CHAPTER OVERVIEW

17: Preparing Datasets and Other Pragmatic Matters

> *The garden of life never seems to confine itself to the plots philosophers have laid out for its convenience. Maybe a few more tractors would do the trick.*
>
> *–Roger Zelazny*[103]

This is a somewhat strange chapter, even by my standards. My goal in this chapter is to talk a bit more honestly about the realities of working with data than you'll see anywhere else in the book. The problem with real world data sets is that they are *messy*. Very often the data file that you start out with doesn't have the variables stored in the right format for the analysis you want to do. Sometimes might be a lot of missing values in your data set. Sometimes you only want to analyse a subset of the data. Et cetera. In other words, there's a lot of ***data manipulation*** that you need to do, just to get all your data set into the format that you need it. The purpose of this chapter is to provide a basic introduction to all these pragmatic topics. Although the chapter is motivated by the kinds of practical issues that arise when manipulating real data, I'll stick with the practice that I've adopted through most of the book and rely on very small, toy data sets that illustrate the underlying issue. Because this chapter is essentially a collection of "tricks" and doesn't tell a single coherent story, it may be useful to start with a list of topics:

- Section 7.1. Tabulating data.
- Section 7.2. Transforming or recoding a variable.
- Section 7.3. Some useful mathematical functions.
- Section 7.4. Extracting a subset of a vector.
- Section 7.5. Extracting a subset of a data frame.
- Section 7.6. Sorting, flipping or merging data sets.
- Section 7.7. Reshaping a data frame.
- Section 7.8. Manipulating text.
- Section 7.9. Opening data from different file types.
- Section 7.10. Coercing data from one type to another.
- Section 7.11. Other important data types.
- Section 7.12. Miscellaneous topics.

As you can see, the list of topics that the chapter covers is pretty broad, and there's a *lot* of content there. Even though this is one of the longest and hardest chapters in the book, I'm really only scratching the surface of several fairly different and important topics. My advice, as usual, is to read through the chapter once and try to follow as much of it as you can. Don't worry too much if you can't grasp it all at once, especially the later sections. The rest of the book is only lightly reliant on this chapter, so you can get away with just understanding the basics. However, what you'll probably find is that later on you'll need to flick back to this chapter in order to understand some of the concepts that I refer to here.