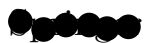


David J. Olive

Theory for Linear Models

July 16, 2021



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 $n \gg p$. Chapter 5 considers Statistical Learning alternatives to least squares when $n \gg p$, 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

1	Introduction	1
1.1	Overview	1
1.2	Response Plots and Response Transformations	4
1.2.1	Response and Residual Plots	5
1.2.2	Response Transformations	8
1.3	A Review of Multiple Linear Regression	13
1.3.1	The ANOVA F Test	16
1.3.2	The Partial F Test	21
1.3.3	The Wald t Test	24
1.3.4	The OLS Criterion	25
1.3.5	The Location Model	27
1.3.6	Simple Linear Regression	28
1.3.7	The No Intercept MLR Model	29
1.4	The Multivariate Normal Distribution	31
1.5	Large Sample Theory	34
1.5.1	The CLT and the Delta Method	34
1.5.2	Modes of Convergence and Consistency	37
1.5.3	Slutsky's Theorem and Related Results	45
1.5.4	Multivariate Limit Theorems	48
1.6	Mixture Distributions	52
1.7	Elliptically Contoured Distributions	53
1.8	Summary	57
1.9	Complements	59
1.10	Problems	60
2	Full Rank Linear Models	71
2.1	Projection Matrices and the Column Space	71
2.2	Quadratic Forms	76
2.3	Least Squares Theory	83
2.3.1	Hypothesis Testing	90
2.4	WLS and Generalized Least Squares	97

2.5	Summary	101
2.6	Complements	103
2.7	Problems	103
3	Nonfull Rank Linear Models and Cell Means Models	113
3.1	Nonfull Rank Linear Models	113
3.2	Cell Means Models	115
3.3	Summary	123
3.4	Complements	128
3.5	Problems	129
4	Prediction and Variable Selection When $n \gg p$	133
4.1	Variable Selection	133
4.1.1	OLS Variable Selection	134
4.2	Large Sample Theory for Some Variable Selection Estimators	143
4.3	Prediction Intervals	148
4.4	Prediction Regions	155
4.5	Bootstrapping Hypothesis Tests and Confidence Regions	161
4.5.1	The Bootstrap	164
4.5.2	Bootstrap Confidence Regions for Hypothesis Testing	167
4.5.3	Theory for Bootstrap Confidence Regions	170
4.5.4	Bootstrapping the Population Coefficient of Multiple Determination	175
4.6	Bootstrapping Variable Selection	178
4.6.1	The Parametric Bootstrap	180
4.6.2	The Residual Bootstrap	181
4.6.3	The Nonparametric Bootstrap	183
4.6.4	Bootstrapping OLS Variable Selection	184
4.6.5	Simulations	188
4.7	Data Splitting	192
4.8	Summary	192
4.9	Complements	195
4.10	Problems	199
5	Statistical Learning Alternatives to OLS	203
5.1	The MLR Model	203
5.2	Forward Selection	210
5.3	Principal Components Regression	213
5.4	Partial Least Squares	216
5.5	Ridge Regression	217
5.6	Lasso	225
5.7	Lasso Variable Selection	229

5.8	The Elastic Net	232
5.9	Prediction Intervals	235
5.10	Cross Validation	240
5.11	Hypothesis Testing After Model Selection, n/p Large ..	244
5.12	Data Splitting	245
5.13	Summary	246
5.14	Complements	251
5.15	Problems	256
6	What if n is not $\gg p$?	265
6.1	Sparse Models	267
6.2	Data Splitting	267
6.3	Summary	268
6.4	Complements	269
6.5	Problems	269
7	Robust Regression	271
7.1	The Location Model	271
7.2	The Multivariate Location and Dispersion Model	273
7.2.1	Affine Equivariance	274
7.2.2	Breakdown	275
7.2.3	The Concentration Algorithm	279
7.2.4	Theory for Practical Estimators	283
7.2.5	Outlier Resistance and Simulations	293
7.2.6	The RMVN and RFCH Sets	302
7.3	Outlier Detection for the MLD Model	304
7.3.1	MLD Outlier Detection if $p > n$	310
7.4	Outlier Detection for the MLR Model	313
7.5	Resistant Multiple Linear Regression	316
7.6	Robust Regression	327
7.6.1	MLR Breakdown and Equivariance	327
7.6.2	A Practical High Breakdown Consistent Estimator	335
7.7	Summary	341
7.8	Complements	342
7.9	Problems	345
8	Multivariate Linear Regression	353
8.1	Introduction	353
8.2	Plots for the Multivariate Linear Regression Model ..	357
8.3	Asymptotically Optimal Prediction Regions	360
8.4	Testing Hypotheses	365
8.5	An Example and Simulations	375
8.5.1	Simulations for Testing	380
8.6	The Robust <code>rmreg2</code> Estimator	383

8.7	Bootstrap	386
	8.7.1 Parametric Bootstrap	386
	8.7.2 Residual Bootstrap	386
	8.7.3 Nonparametric Bootstrap	387
8.8	Data Splitting	387
8.9	Summary	387
8.10	Complements	393
8.11	Problems	394
9	One Way MANOVA Type Models	399
	9.1 Introduction	399
	9.2 Plots for MANOVA Models	402
	9.3 One Way MANOVA	406
	9.4 An Alternative Test Based on Large Sample Theory .	410
	9.5 Summary	413
	9.6 Complements	416
	9.7 Problems	416
10	1D Regression Models Such as GLMs	417
	10.1 Introduction	417
	10.2 Additive Error Regression	422
	10.3 Binary, Binomial, and Logistic Regression	423
	10.4 Poisson Regression	431
	10.5 GLM Inference, n/p Large	437
	10.6 Variable and Model Selection	446
	10.6.1 When n/p is Large	446
	10.6.2 When n/p is Not Necessarily Large	454
	10.7 Generalized Additive Models	457
	10.7.1 Response Plots	459
	10.7.2 The EE Plot for Variable Selection	460
	10.7.3 An EE Plot for Checking the GLM	461
	10.7.4 Examples	461
	10.8 Overdispersion	466
	10.9 Inference After Variable Selection for GLMs	469
	10.9.1 The Parametric and Nonparametric Bootstrap .	469
	10.9.2 Bootstrapping Variable Selection	471
	10.9.3 Examples and Simulations	474
	10.10 Prediction Intervals	479
	10.11 OLS and 1D Regression	487
	10.11.1 Inference for 1D Regression With a Linear Predictor	489
	10.12 Data Splitting	492
	10.13 Complements	492
	10.14 Problems	495

Contents	xi
11 Stuff for Students	497
11.1 R	497
11.2 Hints for Selected Problems	501
11.3 Tables	511
Index	535