

# Table of Contents

Preface .....	xiii
---------------	------

## Part I. R Basics

<b>1. Getting and Installing R .....</b>	<b>3</b>
R Versions .....	3
Getting and Installing Interactive R Binaries .....	3
Windows .....	4
Mac OS X .....	5
Linux and Unix Systems .....	5
<b>2. The R User Interface .....</b>	<b>7</b>
The R Graphical User Interface .....	7
Windows .....	8
Mac OS X .....	8
Linux and Unix .....	9
The R Console .....	11
Command-Line Editing .....	13
Batch Mode .....	13
Using R Inside Microsoft Excel .....	14
RStudio .....	15
Other Ways to Run R .....	17
<b>3. A Short R Tutorial .....</b>	<b>19</b>
Basic Operations in R .....	19
Functions .....	21
Variables .....	22



Introduction to Data Structures	24
Objects and Classes	27
Models and Formulas	28
Charts and Graphics	30
Getting Help	35
<b>4. R Packages</b>	<b>37</b>
An Overview of Packages	37
Listing Packages in Local Libraries	38
Loading Packages	40
Loading Packages on Windows and Linux	40
Loading Packages on Mac OS X	40
Exploring Package Repositories	41
Exploring R Package Repositories on the Web	42
Finding and Installing Packages Inside R	42
Installing Packages From Other Repositories	45
Custom Packages	45
Creating a Package Directory	45
Building the Package	47

## Part II. The R Language

<b>5. An Overview of the R Language</b>	<b>51</b>
Expressions	51
Objects	52
Symbols	52
Functions	52
Objects Are Copied in Assignment Statements	54
Everything in R Is an Object	55
Special Values	55
NA	55
Inf and -Inf	56
NaN	56
NULL	56
Coercion	56
The R Interpreter	58
Seeing How R Works	59
<b>6. R Syntax</b>	<b>63</b>
Constants	63
Numeric Vectors	63
Character Vectors	64
Symbols	65
Operators	66
Order of Operations	67



Assignments	69
Expressions	69
Separating Expressions	69
Parentheses	70
Curly Braces	70
Control Structures	71
Conditional Statements	71
Loops	72
Accessing Data Structures	75
Data Structure Operators	75
Indexing by Integer Vector	76
Indexing by Logical Vector	78
Indexing by Name	79
R Code Style Standards	80
<b>7. R Objects</b>	<b>83</b>
Primitive Object Types	83
Vectors	86
Lists	87
Other Objects	88
Matrices	88
Arrays	89
Factors	89
Data Frames	91
Formulas	92
Time Series	94
Shingles	95
Dates and Times	95
Connections	96
Attributes	96
Class	99
<b>8. Symbols and Environments</b>	<b>101</b>
Symbols	101
Working with Environments	102
The Global Environment	103
Environments and Functions	104
Working with the Call Stack	104
Evaluating Functions in Different Environments	105
Adding Objects to an Environment	107
Exceptions	108
Signaling Errors	108
Catching Errors	109
<b>9. Functions</b>	<b>111</b>
The Function Keyword	111



Arguments	111
Return Values	113
Functions as Arguments	113
Anonymous Functions	114
Properties of Functions	115
Argument Order and Named Arguments	117
Side Effects	118
Changes to Other Environments	118
Input/Output	119
Graphics	119
<b>10. Object-Oriented Programming</b>	<b>121</b>
Overview of Object-Oriented Programming in R	122
Key Ideas	122
Implementation Example	123
Object-Oriented Programming in R: S4 Classes	129
Defining Classes	129
New Objects	130
Accessing Slots	130
Working with Objects	131
Creating Coercion Methods	131
Methods	132
Managing Methods	133
Basic Classes	134
More Help	135
Old-School OOP in R: S3	135
S3 Classes	135
S3 Methods	136
Using S3 Classes in S4 Classes	137
Finding Hidden S3 Methods	137

## Part III. Working with Data

<b>11. Saving, Loading, and Editing Data</b>	<b>141</b>
Entering Data Within R	141
Entering Data Using R Commands	141
Using the Edit GUI	142
Saving and Loading R Objects	145
Saving Objects with save	145
Importing Data from External Files	146
Text Files	146
Other Software	154
Exporting Data	155
Importing Data From Databases	156
Export Then Import	156



Database Connection Packages	156
RODBC	157
DBI	167
TSDBI	172
Getting Data from Hadoop	172
<b>12. Preparing Data</b>	<b>173</b>
Combining Data Sets	173
Pasting Together Data Structures	174
Merging Data by Common Fields	177
Transformations	179
Reassigning Variables	179
The Transform Function	179
Applying a Function to Each Element of an Object	180
Binning Data	185
Shingles	185
Cut	186
Combining Objects with a Grouping Variable	187
Subsets	187
Bracket Notation	188
subset Function	188
Random Sampling	189
Summarizing Functions	190
tapply, aggregate	190
Aggregating Tables with rowsum	193
Counting Values	194
Reshaping Data	196
Data Cleaning	205
Finding and Removing Duplicates	205
Sorting	206

## Part IV. Data Visualization

<b>13. Graphics</b>	<b>213</b>
An Overview of R Graphics	213
Scatter Plots	214
Plotting Time Series	220
Bar Charts	222
Pie Charts	226
Plotting Categorical Data	227
Three-Dimensional Data	232
Plotting Distributions	239
Box Plots	242
Graphics Devices	246
Customizing Charts	247



Common Arguments to Chart Functions	247
Graphical Parameters	247
Basic Graphics Functions	257
<b>14. Lattice Graphics</b>	<b>267</b>
History	267
An Overview of the Lattice Package	268
How Lattice Works	268
A Simple Example	268
Using Lattice Functions	270
Custom Panel Functions	272
High-Level Lattice Plotting Functions	272
Univariate Trellis Plots	273
Bivariate Trellis Plots	297
Trivariate Plots	305
Other Plots	310
Customizing Lattice Graphics	312
Common Arguments to Lattice Functions	312
trellis.skeleton	313
Controlling How Axes Are Drawn	314
Parameters	315
plot.trellis	319
strip.default	320
simpleKey	321
Low-Level Functions	322
Low-Level Graphics Functions	322
Panel Functions	323
<b>15. ggplot2</b>	<b>325</b>
A Short Introduction	325
The Grammar of Graphics	328
A More Complex Example: Medicare Data	333
Quick Plot	342
Creating Graphics with ggplot2	343
Learning More	347
<hr/>	
<b>Part V. Statistics with R</b>	
<b>16. Analyzing Data</b>	<b>351</b>
Summary Statistics	351
Correlation and Covariance	354
Principal Components Analysis	357
Factor Analysis	360
Bootstrap Resampling	361



<b>17. Probability Distributions .....</b>	<b>363</b>
Normal Distribution	363
Common Distribution-Type Arguments	366
Distribution Function Families	366
<b>18. Statistical Tests .....</b>	<b>371</b>
Continuous Data	371
Normal Distribution-Based Tests	372
Non-Parametric Tests	385
Discrete Data	388
Proportion Tests	388
Binomial Tests	389
Tabular Data Tests	390
Non-Parametric Tabular Data Tests	396
<b>19. Power Tests .....</b>	<b>397</b>
Experimental Design Example	397
t-Test Design	398
Proportion Test Design	398
ANOVA Test Design	400
<b>20. Regression Models .....</b>	<b>401</b>
Example: A Simple Linear Model	401
Fitting a Model	403
Helper Functions for Specifying the Model	404
Getting Information About a Model	404
Refining the Model	410
Details About the lm Function	410
Assumptions of Least Squares Regression	412
Robust and Resistant Regression	414
Subset Selection and Shrinkage Methods	416
Stepwise Variable Selection	416
Ridge Regression	417
Lasso and Least Angle Regression	418
elasticnet	419
Principal Components Regression and Partial Least Squares	420
Regression	420
Nonlinear Models	420
Generalized Linear Models	421
glmnet	424
Nonlinear Least Squares	427
Survival Models	428
Smoothing	433
Splines	433
Fitting Polynomial Surfaces	435



Kernel Smoothing	436
Machine Learning Algorithms for Regression	437
Regression Tree Models	439
MARS	450
Neural Networks	455
Project Pursuit Regression	459
Generalized Additive Models	462
Support Vector Machines	464
<b>21. Classification Models</b>	<b>467</b>
Linear Classification Models	467
Logistic Regression	467
Linear Discriminant Analysis	472
Log-Linear Models	476
Machine Learning Algorithms for Classification	477
k Nearest Neighbors	477
Classification Tree Models	478
Neural Networks	482
SVMs	483
Random Forests	483
<b>22. Machine Learning</b>	<b>485</b>
Market Basket Analysis	485
Clustering	490
Distance Measures	490
Clustering Algorithms	491
<b>23. Time Series Analysis</b>	<b>495</b>
Autocorrelation Functions	495
Time Series Models	496
<hr/>	
<b>Part VI. Additional Topics</b>	
<b>24. Optimizing R Programs</b>	<b>503</b>
Measuring R Program Performance	503
Timing	503
Profiling	504
Monitor How Much Memory You Are Using	505
Profiling Memory Usage	506
Optimizing Your R Code	507
Using Vector Operations	507
Lookup Performance in R	509
Use a Database to Query Large Data Sets	516
Preallocate Memory	516



Cleaning Up Memory	516
Functions for Big Data Sets	517
Other Ways to Speed Up R	518
The R Byte Code Compiler	518
High-Performance R Binaries	520
<b>25. Bioconductor .....</b>	<b>525</b>
An Example	525
Loading Raw Expression Data	526
Loading Data from GEO	530
Matching Phenotype Data	532
Analyzing Expression Data	533
Key Bioconductor Packages	537
Data Structures	541
eSet	541
AssayData	543
AnnotatedDataFrame	543
MIAME	544
Other Classes Used by Bioconductor Packages	545
Where to Go Next	546
Resources Outside Bioconductor	546
Vignettes	546
Courses	547
Books	547
<b>26. R and Hadoop .....</b>	<b>549</b>
R and Hadoop	549
Overview of Hadoop	549
RHadoop	554
Hadoop Streaming	568
Learning More	571
Other Packages for Parallel Computation with R	571
Segue	571
doMC	572
Where to Learn More	573
<b>Appendix: R Reference .....</b>	<b>575</b>
<b>Bibliography .....</b>	<b>675</b>
<b>Index .....</b>	<b>677</b>