David J. Olive

Theory for Linear Models

July 16, 2021

[redacted]

Preface

Many statistics departments offer a one semester graduate course in linear model theory. Linear models include multiple linear regression and many experimental design models. Three good books on linear model theory, in increasing order of difficulty, are Myers and Milton (1991), Seber and Lee (2003), and Christensen (2020). Other texts include Agresti (2015), Freedman (2005), Graybill (1976, 2000), Guttman (1982), Harville (2018), Hocking (2013), Monahan (2008), Muller and Stewart (2006), Rao (1973), Rao et al. (2008), Ravishanker and Dey (2002), Rencher and Schaalje (2008), Scheffé (1959), Searle and Gruber (2017), Sengupta and Jammalamadaka (2019), Stapleton (2009), Wang and Chow (1994), and Zimmerman (2020ab). A good summary is Olive (2017a, ch. 11).

The prerequisites for this text are i) a calculus based course in statistics at the level of Chihara and Hesterberg (2011), Hogg et al. (2015), Larsen and Marx (2017), Wackerly et al. (2008), and Walpole et al. (2016). ii) Linear algebra at the level of Anton et al. (2019), and Leon (2015). iii) A calculus based course in multiple linear regression at the level of Abraham and Ledolter (2006), Cook and Weisberg (1999), Kutner et al. (2005), Olive (2010, 2017a), and Weisberg (2014).

This text emphasizes large sample theory over normal theory, and shows how to do inference after variable selection. The text is at a Master's level for the United States. Let n be the sample size and p the number of predictor variables. Chapter 1 reviews some of the material from a calculus based course in multiple linear regression as well as some of the material to be covered in the text. Chapter 1 also covers the multivariate normal distribution and large sample theory. Most of these sections can be skimmed and then reviewed as needed. Chapters 2 and 3 cover full and nonfull rank linear models, respectively, with emphasis on least squares. Chapter 4 considers variable selection when $n >> p$. Chapter 5 considers Statistical Learning alternatives to least squares when $n >> p$, including lasso, lasso variable selection, and the elastic net. Chapter 6 shows how to use data splitting for inference if n/p is not large. Chapter 7 gives theory for robust regression, using results from robust multivariate location and dispersion. Chapter 8 gives theory for the multivariate linear model where there are $m \geq 2$ response variables. Chapter 9 examines the one way MANOVA model, which is a special case of the multivariate linear model. Chapter 10 generalizes much of the material from Chapters 2–6 to many other regression models, including generalized linear models and some survival regression models. Chapter 11 gives some information about *R* and some hints for homework problems.

Chapters 2–4 are the most important for a standard course in Linear Model Theory, along with the multivariate normal distribution and some large sample theory from Chapter 1. Some highlights of this text follow.

- Prediction intervals are given that can be useful even if $n < p$.
- The response plot is useful for checking the model.
- The large sample theory for the elastic net, lasso, and ridge regression is greatly simplified. Large sample theory for variable selection and lasso variable selection is given.
- The bootstrap is used for inference after variable selection if $n \geq 10p$.
- Data splitting is used for inference after variable selection or model building if $n < 5p$.
- Most of the above highlights are extended to many other regression models such as generalized linear models and some survival regression models.

The website (http://parker.ad.siu.edu/Olive/linmodbk.htm) for this book provides *R* programs in the file *linmodpack.txt* and several *R* data sets in the file *linmoddata.txt*. Section 11.1 discusses how to get the data sets and programs into the software, but the following commands will work.

Downloading the book's R functions *linmodpack.txt* and data files *linmoddata.txt* into *R*: The following commands

```
source("http://parker.ad.siu.edu/Olive/linmodpack.txt")
source("http://parker.ad.siu.edu/Olive/linmoddata.txt")
```

can be used to download the *R* functions and data sets into *R*. (*Copy and paste these two commands* into *R* from near the top of the file (http://parker.ad.siu.edu/Olive/linmodhw.txt), which contains commands that are useful for doing many of the *R* homework problems.) Type *ls()*. Over 100 *R* functions from *linmodpack.txt* should appear. Exit *R* with the command *q()* and click *No*.

The *R* software is used in this text. See R Core Team (2016). Some packages used in the text include `glmnet` Friedman et al. (2015), `leaps` Lumley (2009), `MASS` Venables and Ripley (2010), `mgcv` Wood (2017), and `pls` Mevik et al. (2015).

Acknowledgments

Teaching this course in 2014 as Math 583 and in 2019 and 2021 as Math 584 at Southern Illinois University was very useful.

Contents