Preface

Many data analysts use survey data and understand the general purpose of survey weights. However, they may not have studied the details of how weights are computed, nor do they understand the purpose of different steps used in weighting. *Survey Weights: A Step-by-step Guide to Calculation* is intended to fill these gaps in understanding. Throughout the book, we explain the theoretical rationale for why steps are done. Plus, we include many examples that give analysts tools for actually computing weights themselves in Stata.

We assume that the reader is familiar with Stata. If not, Kohler and Kreuter (2012) provide a good introduction.

Finally, we also assume that the reader has some applied sampling experience and knowledge of "lite" theory. Concepts of with-replacement versus without-replacement sampling and single- versus multistage designs should be familiar. Sources for sampling theory and associated applications abound, including Valliant, Dever, and Kreuter (2013), Lohr (2010), and Särndal, Swensson, and Wretman (1992), to name just a few.

Structure of the book

When faced with a new dataset, it is good practice to ask yourself a few questions before analyzing the data. For example,

- Am I dealing with a sample, or does the dataset contain a whole population?
- If it is a sample, how was it selected?
- What is my goal for the analysis? Am I trying to draw inference to the population?
- Do I need to weight my sample to project it to the population?
- Do I need to weight my data to compensate for the fact that the sample does not correctly cover the desired population?

Some datasets you encounter might already contain weights, and it is useful to understand how they were constructed. If you collect data yourself, you might need to construct weights on your own. In both cases, this book will give useful guidance, both for the construction and for the use of survey weights. This book can be read straight through but can also serve as a reference for specific procedures you may need to understand. You can skip around to particular topics and look at the examples for useful code.

We start our book with a general introduction to survey weighting in chapter 1. Weights are intended to project a sample to some larger population. The steps in weight calculation can be justified in different ways, depending on whether a probability or nonprobability sample is used. An overview of the typical steps is given in this chapter, including a flowchart of the steps.

Chapter 2 covers the initial weighting steps in probability samples. The first step is to compute base weights calculated as the inverse of selection probabilities. In some applications, because of inadequate information, it is unclear whether some sample units are actually eligible for the survey. Adjustments can be made to the known eligible units to account for those with an unknown status.

Most surveys suffer from some degree of nonresponse. Chapter 3 reviews methods of nonresponse adjustment. A typical approach is to put sample units into groups (cells) based on characteristics of the units or estimates of the probabilities that units respond to the survey. This chapter also covers another option for cell creation—using machine learning algorithms like CART, random forests, or boosting to classify units.

Chapter 4 covers calibration or adjusting weights so that sample estimates of totals for a set of variables equal their corresponding population totals. Calibration is an important step in correcting coverage problems and nonresponse and, in addition, can also reduce variances.

Chapter 5 discusses options for variance estimation, including exact formulas, linearization, and replication. Using multiple adjustments in weight calculation, as described in the previous chapters, does affect the variance of point estimates of descriptive quantities like means and totals. We illustrate how these multiple effects can be reflected using replication variances.

Not all sets of survey data are selected via probability samples. Even if the initial sample is probability, an investigator often loses control over which units actually provide data. This is especially true in the current climate, in which people, businesses, and institutions are progressively becoming more resistant to cooperating. Chapter 6 describes methods to weight nonprobability samples. The general thinking about estimating propensities of cooperation and using calibration models, covered in chapters 3 and 4, can be adapted to the nonprobability situation.

Chapter 7 covers a few special situations. Normalized weights are scaled so that they sum to the number of units in the sample—not to an estimate of the population size. Although we do not recommend them, normalized weights are used in some applications, particularly in public opinion surveys. Other topics in this chapter include datasets with multiple weights, two-phase sampling, and weights for composite estimation. Some survey datasets come with more than one weight for each case, especially when subsamples of units are selected for different purposes. Two-phase sampling is often used when more intensive efforts are made to convert nonrespondents for a subsample of cases. Composite weighting is used to combine different samples from different frames such as persons with landline telephones and persons with cell phones. This chapter also covers

whether to use survey weights when fitting models. We describe the issues that need to be considered and give some analyses that can be done when deciding whether to use weights in fitting linear and nonlinear models from survey data.

Chapter 8 covers the unexciting but essential procedures needed for quality control when computing survey weights. An orderly system needs to be laid out in advance to guide the sequence of weighting steps, to list quality checks that will be made at every step, and to document the entire process.

Data files and programs for this book

The data and program files used in the examples are available on the Internet. You can access these files from within Stata or by downloading a zip archive. For either method, we suggest that you create a new directory and download the materials there.

- If the machine you are using to run Stata is connected to the Internet, you can download the files from within Stata. To do this, type the following commands in the Stata Command window:

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

  Notice that the statements above are prefaced by "." as in the Stata Results window. We use this convention throughout the book.

- The files are also stored as a zip archive, which you can download by pointing your browser to http://www.stata-press.com/data/svywt/svywt.zip.

  To extract the file `svywt.zip`, create a new folder, for example, `svywt`, copy `svywt.zip` into this folder, and unzip the file `svywt.zip` using any program that can extract zip archives. Make sure to preserve the subdirectory structure contained in the zip file.

Throughout the book, we assume that your current working directory (folder) is the directory where you have stored our files. This is important if you want to reproduce our examples.

Ensure that you do not replace our files with a modified version of the same file; avoid using the command `save, replace` while working with our files.