

David J. Olive

# Prediction and Statistical Learning

July 2, 2023





# Preface

Many statistics departments offer a one semester graduate course in statistical learning theory using texts such as Hastie et al. (2009), Hastie et al. (2015), and James et al. (2021). Also see Berk (2016), Izenman (2008), Kuhn and Johnson (2013), Marden (2006), and Murphy (2012).

The prerequisite for this text is a calculus based course in statistics at the level of Chihara and Hesterberg (2011), Hogg, Tanis, and Zimmerman (2020), Larsen and Marx (2011), Wackerly, Mendenhall and Scheaffer (2008) or Walpole, Myers, Myers and Ye (2016). Linear algebra and one computer programming class are essential. Knowledge of regression would be useful. See Olive (2017a) and Cook and Weisberg (1999). Knowledge of multivariate analysis would be useful. See Olive (2017b) and Johnson and Wichern (2007).

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.
- The large sample theory for some data splitting estimators, variable selection estimators, marginal maximum likelihood estimators, and one component partial least squares will be given. See Olive and Zhang (2023).

**Downloading the book's R functions *slpack.txt* and data files *sl-data.txt* into R:** The commands

```
source("http://parker.ad.siu.edu/Olive/slpack.txt")
source("http://parker.ad.siu.edu/Olive/sldata.txt")
```

The R software is used in this text. See R Core Team (2020). Some packages used in the text include `glmnet` Friedman et al. (2015), `leaps` Lumley (2009), `MASS` Venables and Ripley (2010), and `pls` Mevik et al. (2015).

## Acknowledgements

Teaching the material to Math 583 students at Southern Illinois University in 2017 was very useful. The text was used for a high dimensional statistics course in 2023. Trevor Hastie's website had a lot of useful information. Work by R. Dennis Cook and his coauthors was useful for figuring out OPLS.

# Contents

<b>1</b>	<b>Introduction</b> .....	1
1.1	Overview .....	1
1.2	Response Plots and Response Transformations .....	3
1.2.1	Response and Residual Plots .....	5
1.2.2	Response Transformations .....	8
1.3	The Multivariate Normal Distribution .....	13
1.4	Outlier Detection .....	16
1.4.1	The Location Model .....	17
1.4.2	Outlier Detection with Mahalanobis Distances .	18
1.4.3	Outlier Detection if $p > n$ .....	22
1.5	Large Sample Theory .....	28
1.5.1	The CLT and the Delta Method .....	29
1.5.2	Modes of Convergence and Consistency .....	32
1.5.3	Slutsky's Theorem and Related Results .....	39
1.5.4	Multivariate Limit Theorems .....	42
1.6	Mixture Distributions .....	46
1.7	A Review of Multiple Linear Regression .....	48
1.7.1	The ANOVA F Test .....	51
1.7.2	The Partial F Test .....	56
1.7.3	The Wald t Test .....	59
1.7.4	The OLS Criterion .....	60
1.7.5	The No Intercept MLR Model .....	62
1.8	Summary .....	63
1.9	Complements .....	67
1.10	Problems .....	68
<b>2</b>	<b>Prediction and Variable Selection When <math>n \gg p</math></b> .....	75
2.1	Variable Selection .....	75
2.1.1	OLS Variable Selection .....	76
2.2	Large Sample Theory for Some Variable Selection Estimators .....	85

2.2.1	Some Variable Selection Estimators .....	85
2.2.2	Large Sample Theory for Variable Selection Estimators .....	87
2.3	Prediction Intervals .....	91
2.4	Prediction Regions .....	99
2.4.1	Prediction Regions If $n/p$ Is Small.....	104
2.5	Bootstrapping Hypothesis Tests and Confidence Regions .....	104
2.5.1	The Bootstrap .....	106
2.5.2	Bootstrap Confidence Regions for Hypothesis Testing.....	110
2.5.3	Theory for Bootstrap Confidence Regions .....	114
2.5.4	Bootstrapping the Population Coefficient of Multiple Determination .....	120
2.6	OLS Large Sample Theory .....	123
2.7	Bootstrapping Variable Selection Estimators .....	124
2.7.1	The Parametric Bootstrap .....	127
2.7.2	The Residual Bootstrap.....	128
2.7.3	The Nonparametric Bootstrap .....	130
2.8	Examples and Simulations .....	131
2.8.1	Simulations .....	135
2.9	Data Splitting .....	138
2.10	Summary .....	139
2.11	Complements .....	142
2.12	Problems.....	146
<b>3</b>	<b>Statistical Learning Alternatives to OLS .....</b>	<b>149</b>
3.1	The MLR Model .....	149
3.2	Forward Selection .....	160
3.3	Principal Components Regression .....	163
3.4	Partial Least Squares .....	168
3.5	Ridge Regression.....	170
3.6	Lasso.....	178
3.7	Lasso Variable Selection.....	182
3.8	The Elastic Net .....	185
3.9	OPLS .....	189
3.10	The MMLE .....	191
3.11	$k$ -Component Regression Estimators .....	192
3.12	Prediction Intervals .....	194
3.13	Cross Validation .....	198
3.14	Hypothesis Testing after Model Selection, $n/p$ Large .	203
3.15	What if $n$ is not $\gg p$ ? .....	204
3.15.1	Sparse Models .....	205
3.16	Data Splitting .....	206
3.17	The Multitude of MLR Models .....	208

3.18	<b>Summary</b>	208
3.19	<b>Complements</b>	213
3.20	<b>Problems</b>	218
<b>4</b>	<b>1D Regression Models Such as GLMs</b>	227
4.1	<b>Introduction</b>	227
4.2	<b>Additive Error Regression</b>	232
4.3	<b>Binary, Binomial, and Logistic Regression</b>	233
4.4	<b>Poisson Regression</b>	241
4.5	<b>GLM Inference, <math>n/p</math> Large</b>	246
4.6	<b>Variable and Model Selection</b>	255
4.6.1	<b>When <math>n/p</math> is Large</b>	255
4.6.2	<b>When <math>n/p</math> is Not Necessarily Large</b>	264
4.7	<b>Generalized Additive Models</b>	267
4.7.1	<b>Response Plots</b>	269
4.7.2	<b>The EE Plot for Variable Selection</b>	269
4.7.3	<b>An EE Plot for Checking the GLM</b>	270
4.7.4	<b>Examples</b>	271
4.8	<b>Overdispersion</b>	275
4.9	<b>Inference After Variable Selection for GLMs</b>	278
4.9.1	<b>The Parametric and Nonparametric Bootstrap</b>	279
4.9.2	<b>Bootstrapping Variable Selection</b>	281
4.9.3	<b>Examples and Simulations</b>	283
4.10	<b>Prediction Intervals</b>	289
4.11	<b>Survival Analysis</b>	295
4.11.1	<b>Simulations</b>	298
4.12	<b>Regression Trees</b>	301
4.12.1	<b>Boosting</b>	303
4.13	<b>Data Splitting</b>	304
4.14	<b>Complements</b>	305
4.15	<b>Problems</b>	307
<b>5</b>	<b>Discriminant Analysis</b>	315
5.1	<b>Introduction</b>	315
5.2	<b>LDA and QDA</b>	317
5.2.1	<b>Regularized Estimators</b>	320
5.3	<b>LR</b>	320
5.4	<b>KNN</b>	322
5.5	<b>Some Matrix Optimization Results</b>	324
5.6	<b>FDA</b>	326
5.7	<b>Estimating the Test Error</b>	332
5.8	<b>Some Examples</b>	335
5.9	<b>Classification Trees, Bagging, and Random Forests</b>	338
5.9.1	<b>Pruning</b>	341
5.9.2	<b>Bagging</b>	342

5.9.3	Random Forests	343
5.10	Support Vector Machines	343
5.10.1	Two Groups	343
5.10.2	SVM With More Than Two Groups	346
5.11	Summary	346
5.12	Complements	350
5.13	Problems	351
<b>6</b>	<b>Regularizing a Correlation Matrix</b>	<b>359</b>
6.1	Correlation and Inverse Correlation Matrices	359
6.2	Regularizing a Correlation Matrix	362
6.3	Complements	365
6.4	Problems	366
<b>7</b>	<b>Clustering</b>	<b>367</b>
7.1	Hierarchical and $k$ -Means Clustering	367
7.2	Complements	372
7.3	Problems	372
<b>8</b>	<b>MLR with Heterogeneity</b>	<b>375</b>
8.1	OLS Large Sample Theory	375
8.2	Bootstrap Methods and Sandwich Estimators	376
8.3	Simulations	378
8.4	OPLS in Low and High Dimensions	380
8.5	Summary	380
8.6	Complements	380
8.7	Problems	380
<b>9</b>	<b>High Dimensional Statistics</b>	<b>381</b>
9.1	Introduction	381
9.2	Principle Components Analysis	381
9.3	MANOVA Type Tests	383
9.3.1	Large Sample Theory	384
9.3.2	One Sample Hotelling $T^2$ Type Tests	386
9.3.3	Two Sample Hotelling $T^2$ Type Tests	389
9.4	One Way MANOVA Type Tests	392
9.5	Multivariate Linear Regression	392
9.6	Summary	392
9.7	Complements	392
9.8	Problems	392
<b>10</b>	<b>Multivariate Linear Regression</b>	<b>393</b>
10.1	Introduction	393
10.2	Plots for the Multivariate Linear Regression Model	397
10.3	Asymptotically Optimal Prediction Regions	400
10.4	Testing Hypotheses	405



10.5	<b>An Example and Simulations</b> .....	415
10.5.1	<b>Simulations for Testing</b> .....	420
10.6	<b>The Robust <code>rmreg2</code> Estimator</b> .....	423
10.7	<b>Bootstrap</b> .....	426
10.7.1	<b>Parametric Bootstrap</b> .....	426
10.7.2	<b>Residual Bootstrap</b> .....	426
10.7.3	<b>Nonparametric Bootstrap</b> .....	427
10.8	<b>Data Splitting</b> .....	427
10.9	<b>Ridge Regression, PCR, and Other High Dimensional Methods</b> .....	427
10.10	<b>Summary</b> .....	427
10.11	<b>Complements</b> .....	434
10.12	<b>Problems</b> .....	435
11	<b>Stuff for Students</b> .....	441
11.1	<b>R</b> .....	441
11.2	<b>Hints for Selected Problems</b> .....	444
11.3	<b>Projects</b> .....	445
11.4	<b>Tables</b> .....	448
	<b>Index</b> .....	475