Experimental and Quasi-Experimental Designs

for Generalized Causal Inference

Shadish | Cook | Campbell

Contents

Preface xv

1. EXPERIMENTS AND GENERALIZED CAUSAL INFERENCE 1

Experiments and Causation 3

Defining Cause, Effect, and Causal Relationships 3
Causation, Correlation, and Confounds 7
Manipulable and Nonmanipulable Causes 7
Causal Description and Causal Explanation 9

Modern Descriptions of Experiments 12

Randomized Experiment 13
Quasi-Experiment 13
Natural Experiment 17
Nonexperimental Designs 18

Experiments and the Generalization of Causal Connections 18

Most Experiments Are Highly Local But Have General Aspirations 18
Construct Validity: Causal Generalization as Representation 20
External Validity: Causal Generalization as Extrapolation 21
Approaches to Making Causal Generalizations 22

Experiments and Metascience 26

The Kuhnian Critique 27
Modern Social Psychological Critiques 28
Science and Trust 28
Implications for Experiments 29

A World Without Experiments or Causes? 31

2. STATISTICAL CONCLUSION VALIDITY AND INTERNAL VALIDITY 33

Validity 34
A Validity Typology 37
Threats to Validity 39

Statistical Conclusion Validity 42
Reporting Results of Statistical Tests of Covariation 42
Threats to Statistical Conclusion Validity 45
The Problem of Accepting the Null Hypothesis 52

Internal Validity 53
Threats to Internal Validity 54
Estimating Internal Validity in Randomized Experiments and Quasi-Experiments 61

The Relationship Between Internal Validity and Statistical Conclusion Validity 63

3. CONSTRUCT VALIDITY AND EXTERNAL VALIDITY 64

Construct Validity 64
Why Construct Inferences Are a Problem 66
Assessment of Sampling Particulars 69
Threats to Construct Validity 72
Construct Validity, Preexperimental Tailoring, and Postexperimental Specification 81

External Validity 83
Threats to External Validity 86
Constancy of Effect Size Versus Constancy of Causal Direction 90
Random Sampling and External Validity 91
Purposive Sampling and External Validity 92

More on Relationships, Tradeoffs, and Priorities 93
The Relationship Between Construct Validity and External Validity 93
The Relationship Between Internal Validity and Construct Validity 95
Tradeoffs and Priorities 96

Summary 102

4. QUASI-EXPERIMENTAL DESIGNS THAT EITHER LACK A CONTROL GROUP OR LACK PRETEST OBSERVATIONS ON THE OUTCOME 103

The Logic of Quasi-Experimentation in Brief 104

Designs Without Control Groups 106

The One-Group Posttest-Only Design 106
The One-Group Pretest–Posttest Design 108
The Removed-Treatment Design 111
The Repeated-Treatment Design 113

Designs That Use a Control Group But No Pretest 115

Posttest-Only Design With Nonequivalent Groups 115
Improving Designs Without Control Groups by Constructing Contrasts Other Than With Independent Control Groups 125
The Case-Control Design 128

Conclusion 134

5. QUASI-EXPERIMENTAL DESIGNS THAT USE BOTH CONTROL GROUPS AND PRETESTS 135

Designs That Use Both Control Groups and Pretests 136

The Untreated Control Group Design With Dependent Pretest and Posttest Samples 136
Matching Through Cohort Controls 148

Designs That Combine Many Design Elements 153

Untreated Matched Controls With Multiple Pretests and Posttests, Nonequivalent Dependent Variables, and Removed and Repeated Treatments 153
Combining Switching Replications With a Nonequivalent Control Group Design 154
An Untreated Control Group With a Double Pretest and Both Independent and Dependent Samples 154

The Elements of Design 156

Assignment 156
Measurement 158
Comparison Groups 159
Treatment 160
Design Elements and Ideal Quasi-Experimentation 160

Conclusion 161

Appendix 5.1: Important Developments in Analyzing Data From Designs With Nonequivalent Groups 161

Propensity Scores and Hidden Bias 162
Selection Bias Modeling 166
Latent Variable Structural Equation Modeling 169

6. QUASI-EXPERIMENTS: INTERRUPTED TIME-SERIES DESIGNS 171

What Is a Time Series? 172

Describing Types of Effects 172
Brief Comments on Analysis 174

Simple Interrupted Time Series 175

A Change in Intercept 175
A Change in Slope 176
Weak and Delayed Effects 178
The Usual Threats to Validity 179

Adding Other Design Features to the Basic Interrupted Time Series 181

Adding a Nonequivalent No-Treatment Control Group Time Series 182
Adding Nonequivalent Dependent Variables 184
Removing the Treatment at a Known Time 188
Adding Multiple Replications 190
Adding Switching Replications 192

Some Frequent Problems with Interrupted Time-Series Designs 195

Gradual Rather Than Abrupt Interventions 196
Delayed Causation 197
Short Time Series 198
Limitations of Much Archival Data 203

A Comment on Concomitant Time Series 205

Conclusion 206

7. REGRESSION DISCONTINUITY DESIGNS 207

The Basics of Regression Discontinuity 208

The Basic Structure 208

Examples of Regression Discontinuity Designs 212
Structural Requirements of the Design 216
Variations on the Basic Design 219

Theory of the Regression Discontinuity Design 220

Regression Discontinuities as Treatment Effects in the Randomized Experiment 221
Regression Discontinuity as a Complete Model of the Selection Process 224

Adherence to the Cutoff 227

Overrides of the Cutoff 227
Crossovers and Attrition 228
Fuzzy Regression Discontinuity 229

Threats to Validity 229

Regression Discontinuity and the Interrupted Time Series 229
Statistical Conclusion Validity and Misspecification of Functional Form 230
Internal Validity 237

Combining Regression Discontinuity and Randomized Experiments 238

Combining Regression Discontinuity and Quasi-Experiments 241

Regression Discontinuity—Experiment or Quasi-Experiment? 242

Appendix 7.1: The Logic of Statistical Proofs about Regression Discontinuity 243

8. RANDOMIZED EXPERIMENTS: RATIONALE, DESIGNS, AND CONDITIONS CONDUCIVE TO DOING THEM 246

The Theory of Random Assignment 247

What Is Random Assignment? 248
Why Randomization Works 248
Random Assignment and Units of Randomization 253
The Limited Reach of Random Assignment 256

Some Designs Used with Random Assignment 257

The Basic Design 257
The Pretest-Posttest Control Group Design 261
Alternative-Treatments Design with Pretest 261

Multiple Treatments and Controls with Pretest 262
Factorial Designs 263
Longitudinal Designs 266
Crossover Designs 268

Conditions Most Conducive to Random Assignment 269

When Demand Outstrips Supply 269
When an Innovation Cannot Be Delivered to All Units at Once 270
When Experimental Units Can Be Temporally Isolated: The Equivalent-Time-Samples Design 270
When Experimental Units Are Spatially Separated or Interunit Communication Is Low 271
When Change Is Mandated and Solutions Are Acknowledged to Be Unknown 272
When a Tie Can Be Broken or Ambiguity About Need Can Be Resolved 273
When Some Persons Express No Preference Among Alternatives 273
When You Can Create Your Own Organization 274
When You Have Control over Experimental Units 274
When Lotteries Are Expected 275

When Random Assignment Is Not Feasible or Desirable 276

Discussion 277

9. PRACTICAL PROBLEMS 1: ETHICS, PARTICIPANT RECRUITMENT, AND RANDOM ASSIGNMENT 279

Ethical and Legal Issues with Experiments 280

The Ethics of Experimentation 281
Withholding a Potentially Effective Treatment 283
The Ethics of Random Assignment 286
Discontinuing Experiments for Ethical Reasons 289
Legal Problems in Experiments 290

Recruiting Participants to Be in the Experiment 292

Improving the Random Assignment Process 294

Methods of Randomization 294
What to Do If Pretest Means Differ 303
Matching and Stratifying 304
Matching and Analysis of Covariance 306
The Human Side of Random Assignment 307

Conclusion 311

Appendix 9.1: Random Assignment by Computer 311
SPSS and SAS 311
World Wide Web 313
Excel 313

10. PRACTICAL PROBLEMS 2: TREATMENT IMPLEMENTATION AND ATTRITION 314

Problems Related to Treatment Implementation 315
Inducing and Measuring Implementation 315
Analyses Taking Implementation into Account 320

Post-Assignment Attrition 323
Defining the Attrition Problem 323
Preventing Attrition 324
Analyses of Attrition 334

Discussion 340

11. GENERALIZED CAUSAL INFERENCE: A GROUNDED THEORY 341

The Received View of Generalized Causal Inference: Formal Sampling 342
Formal Sampling of Causes and Effects 344
Formal Sampling of Persons and Settings 346
Summary 348

A Grounded Theory of Generalized Causal Inference 348
Exemplars of How Scientists Make Generalizations 349
Five Principles of Generalized Causal Inferences 353
The Use of Purposive Sampling Strategies 354
Applying the Five Principles to Construct and External Validity 356
Should Experimenters Apply These Principles to All Studies? 371
Prospective and Retrospective Uses of These Principles 372

Discussion 373

12. GENERALIZED CAUSAL INFERENCE: METHODS FOR SINGLE STUDIES 374

Purposive Sampling and Generalized Causal Inference 374
Purposive Sampling of Typical Instances 375
Purposive Sampling of Heterogeneous Instances 376

Purposive Sampling and the First Four Principles 378
Statistical Methods for Generalizing from Purposive Samples 386

Methods for Studying Causal Explanation 389

Qualitative Methods 389
Statistical Models of Causal Explanation 392
Experiments That Manipulate Explanations 414

Conclusion 415

13. GENERALIZED CAUSAL INFERENCE: METHODS FOR MULTIPLE STUDIES 417

Generalizing from Single Versus Multiple Studies 418

Multistudy Programs of Research 418

Phased Models of Increasingly Generalizable Studies 419
Directed Programs of Experiments 420

Narrative Reviews of Existing Research 421

Narrative Reviews of Experiments 422
Narrative Reviews Combining Experimental and Nonexperimental Research 423
Problems with Narrative Reviews 423

Quantitative Reviews of Existing Research 425

The Basics of Meta-Analysis 426
Meta-Analysis and the Five Principles of Generalized Causal Inference 435
Discussion of Meta-Analysis 445

Appendix 13.1: Threats to the Validity of Meta-Analyses 446

Threats to Inferences About the Existence of a Relationship Between Treatment and Outcome 447
Threats to Inferences About the Causal Relationship Between Treatment and Outcome 450
Threats to Inferences About the Constructs Represented in Meta-Analyses 453
Threats to Inferences About External Validity in Meta-Analyses 454

14. A CRITICAL ASSESSMENT OF OUR ASSUMPTIONS 456

Causation and Experimentation 453

Causal Arrows and Pretzels 453
Epistemological Criticisms of Experiments 459
Neglected Ancillary Questions 461

Validity 462

Objections to Internal Validity 462
Objections Concerning the Discrimination Between Construct Validity and External Validity 466
Objections About the Completeness of the Typology 473
Objections Concerning the Nature of Validity 475

Quasi-Experimentation 484

Criteria for Ruling Out Threats: The Centrality of Fuzzy Plausibility 484
Pattern Matching as a Problematic Criterion 485
The Excuse Not to Do a Randomized Experiment 486

Randomized Experiments 488

Experiments Cannot Be Successfully Implemented 488
Experimentation Needs Strong Theory and Standardized Treatment Implementation 489
Experiments Entail Tradeoffs Not Worth Making 490
Experiments Assume an Invalid Model of Research Utilization 493
The Conditions of Experimentation Differ from the Conditions of Policy Implementation 495
Imposing Treatments Is Fundamentally Flawed Compared with Encouraging the Growth of Local Solutions to Problems 497

Causal Generalization: An Overly Complicated Theory? 498

Nonexperimental Alternatives 499

Intensive Qualitative Case Studies 500
Theory-Based Evaluations 501
Weaker Quasi-Experiments 502
Statistical Controls 503

Conclusion 504

Glossary 505

References 514

Name Index 593

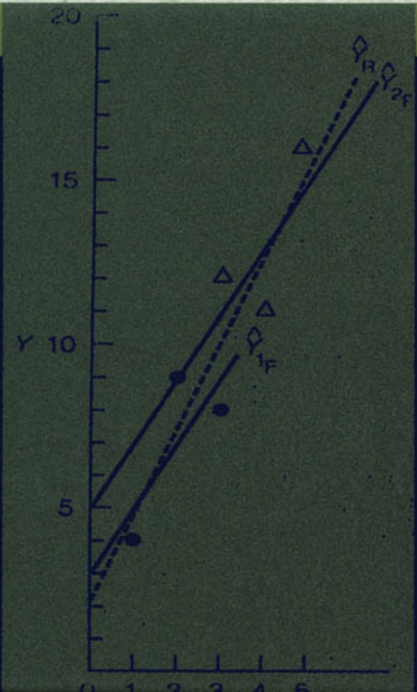
Subject Index 609

CD Enclosed

Designing Experiments and Analyzing Data

Second Edition

A Model Comparison Perspective



Scott E. Maxwell
Harold D. Delaney

TLFeBOOK

Contents

Preface xvii

I CONCEPTUAL BASES OF EXPERIMENTAL DESIGN AND ANALYSIS

1 **The Logic of Experimental Design** 3
The Traditional View of Science 3
Responses to the Criticisms of the Idea of Pure Science 5
Assumptions 5
Modern Philosophy of Science 10
Threats to the Validity of Inferences from Experiments 22
Types of Validity 23
Conceptualizing and Controlling for Threats to Validity 30
Exercises 32

2 **Introduction to the Fisher Tradition** 34
"Interpretation and its Reasoned Basis" 35
A Discrete Probability Example 37
Randomization Test 41
Of Hypotheses and p Values: Fisher Versus Neyman–Pearson 47
Toward Tests Based on Distributional Assumptions 49
Statistical Tests with Convenience Samples 49
The Assumption of Normality 50
Overview of Experimental Designs to be Considered 56
Exercises 59

II MODEL COMPARISONS FOR BETWEEN-SUBJECTS DESIGNS

3 **Introduction to Model Comparisons: One-Way Between-Subjects Designs** 67
The General Linear Model 69
One-Group Situation 71
Basics of Models 71
Proof That $\bar{Y}$ Is the Least-Squares Estimate of μ (*Optional*) 73
Development of the General Form of the Test Statistic 75
Numerical Example 78
Relationship of Models and Hypotheses 80
Two-Group Situation 80
Development in Terms of Models 80

TLFeBOOK

Alternative Development and Identification with Traditional Terminology 83
Tests of Replication (*Optional*) 85
The General Case of One-Way Designs 88
Formulation in Terms of Models 88
Numerical Example 91
A Model in Terms of Effects 93
On Tests of Significance and Measures of Effect 98
Measures of Effect 100
Measures of Effect Size 101
Measures of Association Strength 104
Alternative Representations of Effects 107
Statistical Assumptions 110
Implications for Expected Values 110
Robustness of ANOVA 111
Checking for Normality and Homogeneity of Variance 114
Transformations 117
Power of the *F* Test: One-Way ANOVA 120
Determining Sample Size Using *d* and Table 3.10 123
Pilot Data and Observed Power 124
Exercises 126
Extension: Robust Methods for One-Way Between-Subject Designs 129
Parametric Modifications 131
Nonparametric Approaches 136
Choosing Between Parametric and Nonparametric Tests 137
Two Other Approaches (*Optional*) 143
Why Does the Usual *F* Test Falter with Unequal *n*s When Population Variances Are Unequal? (*Optional*) 145
Exercises 147

4 Individual Comparisons of Means 149
A Model Comparison Approach for Testing Individual Comparisons 150
Preview of Individual Comparisons 150
Relationship to Model Comparisons 150
Derivation of Parameter Estimates and Sum of Squared Errors (*Optional*) 152
Expression of *F* Statistic 153
Numerical Example 155
Complex Comparisons 157
Models Perspective 157
Numerical Example 162
The *t* Test Formulation of Hypothesis Testing for Contrasts 163
Practical Implications 164
Unequal Population Variances 165
Numerical Example 168
Measures of Effect 169
Measures of Effect Size 170
Measures of Association Strength 173
Testing More Than One Contrast 177
How Many Contrasts Should Be Tested? 177
Linear Independence of Contrasts 178
Orthogonality of Contrasts 179

TLFeBOOK

Example of Correlation Between Nonorthogonal Contrasts (*Optional*) 180
Another Look at Nonorthogonal Contrasts: Venn Diagrams 182
Exercises 186
Extension: Derivation of Sum of Squares for a Contrast 190

5 Testing Several Contrasts: The Multiple-Comparison Problem 193
Multiple Comparisons 193
Experimentwise and Per-Comparison Error Rates 193
Simultaneous Confidence Intervals 196
Levels of Strength of Inference 197
Types of Contrasts 198
Overview of Techniques 199
Planned Versus Post Hoc Contrasts 200
Multiple Planned Comparisons 201
Bonferroni Adjustment 202
Modification of the Bonferroni Approach With Unequal Variances 205
Numerical Example 206
Pairwise Comparisons 208
Tukey's WSD Procedure 210
Modifications of Tukey's WSD 212
Numerical Example 213
Post Hoc Complex Comparisons 213
Proof That $SS_{\max} = SS_{\mathrm{B}}$ 215
Comparison of Scheffé to Bonferroni and Tukey 217
Modifications of Scheffé's Method 218
Numerical Example 219
Other Multiple-Comparison Procedures 221
Dunnett's Procedure for Comparisons with a Control 221
Numerical Example 222
Procedures for Comparisons with the Best 223
Numerical Example 227
Fisher's LSD (Protected t) 229
False Discovery Rate 230
Choosing an Appropriate Procedure 234
Exercises 237

6 Trend Analysis 243
Quantitative Factors 243
Statistical Treatment of Trend Analysis 244
The Slope Parameter 246
Numerical Example 248
Hypothesis Test of Slope Parameter 249
Confidence Interval and Other Effect Size Measures for Slope Parameter 251
Numerical Example 251
Testing for Nonlinearity 254
Numerical Example 257
Testing Individual Higher Order Trends 257
Contrast Coefficients for Higher Order Trends 259
Numerical Example 260

TLFeBOOK

Further Examination of Nonlinear Trends 263
Trend Analysis with Unequal Sample Sizes 267
Concluding Comments 269
Exercises 269

7 Two-Way Between-Subjects Factorial Designs 275
The 2 × 2 Design 275
The Concept of Interaction 277
Additional Perspectives on the Interaction 278
A Model Comparison Approach to the General Two-Factor Design 280
Alternate Form of Full Model 280
Comparison of Models for Hypothesis Testing 284
Numerical Example 290
Familywise Control of Alpha Level 291
Measures of Effect 291
Follow-Up Tests 297
Further Investigation of Main Effects 297
Marginal Mean Comparisons Without Homogeneity Assumption (*Optional*) 300
Further Investigation of an Interaction—Simple Effects 301
An Alternative Method for Investigating an Interaction—Interaction Contrasts 309
Statistical Power 317
Advantages of Factorial Designs 319
Nonorthogonal Designs 320
Design Considerations 321
Relationship Between Design and Analysis 321
Analysis of the 2 × 2 Nonorthogonal Design 322
Test of the Interaction 322
Unweighted Marginal Means and Type III Sum of Squares 324
Unweighted Versus Weighted Marginal Means 325
Type II Sum of Squares 327
Summary of Three Types of Sum of Squares 328
Analysis of the General *a* × *b* Nonorthogonal Design 329
Test of the Interaction 329
Test of Unweighted Marginal Means 330
Test of Marginal Means in an Additive Model 331
Test of Weighted Marginal Means 332
Summary of Types of Sum of Squares 333
Which Type of Sum of Squares Is Best? 334
A Note on Statistical Packages for Analyzing Nonorthogonal Designs 335
Numerical Example 337
Final Remarks 343
Exercises 343

8 Higher Order Between-Subjects Factorial Designs 354
The 2 × 2 × 2 Design 354
The Meaning of Main Effects 355
The Meaning of Two-Way Interactions 356

TLFeBOOK

The Meaning of the Three-Way Interaction 357
Graphical Depiction 359
Further Consideration of the Three-Way Interaction 361
Summary of Meaning of Effects 366
The General $A \times B \times C$ Design 367
The Full Model 367
Formulation of Restricted Models 368
Numerical Example 372
Implications of a Three-Way Interaction 374
General Guideline for Analyzing Effects 376
Summary of Results 381
Graphical Depiction of Data 382
Confidence Intervals for Single Degree of Freedom Effects 383
Other Questions of Potential Interest 386
Tests to Be Performed When the Three-Way Interaction Is Nonsignificant 387
Nonorthogonal Designs 389
Higher Order Designs 391
Exercises 392

9 Designs With Covariates: ANCOVA and Blocking 399
ANCOVA 401
The Logic of ANCOVA 401
Linear Models for ANCOVA 403
Two Consequences of Using ANCOVA 414
Assumptions in ANCOVA 420
Numerical Example 428
Measures of Effect 431
Comparisons Among Adjusted Group Means 434
Generalizations of the ANCOVA Model 438
Choosing Covariates in Randomized Designs 439
Sample Size Planning and Power Analysis in ANCOVA 441
Alternate Methods of Analyzing Designs with Concomitant Variables 443
ANOVA of Residuals 444
Gain Scores 444
Blocking 448
Exercises 453
Extension: Heterogeneity of Regression 456
Test for Heterogeneity of Regression 456
Accommodating Heterogeneity of Regression 460

10 Designs with Random or Nested Factors 469
Designs with Random Factors 469
Introduction to Random Effects 469
One-Factor Case 471
Two-Factor Case 474
Numerical Example 481
Alternative Tests and Design Considerations with Random Factors 483
Follow-up Tests and Confidence Intervals 484

TLFeBOOK

Measures of Association Strength 485

Using Statistical Computer Programs to Analyze Designs with Random Factors 489

Determining Power in Designs with Random Factors 490

Designs with Nested Factors 494

Introduction to Nested Factors 494

Example 499

Models and Tests 499

Degrees of Freedom 504

Statistical Assumptions and Related Issues 506

Follow-up Tests and Confidence Intervals 507

Strength of Association in Nested Designs 508

Using Statistical Computer Programs to Analyze Nested Designs 509

Selection of Error Terms When Nested Factors Are Present 510

Complications That Arise in More Complex Designs 512

Exercises 517

III MODEL COMPARISONS FOR DESIGNS INVOLVING WITHIN-SUBJECTS FACTORS

11 One-Way Within-Subjects Designs: Univariate Approach 525

Prototypical Within-Subjects Designs 525

Advantages of Within-Subjects Designs 527

Analysis of Repeated Measures Designs with Two Levels 527

The Problem of Correlated Errors 527

Reformulation of Model 529

Analysis of Within-Subjects Designs with More Than Two Levels 531

Traditional Univariate (Mixed-Model) Approach 532

Comparison of Full and Restricted Models 533

Estimation of Parameters: Numerical Example 534

Assumptions in the Traditional Univariate (Mixed-Model) Approach 539

Homogeneity, Sphericity, and Compound Symmetry 540

Numerical Example 541

Adjusted Univariate Tests 542

Lower-Bound Adjustment 543

$\hat{\varepsilon}$ Adjustment 543

$\tilde{\varepsilon}$ Adjustment 544

Summary of Four Mixed-Model Approaches 545

Measures of Effect 547

Comparisons Among Individual Means 550

Confidence Intervals for Comparisons 551

Confidence Intervals with Pooled and Separate Variances (*Optional*) 553

Considerations in Designing Within-Subjects Experiments 555

Order Effects 556

Differential Carryover Effects 556

Controlling for Order Effects with More Than Two Levels: Latin Square Designs 557

TLFeBOOK

Relative Advantages of Between-Subjects and Within-Subjects Designs 561
Intraclass Correlations for Assessing Reliability 563
Exercises 567

12 Higher-Order Designs with Within-Subjects Factors: Univariate Approach 573
Designs with Two Within-Subjects Factors 573
Omnibus Tests 574
Numerical Example 577
Further Investigation of Main Effects 579
Further Investigation of an Interaction—Simple Effects 581
Interaction Contrasts 582
Statistical Packages and Pooled Error Terms Versus Separate Error Terms 583
Assumptions 584
Adjusted Univariate Tests 588
Confidence Intervals 590
Quasi-F Ratios 590
One Within-Subjects Factor and One Between-Subjects Factor in the Same Design 592
Omnibus Tests 593
Further Investigation of Main Effects 599
Further Investigation of an Interaction—Simple Effects 601
Interaction Contrasts 605
Assumptions 607
Adjusted Univariate Tests 609
More Complex Designs 610
Designs with Additional Factors 610
Latin Square Designs 611
Exercises 616

13 One-Way Within-Subjects Designs: Multivariate Approach 624
A Brief Review of Analysis for Designs with Two Levels 624
Multivariate Analysis of Within-Subjects Designs with Three Levels 626
Need for Multiple D Variables 626
Full and Restricted Models 627
The Relationship Between D_1 and D_2 629
Matrix Formulation and Determinants 630
Test Statistic 632
Multivariate Analysis of Within-Subjects Designs with a Levels 633
Forming D Variables 633
Test Statistic 635
Numerical Example 635
Measures of Effect 638
Choosing an Appropriate Sample Size 639
Choice of D Variables 645
Tests of Individual Contrasts 647
Quantitative Repeated Factors (*Optional*) 649

TLFeBOOK

Multiple-Comparison Procedures: Determination of Critical Values 650
Planned Comparisons 651
Pairwise Comparisons 651
Post Hoc Complex Comparisons 652
Finding $D_{\max}$ (*Optional*) 653
Confidence Intervals for Contrasts 654
The Relationship Between the Multivariate Approach and the Mixed-Model Approach 658
Orthonormal Contrasts 658
Comparison of the Two Approaches 660
Reconceptualization of ε in Terms of $E^*(F)$ (*Optional*) 663
Multivariate and Mixed-Model Approaches for Testing Contrasts 665
Numerical Example 666
The Difference in Error Terms 667
Which Error Term Is Better? 669
A General Comparison of the Multivariate and Mixed-Model Approaches 671
Assumptions 672
Tests of Contrasts 672
Type I Error Rates 672
Type II Error Rates 673
Summary 675
Exercises 676

14 Higher Order Designs with Within-Subjects Factors: Multivariate Approach 682
Two Within-Subjects Factors, Each with Two Levels 682
Formation of Main-Effect D Variables 684
Formation of Interaction D Variables 686
Relationship to the Mixed-Model Approach 688
Multivariate Analysis of Two-Way $a \times b$ Within-Subjects Designs 688
Formation of Main-Effect D Variables 688
Formation of Interaction D Variables 691
Omnibus Tests—Multivariate Significance Tests 693
Measures of Effect 694
Further Investigation of Main Effects 695
Further Investigation of an Interaction—Simple Effects 696
Interaction Contrasts 698
Confidence Intervals for Contrasts 699
The Relationship Between the Multivariate and the Mixed-Model Approaches (*Optional*) 701
Multivariate and Mixed-Model Approaches for Testing Contrasts 703
Comparison of the Multivariate and Mixed-Model Approaches 704
One Within-Subjects Factor and One Between-Subjects Factor in the Same Design 704
Split-Plot Design With Two Levels of the Repeated Factor 704
General $a \times b$ Split-Plot Design 713
Measures of Effect 725

TLFeBOOK

Confidence Intervals for Contrasts 738
The Relationship Between the Multivariate and the Mixed-Model Approaches (*Optional*) 742
Assumptions of the Multivariate Approach 744
Multivariate and Mixed-Model Approaches for Testing Within-Subjects Contrasts 745
Comparison of the Multivariate and Mixed-Model Approaches 746
More Complex Designs (*Optional*) 746
Exercises 752

IV ALTERNATIVE ANALYSIS STRATEGIES

15 An Introduction to Multilevel Models for Within-Subjects Designs 763
Advantages of New Methods 763
Within-Subjects Designs 763
Overview of Remainder of Chapter 764
Within-Subjects Designs 765
Various Types of Within-Subjects Designs 765
Models for Longitudinal Data 765
Review of the ANOVA Mixed-Model Approach 766
Random Effects Models 767
Maximum Likelihood Approach 767
An Example of Maximum Likelihood Estimation (*Optional*) 768
Comparison of ANOVA and Maximum Likelihood Models 770
Numerical Example 773
A Closer Look at the Random Effects Model 778
Graphical Representation of Longitudinal Data 779
Graphical Representation of the Random Intercept Model 781
Coding Random Effects Predictor Variables 785
Random Effects Parameters 786
Numerical Example 788
Graphical Representation of a Model With Random Slope and Intercept 790
Further Consideration of Competing Models 791
Additional Models Deserving Consideration 793
Graphical Representation of a Growth Curve Model 798
Design Considerations 800
An Alternative to the Random Effects Model 802
Additional Covariance Matrix Structures 809
Tests of Contrasts 813
Overview of Broader Model Comparison 814
Complex Designs 816
Factorial Fixed Effects 817
Multiple Variables Measured Over Time 818
Unbalanced Designs 818
Conclusion 820
Exercises 820

TLFeBOOK

16 An Introduction to Multilevel Hierarchical Mixed Models: Nested Designs **828**
Review of the ANOVA Approach **829**
Maximum Likelihood Analysis Models for the Simple Nested Design **831**
Numerical Example—Equal *n* **833**
Numerical Example—Unequal *n* **840**
Maximum Likelihood Analysis Models for Complex Nested Designs **845**
Hierarchical Representation of the Model for a Simple Nested Design **846**
Models With Additional Level 2 Variables **849**
Models with Additional Level 1 Variables **853**
Exercises **867**

Appendixes

A Statistical Tables **A-1**
B Part 1. Linear Models: The Relation Between ANOVA and Regression **B-1**
Part 2. A Brief Primer of Principles of Formulating and Comparing Models **B-26**
C Notes **C-1**
D Solutions to Selected Exercises **D-1**
E References **E-1**
Name Index **N-1**
Subject Index **S-1**

TLFeBOOK

Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Springer

Contents

Preface **vii**

1 Introduction **1**

2 Statistical Learning **15**

2.1 What Is Statistical Learning? 15

2.1.1 Why Estimate f? 17

2.1.2 How Do We Estimate f? 21

2.1.3 The Trade-Off Between Prediction Accuracy and Model Interpretability 24

2.1.4 Supervised Versus Unsupervised Learning 26

2.1.5 Regression Versus Classification Problems 28

2.2 Assessing Model Accuracy 29

2.2.1 Measuring the Quality of Fit 29

2.2.2 The Bias-Variance Trade-Off 33

2.2.3 The Classification Setting 37

2.3 Lab: Introduction to R 42

2.3.1 Basic Commands 42

2.3.2 Graphics 45

2.3.3 Indexing Data 47

2.3.4 Loading Data 48

2.3.5 Additional Graphical and Numerical Summaries . . 49

2.4 Exercises 52

3 Linear Regression 59
3.1 Simple Linear Regression . . . 61
3.1.1 Estimating the Coefficients . . . 61
3.1.2 Assessing the Accuracy of the Coefficient Estimates . . . 63
3.1.3 Assessing the Accuracy of the Model . . . 68
3.2 Multiple Linear Regression . . . 71
3.2.1 Estimating the Regression Coefficients . . . 72
3.2.2 Some Important Questions . . . 75
3.3 Other Considerations in the Regression Model . . . 82
3.3.1 Qualitative Predictors . . . 82
3.3.2 Extensions of the Linear Model . . . 86
3.3.3 Potential Problems . . . 92
3.4 The Marketing Plan . . . 102
3.5 Comparison of Linear Regression with K-Nearest Neighbors . . . 104
3.6 Lab: Linear Regression . . . 109
3.6.1 Libraries . . . 109
3.6.2 Simple Linear Regression . . . 110
3.6.3 Multiple Linear Regression . . . 113
3.6.4 Interaction Terms . . . 115
3.6.5 Non-linear Transformations of the Predictors . . . 115
3.6.6 Qualitative Predictors . . . 117
3.6.7 Writing Functions . . . 119
3.7 Exercises . . . 120

4 Classification 127
4.1 An Overview of Classification . . . 128
4.2 Why Not Linear Regression? . . . 129
4.3 Logistic Regression . . . 130
4.3.1 The Logistic Model . . . 131
4.3.2 Estimating the Regression Coefficients . . . 133
4.3.3 Making Predictions . . . 134
4.3.4 Multiple Logistic Regression . . . 135
4.3.5 Logistic Regression for >2 Response Classes . . . 137
4.4 Linear Discriminant Analysis . . . 138
4.4.1 Using Bayes' Theorem for Classification . . . 138
4.4.2 Linear Discriminant Analysis for $p = 1$ . . . 139
4.4.3 Linear Discriminant Analysis for $p >1$ . . . 142
4.4.4 Quadratic Discriminant Analysis . . . 149
4.5 A Comparison of Classification Methods . . . 151
4.6 Lab: Logistic Regression, LDA, QDA, and KNN . . . 154
4.6.1 The Stock Market Data . . . 154
4.6.2 Logistic Regression . . . 156
4.6.3 Linear Discriminant Analysis . . . 161

4.6.4 Quadratic Discriminant Analysis 163
4.6.5 K-Nearest Neighbors 163
4.6.6 An Application to Caravan Insurance Data 165
4.7 Exercises 168

5 Resampling Methods 175
5.1 Cross-Validation 176
5.1.1 The Validation Set Approach 176
5.1.2 Leave-One-Out Cross-Validation 178
5.1.3 k-Fold Cross-Validation 181
5.1.4 Bias-Variance Trade-Off for k-Fold Cross-Validation 183
5.1.5 Cross-Validation on Classification Problems 184
5.2 The Bootstrap 187
5.3 Lab: Cross-Validation and the Bootstrap 190
5.3.1 The Validation Set Approach 191
5.3.2 Leave-One-Out Cross-Validation 192
5.3.3 k-Fold Cross-Validation 193
5.3.4 The Bootstrap 194
5.4 Exercises 197

6 Linear Model Selection and Regularization 203
6.1 Subset Selection 205
6.1.1 Best Subset Selection 205
6.1.2 Stepwise Selection 207
6.1.3 Choosing the Optimal Model 210
6.2 Shrinkage Methods 214
6.2.1 Ridge Regression 215
6.2.2 The Lasso 219
6.2.3 Selecting the Tuning Parameter 227
6.3 Dimension Reduction Methods 228
6.3.1 Principal Components Regression 230
6.3.2 Partial Least Squares 237
6.4 Considerations in High Dimensions 238
6.4.1 High-Dimensional Data 238
6.4.2 What Goes Wrong in High Dimensions? 239
6.4.3 Regression in High Dimensions 241
6.4.4 Interpreting Results in High Dimensions 243
6.5 Lab 1: Subset Selection Methods 244
6.5.1 Best Subset Selection 244
6.5.2 Forward and Backward Stepwise Selection 247
6.5.3 Choosing Among Models Using the Validation Set Approach and Cross-Validation 248

6.6 Lab 2: Ridge Regression and the Lasso 251
6.6.1 Ridge Regression 251
6.6.2 The Lasso 255
6.7 Lab 3: PCR and PLS Regression 256
6.7.1 Principal Components Regression 256
6.7.2 Partial Least Squares 258
6.8 Exercises 259

7 Moving Beyond Linearity 265
7.1 Polynomial Regression 266
7.2 Step Functions 268
7.3 Basis Functions 270
7.4 Regression Splines 271
7.4.1 Piecewise Polynomials 271
7.4.2 Constraints and Splines 271
7.4.3 The Spline Basis Representation 273
7.4.4 Choosing the Number and Locations of the Knots 274
7.4.5 Comparison to Polynomial Regression 276
7.5 Smoothing Splines 277
7.5.1 An Overview of Smoothing Splines 277
7.5.2 Choosing the Smoothing Parameter λ 278
7.6 Local Regression 280
7.7 Generalized Additive Models 282
7.7.1 GAMs for Regression Problems 283
7.7.2 GAMs for Classification Problems 286
7.8 Lab: Non-linear Modeling 287
7.8.1 Polynomial Regression and Step Functions 288
7.8.2 Splines 293
7.8.3 GAMs 294
7.9 Exercises 297

8 Tree-Based Methods 303
8.1 The Basics of Decision Trees 303
8.1.1 Regression Trees 304
8.1.2 Classification Trees 311
8.1.3 Trees Versus Linear Models 314
8.1.4 Advantages and Disadvantages of Trees 315
8.2 Bagging, Random Forests, Boosting 316
8.2.1 Bagging 316
8.2.2 Random Forests 320
8.2.3 Boosting 321
8.3 Lab: Decision Trees 324
8.3.1 Fitting Classification Trees 324
8.3.2 Fitting Regression Trees 327

8.3.3 Bagging and Random Forests 328
8.3.4 Boosting 330
8.4 Exercises 332

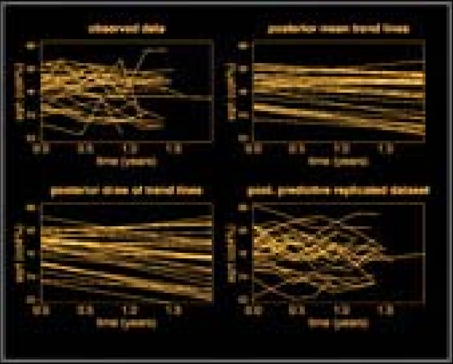
9 Support Vector Machines 337
9.1 Maximal Margin Classifier 338
9.1.1 What Is a Hyperplane? 338
9.1.2 Classification Using a Separating Hyperplane 339
9.1.3 The Maximal Margin Classifier 341
9.1.4 Construction of the Maximal Margin Classifier . . . 342
9.1.5 The Non-separable Case 343
9.2 Support Vector Classifiers 344
9.2.1 Overview of the Support Vector Classifier 344
9.2.2 Details of the Support Vector Classifier 345
9.3 Support Vector Machines 349
9.3.1 Classification with Non-linear Decision Boundaries 349
9.3.2 The Support Vector Machine 350
9.3.3 An Application to the Heart Disease Data 354
9.4 SVMs with More than Two Classes 355
9.4.1 One-Versus-One Classification 355
9.4.2 One-Versus-All Classification 356
9.5 Relationship to Logistic Regression 356
9.6 Lab: Support Vector Machines 359
9.6.1 Support Vector Classifier 359
9.6.2 Support Vector Machine 363
9.6.3 ROC Curves 365
9.6.4 SVM with Multiple Classes 366
9.6.5 Application to Gene Expression Data 366
9.7 Exercises 368

10 Unsupervised Learning 373
10.1 The Challenge of Unsupervised Learning 373
10.2 Principal Components Analysis 374
10.2.1 What Are Principal Components? 375
10.2.2 Another Interpretation of Principal Components . . 379
10.2.3 More on PCA 380
10.2.4 Other Uses for Principal Components 385
10.3 Clustering Methods 385
10.3.1 K-Means Clustering 386
10.3.2 Hierarchical Clustering 390
10.3.3 Practical Issues in Clustering 399
10.4 Lab 1: Principal Components Analysis 401

10.5 Lab 2: Clustering 404
10.5.1 K-Means Clustering 404
10.5.2 Hierarchical Clustering 406
10.6 Lab 3: NCI60 Data Example 407
10.6.1 PCA on the NCI60 Data 408
10.6.2 Clustering the Observations of the NCI60 Data . . . 410
10.7 Exercises 413

Index 419

ANALYTICAL METHODS FOR SOCIAL RESEARCH



Data Analysis Using Regression and Multilevel/Hierarchical Models

ANDREW GELMAN
JENNIFER HILL

CAMBRIDGE

CAMBRIDGE www.cambridge.org/9780521867061

Contents

List of examples *page* xvii

Preface xix

1 Why? 1
1.1 What is multilevel regression modeling? 1
1.2 Some examples from our own research 3
1.3 Motivations for multilevel modeling 6
1.4 Distinctive features of this book 8
1.5 Computing 9

2 Concepts and methods from basic probability and statistics 13
2.1 Probability distributions 13
2.2 Statistical inference 16
2.3 Classical confidence intervals 18
2.4 Classical hypothesis testing 20
2.5 Problems with statistical significance 22
2.6 55,000 residents desperately need your help! 23
2.7 Bibliographic note 26
2.8 Exercises 26

Part 1A: Single-level regression 29

3 Linear regression: the basics 31
3.1 One predictor 31
3.2 Multiple predictors 32
3.3 Interactions 34
3.4 Statistical inference 37
3.5 Graphical displays of data and fitted model 42
3.6 Assumptions and diagnostics 45
3.7 Prediction and validation 47
3.8 Bibliographic note 49
3.9 Exercises 49

4 Linear regression: before and after fitting the model 53
4.1 Linear transformations 53
4.2 Centering and standardizing, especially for models with interactions 55
4.3 Correlation and "regression to the mean" 57
4.4 Logarithmic transformations 59
4.5 Other transformations 65
4.6 Building regression models for prediction 68
4.7 Fitting a series of regressions 73

4.8 Bibliographic note 74
4.9 Exercises 74

5 Logistic regression 79
5.1 Logistic regression with a single predictor 79
5.2 Interpreting the logistic regression coefficients 81
5.3 Latent-data formulation 85
5.4 Building a logistic regression model: wells in Bangladesh 86
5.5 Logistic regression with interactions 92
5.6 Evaluating, checking, and comparing fitted logistic regressions 97
5.7 Average predictive comparisons on the probability scale 101
5.8 Identifiability and separation 104
5.9 Bibliographic note 105
5.10 Exercises 105

6 Generalized linear models 109
6.1 Introduction 109
6.2 Poisson regression, exposure, and overdispersion 110
6.3 Logistic-binomial model 116
6.4 Probit regression: normally distributed latent data 118
6.5 Multinomial regression 119
6.6 Robust regression using the t model 124
6.7 Building more complex generalized linear models 125
6.8 Constructive choice models 127
6.9 Bibliographic note 131
6.10 Exercises 132

Part 1B: Working with regression inferences 135

7 Simulation of probability models and statistical inferences 137
7.1 Simulation of probability models 137
7.2 Summarizing linear regressions using simulation: an informal Bayesian approach 140
7.3 Simulation for nonlinear predictions: congressional elections 144
7.4 Predictive simulation for generalized linear models 148
7.5 Bibliographic note 151
7.6 Exercises 152

8 Simulation for checking statistical procedures and model fits 155
8.1 Fake-data simulation 155
8.2 Example: using fake-data simulation to understand residual plots 157
8.3 Simulating from the fitted model and comparing to actual data 158
8.4 Using predictive simulation to check the fit of a time-series model 163
8.5 Bibliographic note 165
8.6 Exercises 165

9 Causal inference using regression on the treatment variable 167
9.1 Causal inference and predictive comparisons 167
9.2 The fundamental problem of causal inference 170
9.3 Randomized experiments 172
9.4 Treatment interactions and poststratification 178

9.5 Observational studies 181
9.6 Understanding causal inference in observational studies 186
9.7 Do not control for post-treatment variables 188
9.8 Intermediate outcomes and causal paths 190
9.9 Bibliographic note 194
9.10 Exercises 194

10 Causal inference using more advanced models 199
10.1 Imbalance and lack of complete overlap 199
10.2 Subclassification: effects and estimates for different subpopulations 204
10.3 Matching: subsetting the data to get overlapping and balanced treatment and control groups 206
10.4 Lack of overlap when the assignment mechanism is known: regression discontinuity 212
10.5 Estimating causal effects indirectly using instrumental variables 215
10.6 Instrumental variables in a regression framework 220
10.7 Identification strategies that make use of variation within or between groups 226
10.8 Bibliographic note 229
10.9 Exercises 231

Part 2A: Multilevel regression 235

11 Multilevel structures 237
11.1 Varying-intercept and varying-slope models 237
11.2 Clustered data: child support enforcement in cities 237
11.3 Repeated measurements, time-series cross sections, and other non-nested structures 241
11.4 Indicator variables and fixed or random effects 244
11.5 Costs and benefits of multilevel modeling 246
11.6 Bibliographic note 247
11.7 Exercises 248

12 Multilevel linear models: the basics 251
12.1 Notation 251
12.2 Partial pooling with no predictors 252
12.3 Partial pooling with predictors 254
12.4 Quickly fitting multilevel models in R 259
12.5 Five ways to write the same model 262
12.6 Group-level predictors 265
12.7 Model building and statistical significance 270
12.8 Predictions for new observations and new groups 272
12.9 How many groups and how many observations per group are needed to fit a multilevel model? 275
12.10 Bibliographic note 276
12.11 Exercises 277

13 Multilevel linear models: varying slopes, non-nested models, and other complexities 279
13.1 Varying intercepts and slopes 279
13.2 Varying slopes without varying intercepts 283

13.3 Modeling multiple varying coefficients using the scaled inverse-Wishart distribution 284
13.4 Understanding correlations between group-level intercepts and slopes 287
13.5 Non-nested models 289
13.6 Selecting, transforming, and combining regression inputs 293
13.7 More complex multilevel models 297
13.8 Bibliographic note 297
13.9 Exercises 298

14 Multilevel logistic regression 301
14.1 State-level opinions from national polls 301
14.2 Red states and blue states: what's the matter with Connecticut? 310
14.3 Item-response and ideal-point models 314
14.4 Non-nested overdispersed model for death sentence reversals 320
14.5 Bibliographic note 321
14.6 Exercises 322

15 Multilevel generalized linear models 325
15.1 Overdispersed Poisson regression: police stops and ethnicity 325
15.2 Ordered categorical regression: storable votes 331
15.3 Non-nested negative-binomial model of structure in social networks 332
15.4 Bibliographic note 342
15.5 Exercises 342

Part 2B: Fitting multilevel models 343

16 Multilevel modeling in Bugs and R: the basics 345
16.1 Why you should learn Bugs 345
16.2 Bayesian inference and prior distributions 345
16.3 Fitting and understanding a varying-intercept multilevel model using R and Bugs 348
16.4 Step by step through a Bugs model, as called from R 353
16.5 Adding individual- and group-level predictors 359
16.6 Predictions for new observations and new groups 361
16.7 Fake-data simulation 363
16.8 The principles of modeling in Bugs 366
16.9 Practical issues of implementation 369
16.10 Open-ended modeling in Bugs 370
16.11 Bibliographic note 373
16.12 Exercises 373

17 Fitting multilevel linear and generalized linear models in Bugs and R 375
17.1 Varying-intercept, varying-slope models 375
17.2 Varying intercepts and slopes with group-level predictors 379
17.3 Non-nested models 380
17.4 Multilevel logistic regression 381
17.5 Multilevel Poisson regression 382
17.6 Multilevel ordered categorical regression 383
17.7 Latent-data parameterizations of generalized linear models 384

17.8 Bibliographic note 385
17.9 Exercises 385

18 Likelihood and Bayesian inference and computation 387
18.1 Least squares and maximum likelihood estimation 387
18.2 Uncertainty estimates using the likelihood surface 390
18.3 Bayesian inference for classical and multilevel regression 392
18.4 Gibbs sampler for multilevel linear models 397
18.5 Likelihood inference, Bayesian inference, and the Gibbs sampler: the case of censored data 402
18.6 Metropolis algorithm for more general Bayesian computation 408
18.7 Specifying a log posterior density, Gibbs sampler, and Metropolis algorithm in R 409
18.8 Bibliographic note 413
18.9 Exercises 413

19 Debugging and speeding convergence 415
19.1 Debugging and confidence building 415
19.2 General methods for reducing computational requirements 418
19.3 Simple linear transformations 419
19.4 Redundant parameters and intentionally nonidentifiable models 419
19.5 Parameter expansion: multiplicative redundant parameters 424
19.6 Using redundant parameters to create an informative prior distribution for multilevel variance parameters 427
19.7 Bibliographic note 434
19.8 Exercises 434

Part 3: From data collection to model understanding to model checking 435

20 Sample size and power calculations 437
20.1 Choices in the design of data collection 437
20.2 Classical power calculations: general principles, as illustrated by estimates of proportions 439
20.3 Classical power calculations for continuous outcomes 443
20.4 Multilevel power calculation for cluster sampling 447
20.5 Multilevel power calculation using fake-data simulation 449
20.6 Bibliographic note 454
20.7 Exercises 454

21 Understanding and summarizing the fitted models 457
21.1 Uncertainty and variability 457
21.2 Superpopulation and finite-population variances 459
21.3 Contrasts and comparisons of multilevel coefficients 462
21.4 Average predictive comparisons 466
21.5 R^2 and explained variance 473
21.6 Summarizing the amount of partial pooling 477
21.7 Adding a predictor can *increase* the residual variance! 480
21.8 Multiple comparisons and statistical significance 481
21.9 Bibliographic note 484
21.10 Exercises 485

22 Analysis of variance **487**
22.1 Classical analysis of variance 487
22.2 ANOVA and multilevel linear and generalized linear models 490
22.3 Summarizing multilevel models using ANOVA 492
22.4 Doing ANOVA using multilevel models 494
22.5 Adding predictors: analysis of covariance and contrast analysis 496
22.6 Modeling the variance parameters: a split-plot latin square 498
22.7 Bibliographic note 501
22.8 Exercises 501

23 Causal inference using multilevel models **503**
23.1 Multilevel aspects of data collection 503
23.2 Estimating treatment effects in a multilevel observational study 506
23.3 Treatments applied at different levels 507
23.4 Instrumental variables and multilevel modeling 509
23.5 Bibliographic note 512
23.6 Exercises 512

24 Model checking and comparison **513**
24.1 Principles of predictive checking 513
24.2 Example: a behavioral learning experiment 515
24.3 Model comparison and deviance 524
24.4 Bibliographic note 526
24.5 Exercises 527

25 Missing-data imputation **529**
25.1 Missing-data mechanisms 530
25.2 Missing-data methods that discard data 531
25.3 Simple missing-data approaches that retain all the data 532
25.4 Random imputation of a single variable 533
25.5 Imputation of several missing variables 539
25.6 Model-based imputation 540
25.7 Combining inferences from multiple imputations 542
25.8 Bibliographic note 542
25.9 Exercises 543

Appendixes **545**

A Six quick tips to improve your regression modeling **547**
A.1 Fit many models 547
A.2 Do a little work to make your computations faster and more reliable 547
A.3 Graphing the relevant and not the irrelevant 548
A.4 Transformations 548
A.5 Consider all coefficients as potentially varying 549
A.6 Estimate causal inferences in a targeted way, not as a byproduct of a large regression 549

B Statistical graphics for research and presentation **551**
B.1 Reformulating a graph by focusing on comparisons 552
B.2 Scatterplots 553
B.3 Miscellaneous tips 559

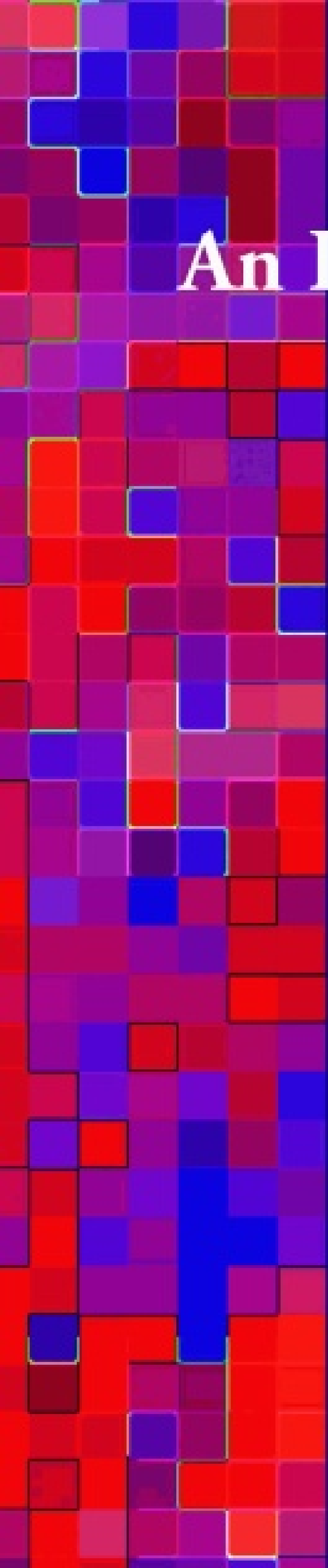
B.4 Bibliographic note 562
B.5 Exercises 563

C Software 565
C.1 Getting started with R, Bugs, and a text editor 565
C.2 Fitting classical and multilevel regressions in R 565
C.3 Fitting models in Bugs and R 567
C.4 Fitting multilevel models using R, Stata, SAS, and other software 568
C.5 Bibliographic note 573

References 575

Author index 601

Subject index 607



An Introduction to R

Second Edition

W. N. Venables, D. M. Smith and the
R Development Core Team

Copyright © 1990 W. N. Venables
Copyright © 1992 W. N. Venables & D. M. Smith
Copyright © 1997 R. Gentleman & R. Ihaka
Copyright © 1997, 1998 M. Maechler
Copyright © 1999–2006 R Development Core Team

Permission is granted to make and distribute verbatim copies of this manual provided the copyright notice and this permission notice are preserved on all copies.

Permission is granted to copy and distribute modified versions of this manual under the conditions for verbatim copying, provided that the entire resulting derived work is distributed under the terms of a permission notice identical to this one.

Permission is granted to copy and distribute translations of this manual into another language, under the above conditions for modified versions, except that this permission notice may be stated in a translation approved by the R Development Core Team.

ISBN 3-900051-12-7

Table of Contents

Preface 1

1 Introduction and preliminaries 2
- 1.1 The R environment 2
- 1.2 Related software and documentation 2
- 1.3 R and statistics 2
- 1.4 R and the window system 3
- 1.5 Using R interactively 3
- 1.6 An introductory session 4
- 1.7 Getting help with functions and features 4
- 1.8 R commands, case sensitivity, etc. 4
- 1.9 Recall and correction of previous commands 5
- 1.10 Executing commands from or diverting output to a file 5
- 1.11 Data permanency and removing objects 5

2 Simple manipulations; numbers and vectors 7
- 2.1 Vectors and assignment 7
- 2.2 Vector arithmetic 7
- 2.3 Generating regular sequences 8
- 2.4 Logical vectors 9
- 2.5 Missing values 9
- 2.6 Character vectors 10
- 2.7 Index vectors; selecting and modifying subsets of a data set 10
- 2.8 Other types of objects 11

3 Objects, their modes and attributes 13
- 3.1 Intrinsic attributes: mode and length 13
- 3.2 Changing the length of an object 14
- 3.3 Getting and setting attributes 14
- 3.4 The class of an object 14

4 Ordered and unordered factors 16
- 4.1 A specific example 16
- 4.2 The function `tapply()` and ragged arrays 16
- 4.3 Ordered factors 17

5 Arrays and matrices 18
- 5.1 Arrays 18
- 5.2 Array indexing. Subsections of an array 18
- 5.3 Index matrices 19
- 5.4 The `array()` function 20
  - 5.4.1 Mixed vector and array arithmetic. The recycling rule 20
- 5.5 The outer product of two arrays 21
- 5.6 Generalized transpose of an array 21
- 5.7 Matrix facilities 22
  - 5.7.1 Matrix multiplication 22

5.7.2 Linear equations and inversion 22
5.7.3 Eigenvalues and eigenvectors 23
5.7.4 Singular value decomposition and determinants 23
5.7.5 Least squares fitting and the QR decomposition 23
5.8 Forming partitioned matrices, `cbind()` and `rbind()` 24
5.9 The concatenation function, `c()`, with arrays 24
5.10 Frequency tables from factors 25

6 Lists and data frames 26
6.1 Lists 26
6.2 Constructing and modifying lists 26
6.2.1 Concatenating lists 27
6.3 Data frames 27
6.3.1 Making data frames 27
6.3.2 `attach()` and `detach()` 27
6.3.3 Working with data frames 28
6.3.4 Attaching arbitrary lists 28
6.3.5 Managing the search path 29

7 Reading data from files 30
7.1 The `read.table()` function 30
7.2 The `scan()` function 31
7.3 Accessing builtin datasets 31
7.3.1 Loading data from other R packages 31
7.4 Editing data 32

8 Probability distributions 33
8.1 R as a set of statistical tables 33
8.2 Examining the distribution of a set of data 33
8.3 One- and two-sample tests 36

9 Grouping, loops and conditional execution 40
9.1 Grouped expressions 40
9.2 Control statements 40
9.2.1 Conditional execution: `if` statements 40
9.2.2 Repetitive execution: `for` loops, `repeat` and `while` 40

10 Writing your own functions 42
10.1 Simple examples 42
10.2 Defining new binary operators 43
10.3 Named arguments and defaults 43
10.4 The '`...`' argument 44
10.5 Assignments within functions 44
10.6 More advanced examples 44
10.6.1 Efficiency factors in block designs 44
10.6.2 Dropping all names in a printed array 45
10.6.3 Recursive numerical integration 45
10.7 Scope 46
10.8 Customizing the environment 48
10.9 Classes, generic functions and object orientation 48

11 Statistical models in R ... 50
- 11.1 Defining statistical models; formulae ... 50
  - 11.1.1 Contrasts ... 52
- 11.2 Linear models ... 53
- 11.3 Generic functions for extracting model information ... 53
- 11.4 Analysis of variance and model comparison ... 54
  - 11.4.1 ANOVA tables ... 54
- 11.5 Updating fitted models ... 54
- 11.6 Generalized linear models ... 55
  - 11.6.1 Families ... 56
  - 11.6.2 The `glm()` function ... 56
- 11.7 Nonlinear least squares and maximum likelihood models ... 58
  - 11.7.1 Least squares ... 58
  - 11.7.2 Maximum likelihood ... 59
- 11.8 Some non-standard models ... 60

12 Graphical procedures ... 62
- 12.1 High-level plotting commands ... 62
  - 12.1.1 The `plot()` function ... 62
  - 12.1.2 Displaying multivariate data ... 63
  - 12.1.3 Display graphics ... 63
  - 12.1.4 Arguments to high-level plotting functions ... 64
- 12.2 Low-level plotting commands ... 65
  - 12.2.1 Mathematical annotation ... 66
  - 12.2.2 Hershey vector fonts ... 66
- 12.3 Interacting with graphics ... 66
- 12.4 Using graphics parameters ... 67
  - 12.4.1 Permanent changes: The `par()` function ... 67
  - 12.4.2 Temporary changes: Arguments to graphics functions ... 68
- 12.5 Graphics parameters list ... 68
  - 12.5.1 Graphical elements ... 69
  - 12.5.2 Axes and tick marks ... 70
  - 12.5.3 Figure margins ... 70
  - 12.5.4 Multiple figure environment ... 72
- 12.6 Device drivers ... 73
  - 12.6.1 PostScript diagrams for typeset documents ... 73
  - 12.6.2 Multiple graphics devices ... 74
- 12.7 Dynamic graphics ... 75

13 Packages ... 76
- 13.1 Standard packages ... 76
- 13.2 Contributed packages and CRAN ... 76
- 13.3 Namespaces ... 76

Appendix A A sample session ... 78

Appendix B Invoking R ... 81
- B.1 Invoking R from the command line ... 81
- B.2 Invoking R under Windows ... 84
- B.3 Invoking R under Mac OS X ... 85
- B.4 Scripting with R ... 85

Appendix C The command-line editor 87

C.1 Preliminaries 87

C.2 Editing actions 87

C.3 Command-line editor summary 87

Appendix D Function and variable index 89

Appendix E Concept index 92

Appendix F References 94

Use R!

Phil Spector

Data Manipulation with R

Springer

Contents

Preface V

1 Data in R 1
- 1.1 Modes and Classes 1
- 1.2 Data Storage in R 2
- 1.3 Testing for Modes and Classes 7
- 1.4 Structure of R Objects 7
- 1.5 Conversion of Objects 8
- 1.6 Missing Values 10
- 1.7 Working with Missing Values 10

2 Reading and Writing Data 13
- 2.1 Reading Vectors and Matrices 13
- 2.2 Data Frames: `read.table` 15
- 2.3 Comma- and Tab-Delimited Input Files 17
- 2.4 Fixed-Width Input Files 17
- 2.5 Extracting Data from R Objects 18
- 2.6 Connections 23
- 2.7 Reading Large Data Files 25
- 2.8 Generating Data 27
  - 2.8.1 Sequences 27
  - 2.8.2 Random Numbers 29
- 2.9 Permutations 30
  - 2.9.1 Random Permutations 30
  - 2.9.2 Enumerating All Permutations 30
- 2.10 Working with Sequences 31
- 2.11 Spreadsheets 33
  - 2.11.1 The `RODBC` Package on Windows 33
  - 2.11.2 The `gdata` Package (All Platforms) 34
- 2.12 Saving and Loading R Data Objects 35
- 2.13 Working with Binary Files 36

2.14 Writing R Objects to Files in ASCII Format 38
2.14.1 The `write` Function 38
2.14.2 The `write.table` function 39
2.15 Reading Data from Other Programs 39

3 R and Databases 43
3.1 A Brief Guide to SQL 43
3.1.1 Navigation Commands 43
3.1.2 Basics of SQL 44
3.1.3 Aggregation 45
3.1.4 Joining Two Databases 46
3.1.5 Subqueries 47
3.1.6 Modifying Database Records 48
3.2 `ODBC` 49
3.3 Using the `RODBC` Package 50
3.4 The `DBI` Package 51
3.5 Accessing a MySQL Database 51
3.6 Performing Queries 52
3.7 Normalized Tables 52
3.8 Getting Data into MySQL 53
3.9 More Complex Aggregations 55

4 Dates 57
4.1 `as.Date` 57
4.2 The `chron` Package 59
4.3 POSIX Classes 60
4.4 Working with Dates 63
4.5 Time Intervals 64
4.6 Time Sequences 65

5 Factors 67
5.1 Using Factors 67
5.2 Numeric Factors 70
5.3 Manipulating Factors 70
5.4 Creating Factors from Continuous Variables 72
5.5 Factors Based on Dates and Times 73
5.6 Interactions 74

6 Subscripting 75
6.1 Basics of Subscripting 75
6.2 Numeric Subscripts 75
6.3 Character Subscripts 75
6.4 Logical Subscripts 76
6.5 Subscripting Matrices and Arrays 77
6.6 Specialized Functions for Matrices 81

6.7 Lists 82
6.8 Subscripting Data Frames 83

7 Character Manipulation 87
7.1 Basics of Character Data 87
7.2 Displaying and Concatenating Character Strings 87
7.3 Working with Parts of Character Values 89
7.4 Regular Expressions in R 90
7.5 Basics of Regular Expressions 91
7.6 Breaking Apart Character Values 93
7.7 Using Regular Expressions in R 94
7.8 Substitutions and Tagging 98

8 Data Aggregation 101
8.1 `table` 101
8.2 Road Map for Aggregation 106
8.3 Mapping a Function to a Vector or List 107
8.4 Mapping a function to a matrix or array 110
8.5 Mapping a Function Based on Groups 113
8.6 The `reshape` Package 120
8.7 Loops in R 126

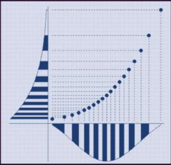
9 Reshaping Data 131
9.1 Modifying Data Frame Variables 131
9.2 Recoding Variables 132
9.3 The `recode` Function 134
9.4 Reshaping Data Frames 135
9.5 The `reshape` Package 140
9.6 Combining Data Frames 142
9.7 Under the Hood of `merge` 146

Index 149

The R Series

Using R for Introductory Statistics

Second Edition



John Verzani

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Preface

These notes are an introduction to using the statistical software package R for an introductory statistics course. They are meant to accompany an introductory statistics book such as Kitchens *"Exploring Statistics"*. The goals are not to show all the features of R, or to replace a standard textbook, but rather to be used with a textbook to illustrate the features of R that can be learned in a one-semester, introductory statistics course.

These notes were written to take advantage of R version 1.5.0 or later. For pedagogical reasons the equals sign, `=`, is used as an assignment operator and not the traditional arrow combination `<-`. This was added to R in version 1.4.0. If only an older version is available the reader will have to make the minor adjustment.

There are several references to data and functions in this text that need to be installed prior to their use. To install the data is easy, but the instructions vary depending on your system. For Windows users, you need to download the "zip" file , and then install from the "packages" menu. In UNIX, one uses the command `R CMD INSTALL packagename.tar.gz`. Some of the datasets are borrowed from other authors notably Kitchens. Credit is given in the help files for the datasets. This material is available as an R package from:

http://www.math.csi.cuny.edu/Statistics/R/simpleR/Simple_0.4.zip for Windows users.
http://www.math.csi.cuny.edu/Statistics/R/simpleR/Simple_0.4.tar.gz for UNIX users.

If necessary, the file can sent in an email. As well, the individual data sets can be found online in the directory

http://www.math.csi.cuny.edu/Statistics/R/simpleR/Simple.

This is version 0.4 of these notes and were last generated on August 22, 2002. Before printing these notes, you should check for the most recent version available from

the CSI Math department (http://www.math.csi.cuny.edu/Statistics/R/simpleR).

Copyright © John Verzani (verzani@math.csi.cuny.edu), 2001-2. All rights reserved.

Contents

Introduction **1**
What is R 1
A note on notation 1

Data **1**
Starting R 1
Entering data with c 2
Data is a vector 3
Problems 7

Univariate Data **8**
Categorical data 8
Numerical data 10
Problems 18

Bivariate Data **19**
Handling bivariate categorical data 20
Handling bivariate data: categorical vs. numerical 21
Bivariate data: numerical vs. numerical 22
Linear regression. 24
Problems 31

Multivariate Data **32**
Storing multivariate data in data frames 32
Accessing data in data frames 33
Manipulating data frames: stack and unstack 34
Using R's model formula notation 35
Ways to view multivariate data 35
The lattice package 40
Problems 40

Random Data **41**
Random number generators in R– the "r" functions. 41
Problems 46

Simulations **47**
The central limit theorem 47
Using simple.sim and functions 49
Problems 51

Exploratory Data Analysis **54**
Our toolbox 54
Examples 54
Problems 58

Confidence Interval Estimation **59**
Population proportion theory 59
Proportion test 61
The z-test 62
The t-test 62
Confidence interval for the median 64
Problems 65

Hypothesis Testing **66**
Testing a population parameter 66
Testing a mean 67
Tests for the median 67
Problems 68

Two-sample tests **68**
Two-sample tests of proportion 68
Two-sample t-tests 69
Resistant two-sample tests 71
Problems 71

Chi Square Tests **72**
The chi-squared distribution 72
Chi-squared goodness of fit tests 72
Chi-squared tests of independence 74
Chi-squared tests for homogeneity 75
Problems 76

Regression Analysis **77**
Simple linear regression model 77
Testing the assumptions of the model 78
Statistical inference 79
Problems 83

Multiple Linear Regression **84**
The model 84
Problems 89

Analysis of Variance **89**
one-way analysis of variance 89
Problems 92

Appendix: Installing R **94**

Appendix: External Packages **94**

Appendix: A sample R session **94**
A sample session involving regression 94
t-tests 97
A simulation example 99

Appendix: What happens when R starts? 100

Appendix: Using Functions 100

The basic template . . . 100
For loops . . . 102
Conditional expressions . . . 103

Appendix: Entering Data into R 103

Using c . . . 104
using scan . . . 104
Using scan with a file . . . 104
Editing your data . . . 104
Reading in tables of data . . . 105
Fixed-width fields . . . 105
Spreadsheet data . . . 105
XML, urls . . . 106
"Foreign" formats . . . 106

Appendix: Teaching Tricks 106

Appendix: Sources of help, documentation 107

Statistics and Computing

Peter Dalgaard

Introductory Statistics with R

Second Edition

Springer

Contents

Preface **vii**

1 Basics **1**

1.1 First steps 1

1.1.1 An overgrown calculator 3

1.1.2 Assignments 3

1.1.3 Vectorized arithmetic 4

1.1.4 Standard procedures 6

1.1.5 Graphics 7

1.2 R language essentials 9

1.2.1 Expressions and objects 9

1.2.2 Functions and arguments 11

1.2.3 Vectors 12

1.2.4 Quoting and escape sequences 13

1.2.5 Missing values 14

1.2.6 Functions that create vectors 14

1.2.7 Matrices and arrays 16

1.2.8 Factors 18

1.2.9 Lists 19

1.2.10 Data frames 20

1.2.11 Indexing 21

1.2.12 Conditional selection 22

1.2.13 Indexing of data frames 23

1.2.14 Grouped data and data frames 25

1.2.15 Implicit loops 26
1.2.16 Sorting 27
1.3 Exercises 28

2 The R environment 31
2.1 Session management 31
2.1.1 The workspace 31
2.1.2 Textual output 32
2.1.3 Scripting 33
2.1.4 Getting help 34
2.1.5 Packages 35
2.1.6 Built-in data 35
2.1.7 `attach` and `detach` 36
2.1.8 `subset`, `transform`, and `within` 37
2.2 The graphics subsystem 39
2.2.1 Plot layout 39
2.2.2 Building a plot from pieces 40
2.2.3 Using `par` 42
2.2.4 Combining plots 42
2.3 R programming 44
2.3.1 Flow control 44
2.3.2 Classes and generic functions 46
2.4 Data entry 46
2.4.1 Reading from a text file 47
2.4.2 Further details on `read.table` 50
2.4.3 The data editor 51
2.4.4 Interfacing to other programs 52
2.5 Exercises 53

3 Probability and distributions 55
3.1 Random sampling 55
3.2 Probability calculations and combinatorics 56
3.3 Discrete distributions 57
3.4 Continuous distributions 58
3.5 The built-in distributions in R 59
3.5.1 Densities 59
3.5.2 Cumulative distribution functions 62
3.5.3 Quantiles 63
3.5.4 Random numbers 64
3.6 Exercises 65

4 Descriptive statistics and graphics 67
4.1 Summary statistics for a single group 67
4.2 Graphical display of distributions 71
4.2.1 Histograms 71

4.2.2 Empirical cumulative distribution 73
4.2.3 Q–Q plots 74
4.2.4 Boxplots 75
4.3 Summary statistics by groups 75
4.4 Graphics for grouped data 79
4.4.1 Histograms 79
4.4.2 Parallel boxplots 80
4.4.3 Stripcharts 81
4.5 Tables 83
4.5.1 Generating tables 83
4.5.2 Marginal tables and relative frequency 87
4.6 Graphical display of tables 89
4.6.1 Barplots 89
4.6.2 Dotcharts 91
4.6.3 Piecharts 92
4.7 Exercises 93

5 One- and two-sample tests 95
5.1 One-sample *t* test 95
5.2 Wilcoxon signed-rank test 99
5.3 Two-sample *t* test 100
5.4 Comparison of variances 103
5.5 Two-sample Wilcoxon test 103
5.6 The paired *t* test 104
5.7 The matched-pairs Wilcoxon test 106
5.8 Exercises 107

6 Regression and correlation 109
6.1 Simple linear regression 109
6.2 Residuals and fitted values 113
6.3 Prediction and confidence bands 117
6.4 Correlation 120
6.4.1 Pearson correlation 121
6.4.2 Spearman's ρ 123
6.4.3 Kendall's τ 124
6.5 Exercises 124

7 Analysis of variance and the Kruskal–Wallis test 127
7.1 One-way analysis of variance 127
7.1.1 Pairwise comparisons and multiple testing . . . 131
7.1.2 Relaxing the variance assumption 133
7.1.3 Graphical presentation 134
7.1.4 Bartlett's test 136
7.2 Kruskal–Wallis test 136
7.3 Two-way analysis of variance 137

7.3.1 Graphics for repeated measurements 140
7.4 The Friedman test 141
7.5 The ANOVA table in regression analysis 141
7.6 Exercises 143

8 Tabular data **145**
8.1 Single proportions 145
8.2 Two independent proportions 147
8.3 *k* proportions, test for trend 149
8.4 $r \times c$ tables 151
8.5 Exercises 153

9 Power and the computation of sample size **155**
9.1 The principles of power calculations 155
9.1.1 Power of one-sample and paired *t* tests 156
9.1.2 Power of two-sample *t* test 158
9.1.3 Approximate methods 158
9.1.4 Power of comparisons of proportions 159
9.2 Two-sample problems 159
9.3 One-sample problems and paired tests 161
9.4 Comparison of proportions 161
9.5 Exercises 162

10 Advanced data handling **163**
10.1 Recoding variables 163
10.1.1 The `cut` function 163
10.1.2 Manipulating factor levels 165
10.1.3 Working with dates 166
10.1.4 Recoding multiple variables 169
10.2 Conditional calculations 170
10.3 Combining and restructuring data frames 171
10.3.1 Appending frames 172
10.3.2 Merging data frames 173
10.3.3 Reshaping data frames 175
10.4 Per-group and per-case procedures 178
10.5 Time splitting 179
10.6 Exercises 183

11 Multiple regression **185**
11.1 Plotting multivariate data 185
11.2 Model specification and output 187
11.3 Model search 190
11.4 Exercises 193

12 Linear models **195**
12.1 Polynomial regression 196
12.2 Regression through the origin 198
12.3 Design matrices and dummy variables 200
12.4 Linearity over groups 202
12.5 Interactions 206
12.6 Two-way ANOVA with replication 207
12.7 Analysis of covariance 208
12.7.1 Graphical description 209
12.7.2 Comparison of regression lines 212
12.8 Diagnostics 218
12.9 Exercises 224

13 Logistic regression **227**
13.1 Generalized linear models 228
13.2 Logistic regression on tabular data 229
13.2.1 The analysis of deviance table 234
13.2.2 Connection to test for trend 235
13.3 Likelihood profiling 237
13.4 Presentation as odds-ratio estimates 239
13.5 Logistic regression using raw data 239
13.6 Prediction 241
13.7 Model checking 242
13.8 Exercises 247

14 Survival analysis **249**
14.1 Essential concepts 249
14.2 Survival objects 250
14.3 Kaplan–Meier estimates 251
14.4 The log-rank test 254
14.5 The Cox proportional hazards model 256
14.6 Exercises 258

15 Rates and Poisson regression **259**
15.1 Basic ideas 259
15.1.1 The Poisson distribution 260
15.1.2 Survival analysis with constant hazard 260
15.2 Fitting Poisson models 262
15.3 Computing rates 266
15.4 Models with piecewise constant intensities 270
15.5 Exercises 274

16 Nonlinear curve fitting **275**
16.1 Basic usage 276
16.2 Finding starting values 278

16.3 Self-starting models 284
16.4 Profiling 285
16.5 Finer control of the fitting algorithm 287
16.6 Exercises 288

A Obtaining and installing R and the `ISwR` package 289

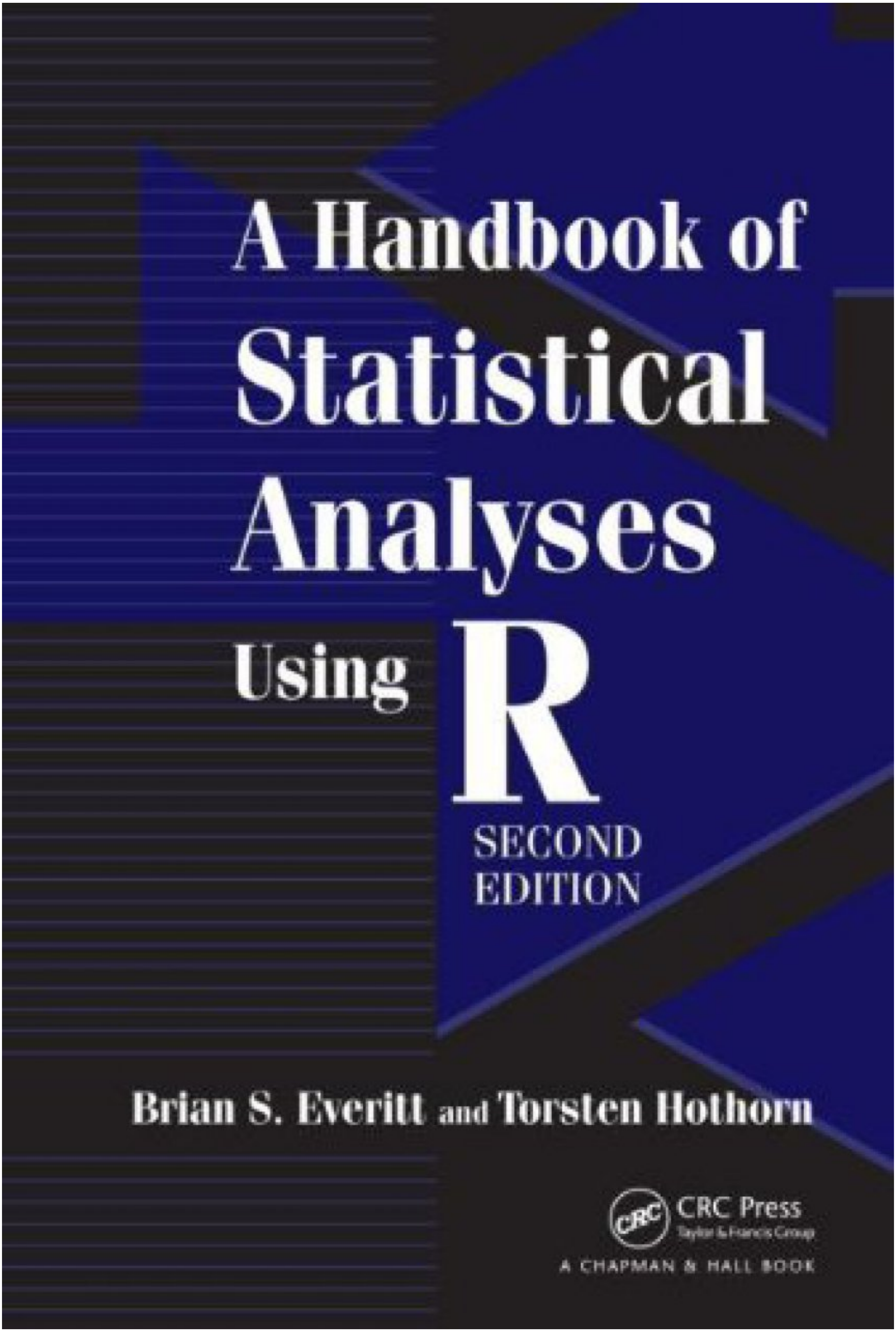
B Data sets in the `ISwR` package 293

C Compendium 325

D Answers to exercises 337

Bibliography 355

Index 357



A Handbook of Statistical Analyses Using R

SECOND EDITION

Brian S. Everitt and Torsten Hothorn

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Contents

1 An Introduction to R 1

1.1 What Is R? 1

1.2 Installing R 2

1.3 Help and Documentation 4

1.4 Data Objects in R 5

1.5 Data Import and Export 9

1.6 Basic Data Manipulation 11

1.7 Simple Summary Statistics 14

1.8 Organising an Analysis 18

1.9 Summary 20

2 Simple Inference 21

2.1 Introduction 21

2.2 Statistical Tests 25

2.3 Analysis Using R 29

2.4 Summary 39

3 Conditional Inference 41

3.1 Introduction 41

3.2 Conditional Test Procedures 44

3.3 Analysis Using R 46

3.4 Summary 53

4 Analysis of Variance 55

4.1 Introduction 55

4.2 Analysis of Variance 58

4.3 Analysis Using R 59

4.4 Summary 71

5 Multiple Linear Regression 73

5.1 Introduction 73

5.2 Multiple Linear Regression 74

5.3 Analysis Using R 76

5.4 Summary 86

© 2006 by Taylor & Francis Group, LLC

6 Logistic Regression and Generalised Linear Models 89
6.1 Introduction 89
6.2 Logistic Regression and Generalised Linear Models 92
6.3 Analysis Using R 94
6.4 Summary 106

7 Density Estimation 109
7.1 Introduction 109
7.2 Density Estimation 111
7.3 Analysis Using R 117
7.4 Summary 125

8 Recursive Partitioning 131
8.1 Introduction 131
8.2 Recursive Partitioning 131
8.3 Analysis Using R 133
8.4 Summary 141

9 Survival Analysis 143
9.1 Introduction 143
9.2 Survival Analysis 144
9.3 Analysis Using R 150
9.4 Summary 157

10 Analysing Longitudinal Data I 159
10.1 Introduction 159
10.2 Analysing Longitudinal Data 162
10.3 Linear Mixed Effects Models 163
10.4 Analysis Using R 165
10.5 Prediction of Random Effects 168
10.6 The Problem of Dropouts 169
10.7 Summary 172

11 Analysing Longitudinal Data II 175
11.1 Introduction 175
11.2 Generalised Estimating Equations 177
11.3 Analysis Using R 179
11.4 Summary 194

12 Meta-Analysis 197
12.1 Introduction 197
12.2 Systematic Reviews and Meta-Analysis 199
12.3 Statistics of Meta-Analysis 201
12.4 Analysis Using R 202
12.5 Meta-Regression 203

© 2006 by Taylor & Francis Group, LLC

12.6 Publication Bias 207
12.7 Summary 211

13 Principal Component Analysis 215
13.1 Introduction 215
13.2 Principal Component Analysis 215
13.3 Analysis Using R 218
13.4 Summary 223

14 Multidimensional Scaling 227
14.1 Introduction 227
14.2 Multidimensional Scaling 227
14.3 Analysis Using R 233
14.4 Summary 239

15 Cluster Analysis 243
15.1 Introduction 243
15.2 Cluster Analysis 245
15.3 Analysis Using R 248
15.4 Summary 253

Bibliography 259

© 2006 by Taylor & Francis Group, LLC

Practical Recipes for Visualizing Data

R Graphics Cookbook

O'REILLY®

Winston Chang

www.it-ebooks.info

Table of Contents

Preface . . . . **ix**

1. R Basics **1**

1.1. Installing a Package 1
1.2. Loading a Package 2
1.3. Loading a Delimited Text Data File 3
1.4. Loading Data from an Excel File 4
1.5. Loading Data from an SPSS File 5

2. Quickly Exploring Data **7**

2.1. Creating a Scatter Plot 7
2.2. Creating a Line Graph 9
2.3. Creating a Bar Graph 11
2.4. Creating a Histogram 13
2.5. Creating a Box Plot 15
2.6. Plotting a Function Curve 17

3. Bar Graphs **19**

3.1. Making a Basic Bar Graph 19
3.2. Grouping Bars Together 22
3.3. Making a Bar Graph of Counts 25
3.4. Using Colors in a Bar Graph 27
3.5. Coloring Negative and Positive Bars Differently 29
3.6. Adjusting Bar Width and Spacing 30
3.7. Making a Stacked Bar Graph 32
3.8. Making a Proportional Stacked Bar Graph 35
3.9. Adding Labels to a Bar Graph 38
3.10. Making a Cleveland Dot Plot 42

4. Line Graphs **49**

www.it-ebooks.info

4.1. Making a Basic Line Graph 49
4.2. Adding Points to a Line Graph 52
4.3. Making a Line Graph with Multiple Lines 53
4.4. Changing the Appearance of Lines 58
4.5. Changing the Appearance of Points 59
4.6. Making a Graph with a Shaded Area 62
4.7. Making a Stacked Area Graph 64
4.8. Making a Proportional Stacked Area Graph 67
4.9. Adding a Confidence Region 69

5. Scatter Plots. 73
5.1. Making a Basic Scatter Plot 73
5.2. Grouping Data Points by a Variable Using Shape or Color 75
5.3. Using Different Point Shapes 77
5.4. Mapping a Continuous Variable to Color or Size 80
5.5. Dealing with Overplotting 84
5.6. Adding Fitted Regression Model Lines 89
5.7. Adding Fitted Lines from an Existing Model 94
5.8. Adding Fitted Lines from Multiple Existing Models 97
5.9. Adding Annotations with Model Coefficients 100
5.10. Adding Marginal Rugs to a Scatter Plot 103
5.11. Labeling Points in a Scatter Plot 104
5.12. Creating a Balloon Plot 110
5.13. Making a Scatter Plot Matrix 112

6. Summarized Data Distributions. 117
6.1. Making a Basic Histogram 117
6.2. Making Multiple Histograms from Grouped Data 120
6.3. Making a Density Curve 123
6.4. Making Multiple Density Curves from Grouped Data 126
6.5. Making a Frequency Polygon 129
6.6. Making a Basic Box Plot 130
6.7. Adding Notches to a Box Plot 133
6.8. Adding Means to a Box Plot 134
6.9. Making a Violin Plot 135
6.10. Making a Dot Plot 139
6.11. Making Multiple Dot Plots for Grouped Data 141
6.12. Making a Density Plot of Two-Dimensional Data 143

7. Annotations. 147
7.1. Adding Text Annotations 147
7.2. Using Mathematical Expressions in Annotations 150

www.it-ebooks.info

7.3. Adding Lines 152
7.4. Adding Line Segments and Arrows 155
7.5. Adding a Shaded Rectangle 156
7.6. Highlighting an Item 157
7.7. Adding Error Bars 159
7.8. Adding Annotations to Individual Facets 162

8. **Axes. 167**
8.1. Swapping X- and Y-Axes 167
8.2. Setting the Range of a Continuous Axis 168
8.3. Reversing a Continuous Axis 170
8.4. Changing the Order of Items on a Categorical Axis 172
8.5. Setting the Scaling Ratio of the X- and Y-Axes 174
8.6. Setting the Positions of Tick Marks 177
8.7. Removing Tick Marks and Labels 178
8.8. Changing the Text of Tick Labels 180
8.9. Changing the Appearance of Tick Labels 182
8.10. Changing the Text of Axis Labels 184
8.11. Removing Axis Labels 185
8.12. Changing the Appearance of Axis Labels 187
8.13. Showing Lines Along the Axes 189
8.14. Using a Logarithmic Axis 190
8.15. Adding Ticks for a Logarithmic Axis 196
8.16. Making a Circular Graph 198
8.17. Using Dates on an Axis 204
8.18. Using Relative Times on an Axis 207

9. **Controlling the Overall Appearance of Graphs. 211**
9.1. Setting the Title of a Graph 211
9.2. Changing the Appearance of Text 213
9.3. Using Themes 216
9.4. Changing the Appearance of Theme Elements 218
9.5. Creating Your Own Themes 221
9.6. Hiding Grid Lines 222

10. **Legends. 225**
10.1. Removing the Legend 225
10.2. Changing the Position of a Legend 227
10.3. Changing the Order of Items in a Legend 229
10.4. Reversing the Order of Items in a Legend 231
10.5. Changing a Legend Title 232
10.6. Changing the Appearance of a Legend Title 235

www.it-ebooks.info

10.7. Removing a Legend Title 236
10.8. Changing the Labels in a Legend 237
10.9. Changing the Appearance of Legend Labels 239
10.10. Using Labels with Multiple Lines of Text 240

11. Facets. 243
11.1. Splitting Data into Subplots with Facets 243
11.2. Using Facets with Different Axes 246
11.3. Changing the Text of Facet Labels 246
11.4. Changing the Appearance of Facet Labels and Headers 250

12. Using Colors in Plots. 251
12.1. Setting the Colors of Objects 251
12.2. Mapping Variables to Colors 252
12.3. Using a Different Palette for a Discrete Variable 254
12.4. Using a Manually Defined Palette for a Discrete Variable 259
12.5. Using a Colorblind-Friendly Palette 261
12.6. Using a Manually Defined Palette for a Continuous Variable 263
12.7. Coloring a Shaded Region Based on Value 264

13. Miscellaneous Graphs. 267
13.1. Making a Correlation Matrix 267
13.2. Plotting a Function 271
13.3. Shading a Subregion Under a Function Curve 272
13.4. Creating a Network Graph 274
13.5. Using Text Labels in a Network Graph 278
13.6. Creating a Heat Map 281
13.7. Creating a Three-Dimensional Scatter Plot 283
13.8. Adding a Prediction Surface to a Three-Dimensional Plot 285
13.9. Saving a Three-Dimensional Plot 289
13.10. Animating a Three-Dimensional Plot 291
13.11. Creating a Dendrogram 291
13.12. Creating a Vector Field 294
13.13. Creating a QQ Plot 299
13.14. Creating a Graph of an Empirical Cumulative Distribution Function 301
13.15. Creating a Mosaic Plot 302
13.16. Creating a Pie Chart 307
13.17. Creating a Map 309
13.18. Creating a Choropleth Map 313
13.19. Making a Map with a Clean Background 317

www.it-ebooks.info

13.20. Creating a Map from a Shapefile 319

14. Output for Presentation. 323
14.1. Outputting to PDF Vector Files 323
14.2. Outputting to SVG Vector Files 325
14.3. Outputting to WMF Vector Files 325
14.4. Editing a Vector Output File 326
14.5. Outputting to Bitmap (PNG/TIFF) Files 327
14.6. Using Fonts in PDF Files 330
14.7. Using Fonts in Windows Bitmap or Screen Output 332

15. Getting Your Data into Shape. 335
15.1. Creating a Data Frame 336
15.2. Getting Information About a Data Structure 337
15.3. Adding a Column to a Data Frame 338
15.4. Deleting a Column from a Data Frame 338
15.5. Renaming Columns in a Data Frame 339
15.6. Reordering Columns in a Data Frame 340
15.7. Getting a Subset of a Data Frame 341
15.8. Changing the Order of Factor Levels 343
15.9. Changing the Order of Factor Levels Based on Data Values 344
15.10. Changing the Names of Factor Levels 345
15.11. Removing Unused Levels from a Factor 347
15.12. Changing the Names of Items in a Character Vector 348
15.13. Recoding a Categorical Variable to Another Categorical Variable 349
15.14. Recoding a Continuous Variable to a Categorical Variable 351
15.15. Transforming Variables 352
15.16. Transforming Variables by Group 354
15.17. Summarizing Data by Groups 357
15.18. Summarizing Data with Standard Errors and Confidence Intervals 361
15.19. Converting Data from Wide to Long 365
15.20. Converting Data from Long to Wide 368
15.21. Converting a Time Series Object to Times and Values 369

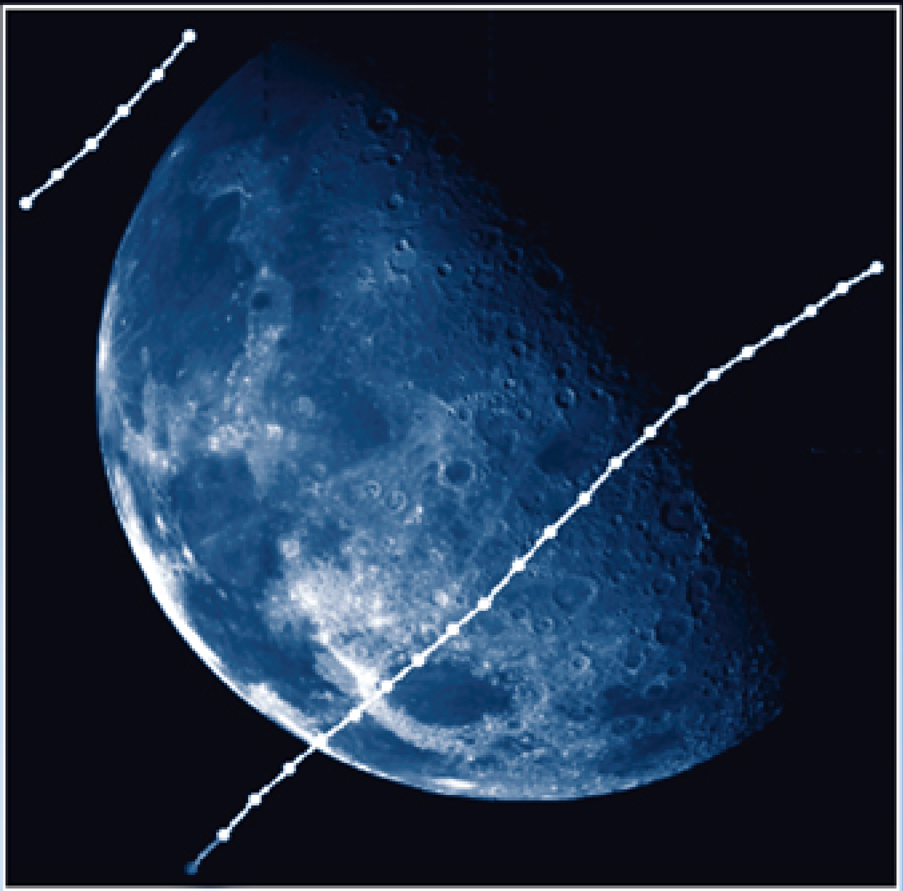
A. Introduction to ggplot2. 373

Index. 385

www.it-ebooks.info

Computer Science and Data Analysis Series

R Graphics



Paul Murrell

Contents

List of Figures

List of Tables

1 An Introduction to R Graphics
- 1.1 R graphics examples
  - 1.1.1 Standard plots
  - 1.1.2 Trellis plots
  - 1.1.3 Special-purpose plots
  - 1.1.4 General graphical scenes
- 1.2 The organization of R graphics
  - 1.2.1 Types of graphics functions
  - 1.2.2 Traditional graphics versus grid graphics
- 1.3 Graphical output formats
  - 1.3.1 Graphics devices
  - 1.3.2 Multiple pages of output
  - 1.3.3 Display lists

I TRADITIONAL GRAPHICS

2 Simple Usage of Traditional Graphics
- 2.1 The traditional graphics model
- 2.2 Plots of one or two variables
  - 2.2.1 Arguments to graphics functions
  - 2.2.2 Standard arguments
- 2.3 Plots of multiple variables
- 2.4 Modern plots and specialized plots
- 2.5 Interactive graphics

3 Customizing Traditional Graphics
- 3.1 The traditional graphics model in more detail
  - 3.1.1 Plotting regions
  - 3.1.2 The traditional graphics state
- 3.2 Controlling the appearance of plots
  - 3.2.1 Colors
  - 3.2.2 Lines

© 2006 by Taylor & Francis Group, LLC

3.2.3 Text
3.2.4 Data symbols
3.2.5 Axes
3.2.6 Plotting regions
3.2.7 Clipping
3.2.8 Moving to a new plot
3.3 Arranging multiple plots
3.3.1 Using the traditional graphics state
3.3.2 Layouts
3.3.3 The split-screen approach
3.4 Annotating plots
3.4.1 Annotating the plot region
3.4.2 Missing values and non-finite values
3.4.3 Annotating the margins
3.4.4 Legends
3.4.5 Axes
3.4.6 Mathematical formulae
3.4.7 Coordinate systems
3.4.8 Bitmap images
3.4.9 Special cases
3.5 Creating new plots
3.5.1 A simple plot from scratch
3.5.2 A more complex plot from scratch
3.5.3 Writing traditional graphics functions

II GRID GRAPHICS

4 Trellis Graphics: the Lattice Package
4.1 The lattice graphics model
4.1.1 Lattice devices
4.2 Lattice plot types
4.2.1 The `formula` argument and multipanel conditioning
4.2.2 A nontrivial example
4.3 Controlling the appearance of lattice plots
4.4 Arranging lattice plots
4.5 Annotating lattice plots
4.5.1 Panel functions and strip functions
4.5.2 Adding output to a lattice plot
4.6 Creating new lattice plots

5 The Grid Graphics Model
5.1 A brief overview of grid graphics
5.1.1 A simple example
5.2 Graphical primitives

© 2006 by Taylor & Francis Group, LLC

5.2.1 Standard arguments
5.3 Coordinate systems
5.3.1 Conversion functions
5.3.2 Complex units
5.4 Controlling the appearance of output
5.4.1 Specifying graphical parameter settings
5.4.2 Vectorized graphical parameter settings
5.5 Viewports
5.5.1 Pushing, popping, and navigating between viewports
5.5.2 Clipping to viewports
5.5.3 Viewport lists, stacks, and trees
5.5.4 Viewports as arguments to graphical primitives
5.5.5 Graphical parameter settings in viewports
5.5.6 Layouts
5.6 Missing values and non-finite values
5.7 Interactive graphics
5.8 Customizing lattice plots
5.8.1 Adding grid output to lattice output
5.8.2 Adding lattice output to grid output

6 The Grid Graphics Object Model
6.1 Working with graphical output
6.1.1 Standard functions and arguments
6.2 Grob lists, trees, and paths
6.2.1 Graphical parameter settings in gTrees
6.2.2 Viewports as components of gTrees
6.2.3 Searching for grobs
6.3 Working with graphical objects off-screen
6.3.1 Capturing output
6.4 Placing and packing grobs in frames
6.4.1 Placing and packing off-screen
6.5 Other details about grobs
6.5.1 Calculating the sizes of grobs
6.5.2 Editing graphical context
6.6 Saving and loading grid graphics
6.7 Working with lattice grobs

7 Developing New Graphics Functions and Objects
7.1 An example
7.1.1 Modularity
7.2 Simple graphics functions
7.2.1 Embedding graphical output
7.2.2 Facilitating annotation
7.2.3 Editing output
7.2.4 Absolute versus relative sizes

© 2006 by Taylor & Francis Group, LLC

7.3 Graphical objects
7.3.1 Overview of creating a new graphical class
7.3.2 Defining a new graphical class
7.3.3 Validating grobs
7.3.4 Drawing grobs
7.3.5 Editing grobs
7.3.6 Sizing grobs
7.3.7 Pre-drawing and post-drawing
7.3.8 Completing the example
7.3.9 Reusing graphical elements
7.3.10 Other details
7.4 Querying grid

A A Brief Introduction to R
A.1 Obtaining and installing R
A.2 An environment for statistical computing and graphics
A.2.1 Batch processing
A.2.2 Data types
A.2.3 Variables
A.2.4 Indexing
A.2.5 Data structures
A.2.6 Formulae
A.2.7 Expressions
A.2.8 Packages
A.2.9 Accessing data sets
A.2.10 Getting help
A.3 A programming language
A.3.1 Debugging
A.4 An object-oriented language

B Combining Traditional Graphics and Grid Graphics
B.1 The `gridBase` package
B.1.1 Annotating base graphics using grid
B.1.2 Embedding base graphics plots in grid viewports
B.1.3 Problems and limitations

Bibliography

© 2006 by Taylor & Francis Group, LLC

List of Figures

1.1 A simple scatterplot
1.2 Some standard plots
1.3 A customized scatterplot
1.4 A Trellis dotplot
1.5 A map of New Zealand produced using R
1.6 Some polar-coordinate plots
1.7 A novel decision tree plot
1.8 A table-like plot
1.9 Didactic diagrams
1.10 A music score
1.11 A piece of clip art
1.12 The structure of the R graphics system

2.1 Four variations on a scatterplot
2.2 Plotting an `lm` object
2.3 Plotting an `agnes` object
2.4 Modifying default `barplot()` and `boxplot()` output
2.5 Standard arguments for high-level functions
2.6 Plotting three variables
2.7 Plotting multivariate data
2.8 Some modern and specialized plots

3.1 The plot regions in traditional graphics
3.2 Multiple figure regions in traditional graphics
3.3 The user coordinate system in the plot region
3.4 Figure margin coordinate systems
3.5 Outer margin coordinate systems
3.6 Predefined and custom line types
3.7 Line join and line ending styles
3.8 Alignment of text in the plot region
3.9 Font families and font faces
3.10 Data symbols available in R
3.11 Basic plot types
3.12 Different axis styles
3.13 Graphics state settings controlling plot regions
3.14 Some basic layouts

© 2006 by Taylor & Francis Group, LLC

3.15 Some complex layouts
3.16 Annotating the plot region
3.17 More examples of annotating the plot region
3.18 Drawing polygons
3.19 Annotating the margins
3.20 Some simple legends
3.21 Customizing axes
3.22 Mathematical formulae in plots
3.23 Custom coordinate systems
3.24 Overlaying plots
3.25 Overlaying output
3.26 Adding a bitmap to a plot
3.27 Special-case annotations
3.28 A panel function example
3.29 Annotating a 3D surface
3.30 A back-to-back barplot
3.31 A graphics function template

4.1 A scatterplot using lattice
4.2 The result of modifying a lattice object
4.3 Plot types available in lattice
4.4 A lattice multipanel conditioning plot
4.5 A complex lattice plot
4.6 Some default lattice settings
4.7 Controlling the layout of lattice panels
4.8 Arranging multiple lattice plots
4.9 Annotating a lattice plot

5.1 A simple scatterplot using grid
5.2 Primitive grid output
5.3 Drawing arrows
5.4 Drawing polygons
5.5 A demonstration of grid units
5.6 Graphical parameters for graphical primitives
5.7 Recycling graphical parameters
5.8 Recycling graphical parameters for polygons
5.9 A diagram of a simple viewport
5.10 Pushing a viewport
5.11 Pushing several viewports
5.12 Popping a viewport
5.13 Navigating between viewports
5.14 Clipping output in viewports
5.15 The inheritance of viewport graphical parameters
5.16 Layouts and viewports
5.17 Layouts and units

© 2006 by Taylor & Francis Group, LLC

5.18 Nested layouts
5.19 Non-finite values for line-tos, polygons, and arrows
5.20 Controlling the size of lattice panels
5.21 Adding grid output to a lattice plot
5.22 Embedding a lattice plot within grid output

6.1 Modifying a `circle` grob
6.2 Editing grobs
6.3 The structure of a gTree
6.4 Editing a gTree
6.5 Using a gTree to group grobs
6.6 Packing grobs by hand
6.7 Calculating the size of a grob
6.8 Grob dimensions by reference
6.9 Editing the graphical context
6.10 Editing the grobs in a lattice plot

7.1 A plot of oceanographic data
7.2 A `grid.imageFun()` function
7.3 Output from the `grid.imageFun()` function
7.4 A `grid.ozFun()` function
7.5 Example output from `grid.ozFun()`
7.6 Annotating `grid.ozFun()` output
7.7 Editing `grid.ozFun()` output
7.8 An `"imageGrob"` class
7.9 Some `validDetails()` methods
7.10 An `"ozGrob"` class
7.11 An `"ozImage"` class
7.12 Some `editDetails()` methods
7.13 Editing an `imageGrob`
7.14 Low-level editing of an `imageGrob`
7.15 Helper functions for a `"ribbonLegend"` class
7.16 A `"ribbonLegend"` class
7.17 An `"ozKey"` class
7.18 A plot of temperature data
7.19 A `splitString()` function
7.20 Performing calculations before drawing
7.21 A `"splitText"` class
7.22 Drawing faces
7.23 Some face functions
7.24 Some face objects

B.1 Annotating a traditional plot with grid
B.2 Embedding a traditional plot within lattice output

© 2006 by Taylor & Francis Group, LLC

List of Tables

1.1 Graphical output formats

3.1 High-level traditional graphics state settings
3.2 Low-level traditional graphics state settings
3.3 Read-only traditional graphics state settings
3.4 Functions to generate color sets
3.5 Font faces
3.6 Font families

4.1 Plotting functions in lattice

5.1 Graphical primitives in grid
5.2 Coordinate systems in grid
5.3 Graphical parameters in grid
5.4 Grid font faces

6.1 Functions for working with grobs

© 2006 by Taylor & Francis Group, LLC