

The Statistical Sleuth in R:

Chapter 9

Linda Loi Kate Aloisio Ruobing Zhang Nicholas J. Horton*

January 23, 2019

Contents

1	Introduction	1
2	Effects of light on meadowfoam flowering	2
2.1	Data coding, summary statistics and graphical display	2
2.2	Multiple linear regression model	4
3	Why do some mammals have large brains?	5
3.1	Data coding and summary statistics	5
3.2	Graphical presentation	6
3.3	Multiple linear regression model	10

1 Introduction

This document is intended to help describe how to undertake analyses introduced as examples in the Third Edition of the *Statistical Sleuth* (2013) by Fred Ramsey and Dan Schafer. More information about the book can be found at <http://www.proaxis.com/~panorama/home.htm>. This file as well as the associated **knitr** reproducible analysis source file can be found at <http://www.math.smith.edu/~nhorton/sleuth3>.

This work leverages initiatives undertaken by Project MOSAIC (<http://www.mosaic-web.org>), an NSF-funded effort to improve the teaching of statistics, calculus, science and computing in the undergraduate curriculum. In particular, we utilize the **mosaic** package, which was written to simplify the use of R for introductory statistics courses. A short summary of the R needed to teach introductory statistics can be found in the mosaic package vignette (<http://cran.r-project.org/web/packages/mosaic/vignettes/MinimalR.pdf>).

To use a package within R, it must be installed (one time), and loaded (each session). The package can be installed using the following command:

*Department of Mathematics and Statistics, Smith College, nhorton@smith.edu

```
> install.packages('mosaic') # note the quotation marks
```

Once this is installed, it can be loaded by running the command:

```
> require(mosaic)
```

This needs to be done once per session.

In addition the data files for the *Sleuth* case studies can be accessed by installing the **Sleuth3** package.

```
> install.packages('Sleuth3') # note the quotation marks
```

```
> require(Sleuth3)
```

We also set some options to improve legibility of graphs and output.

```
> trellis.par.set(theme=col.mosaic()) # get a better color scheme for lattice
> options(digits=3)
```

The specific goal of this document is to demonstrate how to calculate the quantities described in Chapter 9: Multiple Regression using R.

2 Effects of light on meadowfoam flowering

Do different amounts of light affect the growth of meadowfoam (a small plant used to create seed oil)? This is the question addressed in case study 9.1 in the *Sleuth*.

2.1 Data coding, summary statistics and graphical display

We begin by reading the data, clarifying the data, and summarizing the variables.

```
> head(case0901)

  Flowers Time Intensity
1    62.3    1      150
2    77.4    1      150
3    55.3    1      300
4    54.2    1      300
5    49.6    1      450
6    61.9    1      450

> case0901 = transform(case0901, Time = factor(ifelse(case0901$Time > 1, "Early", "Late")))
> summary(case0901)
```

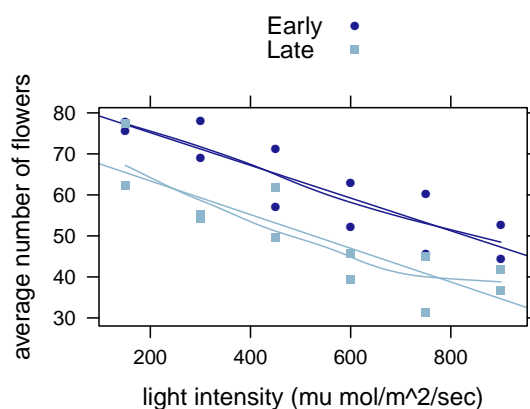
Flowers		Time	Intensity	
Min.	:31.3	Early:12	Min.	:150
1st Qu.:	45.4	Late :12	1st Qu.:	300
Median	:54.8		Median	:525
Mean	:56.1		Mean	:525
3rd Qu.:	64.5		3rd Qu.:	750
Max.	:78.0		Max.	:900


```
> favstats(Flowers ~ Intensity | Time, data=case0901)
```

	Time	min	Q1	median	Q3	max	mean	sd	n	missing
1	150.Early	75.6	76.1	76.7	77.2	77.8	76.7	1.556	2	0
2	300.Early	69.1	71.3	73.5	75.8	78.0	73.5	6.293	2	0
3	450.Early	57.0	60.5	64.0	67.6	71.1	64.0	9.970	2	0
4	600.Early	52.2	54.9	57.5	60.2	62.9	57.5	7.566	2	0
5	750.Early	45.6	49.3	53.0	56.6	60.3	53.0	10.394	2	0
6	900.Early	44.4	46.4	48.5	50.6	52.6	48.5	5.798	2	0
7	150.Late	62.3	66.1	69.8	73.6	77.4	69.8	10.677	2	0
8	300.Late	54.2	54.5	54.8	55.0	55.3	54.8	0.778	2	0
9	450.Late	49.6	52.7	55.8	58.8	61.9	55.8	8.697	2	0
10	600.Late	39.4	41.0	42.5	44.1	45.7	42.5	4.455	2	0
11	750.Late	31.3	34.7	38.1	41.5	44.9	38.1	9.617	2	0
12	900.Late	36.8	38.1	39.3	40.6	41.9	39.3	3.606	2	0
13	Early	44.4	52.5	61.6	72.2	78.0	62.2	12.117	12	0
14	Late	31.3	41.3	47.7	56.9	77.4	50.1	12.919	12	0

A total of 24 meadowfoam plants were included in this data. There were 12 treatment groups - 6 light intensities at each of the 2 timing levels (Display 9.2, page 239 of the *Sleuth*). The following code generates the scatterplot of the average number of flowers per plant versus the applied light intensity for each of the 12 experimental units akin to Display 9.3 on page 240.

```
> xyplot(Flowers ~ Intensity, groups=Time, type=c("p", "r", "smooth"),
+       data=case0901, auto.key=TRUE,
+       xlab="light intensity (mu mol/m^2/sec)", ylab="average number of flowers")
```



2.2 Multiple linear regression model

We next fit a multiple linear regression model that specifies parallel regression lines for the mean number of flowers as a function of light intensity as interpreted on page 239.

```
> lm1 = lm(Flowers ~ Intensity+Time, data=case0901)
> summary(lm1)
```

Call:

```
lm(formula = Flowers ~ Intensity + Time, data = case0901)
```

Residuals:

Min	1Q	Median	3Q	Max
-9.65	-4.14	-1.56	5.63	12.16

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	83.46417	3.27377	25.49	< 2e-16
Intensity	-0.04047	0.00513	-7.89	1e-07
TimeLate	-12.15833	2.62956	-4.62	0.00015

Residual standard error: 6.44 on 21 degrees of freedom

Multiple R-squared: 0.799, Adjusted R-squared: 0.78

F-statistic: 41.8 on 2 and 21 DF, p-value: 4.79e-08

```
> confint(lm1, level=.95) # 95% confidence intervals
```

	2.5 %	97.5 %
(Intercept)	76.6560	90.2723
Intensity	-0.0511	-0.0298
TimeLate	-17.6268	-6.6899

We can also fit a multiple linear regression with an interaction between light intensity and timing of its initiation as shown in Display 9.14 (page 260) and interpreted on page 239.

```
> lm2 = lm(Flowers ~ Intensity*Time, data=case0901)
> summary(lm2)
```

Call:
lm(formula = Flowers ~ Intensity * Time, data = case0901)

Residuals:

Min	1Q	Median	3Q	Max
-9.52	-4.28	-1.42	5.47	11.94

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	83.14667	4.34330	19.14	2.5e-14
Intensity	-0.03987	0.00744	-5.36	3.0e-05
TimeLate	-11.52333	6.14236	-1.88	0.075
Intensity:TimeLate	-0.00121	0.01051	-0.12	0.910

Residual standard error: 6.6 on 20 degrees of freedom
Multiple R-squared: 0.799, Adjusted R-squared: 0.769
F-statistic: 26.5 on 3 and 20 DF, p-value: 3.55e-07

3 Why do some mammals have large brains?

What characteristics predict large brains in mammals? This is the question addressed in case study 9.2 in the *Sleuth*.

3.1 Data coding and summary statistics

We begin by reading the data and summarizing the variables.

```
> case0902 = transform(case0902, logbrain = log(Brain))
> case0902 = transform(case0902, logbody = log(Body))
> case0902 = transform(case0902, loggest = log(Gestation))
> case0902 = transform(case0902, loglitter = log(Litter))
```

```
> summary(case0902)
```

	Species	Brain	Body	Gestation
Aardvark	: 1	Min. : 0	Min. : 0	Min. : 16
Acouchis	: 1	1st Qu.: 13	1st Qu.: 2	1st Qu.: 63

```

African elephant: 1   Median : 74   Median : 9   Median :134
Agoutis           : 1   Mean    : 219   Mean    : 108   Mean    :151
Axis deer         : 1   3rd Qu.: 260   3rd Qu.: 95   3rd Qu.:226
Badger            : 1   Max.    :4480   Max.    :2800   Max.    :655
(Other)           :90

      Litter      logbrain      logbody      loggest
Min.    :1.00    Min.    :-0.80    Min.    :-4.07    Min.    :2.77
1st Qu.:1.00    1st Qu.: 2.53    1st Qu.: 0.73    1st Qu.:4.14
Median :1.20    Median : 4.30    Median : 2.19    Median :4.89
Mean    :2.31    Mean     : 3.86    Mean     : 2.13    Mean     :4.71
3rd Qu.:3.20    3rd Qu.: 5.56    3rd Qu.: 4.55    3rd Qu.:5.42
Max.    :8.00    Max.     : 8.41    Max.     : 7.94    Max.     :6.48

      loglitter
Min.    :0.000
1st Qu.:0.000
Median :0.182
Mean    :0.598
3rd Qu.:1.162
Max.    :2.079

```

A total of 96 mammals were included in this data. The average values of brain weight, body weight, gestation length, and litter size for each of the species were calculated and presented in Display 9.4 (page 241 of the *Sleuth*).

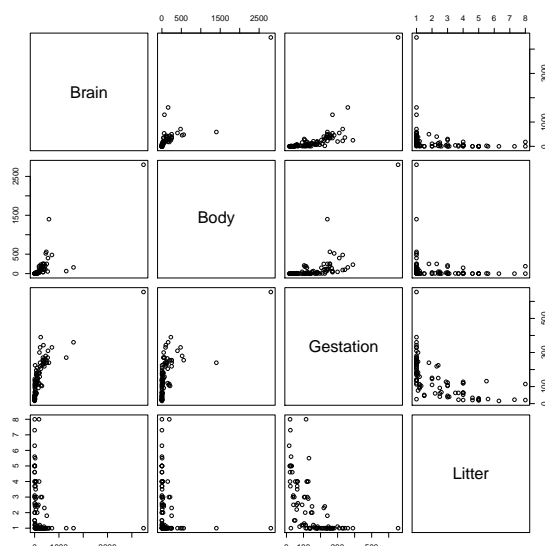
3.2 Graphical presentation

The following displays a simple (unadorned) pairs plot, akin to Display 9.10 on page 255.

```

> smalllds = subset(case0902, select=c("Brain", "Body", "Gestation", "Litter"))
> pairs(smalllds)

```

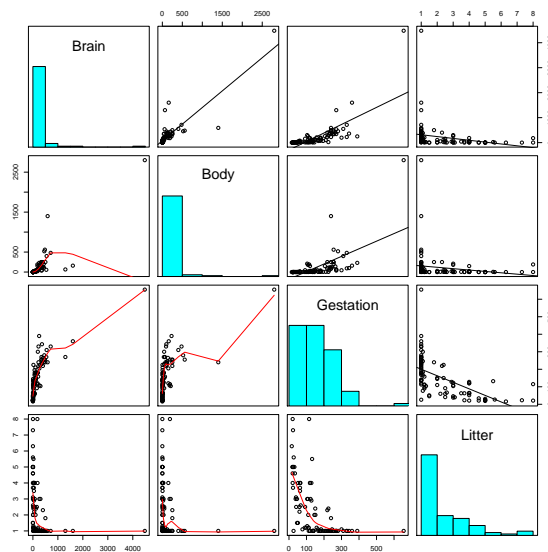


We can make it fancier if we like.

```
> panel.hist = function(x, ...)
+ {
+   usr = par("usr"); on.exit(par(usr))
+   par(usr = c(usr[1:2], 0, 1.5) )
+   h = hist(x, plot=FALSE)
+   breaks = h$breaks; nB = length(breaks)
+   y = h$counts; y = y/max(y)
+   rect(breaks[-nB], 0, breaks[-1], y, col="cyan", ...)
+ }
>
> panel.lm = function(x, y, col=par("col"), bg=NA,
+   pch=par("pch"), cex=1, col.lm="red", ...)
+ {
+   points(x, y, pch=pch, col=col, bg=bg, cex=cex)
+   ok = is.finite(x) & is.finite(y)
+   if (any(ok))
+     abline(lm(y[ok] ~ x[ok]))
+ }
```

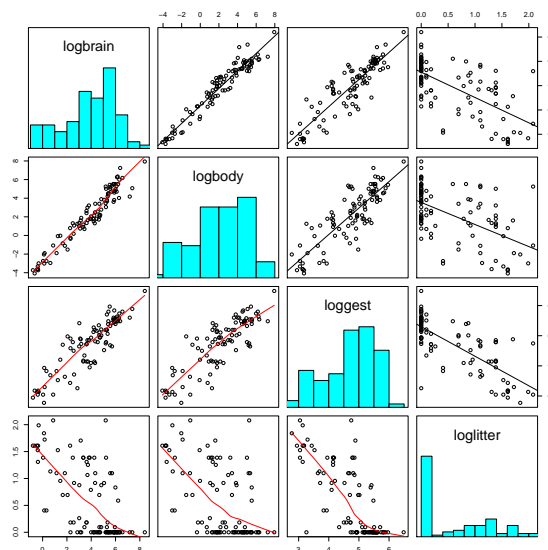
Below is a somewhat fancier pairs plot.

```
> pairs(~ Brain+Body+Gestation+Litter,
+   lower.panel=panel.smooth, diag.panel=panel.hist,
+   upper.panel=panel.lm, data=case0902)
```



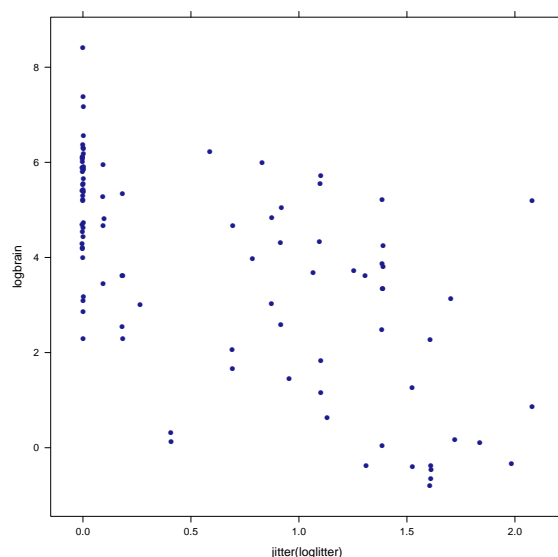
Here is an even fancier pairs plot using the log-transformed variables, akin to Display 9.11 on page 256.

```
> pairs(~ logbrain+logbody+loggest+loglitter,
+       lower.panel=panel.smooth, diag.panel=panel.hist,
+       upper.panel=panel.lm, data=case0902)
```



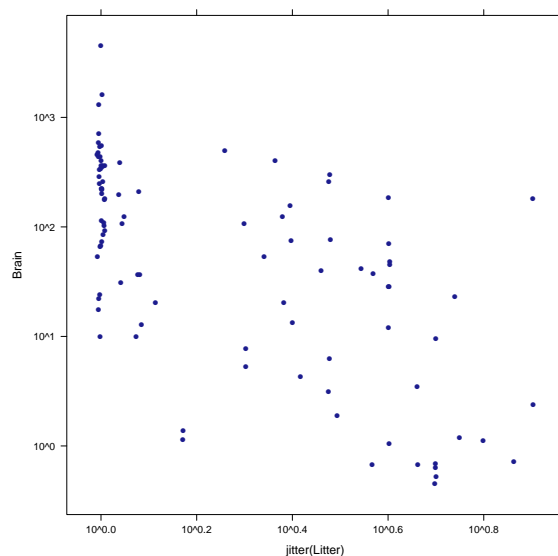
The following displays a jittered scatterplot of log brain weight as a function of log litter size, akin to Display 9.12 on page 258.

```
> xyplot(logbrain ~ jitter(loglitter), data=case0902)
```

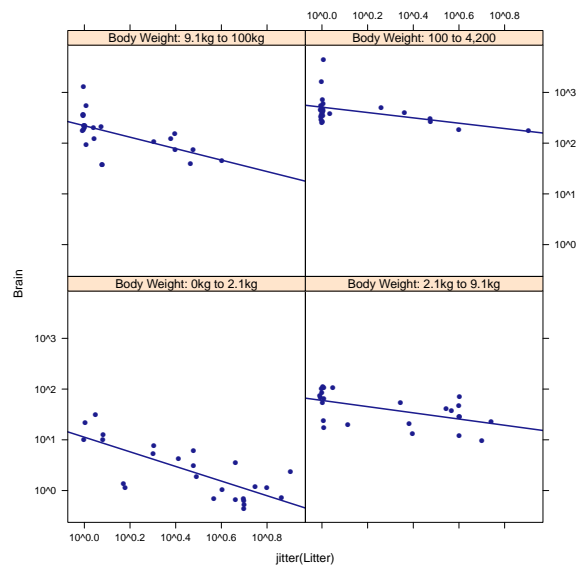
Below displays a jittered scatterplot using the original data on a log-transformed axis, akin to Display 9.12 on page 258.

```
> xyplot(Brain ~ jitter(Litter), scales=list(y=list(log=TRUE),
+                                           x=list(log=TRUE)), data=case0902)
```



The following displays a jittered scatterplot using the original data stratified by body weight on a log-transformed axis, akin to Display 9.13 on page 259.

```
> case0902$weightcut = cut(case0902$Body, breaks=c(0, 2.1, 9.1, 100, 4200), labels=c("Body Wei
> xyplot(Brain ~ jitter(Litter) | weightcut,
+       scales=list(y=list(log=TRUE), x=list(log=TRUE)), type=c("p", "r"), data=case0902)
```



3.3 Multiple linear regression model

The following model is interpreted on page 240 and shown in Display 9.15 (page 260).

```
> lm1 = lm(logbrain ~ logbody+loggest+loglitter, data=case0902)
> summary(lm1)
```

Call:

```
lm(formula = logbrain ~ logbody + loggest + loglitter, data = case0902)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.9541	-0.2964	-0.0311	0.2811	1.5749

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.8548	0.6617	1.29	0.1996
logbody	0.5751	0.0326	17.65	<2e-16
loggest	0.4179	0.1408	2.97	0.0038
loglitter	-0.3101	0.1159	-2.67	0.0089

Residual standard error: 0.475 on 92 degrees of freedom

Multiple R-squared: 0.954, Adjusted R-squared: 0.952

F-statistic: 632 on 3 and 92 DF, p-value: <2e-16