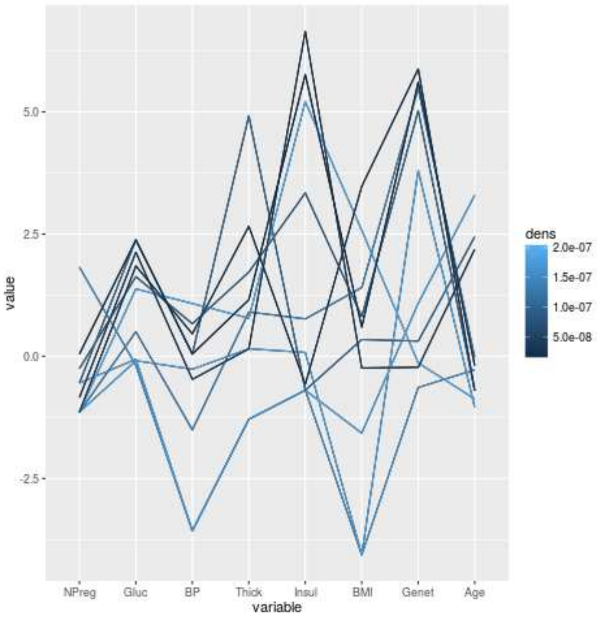
Statistical Regression and Classification

From Linear Models to Machine Learning

Norman Matloff

University of California, Davis



Outlier Hunt

Contents

Preface **xix**

1 Setting the Stage **1**

1.1 Example: Predicting Bike-Sharing Activity 1

1.2 Example of the Prediction Goal: Body Fat 2

1.3 Example: Who Clicks Web Ads? 3

1.4 Approach to Prediction 4

1.5 A Note about E(), Samples and Populations 5

1.6 Example of the Description Goal: Do Baseball Players Gain Weight As They Age? 6

1.6.1 Prediction vs. Description 7

1.6.2 A First Estimator 9

1.6.3 A Possibly Better Estimator, Using a Linear Model 10

1.7 Parametric vs. Nonparametric Models 15

1.8 Example: Click-Through Rate 15

1.9 Several Predictor Variables 17

1.9.1 Multipredictor Linear Models 18

1.9.1.1 Estimation of Coefficients 18

1.9.1.2 The Description Goal 19

1.9.2 Nonparametric Regression Estimation: k-NN 19

1.9.2.1 Looking at Nearby Points 20

1.9.2.2 Measures of Nearness 20

1.9.2.3 The k-NN Method, and Tuning Parameters 21

1.9.2.4 Nearest-Neighbor Analysis in the regtools Package 21

1.9.2.5 Example: Baseball Player Data 22

1.10 After Fitting a Model, How Do We Use It for Prediction? . 22

1.10.1 Parametric Settings 22

1.10.2 Nonparametric Settings 23

1.10.3 The Generic predict() Function 23

1.11 Overfitting, and the Variance-Bias Tradeoff 24

1.11.1 Intuition 24

1.11.2 Example: Student Evaluations of Instructors 26

1.12 Cross-Validation 26

1.12.1 Linear Model Case 28

1.12.1.1 The Code 28

1.12.1.2 Applying the Code 29

1.12.2 k-NN Case 29

1.12.3 Choosing the Partition Sizes 31

1.13 Important Note on Tuning Parameters 31

1.14 Rough Rule of Thumb 32

1.15 Example: Bike-Sharing Data 32

1.15.1 Linear Modeling of $\mu(t)$ 33

1.15.2 Nonparametric Analysis 38

1.16 Interaction Terms, Including Quadratics 38

1.16.1 Example: Salaries of Female Programmers and Engineers 39

1.16.2 Fitting Separate Models 42

1.16.3 Saving Your Work 43

1.16.4 Higher-Order Polynomial Models 43

1.17 Classification Techniques 44

1.17.1 It's a Regression Problem! 44

1.17.2 Example: Bike-Sharing Data 45

1.18 Crucial Advice: Don't Automate, Participate! 47

1.19 Mathematical Complements 48

1.19.1 Indicator Random Variables 48

1.19.2 Mean Squared Error of an Estimator 48

1.19.3 μ(t) Minimizes Mean Squared Prediction Error . . . 49

1.19.4 $\mu(t)$ Minimizes the Misclassification Rate 50

1.19.5 Kernel-Based Nonparametric Estimation of Regression Functions 52

1.19.6 General Nonparametric Regression 53

1.19.7 Some Properties of Conditional Expectation 54

1.19.7.1 Conditional Expectation As a Random Variable 54

1.19.7.2 The Law of Total Expectation 55

1.19.7.3 Law of Total Variance 55

1.19.7.4 Tower Property 56

1.19.7.5 Geometric View 56

1.20 Computational Complements 57

1.20.1 CRAN Packages 57

1.20.2 The Function **tapply()** and Its Cousins 58

1.20.3 The Innards of the k-NN Code 60

1.20.4 Function Dispatch 61

1.21 Centering and Scaling 62

1.22 Exercises: Data, Code and Math Problems 63

2 Linear Regression Models 67

2.1 Notation 67

2.2 The “Error Term” 69

2.3 Random- vs. Fixed-X Cases 69

2.4 Least-Squares Estimation 70

2.4.1 Motivation 70

2.4.2 Matrix Formulations 72

2.4.3 (2.18) in Matrix Terms 73

2.4.4 Using Matrix Operations to Minimize (2.18) 73

2.4.5 Models without an Intercept Term 74

2.5 A Closer Look at lm() Output 75

2.5.1 Statistical Inference 76

2.6 Assumptions 77

2.6.1 Classical 77

2.6.2 Motivation: the Multivariate Normal Distribution Family 78

2.7 Unbiasedness and Consistency 81

2.7.1 $\widehat{\beta}$ Is Unbiased 81

2.7.2 Bias As an Issue/Nonissue 82

2.7.3 $\widehat{\beta}$ Is Statistically Consistent 82

2.8 Inference under Homoscedasticity 83

2.8.1 Review: Classical Inference on a Single Mean 83

2.8.2 Back to Reality 84

2.8.3 The Concept of a Standard Error 85

2.8.4 Extension to the Regression Case 85

2.8.5 Example: Bike-Sharing Data 88

2.9 Collective Predictive Strength of the $X^{(j)}$ 90

2.9.1 Basic Properties 90

2.9.2 Definition of R^2 92

2.9.3 Bias Issues 93

2.9.4 Adjusted-R^2 94

2.9.5 The “Leaving-One-Out Method” 96

2.9.6 Extensions of LOOM 97

2.9.7 LOOM for k-NN 97

2.9.8 Other Measures 98

2.10 The Practical Value of p-Values — Small OR Large 98

2.10.1 Misleadingly Small p-Values 99

2.10.1.1 Example: Forest Cover Data 99

2.10.1.2 Example: Click Through Data 100

2.10.2 Misleadingly LARGE p-Values 101

2.10.3 The Verdict 102

2.11 Missing Values 102

2.12 Mathematical Complements 103

2.12.1 Covariance Matrices 103

2.12.2 The Multivariate Normal Distribution Family 105

2.12.3 The Central Limit Theorem 106

2.12.4 Details on Models Without a Constant Term 106

2.12.5 Unbiasedness of the Least-Squares Estimator 107

2.12.6 Consistency of the Least-Squares Estimator 108

2.12.7 Biased Nature of S 110

2.12.8 The Geometry of Conditional Expectation 110

2.12.8.1 Random Variables As Inner Product Spaces 110

2.12.8.2 Projections 111

2.12.8.3 Conditional Expectations As Projections . 112

2.12.9 Predicted Values and Error Terms Are Uncorrelated 113

2.12.10 Classical "Exact" Inference 114

2.12.11 Asymptotic $(p+1)$-Variate Normality of $\widehat{\beta}$ 115

2.13 Computational Complements 117

2.13.1 Details of the Computation of (2.28) 117

2.13.2 R Functions for the Multivariate Normal Distribution Family 118

2.13.2.1 Example: Simulation Computation of a Bivariate Normal Quantity 118

2.13.3 More Details of 'lm' Objects 120

2.14 Exercises: Data, Code and Math Problems 122

3 Homoscedasticity and Other Assumptions in Practice 125

3.1 Normality Assumption 126

3.2 Independence Assumption — Don't Overlook It 127

3.2.1 Estimation of a Single Mean 127

3.2.2 Inference on Linear Regression Coefficients 128

3.2.3 What Can Be Done? 128

3.2.4 Example: MovieLens Data 129

3.3 Dropping the Homoscedasticity Assumption 132

3.3.1 Robustness of the Homoscedasticity Assumption . . 133

3.3.2 Weighted Least Squares 135

3.3.3 A Procedure for Valid Inference 137

3.3.4 The Methodology 137

3.3.5 Example: Female Wages 138

3.3.6 Simulation Test 139

3.3.7 Variance-Stabilizing Transformations 139

3.3.8 The Verdict 141

3.4 Further Reading 141

3.5 Computational Complements 142

3.5.1 The R merge() Function 142

3.6 Mathematical Complements 143

3.6.1 The Delta Method 143

3.6.2 Distortion Due to Transformation 144

3.7 Exercises: Data, Code and Math Problems 145

4 Generalized Linear and Nonlinear Models 149

4.1 Example: Enzyme Kinetics Model 150

4.2 The Generalized Linear Model (GLM) 152

4.2.1 Definition 152

4.2.2 Poisson Regression 153

4.2.3 Exponential Families 154

4.2.4 GLM Computation 155

4.2.5 R's **glm()** Function 156

4.3 GLM: the Logistic Model 157

4.3.1 Motivation 157

4.3.2 Example: Pima Diabetes Data 161

4.3.3 Interpretation of Coefficients 162

4.3.4 The predict() Function Again 164

4.3.5 Overall Prediction Accuracy 164

4.3.6 Example: Predicting Spam E-mail 165

4.3.7 Linear Boundary 167

4.4 GLM: the Poisson Regression Model 167

4.5 Least-Squares Computation 169
4.5.1 The Gauss-Newton Method 169
4.5.2 Eicker-White Asymptotic Standard Errors 171
4.5.3 Example: Bike Sharing Data 174
4.5.4 The "Elephant in the Room": Convergence Issues 175
4.6 Further Reading 175
4.7 Computational Complements 176
4.7.1 R Factors 176
4.8 Mathematical Complements 177
4.8.1 Maximum Likelihood Estimation 177
4.9 Exercises: Data, Code and Math Problems 177
5 Multiclass Classification Problems 181
5.1 Key Notation 181
5.2 Key Equations 182
5.3 Estimating the Functions $\mu_i(t)$ 184
5.4 How Do We Use Models for Prediction? 184
5.5 One vs. All or All vs. All? 185
5.5.1 Which Is Better? 186
5.5.2 Example: Vertebrae Data 186
5.5.3 Intuition 187
5.5.4 Example: Letter Recognition Data 188
5.5.5 Example: k-NN on the Letter Recognition Data . . 189
5.5.6 The Verdict 190
5.6 Fisher Linear Discriminant Analysis 190
5.6.1 Background 191
5.6.2 Derivation 191

5.6.3 Example: Vertebrae Data 192
5.6.3.1 LDA Code and Results 192
5.7 Multinomial Logistic Model 193
5.7.1 Model 193
5.7.2 Software 194
5.7.3 Example: Vertebrae Data 194
5.8 The Issue of "Unbalanced" (and Balanced) Data 195
5.8.1 Why the Concern Regarding Balance? 196
5.8.2 A Crucial Sampling Issue 197
5.8.2.1 It All Depends on How We Sample 197
5.8.2.2 Remedies 199
5.8.3 Example: Letter Recognition 201
5.9 Going Beyond Using the 0.5 Threshold 202
5.9.1 Unequal Misclassification Costs 203
5.9.2 Revisiting the Problem of Unbalanced Data 204
5.9.3 The Confusion Matrix and the ROC Curve 204
5.9.3.1 Code 205
5.9.3.2 Example: Spam Data 205
5.10 Mathematical Complements 206
5.10.1 Classification via Density Estimation 206
5.10.1.1 Methods for Density Estimation 207
5.10.2 Time Complexity Comparison, OVA vs. AVA 208
5.10.3 Optimal Classification Rule for Unequal Error Costs 208
5.11 Computational Complements 209
5.11.1 R Code for OVA and AVA Logit Analysis 209
5.11.2 ROC Code 213

5.12 Exercises: Data, Code and Math Problems 214

6 Model Fit Assessment and Improvement 217

6.1 Aims of This Chapter 217

6.2 Methods 218

6.3 Notation 218

6.4 Goals of Model Fit-Checking 219

6.4.1 Prediction Context 219

6.4.2 Description Context 220

6.4.3 Center vs. Fringes of the Data Set 220

6.5 Example: Currency Data 221

6.6 Overall Measures of Model Fit 222

6.6.1 R-Squared, Revisited 223

6.6.2 Cross-Validation, Revisited 224

6.6.3 Plotting Parametric Fit Against a Nonparametric One 224

6.6.4 Residuals vs. Smoothing 225

6.7 Diagnostics Related to Individual Predictors 226

6.7.1 Partial Residual Plots 227

6.7.2 Plotting Nonparametric Fit Against Each Predictor 229

6.7.3 The freqparcoord Package 231

6.7.3.1 Parallel Coordinates 231

6.7.3.2 The freqparcoord Package 232

6.7.3.3 The regdiag() Function 233

6.8 Effects of Unusual Observations on Model Fit 234

6.8.1 The influence() Function 234

6.8.1.1 Example: Currency Data 235

6.8.2 Use of freqparcoord for Outlier Detection 237

6.9 Automated Outlier Resistance 238

6.9.1 Median Regression 238

6.9.2 Example: Currency Data 240

6.10 Example: Vocabulary Acquisition 241

6.11 Classification Settings 244

6.11.1 Example: Pima Diabetes Study 244

6.12 Improving Fit 247

6.12.1 Deleting Terms from the Model 249

6.12.2 Adding Polynomial Terms 249

6.12.2.1 Example: Currency Data 249

6.12.2.2 Example: Programmer/Engineer Census Data250

6.12.3 Boosting 253

6.12.3.1 View from the 30,000-Foot Level 254

6.12.3.2 Performance 256

6.13 A Tool to Aid Model Selection 256

6.14 Special Note on the Description Goal 258

6.15 Computational Complements 258

6.15.1 Data Wrangling for the Currency Dataset 258

6.15.2 Data Wrangling for the Word Bank Dataset 259

6.16 Mathematical Complements 260

6.16.1 The Hat Matrix 260

6.16.2 Matrix Inverse Update 261

6.16.3 The Median Minimizes Mean Absolute Deviation 262

6.17 Exercises: Data, Code and Math Problems 263

7 Disaggregating Regressor Effects 267

7.1 A Small Analytical Example 268

7.2 Example: Baseball Player Data 270

7.3 Simpson's Paradox 274

7.3.1 Example: UCB Admissions Data (Logit) 274

7.3.2 The Verdict 278

7.4 Unobserved Predictor Variabless 278

7.4.1 Instrumental Variables (IVs) 279

7.4.1.1 The IV Method 280

7.4.1.2 2 Stage Least Squares: 282

7.4.1.3 Example: Years of Schooling 284

7.4.1.4 Multiple Predictors 285

7.4.1.5 The Verdict 286

7.4.2 Random Effects Models 286

7.4.2.1 Example: Movie Ratings Data, Random Effects 287

7.4.3 Multiple Random Effects 288

7.4.4 Why Use Random/Mixed Effects Models? 289

7.5 Regression Function Averaging 289

7.5.1 Estimating the Counterfactual 290

7.5.1.1 Example: Job Training 290

7.5.2 Small Area Estimation: "Borrowing from Neighbors" 291

7.5.3 The Verdict 295

7.6 Multiple Inference 295

7.6.1 The Frequent Occurence of Extreme Events 295

7.6.2 Relation to Statistical Inference 296

7.6.3 The Bonferroni Inequality 297

7.6.4 Scheffe's Method 298
7.6.5 Example: MovieLens Data 299
7.6.6 The Verdict 302
7.7 Computational Complements 303
7.7.1 Movie Lens Data Wrangling 303
7.7.2 More Data Wrangling in the MovieLens Example . . 303
7.8 Mathematical Complements 305
7.8.1 Iterated Projections 305
7.8.2 Standard Errors for RFA 307
7.8.3 Asymptotic Chi-Square Distributions 308
7.9 Exercises: Data, Code and Math Problems 309

8 Shrinkage Estimators 311
8.1 Relevance of James-Stein to Regression Estimation 312
8.2 Multicollinearity 313
8.2.1 What's All the Fuss About? 313
8.2.2 A Simple Guiding Model 313
8.2.2.1 "Wrong" Signs in Estimated Coefficients . 314
8.2.3 Checking for Multicollinearity 315
8.2.3.1 The Variance Inflation Factor 315
8.2.3.2 Example: Currency Data 315
8.2.4 What Can/Should One Do? 316
8.2.4.1 Do Nothing 316
8.2.4.2 Eliminate Some Predictors 316
8.2.4.3 Employ a Shrinkage Method 316
8.3 Ridge Regression 316
8.3.1 Alternate Definitions 317

8.3.2 Yes, It Is Smaller 318
8.3.3 Choosing the Value of λ 319
8.3.4 Example: Currency Data 320
8.4 The LASSO 322
8.4.1 Definition 322
8.4.2 The lars Package 323
8.4.3 Example: Currency Data 324
8.4.4 The Elastic Net 325
8.5 Cases of Exact Multicollinearity, Including $p > n$ 326
8.5.1 Why It May Work 326
8.5.2 Example: R mtcars Data 326
8.5.2.1 Additional Motivation for the Elastic Net . 328
8.6 Bias, Standard Errors and Signficance Tests 328
8.7 Generalized Linear Models 328
8.7.1 Example: Vertebrae Data 329
8.8 Other Terminology 330
8.9 Further Reading 330
8.10 Mathematical Complements 330
8.10.1 James-Stein Theory 330
8.10.1.1 Definition 331
8.10.1.2 Theoretical Properties 331
8.10.1.3 When Might Shrunken Estimators Be Helpful? 331
8.10.2 Ridge Action Increases Eigenvalues 332
8.11 Computational Complements 332
8.11.1 Code for ridgelm() 332
8.12 Exercises: Data, Code and Math Problems 334

9 Variable Selection and Dimension Reduction 337

9.1 A Closer Look at Under/Overfitting 339

9.1.1 A Simple Guiding Example 340

9.2 How Many Is Too Many? 342

9.3 Fit Criteria 342

9.3.1 Some Common Measures 343

9.3.2 No Panacea! 346

9.4 Variable Selection Methods 347

9.5 Simple Use of p-Values: Pitfalls 347

9.6 Asking "What If" Questions 347

9.7 Stepwise Selection 349

9.7.1 Basic Notion 349

9.7.2 Forward vs. Backward Selection 350

9.7.3 R Functions for Stepwise Regression 350

9.7.4 Example: Bodyfat Data 350

9.7.5 Classification Settings 355

9.7.5.1 Example: Bank Marketing Data 355

9.7.5.2 Example: Vertebrae Data 359

9.7.6 Nonparametric Settings 360

9.7.6.1 Is Dimension Reduction Important in the Nonparametric Setting? 360

9.7.7 The LASSO 362

9.7.7.1 Why the LASSO Often Performs Subsetting 362

9.7.7.2 Example: Bodyfat Data 364

9.8 Post-Selection Inference 365

9.9 Direct Methods for Dimension Reduction 367

9.9.1 Informal Nature 367

9.9.2 Role in Regression Analysis 368
9.9.3 PCA 368
9.9.3.1 Issues 369
9.9.3.2 Example: Bodyfat Data 369
9.9.3.3 Example: Instructor Evaluations 373
9.9.4 Nonnegative Matrix Factorization (NMF) 375
9.9.4.1 Overview 375
9.9.4.2 Interpretation 375
9.9.4.3 Sum-of-Parts Property 376
9.9.4.4 Example: Spam Detection 376
9.9.5 Use of freqparcoord for Dimension Reduction 378
9.9.5.1 Example: Student Evaluations of Instructors 378
9.9.5.2 Dimension Reduction for Dummy/R Factor Variables 378
9.10 The Verdict 380
9.11 Further Reading 380
9.12 Computational Complements 381
9.12.1 Computation for NMF 381
9.13 Mathematical Complements 383
9.13.1 MSEs for the Simple Example 383
9.14 Exercises: Data, Code and Math Problems 384

10 Partition-Based Methods 389

10.1 CART 390
10.2 Example: Vertebral Column Data 392
10.3 Technical Details 396
10.4 Statistical Consistency 396
10.5 Tuning Parameters 397

10.6 Random Forests 397
10.6.1 Bagging 397
10.6.2 Example: Vertebrae Data 398
10.6.3 Example: Letter Recognition 399
10.7 Other Implementations of CART 400
10.8 Exercises: Data, Code and Math Problems 401

11 Semi-Linear Methods 403

11.1 k-NN with Linear Smoothing 405
11.1.1 Extrapolation Via lm() 405
11.1.2 Multicollinearity Issues 408
11.1.3 Example: Bodyfat Data 408
11.1.4 Tuning Parameter 408
11.2 Linear Approximation of Class Boundaries 409
11.2.1 SVMs 409
11.2.1.1 Geometric Motivation 410
11.2.1.2 Reduced Convex Hulls 412
11.2.1.3 Tuning Parameter 416
11.2.1.4 Nonlinear Boundaries 417
11.2.1.5 Statistical Consistency 418
11.2.1.6 Example: Letter Recognition Data 418
11.2.2 Neural Networks 419
11.2.2.1 Example: Vertebrae Data 419
11.2.2.2 Tuning Parameters and Other Technical Details 422
11.2.2.3 Dimension Reduction 422
11.2.2.4 Statistical Consistency 422
11.3 The Verdict 423

11.4 Mathematical Complements 424
11.4.1 Edge Bias with k-NN and Kernel Methods 424
11.4.2 Dual Formulation for SVM 425
11.4.3 The Kernel Trick 427
11.5 Further Reading 428
11.6 Exercises: Data, Code and Math Problems 429
12 Regression and Classification in Big Data 431
12.1 Solving the Big-n Problem 432
12.1.1 Software Alchemy 432
12.1.2 Example: Flight Delay Data 433
12.1.3 More on the Insufficient Memory Issue 436
12.1.4 Deceivingly “Big” n 437
12.1.5 The Independence Assumption in Big-n Data 437
12.2 Addressing Big-p 438
12.2.1 How Many Is Too Many? 438
12.2.1.1 Toy Model 439
12.2.1.2 Results from the Research Literature . . . 440
12.2.1.3 A Much Simpler and More Direct Approach 441
12.2.1.4 Nonparametric Case 441
12.2.1.5 The Curse of Dimensionality 443
12.2.2 Example: Currency Data 443
12.2.3 Example: Quiz Documents 444
12.2.4 The Verdict 446
12.3 Mathematical Complements 447
12.3.1 Speedup from Software Alchemy 447
12.4 Computational Complements 448

12.4.1 The partools Package 448

12.4.2 Use of the tm Package 449

12.5 Exercises: Data, Code and Math Problems 450

A Matrix Algebra 451

A.1 Terminology and Notation 451

A.1.1 Matrix Addition and Multiplication 452

A.2 Matrix Transpose 453

A.3 Linear Independence 454

A.4 Matrix Inverse 454

A.5 Eigenvalues and Eigenvectors 455

A.6 Rank of a Matrix 456

A.7 Matrices of the Form B'B 456

A.8 Partitioned Matrices 457

A.9 Matrix Derivatives 458

A.10 Matrix Algebra in R 459

A.11 Further Reading 462

Preface

Why write yet another regression book? There is a plethora of books out there already, written by authors whom I greatly admire, and whose work I myself have found useful. I might cite the books by Harrell [60] and Fox [49], among many, many excellent examples. Note that I am indeed referring to general books on regression analysis, as opposed to more specialized work such as [65] and [75], which belong to a different genre. My book here is intended for a traditional (though modernized) regression course, rather than one on statistical learning.

Yet, I felt there is an urgent need for a different kind of book. So, why is this regression book different from all other regression books? First, **it modernizes the standard treatment of regression methods**. In particular:

- The book supplements classical regression models with introductory material on machine learning methods.
- Recognizing that these days, classification is the focus of many applications, the book covers this topic in detail, especially the multiclass case.
- In view of the voluminous nature of many modern datasets, there is a chapter on Big Data.
- There is much more hands-on involvement of computer usage.

Other major senses in which this book differs from others are:

- Though presenting the material in a mathematically precise manner, the book aims to provide much needed practical insight for the practicing analyst, remedying the "too many equations, too few explanations" problem.

For instance, the book not only shows how the math works for transformations of variables, but also raises points on why one might refrain from applying transformations.

- The book features a recurring interplay between parametric and nonparametric methods. For instance, in an example involving currency data, the book finds that the fitted linear model predicts substantially more poorly than a k-nearest neighbor fit, suggesting deficiencies in the linear model. Nonparametric analysis is then used to further investigate, providing parametric model assessment in a manner that is arguably more insightful than classical residual plots.

- For those interested in computing issues, many of the book's chapters include optional sections titled Computational Complements, on topics such as data wrangling, views of package source code, parallel computing and so on.

  Also, many of the exercises are code-oriented. In particular, in such exercises the reader is asked to write "mini-CRAN" functions,[1] short but useful library functions that can be applied to practical regression analysis. Here is an example exercise of this kind:

  > Write a function stepAR2() that works similarly to stepAIC(), except that this new function uses adjusted R^2 as its criterion for adding or deleting a predictor. The call form will be
  >
  > ```
  > stepAR2(lmobj,direction='fwd',
  >    nsteps=ncol(lmobj$model)-1)
  > ```
  >
  > where the arguments are...Predictors will be added/deleted one at a time, according to which one maximizes adjusted R^2. The return value will be an S3 object of type 'stepr2', with sole component a data frame of $\widehat{\beta}_i$ values (0s meaning the predictor is not currently in the prediction equation), one row per model. There will also be an R^2 column. Write a summary() function for this class that shows the actions taken at each step of the process.

- For those who wish to go into more depth on mathematical topics, there are Mathematical Complements sections at the end of most chapters, and math-oriented exercises. The material ranges from straightforward computation of mean squared error to esoteric topics such as a proof of the Tower Property, $E\left[E(V|U_1, U_2) \mid U_1\right] = E(V \mid U_1)$, a result that is used in the text.

[1] CRAN is the online repository of user-contributed R code.

As mentioned, **this is still a book on traditional regression analysis.** In contrast to [65], this book is aimed at a traditional regression course. Except for Chapters 10 and 11, the primary methodology used is linear and generalized linear parametric models, covering both the Description and Prediction goals of regression methods. We are just as interested in Description applications of regression, such as measuring the gender wage gap in Silicon Valley, as we are in forecasting tomorrow's demand for bike rentals. An entire chapter is devoted to measuring such effects, including discussion of Simpson's Paradox, multiple inference, and causation issues. The book's examples are split approximately equally in terms of Description and Prediction goals. Issues of model fit play a major role.

The book includes more than 75 full examples, using real data. But concerning the above comment regarding "too many equations, too few explanations,", merely including examples with real data is not enough to truly tell the story in a way that will be useful in practice. Rather few books go much beyond presenting the formulas and techniques, thus leaving the hapless practitioner to his own devices. Too little is said in terms of what the equations really mean in a practical sense, what can be done with regard to the inevitable imperfections of our models, which techniques are too much the subject of "hype," and so on.

As a nonstatistician, baseball great Yogi Berra, put it in his inimitable style, "In theory there is no difference between theory and practice. In practice there is." This book aims to remedy this gaping deficit. It develops the material in a manner that is *mathematically precise* yet always maintains as its top priority — borrowing from a book title of the late Leo Breiman — *a view toward applications.*

In other words:

> The philosophy of this book is to not only prepare the analyst to know *how* to do something, but also to understand *what* she is doing. For successful application of data science techniques, the latter is just as important as the former.

Some further examples of how this book differs from the other regression books:

Intended audience and chapter coverage:

This book is aimed at both practicing professionals and use in the classroom. It aims to be both accessible and valuable to this diversity of readership.

In terms of classroom use, with proper choice of chapters and appendices, the book could be used as a text tailored to various discipline-specific audiences and various levels, undergraduate or graduate. I would recommend that the core of any course consist of most sections of Chapters 1-4 (excluding the Math and Computational Complements sections), with coverage of at least introductory sections of Chapters 5, 6, 7, 8 and 9 for all audiences. Beyond that, different types of disciplines might warrant different choices of further material. For example:

- **Statistics students:** Depending on level, at least some of the Mathematical Complements and math-oriented exercises should be involved. There might be more emphasis on Chapters 6, 7 and 9.
- **Computer science students:** Here one would cover more of classification, machine learning and Big Data material, Chapters 5, 8, 10, 11 and 12. Also, one should cover the Computational Complements sections and associated "mini-CRAN" code exercises.
- **Economics/social science students:** Here there would be heavy emphasis on the Description side, Chapters 6 and 7, with special emphasis on topics such as Instrumental Variables and Propensity Matching in Chapter 7. Material on generalized linear models and logistic regression, in Chapter 4 and parts of Chapter 5, might also be given emphasis.
- **Student class level:** The core of the book could easily be used in an undergraduate regression course, but aimed at students with background in calculus and matrix algebra, such as majors in statistics, math or computer science. A graduate course would cover more of the chapters on advanced topics, and would likely cover more of the Mathematical Complements sections.
- **Level of mathematical sophistication:** In the main body of the text, i.e., excluding the Mathematical Complements sections, basic matrix algebra is used throughout, but use of calculus is minimal. As noted, for those instructors who want the mathematical content, it is there in the Mathematical Complements sections, but the main body of the text requires only the matrix algebra and a little calculus.

The reader must of course be familiar with terms like *confidence interval*, *significance test* and *normal distribution*. Many readers will have had at least some prior exposure to regression analysis, but this is not assumed, and the subject is developed from the beginning.

The reader is assumed to have some prior experience with R, but at a minimal level: familiarity with function arguments, loops, if-else and vector/matrix operations and so on. For those without such background, there are many gentle tutorials on the Web, as well as a leisurely introduction in a statistical context in [21]. Those with programming experience can also read the quick introduction in the appendix of [102]. My book [95] gives a detailed treatment of R as a programming language, but that level of sophistication is certainly not needed for the present book.

A comment on the field of machine learning:

Mention should be made of the fact that this book's title includes both the word *regression* and the phrase *machine learning.* The latter phrase is included to reflect that the book includes some introductory material on machine learning, in a regression context.

Much has been written on a perceived gap between the statistics and machine learning communities [24]. This gap is indeed real, but work has been done to reconcile them [16], and in any event, the gap is actually not as wide as people think.

My own view is that machine learning (ML) consists of the development of regression models with the Prediction goal. Typically nonparametric (or what I call semi-parameteric) methods are used. Classification models are more common than those for predicting continuous variables, and it is common that more than two classes are involved, sometimes a great many classes. All in all, though, it's still regression analysis, involving the conditional mean of Y given X (reducing to $P(Y = 1|X)$ in the classification context).

One often-claimed distinction between statistics and ML is that the former is based on the notion of a sample from a population whereas the latter is concerned only with the content of the data itself. But this difference is more perceived than real. The idea of cross-validation is central to ML methods, and since that approach is intended to measure how well one's model generalizes beyond our own data, it is clear that ML people do think in terms of samples after all. Similar comments apply to ML's citing the variance-vs.-bias tradeoff, overfitting and so on

So, at the end of the day, we all are doing regression analysis, and this book takes this viewpoint.

Code and software:

The book also makes use of some of my research results and associated software. The latter is in my package **regtools**, available from CRAN [98].

A number of other packages from CRAN are used. Note that typically we use only the default values for the myriad arguments available in many functions; otherwise we could fill an entire book devoted to each package! Cross-validation is suggested for selection of tuning parameters, but with a warning that it too can be problematic.

In some cases, the **regtools** source code is also displayed within the text, so as to make clear exactly what the algorithms are doing. Similarly, data wrangling/data cleaning code is shown, not only for the purpose of "hands-on" learning, but also to highlight the importance of those topics.

Thanks:

Conversations with a number of people have directly or indirectly enhanced the quality of this book, among them Charles Abromaitis, Stuart Ambler, Doug Bates, Oleksiy Budilovsky, Yongtao Cao, Tony Corke, Tal Galili, Frank Harrell, Harlan Harris, Benjamin Hofner, Jiming Jiang, Hyunseung Kang, Martin Mächler, Erin McGinnis, John Mount, Richard Olshen, Pooja Rajkumar, Ariel Shin, Chuck Stone, Jessica Tsoi, Yu Wu, Yihui Xie, Yingkang Xie, Achim Zeileis and Jiaping Zhang.

A seminar presentation by Art Owen introduced me to the application of random effects models in recommender systems, a provocative blend of old and new. This led to the MovieLens examples and other similar examples in the book, as well as a vigorous new research interest for me. Art also led me to two Stanford statistics PhD students, Alex Chin and Jing Miao, who each read two of the chapters in great detail. Special thanks also go to Nello Cristianini, Hui Lin, Ira Sharenow and my old friend Gail Gong for their detailed feedback.

Thanks go to my editor, John Kimmel, for his encouragement and much-appreciated patience, and to the internal reviewers, David Giles and Robert Gramacy. Of course, I cannot put into words how much I owe to my wonderful wife Gamis and our daughter Laura, both of whom inspire all that I do, including this book project.

Website:

Code, errata, extra examples and so on are available at

http://heather.cs.ucdavis.edu/regclass.html.

A final comment:

My career has evolved quite a bit over the years. I wrote my dissertation in abstract probability theory [104], but turned my attention to applied statistics soon afterward. I was one of the founders of the Department of

Statistics at UC Davis. Though a few years later I transferred into the new Computer Science Department, I am still a statistician, and much of my CS research has been statistical, e.g., [100]. Most important, my interest in regression has remained strong throughout those decades.

I published my first research papers on regression methodology way back in the 1980s, and the subject has captivated me ever since. My long-held wish has been to write a regression book, and thus one can say this work is 30 years in the making. I hope you find its goals both worthy and attained. Above all, I simply hope you find it an interesting read.

List of Symbols and Abbreviations

Y: the response variable
X: vector of predictor variables
$\widetilde{X}$: X with a 1 prepended
$X^{(j)}$: the j^{th} predictor variable
n: number of observations
p: number of predictors
Y_i: value of the response variable in observation i
X_i: vector of predictors in observation i
$X_i^{(j)}$: value of the j^{th} predictor variable in observation i
A: $n \times (p+1)$ matrix of the predictor data in a linear model
D: length-n vector of the response data in a linear model
H: the *hat matrix*, $A(A'A)^{-1}A'$
$\mu(t)$: the regression function $E(Y|X=t)$
$\sigma^2(t)$: $Var(Y|X=t)$
$\widehat{\mu}(t)$: estimated value of $\mu(t)$
β: vector of coefficients in a linear/generalized linear model
$\widehat{\beta}$: estimated value of β
$'$: matrix transpose
I: multiplicative identity matrix
k-NN: k-Nearest Neighbor method
MSE: Mean Squared (Estimation) Error
MSPE: Mean Squared Prediction Error
CART: Classification and Regression Trees
SVM: Support Vector Machine
NN: neural network
PCA: Principal Components Analysis
NMF: Nonnegative Matrix Factorization
OVA: One vs. All classification

AVA: All vs. All classification
LDA: Linear Discriminant Analysis