

# Exercises that Practice and Extend Skills with R

John Maindonald

April 15, 2009

**Note:** Asterisked exercises (or in the case of “IV:  $\hat{L}$ Examples that Extend or Challenge”, set of exercises) are intended for those who want to explore more widely or to be challenged. The subdirectory **scripts** at <http://www.math.anu.edu.au/~courses/r/exercises/scripts/> has the script files.

Also available are Sweave (**.Rnw**) files that can be processed through R to generate the  $\text{\LaTeX}$  files from which pdf's for all or some subset of exercises can be generated. The  $\text{\LaTeX}$  files hold the R code that is included in the pdf's, output from R, and graphics files.

There is extensive use of datasets from the *DAAG* and *DAAGxtras* packages. Other required packages, aside from the packages supplied with all binaries, are:

`randomForest` (XII:rdiscrim-lda; XI:rdiscrim-ord; XIII: rdiscrim-trees; XVI:r-largish), `mlbench` (XIII:rdiscrim-ord), `e1071` (XIII:rdiscrim-ord; XV:rdiscrim-trees), `ggplot2` (XIII: rdiscrim-ord), `ape` (XIV: r-ordination), `mcclust` (XIV: r-ordination), `oz` (XIV: r-ordination).

## Contents

<b>I</b>	<b>R Basics</b>	<b>5</b>
1	Data Input	5
2	Missing Values	5
3	Useful Functions	6
4	Subsets of Dataframes	6
5	Scatterplots	7
6	Factors	8
7	Dotplots and Stripplots ( <i>lattice</i> )	8
8	Tabulation	9
9	Sorting	9
10	For Loops	10
11	The <code>paste()</code> Function	10
12	A Function	10
<b>II</b>	<b>Further Practice with R</b>	<b>11</b>
1	Information about the Columns of Data Frames	11
2	Tabulation Exercises	11
3	Data Exploration – Distributions of Data Values	12
4	The <code>paste()</code> Function	12
5	Random Samples	13
6	*Further Practice with Data Input	14

<b>III</b>	<b>Informal and Formal Data Exploration</b>	<b>15</b>
1	Rows with Missing Data – Are they Different	15
2	Comparisons Using Q-Q Plots	16
<b>IV</b>	<b>*Examples that Extend or Challenge</b>	<b>17</b>
1	Further Practice with Data Input	17
2	Graphs with logarithmic scales	17
3	Information on Workspace Objects	18
4	Different Ways to Do a Calculation – Timings	18
5	Functions – Making Sense of the Code	19
6	A Regression Estimate of the Age of the Universe	20
7	Use of <code>sapply()</code> to Give Multiple Graphs	21
8	The Internals of R – Functions are Pervasive	21
<b>V</b>	<b>Data Summary – Traps for the Unwary</b>	<b>23</b>
1	Multi-way Tables	23
2	Weighting Effects – Example with a Continuous Outcome	25
3	Extraction of <code>nassCDS</code>	26
<b>VI</b>	<b>Populations &amp; Samples – Theoretical &amp; Empirical Distributions</b>	<b>27</b>
1	Populations and Theoretical Distributions	27
2	Samples and Estimated Density Curves	28
3	*Normal Probability Plots	30
4	Boxplots – Simple Summary Information on a Distribution	31
<b>VII</b>	<b>Informal Uses of Resampling Methods</b>	<b>33</b>
1	Bootstrap Assessments of Sampling Variability	33
2	Use of the Permutation Distribution as a Standard	34
<b>VIII</b>	<b>Sampling Distributions, &amp; the Central Limit Theorem</b>	<b>35</b>
1	Sampling Distributions	35
2	The Central Limit Theorem	38
<b>IX</b>	<b>Simple Linear Regression Models</b>	<b>41</b>
1	Fitting Straight Lines to Data	41
2	Multiple Explanatory Variables	42
<b>X</b>	<b>Extending the Linear Model</b>	<b>43</b>
1	A One-way Classification – Eggs in the Cuckoo’s Nest	43
2	Regression Splines – one explanatory variable	45
3	Regression Splines – Two or More Explanatory Variables	46
4	Errors in Variables	47
<b>XI</b>	<b>Multi-level Models</b>	<b>49</b>
1	Description and Display of the Data	49

2	Multi-level Modeling	51
3	Multi-level Modeling – Attitudes to Science Data	53
4	*Additional Calculations	53
5	Notes – Other Forms of Complex Error Structure	54
<b>XII</b>	<b>Linear Discriminant Analysis vs Random Forests</b>	<b>55</b>
1	Accuracy for Classification Models – the Pima Data	55
2	Logistic regression – an alternative to lda	60
3	Data that are More Challenging – the crx Dataset	61
4	Use of Random Forest Results for Comparison	62
5	Note – The Handling of NAs	63
<b>XIII</b>	<b>Discriminant Methods &amp; Associated Ordinations</b>	<b>65</b>
1	Discrimination with Multiple Groups	65
<b>XIV</b>	<b>Ordination</b>	<b>71</b>
1	Australian road distances	71
2	If distances must first be calculated ...	73
3	Genetic Distances	73
4	*Distances between fly species	75
5	*Rock Art	76
<b>XV</b>	<b>Trees, SVM, and Random Forest Discriminants</b>	<b>77</b>
1	rpart Analyses – the Pima Dataset	77
2	rpart Analyses – Pima.tr and Pima.te	79
3	Analysis Using <i>svm</i>	81
4	Analysis Using <i>randomForest</i>	82
5	Class Weights	83
6	Plots that show the “distances” between points	83
7	Further Examples	84
<b>XVI</b>	<b>Data Exploration and Discrimination – Largish Dataset</b>	<b>85</b>
1	Data Input and Exploration	85
2	Tree-Based Classification	88
3	Use of <code>randomForest()</code>	89
4	Further comments	90
<b>A</b>	<b>Appendix – Use of the Sweave (.Rnw) Exercise Files</b>	<b>91</b>