

The Statistical Sleuth in R:

Chapter 9

Kate Aloisio

Ruobing Zhang

Nicholas J. Horton*

January 21, 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	11

1 Introduction

This document is intended to help describe how to undertake analyses introduced as examples in the Second Edition of the *Statistical Sleuth* (2002) 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.amherst.edu/~nhorton/sleuth>.

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, Amherst College, nhorton@amherst.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 **Sleuth2** package.

```
> install.packages('Sleuth2') # note the quotation marks
```

```
> require(Sleuth2)
```

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 and summarizing the variables.

```
> summary(case0901)
```

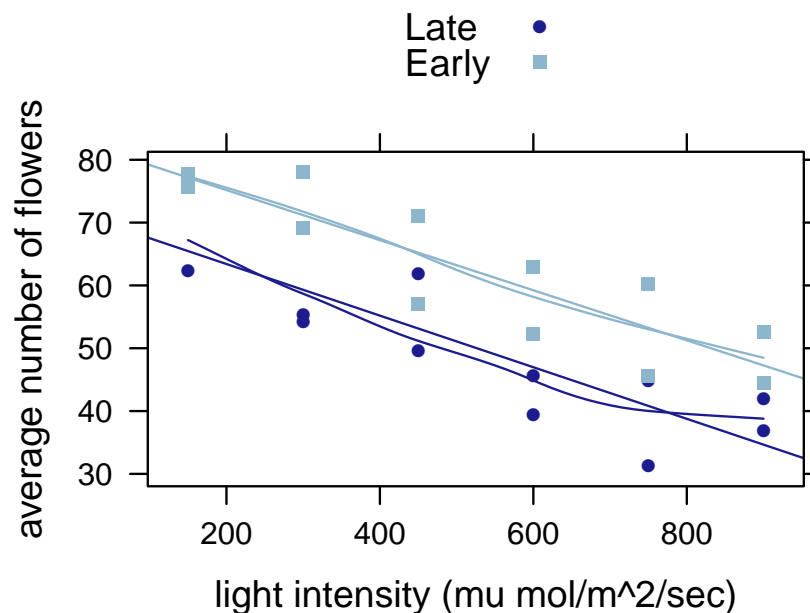
Flowers		Time	Intens	
Min.	:31.3	Late :12	Min.	:150
1st Qu.	:45.4	Early:12	1st Qu.	:300
Median	:54.8		Median	:525
Mean	:56.1		Mean	:525
3rd Qu.	:64.5		3rd Qu.	:750
Max.	:78.0		Max.	:900

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

	Time	min	Q1	median	Q3	max	mean	sd	n	missing
1	150.Late	62.3	66.1	69.9	73.6	77.4	69.9	10.677	2	0
2	300.Late	54.2	54.5	54.8	55.0	55.3	54.8	0.778	2	0
3	450.Late	49.6	52.7	55.8	58.8	61.9	55.8	8.697	2	0
4	600.Late	39.4	41.0	42.6	44.1	45.7	42.6	4.455	2	0
5	750.Late	31.3	34.7	38.1	41.5	44.9	38.1	9.617	2	0
6	900.Late	36.8	38.1	39.4	40.6	41.9	39.4	3.606	2	0
7	150.Early	75.6	76.1	76.7	77.3	77.8	76.7	1.556	2	0
8	300.Early	69.1	71.3	73.5	75.8	78.0	73.5	6.293	2	0
9	450.Early	57.0	60.5	64.0	67.6	71.1	64.0	9.970	2	0
10	600.Early	52.2	54.9	57.6	60.2	62.9	57.6	7.566	2	0
11	750.Early	45.6	49.3	52.9	56.6	60.3	52.9	10.394	2	0
12	900.Early	44.4	46.5	48.5	50.5	52.6	48.5	5.798	2	0
13	Late	31.3	41.3	47.6	56.9	77.4	50.1	12.919	12	0
14	Early	44.4	52.5	61.6	72.2	78.0	62.2	12.117	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 237 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 238.

```
> xyplot(Flowers ~ Intens, 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 237.

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

Call:
lm(formula = Flowers ~ Intens + 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)	71.30583	3.27377	21.78	6.8e-16
Intens	-0.04047	0.00513	-7.89	1.0e-07
TimeEarly	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)	64.4977	78.1140
Intens	-0.0511	-0.0298
TimeEarly	6.6899	17.6268

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 256) and interpreted on page 237.

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

Call:
lm(formula = Flowers ~ Intens * 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)	71.62333	4.34330	16.49	4.1e-13
Intens	-0.04108	0.00744	-5.52	2.1e-05
TimeEarly	11.52333	6.14236	1.88	0.075
Intens:TimeEarly	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
Length:96	Min. : 0	Min. : 0	Min. : 16
Class :character	1st Qu.: 13	1st Qu.: 2	1st Qu.: 63
Mode :character	Median : 74	Median : 9	Median :134
	Mean : 219	Mean : 108	Mean :151
	3rd Qu.: 260	3rd Qu.: 95	3rd Qu.:226
	Max. :4480	Max. :2800	Max. :655

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

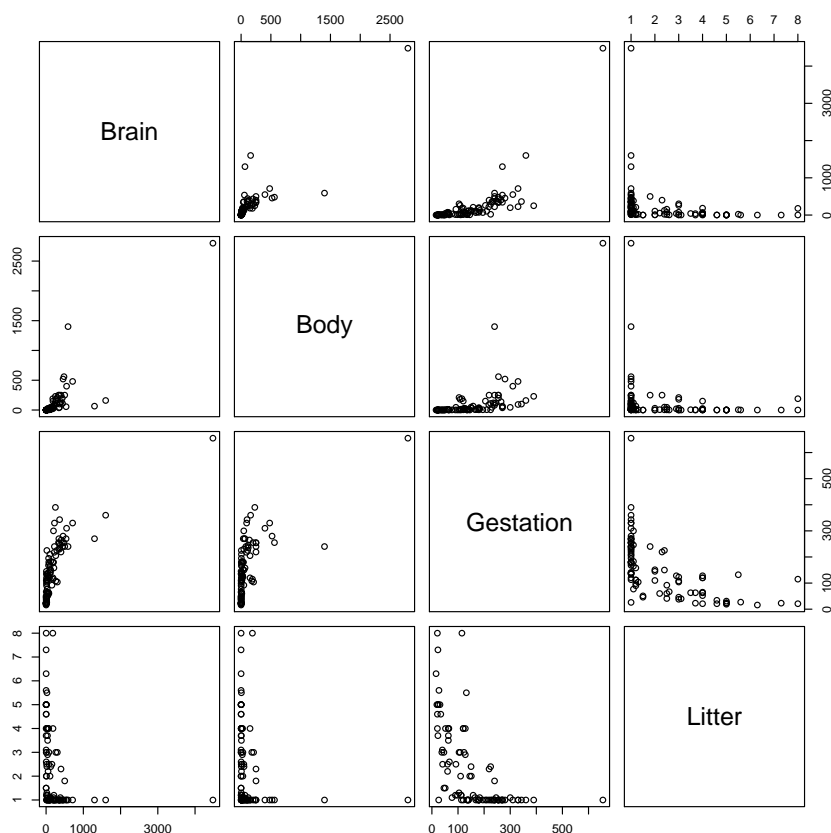
```
Mean    :0.598
3rd Qu.:1.162
Max.    :2.079
```

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 239 of the *Sleuth*).

3.2 Graphical presentation

The following displays a simple (unadorned) pairs plot, akin to Display 9.10 on page 252.

```
> pairs(case0902[c("Brain", "Body", "Gestation", "Litter")])
```



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]))
+ }

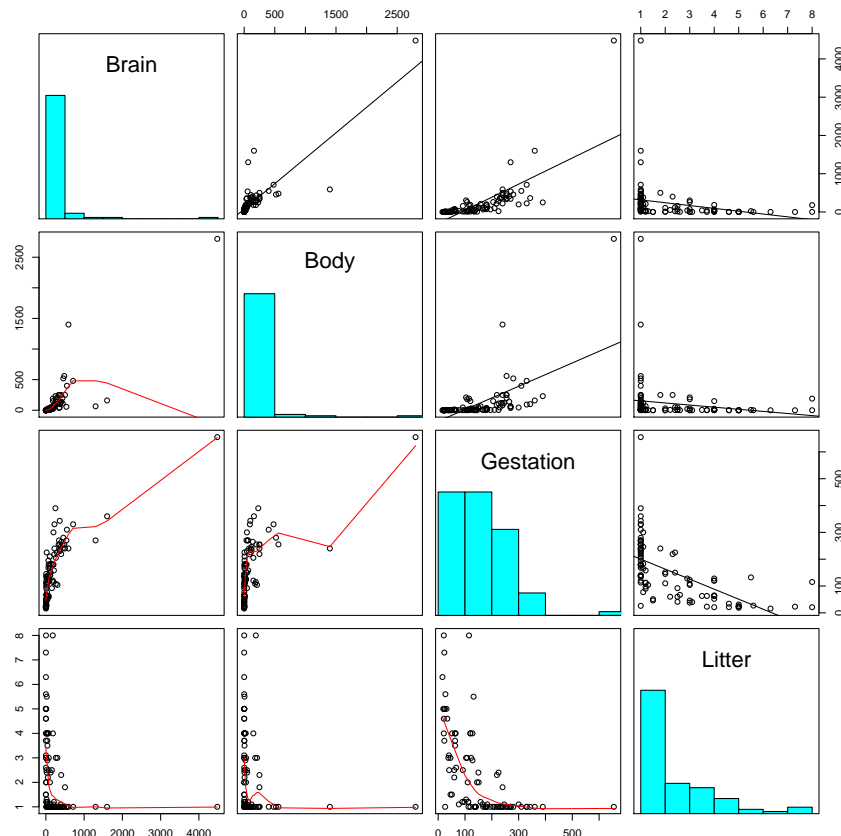
```

Below is a somewhat fancier pairs plot.

```

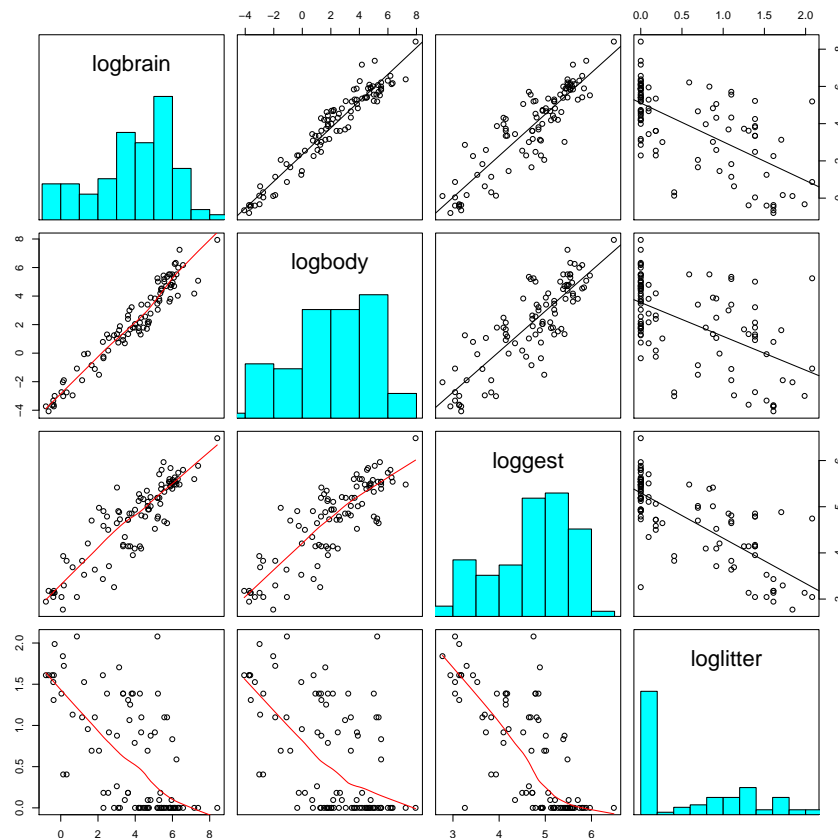
> pairs(~ Brain+Body+Gestation+Litter,
+   lower.panel=panel.smooth, diag.panel=panel.hist,
+   upper.panel=panel.lm, data=case0902)

```



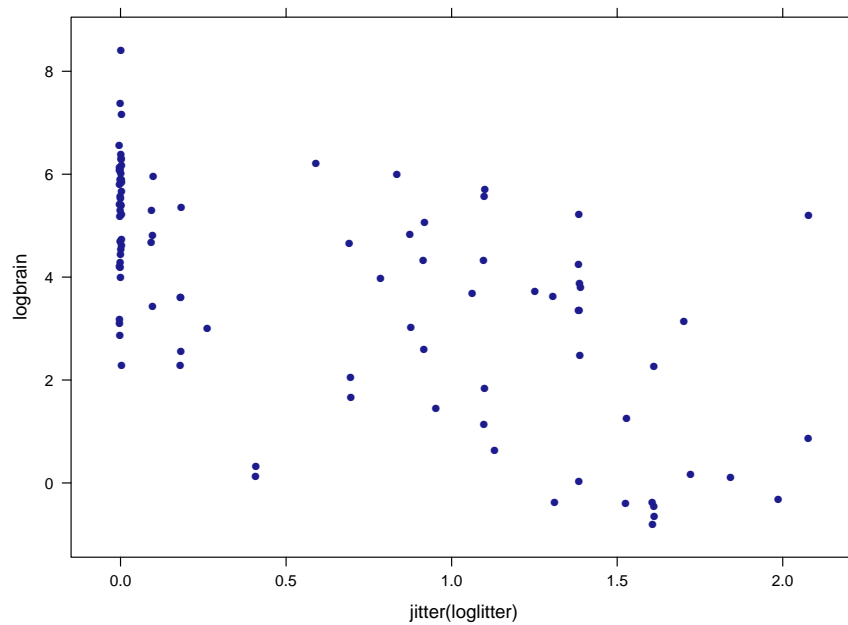
Here is an even fancier pairs plot using the log-transformed variables, akin to Display 9.11 on page 253.

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



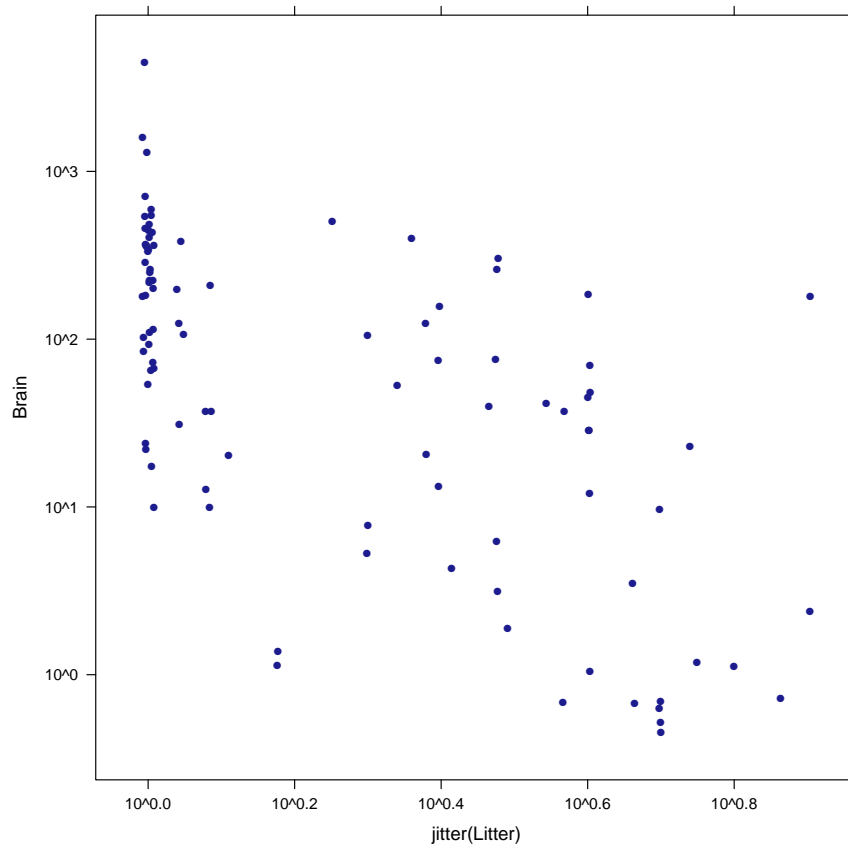
The following displays a jittered scatterplot of log brain weight as a function of log litter size, akin to Display 9.12 on page 254.

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

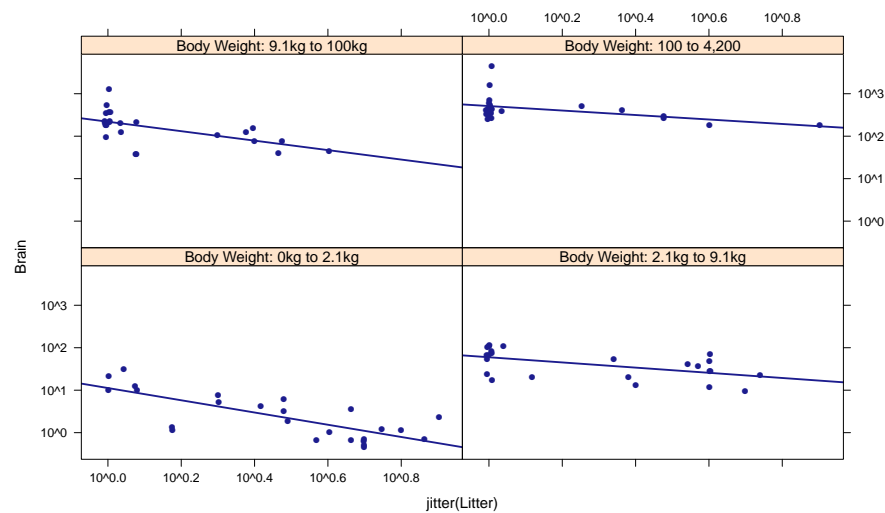
Below displays a jittered scatterplot using the original data on a log-transformed axis, akin to Display 9.12 on page 254.

```
> 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 255.

```
> case0902$weightcut = cut(case0902$Body, breaks=c(0, 2.1, 9.1, 100, 4200),
+   labels=c("Body Weight: 0kg to 2.1kg", "Body Weight: 2.1kg to 9.1kg",
+   "Body Weight: 9.1kg to 100kg", "Body Weight: 100 to 4,200"))
> 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 238 and shown in Display 9.15 (page 256).

```
> 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