# 1 Introduction

## 1.1 What is this book about?

Our book shows you efficient and effective ways to use regression models for categorical and count outcomes. It is a book about data analysis and is not a formal treatment of statistical models. To be effective in analyzing data, you want to spend your time thinking about substantive issues and not laboring to get your software to generate the results of interest. Accordingly, good data analysis requires good software and good technique.

Although we believe that these points apply to all data analysis, they are particularly important for the regression models that we examine. The reason is that these models are *nonlinear*, and consequently the simple interpretations that are possible in linear models are no longer appropriate. In nonlinear models, the effect of each variable on the outcome depends on the level of *all* variables in the model. Because of this nonlinearity, which we discuss in more detail in chapter 3, no method of interpretation can fully describe the relationships among the independent variables and the outcome. Rather, a series of *postestimation* explorations are needed to uncover the most important aspects of these relationships. In general, if you limit your interpretations to the standard output, that output constrains and can even distort how you understand your results.

In the linear regression model (LRM), most of the work of interpretation is complete once the estimates are obtained. You simply read off the coefficients, which can be interpreted as for a unit increase in $x_k$, $y$ is expected to increase by $\beta_k$ units, holding all other variables constant. In nonlinear models, such as logit or negative binomial regression, much more computation is necessary after the estimates are obtained. With few exceptions, the software that fits regression models does not provide much help with these analyses. Consequently, the computations are tedious, time consuming, and error prone. All in all, it is not fun work. In this book, we show how postestimation analysis can be accomplished easily using Stata and the set of new commands that we have written. These commands make sophisticated postestimation analysis routine and even enjoyable. With the tedium removed, the data analyst can focus on the substantive issues.

## 1.2 Which models are considered?

Regression models analyze the relationship between an explanatory variable and an outcome variable while controlling for the effects of other variables. The linear regression model (LRM) is probably the most commonly used statistical method in the social sciences. As mentioned, a key advantage of the LRM is the ease of interpreting results. Unfortunately, this model applies only to cases in which the dependent variable is continuous.[1] Using the LRM when it is not appropriate produces coefficients that are biased and inconsistent, and there is nothing advantageous about the simple interpretation of results that are incorrect.

Fortunately, many appropriate models exists for categorical outcomes, and these models are the focus of our book. We cover cross-sectional models for four kinds of dependent variables.

*Binary* outcomes (dichotomous or dummy variables) have two values, such as whether a citizen voted in the last election, whether a patient was cured after receiving some medical treatment, or whether a respondent attended college.

*Ordinal* or *ordered* outcomes have more than two categories, and these categories are assumed to be ordered. For example, a survey might ask if you would be "very likely", "somewhat likely", or "not at all likely" to take a new subway to work, or if you agree with the president on "all issues", "most issues", "some issues", or "almost no issues".

*Nominal* outcomes also have more than two categories but are not ordered. Examples include the mode of transportation a person takes to work (e.g., bus, car, train) or an individual's employment status (e.g., employed, unemployed, out of the labor force).

Finally, *count* variables count the number of times something has happened, such as the number of articles written by a student after receiving the Ph.D., or the number of patents a biotechnology company has obtained.

The specific cross-sectional models that we consider, along with the corresponding Stata commands, are

**Binary outcomes:** binary logit (`logit`), binary probit (`probit`), and the complementary log-log regression model (`cloglog`)

**Ordinal and nominal outcomes:** ordered logit (`ologit`) and ordered probit (`oprobit`) for ordinal outcomes; the stereotype logistic regression (`slogit`) for ordinal and nominal outcomes; multinomial logit (`mlogit`), multinomial probit with uncorrelated errors (`mprobit`), conditional logit (`clogit`), and alternative-specific multinomial probit with correlated errors (`asmprobit`) for nominal outcomes; and rank-ordered logit (`rologit`) for ranked outcomes

1. Using the LRM with binary dependent variables leads to the linear probability model (LPM). We do not consider the LPM further, given the advantages of models such as logit and probit. See Long (1997, 35–40) for details.

**Count outcomes:** Poisson regression (`poisson`), negative binomial regression (`nbreg`), zero-truncated Poisson (`ztp`), zero-truncated negative binomial (`ztnb`), hurdle regression, zero-inflated Poisson regression (`zip`), and zero-inflated negative binomial regression (`zinb`)

Although this book covers models for many different types of outcomes, they are all models for cross-sectional data. We do not consider models for survival or event-history data, even though Stata has a powerful set of commands for dealing with these data (see the entry for `st` in the *Survival Analysis Reference Manual*). Likewise, we do not consider any models for panel data, even though Stata contains commands for fitting these models (see the entry for `xt` in the *Longitudinal/Panel Data Reference Manual*).

## 1.3 Whom is this book for?

We expect that readers of this book will vary considerably in their knowledge of both statistics and Stata. With this in mind, we have tried to structure the book to accommodate the diversity of our audience. Minimally, however, we assume that readers have a solid familiarity with ordinary least-squares regression for continuous dependent variables and that they are comfortable using the basic features of the operating system of their computer. Although we have provided sufficient information about each model so that you can read each chapter without prior exposure to the models discussed, we strongly recommend that you do *not* use this book as your sole source of information on the models (section 1.6 recommends more readings). Our book will be most useful if you have already studied the models considered or are studying these models in conjunction with this book.

We assume that you have access to a computer that is running Stata 9 or later and that you have access to the Internet to download commands, datasets, and sample programs that we have written (see section 1.5 for details on obtaining these). For information about obtaining Stata, see the StataCorp web site at *http://www.stata.com*. Although most of the commands in later chapters also work in Stata 8 and Stata 7, there are some differences. For details, see our web site at *http://www.indiana.edu/~jslsoc/spost.htm*.

## 1.4 How is the book organized?

Chapters 2 and 3 introduce materials that are necessary for working with the models we present in the later chapters:

**Chapter 2: Introduction to Stata** reviews the basic features of Stata that are necessary to get new or inexperienced users up and running with the program. This introduction is by no means comprehensive, so we include information on how to get more help. New users should work through the brief tutorial that we provide in section 2.17. Users already skilled with Stata can skip this chapter, although even these readers might benefit from quickly reading it.

**Chapter 3: Estimation, testing, fit, and interpretation** provides a review of using Stata for regression models. It includes details on how to fit models, test hypotheses, compute measures of model fit, and interpret results. We focus on those issues that apply to all the models considered in part II. We also provide detailed descriptions of the add-on commands that we have written to make these tasks easier. Even if you are an advanced user, we recommend that you look over this chapter before jumping ahead to the chapters on specific models.

Chapters 4–8 cover models for a different type of outcome:

**Chapter 4: Models for binary outcomes** begins with an overview of how the binary logit and probit models are derived and how they can be fitted. After the model has been fitted, we show how Stata can be used to test hypotheses, compute residuals and influence statistics, and calculate scalar measures of model fit. Then we describe postestimation commands that assist in interpretation using predicted probabilities, discrete and marginal change in the predicted probabilities, and, for the logit model, odds ratios. Because binary models provide a foundation on which some models for other kinds of outcomes are derived, and because chapter 4 provides more detailed explanations of common tasks than later chapters do, we recommend reading this chapter even if you are interested mainly in another type of outcome.

**Chapter 5: Models for ordinal outcomes** introduces the ordered logit and ordered probit models. We show how these models are fitted and how to test hypotheses about coefficients. We also consider two tests of the parallel regression assumption. In interpreting results, we discuss similar methods as those described in chapter 4, as well as interpretation in terms of a latent dependent variable.

**Chapter 6: Models for nominal outcomes with case-specific data** focuses on the multinomial logit model. We show how to test a variety of hypotheses that involve multiple coefficients and discuss two tests of the assumption of the independence of irrelevant alternatives. Although the methods of interpretation are again similar to those presented in chapter 4, the model's many parameters often complicate interpretation. To deal with this complexity, we present two graphical methods of representing results. The multinomial probit model without correlated errors is discussed briefly, and then the multinomial logit model is used to explain the stereotype logit model. This model, which is often used with ordinal outcomes, also has applications with nominal outcomes.

**Chapter 7: Models for nominal outcomes with alternative-specific data** introduces models for situations in which you have at least some variables that vary over the alternatives for each individual, such as an individual's similarity to each candidate in an election. We first show you how to rearrange data into the format required for these models. Then we describe the conditional logit model, which is equivalent to the multinomial logit model when only case-specific regressors are used, but which also allows alternative-specific regressors. Then

we discuss the alternative-specific multinomial probit model, which is more than just a probit version of the conditional logit model because it allows correlations between alternative-specific error terms, thus relaxing the assumption of the independence of irrelevant alternatives. Last, we present the rank-ordered logistic regression model, which can be used when you have information about the ranking of outcomes as opposed to only information about the selected or most preferred outcome.

**Chapter 8: Models for count outcomes** begins with the Poisson and negative binomial regression models, including a test to determine which model is appropriate for your data. We also show how to incorporate differences in exposure time into the estimation. Next we consider interpretation for changes in the predicted rate and changes in the predicted probability of observing a given count. The rest of the chapter deals with models that specifically address problems associated with having too many zeros or none at all. We start with zero-truncated models for which zeros are missing from the outcome variable, perhaps due to the way the data were collected. We then merge a binary model and a zero-truncated model to create the hurdle model. The rest of the chapter considers fitting and interpreting zero-inflated count models, which are designed to account for the many zero counts often found in count outcomes.

Chapter 9 returns to issues that affect all models.

**Chapter 9: More topics** deals with several topics, but the primary concern is with complications among independent variables. We consider the use of ordinal and nominal independent variables, nonlinearities among the independent variables, and interactions. The proper interpretation of the effects of these types of variables requires special adjustments to the commands considered in earlier chapters. Many of these examples involve writing small programs with macros and loops. We then comment briefly on how to modify our commands to work with other estimation commands. Finally, we discuss several features in Stata that we think make data analysis easier and more enjoyable.

## 1.5 What software do you need?

To get the most out of this book, you should read it while you are at a computer where you can experiment with the commands as they are introduced. We assume that you are using Stata 9 or later. If you are running Stata 8 or Stata 7, most of the commands work in the same way, but a few options will not work and the output might look different. The SPost commands that ran in Stata 8 (and usually with Stata 7) will continue to be available if you install the package `spostado`, with examples contained in `spostst8`.[2] The version 9 commands require Stata 9 or later.

2. `spostado` and `spostst8` are the names used in the prior edition of our book. We are keeping the names and contents of those packages unchanged so that they will continue to work as described in that edition. For Stata 9, the packages are called `spost9_ado` and `spost9_do`, following a naming scheme we plan to keep as new versions of Stata are released.

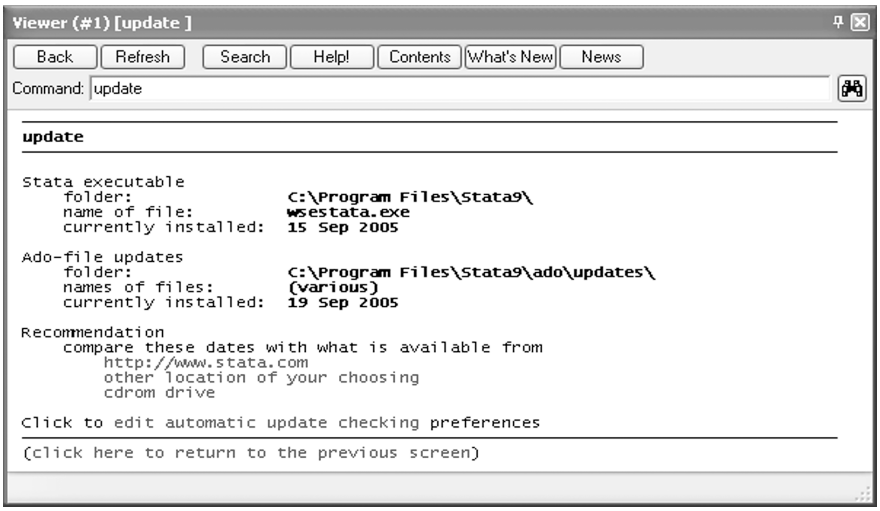
---

**Advice to new Stata users** If you have never used Stata, you might find the instructions in this section to be confusing. It might be easier if you only skim the material now and return to it after you have read the introductory sections of chapter 2.

---

### 1.5.1 Updating Stata 9

Before you work through our examples, we strongly recommend that you have the latest version of `wstata.exe` (or `wsestata.exe` if you are using Stata/SE) and the official Stata ado-files. *You should do this even if you have just installed Stata* because the CD that you received might not have the latest changes to the program. If you are connected to the Internet and are in Stata, you can update Stata by selecting Official Updates from the Help menu. Stata responds with the following screen:




This screen tells you the current dates of your files. If you do not have Stata set up to automatically update, you can update your files to the latest versions by clicking on `http://www.stata.com`. We suggest that you do this every few months. Or, if you encounter something that you think is a bug in Stata or in our commands, update your copy of Stata to see if the problem has been resolved. After you update, read the resulting screen carefully. You might need to click on `update swap` to complete the update.

## 1.5.2 Installing SPost

From our point of view, one of the best things about Stata is how easy it is to add your own commands. This means that if Stata does not have a command you need or some command does not work the way you like, you can program the command yourself, and it will work as if it were part of official Stata. Indeed, we have created a suite of programs, referred to collectively as SPost (for Stata Postestimation commands), for the postestimation interpretation of regression models. These commands must be installed before you can try the examples in later chapters.

---

**What is an ado-file?** Programs that add commands to Stata are contained in files that end in the extension `.ado` (hence the name). For example, `prvalue.ado` is the program for the command `prvalue`. Hundreds of ado-files are included with the official Stata package, but experienced users can write their own ado-files to add new commands. However, for Stata to use a command implemented as an ado-file, *the ado-file must be located in one of the directories where Stata looks for ado-files*. If you type the command `sysdir`, Stata lists the directories that Stata searches for ado-files in the order that it searches them. However, if you follow our instructions below, you should not have to worry about managing these directories.

---

### Installing SPost using search

Installation should be simple if you are connected to the Internet. If you have installed a prior version of SPost, we suggest that you begin by uninstalling it. To do this, enter the command `ado`. This will list all packages (a package is a collection of related files) that have been installed, with each package marked with a number in square brackets. Scan the list and record the number of the package containing the SPost ado-files and any other packages related to SPost. Although the ado-files are now contained in the package `spost9_ado`, other names were used in the past. Uninstall these packages with the command `ado uninstall [#]`, where # is the package number (you must include the brackets; for example, `ado uninstall [3]`). You must run `ado uninstall` once for each package to be uninstalled.

The `search` *word*`, net` command searches an online database that StataCorp uses to keep track of user-written additions to Stata. Typing `search spost, net` brings up the names and descriptions of several packages in the Results window. One of these packages should be labeled `spost9_ado from http://www.indiana.edu/~jslsoc/stata`. The label is in blue, which means that it is a clickable link.[3] After you click on the link, a new Viewer window opens with information about our commands and another link saying "click here to install". If you click on this link, Stata attempts to install the

3. If you click on a link and immediately get a beep with an error message saying that Stata is busy, Stata is probably waiting for you to press a key. Most often, this occurs when Stata is displaying output that does not fit on one screen.

package. After a delay during which files are downloaded, Stata responds with one of the following messages:

`installation complete` means that the package has been successfully installed and that you can now use the commands. Just above the "installation complete" message, Stata tells you the directory where the files were installed.

`all files already exist and are up-to-date` means that your system already has the latest version of the package. You can stop.

`the following files exist and are different` indicates that your system already has files with the same names as those in the package being installed and that the existing files differ from those in the package. The names of those files are listed and you are given several options. Assuming that the files listed are earlier versions of our programs, you should select the option "Force installation replacing already-installed files". This might sound ominous, but it is not. Because the files on our web site are the latest versions, you want to replace your current files with these new files. After you accept this option, Stata updates your files to newer versions.

`cannot write in directory` *directory-name* means that you do not have write privileges to the directory where Stata wants to install the files. Usually, this occurs only when you are using Stata on a network. Then we recommend that you contact your network administrator and ask if our commands can be installed using the instructions given above. If you cannot wait for a network administrator to install the commands or to give you the needed write access, you can install the programs to any directory where you have write permission, including a zip disk or your directory on a network. For example, suppose that you want to install SPost to your directory called `d:\username` (which can be any directory where you have write access). You should use the following commands:

```
. cd d:\username
d:\username
. mkdir ado
. sysdir set PERSONAL "d:\username\ado"
. net set ado PERSONAL
. net search spost
(contacting http://www.stata.com)
```

Then follow the installation instructions that we provided earlier for installing SPost. If you get the error "could not create directory" after typing `mkdir ado`, you probably do not have write privileges to the directory.

If you install ado-files to your own directory, each time you begin a new session you must tell Stata where these files are located. You do this by typing `sysdir set PERSONAL` *directory*, where *directory* is the location of the ado-files you have installed. For example,

```
. sysdir set PERSONAL d:\username\ado
```

#### Installing SPost using net install

You can also install the commands entirely from the Command window. (If you have already installed SPost, you do not need to read this section.) While you are online, enter

```
. net from http://www.indiana.edu/~jslsoc/stata/
```

The available packages will be listed. To install `spost9_ado`, type

```
. net install spost9_ado
```

`net get` can be used to download supplementary files (e.g., datasets, sample do-files) from our web site. For example, to download the package `spost9_do`, type

```
. net get spost9_do
```

These files are placed in the current working directory (see chapter 2 for a full discussion of the working directory).

### 1.5.3 What if commands do not work?

This section assumes that you have installed SPost but that some of the commands do not work. Here are some things to consider:

1. If you get the error message `unrecognized command`, there are several possibilities.
   a. If you discover that commands that used to work do not work anymore, you could be working on a different computer or on a different station in a computer lab. Because user-written ado-files work seamlessly in Stata, you might not realize that these programs must be installed on each machine that you use.
   b. If you sent a do-file that contains SPost commands to another person, and they cannot get the commands to work, let them know that they need to install SPost.
   c. If you get the error message `unrecognized command:` *strangename* after typing one of our commands, where *strangename* is not the name of the command that you typed, it means that Stata cannot find an ancillary ado-file that the command needs. We recommend that you install the SPost files again.
2. If you are getting an error message that you do not understand, click on the blue return code beneath the error message for more information about the error.
3. Make sure that Stata is properly installed and up to date. Typing `verinst` will verify that Stata has been properly installed. Typing `update query` will tell you if the version you are running is up to date and what you need to type to update it. If you are running Stata over a network, your network administrator may need to do this for you.

4. Often what appears to be a problem with one of our commands is actually a mistake you have made (we know because we make them, too). For example, make sure that you are not using `=` when you should be using `==`.
5. Because our commands work after you have fitted a model, make sure that there were no problems with the last model fitted. If Stata was not successful in fitting your model, our commands will not have the information needed to operate properly.
6. Irregular value labels can cause Stata programs to fail. We recommend using labels that have fewer than eight characters and contain no spaces or special characters other than underscores (_). If your variables (especially your dependent variable) do not meet this standard, try changing your value labels with the `label` command (details are given in section 2.15).
7. Unusual values of the outcome categories can also cause problems. For ordinal or nominal outcomes, some of our commands require that all the outcome values be integers between 0 and 99. For these types of outcomes, we recommend using consecutive integers starting with 1.

Before attempting to contact us about a problem, please check our web site *http://www.indiana.edu/~jslsoc/spost.htm* for new information about SPost. You should also read the information at *http://www.indiana.edu/~jslsoc/spost_help.htm* and check page xxxi in the Preface.

### 1.5.4 Uninstalling SPost

Stata keeps track of the packages that it has installed, which makes it easy for you to uninstall them in the future. If you want to uninstall our commands, simply type `ado uninstall spost9_ado`.

### 1.5.5 Using spex to load data and run examples

Experimenting with the postestimation commands that we discuss requires that you have fitted the appropriate model. In our examples, we show you how to open a dataset and fit models as you would when you were working with your own data. Accordingly, we begin with a `use` command to load the data and then use an estimation command, such as `logit`, to fit the model. To make it simpler to experiment with the methods in later chapters, we have written the command `spex` (Stata postestimation examples). If you type `spex logit`, for example, it will automatically load the data and fit the model that serves as our main logit example. Typing `spex` *commandname* will produce our primary example for that estimation command. Or, you can specify the name of any dataset that we use, `spex` *datasetname*, and `spex` will load those data but not fit any model. By default, `spex` looks for the dataset on our web site. If it does not find it there, it will look in the current working directory and all the directories where Stata searches for ado-files. Specifying the option `user` tells `spex` to look in your current working directory first. For more information, type `help spex`.

### 1.5.6 More files available on the web site

In addition to the SPost commands, we have provided other packages that you might find useful. For example, the package called `spost9_do` contains the do-files and datasets needed to reproduce the examples from this book. The package `spoststst8` contains the do-files and datasets to reproduce the results from Long (1997). To obtain these packages, type `net search spost` and follow the instructions you will be given. **Important**: if a package does not contain ado-files, Stata will download the files to the current working directory. Consequently, you need to change your working directory to wherever you want the files to go *before* you select "click here to get". More information about working directories and changing your working directory is provided in section 2.5.

## 1.6 Where can I learn more about the models?

There are many valuable sources for learning more about the regression models that are covered in this book. Not surprisingly, we recommend

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage. This book provides more details about the models discussed in our book.

We also recommend the following:

Cameron, A. C. and P. K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press. This graduate textbook provides an excellent introduction to the methods and models discussed in this book.

Cameron, A. C. and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. This is the definitive reference for count models.

Greene, W. H. 2003. *Econometric Analysis*. 5th ed. Upper Saddle River, NJ: Prentice Hall. Although this book focuses on models for continuous outcomes, several later chapters deal with models for categorical outcomes.

Hardin, J. and J. Hilbe. 2001. *Generalized Linear Models and Extensions*. College Station, TX: Stata Press. This is a thorough review of the generalized linear model or GLM approach to modeling and includes detailed information on the use of these models with Stata.

Hosmer, Jr., D. W., and S. Lemeshow. 2000. *Applied Logistic Regression*. 2nd ed. New York: Wiley. This book, written primarily for biostatisticians and medical researchers, provides much useful information about logit models for binary, ordinal, and nominal outcomes. Often the authors discuss how their recommendations can be executed using Stata.

Powers, D. A. and Y. Xie. 2000. *Statistical Methods for Categorical Data Analysis*. San Diego: Academic Press. This book considers all the models discussed in our book, with the exception of count models, and includes loglinear models and models for event history analysis.

Train, K. 2003. *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press. This is an outstanding review of models for a wide range of models for discrete choice and includes details on new methods of estimation using simulation.