most important table in the output because it contains the results of the independent variable after controlling for the various covariates. Notice there is no constant, as in both linear and logistic regression, because the baseline hazard function is unspecified. The table contains the regression slope ($B$), standard error ($SE$), Wald statistic, degrees of freedom ($df$), $p$-value ($Sig.$), hazard ratio ($Exp(B)$) as well as the 95% confidence interval around the hazard ratio ($95.0\% CI \text{ for } Exp(B)$).

<table>
<thead>
<tr>
<th></th>
<th>$B$</th>
<th>$SE$</th>
<th>Wald</th>
<th>$df$</th>
<th>$Sig.$</th>
<th>$Exp(B)$</th>
<th>95.0% CI for Exp($B$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0</td>
<td>-0.004</td>
<td>0.008</td>
<td>0.223</td>
<td>1</td>
<td>0.633</td>
<td>0.996</td>
<td>0.980 - 1.012</td>
</tr>
<tr>
<td>Smk</td>
<td>0.728</td>
<td>0.778</td>
<td>0.878</td>
<td>1</td>
<td>0.368</td>
<td>2.072</td>
<td>0.451 - 9.523</td>
</tr>
<tr>
<td>Wt</td>
<td>0.045</td>
<td>0.027</td>
<td>2.703</td>
<td>1</td>
<td>0.100</td>
<td>1.056</td>
<td>0.006 - 1.009</td>
</tr>
<tr>
<td>Arx</td>
<td>-0.259</td>
<td>0.168</td>
<td>2.362</td>
<td>1</td>
<td>0.125</td>
<td>0.772</td>
<td>0.554 - 1.075</td>
</tr>
<tr>
<td>Tx</td>
<td>7.498</td>
<td></td>
<td></td>
<td>2</td>
<td>0.024</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tx(1)</td>
<td>-1.830</td>
<td>0.911</td>
<td>4.065</td>
<td>1</td>
<td>0.044</td>
<td>0.159</td>
<td>0.027 - 0.960</td>
</tr>
<tr>
<td>Tx(2)</td>
<td>-2.434</td>
<td>0.948</td>
<td>6.595</td>
<td>1</td>
<td>0.010</td>
<td>0.088</td>
<td>0.014 - 0.562</td>
</tr>
</tbody>
</table>

**Figure 44.8**

Next, you are presented a table titled *Covariate Means and Pattern Values* and two plots titled *Survival Function at mean of covariates* and *Hazard Function at mean of covariates*. These plots present the survival or hazard function for a hypothetical participant who is at the mean of all independent variables and covariates. These values are shown in the *Mean* column of the table.

Finally, the *Survival Function for patterns 1 – 3* and *Hazard Function for patterns 1 – 3* are shown in Figure 44.9. These plots are interpreted similarly to the Kaplan-Meier curve. Notice that in both plots, the treatment group is experiencing the event (i.e., falling asleep) at a greater rate than the other two groups.

**Figure 44.9**
Interpretation

Interpretation begins by evaluating the absence of outliers assumption, which requires the calculation of deviance residuals. Because SPSS does not calculate deviance residuals directly, we need to calculate the residual values based on the Cox-Snell residuals we requested in Step 6d above. This process requires syntax commands. In the dataset, a new variable was saved named HAZ_2 (note that HAZ_1 was used when testing the form of continuous variables above). If you requested additional Cox-Snell residuals previously, the variable name might be different (e.g., HAZ_3, HAZ_4, etc). If this is the case, simply change the name in the syntax commands.

1. Click **File**, then choose **New**, and then click **Syntax**.... A new syntax window will pop-up automatically.
2. Copy and paste (or type) the syntax commands below into the syntax window as shown in Figure 44.10 on the next page.

   ```
   COMPUTE Mart_2 = (Asleep = 1) - HAZ_2.
   COMPUTE Dev_2 = SQRT(-2*(Mart_2+(Asleep=1)*LN((Asleep=1)-Mart_2))).
   IF Mart_2 < 0 Dev_2 = -Dev_2.
   FORMATS Mart_2 Dev_2 (f8.5).
   EXECUTE.
   ```

3. Click **Run** and then click **All**....

Two new variables are now created in the dataset named Mart_2 and Dev_2. I used the name Dev_2 to keep the names consistent and to remind you from which Martingale residuals the deviance residuals were calculated. To evaluate for outliers, we simply print a scatterplot of the deviance residual against the participant’s ID variable. That is, plot the deviance residual for each participant.

4. Click **Graphs**, then choose **Legacy Dialogs**, and then click **Scatter/Dot**.
5. Click **Simple Scatter** and then click **Define**.
6. Click the Dev_2 variable and then click the right arrow (►) next to the **Y Axis** box.
7. Click the ID variable and then click the right arrow (►) next to the **X Axis** box.
8. Click **OK**.
When you click OK, SPSS will produce an Output screen displaying the results. Click the Output window to view your results (if it does not pop up automatically). The scatterplot should appear as shown in Figure 44.11. No deviance residuals are extremely large (e.g., > 3.29 or < -3.29) and no residual values are disconnected. Thus, we can consider the assumption satisfied.

With all assumptions satisfied, we can now begin interpreting the results of the multiple Cox regression analysis. Interpretation begins by evaluating the Omnibus Tests of Model Coefficients table in the Block 1: Method=Enter section shown in Figure 44.7. Consider only the Change From Previous Block columns. The Chi-square value of 33.479 is statistically significant as \( p < .0005 \). Note that this value was calculated as the difference between the -2 Log Likelihood from Block 0: Beginning Block (i.e., 86.990) and the -2 Log Likelihood from Block 1: Method=Enter (i.e., 53.511) sections. This value, termed the likelihood ratio statistic (or \( G^2 \)) is used when calculating the effect size estimate. As stated above, the effect size estimate is calculated as: \( 1 - \exp(-G^2/n) \). Thus, the \( R^2 \) estimate equals .672 (i.e., \( 1 - \exp(-33.479/30) \)), which represents the relative association between survival time and the independent variable and covariates as a set. In addition, degrees of freedom are calculated as the difference in the number of parameters between models; that is, the model in Block 1 has six more parameters than the Block 0.

Next, you evaluate the effect of the independent variable, which is found in the Variables in the Equation table. Evaluate the regression coefficient (\( B \)) and standard error (\( S.E. \)) as well as the hazard ratio (\( \exp(B) \)) and confidence interval (95.0\% C.I.for \( \exp(B) \)) for the \( Tx \) variable. Large standard errors and wide confidence intervals indicate poor measurement. In addition, a 95\% confidence interval that contains 1 indicates the variable is not statistically significant.

The interpretation of log-hazards are similar to log-odds in logistic regression and slopes in linear regression—a one-unit increase in the independent variable results in an increase (+) or decrease (-) in the log-hazards of experiencing the category coded 1 on the dependent variable. Here, because the independent variable is categorical (see Chapter 43 for interpretation when you have a continuous independent variable), a one-unit increase indicates going from either the new sleep treatment group (coded 0) to the placebo group (coded 1) or the new sleep treatment group (coded 0) to the OTC group (coded 1). Thus, being in the group receiving the placebo treatment...
resulted in a 1.836 decrease in the log-hazards of falling asleep within one hour compared to the new sleep treatment group. Additionally, being in the OTC group resulted in a 2.434 decrease in the log-hazards of falling asleep within one hour compared to the new sleep treatment group.

When interpreting hazard ratios, you can indicate which group is more likely to experience the event. This is primarily due to the fact that hazard ratios are based on risk, which are based on the number of participants not experiencing the event up to a given time point (i.e., a rate ratio). Thus, from Figure 44.8, the hazard ratio of .159 for the placebo group indicates that participants who received placebo were 84.1% (i.e., \([1 - .159]\times 100\)) less likely to fall asleep within one hour compared to those who received the new sleep treatment. Alternatively, those who received the OTC treatment were 91.2% (i.e., \([1 - .088]\times 100\)) less likely to fall asleep within one hour compared to those who received the new sleep treatment.

Finally, the survival and hazard functions presented in Figure 44.9 should be provided in the manuscript, with the choice of which plot you present to your audience being completely up to you. Be sure to note what value of the independent variable and covariates was used when calculating the plots. The interpretation of the plots differs slightly for survival and hazard functions. The survival function ranges from 0 to 1. A value of 0 indicates everyone has experienced the event (or been censored) at a specific time point, whereas a value of 1 indicates no one has experienced the event at a specific time point. It is always important to consider the time aspect, as the survival function represents the probability of not experiencing the event in the next time interval, which for this example is the next minute. Alternatively, the hazard function ranges from 0 to infinity, with greater values indicating greater likelihood of experiencing the event in the next time unit.

**Example Results Section**

Prior to analysis, no differences between withdrawn and remaining participants were indicated and the probability of survival was assumed constant over the entire study period. Further, the linear form of all continuous covariates was assured by examining Martingale residuals and no outliers were indicated via deviance residuals. The proportionality of hazards assumption was assured for all variables as no statistically significant variable-by-time interaction was indicated. In addition, no multivariate outliers were indicated by Mahalanobis distance using critical chi-square of 22.458 (p < .001). Finally, no multicollinearity was evident using tolerance and VIF values.

The results of a multiple Cox regression analysis indicated the independent variable and covariates, as a set, significantly predicted time to fall asleep during the 60-minute study period, \(\chi^2\) = 33.479, \(p < .05\), \(R^2 = .672\). More specifically, a statistically significant association between time to fall asleep and treatment group was indicated. Participants receiving the placebo treatment had a 1.836 decrease in the log-hazards and were 84.1% less likely to fall asleep during the study period compared to participants receiving the new sleep treatment after adjusting for caffeine intake, smoking status, body weight, and anxiety. Further, participants receiving the OTC treatment had a 2.434 decrease in the log-hazards and were 91.2% less likely to fall asleep during the study period compared to participants receiving the new sleep treatment after adjusting for caffeine intake, smoking status, body weight, and anxiety. This effect is evidenced by both the survival and hazard plots.
Chapter 45
SIMPLE COX REGRESSION WITH A TIME-VARYING COVARIATE

As stated in Chapters 43 and 44, a violation of the proportionality of hazards assumption indicates that the effect of the independent variable (or covariate) interacts with time. That is, the risk of experiencing the event is \textit{conditional} on the time when risk was calculated. The interaction effect is characterized differently depending on the scale of the independent variable.

For both categorical and continuous independent variables or covariates, you need to include the variable-by-time interaction in your Cox model. The interpretation of the interaction effect, then, is how the passage of time directly influences the main effect of the independent variable. Interpretation of interaction effects in a regression model is slightly different from those discussed in ANOVA. In ANOVA, it is often recommended that you should not interpret main effects given a statistically significant interaction effect. This was the procedure described in Chapter 30 and is primarily due to the use of marginal effects in ANOVA. In a regression analysis, however, main effects can and should be interpreted in the presence of a statistically significant interaction effect. However, care must be taken, as the correct way to interpret main effects is \textit{conditional on their interacting predictor}. That is, the main effect is now interpreted as the \textit{simple (conditional) main effect} when the interacting predictor is 0. The interaction effect, then, only serves to modify the simple main effects. That is, the interaction effect can adjust the simple main effect to be more positive, less positive, more negative, or less negative. This probably seems confusing, and it can be, but as you progress through the Chapter these four alternatives should become clearer. I promise you once you see the pattern interaction effects have on main effects you cannot un-see it!

As an example of a categorical independent variable, say you want to know whether participants who take your new sleep treatment fall asleep faster during a 60-minute study period compared to participants who take placebo. As an example of a continuous independent variable, say you want to know whether body weight in pounds significantly predicts how fast participants fall asleep during the 60-minute study period. The collected data is presented below:

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|c|c|}
\hline
\textbf{ID} & \textbf{Tx} & \textbf{Wt} & \textbf{Asleep} & \textbf{Mins} & \textbf{ID} & \textbf{Tx} & \textbf{Wt} & \textbf{Asleep} & \textbf{Mins} & \textbf{ID} & \textbf{Tx} & \textbf{Wt} & \textbf{Asleep} & \textbf{Mins} \\
\hline
1 & 0 & 130 & 1 & 35 & 11 & 0 & 184 & 0 & 60 & 21 & 1 & 122 & 1 & 26 \\
2 & 0 & 165 & 0 & 60 & 12 & 0 & 160 & 1 & 56 & 22 & 1 & 129 & 1 & 12 \\
3 & 0 & 205 & 1 & 12 & 13 & 0 & 193 & 0 & 60 & 23 & 1 & 197 & 1 & 9 \\
4 & 0 & 155 & 0 & 60 & 14 & 0 & 125 & 1 & 44 & 24 & 1 & 201 & 0 & 60 \\
5 & 0 & 179 & 1 & 36 & 15 & 0 & 205 & 0 & 60 & 25 & 1 & 150 & 1 & 13 \\
6 & 0 & 131 & 1 & 51 & 16 & 1 & 144 & 1 & 24 & 26 & 1 & 187 & 0 & 60 \\
7 & 0 & 181 & 0 & 60 & 17 & 1 & 135 & 1 & 17 & 27 & 1 & 170 & 1 & 25 \\
8 & 0 & 157 & 1 & 48 & 18 & 1 & 199 & 1 & 19 & 28 & 1 & 175 & 0 & 60 \\
9 & 0 & 215 & 0 & 60 & 19 & 1 & 235 & 0 & 60 & 29 & 1 & 165 & 1 & 11 \\
10 & 0 & 150 & 0 & 60 & 20 & 1 & 195 & 0 & 60 & 30 & 1 & 159 & 0 & 60 \\
\hline
\end{tabular}
\caption{Data for Simple Cox Regression with Time-Varying Covariate}
\end{table}

Following the data entry procedures described in Chapters 1 and 2, label the first variable \textit{ID}, the second variable \textit{Tx} (1 = Treatment; 0 = Placebo) for treatment group, the third variable \textit{Wt} for body weight, the fourth variable \textit{Asleep} (1 = Yes; 0 = No) to indicate whether the participants fell asleep within the first hour, and the final variable \textit{Minutes} for minutes required to fall asleep. Enter the data appropriately.
Assumptions

Aside from the proportionality of hazards assumption, which this Chapter assumes is violated (if not, use procedures described in Chapters 43 and 44), the assumptions for a Cox model with time-varying covariates are identical to a simple Cox regression without time-varying covariates. These assumptions include no differences between withdrawn and remaining participants, constant probability of survival over the entire study period, that the event occurred at the time specified, independence of residuals, that the continuous covariate assumes a linear form, and absence of outliers. All of these assumptions were described in detail in Chapters 43 and 44; thus, only brief descriptions of previously defined assumptions are provided here.

No differences between withdrawn and remaining participants require that participants who are censored (or lost to follow-up) do not differ systematically from those who experience the event. This assumption is testable using a binary logistic regression (Chapters 40 or 41) to evaluate for differences between withdrawn and remaining participants. Having constant probability of survival over the study period requires (1) that the same factors affecting survival are present throughout the entire study period and (2) participants entering the study at different time points have equal probability of survival. Assuring that the event occurred at the time specified can be difficult, especially for left- or interval-censored participants (e.g., participant fell asleep 10-15 minutes after dosing, but the researcher is not exactly sure when). The assumption of independence of residuals is technically a design issue and can usually be satisfied by data that is not clustered and by not measuring participants repeatedly.

The first four assumptions mentioned above were mostly assured via proper experimental design and protocol implementation. The remaining assumptions are strictly related to the regression component. In Cox regression, continuous covariates are assumed to have a linear form, an assumption similar to the linearity in the logit assumption for logistic regression. There are several methods for testing this assumption, but we have focused primarily on the use of Martingale residuals. This assumption is tested by estimating a Cox regression model without the continuous independent variable (i.e., a baseline model), plotting the Martingale residuals on the $y$-axis with the previously omitted continuous variable on the $x$-axis, and finally, requesting a loess line to determine the functional form of the variable. The absence of outliers assumption was evaluated by plotting deviance residuals for each participant; however, for a time-varying covariate, the residuals are also time-dependent. Thus, they cannot be plotted and the assumption is left untestable.

Categorical Independent Variable

Evaluating the Proportionality of Hazards Assumption

A violation of the proportionality of hazards assumption is only reason why you would ever need to consider treating the independent variable as time varying. As stated in Chapter 43, for a categorical independent variable there are two methods for testing the proportionality of hazards assumption—one graphical, one statistical. I will not go into detail discussing the procedure to test the assumption (see Chapter 43), but the results are presented below in Figure 45.1. In the vast majority of situations, the methods should produce similar results, and while the results below are no exception, there is some subjectivity. You can see in the top of Figure 45.1 that the ln-ln survival functions are relatively parallel, but do exhibit some divergence. The
statistical method confirms this divergence as indicated by the statistically significant interaction between the treatment group variable and time. Reasonable people will often disagree about what steps to take from here; however, we will proceed based on the results of the statistical method, and consider the $Tx$ variable to be time varying.

Figure 45.1

<table>
<thead>
<tr>
<th>Variables in the Equation</th>
<th>B</th>
<th>SE</th>
<th>Wald</th>
<th>df</th>
<th>Sig.</th>
<th>Exp(B)</th>
<th>95.0% CI for Exp(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Tx$</td>
<td>11.296</td>
<td>5.536</td>
<td>4.164</td>
<td>1</td>
<td>.041</td>
<td>0.0516857</td>
<td>1.562</td>
</tr>
<tr>
<td>T_COV_*$Tx$</td>
<td>-3.276</td>
<td>1.684</td>
<td>3.876</td>
<td>1</td>
<td>.049</td>
<td>.038</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Analysis

With the proportionality of hazards assumption violated, but assuming no differences between withdrawn and remaining participants, constant probability of survival over the entire study period, that the event occurred at the time specified, and independence of residuals, to conduct a simple Cox regression with a categorical, time-varying independent variable follow the steps below. Note that several options are not available when using time-varying covariates. This results specifically from the inclusion of the variable-by-time interaction.

1. Click **Analyze**, then choose **Survival**, and then click **Cox w/ Time-Dep Cov...** to bring up the **Compute Time-Dependent Covariate** dialog box.

2. The initial dialog box is used specifically to transform the time variable. You will notice that all the variables in your dataset are provided on the left hand side of this dialog box. However, notice that the first variable is new and listed as $Time [T_]$. This is an internal time variable created by SPSS to be used for all time related transformations. Similar to when we tested the proportionality of hazards assumption, use the natural log of time.

   a. In the **Expression for T_COV_**; box, type: $\ln(T_)$.

   b. Click the **Model...** button on the right hand side of the dialog box to bring up the **Cox Regression** dialog box.
3. You will immediately see all of the variables in your dataset listed on the left hand side of the dialog box. You will be concerned primarily with three boxes labeled \( \text{Time} \), \( \text{Status} \), and \( \text{Covariates} \). The \( \text{Time} \) box is fairly self-explanatory. Here, you will place the dependent variable that measures time-to-event. The \( \text{Status} \) box is essentially the censoring variable. That is, the variable that identifies who experienced the event or not. Finally, the \( \text{Covariates} \) box is where the categorical independent variable and variable-by-time interaction will be placed. In addition, the \( \text{Strata} \) box is available, but is not used for simple Cox regression.

a. Click to highlight the **Mins** variable, and then click the right arrow ( ) next to the \( \text{Time} \) box.

b. Click to highlight the **Asleep** variable, and then click the right arrow ( ) next to the \( \text{Status} \) box.

i. Click the **Define Event**… button to bring up the *Cox Regression: Define Event for Status Variable* dialog box. In this dialog box, you have three options.

1. First, you can specify a **Single Value**. Although you may have multiple values coded for various events, the use of this box allows you to specify one value as the event and all other values as non-events.

   a. For the example data above, type 1 in this box.

2. Alternatively, you can specify a **Range of values**. This option is useful if you have multiple values coded for various events, and you are interested in two or more events. However, note that the events must be coded in numerical order. That is, say you are interested only in the events coded 1 and 3, not event 2. By specifying a range of values, you would necessarily include the event coded 2, which is not of interest to you.

3. Alternatively, you can specify a **List of values**. This option can be used if you have any number of events, coded in any numerical order. Thus, if you are interested in events coded 1 and 3, you can specify each event individually by clicking the radio button, typing the event number of interest and clicking **Add**.

When you have specified the event(s) of interest, click **Continue**.

4. Click to highlight the **Tx** variable, and then click the right arrow ( ) next to the \( \text{Covariates} \) box.

5. Click the **Tx** variable, press and hold the **Ctrl** button on your keyboard, and then click the \( \text{T_COV}[\text{T_COV}] \) variable. With both variables highlighted, click the \( \text{>}\*\text{b}> \) button ( ) next to the \( \text{Covariates} \) box. This will place the interaction term in this box, which should appear as \( \text{T_COV}_*\text{Tx} \).

6. Click the **Categorical…** button. Here, only the variables you entered in the \( \text{Covariates} \) box in Steps 5 and 6 above are shown on the left hand side. You will use this option when you have *any* categorical independent variable or covariate. Not specifying the
independent variable as categorical will not influence your results with only two groups. In addition, the categorical option removes the need for you to create new dummy variables as you did for linear regression for variables consisting of three or more groups.

a. Click to highlight the $\text{Tx}$ variable and then click the right arrow (►) next to the Categorical Covariates: box.

b. Notice the Change Contrast section is now available. There are seven Contrast: available from the drop down list. These include:

   i. **Indicator**: Compares the presence (i.e., 1) or absence (i.e., 0) of the independent variable to a reference category. Note that Indicator is a synonym for dummy coding, the option we want to use for this example and described in detail in Chapter 23.

   ii. **Simple**: Each category of the independent variable is compared to the unweighted average of all categories.

   iii. **Difference**: Each category of the independent variable, except lowest coded category, is compared to the average effect of the previous categories. Also known as reverse Helmert contrasts, because...

   iv. **Helmert**: Each category of the independent variable, except highest coded category, is compared to the average effect of the previous categories.

   v. **Repeated**: Each category of the independent variable, except lowest coded category, is compared to previous category (e.g., 2 to 1; 3 to 2, etc).

   vi. **Polynomial**: Available only for continuous independent variables; uses orthogonal polynomial contrasts when categories are equally spaced. Think, trend-type analysis.

   vii. **Deviation**: Each category of the independent variable, except the reference category, is compared to the unweighted (i.e., mean) overall effect.

c. For several of the options listed above you need to identify the Reference Category:. You can select either the First or the Last category as reference. For variables with three or more groups, if you are more interested in a middle category, the variable will need to be recoded (see Chapter 8). For this example, however, we only have two groups, so it does not matter.

   i. Select the **First** category as reference and then click the Change button.

When you are satisfied with your selections, click **Continue**.

7. Click the **Save** button to bring up the Cox Regression: Save dialog box.

   a. For Cox regression with time-varying covariates, you only have one option, $\text{DfBeta(s)}$, which saves the estimated change in the coefficient value if the participant was removed from analysis.

   b. In addition, you can Export Model Information to XML File by identifying a previously created XML file via the Browse... button. When you are satisfied with your selections, click **Continue**.
8. Finally, click the **Options**… button to bring up the *Cox Regression: Options* dialog box. Here, you can select several *Model Statistics* you want printed as well as information regarding the required *Probability for Stepwise* entry or removal.

   a. Under the *Model Statistics* section, you can choose to print the confidence interval for the hazard ratios (**CI for exp(B)**), as well as set the confidence interval level, with the default being 95%. You can also print the *Correlation of estimates*, but this option is only appropriate when you have more than one independent variable. Finally, you have the choice to *Display Model Information At each step* or *At last step*.

   b. Under the *Probability for Stepwise* section, you can indicate how you want SPSS to use *p*-values to enter (*Entry:* ) or remove (*Removal:* ) variables in a stepwise regression analysis. Because we are not conducting stepwise regression in the example, these options can be ignored.

   c. Finally, because Cox regression uses partial likelihoods, it uses an iterative estimation process to maximize the likelihood that the estimates are correct. You can set the number of *Maximum Iterations:* which essentially controls how long the procedure will search for a solution.

When you are satisfied with your selections, click **Continue**.

9. That’s it! Click **OK** to conduct the analysis.

**Output**

When you click **OK**, SPSS will produce an *Output* screen displaying the results. Click the *Output* window to view your results (if it does not pop up automatically). The first table you see is titled *Case Processing Summary*, which contains important information about your sample including the frequency and percentage of participants experiencing the event (*Event*) and *Censored*, the *Total* number of participants, as well as participants with missing values, negative time, and those that were left-censored.

The next table, titled *Categorical Variable Codings*, is presented in Figure 45.2. This table contains information that is incredibly important to interpretation as it indicates how the categorical independent variable was coded as well as the number of participants in each category.

<table>
<thead>
<tr>
<th>Categorical Variable Codings&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Frequency</th>
<th>(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tx&lt;sup&gt;a&lt;/sup&gt; 0=Placebo</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>1=Treatment</td>
<td>15</td>
<td>1</td>
</tr>
</tbody>
</table>

<sup>a</sup> Indicator Parameter Coding

<sup>b</sup> Category variable: Tx

**Figure 45.2**
The next section of the output is titled *Block 0: Beginning Block*. This block can be thought of as similar to the Block 0 from a logistic regression analysis. That is, it contains information for the model without the independent variable and serves as the comparison or reference model. The table in this section is titled *Omnibus Tests of Model Coefficients* and contains the -2 Log Likelihood value.

Next, the *Block 1: Method=Enter* section is shown which contains the results of the Cox regression analysis. The first table, titled *Omnibus Tests of Model Coefficients*, shown in Figure 45.3, contains the overall test of the model based on a statistical comparison between Block 0 and Block 1. This table contains the -2 Log Likelihood value as well as omnibus tests for the Overall (score) model as well as Change from Previous Step and Change from Previous Block. Note that the latter two sections of this table provide identical results because we did not use hierarchical, sequential, or stepwise Cox regression. Further, note that the Chi-square value from the Change From Previous Block section is the likelihood ratio statistic (i.e., $G^2$) and is used when calculating effect size.

<table>
<thead>
<tr>
<th>Omnibus Tests of Model Coefficients&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Overall (score)</th>
<th>Change From Previous Step</th>
<th>Change From Previous Block</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2 Log Likelihood</td>
<td>Chi-square</td>
<td>df</td>
<td>Sig.</td>
</tr>
<tr>
<td>90.149</td>
<td>7.258</td>
<td>2</td>
<td>.026</td>
</tr>
</tbody>
</table>

<sup>a</sup> Beginning Block Number 1. Method = Enter

The next table is titled *Variables in the Equation* and is identical to table presented in Figure 45.1. This is the most important table in the output because it contains the results of the independent variable. You probably noticed that there is no constant (i.e., $y$-intercept), as provided in both linear and logistic regression. This is because the baseline hazard function is unspecified, or, stated another way, the baseline function is dependent on the specific time point. This table contains the regression slope ($B$), standard error ($SE$), Wald statistic, degrees of freedom ($df$), $p$-value ($Sig.$), hazard ratio ($Exp(B)$) as well as the 95% confidence interval around the hazard ratio (95.0% CI for $Exp(B)$).

Finally, you are presented a table titled *Covariate Means*, which presents the mean for each covariate in the analysis. This table would normally be used to calculate and interpret the survival or hazard functions. However, because the shape of the function changes as time passes, they are not calculated.

**Interpretation**

Because all assumptions for a categorical independent variable were evaluated prior to analysis, interpretation begins by evaluating the *Omnibus Tests of Model Coefficients* table in the *Block 1: Method=Enter* section shown in Figure 45.3. Consider only the omnibus test found in the Change From Previous Block columns. The Chi-square value of 8.858 is statistically significant as $p = .012$. Note that this value was calculated as the difference between the -2 Log Likelihood from Block 0: Beginning Block (i.e., 99.007) and the -2 Log Likelihood from Block 1: Method=Enter (i.e., 90.149) sections. This value, termed the likelihood ratio statistic (or $G^2$) is
used when calculating the effect size estimate. As stated above, the effect size estimate is calculated as: $1 - \exp(-G^2/n)$. Thus, the $R^2$ estimate equals .256 (i.e., $1 - \exp(-8.858/30)$), which represents the relative association between survival time and the treatment group. In addition, degrees of freedom are calculated as the difference in the number of parameters between models; that is, the model in Block 1 has two more parameters (i.e., the independent variable and variable-by-time interaction) than the Block 0.

Next, you evaluate the effect of the independent variable and variable-by-time interaction found in the Variables in the Equation table shown in Figure 45.1. Evaluate the regression coefficient ($B$) and standard error ($S.E.$) as well as the hazard ratio ($\exp(B)$) and confidence interval ($95.0\% \text{ C.I. for } \exp(B)$). Large standard errors and wide confidence intervals indicate poor measurement (look at the width of the CI for the $Tx$ variable!). In addition, a 95% confidence interval that contains 1 indicates the variable is not statistically significant.

**Interpreting Interactions between a Categorical Independent Variable and Time**

Interaction effects adjust main effects, a notion that should become clearer as we work through the example. For categorical independent variables, interaction effects result in an increase or decrease in the difference between groups across time (i.e., log-hazards). This can happen in four ways. A positive interaction effect can make a positive main effect more positive (larger difference across time) or a negative main effect less negative (smaller difference across time). Alternately, a negative interaction effect can make a positive main effect less positive (smaller difference across time) or a negative main effect more negative (larger difference across time). This should make more sense after we consider the regression equation and graph.

The regression equation for this Cox model is:

Remember, there is no baseline function specified for semiparametric Cox regression and the treatment group variable, $Tx$, is coded 1 = treatment group and 0 = placebo. With this information in mind, the value of 11.296 represents the simple main effect of the difference in log-hazard between the placebo and treatment group specifically when time = 0. That is, setting time to 0 (i.e., ln(1)) removes the interaction term (i.e., multiplying anything by 0 equals 0). Thus, at ln(1), there is a 11.296 difference in log-hazards favoring the treatment group, where the hazard ratio was 80518.657. This difference remained statistically significant after adjusting for the interaction between treatment group and time. The statistically significant interaction effect indicates that the difference in log-hazards favoring the treatment group gets smaller (less positive) by 3.276 for every natural log minute increase. The natural log is used because that is what we specified for time in Step 1a above.

Interpretation can be facilitated by examining a plot of hazard ratios provided below in Figure 45.4. This graph was created in Microsoft Excel and the scale of the $x$-axis is the natural log of minutes because that is how the $T_{COV}$ variable involved in the $Tx$-by-time interaction was scaled in Step 1a above. Figure 45.4 presents the hazard ratio across time as opposed to the log-hazards for each group individually because there is no baseline log-hazard value to calculate the log-hazards for the reference (i.e., Placebo) group. From Figure 45.4, you can see that while the hazard ratio is excessively large, favoring the treatment group, at time 0, it decreases quickly and eventually favors the placebo group by study end.

310
Considering Figure 45.4, you can see that the hazard ratio crosses 1 between 3.43 and 3.47 (i.e., between minutes 31 and 32; exp(3.43) and exp(3.47)). That is, the value of 1 is where the hazard ratio shifts from favoring the treatment group to favoring the placebo group. Based on the Figure, we have no way of determining whether differences at any other value other than time 0 are statistically significant (i.e., we only know the hazard ratio for the simple main effect). This is because we are not provided with the standard errors at other time points necessary to conduct a statistical test. However, we could center the time variable; that is, make the 0 point take on another value by subtracting a constant. For this example data, we could center the \textit{Mins} variable at 40 minutes (i.e., \textit{Mins} – 40) using the Compute function (Chapter 13) and using the centered variable (e.g., \textit{Mins40}) in the analysis instead of \textit{Mins}. There is no wrong answer when centering variables (it will not alter slopes), so you could use whatever value you believe is meaningful. However, be careful not to allow the centered variable to take on negative values, especially when using \textit{ln(Mins)} as the time variable, because the natural log of negative numbers does not exist.

\textbf{Example Results Section}

Prior to analysis, no differences between withdrawn and remaining participants were indicated and the probability of survival was assumed constant over the entire study period. Further, independence of residuals was considered satisfied. However, the proportionality of hazards assumption was violated as evidenced by both log-minus-log survival plots and a statistically significant interaction between treatment group and time ($p = .049$). Thus, treatment group was treated as a time-varying covariate.

The results of a simple Cox regression with a time-varying covariate indicated the model significantly predicted time to fall asleep during the 60-minute study period, $\chi^2_2 = 8.858$, $p < .05$, $R^2 = .256$. Specifically at time 0, participants receiving the new sleep treatment had a 11.296 increase in the log-hazards and were 80518.657 times more likely to fall asleep during the study period compared to participants receiving placebo. However, this effect diminished quickly as each natural log minute increase decreased the difference in log-hazard by 3.276 units. This effect is presented graphically in Figure 45.4.
Continuous Independent Variable

Evaluating the Form of the Independent Variable

As stated in Chapter 43, Cox regression assumes the form of a continuous independent variable is linear. More importantly, a violation of this assumption can result in an erroneous violation of the proportionality of hazards assumption. Thus, you should always check this assumption prior to evaluating proportionality of hazards. For a full description of the methods involved in testing this assumption, please refer to Chapter 43. Briefly, to test this assumption we need to plot the Martingale residuals from a baseline model excluding the independent variable against the values of the independent variable and then requesting a smoothing line. Results are shown in Figure 45.5. Although there is a slight deceleration for participants with average body weight, this is not considered fatal to analysis, and the body weight variable is considered to have linear form.

Figure 45.5

Evaluating the Proportionality of Hazards Assumption

With the independent variable shown to have a linear form, the proportionality of hazards assumption can now be evaluated. As stated in Chapter 43, for a continuous independent variable only the statistical method can be used to evaluate the assumption. Similar to above, I will not go into detail discussing the procedure to test the assumption (see Chapter 43), but the results are presented below in Figure 45.6 indicating a statistically significant interaction between body weight and time. Thus, the proportionality of hazards assumption has been violated, and \( W_t \) will be treated as a time-varying covariate.

<table>
<thead>
<tr>
<th>Variables in the Equation</th>
<th>B</th>
<th>SE</th>
<th>Wald</th>
<th>df</th>
<th>Sig.</th>
<th>Exp(B)</th>
<th>95.0% CI for Exp(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( W_t )</td>
<td>.122</td>
<td>.053</td>
<td>4.452</td>
<td>1</td>
<td>.035</td>
<td>1.129</td>
<td>1.009 – 1.264</td>
</tr>
<tr>
<td>T_COV_( W_t )</td>
<td>-.048</td>
<td>.013</td>
<td>6.157</td>
<td>1</td>
<td>.013</td>
<td>.963</td>
<td>.917 – .990</td>
</tr>
</tbody>
</table>

Figure 45.6
Analysis

With the proportionality of hazards assumption violated, but assuming no differences between withdrawn and remaining participants, constant probability of survival over the entire study period, that the event occurred at the time specified, and independence of residuals, to conduct a simple Cox regression with a continuous, time-varying independent variable follow the steps below. You will notice that several options are not available when using time-varying covariates. This results specifically from the inclusion of the variable-by-time interaction.

1. Click **Analyze**, then choose **Survival**, and then click **Cox w/ Time-Dep Cov...** to bring up the **Compute Time-Dependent Covariate** dialog box.

2. The initial dialog box is used specifically to transform the time variable. You will notice that all the variables in your dataset are provided on the left hand side of this dialog box. However, notice that the first variable is new and listed as \( T \_ \). This is an internal time variable created by SPSS to be used for all time related transformations. Similar to testing the proportionality of hazards assumption, you need to use the natural log of time.
   a. In the **Expression for \( T \_ COV \_ \)**: box, type: \( \ln(T) \).
   b. Click the **Model...** button on the right hand side of the dialog box to bring up the **Cox Regression** dialog box.

3. You will immediately see all of the variables in your dataset listed on the left hand side of the dialog box. You will be concerned primarily with three boxes labeled **Time:**, **Status:**, and **Covariates:**. The **Time:** box is fairly self-explanatory. Here, you will place the dependent variable that measures time-to-event. The **Status:** box is essentially the censoring variable. That is, the variable that identifies who experienced the event or not. Finally, the **Covariates:** box is where the categorical independent variable and variable-by-time interaction will be placed. In addition, the **Strata:** box is available, but is not used for simple Cox regression.
   a. Click to highlight the **Mins** variable, and then click the right arrow ( ) next to the **Time:** box.
   b. Click to highlight the **Asleep** variable, and then click the right arrow ( ) next to the **Status:** box.
   i. Click the **Define Event...** button to bring up the **Cox Regression: Define Event for Status Variable** dialog box. In this dialog box, you have three options.
      1. First, you can specify a **Single Value:**. Although you may have multiple values coded for various events, the use of this box allows you to specify one value as the event and all other values as non-events.
         a. For the example data above, type 1 in this box.
      2. Alternatively, you can specify a **Range of values:**. This option is useful if you have multiple values coded for various events, and you are interested in two or more events. However, note that the events must be coded in numerical order. That is, say you are
314

interested only in the events coded 1 and 3, not event 2. By
specifying a range of values, you would necessarily include
the event coded 2, which is not of interest to you.

3. Alternatively, you can specify a list of values. This option can be
used if you have any number of events, coded in any numerical
order. Thus, if you are interested in events coded 1 and 3, you can
specify each event individually by clicking the radio button, typing
the event number of interest and clicking Add.

When you have specified the event(s) of interest, click Continue.

4. Click to highlight the Wt variable, and then click the right arrow (➡) next to the
Covariates: box.

5. Click the Wt variable, press and hold the Ctrl button on your keyboard, and then click
the T_COV_[T_COV_] variable. With both variables highlighted, click the >a*b<
button (➡a*b) next to the Covariates: box. This will place the interaction term in this box,
which should appear as T_COV_*Wt.

6. Click the Save button to bring up the Cox Regression: Save dialog box.
   a. For Cox regression with time-varying covariates, you only have one option,
      DfBeta(s), which saves the estimated change in the coefficient value if the
      participant was removed from analysis. A new variable is saved for each
      independent variable in the model.
   b. In addition, you can Export Model Information to XML File by identifying a
      previously created XML file via the Browse... button. When you are satisfied
      with your selections, click Continue.

7. Finally, click the Options... button to bring up the Cox Regression: Options dialog box.
   Here, you can select several Model Statistics you want printed as well as information
   regarding the required Probability for Stepwise entry or removal.
   a. Under the Model Statistics section, you can choose to print the confidence interval
      for the hazard ratios (CI for exp(B)), as well as set the confidence interval level,
      with the default being 95%. You can also print the Correlation of estimates, but
      this option is only appropriate when you have more than one independent
      variable. Finally, you have the choice to Display Model Information At each step
      or At last step.
   b. Under the Probability for Stepwise section, you can indicate how you want SPSS
      to use p-values to enter (Entry:) or remove (Removal:) variables in a stepwise
      regression analysis. Because we are not conducing stepwise regression in the
      example, these options can be ignored.
   c. Finally, because Cox regression uses partial likelihoods, it uses an iterative
      estimation process to maximize the likelihood that the estimates are correct. You
      can set the number of Maximum Iterations: which essentially controls how long
      the procedure will search for a solution.

When you are satisfied with your selections, click Continue.

8. That’s it! Click OK to conduct the analysis.
Output

When you click OK, SPSS will produce an Output screen displaying the results. Click the Output window to view your results (if it does not pop up automatically). The first table you see is titled Case Processing Summary, which contains important information about your sample including the frequency and percentage of participants experiencing the event (Eventa) and Censored, the Total number of participants, as well as participants with missing values, negative time, and those that were left-censored.

The next section of the output is titled Block 0: Beginning Block. This block can be thought of as similar to the Block 0 from a logistic regression analysis. That is, it contains information for the model without the independent variable and serves as the comparison or reference model. The table in this section is titled Omnibus Tests of Model Coefficients and contains the -2 Log Likelihood value.

Next, the Block 1: Method=Enter section is shown which contains the results of the Cox regression analysis. The first table, titled Omnibus Tests of Model Coefficients, shown in Figure 45.7, contains the overall test of the model based on a statistical comparison between Block 0 and Block 1. This table contains the -2 Log Likelihood value as well as omnibus tests for the Overall (score) model as well as Change from Previous Step and Change from Previous Block. Note that the latter two sections of this table provide identical results because we did not use hierarchical, sequential, or stepwise Cox regression. Further, note that the Chi-square value from the Change From Previous Block section is the likelihood ratio statistic (i.e., $G^2$) and is used when calculating effect size.

![Image of Omnibus Tests of Model Coefficients]

Figure 45.7

The next table is titled Variables in the Equation and is identical to table presented in Figure 45.6. This is the most important table in the output because it contains the results of the independent variable. You probably noticed that there is no constant (i.e., y-intercept), as provided in both linear and logistic regression. This is because the baseline hazard function is unspecified, or, stated another way, the baseline function is dependent on the specific time point. This table contains the regression slope ($B$), standard error ($SE$), Wald statistic, degrees of freedom ($df$), p-value ($Sig.$), hazard ratio ($Exp(B)$) as well as the 95% confidence interval around the hazard ratio (95.0% CI for $Exp(B)$).

Finally, you are presented a table titled Covariate Means, which presents the mean for each covariate in the analysis. This table would normally be used to calculate and interpret the survival or hazard functions. However, because the shape of the function changes as time passes, they are not presented.
**Interpretation**

At this point, we would normally calculate the deviance residuals as described in Chapter 43. However, because body weight is time dependent, the deviance residuals are also time dependent and cannot be evaluated by a static plot. Thus, this assumption cannot be evaluated appropriately and is considered satisfied.

Continuing forward, because all testable assumptions for a continuous independent variable were evaluated prior to analysis, interpretation begins by evaluating the *Omnibus Tests of Model Coefficients* table in the *Block 1: Method=Enter* section shown in Figure 45.7. Consider only the omnibus test found in the *Change From Previous Block* columns. The Chi-square value of 13.140 is statistically significant as \( p = .001 \). Note that this value was calculated as the difference between the -2 Log Likelihood from *Block 0: Beginning Block* (i.e., 99.007) and the -2 Log Likelihood from *Block 1: Method=Enter* (i.e., 85.866) sections. This value, termed the likelihood ratio statistic (or \( G^2 \)) is used when calculating the effect size estimate. As stated above, the effect size estimate is calculated as: \( 1 - \exp(-G^2/n) \). Thus, the \( R^2 \) estimate equals .355 (i.e., \( 1 - \exp(-13.140/30) \)), which represents the relative association between survival time and the body weight. In addition, degrees of freedom are calculated as the difference in the number of parameters between models; that is, the model in Block 1 has two more parameters (i.e., the independent variable and variable-by-time interaction) than the Block 0.

Next, you evaluate the effect of the independent variable and variable-by-time interaction found in the *Variables in the Equation* table shown in Figure 45.6. Evaluate the regression coefficient (\( B \)) and standard error (S.E.) as well as the hazard ratio (\( \exp(B) \)) and confidence interval (95.0% C.I. for \( \exp(B) \)). Large standard errors and wide confidence intervals indicate poor measurement. In addition, a 95% confidence interval that contains 1 indicates the variable is not statistically significant.

**Interpreting Interactions between a Continuous Independent Variable and Time**

Interaction effects adjust main effects, a notion that should become clearer as we work through the example. For continuous independent variables, interaction effects result in an increase or decrease in the slope (i.e., log-hazards) across time. This can happen in four ways. A positive interaction effect can make a positive main effect more positive (increase the slope across time) or a negative main effect less negative (decrease the slope across time). Alternately, a negative interaction effect can make a positive main effect less positive (decrease the slope across time) or a negative main effect more negative (increase the slope across time). This should make more sense after we consider the regression equation and graph.

The regression equation for this Cox model is:

Remember, there is no baseline function specified for semiparametric Cox regression. The value of .122 represents the simple main effect of the slope for body weight specifically when time = 0. That is, setting time to 0 (i.e., ln(1)) removes the interaction term (i.e., multiplying anything by 0 equals 0). Thus, at ln(1), every one-unit increase in body weight resulted in a .122 increase in the log-hazards of falling asleep, where an associated hazard ratio of 1.129. This slope remained statistically significant after adjusting for the interaction between body weight and time. The statistically significant interaction effect indicates that the log-
hazards get smaller (less positive) by .048 for every natural log minute increase. The natural log is used because that is what we specified for time in Step 1a above.

Interpretation can be facilitated by examining a plot of hazard ratios provided below in Figure 45.8. This graph was created in Microsoft Excel and the scale of the $x$-axis is the natural log of minutes because that is how the $T_{COV}$ variable involved in the $Wt$-by-time interaction was scaled in Step 1a above. Figure 45.8 presents the hazard ratio across time. From this Figure, you can see that while the hazard ratio is above 1 at time 0, it decreases quickly and eventually favors those weighing less by study end.

More specifically, you can see that the hazard ratio crosses 1 between 2.48 and 2.56 (i.e., between minutes 12 and 13; $\exp(2.48)$ and $\exp(2.56)$). That is, the value of 1 is where the hazard ratio shifts from favoring heavier participants to lighter participants. Based on the Figure, we have no way of determining whether differences at any other value other than at specifically time 0 are statistically significant (i.e., we only know the hazard ratio for the simple main effect). This is because we are not provided with the standard errors at other time points necessary to conduct a statistical test. However, we could center the time variable; that is, make the 0 point take on another value by subtracting a constant. For this example data, we could center the $Mins$ variable at 13 minutes (i.e., $Mins - 13$) using the Compute function (Chapter 13) and using the centered variable (e.g., $Mins13$) in the analysis instead of $Mins$. There is no wrong answer when centering variables (it will not alter slopes), so you could use whatever value you believe is meaningful. However, be careful not to allow the centered variable to take on negative values, especially when using $\ln(Mins)$ as the time variable, because the natural log of negative numbers does not exist.

**Example Results Section**

Prior to analysis, no differences between withdrawn and remaining participants were indicated and the probability of survival was assumed constant over the entire study period. Further, independence of residuals was considered satisfied. However, the proportionality of hazards assumption as violated as the interaction between body weight and time was statistically significant ($p = .013$).

The results of a simple Cox regression with a time-varying covariate indicated the model significantly predicted time to fall asleep during the 60-minute study period, $\chi^2 = 13.140, p < .05, R^2 = .355$. Specifically at time 0, every
one-unit increase in body weight resulted in a .122 increase in the log-hazards of falling asleep during the study period. The associated hazard ratio was 1.129. However, this effect diminished quickly as time passed and eventually favored participants with lower body weights, as evidenced the plot of hazard ratios presented in Figure 45.8.
Chapter 46
MULTIPLE COX REGRESSION WITH TIME-VARYING COVARIATES

As stated in Chapters 43 and 44, a violation of the proportionality of hazards assumption indicates that the effect of the independent variable (or covariate) interacts with time. That is, the risk of experiencing the event is conditional on the time when risk was calculated. The interaction effect is characterized differently depending on the scale of the independent variable, as described in detail in Chapter 45. Please re-read Chapter 45 if necessary.

The current Chapter pertains to Cox models with multiple independent variables or covariates. As stated in the previous Chapter, a violation of the proportionality of hazards assumption for both categorical and continuous independent variables or covariates, simply include the variable-by-time interaction in your Cox model. The interpretation of the interaction effect, then, is how the passage of time directly affects the main effect of the independent variable or covariate rendering them conditional on the value of their interacting predictor. That is, the main effect becomes the simple main effect when the interacting predictor is 0.

While the procedure above is an option for multiple Cox regression models, an additional option is available for categorical covariates where the analysis is stratified across the covariate’s categories. This analysis is termed the Stratified Cox Model and is only available when the offending variable is not of specific research interest. A stratified Cox model calculates different baseline functions for each category, but does not print statistics for the covariate. That is, you are given no interpretable information about the covariate (e.g., slopes, hazard ratios). However, interpretation of the independent variable is still adjusted for the stratification variable.

Continuing with the example from Chapter 45, say you want to know whether participants who take your new sleep treatment fall asleep faster during a 60-minute study period compared to participants who take the leading over-the-counter treatment or placebo. Because the amount of sleep per night can be affected by numerous variables, you decide to control for caffeine consumption after 2pm (in milligrams), whether the participant smokes, body weight in pounds, and level of global anxiety as measured by the Hamilton Anxiety Rating Scale (HAM-A). The collected data is presented below:

<table>
<thead>
<tr>
<th>ID</th>
<th>Tx</th>
<th>Caff</th>
<th>Smk</th>
<th>Wt</th>
<th>Anx</th>
<th>Asleep</th>
<th>Mins</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>89</td>
<td>0</td>
<td>146</td>
<td>8</td>
<td>1</td>
<td>35</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>191</td>
<td>1</td>
<td>160</td>
<td>22</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>115</td>
<td>1</td>
<td>172</td>
<td>14</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>196</td>
<td>0</td>
<td>153</td>
<td>18</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>122</td>
<td>0</td>
<td>189</td>
<td>15</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>59</td>
<td>0</td>
<td>139</td>
<td>13</td>
<td>1</td>
<td>51</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>140</td>
<td>0</td>
<td>171</td>
<td>11</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>210</td>
<td>0</td>
<td>167</td>
<td>10</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>180</td>
<td>1</td>
<td>168</td>
<td>20</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>103</td>
<td>1</td>
<td>199</td>
<td>16</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>125</td>
<td>0</td>
<td>184</td>
<td>14</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>99</td>
<td>0</td>
<td>170</td>
<td>9</td>
<td>1</td>
<td>56</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>101</td>
<td>1</td>
<td>149</td>
<td>19</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>137</td>
<td>0</td>
<td>144</td>
<td>13</td>
<td>1</td>
<td>44</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>215</td>
<td>1</td>
<td>192</td>
<td>15</td>
<td>0</td>
<td>60</td>
</tr>
</tbody>
</table>
Following the data entry procedures described in Chapters 1 and 2, label the first variable $ID$, the second variable $Tx$ (1 = Treatment; 0 = Placebo) for treatment group, the third variable $Caff$ for caffeine intake, the fourth variable $Smk$ (1 = Yes; 0 = No) for smoking status, the fifth variable $Wt$ for body weight, the sixth variable $Anx$ for anxiety, the seventh variable $Asleep$ (1 = Yes; 0 = No) to indicate whether the participants fell asleep within the first hour, and the eighth variable $Mins$ for minutes required to fall asleep. Enter the data appropriately.

**Assumptions**

Aside from the proportionality of hazards assumption, which this Chapter assumes is violated (if not, use the procedures described in Chapter 44), the assumptions for a Cox model with time-varying covariates are identical to a Cox model without time-varying covariates. These assumptions include no differences between withdrawn and remaining participants, constant probability of survival over the entire study period, that the event occurred at the time specified, independence of residuals, that the continuous covariate assumes a linear form, and absence of outliers. Further, if you are conducting a multiple Cox regression absence of multicollinearity and multivariate outliers is assumed. In addition, the stratified Cox model assumes no interaction between the stratification variable and any remaining independent variables and covariates. All of these assumptions, except the last, were described in detail in Chapters 43 and 44; thus, only brief descriptions of previously defined assumptions are provided here.

No differences between withdrawn and remaining participants require that participants who are censored (or lost to follow-up) do not differ systematically from those who experience the event. This assumption is testable using a binary logistic regression (Chapters 40 or 41) to evaluate for differences between withdrawn and remaining participants. Having constant probability of survival over the study period requires (1) that the same factors affecting survival are present throughout the entire study period and (2) participants entering the study at different time points have equal probability of survival. Assuring that the event occurred at the time specified can be difficult, especially for left- or interval-censored participants (e.g., participant fell asleep 10-15 minutes after dosing, but the researcher is not exactly sure when). The assumption of independence of residuals is technically a design issue and can usually be satisfied by data that is not clustered and by not measuring participants repeatedly.

The first four assumptions mentioned above were mostly assured via proper experimental design and protocol implementation. The remaining assumptions are strictly related to the regression component. In Cox regression, continuous covariates are assumed to have a linear form, an assumption similar to the linearity in the logit assumption for logistic regression. There are several methods for testing this assumption, but we have focused primarily on the use of Martingale residuals. This assumption is tested by estimating a Cox regression model without the continuous independent variable (i.e., a baseline model), plotting the Martingale residuals on the y-axis with the previously omitted continuous variable on the x-axis, and finally, requesting a loess line to determine the functional form of the variable. The absence of outliers assumption was evaluated by plotting deviance residuals for each participant; however, with time-varying covariates the residuals are also time-dependent. Thus, they cannot be plotted and the assumption is left untestable. With multiple independent variables or covariates, absence of multicollinearity and absence of multivariate outliers are required. Both are evaluated using the linear regression procedure described in Chapter 23. Multicollinearity is evaluated via tolerance and VIF values, whereas multivariate outliers assumption are identified via Mahalanobis distance.
A stratified Cox model requires no interaction between the stratification variable and the other independent variables and covariates remaining in the model. This assumption is tested by conducting a likelihood-ratio test between the model without the stratification variable-by-independent variable (or covariate) interactions and the model with these interactions. The nature of the log-likelihood is such that a model with more parameters will have a lower log-likelihood; thus, the statistical test is whether including the interactions improved relative model fit enough to matter statistically. If the result is statistically significant, the interaction(s) must be retained and interpretations of the independent variable or covariate main effects become simple main effects that are conditional on the level of the stratification variable. This procedure will be described in detail below.

**Evaluating the Form of Continuous Independent Variables**

As stated in Chapter 43, Cox regression assumes the form of a continuous independent variable is linear. More importantly, a violation of this assumption can result in an erroneous violation of the proportionality of hazards assumption. Thus, you should always check this assumption prior to evaluating proportionality of hazards. For a full description of the methods involved in testing this assumption, please refer to Chapter 44. Briefly, to test this assumption we need to plot the Martingale residuals from a baseline model excluding any independent variables or covariates against the values of the independent variable and then requesting a smoothing line. Note that the assumption is evaluated for all continuous covariates (e.g., Caff, Wt, and Anx from the example data above). Results are shown in Figure 46.1. We are concerned only with the top row of the Figure, with the Martingale residuals (i.e., Mart_1) on the y-axis and the continuous covariates on the x-axis. Although not perfect, each loess lines is approximately linear; thus, the assumption is considered satisfied.

![Figure 46.1](image)

**Evaluating the Proportionality of Hazards Assumption**

With all continuous covariates shown to have a linear form, the proportionality of hazards assumption can now be evaluated. As stated in Chapter 44, for multiple Cox regression, all covariates are evaluated statistically in the same analysis. Similar to above, I will not go into
detail discussing the procedure to test the assumption (see Chapter 44), but the results are presented below in Figure 46.2. Results indicated a statistically significant interaction between time and smoking status, body weight, and treatment group. Thus, the proportionality of hazards assumption has been violated for these variable and they will be treated as time-varying covariates in all subsequent analyses.

<table>
<thead>
<tr>
<th>Variables in the Equation</th>
<th>B</th>
<th>SEM</th>
<th>Wald</th>
<th>df</th>
<th>Sig.</th>
<th>Exp(B)</th>
<th>95.0% CI for Exp(B) Lower</th>
<th>95.0% CI for Exp(B) Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caff</td>
<td>0.015</td>
<td>0.030</td>
<td>2.40</td>
<td>1</td>
<td>0.124</td>
<td>1.015</td>
<td>0.956</td>
<td>1.077</td>
</tr>
<tr>
<td>Wt</td>
<td>0.293</td>
<td>0.130</td>
<td>5.006</td>
<td>1</td>
<td>0.024</td>
<td>1.340</td>
<td>1.039</td>
<td>1.720</td>
</tr>
<tr>
<td>Arx</td>
<td>-0.278</td>
<td>0.186</td>
<td>0.327</td>
<td>1</td>
<td>0.567</td>
<td>0.757</td>
<td>0.202</td>
<td>1.964</td>
</tr>
<tr>
<td>TX</td>
<td>18.143</td>
<td>6.092</td>
<td>8.869</td>
<td>1</td>
<td>0.003</td>
<td>7.579E7</td>
<td>494.122</td>
<td>1.16E13</td>
</tr>
<tr>
<td>Caff*T_COV_</td>
<td>-0.005</td>
<td>0.010</td>
<td>0.222</td>
<td>1</td>
<td>0.638</td>
<td>0.995</td>
<td>0.977</td>
<td>1.015</td>
</tr>
<tr>
<td>Smk*T_COV_</td>
<td>-8.137</td>
<td>3.723</td>
<td>4.812</td>
<td>1</td>
<td>0.028</td>
<td>0.039</td>
<td>0.000</td>
<td>0.419</td>
</tr>
<tr>
<td>T_COV_*Wt</td>
<td>-0.094</td>
<td>0.042</td>
<td>5.042</td>
<td>1</td>
<td>0.025</td>
<td>0.911</td>
<td>0.839</td>
<td>0.988</td>
</tr>
<tr>
<td>Arx*T_COV_</td>
<td>0.034</td>
<td>0.156</td>
<td>0.048</td>
<td>1</td>
<td>0.827</td>
<td>1.035</td>
<td>0.762</td>
<td>1.404</td>
</tr>
<tr>
<td>T_COV_*Rk</td>
<td>-5.132</td>
<td>1.022</td>
<td>7.905</td>
<td>1</td>
<td>0.006</td>
<td>0.006</td>
<td>0.000</td>
<td>0.212</td>
</tr>
</tbody>
</table>

**Figure 46.2**

**Analysis**

With the proportionality of hazards assumption violated for smoking status, body weight, and treatment group, but assuming no differences between withdrawn and remaining participants, constant probability of survival over the entire study period, that the event occurred at the time specified, independence of residuals, and no multicollinearity or multivariate outliers, to conduct a multiple Cox regression with time-varying covariates follow the steps below. Similar to Chapter 45, you will notice that several options are not available when using time-varying covariates, a result specifically due to the inclusion of the variable-by-time interactions.

1. Click **Analyze**, then choose **Survival**, and then click **Cox w/ Time-Dep Cov…** to bring up the **Compute Time-Dependent Covariate** dialog box.

2. The initial dialog box is used specifically to transform the time variable. You will notice that all the variables in your dataset are provided on the left hand side of this dialog box. However, notice that the first variable is new and listed as **Time [T]**. This is an internal time variable created by SPSS to be used for all time related transformations. Similar to testing the proportionality of hazards assumption, we need to use the natural log of time.
   a. In the **Expression for T_COV_** box, type: **ln(T)**.
   b. Click the **Model...** button on the right hand side of the dialog box to bring up the **Cox Regression** dialog box.
3. You will immediately see all of the variables in your dataset listed on the left hand side of the dialog box. You will be concerned primarily with three boxes labeled *Time*, *Status*, and *Covariates*. The *Time* box is fairly self-explanatory. Here, you will place the dependent variable that measures time-to-event. The *Status* box is essentially the censoring variable. That is, the variable that identifies who experienced the event or not. Finally, the *Covariates* box is where the categorical independent variable and variable-by-time interaction will be placed. In addition, the *Strata* box is available, but is not used for simple Cox regression.

   a. Click to highlight the *Mins* variable, and then click the right arrow (→) next to the *Time* box.

   b. Click to highlight the *Asleep* variable, and then click the right arrow (→) next to the *Status* box.

   i. Click the **Define Event…** button to bring up the *Cox Regression: Define Event for Status Variable* dialog box. In this dialog box, you have three options.

   1. First, you can specify a **Single Value**: Although you may have multiple values coded for various events, the use of this box allows you to specify one value as the event and all other values as non-events.

   b. For the example data above, type 1 in this box.

   2. Alternatively, you can specify a **Range of values**: This option is useful if you have multiple values coded for various events, and you are interested in two or more events. However, note that the events must be coded in numerical order. That is, say you are interested only in the events coded 1 and 3, not event 2. By specifying a range of values, you would necessarily include the event coded 2, which is not of interest to you.

   3. Alternatively, you can specify a **List of values**: This option can be used if you have any number of events, coded in any numerical order. Thus, if you are interested in events coded 1 and 3, you can specify each event individually by clicking the radio button, typing the event number of interest and clicking **Add**.

   When you have specified the event(s) of interest, click **Continue**.

4. Click to highlight the *Caff* variable, press and hold the **Ctrl** button on your keyboard, and then click the *Smk*, *Wt*, and *Anx* variables. Click the right arrow (→) next to the *Covariates* box.

5. Next, click the *Smk* variable, press and hold the **Ctrl** button on your keyboard, and then click the *T_COV_[T_COV_]* variable. With both variables highlighted, click the >a*b> button (→) next to the *Covariates* box. This will place the interaction term in this box, which should appear as *T_COV_* *Smk*. Follow this same procedure to include the interaction for *Wt*. 

323
6. With all covariates entered, you can now enter your independent variable and interaction term. Click to highlight the \( Tx \) variable, and then click the right arrow (\( \rightarrow \)) next to the \textit{Covariates:} box.

7. Click to highlight the \( Tx \) variable, press and hold the \texttt{Ctrl} button on your keyboard, and then click the \texttt{T_COV[T_COV]} variable. With both variables highlighted, click the \texttt{\textgreater a*b\textless} button (\( \rightarrow \)) next to the \textit{Covariates:} box.

8. Click the \textit{Categorical…} button. Here, only the variables you entered in the \textit{Covariates:} box in Steps 4 and 6 above are shown on the left hand side. You will use this option when you have any categorical independent variable or covariate. Not specifying the independent variable as categorical will not influence your results with only two groups, but you will not be able to specify a particular reference category. In addition, the categorical option removes the need for you to create new dummy variables as you did with linear regression for variables consisting of three or more groups.

   a. Click to highlight the \( Tx \) variable, press and hold the \texttt{Ctrl} button on your keyboard, click the \texttt{Smk} variable, and then click the right arrow (\( \rightarrow \)) next to the \textit{Categorical Covariates:} box.

   b. Notice the \textit{Change Contrast} section is now available. There are seven \textit{Contrast:} options available from the drop down list. These include:

      i. \textbf{Indicator}: Compares the presence (i.e., 1) or absence (i.e., 0) of the independent variable to a reference category. Note that \textit{Indicator} is a synonym for dummy coding, the option we want to use for this example and described in detail in Chapter 23.

      ii. \textit{Simple}: Each category of the independent variable is compared to the unweighted average of all categories.

      iii. \textit{Difference}: Each category of the independent variable, except lowest coded category, is compared to the average effect of the previous categories. Also known as reverse Helmert contrasts, because…

      iv. \textit{Helmert}: Each category of the independent variable, except highest coded category, is compared to the average effect of the previous categories.

      v. \textit{Repeated}: Each category of the independent variable, except lowest coded category, is compared to previous category (e.g., 2 to 1; 3 to 2, etc).

      vi. \textit{Polynomial}: Available only for continuous independent variables; uses orthogonal polynomial contrasts when categories are equally spaced. Think, trend-type analysis.

      vii. \textit{Deviation}: Each category of the independent variable, except the reference category, is compared to the unweighted (i.e., mean) overall effect.

   c. For several of the options listed above you need to identify the \textit{Reference Category:}. You can select either the \textit{First} or the \textit{Last} category as reference. For variables with three or more groups, if you are more interested in a middle category, the variable will need to be recoded (see Chapter 8). For this example, however, we only have two groups, so it does not matter.
i. In the *Categorical Covariates:* box, click to highlight the *Tx(Indicator)* variable. Then, in the *Change Contrast* section, click the *First* radio button to select category 0 (i.e., placebo group) as reference and then click the *Change* button.

When you are satisfied with your selections, click **Continue.**

9. Clicking the **Save** button will bring up the *Cox Regression: Save* dialog box.
   a. For Cox regression with time-varying covariates, you only have one option, *DfBeta(s)*, which saves the estimated change in the coefficient value if the participant was removed from analysis. A new variable is saved for each independent variable in the model.
   b. In addition, you can *Export Model Information to XML File* by identifying a previously created XML file via the *Browse…* button.

When you are satisfied with your selections, click **Continue.**

10. Finally, click the **Options…** button to bring up the *Cox Regression: Options* dialog box. Here, you can select several *Model Statistics* you want printed as well as information regarding the required *Probability for Stepwise* entry or removal.
   a. Under the *Model Statistics* section, you can choose to print the confidence interval for the hazard ratios (*CI for exp(B)*), as well as set the confidence interval level, with the default being 95%. You can also print the *Correlation of estimates.* Finally, you have the choice to *Display Model Information At each step* or *At last step.* Because we are not conducting stepwise regression in the example, this last option can be ignored.
   b. Under the *Probability for Stepwise* section, you can indicate how you want SPSS to use *p*-values to enter (*Entry:* ) or remove (*Removal:* ) variables in a stepwise regression analysis. Because we are not conducting stepwise regression in the example, these options can be ignored.
   c. Finally, because Cox regression uses partial likelihoods, it uses an iterative estimation process to maximize the likelihood that the estimates are correct. You can set the number of *Maximum Iterations:* which essentially controls how long the procedure will search for a solution.

When you are satisfied with your selections, click **Continue.**

11. That’s it! Click **OK** to conduct the analysis.

**Output**

When you click **OK**, SPSS will produce an *Output* screen displaying the results. Click the *Output* window to view your results (if it does not pop up automatically). The first table you see is titled *Case Processing Summary,* which contains important information about your sample including the frequency and percentage of participants experiencing the event (*Event*) and *Censored,* the *Total* number of participants, as well as participants with missing values, negative time, and those that were left-censored.
The next table, titled *Categorical Variable Codings*, is presented in Figure 46.3. This table contains information that is incredibly important to interpretation as it indicates how each categorical covariate was coded as well as the number of participants within each category.

<table>
<thead>
<tr>
<th>Categorical Variable Codings&lt;sup&gt;c,d&lt;/sup&gt;</th>
<th>Frequency</th>
<th>(1)&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tx&lt;sup&gt;a&lt;/sup&gt; 0=Placebo</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>1=Treatment</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>Smk&lt;sup&gt;a&lt;/sup&gt; 0=No</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>1=Yes</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

a. Indicator Parameter Coding  
b. The (0,1) variable has been recoded, so its coefficients will not be the same as for indicator (0,1) coding.  
c. Category variable: Tx  
d. Category variable: Smk

Figure 46.3

The next section of the output is titled *Block 0: Beginning Block*. This block can be thought of as similar to the Block 0 from a logistic regression analysis. That is, it contains information for the model without the independent variable and serves as the comparison or reference model. The table in this section is titled *Omnibus Tests of Model Coefficients* and contains the -2 Log Likelihood value.

Next, the *Block 1: Method=Enter* section is shown which contains the results of the Cox regression analysis. The first table, titled *Omnibus Tests of Model Coefficients*, shown in Figure 46.4, contains the overall test of the model based on a statistical comparison between Block 0 and Block 1. This table contains the -2 Log Likelihood value as well as omnibus tests for the Overall (score) model as well as Change from Previous Step and Change from Previous Block. Note that the latter two sections of this table provide identical results because we did not use hierarchical, sequential, or stepwise Cox regression. Further, note that the Chi-square value from the Change From Previous Block section is the likelihood ratio statistic (i.e., $G^2$) and is used when calculating effect size.

<table>
<thead>
<tr>
<th>Omnibus Tests of Model Coefficients&lt;sup&gt;5&lt;/sup&gt;</th>
<th>Overall (score)</th>
<th>Change From Previous Step</th>
<th>Change From Previous Block</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2 Log Likelihood</td>
<td>Chi-square</td>
<td>df</td>
<td>Sig.</td>
</tr>
<tr>
<td>86.653</td>
<td>31.129</td>
<td>8</td>
<td>.000</td>
</tr>
</tbody>
</table>

a. Beginning Block Number 1. Method = Enter

Figure 46.4
The next table is titled *Variables in the Equation* and is presented in Figure 46.5. This is the most important table in the output because it contains the results of the independent variable after controlling for the covariates. You probably noticed that there is no constant (i.e., \( y \)-intercept), as provided in both linear and logistic regression. This is because the baseline hazard function is unspecified, or, stated another way, the baseline function is dependent on the specific time point. This table contains the regression slope (\( B \)), standard error (\( SE \)), *Wald* statistic, degrees of freedom (\( df \)), \( p \)-value (\( Sig. \)), hazard ratio (\( Exp(B) \)) as well as the 95% confidence interval around the hazard ratio (95.0% CI for \( Exp(B) \)).

<table>
<thead>
<tr>
<th>Variables in the Equation</th>
<th>B</th>
<th>SE</th>
<th><em>Wald</em></th>
<th>df</th>
<th>Sig.</th>
<th>Exp(B)</th>
<th>95.0% CI for Exp(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calf</td>
<td>.001</td>
<td>.006</td>
<td>.023</td>
<td>1</td>
<td>.830</td>
<td>1.001</td>
<td>.990</td>
</tr>
<tr>
<td>Smk</td>
<td>-25.661</td>
<td>10.363</td>
<td>6.143</td>
<td>1</td>
<td>.013</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>Wt</td>
<td>3.06</td>
<td>1.30</td>
<td>5.567</td>
<td>1</td>
<td>.018</td>
<td>1.358</td>
<td>1.063</td>
</tr>
<tr>
<td>Anx</td>
<td>-1.101</td>
<td>.099</td>
<td>3.372</td>
<td>1</td>
<td>.066</td>
<td>.834</td>
<td>.666</td>
</tr>
<tr>
<td>Tx</td>
<td>17.564</td>
<td>5.861</td>
<td>9.012</td>
<td>1</td>
<td>.003</td>
<td>2.426E7</td>
<td>444.478</td>
</tr>
<tr>
<td>Smk*T_COV</td>
<td>7.988</td>
<td>3.635</td>
<td>4.722</td>
<td>1</td>
<td>.030</td>
<td>2932.693</td>
<td>2.169</td>
</tr>
<tr>
<td>T_COV*Wt</td>
<td>.088</td>
<td>.042</td>
<td>5.685</td>
<td>1</td>
<td>.018</td>
<td>.006</td>
<td>.036</td>
</tr>
<tr>
<td>T_COV*Tx</td>
<td>-4.938</td>
<td>1.730</td>
<td>8.145</td>
<td>1</td>
<td>.004</td>
<td>.007</td>
<td>.000</td>
</tr>
</tbody>
</table>

Finally, you are presented a table titled *Covariate Means*, which presents the mean for each covariate in the analysis. This table would normally be used to calculate and interpret the survival or hazard functions; however, because the shape of the function changes as time passes, these functions are not presented.

**Interpretation**

At this point, we would normally calculate the deviance residuals as described in Chapter 43. However, because we have included time dependent covariates, the deviance residuals are also time dependent and cannot be evaluated by a static plot. Thus, this assumption cannot be evaluated appropriately and is considered satisfied.

Continuing forward, because all testable assumptions for a this multiple Cox regression with time-varying covariates were evaluated prior to analysis, interpretation begins by evaluating the *Omnibus Tests of Model Coefficients* table in the *Block 1: Method=Enter* section shown in Figure 46.4. Consider only the omnibus test found in the *Change From Previous Block* columns. The *Chi-square* value of 37.632 is statistically significant as \( p < .0005 \). Note that this value was calculated as the difference between the -2 *Log Likelihood* from *Block 0: Beginning Block* (i.e., 104.285) and the -2 *Log Likelihood* from *Block 1: Method=Enter* (i.e., 66.653) sections. This value, termed the likelihood ratio statistic (or \( G^2 \)), is used when calculating the effect size estimate. As stated above, the effect size estimate is calculated as: \( 1 - \exp(-G^2/n) \). Thus, the \( R^2 \)
estimate equals .715 (i.e., $1 - \exp(-37.632/30)$), an enormous value representing the relative
association between survival time and the covariates. In addition, degrees of freedom are
calculated as the difference in the number of parameters between models; that is, the model in
Block 1 has eight more parameters (i.e., five covariates and three variable-by-time interactions)
than the Block 0.

Next, you evaluate the effect of the independent variable and variable-by-time interaction
found in the Variables in the Equation table shown in Figure 46.5. Because we are only
concerned with the difference between treatment group after adjusting for the other covariates,
we evaluate the regression coefficient ($B$) and standard error (S.E.) as well as the hazard ratio
($\exp(B)$) and confidence interval (95.0% C.I. for $\exp(B)$) only for this variable and its interaction
with time. Both the main effect and interaction are statistically significant. However, note that
large standard errors and wide confidence intervals indicate poor measurement (look at the
standard error for the $Tx$ variable, which is more than likely a direct result of using a small
sample size; remember, standard errors are greatly influenced by sample size). In addition, note
that a 95% confidence interval for both variables does not contain 1. This indicates that the
variables are statistically significant at $p < .05$.

As stated in Chapter 45, interaction effects adjust main effects. Because our independent
variable of interest is categorical, the treatment group-by-time interaction produces either an
increase or decrease in the difference between groups across time. Please refer to Chapter 45 for
a full description of interpreting interaction effects. It is important to note that the primary
difference between the interaction effects discussed in Chapter 45 and those in this Chapter is
that now the interaction effects are adjusted for all other covariates in the model. This additional
step in interpretation should make sense provided you understand the multiple regression
procedures from previous Chapters (i.e., Chapters 23 and 41). Considering this information, we
move on to the interpretation.

Before we begin, it is important to remember, there is no baseline function specified for
semiparametric Cox regression and the treatment group variable, $Tx$, is coded 1 = treatment
group and 0 = placebo. With this information in mind, the $Tx$ value of 17.564 represents the
simple main effect of the difference in log-hazard between the placebo and treatment group
specifically when time = 0 (i.e., ln(1)) after adjusting for the interaction between treatment group
and time as well as the other covariates in the model. That is, setting time to 0 removes the
interaction term (i.e., multiplying anything by 0 equals 0). Thus, at ln(1), there is a 17.564
difference in log-hazards favoring the treatment group, with a hazard ratio of 4246000. Because
the $Tx$ variable is time dependent, the difference in log-hazards favoring the treatment group gets
smaller (less positive) by 4.938 for every natural log minute increase after controlling for the
other covariates. Remember, the natural log is used because that is what we specified for time in
Step 1a above.

Similar to Chapter 45, interpretation can be facilitated by examining a plot of hazard
ratios provided below in Figure 46.6. This graph was created in Microsoft Excel and the scale of
the x-axis is the natural log of minutes because that is how the $T\_COV\_x$ variable was scaled to be
included in the $Tx$-by-time interaction in Step 1a above. Figure 46.6 presents the hazard ratio
across time as opposed to the log-hazards for each group individually because there is no
baseline log-hazard value to calculate the log-hazards for the reference (i.e., Placebo) group.
Note that the Figure assumes the value of each covariate is 0; that is, the Figure is calculated for
a participant who did not ingest caffeine after 2pm, does not smoke, weighs 0 pounds, and has no
anxiety. Using 0s also removes the smoking status-by-time interaction. Stated another way, this
Figure is for an individual who does not exist in our sample, offering another situation where centering the covariates would be effective; that is, creating a meaningful 0 point for each continuous covariate allowing relevant interpretation.

From Figure 45.6, you can see that while the hazard ratio is excessively large, favoring the treatment group at time 0, it decreases quickly and eventually favors the placebo group by study end.

More specifically, you can see that the hazard ratio crosses 1 between 3.56 and 3.58 (i.e., between minutes 35 and 36; exp(3.56) and exp(3.58)). That is, the value of 1 is where the hazard ratio shifts from favoring the treatment group to favoring the placebo group. Based on the Figure, we have no way of determining whether differences at any value other than time 0 are statistically significant (i.e., we only know the hazard ratio for the simple main effect). This is because we are not provided with the standard errors at other time points necessary to conduct a statistical test. However, we could center the time variable; that is, make the 0 point take on another value by subtracting a constant. For this example data, we could center the Mins variable at 35 minutes (i.e., Mins – 35) using the Compute function (Chapter 13) and using the centered variable (e.g., Mins35) in the analysis instead of Mins. There is no wrong answer when centering variables (it will not alter results), so you could use whatever value you believe is meaningful. However, be careful not to allow the centered variable to take on negative values, especially when using ln(Mins) as the time variable, because the natural log of negative numbers does not exist.

**Example Results Section**

A violation of the proportionality of hazards assumption was indicated as smoking status, body weight, and treatment group interacted significantly with time. Thus, these interactions were retained in all subsequent analyses. Prior to analysis, no differences between withdrawn and remaining participants was indicated, the probability of survival was assumed constant over the entire study period, the event occurred at the time specified, independence of residuals was assured, and no multicollinearity or multivariate outliers were indicated.
The results of a multiple Cox regression with time-varying covariates indicated the variables as a set provided statistically significant prediction of the time required to fall asleep during the 60-minutes study period, $\chi^2_8 = 37.632, p < .05, R^2 = .715$. Specifically at time 0, participants receiving the new sleep treatment had a 17.564 increase in the log-hazards and were 42460000 times more likely to fall asleep during the study period compared to participants receiving placebo after adjusting for caffeine intake, body weight, smoking status, and anxiety level. However, this effect diminished quickly as each natural log minute increase decreased the difference in log-hazard by 4.938 units. This effect is presented graphically in Figure 46.6.

**Stratified Cox Model**

The stratified Cox model is a modification of Cox regression that uses stratification to control for variables violating the proportionality of hazards assumption. As stated in the introduction to this Chapter, a stratified model is only used when the offending variable is not of specific research interest (i.e., the variable is a true covariate). Further, this type of Cox model calculates different baseline functions for each stratification category, but does not print statistics for the stratifying covariate. That is, you are given no interpretable information about the covariate (e.g., slopes, hazard ratios), but interpretation of the independent variable(s) included in the model are still adjusted for the stratification variable.

While it makes more sense for the stratification variable to be categorical, you can stratify continuous variables by categorizing them prior to analysis. However, keep in mind that any time you categorize data you lose information, so I recommend modeling the variable-by-time interaction to control for offending continuous variables as described above. Therefore, this section will only pertain to categorical covariates.

Before we proceed, it is important to note that SPSS only allows one stratification variable to be modeled. With that said, if you have multiple categorical variables that violate the proportionality of hazards assumption, you have two choices. First, you can select one variable to stratify over and then model the variable-by-time interaction for the other. Second, you can create a new categorical variable (i.e., recode; Chapter 8) that describes all groups across the two variables. For example, say you have two variables violating the assumption—treatment group with three levels (e.g., treatment, OTC, and placebo) and smoking status with two levels (e.g., smoke and do not smoke). You would then create a new variable with six levels to describe each participant in your sample (e.g., treatment/smoke, treatment/do not smoke, OTC/smoke, OTC/do not smoke, placebo/smoke, placebo/do not smoke). This new six-category variable would then be used as the stratification variable.

Finally, it is important to note that the stratified Cox model has one additional assumption over and above those already discussed requiring no interaction between the stratification variable and the other independent variables and covariates remaining in the model. This assumption is tested by conducting a likelihood-ratio test between the model without the stratification variable-by-independent variable (or covariate) interactions and the model with these interactions. You do not want this test to be statistically significant. However, if it is statistically significant, the statistically significant interactions must be included in all subsequent analyses. Testing this assumption is not difficult, but requires several steps and the procedure is described in detail below.
Evaluating the No Interaction Assumption

For this example, we will use the data presented above. Note that all of the assumptions for multiple Cox regression apply to stratified Cox regression. Thus, because all assumptions were tested and satisfied (except proportionality of hazards, of course) for this data previously we will work through testing the no interaction assumption and then proceed to interpretation. For these data, both \( T_x \) and \( S_mk \) violated the proportionality of hazards assumption. Because we are interested specifically in \( T_x \) (i.e., we want the test statistics) and \( S_mk \) is not of research interest, we will use \( S_mk \) as the stratification variable.

The no interaction assumption is evaluated by comparing two stratified Cox regression models, one with and one without the stratification variable-by-covariate interactions. The statistical test of this assumption is the likelihood-ratio test, which can be calculated easily by hand; however, we will have SPSS conduct the test for us using the exact same procedure described above used to evaluate the omnibus effect (i.e., \(-2LL_{\text{bigger}} - (-2LL_{\text{smaller}})\)).

Because no test statistics are produced for the stratification variable, we are unable to specify the required interactions directly in the Cox regression dialog boxes. Thus, prior to analysis, we need to create our interaction variables using the Compute procedure.

1. Click **Transform**, and then click **Compute Variable**....
2. In the **Target Variable**: box, type **SmkTx**.
3. In the **Numeric Expression**: box, type **Smk*Tx** to multiply the smoking status and treatment group variables.
4. Click **OK**.
5. Create the remaining interactions by repeating Steps 1 through 4, replacing \( T_x \) with \( C_aфф, W_t, \) and \( A_nx \). Note I named these remaining variables **SmkCaff**, **SmkWt**, and **SmkAnx**. If you completed the steps correctly, you will have four new variables in the dataset.

With the interactions created successfully, we will now evaluate whether the interactions provide a statistically significant contribution to the stratified Cox model. Again, we are hoping for non-statistically significant results. To test the assumption we will be using Cox regression with time-varying covariates because the \( T_x \) and \( W_t \) variables violated the proportionality of hazards assumption. Note that you may have to go through and click **Reset** both at the **Compute Time-Dependent Covariate** and **Cox Regression** dialog boxes to get the analysis to run. That is, we need to start fresh. This is a glitch in SPSS and a definite pain!!

6. Click **Analyze**, then choose **Survival**, and then click **Cox w/ Time-Dep Cov**... to bring up the **Compute Time-Dependent Covariate** dialog box.
   a. In the **Expression for T_COV_**: box, type **ln(T_)** and click **Model**....
7. Click to highlight the **Mins** variable, and then click the right arrow ( ) next to the **Time**: box.
8. Click to highlight the **Asleep** variable, and then click the right arrow ( ) next to the **Status**: box.
   a. Click the **Define Event**... button. In the **Single Value**: box type **1** and then click **Continue**.
9. Click to highlight the Smk variable, and then click the right arrow ( ) next to the Strata: box. Note that this variable was removed completely from the variable list.

10. Click to highlight the Tx variable, press and hold the Ctrl button on your keyboard, and then click the Caff, Wt, and Anx variables. Click the right arrow ( ) next to the Covariates: box.

11. Next, click the Tx variable, press and hold the Ctrl button on your keyboard, and then click the T_COV_T_COV variable. With both variables highlighted, click the >a*b< button ( ) next to the Covariates: box. Follow this same procedure to include the interaction for Wt.

12. In the Block 1 of 1 section, click the Next button.

13. Repeat Steps 10 through 12.

14. Click to highlight the SmkTx variable, press and hold the Ctrl button on your keyboard, and then click the SmkCaff, SmkWt, and SmkAnx variables. Click the right arrow ( ) next to the Covariates: box.

15. That’s it! Click OK to conduct the analysis.

When you click OK, SPSS will produce an Output screen displaying the results. Click the Output window to view your results (if it does not pop up automatically). The steps above conducted what is termed a sequential Cox regression analysis with time-varying covariates. The sequential portion indicates that there are different variables added in different block or steps allowing you to evaluate the effect of the additional variables over and above the effect of the variables already in the model. In the output, we are primarily concerned with the Omnibus Tests of Model Coefficients table in the Block 2: Method = Enter section, shown in Figure 46.7.

![Figure 46.7](image.png)

Similar to the Cox models we discussed above, we are interested in the likelihood-ratio test provided in the Change from Previous Block columns. The Chi-Square value of 4.766 is the difference in -2 Log Likelihoods from Block 1 and Block 2 (i.e., 59.285 – 54.519 = 4.766). Remember, any time you add variables to the model, the log-likelihood decreases. Thus, likelihood-ratio test answers the question of whether the four interactions between smoke and the remaining covariates contribute significantly to the model. Because the omnibus test is not statistically significant (i.e., \( p = .312 \)), the four interactions did not contribute statistically to the model and can be removed. Thus, the no interaction assumption is considered satisfied.

If, however, the likelihood-ratio test had been statistically significant, you would evaluate the Variables in the Equation table to determine which interaction(s) was statistically significant (remember, the likelihood-ratio test is an omnibus test). All statistically significant interactions would then be included in the subsequent analyses, whereas interactions that were not statistically significant would be removed.
Analysis

With all assumptions evaluated and considered prior to analysis, to conduct a stratified Cox regression with time-varying covariates follow the steps below. Note that this procedure is identical to the multiple Cox regression with time-varying covariates procedure described earlier in this Chapter; thus, abbreviated instructions are presented here. The reader is referred back to the beginning of the Chapter for a full description of the options available.

1. Click **Analyze**, then choose **Survival**, and then click **Cox w/ Time-Dep Cov…** to bring up the **Compute Time-Dependent Covariate** dialog box.
   a. In the **Expression for T_COV_** box, type: \(\ln(T)\) and click **Model…**
2. Click to highlight the **Mins** variable, and then click the right arrow (\(\Rightarrow\)) next to the **Time:** box.
3. Click to highlight the **Asleep** variable, and then click the right arrow (\(\Rightarrow\)) next to the **Status:** box.
   b. Click the **Define Event…** button. In the **Single Value:** box type \(1\) and then click **Continue**.
4. Click to highlight the **Smk** variable, and then click the right arrow (\(\Rightarrow\)) next to the **Strata:** box. Note that this variable was removed completely from the variable list.
5. Click to highlight the **Tx** variable, press and hold the **Ctrl** button on your keyboard, and then click the **Caff, Wt, and Anx** variables. Click the right arrow (\(\Rightarrow\)) next to the **Covariates:** box.
6. Next, click the **Tx** variable, press and hold the **Ctrl** button on your keyboard, and then click the **T_COV_[T_COV_]** variable. With both variables highlighted, click the \(>a*b>\) button (\(\Rightarrow\)) next to the **Covariates:** box. Follow this same procedure to include the interaction for **Wt**. That’s it! Click **OK** to conduct the analysis.
7. Click the **Categorical…** button.
   a. Click to highlight the **Tx** variable, click the right arrow (\(\Rightarrow\)) next to the **Categorical Covariates:** box. In the **Change Contrast** section, click the **First** radio button and then click **Change**. Click **Continue**.
8. Click the **Options…** button.
   a. Under the **Model Statistics** section, click **CI for exp(B)** and leave the default confidence interval level at **95%**. Click **Continue**.
9. That’s it! Click **OK** to conduct the analysis.

Output

When you click **OK**, SPSS will produce an **Output** screen displaying the results. Click the **Output** window to view your results (if it does not pop up automatically). The first table you see is titled **Case Processing Summary**, which contains important information about your sample including the frequency and percentage of participants experiencing the event (**Event**).
Censored, the Total number of participants, as well as participants with missing values, negative time, and those that were left-censored.

The next table is titled Stratum Status and is presented in Figure 46.8. This table contains frequency counts for participants who experienced the event or were censored within each stratum. Here, because Smk had two levels, there are two rows of data.

<table>
<thead>
<tr>
<th>Stratum</th>
<th>Strata label</th>
<th>Event</th>
<th>Censored</th>
<th>Censored Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No</td>
<td>12</td>
<td>8</td>
<td>40.0%</td>
</tr>
<tr>
<td>1</td>
<td>Yes</td>
<td>5</td>
<td>5</td>
<td>50.0%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>17</td>
<td>13</td>
<td>43.3%</td>
</tr>
</tbody>
</table>

a. The strata variable is Smk

**Figure 46.8**

The next table, titled Categorical Variable Codings, is presented in Figure 46.9. This table contains information that is incredibly important to interpretation as it indicates how each categorical covariate was coded as well as the number of participants within each category. Notice that while the smoking status variable was considered categorical in Figure 46.8, it is not in this table.

<table>
<thead>
<tr>
<th>Categorical Variable Codings</th>
<th>Frequency</th>
<th>(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tx^a 0=Placebo</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>1=Treatment</td>
<td>15</td>
<td>1</td>
</tr>
</tbody>
</table>

a. Indicator Parameter Coding

b. Category variable: Tx

**Figure 46.9**

The next section of the output is titled Block 0: Beginning Block. This block can be thought of as similar to the Block 0 from a logistic regression analysis. That is, it contains information for the model without the independent variable and serves as the comparison or reference model. The table in this section is titled Omnibus Tests of Model Coefficients and contains the -2 Log Likelihood value.

Next, the Block 1: Method=Enter section is shown which contains the results of the Cox regression analysis. The first table, titled Omnibus Tests of Model Coefficients, shown in Figure 46.10, contains the overall test of the model based on a statistical comparison between Block 0 and Block 1. This table contains the -2 Log Likelihood value as well as omnibus tests for the Overall (score) model as well as Change from Previous Step and Change from Previous Block. Note that the latter two sections of this table provide identical results because we did not use hierarchical, sequential, or stepwise Cox regression. Further, note that the Chi-square value from the Change From Previous Block section is the likelihood ratio statistic (i.e., $G^2$) and is used when calculating effect size.
The next table is titled *Variables in the Equation* and is presented in Figure 46.11. This is the most important table in the output because it contains the results of the independent variable after controlling for the covariates. You will notice that there are no parameter estimates for the *Smk* variable. Remember, when we set smoking status as a stratification variable, we knew no test statistics would be provided for this variable. Further, you will notice that there is no constant (i.e., *y*-intercept), as provided in both linear and logistic regression. This is because the baseline hazard function is unspecified, or, stated another way, the baseline function is dependent on the specific time point. This table contains the regression slope (*B*), standard error (*SE*), *Wald* statistic, degrees of freedom (*df*), *p*-value (*Sig.*), hazard ratio (*Exp(B)*) as well as the 95% confidence interval around the hazard ratio (*95.0% CI for Exp(B)*).

![Figure 46.10](image)

<table>
<thead>
<tr>
<th>Variables in the Equation</th>
<th>B</th>
<th>SE</th>
<th><em>Wald</em></th>
<th><em>df</em></th>
<th><em>Sig.</em></th>
<th><em>Exp(B)</em></th>
<th>95.0% CI for Exp(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tx</strong></td>
<td>17.229</td>
<td>5.807</td>
<td>8.802</td>
<td>1</td>
<td>.003</td>
<td>3.036E7</td>
<td>346.095 2.664E12</td>
</tr>
<tr>
<td><strong>Caff</strong></td>
<td>.001</td>
<td>.006</td>
<td>.022</td>
<td>1</td>
<td>.881</td>
<td>1.001</td>
<td>.990 1.012</td>
</tr>
<tr>
<td><strong>Wt</strong></td>
<td>307</td>
<td>.120</td>
<td>5.729</td>
<td>1</td>
<td>.017</td>
<td>1.359</td>
<td>1.057 1.747</td>
</tr>
<tr>
<td><strong>Anx</strong></td>
<td>-181</td>
<td>.988</td>
<td>3.377</td>
<td>1</td>
<td>.066</td>
<td>.335</td>
<td>.688 1.012</td>
</tr>
<tr>
<td><strong>T_COV_Wt</strong></td>
<td>-0.98</td>
<td>.041</td>
<td>5.712</td>
<td>1</td>
<td>.017</td>
<td>.906</td>
<td>.836 .982</td>
</tr>
<tr>
<td><strong>T_COV_Tx</strong></td>
<td>-4.855</td>
<td>1.722</td>
<td>7.948</td>
<td>1</td>
<td>.005</td>
<td>.008</td>
<td>.000 .228</td>
</tr>
</tbody>
</table>

![Figure 46.11](image)

Finally, you are presented a table titled *Covariate Means*, which presents the mean for each covariate in the analysis. This table would normally be used to calculate and interpret the survival or hazard functions; however, because the shape of the function changes as time passes, these functions are not presented.

**Interpretation**

At this point, we would calculate the deviance residuals as described in Chapter 43. However, because we have included time dependent covariates, the deviance residuals are also time dependent and cannot be evaluated by a static plot. Thus, this assumption cannot be evaluated appropriately and is considered satisfied.
Continuing forward, because all testable assumptions were evaluated prior to analysis, interpretation begins by evaluating the Omnibus Tests of Model Coefficients table in the Block 1: Method=Enter section shown in Figure 46.10. Consider only the omnibus test found in the Change From Previous Block columns. The Chi-square value of 25.078 is statistically significant as \( p < .0005 \). Note that this value was calculated as the difference between the -2 Log Likelihood from Block 0: Beginning Block (i.e., 84.363) and the -2 Log Likelihood from Block 1: Method=Enter (i.e., 59.285) sections. This value, termed the likelihood ratio statistic (or \( G^2 \)), is used when calculating the effect size estimate. As stated above, the effect size estimate is calculated as: 
\[
1 - \exp\left(-\frac{G^2}{n}\right)
\]
Thus, the \( R^2 \) estimate equals .567 (i.e., \( 1 - \exp(-25.078/30) \)), representing the relative association between survival time and the set of covariates. In addition, degrees of freedom are calculated as the difference in the number of parameters between models; that is, the model in Block 1 has eight six parameters (i.e., four covariates and two variable-by-time interactions) than the Block 0.

Next, you evaluate the effect of the independent variable and variable-by-time interaction found in the Variables in the Equation table shown in Figure 46.11. Because we are only concerned with the difference between treatment group after adjusting for the other covariates, we evaluate the regression coefficient (\( B \)) and standard error (S.E.) as well as the hazard ratio (\( \exp(B) \)) and confidence interval (95.0% C.I. for \( \exp(B) \)) only for this variable and its interaction with time. Both the main effect and interaction are statistically significant. However, note that large standard errors and wide confidence intervals indicate poor measurement. This statement is confirmed here as the standard error for the \( Tx \) variable is massive, which is more than likely a direct result of using a small sample size. Remember, standard errors are greatly influenced by sample size. In addition, note that a 95% confidence interval for both variables does not contain 1, indicating the variables are statistically significant at \( p < .05 \).

As stated in Chapter 45, interaction effects adjust main effects. Because our independent variable of interest is categorical, the treatment group-by-time interaction produces either an increase or decrease in the difference between groups across time. Please refer to Chapter 45 for a full description of interpreting interaction effects. It is important to note that the primary difference between the interaction effects discussed in Chapter 45 and those in this Chapter is that now the interaction effects are adjusted for all other covariates in the model, including the stratification variable.

Before we begin, it is important to remember, there is no baseline function specified for semiparametric Cox regression and the treatment group variable, \( Tx \), is coded \( 1 \) = treatment group and \( 0 \) = placebo. With this information in mind, the \( Tx \) value of 17.229 represents the simple main effect of the difference in log-hazard between the placebo and treatment group specifically when time = 0 after adjusting for all covariates. That is, setting time to 0 (i.e., \( \ln(1) \)) removes the interaction term (i.e., multiplying anything by 0 equals 0). Thus, at \( \ln(1) \), there is a 17.229 difference in log-hazards favoring the treatment group, where the hazard ratio was 30360000. This difference remained statistically significant after adjusting for the interaction between treatment group and time as well as the other covariates in the model. Because the \( Tx \) variable is time dependent, the difference in log-hazards favoring the treatment group gets smaller (less positive) by 4.855 for every natural log minute increase after controlling for the other covariates. Remember, the natural log is used because that is what we specified for time in Step 1a above.

Although I provided a plot of hazard ratios in most of the examples above, I will not do so here. This is primarily due to the differing baseline hazard functions that were calculated.
across the stratification variable. However, note the similarity between the results produced from the multiple Cox regression with time-varying covariates and the stratified Cox model with time-varying covariates. They are nearly identical! I point this out to show you that it does not really matter which method you choose; they both produce similar results. Only when you have a continuous covariate that violates the proportionality of hazards assumption will the two models have a possibility to produce divergent results, which would be primarily due to the loss of information from categorizing a continuous variable to fit a stratified model. However, this is not recommended, with the more appropriate approach being to model the variable-by-time interaction(s) in a multiple Cox regression analysis.

Example Results Section

A violation of the proportionality of hazards assumption was indicated as smoking status, body weight, and treatment group interacted significantly with time. Because smoking status was not of direct research interest, it was used as a stratification variable in all subsequent analyses. The interaction between body weight and time and treatment group and time were also included. Prior to analysis, no differences between withdrawn and remaining participants were indicated, the probability of survival was assumed constant over the entire study period, the event occurred at the time specified, independence of residuals was assured, and no multicollinearity or multivariate outliers were indicated. Further, no statistically significant interactions were indicated between the stratification variable and covariates.

The results of a stratified Cox regression with time-varying covariates indicated the variables as a set provided statistically significant prediction of the time required to fall asleep during the 60-minutes study period, $\chi^2_{6} = 25.078, p < .05, R^2 = .567$. Specifically at time 0, participants receiving the new sleep treatment had a 17.229 increase in the log-hazards and were 30360000 times more likely to fall asleep during the study period compared to participants receiving placebo after adjusting for caffeine intake, body weight, smoking status, and anxiety level. However, this effect diminished quickly as each natural log minute increase decreased the difference in log-hazard by 4.855 units.