Chapter 18

Beyond

> "End? No, the journey doesn't end here."
>
> — **J.R.R. Tolkien**

After reading this chapter you will be able to:

- Understand the roadmap to continued education about models and the `R` programming language.

18.1 What's Next

So you've completed STAT 420, where do you go from here? Now that you understand the basics of linear modeling, there is a wide world of applied statistics waiting to be explored. We'll briefly detail some resources and discuss how they relate to what you have learned in STAT 420.

18.2 RStudio

RStudio has recently released version 1.0! This is exciting for a number of reason, especially the release of `R` Notebooks. `R` Notebooks combine the RMarkdown you have already learned with the ability to work interactively.

18.3 Tidy Data

In this textbook, much of the data we have seen has been nice and tidy. It was rectangular where each row is an observation and each column is a variable.

This is not always the case! Many packages have been developed to deal with data, and force it into a nice format, which is called tidy data, that we can then use for modeling. Often during analysis, this is where a large portion of your time will be spent.

The R community has started to call this collection of packages the Tidyverse. It was once called the Hadleyverse, as Hadley Wickham has authored so many of the packages. Hadley is writing a book called R for Data Science which describes the use of many of these packages. (And also how to use some to make the modeling process better!) This book is a great starting point for diving deeper into the R community. The two main packages are dplyr and tidyr both of which are used internally in RStudio.

18.4 Visualization

In this course, we have mostly used the base plotting methods in R. When working with tidy data, many users prefer to use the ggplot2 package, also developed by Hadley Wickham. RStudio provides a rather detailed "cheat sheet" for working with ggplot2. The community maintains a graph gallery of examples.

Use of the manipulate package with RStudio gives the ability to quickly change a static graphic to become interactive.

18.5 Web Applications

RStudio has made it incredible easy to create data products through the use of Shiny, which allows for the creation of web applications with R. RStudio maintains an ever-growing tutorial and gallery of examples.

18.6 Experimental Design

In the ANOVA chapter, we briefly discussed experimental design. This topic could easily be its own class, and is currently an area of revitalized interest with the rise of A/B testing. Two more classic statistical references include *Statistics for Experimenters* by Box, Hunter, and Hunter as well as *Design and Analysis of Experiments* by Douglas Montgomery. There are several R packages for design of experiments, list in the CRAN Task View.

18.7 Machine Learning

Using models for prediction is the key focus of machine learning. There are many methods, each with its own package, however `R` has a wonderful package called `caret`, *Classification And REgression Training,* which provides a unified interface to training these models. It also contains various utilities for data processing and visualization that are useful for predictive modeling.

Applied Predictive Modeling by Max Kuhn, the author of the `caret` package is a good general resource for predictive modeling, which obviously utilizes `R`. *An Introduction to Statistical Learning* by James, Witten, Hastie, and Tibshirani is a gentle introduction to machine learning from a statistical perspective which uses `R` and picks up right where this courses stops. This is based on the often referenced *The Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman. Both are freely available online.

18.7.1 Deep Learning

While, it probably isn't the best tool for the job, `R` now has the ability to train deep neural networks via TensorFlow.

18.8 Time Series

In this class we have only considered independent data. What if data is dependent? Time Series is the area of statistics which deals with this issue, and could easily span multiple courses.

The primary textbook for STAT 429: Time Series Analysis at the University of Illinois that is free is:

- *Time Series Analysis and Its Applications: With R Examples* by Shumway and Stoffer

Some tutorials:

- Little Book of R for Time Series
- Quick `R`: Time Series and Forecasting
- TSA: Start to Finish Examples

When performing time series analysis in `R` you should be aware of the many packages that are useful for analysis. It should be hard to avoid the `forecast` and `zoo` packages. Often the most difficult part will be dealing with time and date data. Make sure you are utilizing one of the many packages that help with this.

18.9 Bayesianism

In this class, we have worked within the frequentist view of statistics. There is an entire alternative universe of Bayesian statistics.

Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan by John Kruschke is a great introduction to the topic. It introduces the world of probabilistic programming, in particular Stan, which can be used in both `R` and Python.

18.10 High Performance Computing

Often `R` will be called a "slow" language, for two reasons. One, because many do not understand `R`. Two, because sometimes it really is. Luckily, it is easy to extend `R` via the `Rcpp` package to allow for faster code. Many modern `R` packages utilize `Rcpp` to achieve better performance.

18.11 Further `R` Resources

Also, don't forget that previously in this book we have outlined a large number of `R` resources. Now that you've gotten started with `R` many of these will be much more useful.

If any of these topics interest you, and you would like more information, please don't hesitate to start a discussion on the forums!

:)