Handbook of Statistics
Volume 32
Computational Statistics with R

Edited by
Marepalli B. Rao
Division of Biostatistics and Epidemiology,
Department of Environmental Health,
University of Cincinnati, Cincinnati, Ohio, USA

C.R. Rao
C.R. Rao AIMSCS,
University of Hyderabad Campus,
Hyderabad, India
1. Introduction to R

Chaitra H. Nagaraja

1 Introduction

2 Setting Up R
   2.1 Installing and Starting R
   2.2 Memory
   2.3 Saving Your Code and Workspace
   2.4 R Packages

3 Basic R Objects and Commands
   3.1 Numbers, Character Strings, and Logicals
   3.2 Scalars, Vectors, Matrices, and Arrays
   3.3 Data Frames and Lists
   3.4 Strings and Factors

4 Writing Programs
   4.1 Conditional Statements
   4.2 if/else Statements
   4.3 for Loops
   4.4 while Loops
   4.5 Functions
   4.6 Debugging and Efficiency

5 Input and Output

6 Data Processing

7 Exploratory Data Analysis

8 Statistical Inference and Modeling
   8.1 Hypothesis Testing
   8.2 Regression

9 Simulation

10 Numerical Techniques

11 Annotated References
   Set Up
   Text Editors
   Introductory Resources and Books
2. R Graphics

Deepayan Sarkar

1 Introduction
1.1 Origins
1.2 Principles of Data Graphics

2 Traditional Graphics
2.1 The plot() Function
2.2 Other Common High-Level Functions
2.3 Visualizations for Time Series Data
2.4 Customizing Plots Using Low-Level Functions
2.5 Limitations of Traditional Graphics

3 Grid Graphics
3.1 Viewports
3.2 Units and Primitives
3.3 First Attempt

4 Lattice
4.1 Overview
4.2 Common High-Level Functions
4.3 Bar Charts and Dot Plots for Tabular Data
4.4 Scatterplots and Custom Displays
4.5 The "trellis" Object

5 ggplot

6 Further Reading
References

3. Graphics Miscellanea

Palash Mallick and Marepalli B. Rao

1 Introduction
2 The plot() Command
2.1 Features that Can Be Included in a Scatter Plot
3 Scatter Plots
3.1 Regression Analysis with Scatter Plots
3.2 Multiple Regression Analysis with Scatterplot Matrices
3.3 Scatterplot Matrices of Data Segregated by a Categorical Variable

4 Time Series Plots
4.1 Three Graphs in a Single Frame
4.2 Two Different Time Series Data Sets in a Single Plot

5 Pie Charts
6 Special Box Plots
7 xy Plots
8 Curves
9 LOWESS
10 Sunflower Plots
11 Violin Plots
12 Bean Plots
13 Bubble Charts
14 3D Surface Plot 130
15 Chernoff Faces—Graphical Presentation of Multivariate Data 133
16 Maps 137
  16.1 Drawing Common Maps 137
  16.2 Creating a Choropleth Map 139
References 142

4. Matrix Algebra Topics in Statistics and Economics Using R 143
Hrishikesh D. Vinod

1 Introduction 143
2 Basic Matrix Manipulations in R 144
3 Descriptive Statistics 146
  3.1 Outlier Detection and Normality Tests 148
  3.2 Multivariate Normality Tests 148
4 Matrix Transformations, Invariance, and Equivariance 148
  Affine Transformations Defined 149
  Desirable Invariance and Equivariance 149
  4.1 Data Standardization 149
  4.2 Limitations of the Usual Standardization 151
  4.3 Mahalanobis Distance and Outlier Detection 153
5 Payoff Matrices in Decision Analysis 154
6 Matrix Algebra in Regression Models 156
  6.1 Matrix QR Decomposition 157
  6.2 Collinearity and Singular Value Decomposition 158
  6.3 Heteroscedastic and Autocorrelated Errors 159
7 Correlation Matrices and Generalizations 160
  Bounds on the Cross-Correlation 160
  7.1 New Asymmetric Generalized Correlation Matrix 161
8 Matrices for Population Dynamics 165
9 Multivariate Components Analysis 168
  9.1 Projection Matrix: Generalized Canonical Correlations 168
  9.2 Invariant Coordinate Selection 169
10 Sparse Matrices 172
References 175

5. Sample Size Calculations with R: Level 1 177
Marepalli B. Rao and Subramanyam Kasala

1 Introduction 177
  1.1 Goals 178
  1.2 Why Did We Choose R? 178
2 General Ideas on Sample Size Calculations 178
  2.1 Example 179
  2.2 FAQ and Pointers 180
  2.3 Signal-to-Noise Ratio 181
  2.4 Some Features of the Normal Distribution 181
3 Single-Sample Problems

3.1 Quantitative 184
3.2 Testing of Hypotheses Environment 184
3.3 Specifications 185
3.4 Formula for Sample Size 186
3.5 Comments 190
3.6 The Other Type of One-Sided Alternative 190
3.7 The Case of Two-Sided Alternative 190
3.8 Comments 194
3.9 One-Sided Alternative 194
3.10 Two-Sided Alternative 194
3.11 The Case When the Population Standard Deviation $\sigma$ Is Unknown 194
3.12 The Case of One-Sided Alternative 194
3.13 Specifications 195
3.14 Comments 198
3.15 One-Sample Problem: One-Sided Alternative: $\sigma$ Is Known 198
3.16 One-Sample Problem: One-Sided Alternative: $\sigma$ Is Unknown 198
3.17 R Code 199
3.18 One-Sample Problem 201
3.19 Specifications 202
3.20 Example 202
3.21 An Alternative Approach 202
3.22 Example 202
3.23 Specifications 204

4 Two-Sample Problems: Quantitative Responses

4.1 Scenario 1 205
4.2 Specifications 205
4.3 Scenario 2 206
4.4 One-Sided Alternative 206
4.5 Specifications 206
4.6 Scenario 3 207
4.7 One-Sided Alternative 207
4.8 Specifications 207
4.9 Illustration 207
4.10 Two-Sided Alternative 208
4.11 Specifications 208
4.12 An Illustration 208
4.13 Scenario 4 211
4.14 Estimation Perspective 211
4.15 Scenario 1 211
4.16 Specifications 211
4.17 Example 212
4.18 Scenario 2 212
4.19 Specifications 212
4.20 Example 213
4.21 Scenario 3 213
4.22 Paired t-Test 213
4.23 Specifications 214
5 Multisample Problem—Quantitative Responses—Analysis of Variance
   5.1 Specifications 215
   5.2 Examples 216
   5.3 Structure of the Data 216
   5.4 Specifications 217
   5.5 Specifications 217
   5.6 Some Guidelines from the Social Sciences and Psychology 218
   5.7 Comments 220
References 220

6. Sample Size Calculations with R: Level 2 221
   Marepalli B. Rao and Hansen Bannerman-Thompson
   1 Single Proportions 221
      1.1 Problem 221
   2 Two-Sample Proportions 232
      2.1 Traditional Test 233
      2.2 Arcsine Square Root Transformation 234
   3 Effect Sizes 237
      3.1 The Case of Proportions 237
      3.2 The Case of t-Test 237
      3.3 The Case of Correlation 238
      3.4 Analysis of Variance 238
   4 Multisample Proportions 239
      4.1 Testing Equality of Several Population Proportions 239
   5 McNemar TEST 242
   6 Correlations 244
   7 Hazard Ratio in Survival Analysis 247
      7.1 A Pilot Study 249
   8 Multiple Regression 251
References 255

7. Binomial Regression in R 257
   John Muschelli, Joshua Betz, and Ravi Varadhan
   1 Binomial Regression in the Generalized Linear Model 258
   2 Standard Logistic Regression 259
   3 Assumptions Involved in the Standard Logistic Regression Model 261
   4 Residuals 261
      4.1 Interpreting Residuals 263
      4.2 Influential Points 266
   5 Overdispersion 268
      5.1 Estimation Using Quasilikelihood 269
      5.2 Adding Explanatory Terms to the Model 271
   6 Hypothesis Testing and Inference 273
   7 Model Performance 275
      7.1 ROC Curves/Sensitivity/Specificity/Accuracy 275
Contents

3 Current Knowledge of Second Malignancy 340
4 Clinical Trial Database 342
  4.1 Laboratory Test Data Analysis 343
  4.2 Medical History and Concomitant Medicines 345
  4.3 Efficacy Data Consideration 347
5 Integrated Analysis 347
6 Assessing Model Adequacy 353
7 Summary 355
References 356

10. Bayesian Networks 357
    Marepalli B. Rao and C. R. Rao

1 Introduction 357
2 Joint and Conditional Distributions 358
3 Generalities and Issues 362
4 Graph Theory 365
5 A Case Study 367
    Model Selection 372
6 Network Model Fitting 378
7 Learning Algorithm 383
References 385

Index 387