1.1 Read me first!

This book is filled with examples I hope you will try on your own. This section shows how you can download the example datasets and supplemental programs.

1.1.1 Downloading the example datasets and programs

All the example datasets and programs can be downloaded from within Stata using the following commands:

```
. net from http://www.stata-press.com/data/sbs
. net install sbs
. net get sbs
```

The `net install` command downloads the programs, I wrote for the book. These programs are listed below with a brief description and the chapters in which they are used:

- The `showcoding` command shows the coding of an original variable as compared with the recoded version. This program is used in chapter 13.
- The `power multreg` command computes power for simple and multiple regression analyses. It is used in chapter 19.
- The `power nestreg` command computes power for a nested regression. It is used in chapter 19.

The `net get` command downloads the example datasets. I encourage you to download these so that you can reproduce and extend the examples illustrated in this book.

1.1.2 Other user-written programs

The book also uses a number of user-written programs. All of these user-written programs are stored at the Statistical Software Components (SSC) repository and can be downloaded using the `ssc` command. The programs are described below along with the `ssc` command that you can use to download the program.

The fre command

This program shows frequencies of a variable with the values of the variable and the value labels. It is an alternative to the `tabulate` command for one-way tabulations. You can download it by typing

```
. ssc install fre, replace
```

The `fre` command, written by Ben Jann (2007a), is used in many chapters throughout the book.

The esttab command

This program creates formatted estimation tables. It is an extended version of the Stata `estimates table` command. You can download it by typing[1]

```
. ssc install estout, replace
```

The `esttab` program is extensively used in chapter 16, showing how to create presentation-quality regression tables. This program was also written by Jann (2007b).

The extremes command

The `extremes` command displays extreme values for a variable. You can download it from the SSC repository using the `ssc` command as shown below:

```
. ssc install extremes, replace
```

This program, written by Nicholas J. Cox (2003), is used in chapter 18 on regression diagnostics.

I am grateful to Jann and Cox for their kind permission to use their programs in my book. These programs not only helped to make my book better but also showed the kinds of user-written programs that have been created by the skilled and generous members of the Stata community. I touch on this point in section 1.2.7 later in this chapter.

1.2 Why use Stata?

I have extensively used and supported many statistical packages, both general and specialized. As someone whose background is in psychology and who feels strong connections to the behavioral sciences, I think there are many reasons for behavioral scientists to choose Stata as their first statistical package or to switch from their current package over to Stata, as described below.

1.2.1 ANOVA

Analysis of variance (ANOVA) is a cornerstone statistical technique of the behavioral sciences, especially factorial ANOVA with its ability to dissect the interactions to answer very meaningful substantive hypotheses. Stata offers exceptionally powerful (yet easy-to-use) tools that allow you to analyze and dissect results from ANOVA. Let me illustrate with an example.

1. The command `esttab` is part of the `estout` package of programs, so it is downloaded by typing `ssc install estout`.

Imagine a study in which subjects either are assigned to a control group or will receive a kind of therapy called optimism therapy, which is geared toward increasing optimism. All subjects in the study were screened for depression and only nondepressed subjects were included. At the conclusion of the study, the optimism of the participants was measured, and optimism was found to be greater in the optimism therapy group than in the control group. As we can see in figure 1.1, the optimism of those who received optimism therapy is 10 units greater than the optimism of the control group. The effect of optimism therapy is that it boosted optimism by 10 units.

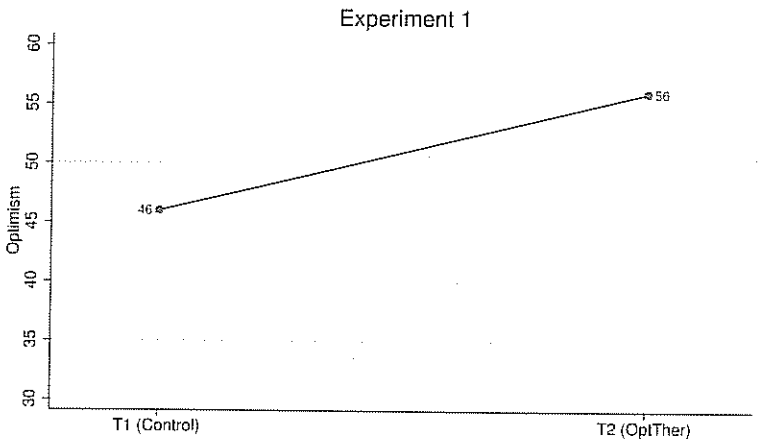

Figure 1.1: Results of experiment 1

The experimenter decides to replicate the previous study but with a twist: in addition to recruiting nondepressed participants, the experimenter will also recruit participants who are depressed. The experimenter expects that among those who are nondepressed, he or she will find the same kind of pattern as was found in the first study (that is, shown in figure 1.1). The experimenter wonders how those who are depressed will respond to this therapy. Will they equally benefit from the therapy? Will they benefit from the therapy but not as much as those who are nondepressed? Will they not benefit from the therapy at all? These three possible patterns are described further below and also illustrated in figure 1.2.

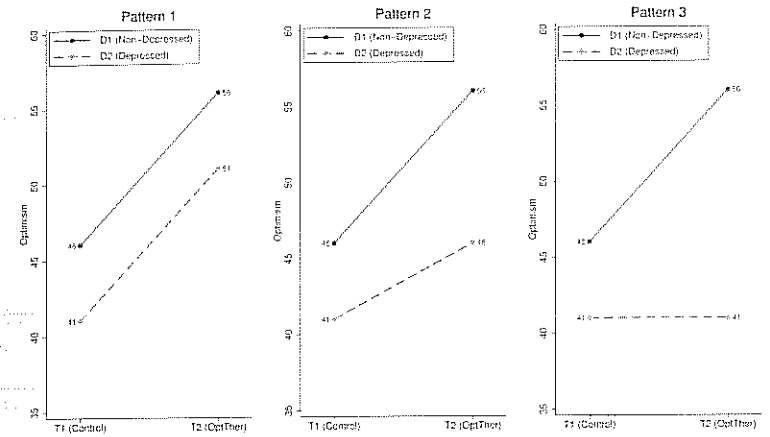

Figure 1.2: Three possible patterns of results from a 2 by 2 factorial design

Pattern 1. Optimism therapy benefits those who are depressed and nondepressed equally alike. Among those who are nondepressed, the effect of optimism therapy is 10 units (56 versus 46), the same gain as for those who are depressed (51 versus 41).

Pattern 2. Optimism therapy boosts optimism by 5 points for those who are depressed but does not yield the 10-point boost experienced by nondepressed participants. Optimism therapy is effective for those who are depressed but just not to the same degree as for those who are not depressed.

Pattern 3. Optimism therapy has no effect for those who are depressed (yielding no change from the control group, 41 versus 41), while optimism therapy delivered a 10-point boost in optimism for nondepressed participants.

The comparisons to distinguish patterns 1, 2, and 3 involve tests of interactions and tests to dissect the specific pattern of the interaction. As illustrated in chapter 7, Stata has exceptionally powerful, yet easy-to-use, tools for dissecting interactions with surgical precision.

But this is not the only reason I think Stata is an outstanding statistical package for behavioral scientists. Here are some other reasons.

1.2.2 Supercharging your ANOVA

Suppose the outcome of the previous study was changed to a binary outcome, necessitating the use of a logistic regression. Not to worry, because if you are using Stata, you can perform a two-by-two factorial logistic regression drawing upon the same concepts and tools as you could in performing a two-by-two ANOVA. The same can be said if you were using a count outcome and wanted to run a two-by-two Poisson regression model. In fact, Stata has brought ANOVA-like analysis technology to virtually all of its analysis commands. As a behavioral scientist, you can apply these familiar designs to an exceptionally wide variety of statistical models.

1.2.3 Stata is economical

Stata has very attractive academic pricing, including student pricing, that can put Stata into your hands (at the time this book was written) for $125 a year (or $54 a year if you use very small datasets). In addition, there is no extra price for extra modules. Some statistical software packages charge extra for modules that address missing data, the analysis of complex survey data, structural equation modeling, or bootstrapping. In fact, there are some specialized statistical packages that people buy to add these features to their existing statistical package. With Stata, all of these features are included at no extra cost.

Note: Stat/Transfer

If you want the ability to translate datasets from a variety of formats (for example, from Stata to SAS, from SPSS to Stata), then I suggest you consider purchasing Stat/Transfer, which you can obtain directly from StataCorp. Stat/Transfer can convert virtually any kind of dataset from one format to another format (for example, from Stata to SAS, from SPSS to Stata, from SAS to SPSS, and from Access to Excel). I have used Stat/Transfer for many years and been consistently impressed with its ease of use and how well it works. Speaking for myself, I cannot imagine doing my daily work without Stat/Transfer.

1.2.4 Statistical powerhouse

Stata is a statistical powerhouse in terms of its statistical features. Rather than listing all the features, I will point you to http://www.stata.com/features/ so that you can look at them yourself.

Tip: Stata/MP for multiprocessing

If you want to really unleash the power of Stata, consider Stata/MP. Taking advantage of multicore and multiprocessor computers, Stata/MP speeds up computations by using a divide and conquer strategy. Using 8 cores, the median time to complete an estimation command is 4.1 times faster. The *Stata/MP Performance Report* (available at http://www.stata.com/statamp/statamp.pdf) describes the concepts involved in parallel processing and details the performance gains achieved across Stata commands. Consider these performance gains that can be achieved on a 16-core machine—a linear regression using the `regress` command runs 16.5 times faster, a 2-way ANOVA using the `anova` command runs 12.7 times faster, and a logistic regression using the `logistic` command runs 11.1 times faster.

1.2.5 Easy to learn

Looking at the list of statistical features, you might feel overwhelmed wondering how you could learn all of these different commands. The amazing thing is how similarly the commands work. To run a regression predicting `y` from `x1`, `x2`, and `x3`, you type

```
. regress y x1 x2 x3
```

To run a robust regression predicting `y` from `x1`, `x2`, and `x3`, you type

```
. rreg y x1 x2 x3
```

To run a logistic regression predicting `y` from `x1`, `x2`, and `x3`, you type

```
. logistic y x1 x2 x3
```

To run a Poisson regression predicting `y` from `x1`, `x2`, and `x3`, you type

```
. poisson y x1 x2 x3
```

If Stata had a regression style command called `xyzreg`, you would likely be able to use it to predict `y` from `x1`, `x2`, and `x3` by typing

```
. xyzreg y x1 x2 x3
```

What could be simpler?

1.2.6 Simple and powerful data management

With its extensive statistical capabilities, I think many overlook the power of Stata for data management. Stata features many specialized commands that are like data management shortcuts, directly handling commonly difficult data management tasks. Example commands include `reshape long`, `reshape wide`, `egen`, `collapse`, and `merge`.

1.2.7 Access to user-written programs

One of the greatest virtues of Stata is the way it dovetails the ease of developing add-on programs with a great support structure for finding and downloading these programs. This virtue has led to a rich and diverse network of user-written Stata programs that extend the capabilities of Stata. As a result, the power of Stata is greatly extended and enhanced by these user contributions. Such programs are easily found, downloaded, and installed with the `search` command.

The `search` command connects to Stata's own search engine, which indexes user-written Stata programs from all around the world. Typing, for example, `search regression` searches for and displays Stata resources associated with the keyword `regression`. The resources searched include the Stata online help, Stata frequently asked questions (FAQs), the *Stata Journal* and its predecessor, the *Stata Technical Bulletin*, and programs posted on the websites of Stata users from around the world. All of these results are culled together and displayed in the Stata Viewer window. You can then point to and click on the programs you want to download and install.

> **Video tutorial: Downloading user-written programs**
>
> See a video demonstration of how to find and download user-written programs at http://www.stata.com/sbs/user-written.

Many of these programs are hosted at the SSC archive. This repository makes it easy for people to contribute programs to the Stata community and makes it easy for end users like you and me to easily download such programs.

You can see the newest additions to this archive by typing

```
. ssc new
```

You can also see the most popular downloads by typing

```
. ssc hot
```


> **Even more information: Downloading user-written programs**
>
> The *Getting Started* manual has more information about finding and downloading user-written programs. Just type `help gs`, and see the chapter titled "Updating and extending Stata—Internet functionality"

1.2.8 Point and click or commands: Your choice

You can use Stata with a point-and-click interface or by typing in commands. In this book, I focus exclusively on showing commands, but at any time, you can explore Stata via the drop-down menus, which give you point-and-click access to data management commands (via the **Data** drop-down menu), graphics commands (via the **Graphics** drop-down menu), and statistics commands (via the **Statistics** drop-down menu). Whenever you execute commands via the menus, Stata will display the command equivalent in the output window, teaching you the commands even as you use the point-and-click interface.

1.2.9 Powerful yet simple

In life, you often face the dilemma of choosing power or simplicity. With Stata, you can have both. It offers the simplicity of a point-and-click interface and commands that are simple to use; it also offers the power of being able to write your own programs using its Mata programming language.

1.2.10 Access to Stata source code

Stata is not just a statistical software program; it is a statistical programming environment. You can write your own Stata commands and programs, and you can view the source code for nearly all Stata commands. For example, would you like to see the source code for the `ttest` command? If so, just type

```
. viewsource ttest.ado
```

You can view the source code for nearly all Stata commands in this way. This allows you to see how every command works. Moreover, it means that you could even make your own version of any such Stata command with your own personal customizations.

1.2.11 Online resources for learning Stata

Another reason to use Stata is that there are so many terrific online resources to help you learn Stata.

The Stata resources and support page provides a comprehensive list of online resources available for Stata. It lists official resources available from StataCorp as well as from the Stata community. See http://www.stata.com/support/.

The Stata Resource links page provides a list of resources created by the Stata community to help you learn and use Stata; see http://www.stata.com/links/. Among the links included there, I highly recommend the UCLA Institute for Digital Research and Education (IDRE) Stata web resources at http://www.ats.ucla.edu/stat/stata/, which include FAQs, annotated Stata output, textbook examples solved in Stata, and online classes and seminars about Stata.

The Video tutorials on using Stata page contains links to numerous videos illustrating a wide variety of topics about Stata. These videos uniquely exploit the ability to show you about the use of Stata in a way that a written explanation cannot convey. Further, the videos are brief and to the point (usually lasting between two and five minutes). While you can find the videos on the StataCorp YouTube channel at http://www.youtube.com/user/statacorp, the "Video tutorials on using Stata" page (at http://www.stata.com/links/video-tutorials/) shows the videos organized by topic.

The Stata Frequently Asked Questions page is special because it not only contains many frequently asked questions but also includes answers! The FAQs cover common questions (for example, How do I export tables from Stata?) as well as esoteric questions (for example, How are estimates of rho outside the bounds $[-1, 1]$ handled in the two-step Heckman estimator?). You can search the FAQs using keywords, or you can browse the FAQs by topic. See http://www.stata.com/support/faqs/.

Statalist is an independently run web forum that connects Stata users from all over the world. It began in 1994 as a listserv (hence the name "Statalist") and was relaunched in March 2014 as a web forum. The community is both extremely knowledgeable and friendly, welcoming questions from newbies and experts alike. Even if you never post a question, you can learn quite a bit by reading the questions and answers posted by others. You can visit the web forum at http://www.statalist.org/. And if you wish to read questions and answers that predate the web forum (going all the way back to 2002), you can visit the "Statalist archives" page at http://www.stata.com/statalist/archive/.

The Stata Blog covers many interesting and technical aspects of Stata. Entries are written by Stata's developers and technical support team; see http://blog.stata.com/.

The Stata Journal is published quarterly with articles that integrate various aspects of statistical practice with Stata. Although current issues and articles are available by subscription, articles over three years old are available for free online as PDF files. See http://www.stata-journal.com/.

The Stata Technical Bulletin is the predecessor of the *Stata Journal*. All of these issues are available for free online. Although many articles may be out of date, there are many gems that contain timeless information. For more information, see http://www.stata.com/bookstore/individual-stata-technical-bulletin-issues/.

> **Note: Help menu**
>
> It is easy to overlook or forget that the Stata **Help** drop-down menu is a central hub for directing you to many helpful resources, including the Stata documentation in PDF, the help files organized by content, and information about what's new in Stata. My favorite is the **Resources** item, which brings up *Resources for learning more about Stata*, which provides a concise and comprehensive list of resources to help you learn and use Stata. You can also access this help file by typing `help resources`.

1.2.12 And yet there is more!

For even more information and more reasons why you would enjoy using Stata, see the Stata webpage titled "Why use Stata" at http://www.stata.com/why-use-stata/.

1.3 Overview of the book

The book is divided into five parts, described below.

1.3.1 Part I: Warming up

As implied by the title, this part warms us up for the more substantial parts. This includes the current chapter you are reading. The next chapter, chapter 2, covers descriptive statistics such as tabulations (frequency distributions), summary statistics, cross-tabulations, and summary statistics for specific subgroups. This part concludes with chapter 3, which introduces basic inferential statistics such as two-sample t tests, one-sample t tests, and one- and two-sample tests of proportions.

1.3.2 Part II: Between-subjects ANOVA models

This part covers between-subjects ANOVA models, beginning with chapter 4, which covers one-way between-subjects ANOVA. This is followed by chapter 5, which illustrates contrasts that you can use for making comparisons among groups in a one-way ANOVA. Next, chapter 6 covers analysis of covariance (ANCOVA), illustrating its use in experimental designs (to increase power) and in nonexperimental designs (to attempt to statistically control for confounding variables). Chapter 7 introduces factorial designs,

covering two-by-two designs, two-by-three designs, and three-by-three designs. This chapter emphasizes how to visualize and interpret the interactions. It also illustrates how to dissect two-way interactions using simple effects, simple contrasts, partial interactions, and interaction contrasts. The `contrast` command is illustrated for dissecting the interactions, and the `margins` and `marginsplot` commands are used to display and graph the means associated with the interactions. Chapter 8 illustrates ANCOVA-type analyses with the focus on interactions of the independent variable (IV) and the covariate (in other words, categorical by continuous variable interactions). Chapter 9 covers factorial models with three IVs. This chapter, like chapter 7, emphasizes the visualization and interpretation of the interactions, in this case focusing on the three-way interactions. Figures are used to visually understand the three-way interactions, and a variety of analytic and graphical methods is illustrated to also help understand them. Chapter 10 shows how you can extend the power of ANOVA by blending the ANOVA designs with regression commands. This chapter illustrates how you can extend your ANOVA design to analyze data that come from complex surveys, data that violate the homogeneity of variance assumption, or data with influential observations (via robust regression or quantile regression). This part concludes with chapter 11, which illustrates power analysis for ANOVA and ANCOVA.

1.3.3 Part III: Repeated measures and longitudinal models

This part covers two different strategies for analyzing designs with multiple observations on the same subject.

Chapter 12 covers repeated measures ANOVA designs. In such designs, participants are observed at more than one time point. All participants are observed according to the same time schedule. This chapter shows three examples illustrating the analysis of repeated measures designs. The first example includes a single repeated measures IV (see section 12.2). The second example illustrates a two-by-three between-within design where the between-subjects IV has two levels and the repeated measures IV has three levels (see section 12.3). The third example illustrates a three-by-three between-within design where the between-subjects IV has three levels and the repeated measures IV also has three levels (see section 12.4).

Chapter 13 covers longitudinal models. These models, in contrast to repeated measures designs, typically have a larger number of observations per subject, and the time gaps between the repeated measures can vary between people. This chapter includes four examples, all of which use multilevel modeling as the main analysis strategy. The first example models the dependent variable (DV) as a linear function of time (see section 13.2). The second example adds a between-subjects IV, which allows us to model the linear effect of time and explore IV by time interaction (see section 13.3). This example is similar to an ANCOVA with a treatment by covariate interaction (for instance, like the examples in chapter 8). The third example includes time as the only predictor but uses a piecewise modeling strategy for the effect of time (see section 13.4). The fourth example adds a between-subjects IV to the third example, modeling the interaction of the IV with the piecewise effects of time (see section 13.5).

1.3.4 Part IV: Regression models

This part illustrates how to use Stata commands to fit regression models. The chapters are ordered like a meal in which you decide to eat dessert first. The sweet and delicious chapters are presented first, deferring nutritional topics such as regression diagnostics and power analysis to the end. This part begins with chapter 14, which shows you how to perform multiple regression using Stata (showing you how to fit a simple linear regression model and multiple regression models) and how to test multiple coefficients within a multiple regression model. Chapter 15 covers more details about using the `regress` command, showing options for customizing output and how to create summary statistics based on the sample of observations included in the most recent regression analysis. This chapter also shows you how to store results of regression models for use later in your Stata session. This feature is used in chapter 16, which shows tools that you can use to create formatted regression tables. The chapter illustrates how to create such tables for display on the screen and how to create customized formatted output that can be used within a word processor like Word to create presentation-quality regression tables. Chapter 17 illustrates model-building tools, showing you how to fit multiple models using the same sample of observations. The chapter also shows you how to fit nested regression models and perform stepwise regression models. Chapter 18 illustrates commands for performing regression diagnostics, demonstrating analytic and graphical methods for identifying outliers. The chapter also illustrates analytic and graphical methods that you can use for testing for nonlinearity and how you can detect multicollinearity, assess the homoskedasticity assumption, and evaluate the normality of the residuals. This part concludes with chapter 19, which illustrates how to perform power analysis for a simple regression model and a multiple regression model and how to compute power for a nested multiple regression model.

1.3.5 Part V: Stata overview

The previous parts have focused on different statistical techniques, providing examples of how to perform analyses using those techniques in Stata. This part provides command-centric information, offering an overview of the use of Stata.

The first chapter (chapter 20) shows common features of estimation commands. Even though Stata has a very large number of estimation commands, they share a number of common features. This is by design, not by accident. Because estimation commands work similarly, what you learn about the behavior of one estimation command translates over to the use of other estimation commands. This chapter is about the features (behaviors) that estimation commands share.

Chapter 21 discusses a special set of commands called postestimation commands. They are called this because they are used after an estimation command (for example, after the `anova` or `regress` command). In particular, this chapter provides additional details about the `contrast`, `margins`, `marginsplot`, and `pwcompare` commands.

Chapter 22 provides brief information about basic data management commands in Stata. It illustrates reading data into Stata, keeping and dropping variables and observations, labeling data, creating variables, appending datasets, merging datasets, and reshaping datasets wide to long and long to wide.

The final chapter, chapter 23, recognizes that many readers of this book might be familiar with IBM® SPSS®. If you are such a reader, you might find yourself asking, for a given SPSS command, What is the equivalent Stata command? To answer such questions, this chapter lists commonly used SPSS commands (in alphabetical order) and shows the equivalent (or near equivalent) Stata command along with a brief example of the Stata command.

1.3.6 The GSS dataset

One of the commonly used datasets in this book is based on the General Social Survey (GSS). The GSS dataset was created and is collected by the National Opinion Research Center (NORC). To learn more, see http://www.norc.uchicago.edu/GSS+Website/. The GSS is a unique survey and dataset. It contains numerous variables measuring demographics and societal trends from 1972 to 2012. This is a cross-sectional dataset; thus the data for each year represent different respondents. (Note that the GSS does have a panel dataset for 2006, 2008, and 2010, but this is not used here.) In some years, certain demographic groups were oversampled. For simplicity, I will overlook this and treat the sample as though simple random sampling was used.

The version of the dataset we will use for the book is based on the GSS from 2012. This dataset was accessed by visiting http://www3.norc.org/GSS+Website/Download/STATA+v8.0+Format/ and looking under the heading "Download Individual Year Data Sets (cross-section only)" and the subheading "GSS 1972–2012 Release 6". Clicking on the link *2012*, I downloaded a file named `2012_stata.zip`, and unzipping that file yielded the dataset named `GSS2012.dta`. I created a Stata do-file that subsets and recodes the variables to create the analytic data file we will use, named `gss2012_sbs.dta`. This dataset is used below.

```
. use gss2012_sbs
```


The `describe` command shows that the dataset contains 1,974 observations and 42 variables.

```
. describe, short
Contains data from gss2012_sbs.dta
  obs:         1,974
 vars:            42                          17 Jul 2015 09:41
 size:       157,920
Sorted by:
```

Note that http://www3.norc.org/GSS+Website/Download/STATA+v8.0+Format/ provides some key information about missing value codes.

There are four missing values in the data:

- `.c`: Cannot choose.
- `.i`: Inapplicable. Respondents who are not asked to answer a specific question are assigned to IAP.
- `.d`: Don't know.
- `.n`: No answer.

This suggests that the special missing value code of `.c` indicates that the value is missing because the respondent could not choose a rating that reflected his or her happiness. The missing value code of `.i` indicates that the response is missing because the question was not asked of the respondent. (Note that some groups of questions are asked of only some randomly chosen respondents.) The missing value code of `.d` means that the answer is missing because the respondent did not know. Finally, a missing value code of `.n` means that there is no answer (for example, the respondent preferred not to respond). For more documentation about the GSS, you can visit http://www3.norc.org/GSS+Website/Documentation/. You can also learn more about missing value codes in Stata by typing `help missing`.

1.3.7 Language used in the book

I would like to comment on the language used in this book. I use language from the tradition of experimental design and the behavioral sciences. Here are some of the terms that I will be using and my intended meaning.

Independent and dependent variables. In ANOVA, it is traditional to call the categorical predictor an IV and the outcome the DV. This usage reflects the tradition that ANOVA was most commonly used for the analysis of designed experiments. I will continue to describe the categorical predictor as the IV and the outcome as the DV—even if the study is not a designed experiment and even if the design is not an ANOVA design. I choose this terminology to emphasize that the role these variables play (in a statistical sense) are the same, even if they do not arise from a designed experiment and even if they are not used in a traditional ANOVA analysis.

Note: Factor variables

Sometimes, when referring to an IV or a categorical variable, I will also use the term "factor variable". This is a Stata-specific term that refers to a categorical variable that has been entered into an ANOVA model or into a regression model by using the `i.` prefix (for example, `i.race`). Stata treats factor variables differently, knowing that they are categorical variables, and most Stata commands understand how to treat such variables differently from continuous variables (for example, age).

Covariate. In an ANCOVA design, a covariate is a continuous predictor, in addition to the IV in the prediction of the DV.

Effect. It can be very parsimonious to talk about a "treatment effect" when talking about the difference in the means for a treatment versus control group. Or in the context of regression, it can be useful to call the regression coefficient for a variable (such as age) the "effect of age". Whenever I use this term, I am not using it in the context of cause and effect but to describe an observed statistical relationship.

Note! Internal validity

A key question in many studies is whether statistically significant associations reflect underlying causal relationships. The issue of causal inference concerns the scientific integrity study design, and the ability to draw such causal conclusions is often described as the "internal validity" of the study. In this book, I will sidestep such issues but instead refer you to your favorite book in experimental methods for more information about the conditions that are necessary for drawing causal conclusions regarding statistically significant findings.

1.3.8 Online resources for this book

The online resources for this book can be found at the book's website:

http://www.stata-press.com/books/sbs.html

Resources you will find there include the following:

- All the datasets used in the book. I encourage you to download the datasets, reproduce the examples, and try variations on your own. You can download all the datasets into your current working directory from within Stata by typing

```
. net from http://www.stata-press.com/data/sbs
. net get sbs
```

- Errata (which I hope will be short or blank). Although I have tried hard to make this book error free, I know that some errors will be found, and they will be listed in the errata.
- Other resources that may be placed on the site after this book goes to press. Be sure to visit the site to see what else may appear there.

1.4 Recommended resources and books

This book focuses on how to analyze data from a behavioral science perspective and is not a general purpose book about the overall use of Stata. It omits topics such as an overall introduction to Stata and general principles of using Stata and provides very little details about data management or graphics. I made this deliberate choice because there are so many other resources that cover these topics and because the coverage of these topics is not specific to a behavioral scientist. Thus here I provide recommendations for resources to help you acquire this information.

1.4.1 Getting started

If you are new to Stata, I highly recommend the Stata *Getting Started* manual. There is a unique version of the *Getting Started* manual that shows what Stata will look like and how it works on your platform. (The manual comes in three separate versions written for Windows, Mac, and Unix.) You can access the *Getting Started* manuals by typing

```
. help gs
```

To get you started, I suggest you read the chapter titled "Introducing Stata—sample session". In addition, the chapters titled "The Stata user interface", "Using the Viewer", and "Getting help" should help you quickly feel comfortable in the Stata environment.

Video tutorial: The Stata interface

Take a video tour of the Stata interface at http://www.stata.com/sbs/interface.

1.4.2 Data management in Stata

The *Getting Started* manual shows how to get data into Stata in the chapters titled "Opening and saving Stata datasets" and "Importing data". It also covers "Creating new variables" and "Deleting variables and observations".

For more information about general topics in data management, I recommend the IDRE (formerly Academic Technology Services [ATS]) UCLA website. There is a special page devoted to the topic of data management at http://www.ats.ucla.edu/stat/stata/topics/data_management.htm.

For comprehensive coverage of the topic of data management (from basic tasks such as labeling variables and recoding variables to advanced tasks such as merging or reshaping datasets), I recommend my book titled *Data Management Using Stata: A Practical Handbook*, which is available at http://www.stata.com/bookstore/data-management-using-stata/.

Video tutorial: Getting help in Stata

See a video demonstration of how to get help in Stata at http://www.stata.com/sbs/help.

1.4.3 Reproducing your results

One topic that I have not addressed but is extremely important concerns how you can reproduce your results. This book illustrates commands that you can use for performing your analyses, but it does not show you how to create a procedure for saving these commands so you can easily execute them again. A related topic is how to save the results of your commands so that you can refer to the results in the future. The *Getting Started* manual has an excellent introduction to those topics. The chapter "Using the Do-file Editor—automating Stata" shows you how to save a sequence of Stata commands in a file called a do-file, which you can execute at a later time.[2] The chapter "Saving and printing results by using logs" shows you how to can save your results in a log file, which provides you a transcript of your commands and output from previous analyses. These two features can be combined so that your do-files automatically generate log files.

2. For IBM® SPSS® users, this is the equivalent of an SPSS syntax file.

The topic of creating and using do-files is also covered in my book *Data Management Using Stata: A Practical Handbook*.

Video tutorial: PDF documentation in Stata

Did you know that Stata has well over 12,000 pages of documentation that are just one click away? From the **Help** menu, click on **PDF documentation**. You can also see a video demonstration of how you can access PDF documentation within Stata at http://www.stata.com/sbs/pdf-documentation.

1.4.4 Recommended Stata Press books

StataCorp has a publishing arm called Stata Press, which issues books like this one. You can see a list of all the books at http://www.stata-press.com/catalog/. At that site, you can see a description of each book, including a detailed table of contents, comments from the Stata technical group, and a sample chapter. Every Stata Press book I have read has excelled in providing useful information about the use of Stata for researchers. Among these books, I particularly recommend the following as books that build upon what is presented in this book:

- *Discovering Structural Equation Modeling Using Stata, Revised Edition* by Alan C. Acock (2013). This book provides an excellent introduction to the use of structural equation modeling using Stata.
- *An Introduction to Survival Analysis Using Stata, Third Edition* by Mario Cleves, William Gould, Roberto G. Gutierrez, and Yulia V. Marchenko (2010). This book provides excellent and detailed information about how to perform survival analysis using Stata.
- *Multilevel and Longitudinal Modeling Using Stata, Third Edition (Volumes I and II)* by Sophia Rabe-Hesketh and Anders Skrondal (2012b). This two-volume set provides extensive and very detailed information about fitting multilevel and longitudinal models using Stata.
- *An Introduction to Stata Programming* by Christopher F. Baum (2009). This is an excellent book to help you learn and explore the power of Stata programming.

I would be remiss if I did not mention my other books published by Stata Press, listed below.

- *Data Management Using Stata: A Practical Handbook* by Michael N. Mitchell (2010). This book provides data management information to complement the statistical examples shown in the book you are holding.
- *A Visual Guide to Stata Graphics, Third Edition* by Michael N. Mitchell (2012b). This book visually illustrates the use of Stata graphics.

- *Interpreting and Visualizing Regression Models Using Stata* by Michael N. Mitchell (2012a). This book overlaps with the book you are holding, covering many of the same topics but discussing them in a way that would appeal to a more general Stata audience. It discusses modeling of categorical, continuous, and categorical and continuous interactions in a much more general fashion than the book you are holding.

2 Descriptive statistics

2.1 Chapter overview

This chapter introduces how to perform descriptive statistics. It begins with section 2.2, which introduces the General Social Survey (GSS) dataset as well as some of the Stata commands you can use to become familiar with a new dataset. This is followed by a series of sections that illustrate how to perform different kinds of descriptive statistics, namely, one-way tabulations (section 2.3), summary statistics (section 2.4), summary statistics by one group (section 2.5), two-way tabulations (section 2.6), and summary statistics by two groups (section 2.7).

2.2 Using and describing the GSS dataset

The examples from this chapter are based on analyses of the GSS from the year 2012 using a dataset named `gss2012_sbs.dta`. Once you have downloaded the datasets for the book (as described in section 1.1), you can load this dataset into Stata by typing

```
. use gss2012_sbs
```

We can use the `describe` command to obtain information about the dataset, including the number of observations, the number of variables, and a listing of all the variables and labels. The command is shown below, but the output is very long, so I have omitted it to save space. I suggest you try the command so that you can see the output for yourself.