Corrections (as of August 24, 2006) for the first printing of

Data Analysis and Graphics Using R – An Example-Based Approach

John Maindonald (email: john.maindonald@anu.edu.au) and John Braun

Readers who require corrections that are additional to the list dated March 19 2004 are advised to check the (very small) list of corrections for the second printing.

Note:

This document is designed to be printed in color, with replacement or new text in blue.

General

As a result of changes to R since this first edition was published:

- Code that uses data() to bring datasets in attached packages into the workspace can now, for packages used in this text, be omitted. (It may still be useful to use an explicit call to data() to overwrite a dataset, or in some cases multiple datasets, already present in the workspace.)
- The *date* package, described in Section 12.2.3, has been superseded by **as.Date()** and allied functions that are in *base* R.
- The description of the handling of missing values in subscripts, in Section 12.5 on page 318, is no longer accurate.
- While the code given in the text should all run, there are now more direct ways to handle some computations. This is especially relevant to the use of the *lattice* package.

See the web page http://www.maths.anu.edu.au/~johnm/r-book.html for additional details.

Preface

p.xviii: lines 9 & 10

Remove the two stray appearances of ss: responsibility ss and ss of

p.xviii, lines -6 to -5: Replace with

a CRAN (Comprehensive R Archive Network) site; go to http://cran.r-project.org and click on <u>Mirrors</u>.

Chapter 1

p.1, line 15: Replace pages http://cbis.anu.edu.au/DAAG and http://www.stats.uwo.ca/DAAG with page http://www.stats.uwo.ca/DAAG, or from http://cran.r-project.org.

p.5

line 4 of Section 1.1.6: Replace working directory by workspace
line 9 of Section 1.1.6: Replace workshop by workspace.

p.13, line -10

This line should read

> data(possum) # Precede, if necessary, with library(DAAG)

p.14

p.14, line 5: Omit , eda, ts (time series)line 19 and line 22: Replace working directory with workspace

p.16, lines -6 to -5

Omit the sentence: We have already used ... graphs on a page.

p.18, line4

left adjusted would be better left-justified

p.19, line 16

Replace default by current

p.22, lines 12-13

Replace grid.points(), grid.lines(), grid.text(), etc. by llines(), lpoints(), ltext(), lplot(), larrows() and lsegments().

p.23, fnote 3, line 2

Replace each of the 4 occurrences of `` with ".

p.24, line 8

The sentence starting "Rounding ..." should start a new paragraph.

p.27, Exercise 4

In the final line of this exercise, replace \checkmark with ".

Chapter 2

p.33, fnote 7

Add two further lines:

See Section 4.5 for a description of stripplot()

See Subsection 12.2.1 for a description of strsplit()

p.35, line 4

Omit (in the *modreg* package).

[The function lowess() is now in *stats*; in an out-of-the-box installation this is attached at startup.]

p.37

fnote 12: Replace lines 2 & 3 with

Arguments 1 & 2 to matplot() can be vector or matrix or data frame

fnote 13: Add following line 2

stack() concatenates selected data frame columns into a single # column, & adds a factor that has the column names as levels

Add as the final two lines of fnote 13:

The parameter relation="sliced" causes the length of the # relevant scale(s) ('slice(s)') to be the same for all panels

p.40

lines 11-12: panel.superpose is illegally split across the two lines.

(Actually panel=panel.superpose, is not necessary and could be omitted. Note also the parameter setting auto.key=TRUE that might usefully appear.)

p.45, fnote 17

Add the further line

Subsection 12.6.7 has information on split()

p.46, lines -16

A closing) should appear following method="spearman"

p.50, Section 2.6, Exercise 1

Replace the final line (Comment ... present.) by

Use table(cut(possum\$age, breaks=c(0,1.5,3,4.5,6,7.5,9)) to obtain the table of counts. In which interval are possums with age=3 included; in (1.5,3] or in (3,4.5]? List the values of age that are included in each successive interval. Explain why setting breaks=c(0,1.5,3,4.5,6,7.5,9) leads to a histogram that is misleading.

Chapter 3

p.52, line 11 Replace natural by physical

```
p.57, lines 3 and 7
```

Replace } by)

p.59, line -4

Replace 1.645 by 1.28

p.69, Exercise 5

Replace mfrow by the parameter mfrow to par()

Replace rows by panels, in the three places where it appears.

Chapter 4

p.74, fnote 2 Modify the final line to read

p.77, Figure 4.4

Replace pnorm(0.975) by qnorm(0.975), and replace pt(0.975, 8) by qt(0.975, 8)

p.83, line-10: Omit in the *ctest* package

line-2 to -1: Replace This methodology is not, at the time of writing, implemented in R. with The function cor.test() implements this methodology.

p.87, line 10: Replace somewhat with substantially

p.89, line -9: Replace default). with default, before fitting the model).

p.92, line1-4 of Subsection 4.5.3:

fnote 22: The code should read

(The first argument to strsplit() must be a character string. Hence we replace rice\$trt, in lines 1 and 3, by as.character(rice\$trt). In the second to final line, replace anu843 by variety.)

p.92, lines 12-13

Replace In this instance it turns out that there is no detectable block effect, which is why we could ignore it. by In this instance it turns out that the block effect is small, and does not much affect the standard error estimates.

p.98

line -3: Replace standard errors by confidence intervals.

fnote 24, line6: Replace z <- array(, 2000) by z <- numeric(2000)</pre>

p.99

fnote 25: Replace fnote 25 by

Use boot.wren, calculated on p.74, as a starting point boot.ci(wren.boot, type="perc")

fnote 26, line 1: Replace poss o m by poss u m

p.102, Further reading

The author of the *boot* package, whom we inadvertently failed to acknowledge, is Angelo Canty. See Canty (2002) for a readable and helpful overview of this package.

lines -13 to -11: Omit the sentence Gigerenzer (2002) has helpful insights on ... violent behaviour.

(It is scrambled and is anyway redundant.)

p.103, References for Further reading

Add the reference:

Canty, A.J. 2002. Resampling methods in R: the boot package. R News 2/3: 2-7.

p.104

p.104, line4: Omit library(ts)

exercise 7, line2: Omit that is in the *ctest* package

Exercise 3: Add

Use qqplot(), with the parameter setting x = qunif(ppoints(50)), to compare the distribution of the *p*-values with that of a uniform random variable, on the interval [0, 1]. Comment on the plot. **Exercise** 7: This should be asterisked.

Chapter 5

page 112

Figure 5.4: Note that Exercise 6 on page 133 has R code for panels B and D.)

p.112, line 10: Replace a strong indication with an indication

line -9: Replace Figure 5.5 shows the fitted loess curve with

Figure 5.5 shows the plot of residuals from the loess fit versus the predictor

page 113, Figure 5.5

Add to the caption Note that Exercise 7 on page 133 has the R code.

page 114, fnote 8

Insert as the first line

data(softbacks)

p.116, line 6: Replace in the lqs package with (also in MASS)

page 122, fnote 14, line 3

Replace the semicolon by a comma, i.e. cv.lm(houseprices, houseprices.lm\$call)

p.132, Section 5.11, Exercises

Exercise 2: Replace the final three lines with:

Use the robust regression function rlm() from the *MASS* package to fit lines to the data in elastic1 and elastic2. Compare the results with those from use of lm(). Compare regression coefficients, standard errors of coefficients, and plots of residuals against fitted values.

Exercise 4, line 3: Replace the final three lines with:

[The relevant theory is that associated with the Claudius-Clapeyron equation, by which the logarithm of the vapor pressure is approximately inversely proportional to the absolute temperature. For further details of the Claudius-Clapeyron equation, search on the internet, or look in a suitable reference text.]

p.133, line-9: Omit library(modreg), i.e., omit this line.

Chapter 6

p.136, line 16

The 95% confidence interval is for the coefficient of volume.

p.136, line 18

The confidence interval is for the coefficient of volume.

p. 144, fnote 13

Move line -4, i.e. par(mfrow = c(1,1)), to be the final line of the footnote.

p. 145, lines 6-7

The split between lines 6 & 7 is at the wrong place. Replace with

p.166, line 15

Omit the final 1.

p.173, Section 6.10, Exercise 6

Reword thus:

Section 6.1 used lm() to analyze the allbacks data that are presented in Figure 6.1. Repeat the analysis using (1) the function rlm() in the *MASS* package, and (2) the function lqs() in the *lqs* package. Compare the two sets of results with the results in Section 6.1.

p.173, line-12 to -11: Replace in the lqs package with (MASS)

p.174, Exercise 9

Remove the asterisk. Omit linear in line 3 and non-linear in line 5. Omit from Are there any advantages ... to the end of the question.

Chapter 7

p.179

line 19: Replace three by four.

p.183

line -10 (final line under "Model 3)": Replace " $(z_2 = 1 \text{ and } z_3 = 0)$ " by $(z_2 = 0 \text{ and } z_3 = 1)$ line -6 (final line under "Model 4)": Replace " $(z_2 = 1 \text{ and } z_3 = 0)$ " by $(z_2 = 0 \text{ and } z_3 = 1)$

p.191, line 4: Omit , implemented in the modreg package,

p.194, lines 4-5: Omit and references on the help pages for the *modreg* and *mgcv* packages with references on the help pages for the *mgcv* package, and for the smoothing functions used in this chapter

pp.194-196, Section 7.7, Exercises Exercise 1, Section 7.7: Replace "using the log(weight), i.e., omit the.

Exercise 2: Replace "Section 6.7" by "Section 7.3".

Exercise 3(c), line -2 on page: Replace Minimizing by Maximizing.

Exercises 9 & 10: Replace "Exercise 6 by "Exercise 7.

Exercise 11, lines 3-4: The web page is now http://www.antarctica.ac.uk/met/jds/ozone/

Chapter 8

p.202

line 4: Replace axes=FAlSE by axes=FALSE

line -6: Replace meanmin by meanmax

p.214-215

```
p. 214, lines 14-15
Replace
```

In our example, any comparisons with Bank, which is the default reference level, will be inaccurate, as it has only one by

In the data as they stand, the initial level of habitat, which is Bank, will by default be used as the reference level. No moths of the first species (A) were found in this habitat. This creates problems for the calculation of standard errors, as will shortly become apparent.

p. 215, line 10

Replace "have large standard errors." with "have large and meaningless standard errors. The standard generalized linear model approximation to the standard error breaks down when, as here, the fitted value for the reference level is zero."

Here is some further commentary. The following results are obtained from version 2.0.1 of R.

```
> A1.glm <- glm(A ~ log(meters) + habitat, family=quasipoisson, data=moths)
> A2.glm <- glm(A ~ log(meters) + habitat, family=quasipoisson, data=moths,
+ subset=habitat!="Bank")
> b1 <- coef(A1.glm)</pre>
> b1
     (Intercept)
                      log(meters) habitatDisturbed habitatLowerside
         -15.696
                                             15.622
                                                               16.906
                             0.129
   habitatNEsoak
                    habitatNWsoak
                                      habitatSEsoak
                                                       habitatSWsoak
          16.084
                            18.468
                                             16.968
                                                               17.137
habitatUpperside
          16.743
> ## Notice that the above differ from the coefficients given on p.214.
> ## The reason is that, since the analysis reported in the book was done,
> ## glm()'s convergence criteria have been tightened
> ## Instead of the former exp(-6.696) on p.214 as an approximation to zero,
> ## there is now, with the tighter convergence criterion, exp(-15.67) as
> ## an approximation to zero.
> ##
> b2 <- coef(A2.glm)
> options(digits=3)
> b1[1]+b1[3:9]
habitatDisturbed habitatLowerside
                                      habitatNEsoak
                                                       habitatNWsoak
         -0.0734
                            1,2098
                                             0.3881
                                                               2.7722
   habitatSEsoak
                    habitatSWsoak habitatUpperside
          1,2719
                            1.4412
                                             1.0475
> b1[1]+b1[4:9]
habitatLowerside
                                      habitatNWsoak
                    habitatNEsoak
                                                       habitatSEsoak
                                              2.772
                                                                1.272
           1.210
                            0.388
   habitatSWsoak habitatUpperside
           1.441
                             1.047
> b2[1]+b2[3:8]
habitatLowerside
                                      habitatNWsoak
                    habitatNEsoak
                                                       habitatSEsoak
           1.210
                            0.388
                                              2.772
                                                                1.272
   habitatSWsoak habitatUpperside
           1.441
                             1.047
```

Notice that b1[1]+b1[4:9] and b1[1]+b1[3:8] are, essentially, identical, i.e., the fitted values from the two models are the same. The same fitted values would be obtained from the coefficients in the book.

Corrections

One set of nonsensical SEs has been replaced by another (as it happens, even less believable). As noted above, the standard GLM approximations have broken down.

p.221, line -5

Replace probit by logit

p.223, Section 8.9, Exercise 2

Replace the final sentence by:

Patients whose risk is sufficiently high will be sent for CT (computed tomography). Using a risk threshold of 0.025 (2.5%), turn the result into a decision rule for use of CT.

Chapter 9

p.227

line 12: Following These values can be gleaned from VarCorr(science.lme), add

, or alternatively by squaring output from intervals(science.lme, which="var-cov")

Thus the sentence now reads

"These values can be gleaned from VarCorr(science.lme, or alternatively by squaring output from intervals(science.lme, which="var-cov")"

lines -10 & -9: In line -10, omit

which we give in the form that we obtain from the output,

line -9: Omit very. Following the full stop, insert

In the code given in footnote 4, the values have been squared, and are variances, not SDs.

fnote 3, line 2: Indent the code to align following the round bracket on the previous line

fnote 3, final line: Omit

fnote 4: Replace intervals.lme() by intervals()

[intervals() is the generic function, which calls intervals.lme(). It is not now possible to call intervals.lme() directly.]

p.238

line 1 (first line below the graphs): (near end of line): Replace level= 0 by level=2

line 3, near end of line: Replace level= 2 by level=0

line 4: Replace

"fitted values at level 2 and fitted values at level 1"

by "fitted values at level 1 and fitted values at level 0"

line 5: Replace

"fitted values at level 1 and fitted values at level 0"

by "fitted values at level 2 and fitted values at level 1"

(The text (but not the code) has, effectively, interchanged the labels 0 and 2.)

fnote 12: Replace the final line (# i.e., the highest ...) by

i.e., they are adjusted for plot and block effects.

p.240, final three lines

Replace the final The at the end of line -3 by Examination of the table of coefficients that we give below shows that the

Replace agegpsenior, in the two places where it appears, by agegp0lder

With version 0.84 of DAAG or later, this labeling of output can be reproduced by preceding the use of the code on page 240 with:

tinting\$agegp <- capstring(tinting\$agegp)</pre>

```
p.242, line 12: Omit , in the ts package,
```

footnote 15, line 1: Omit this line (starting library(ts))

p.245, fnote 19

Following line 4, insert the line

LH.yw\$ar # autocorrelation estimate of alpha

Following line 7, insert the line

p.246, line -12 (third bullet point)

With $b_1 = b_2 = 0$, $b_3 = 0.5$, $b_4 = 0$, $b_5 = 0.5$ (q = 5) there are spikes at lags 2, 3 and 5.

p.246, final line Omit this line.

p.252, line 12

Add normality at the end of the line (also one # is enough, here and elsewhere on this page), i.e.

qqnorm(resid(soi.lme, type="normalized")) # Examine normality

p.258, Section 9.10, Exercise 3

In the final sentence, prior to "would it be too narrow or too wide?", insert with σ is defined as in Section 9.5.3,

Chapter 10

p.262, fnote 2

The code that follows library(rpart) should start a new line.

p.268, line 9

Omit from the R modreg package. The function lowess() is now in the stats package.

p.275, line 4

Replace nsplit = 12 by nsplit = 16

p.277

In the output on the final three lines of p.277, replace 320 by 321, and replace 240 by 239.

Note: The final five lines of page 277 continue the output from line -2 on page 276, i.e., from > print(mifemb.rpart).

p.280, Exercise 3

Replace nsw74p re d1 by nsw74psid1

Chapter 11

p.284, line 20: Omit this line (starting library(mva))

p.293, fnote 7, line 1

Replace MASS by DAAG.

lines 16-17: Replace Both are males by Both are females.

Omit the next sentence, i.e., omit

Examination of the graph where different levels of employment are identified makes it clear that both rely on parent

p.294, fnote 10, lines 2-4

These require correction, to read

```
plot(BDI[not.na] ~ ss.pr$scores[ ,1], col=as.numeric(gender[not.na]),
        pch=as.numeric(gender[not.na]), xlab ="1st principal component",
        ylab="BDI")
topleft <- par()$usr[c(1,4)]</pre>
```

(A consequence of the failure of the code on p.294 to subset **gender** is that the assignment of symbols to points in Figure 11.5 is wrong.)

p.298, Section 11.6, Exercises

Exercise 1: Replace the second sentence (Compare ...) by

For each of the first and second principal components, plot the loadings for females against the loadings for all data combined, and similarly for males. Are there any striking differences?

Exercise 2: Replace principal component , in each of the three places where it appears, by discriminant function.

In line 3, replace the second occurrence of second by first

Exercise 4: Sentences after the first should be replaced by

The variable altitude has the same value for all observations within a site. What are the implications for any attempt to use altitude as an explanatory variable in the linear discriminant function calculations of Subsection 11.2.1?

Chapter 12 p.303, line 6 Replace per by par

p.305, line -3: Replace String functions by Character string functions
line -2: Replace text strings by character strings.

p.307, line 9

Delete > attach(cabbages) . As the final line of the computer output, insert > detach(cabbages)

p.309, line 20

Following sapply(elastic1, mean) , add # elastic1 is from DAAG

p.310, line 5: Replace These functions are suitable only for the simplest applications. by See however help(as.Date) for alternatives to the use of the *date* package.

p.314

lines -12 and -2: Replace elasticband by elastic1

line -11: Omit **#**, and move stretch and distance to be above the columns for which they are headings.

p.315, Section 12.4: Factors

line -19: To extract the codes 1, 2, ..., specify unclass(country).

lines -12 and -11 (consequent on changes to R): If the labels argument to text() is a factor, as of 2.0.0 it is the levels that are plotted.

p.318

line -14: Omit (but not in S-PLUS)

line -12: Immediately prior to this line of computer code, insert

y <- c(NA, 2, 4, 1, 7)

line -12 (end of line): Replace see below by see page 341

p.321, line 20

Replace y.names by y.names

p.323, lines 8-9

To avoid ambiguity, this sentence should read

Because a data frame has the structure of a list of columns, sapply() and lapply() can be used to apply a function to each of its columns in turn.

Corrections

p.327, lines1 Replace rep(1:5, 6) by rep(1:6,5)

p.330, line 21: [The names and contents of the base packages changed in R-1.9.0, and again in R-2.2.0. Partly as a result, the names of the packages that are attached at startup are now different.]

Replace that *methods*, *ctest* and several other packages are by that a number of packages were

p.335, Exercise 8

Modify this question to read

Save the objects in your workspace, into an image (**.RData**) file with the name **archive.RData**. Then remove all objects from the workspace. Demonstrate how, without loading the image file, it is possible to list the objects that were included in **archive.RData** and to recover a deleted object that is again required.

pp.336-337, Exercise 11

Reword this question thus

Try the following, for a range of values of n between, e.g., 2×10^5 and 10^7 . (On systems that are unable to cope with such large numbers of values, adjust the range of values of n as necessary.)

n <- 10000; system.time(sd(rnorm(n)))</pre>

The first output number is the user cpu time, while the third output number is the elapsed time. Plot each of these numbers, separately, against n. Comment on the graphs. Is the elapsed time roughly linear with n? Try the computations both for an otherwise empty workspace, and with large data objects (e.g., with 10^7 or more elements) in the workspace.

Epilogue

p.339, lines 22

Replace 90% by 10%

p.340, lines 1-2

Between the first and second line, insert the words the motivation.

References

pp.346-349

There are more up to date versions of some monographs that are noted in the references:

Chatfield, C. 2002. Confessions of a statistician. The Statistician 51: 1-20.

Chatfield, C. 2003. The Analysis of Time Series: an Introduction, 6th edn. Chapman and Hall.

Chatfield, C. 2003. Problem Solving. A Statistician's Guide, 2nd edn. Chapman and Hall/CRC.

Cleveland, W.S. 1994. The Elements of Graphing Data, revised edn. Hobart Press.

Collett, D. 2003. Modelling Survival Data in Medical Research, 2nd edn. Chapman and Hall.

Diggle, P. 1990. Time Series: a Biostatistical Introduction. Clarendon Press.

Krzanowski, W.J. 2000. *Principles of Multivariate Analysis. A User's Perspective*, revised edn. Clarendon Press.

R Core Development Team, updated regularly. An Introduction to R. Available from CRAN sites (for a list of CRAN sites, go to http://cran.r-project.org).

R Core Development Team, updated regularly. R Language Definition. Available from CRAN sites.

Steel R.G.D., Torrie J.H. and Dickie, D.A. 1993. Principles and Procedures of Statistics. A Biometrical Approach, 3rd edn. McGraw-Hill, New York.

Williams, E.R., Matheson, A.C. and Harwood, C.E. 2002. Experimental Design and Analysis for Use in Tree Improvement, revised edn. CSIRO Information Services.

See p.194 for the Eubank reference.

Note also that Venables and Ripley's *Modern Applied Statistics with* S-PLUS, 4th edn. should be *Modern Applied Statistics with* S, 4th edn.

p.351, Web sites

line 6: Data set ozone Change the url to http://www.antarctica.ac.uk/met/jds/ozone/

Indices

pp.352-355, Index of R Symbols & Functions

Under additions (in text) replace 333 by 332 Under assign the only page number that should appear is 345. Replace I... by I (wrapper function) ... Under library, omit 343.3 Under log, omit 343.3 Under plot, 342.1 should be 342 Under text, replace 342.1, 342.4 by 342.

pp.356-360, Index of Terms

Under <u>AIC</u>, the page numbers should be 150-151, 152-153, 240, 245, 251, 278, see also BIC

Add the main item and subitem (p.357, column 2) dates

calculations with, $310\,$

Under <u>function</u>, add the subitem issues for writing and use, 308

Under histograms, run the lines on.

Under <u>plots</u>, add the subitem component plus residual, 143, 144, 145

Add the main item (p.359, column 2) random numbers, 60-62, see also *Index of R symbols and Functions*, under rbinom,

Corrections

rchisq, rexp, rnorm, rpois, runif and sample Under regression, influence, the page numbers should be: 66, 114, 115, 116, 139, 146, 194, 195 Under regression, add the subitems Cook's distance, 115, 139, 143, 146 ordinal logistic (POLR), 216-220 outliers, 41, 48, 49, 66, 115, 143, 146, 148, 149, 152, 157, 171, 189, 216, 262, 292, 293, 298 propensity scores, 161, 295, 296, 297 variance inflation factor (VIF), 167, 168 Under residuals add the subitem from two-way table, 87 Under variance, add the subitem heterogeneity, 112, 169, 170 Add the main items

working directory, 5, 7, 331 workspace, 5,7, 25, 330, 331, 335, 343

p.361, Index of Names

Add Eubank, R.L., 194. Add Williams, E.R., 256

Figures

Figures 2.10, 2.11, 8.8, 11.1, 11.2 and 12.2 should be in color.