Chapter 9: MLR models with examples

Mulitple linear regression (MLR) GOALS:

- Find a good fitting model for the mean response $\mu\{Y|X_1, X_2, ...\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...$
- State the scientific questions of interest in terms of the model parameters $\beta_0, \beta_1, \beta_2, ...$
- Estimate the parameters with available data via least squares regression
- Employ appropriate inferential tools (tests and CIs) for answering the questions of interest and for expressing the uncertainty in the answers.

Four basic MLR models form a roadmap for Chapter 9

- **Model I.** (9.3.2) A model of a response Y as a function of a predictor (or covariate or explanatory variable or independent continuous variable X_1) and a factor (a categorical variable X_2) that has only two levels $X_2 = 0$ or $X_2 = 1$):

$$\mu\{Y|X_1, X_2\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{1}$$

  This is an *equal slopes* model, because the two lines being estimated have the same slope.

  The R-code to fit this model to data is

  `lm(Y ~ X1 + as.factor(X2)).`

  The `as.factor()` function is not required if the levels (or categories) of `X2` are character strings.

- **Model II.** (9.3.4) A model of a response Y with a predictor (X_1) and a factor (X_2) that has only two levels ($X_2 = 0$ or $X_2 = 1$):

$$\mu\{Y|X_1, X_2\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \tag{2}$$

  This is a *separate slopes* model, because the two lines being estimated are allowed to have different slopes.

  The R-code to fit this model to data is

  `lm(Y ~ X1 * as.factor(X2)).`

- **Model III.** (9.2.1) A model of a response Y with two predictors (X_1 and X_2):

$$\mu\{Y|X_1, X_2\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{3}$$

  The R-code to fit this model to data is

  `lm(Y ~ X1 + X2).`

- **Model IV.** We can add an interaction term to Model III as well:

$$\mu\{Y|X_1, X_2\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

  In R:

  `lm(Y ~ X1 * X2).`

Assumptions

To reinforce that the assumptions of an MLR are the same as for SLR, we can rewrite Model I as

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

where $\varepsilon \sim N(0, \sigma)$. In fact, any of the Models I-IV, and any other MLR model, can be rewritten as

$$y = \mu\{Y|X_1, X_2, ...\} + \varepsilon$$

where $\varepsilon \sim N(0, \sigma)$. Therefore, the MLR assumptions are:

- **normality**: the residuals in an MLR are normally distributed about the model $\mu\{Y|X_1, X_2\}$
- **constant variance**: the residuals in an MLR have constant variance σ^2
- **linearity**: the residuals have zero mean
- **random sample**: the residuals are independent

We will check these assumptions using graphical approaches just as we did for ANOVA and SLR.

9.3.2 An example of Model I

9.1.1 Effects of light on flowers

Meadowfoam (*Limananthes alba*) has a unique seed oil that is similar to the oil from sperm whales (long carbon chains) in that it is non-greasy and stable. A randomized experiment was conducted to explore the relationship between light intensity (that your book calls X_1 on p. 242, with 6 intensities investigated: 150, 300, 450, 600, 750, 900 $\mu mol/m^2/sec$); and the timing of the onset of the light treatment (X_2) at either photoperiodic floral induction (PFI) ($X_2 = 1$ when Timing = "at-PFI") or 24 days prior to PFI ($X_2 = 2$ when Timing = "before-PFI").

There were $6 \times 2 = 12$ treatment combinations. There were $n_i = 2$ randomly assigned replications to each treatment combination. The response variable is the number of flowers per plant (Y).

Your book (p. 248) writes Model I (equation (1)) as

$$\mu\{\text{Flowers}|\text{Intensity}, \text{Time}\} = \beta_0 + \beta_1 \text{Intensity} + \beta_2 \text{Dummy}_2(\text{Time}).$$

The function $\text{Dummy}_L()$, which we also will write as $\text{D}_L()$ to save space, is also called an indicator function or a delta function. It converts the levels of any categorical variable to 0's and 1's. The level that is equal to L is assigned a 1, all other levels are assigned to 0. For example,

$$\text{D}_2(\text{Time}) = \begin{cases} 1 & \text{if Time} = 2 \\ 0 & \text{otherwise (i.e., if Time} = 1) \end{cases}$$

An equivalent model could be written with respect to the categorical variable Timing as

$$\mu\{\text{Flowers}|\text{Intensity}, \text{Timing}\} = \beta_0 + \beta_1 \text{Intensity} + \beta_2 \text{D}_{\text{before}-\text{PFI}}(\text{Timing}).$$

so now $\text{D}_{\text{before}-\text{PFI}}(\text{Timing}) = \begin{cases} 1 & \text{if Timing} = \text{"before} - \text{PFI"} \\ 0 & \text{otherwise (i.e., if Timing} = \text{"at} - \text{PFI"}) \end{cases}$

The last model allows us to write out the equation for the line for the timing condition Timing = "at-PFI"

$$\mu\{\text{Flowers}|\text{Intensity}, \text{Timing} = \text{"at} - \text{PFI"}\} = \beta_0 + \beta_1 \text{Intensity}.$$

A second line with the same slope (β_1) for the timing condition Timing = "before-PFI" is

$$\mu\{\text{Flowers}|\text{Intensity}, \text{Timing} = \text{“before} - \text{PFI”}\} = (\beta_0 + \beta_2) + \beta_1 \text{Intensity}$$

Unfortunately, it is common to abuse notation and drop the notation for the Dummy variable Dummy() when writing an MLR model. It is understood that if you include a factor X_2 into a MLR, what you really mean to add to the model is Dummy(X_2). Be careful with R's `lm()` output when using factors because:

- if your factor levels have numeric labels, R may include it as a covariate instead of a factor unless you use the `as.factor()` function.
- R uses a Dummy variable that sets the level that comes alpha-numerically first to 0 just as in the example above. You can reset the reference or base level using the `relevel()` command that we used earlier.

QUESTIONS OF INTEREST:

1. Do differences in intensity affect flowering production? State this question in terms of the parameters in the model above.

2. Does timing affect production? State this question in terms of the parameters in the model above.

Let's see about answering these questions after fitting the model $\mu\{\text{Flowers}|\text{Intensity}, \text{Time}\} = \beta_0 + \beta_1 \text{Intensity} + \beta_2 \text{Time}$ to the data.

```
library(Sleuth3)
source("http://www.math.montana.edu/parker/courses/STAT411/diagANOVA.r")
d1=case0901
summary(d1)
```

```
##     Flowers           Time       Intensity
##  Min.   :31.30   Min.   :1.0   Min.   :150
##  1st Qu.:45.42   1st Qu.:1.0   1st Qu.:300
##  Median :54.75   Median :1.5   Median :525
##  Mean   :56.14   Mean   :1.5   Mean   :525
##  3rd Qu.:64.45   3rd Qu.:2.0   3rd Qu.:750
##  Max.   :78.00   Max.   :2.0   Max.   :900
```

```
dim(d1)   # n=24
```

```
## [1] 24  3
```

```
# Let's add a categorical variable called Timing
d1$Timing=character(24)
d1$Timing[d1$Time==1]="at-PFI"
d1$Timing[d1$Time==2]="before-PFI"
d1$Timing = as.factor(d1$Timing)

# Look at rows 1-3, 12-15 of the data
d1[c(1:3,12:15),]
```

```
##    Flowers Time Intensity     Timing
## 1     62.3    1       150     at-PFI
## 2     77.4    1       150     at-PFI
```

```
## 3     55.3    1       300     at-PFI
## 12    41.9    1       900     at-PFI
## 13    77.8    2       150 before-PFI
## 14    75.6    2       150 before-PFI
## 15    69.1    2       300 before-PFI
```

```r
# A scatterplot
plot(Flowers ~ Intensity,pch=Time,col=Time,data=d1)

# Let's fit Model I to the data wrt the factor Time
m1 = lm(Flowers ~ Intensity + as.factor(Time),data=d1)
summary(m1)
```

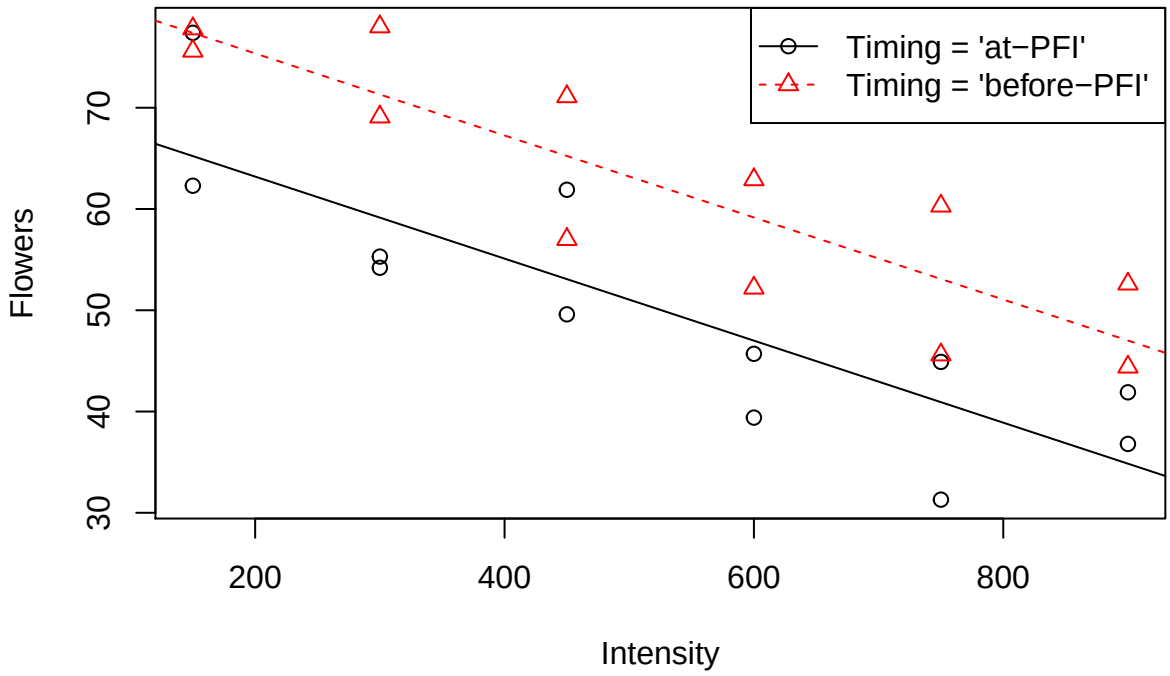
```
##
## Call:
## lm(formula = Flowers ~ Intensity + as.factor(Time), data = d1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.652 -4.139 -1.558  5.632 12.165
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       71.305833   3.273772  21.781 6.77e-16 ***
## Intensity         -0.040471   0.005132  -7.886 1.04e-07 ***
## as.factor(Time)2 12.158333   2.629557   4.624 0.000146 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.441 on 21 degrees of freedom
## Multiple R-squared:  0.7992, Adjusted R-squared:   0.78
## F-statistic: 41.78 on 2 and 21 DF,  p-value: 4.786e-08
```

```r
# Here's Model I again, but now wrt the factor Timing
m2=lm(Flowers ~ Intensity + Timing,data=d1)
summary(m2)
```

```
##
## Call:
## lm(formula = Flowers ~ Intensity + Timing, data = d1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.652 -4.139 -1.558  5.632 12.165
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        71.305833   3.273772  21.781 6.77e-16 ***
## Intensity          -0.040471   0.005132  -7.886 1.04e-07 ***
## Timingbefore-PFI 12.158333   2.629557   4.624 0.000146 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.441 on 21 degrees of freedom
## Multiple R-squared:  0.7992, Adjusted R-squared:   0.78
```

```
## F-statistic: 41.78 on 2 and 21 DF,  p-value: 4.786e-08
```

```r
# Add the model fit to the scatterplot
abline(71.3,-0.0405,col=1,lty=1)
abline(71.3+12.16,-0.0405,col=2,lty=2)
legend("topright",legend=c("Timing = 'at-PFI'","Timing = 'before-PFI'"),
       pch=c(1,2),lty=c(1,2),col=c(1,2))
```



The assumptions for MLR are the same as the assumptions for SLR. The plots from the `diagANOVA()` command below assess these assumptions (i.e., model fit).

QUESTIONS:

1. Do the MLR assumptions appear to be satisfied?

2. Do differences in intensity affect flowering production?

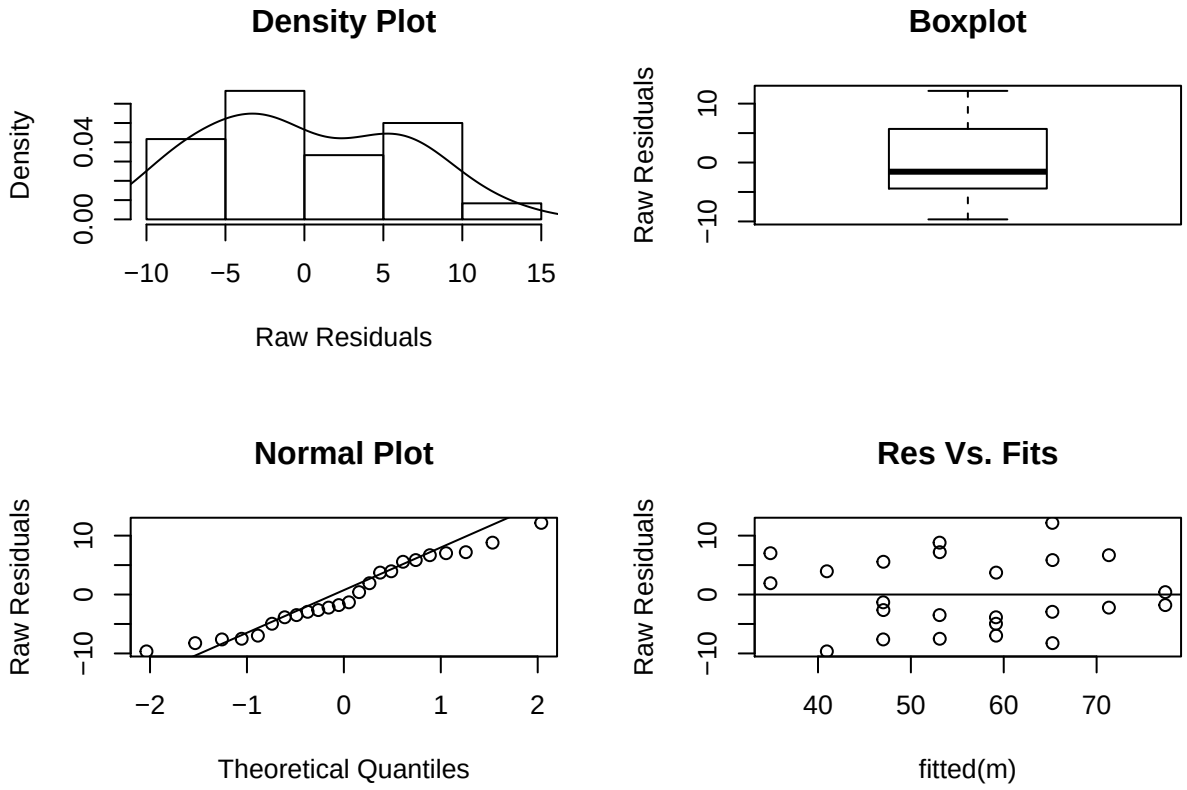
3. Does timing affect production?

4. Give an estmate of σ, the constant SD

5. Give the value of R^2 as a quantitiative measure of the model's fit.

6. Does this MLR fit better than the null model $\mu\{Flowers|Intensity, Timing\} = \beta_0$?

```
diagANOVA(m2)
```



9.3.4 An example of Model II

For the flowering experiment, what if we wanted to answer the question:

Does the change in flowering production as a function of intensity depend on the timing?

Another way to ask this question is whether there is an interaction between intensity and timing that affects flowering. Let us consider Model II to estimate the interaction.

Your book (p. 250) writes Model II (equation (2)) as

$$\mu\{\text{Flowers}|\text{Intensity}, \text{Timing}\} = \beta_0 + \beta_1 \text{Intensity} + \beta_2 \text{D}_{\text{before}-\text{PFI}}(\text{Timing}) + \beta_3 \text{Intensty} \times \text{D}_{\text{before}-\text{PFI}}(\text{Timing}).$$

This model allows us to write the equation for a line for the timing condition at Timing = "at-PFI" as:

$$\mu\{\text{Flowers}|\text{Intensity}, \text{Timing} = \text{“at} - \text{PFI”}\} = \beta_0 + \beta_1 \text{Intensity}$$

The second line for the timing condition at Timing = "before-PFI" is

$$\mu\{\text{Flowers}|\text{Intensity}, \text{Timing} = \text{“before} - \text{PFI”}\} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\text{Intensity}$$

QUESTION: Which parameter is of interest to answer the queston: Does the change in flowering production as a function of intensity depend on the timing?

Let's fit Model II to the flowering data:

```
m3 = lm(Flowers ~ Intensity*Timing,data=d1)
summary(m3)
```

```
##
## Call:
## lm(formula = Flowers ~ Intensity * Timing, data = d1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.516 -4.276 -1.422  5.473 11.938
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 71.623333   4.343305  16.491 4.14e-13 ***
## Intensity                   -0.041076   0.007435  -5.525 2.08e-05 ***
## Timingbefore-PFI            11.523333   6.142360   1.876   0.0753 .
## Intensity:Timingbefore-PFI  0.001210   0.010515   0.115   0.9096
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.598 on 20 degrees of freedom
## Multiple R-squared:  0.7993, Adjusted R-squared:  0.7692
## F-statistic: 26.55 on 3 and 20 DF,  p-value: 3.549e-07
```

```
# A scatterplot with model fit - doesn't look any different than scatterplot on p.5
#plot(Flowers ~ Intensity,pch=Time,col=Time,data=d1)
#abline(71.6,-0.0411,col=1,lty=1)
#abline(71.6+11.52,-0.0411+.0012,col=2,lty=2)
#legend("topright",legend=c("Timing = 'at-PFI'","Timing = 'before-PFI'"),
#       pch=c(1,2),lty=c(1,2),col=c(1,2))
#diagANOVA(m3)

# Compare the two models using an extra sum of squares test
anova(m2,m3)
```

```
## Analysis of Variance Table
##
## Model 1: Flowers ~ Intensity + Timing
## Model 2: Flowers ~ Intensity * Timing
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     21 871.24
## 2     20 870.66  1   0.57604 0.0132 0.9096
```

QUESTIONS:

1. Does the change in flowering production as a function of intensity depend on the timing?

2. Report the R^2 value for this last Model II fit to the data. How does it compare to the Model I fit to the data?

3. Report and interpret the extra sum of squares test.

An example of MLR Model III

9.1.2 Why do some mammals have larger brain size?

Data on brain weight (grams), body weight (kilograms), litter size, and gestation (i.e., length of pregnancy in days) were available on 96 mammals (see Display 9.4 on page 241). Brain weight is the response variable and the other three variables are explanatory variables. Since brain size is obviously related to to body size, the question of interest is: *Are gestation length and/or litter size associated with brain weight after accounting for body size?*

```
# Get the data
summary(case0902)
```

```
##              Species       Brain             Body
##  Aardvark        : 1   Min.   :   0.45   Min.   :   0.017
##  Acouchis        : 1   1st Qu.:  12.60   1st Qu.:   2.075
##  African elephant: 1   Median :  74.00   Median :   8.900
##  Agoutis         : 1   Mean   : 218.98   Mean   : 108.328
##  Axis deer       : 1   3rd Qu.: 260.00   3rd Qu.:  94.750
##  Badger          : 1   Max.   :4480.00   Max.   :2800.000
##  (Other)         :90
##    Gestation         Litter
##  Min.   : 16.0   Min.   :1.00
##  1st Qu.: 63.0   1st Qu.:1.00
##  Median :133.5   Median :1.20
##  Mean   :151.3   Mean   :2.31
##  3rd Qu.:226.2   3rd Qu.:3.20
##  Max.   :655.0   Max.   :8.00
##
```

```
dim(case0902)   # n=96 mammals
```

```
## [1] 96  5
```

```
# Look at first 10 rows
case0902[1:10,]
```

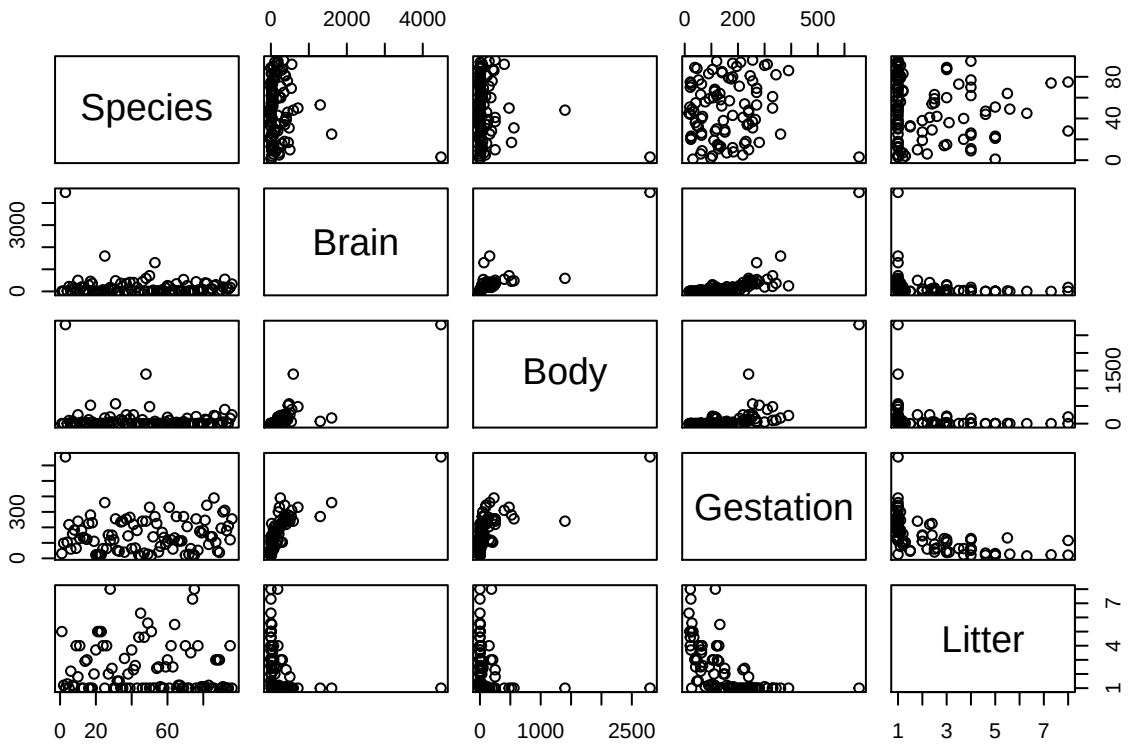
```
##              Species  Brain    Body Gestation Litter
## 1           Aardvark    9.6    2.20        31    5.0
## 2           Acouchis    9.9    0.78        98    1.2
## 3  African elephant 4480.0 2800.00       655    1.0
## 4           Agoutis   20.3    2.80       104    1.3
## 5         Axis deer  219.0   89.00       218    1.0
## 6            Badger   53.0    6.00        60    2.2
## 7     Barbary sheep  210.0   66.00       158    1.2
## 8      Barking deer  124.0   16.00       183    1.1
## 9     Bat-eared fox   28.5    3.20        65    4.0
## 10     Beaked whale  500.0  250.00       240    1.8
```

9.5.1 Graphical investigation by a matrix plot

A *matrix plot* is a single view of all possible pairwise scatterplots from a set of variables. The pairwise relationships observed in a matrix plot do not necessarily indicate the simultaneous effect of the exlanatory variables on the response. Nonetheless, a matrix plot is very useful for suggesting possible transformations of the variables to include in the regression model. The model fit must still be assessed by residual plots.

A first matrix plot for the Brain size data:

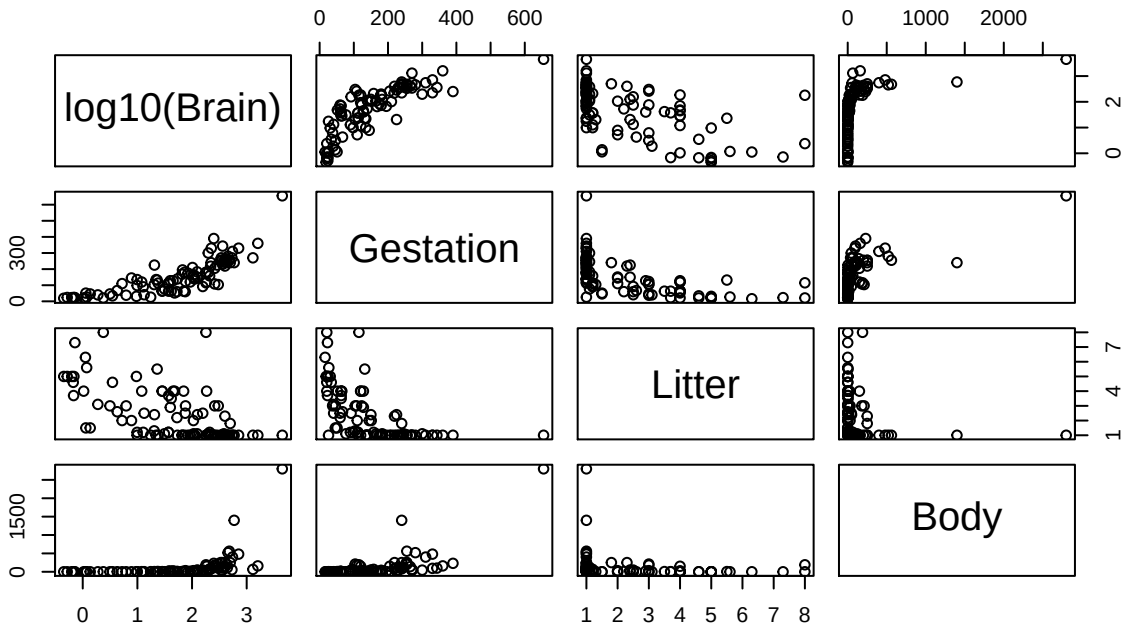
```
# A matrix plot - compare with Display 9.10
pairs(case0902)
```



The data are bunched up for small values of Brain size with an increasing variance as Body Size and Gestation increase. This suggests a log-transform of Brain size, probably a log-transform of some of the other explanatory

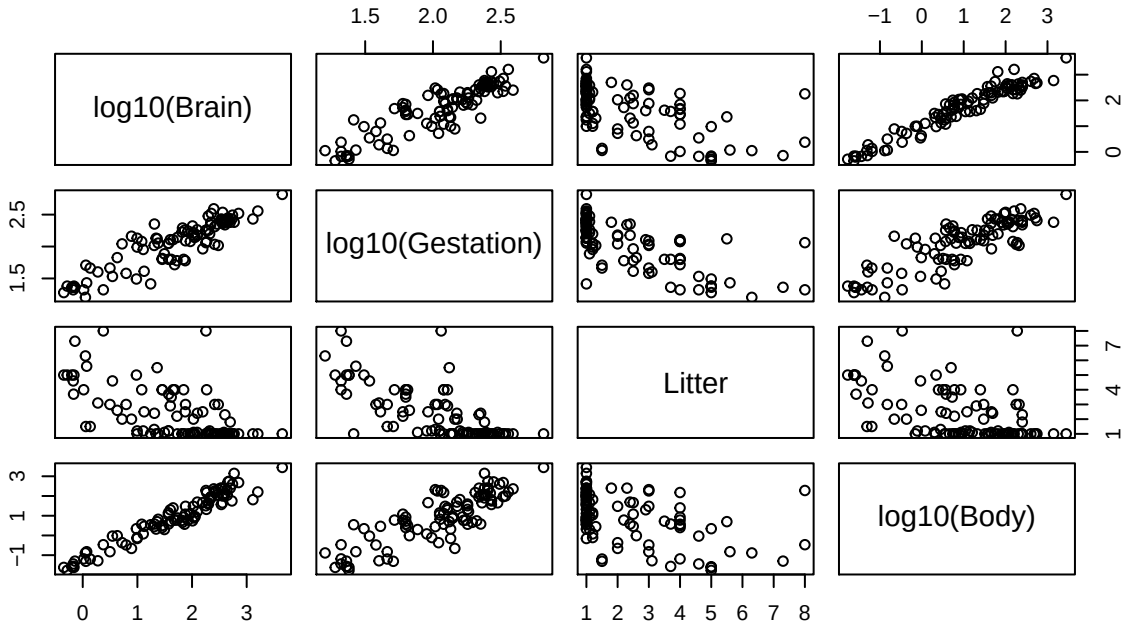
variables as well. To start, let's $\log_{10}$-transform Brain size only then use another matrix plot (leaving out Species this time) to assess:

```
pairs(log10(Brain) ~ Gestation + Litter + Body,data=case0902)
```



This last matrix plot indicates a non-linear reltionship between the mean log-Brain size and each of Gestation and Body size. Consider log-transforms of these explanatory variables next:

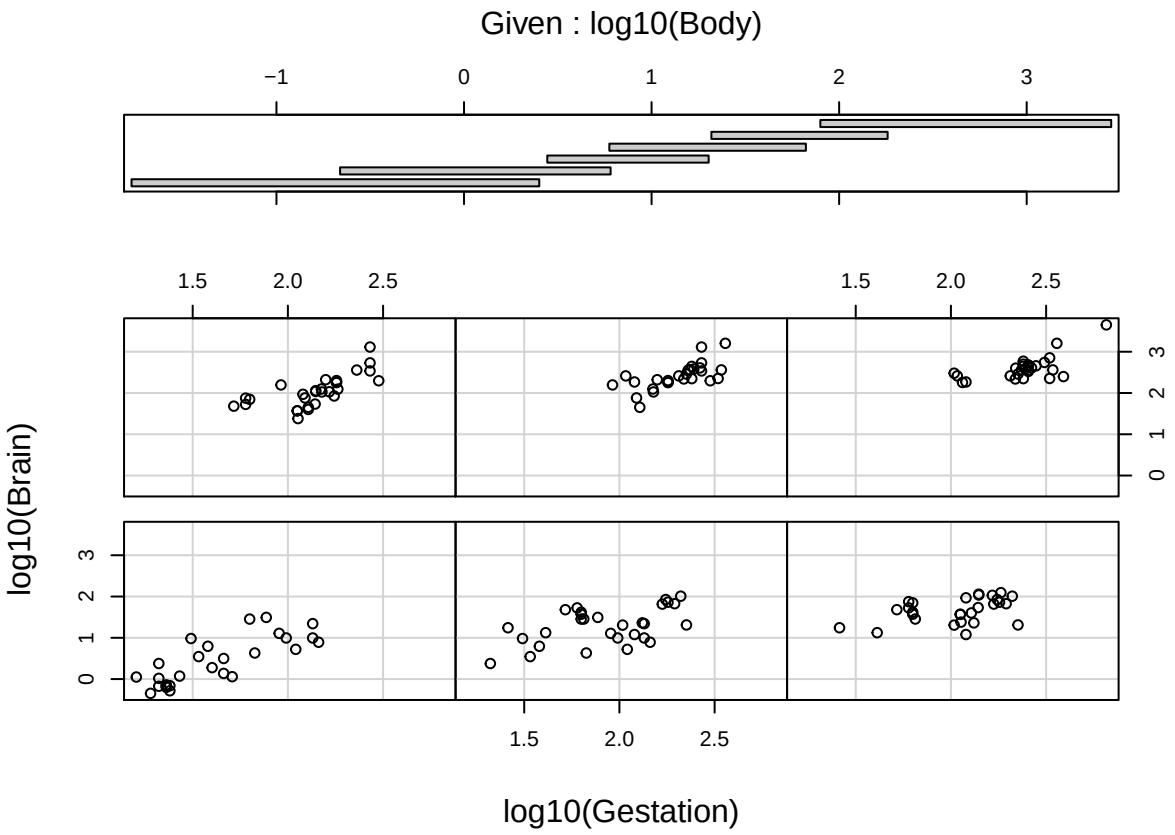
```
pairs(log10(Brain) ~ log10(Gestation) + Litter + log10(Body),data=case0902)
```



9.5.4 Graphical investigation by Trellis plot

Three dimensions are needed to view the simultaneous effect of Gestation and Body size on the response Brain size. One way to visualize this 3 dimensional relationship is to use a *Trellis plot*. The Trellis plot shown next breaks Body size into different chunks where each chunk is like a level of a factor. For each chunk, a different pane shows Brain size as a function of Gestation.

```
# A Trellis graph - compare with Display 9.13
coplot(log10(Brain) ~ log10(Gestation)|log10(Body),data=case0902)
```



From p. 257: The Trellis plot suggests that there's a positive relationship between Brain size and Gestation even after accounting for Body size. **Very importantly, the slope between log-Brain size and log-Gestation in each pane appears to be about the same, which fails to suggest an interaction between Gestation and Body size.** We will investigate more critically whether there is an interaction for these data later.

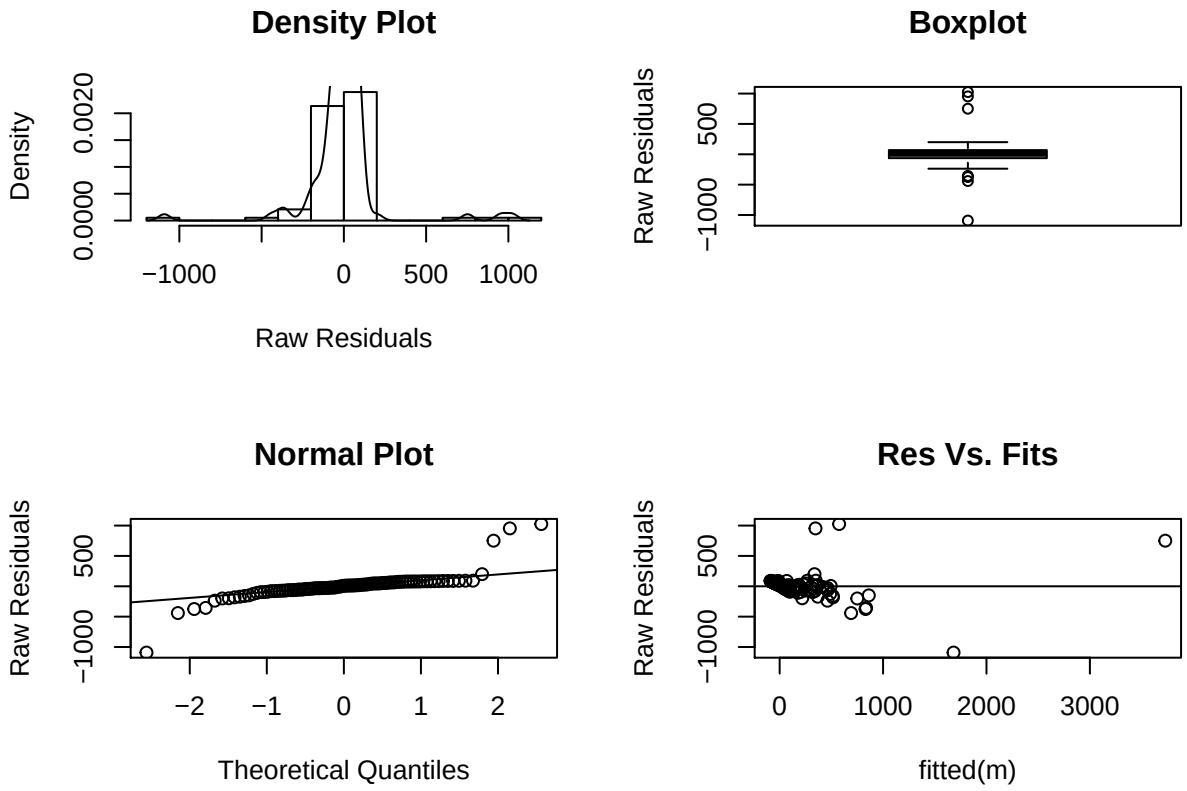
9.2.1 Fitting Model III

Armed with the graphical assessment above, we already know that fitting Model III to Brain size as a linear function of Gestation (X_1) and body size (X_2) (as your book does on bottom of page 245) results in a bad fit. But it will be instructive to try anyways. We'll throw Litter size into the mix later:

$$\mu\{\text{Brain}|\text{Gestation}, \text{Body}\} = \beta_0 + \beta_1 \text{Gestation} + \beta_2 \text{Body}$$

Let's fit this MLR and assess residual plots:

```
m.BAD = lm(Brain ~ Gestation + Body,data=case0902)
diagANOVA(m.BAD)
```



QUESTION: Which MLR assumptions appear violated in the above normal probability and residual plots?

So both the residual plots and the matrix plots suggest a deviation from MLR assumptions. Unlike the residual plots however, the matrix plot suggests the next MLR to try: log-transforming Brain size, Gestation and Body Size:

$$\mu\{\log_{10}(\text{Brain})|\text{Gestation}, \text{Body}\} = \beta_0 + \beta_1 \log_{10}(\text{Gestation}) + \beta_2 \log_{10}(\text{Body}) \tag{4}$$

For any fixed Body size, this shows a linear relationship between the mean log-Brain size and log-Gestation. Re-arranging the terms a little we get the "$y = b + mx$ form" of the line:

$$\mu\{\log_{10}(\text{Brain})|\text{Gestation}, \text{Body}\} = (\beta_0 + \beta_2 \log_{10}(\text{Body})) + \beta_1 \log_{10}(\text{Gestation}).$$

The y-intercept $= \beta_0 + \beta_2 \log_{10}(\text{Body})$ but the slope is a constant β_1 regardless of Body size.

QUESTIONS OF INTEREST:

1. Is gestation length associated with brain size after accounting for body size? State this question in terms of parameter(s) in the model above.

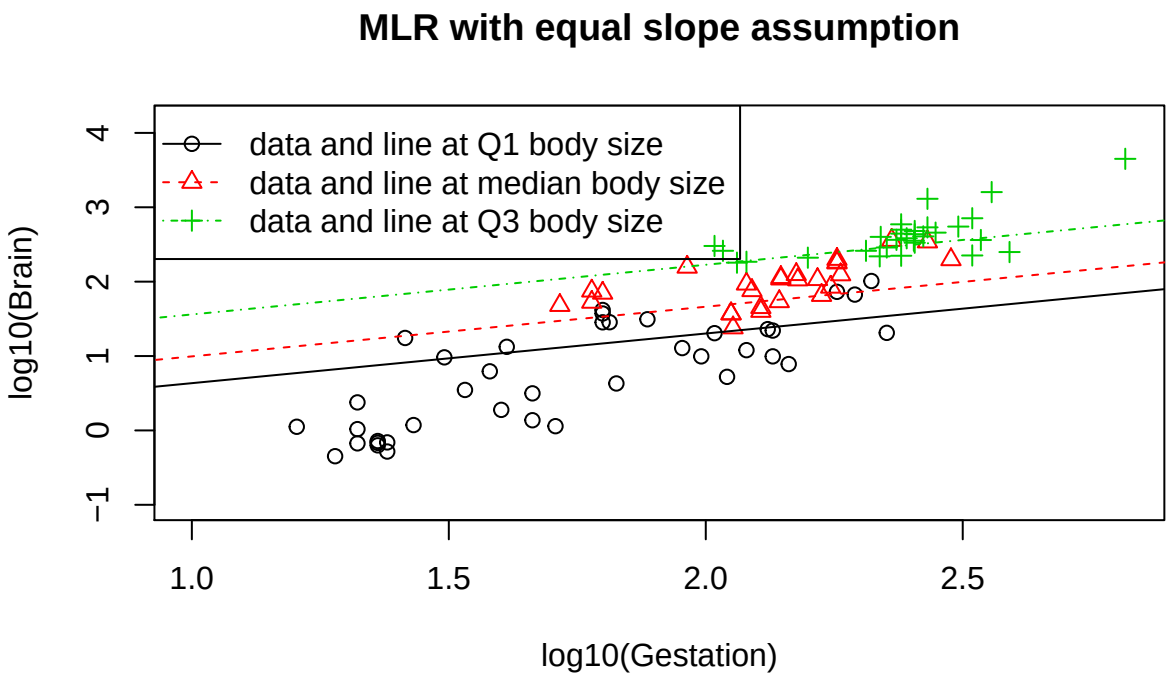
2. Is body size associated with brain size after accounting for Gestation length? State this question in terms of parameter(s) in the model above.

Let's see about answering these questions after fitting Model III in equation (4) to the data:

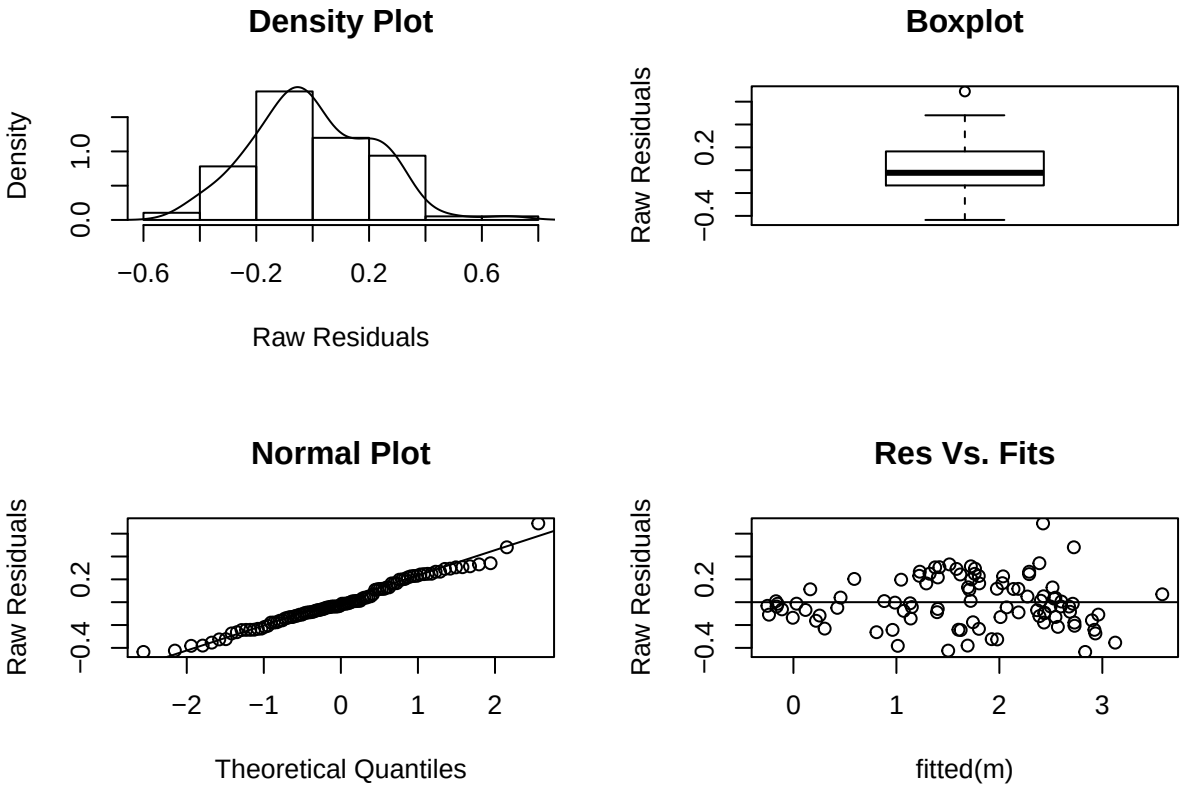
```
# Fit the model
m3 = lm(log10(Brain) ~ log10(Gestation) + log10(Body),data=case0902)
summary(m3)
```

```
##
## Call:
## lm(formula = log10(Brain) ~ log10(Gestation) + log10(Body), data = case0902)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43554 -0.13190 -0.02276  0.16438  0.68961
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.19860    0.19912  -0.997    0.321
## log10(Gestation)  0.66782    0.10875   6.141    2e-08 ***
## log10(Body)       0.55117    0.03236  17.033   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2129 on 93 degrees of freedom
## Multiple R-squared:  0.9501, Adjusted R-squared:  0.949
## F-statistic: 885.2 on 2 and 93 DF,  p-value: < 2.2e-16
```

```
# Lets try to visualize the model fit to the data
WtQ1 = case0902$Body<=5.5
WtMedian = case0902$Body>5.5&case0902$Body<=52
WtQ3 = case0902$Body>52
plot(log10(Brain) ~ log10(Gestation),data=case0902[WtQ1,],pch=1,col=1,xlim =
       c(1,log10(660)),ylim=c(-1,log10(4600)+.5),main="MLR with equal slope assumption")
points(log10(Brain) ~ log10(Gestation),data=case0902[WtMedian,],pch=2,col=2)
points(log10(Brain) ~ log10(Gestation),data=case0902[WtQ3,],pch=3,col=3)
gest = 0:2800
bod = 2   # 25th percentile for body size
lines(log10(gest),-0.19860 + 0.66782*log10(gest) + 0.55117*log10(bod),col=1,lty=1)
bod = 9   # Median for body size
lines(log10(gest),-0.19860 + 0.66782*log10(gest) + 0.55117*log10(bod),col=2,lty=2)
bod = 95   # 75th percentile for body size
lines(log10(gest),-0.19860 + 0.66782*log10(gest) + 0.55117*log10(bod),col=3,lty=4)
legend("topleft",legend=c("data and line at Q1 body size","data and line at median body size","data and
```




```
# Assess model fit
diagANOVA(m3)
```



The paralleism between the lines in the visualization in the scatterplot above is a consequence of using the equal slopes model. Because of the positive association (as expected) between Body size and Brain size, the y-intercepts of the lines increase as Body size increases. The colored lines are actually contour lines that help us visualize the plane that defines the 3D relationship between log-Brain size, log-Gestation and log-Body size.

QUESTIONS:

1. Do the assumptions for the MLR in equation (4) appear to be satisfied?

2. Is gestation length associated with median brain size after accounting for body size?

3. Is body size associated with median brain size after accounting for Gestation length?

4. Report the results of the extra sum of squares test that compares the MLR to the single mean null model.

5. Give an estmate of σ, the constant SD

6. Give the value of R^2 as a quantitiative measure of the model's fit.

An example of Model IV - including an interaction

For the observational study of Brain size, Model IV is useful if we wanted to answer the question:

Does the change in median Brain size as a function of Gestation depend on the Body size?

Another way to ask this question is whether there is an interaction between Gestation and Body size that affects the median Brain size. Model IV allows us to estimate this interaction:

$$\mu\{\log_{10}(\text{Brain})|\text{Gestation}, \text{Body}\} = \beta_0 + \beta_1 \log_{10}(\text{Gestation}) + \beta_2 \log_{10}(\text{Body}) + \beta_3 \log_{10}(\text{Gestation}) \times \log_{10}(\text{Body})$$

Let's rearrange the terms in this equation so that it is clear how Body size affects the slope and y-intercept:

$$\mu\{\log_{10}(\text{Brain})|\text{Gestation}, \text{Body}\} = (\beta_0 + \beta_2 \log_{10}(\text{Body})) + (\beta_1 + \beta_3 \log_{10}(\text{Body})) \times \log_{10}(\text{Gestation}).$$

QUESTION: Which parameter is of interest to answer the queston: Does the change in median Brain size as a function of Gestation depend on the Body size?

Let's fit the interaction Model IV to the Brain size data:

```
m4 = lm(log10(Brain) ~ log10(Gestation)*log10(Body),data=case0902)
summary(m4)
```

```
##
## Call:
## lm(formula = log10(Brain) ~ log10(Gestation) * log10(Body), data = case0902)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45621 -0.12670 -0.01429  0.13758  0.70269
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -0.23608    0.19800  -1.192   0.2362
## log10(Gestation)              0.70301    0.10934   6.430 5.61e-09 ***
## log10(Body)                   0.70920    0.09477   7.483 4.22e-11 ***
## log10(Gestation):log10(Body) -0.07887    0.04452  -1.772   0.0798 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2105 on 92 degrees of freedom
## Multiple R-squared:  0.9517, Adjusted R-squared:  0.9502
## F-statistic: 604.8 on 3 and 92 DF,  p-value: < 2.2e-16
```

```
# Perform an extra sum of squares test
anova(m3,m4)
```

```
## Analysis of Variance Table
##
## Model 1: log10(Brain) ~ log10(Gestation) + log10(Body)
## Model 2: log10(Brain) ~ log10(Gestation) * log10(Body)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     93 4.2152
## 2     92 4.0762  1   0.13904 3.1382 0.07979 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# This visualization looks basically the same as for Model III without interaction
#plot(log10(Brain) ~ log10(Gestation),data=case0902[WtQ1,],pch=1,col=1,xlim =
#c(1,log10(660)),ylim=c(-1,log10(4600)+2.5),main="MLR with equal slope assumption")
#points(log10(Brain) ~ log10(Gestation),data=case0902[WtMedian,],pch=2,col=2)
#points(log10(Brain) ~ log10(Gestation),data=case0902[WtQ3,],pch=3,col=3)
#gest = 0:2800
#bod = 2   # 25th percentile for body size
#lines(log10(gest),-0.23608 + 0.70301*log10(gest) + 0.70920*log10(bod)
#     -0.07887*log10(gest)*log10(bod),col=1,lty=1)
#
#bod = 9   # Median for body size
```

```
#lines(log10(gest),-0.23608 + 0.70301*log10(gest) + 0.70920*log10(bod)
#    -0.07887*log10(gest)*log10(bod),col=2,lty=2)
#
#bod = 95   # 75th percentile for body size
#lines(log10(gest),-0.23608 + 0.70301*log10(gest) + 0.70920*log10(bod)
#     -0.07887*log10(gest)*log10(bod),col=3,lty=4)
#
#legend("topleft",legend=c("data around Q1 for body size","line at Q1 for body size",
#   "data around median body size","line at #median body size","data around Q3 for body size",
#   "line at Q3 for body size"),col=c(1,1,2,2,3,3),lty=c(0,1,0,2,0,4),pch=c(1,-1,2,-1,3,-1))


# The residual plots look basically the same as for Model III without interaction
#diagANOVA(m4)
```

QUESTIONS:

1. Does the change in median Brain size as a function of Gestation depend on the Body size?

2. Report the R^2 value for this last Model IV fit to the data. How does it compare to the Model III fit to the data?

3. Report and interpret the extra sum of squares test.

Adding in more than two predictors and/factors

The Models I-IV that we have investigated considered the mean of a response Y as a linear function of two explanatory variables, but these were just building blocks. If required for the scientific questions of interest, it is desirable to add in more than two explanatory variables into the regression model.

Consider the Brain size data where we actually had 3 explanatory variables: Gestation, Body size and Litter size. Let's consider a model that investgates their simultaneous affect on Brain size. Our graphical assessment using matrix plots suggests using Litter directly in the model:

$$\mu\{\log_{10}(\text{Brain})|\text{Gestation},\text{Body}\} = \beta_0 + \beta_1 \log_{10}(\text{Gestation}) + \beta_2 \log_{10}(\text{Body}) + \beta_3 \text{Litter}. \quad (5)$$

Or how about including a 2-way interaction? For example:

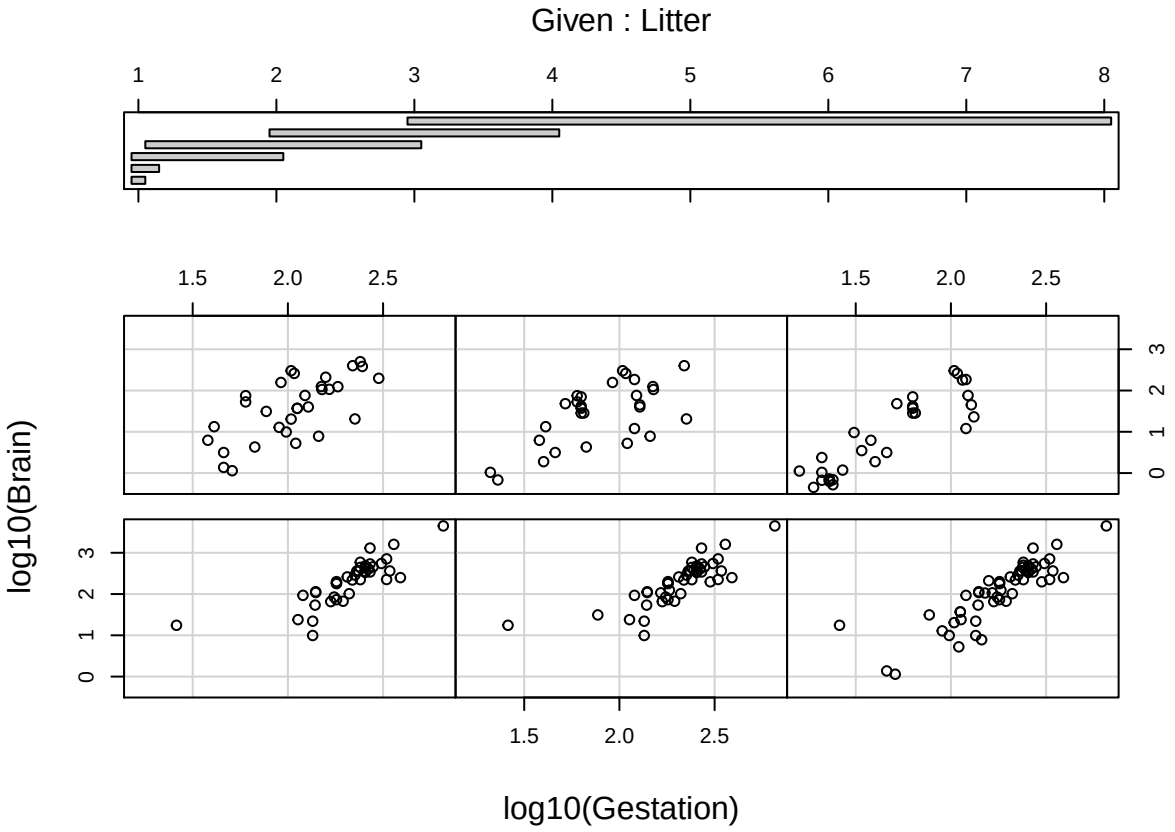
$$\begin{aligned}\mu\{\log_{10}(\text{Brain})|\text{Gestation},\text{Body}\} = \beta_0 \quad &+ \quad \beta_1 \log_{10}(\text{Gestation}) + \beta_2 \log_{10}(\text{Body}) + \beta_3 \text{Litter} \\ &+ \quad \beta_4 \log_{10}(\text{Gestation}) \times \log_{10}(\text{Body})\end{aligned}$$

or

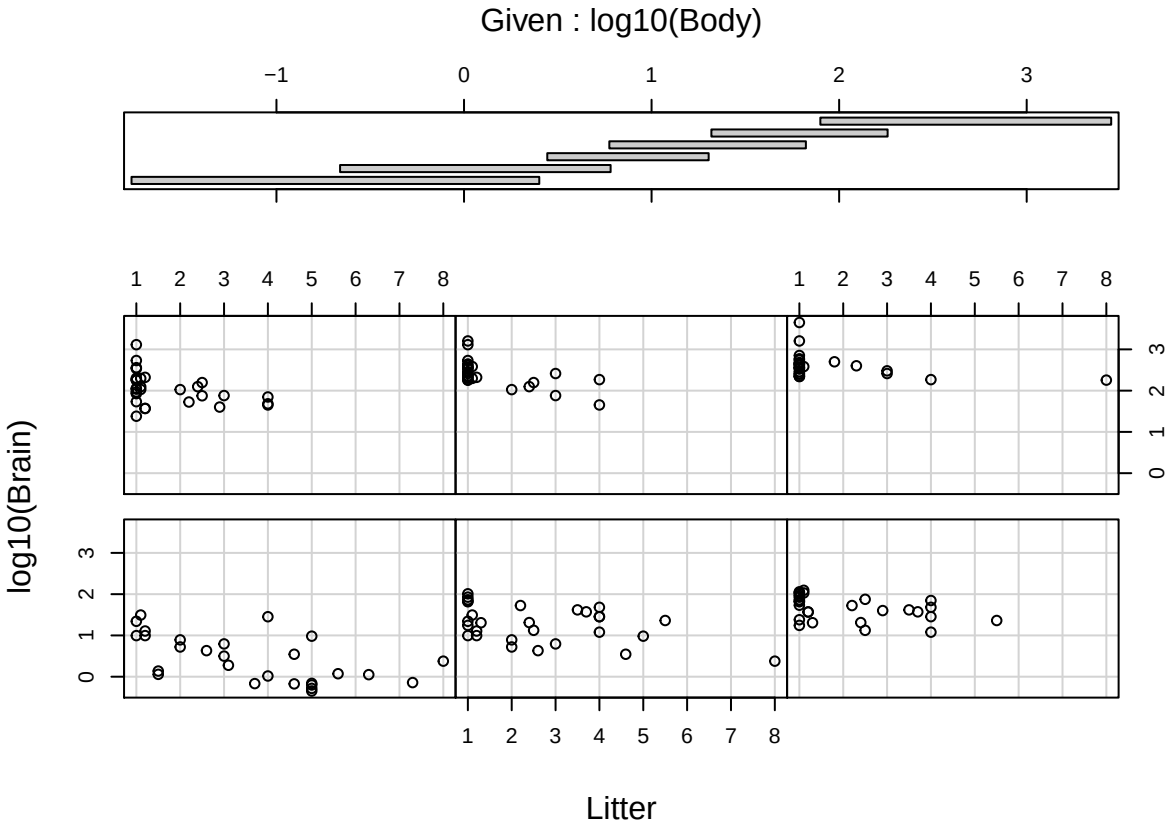
$$\begin{aligned}\mu\{\log_{10}(\text{Brain})|\text{Gestation},\text{Body}\} = \beta_0 \quad &+ \quad \beta_1 \log_{10}(\text{Gestation}) + \beta_2 \log_{10}(\text{Body}) + \beta_3 \text{Litter} \\ &+ \quad \beta_4 \log_{10}(\text{Gestation}) \times \text{Litter}\end{aligned}$$

or should we include some other set of interactions? There are 3 two-way interactions between Gestation, Body size and Litter; then there is the choice of how many of these 3 to include in any given model; and then there is the choice of whether to include the 3-way interaction $\beta_4 \log_{10}(\text{Gestation}) \times \log_{10}(\text{Body}) \times \text{Litter}$. We can consider a few plots to help decide on interactions:

```
coplot(log10(Brain) ~ log10(Gestation)|Litter,data=case0902)
```




```
coplot(log10(Brain) ~ Litter|log10(Body),data=case0902)
```



Let's fit the simpler model with no interactions in equation (5) as well as the most complicated model that includes all interactions:

```
m.simple = lm(log10(Brain)~log10(Gestation) + log10(Body) + Litter,data=case0902)
m.complex = lm(log10(Brain)~log10(Gestation)*log10(Body)*Litter,data=case0902)
anova(m.simple,m.complex)
```

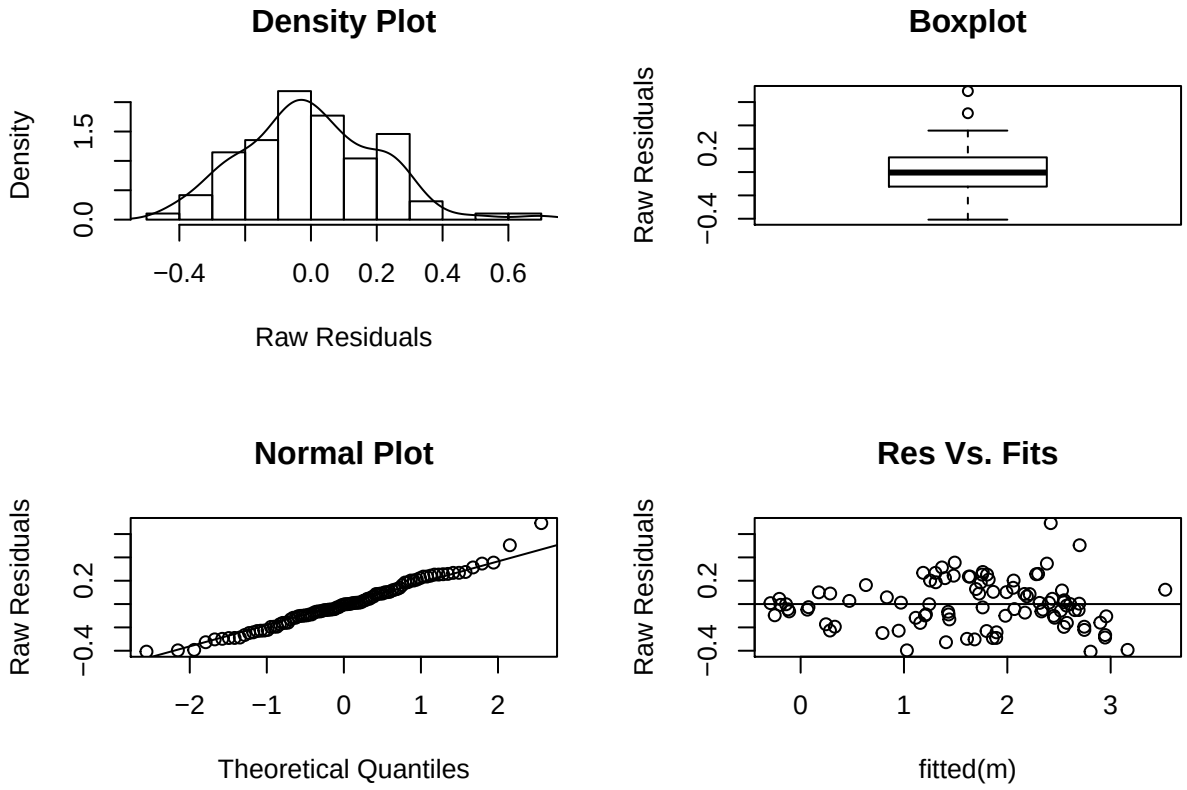
```
## Analysis of Variance Table
##
## Model 1: log10(Brain) ~ log10(Gestation) + log10(Body) + Litter
## Model 2: log10(Brain) ~ log10(Gestation) * log10(Body) * Litter
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     92 3.9243
## 2     88 3.6242  4   0.30007 1.8215 0.1318
```

```
summary(m.simple)
```

```
##
## Call:
## lm(formula = log10(Brain) ~ log10(Gestation) + log10(Body) +
##     Litter, data = case0902)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40778 -0.12126 -0.00403  0.12441  0.69375
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)       0.35759    0.28753   1.244  0.21678
## log10(Gestation)  0.43964    0.13698   3.210  0.00183 **
## log10(Body)       0.57455    0.03264  17.601  < 2e-16 ***
## Litter           -0.04794    0.01836  -2.611  0.01053 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2065 on 92 degrees of freedom
## Multiple R-squared:  0.9535, Adjusted R-squared:  0.952
## F-statistic: 629.4 on 3 and 92 DF,  p-value: < 2.2e-16
```

```
diagANOVA(m.simple)
```



QUESTIONS:

Write out the fit regression equation.

Interpret the coefficient for Litter in terms of the problem.

Are any of the explanatory variables Gestation, Body size or Litter size not useful in predicting Brain size?

Which is preferred, `m.simple` or `m.complex`? Answer using results from the extra sum of squares test, R^2 values, the graphical assessment of the simple model fit (via `diagANOVA()`) and the graphical assessment of 2-way interactions (via `coplot()`).

9.3.3 Models that include factors with more than 2 categories

We now consider the more general scenario where the mean response is modeled as a function of a quantitative predictor and a factor that has more than 2 levels. Models I and II considered the more simplified case where the mean response is modeled as a function of a predictor and a factor that has only two levels.

We will use this model to analyze the study data from Exercise 7.22 on page 200 of the textbook. The reseachers in this study measured the closing force and propodus heights of claws for 3 species of crabs: *Hemigrapsus nudus, Lophopanopeus bellus, Cancer productus*. These are abbreviated as *H.n., L.b. and C.p.* below. As your textbook does, we will model log-Force as a linear function of log-Height for each Species by simply extending Models I and II .

```
# Initial view of the data:
summary(ex0722)
```

```
##      Force           Height                       Species
##  Min.   : 2.00   Min.   : 5.000   Cancer productus    :12
##  1st Qu.: 4.00   1st Qu.: 7.025   Hemigrapsus nudus   :14
##  Median : 8.70   Median : 8.250   Lophopanopeus bellus:12
##  Mean   :12.13   Mean   : 8.813
##  3rd Qu.:19.60   3rd Qu.:10.650
##  Max.   :29.40   Max.   :13.100
```

```
dim(ex0722)  # n=38
```

```
## [1] 38  3
```

```
# Select a few rows corresponding to crabs of all 3 species
ex0722[c(13:16,25:28),]
```

```
##    Force Height              Species
## 13   4.0   12.1    Hemigrapsus nudus
## 14   5.2   12.2    Hemigrapsus nudus
## 15   2.1    5.1 Lophopanopeus bellus
## 16   8.7    5.9 Lophopanopeus bellus
## 25  27.4    8.2 Lophopanopeus bellus
## 26  29.4   11.0 Lophopanopeus bellus
## 27   5.0    6.7     Cancer productus
## 28   7.8    7.1     Cancer productus
```

The model without an interaction is:

$$\mu\{\log_{10}(\text{Force})|\text{Height}\} = \beta_0 + \beta_1\text{Height} + \beta_2\text{D}_{\text{H.n.}}(\text{Species}) + \beta_3\text{D}_{\text{L.b.}}(\text{Species}).$$

This is an *equal slope* model because we are simultaneously modeling three lines with the same slope for each of the 3 species:

$$\mu\{\log_{10}(\text{Force})|\text{Height}\} = \begin{cases} \beta_0 + \beta_1 \log_{10}(\text{Height}) & \text{for } C.p. \\ (\beta_0 + \beta_2) + \beta_1 \log_{10}(\text{Height}) & \text{for } H.n. \\ (\beta_0 + \beta_3) + \beta_1 \log_{10}(\text{Height}) & \text{for } L.b \end{cases}$$

The model with an interaction is:

$$\begin{aligned} \mu\{\log_{10}(\text{Force})|\text{Height}\} = \beta_0 \quad &+ \quad \beta_1 \log_{10}(\text{Height}) + \beta_2 \text{D}_{\text{H.n.}}(\text{Species}) + \beta_3 \text{D}_{\text{L.b.}}(\text{Species}) \\ &+ \quad \beta_4 \log_{10}(\text{Height}) \times \text{D}_{\text{H.n.}}(\text{Species}) \\ &+ \quad \beta_5 \log_{10}(\text{Height}) \times \text{D}_{\text{L.b.}}(\text{Species}). \end{aligned}$$

This is a *separate slopes* model that simultaneously models a set of lines, each with a different slope and y-intercept for each of the 3 species:

$$\mu\{\log_{10}(\text{Force})|\text{Height}\} = \begin{cases} \beta_0 + \beta_1 \log_{10}(\text{Height}) & \text{for } C.p. \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_4) \log_{10}(\text{Height}) & \text{for } H.n. \\ (\beta_0 + \beta_3) + (\beta_1 + \beta_5) \log_{10}(\text{Height}) & \text{for } L.b \end{cases}$$

We will fit two models: one without and one with the interaction.

```
# Fit the model without any interaction
m.crab = lm(log10(Force) ~ log10(Height) + Species,data=ex0722)
#summary(m.crab)
#diagANOVA(m.crab)

# Fit the model with an interaction
m.crab.int = lm(log10(Force) ~ log10(Height)*Species,data=ex0722)
anova(m.crab,m.crab.int)
```

```
## Analysis of Variance Table
##
## Model 1: log10(Force) ~ log10(Height) + Species
## Model 2: log10(Force) ~ log10(Height) * Species
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     34 1.5809
## 2     32 1.1311  2   0.44973 6.3615 0.00472 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

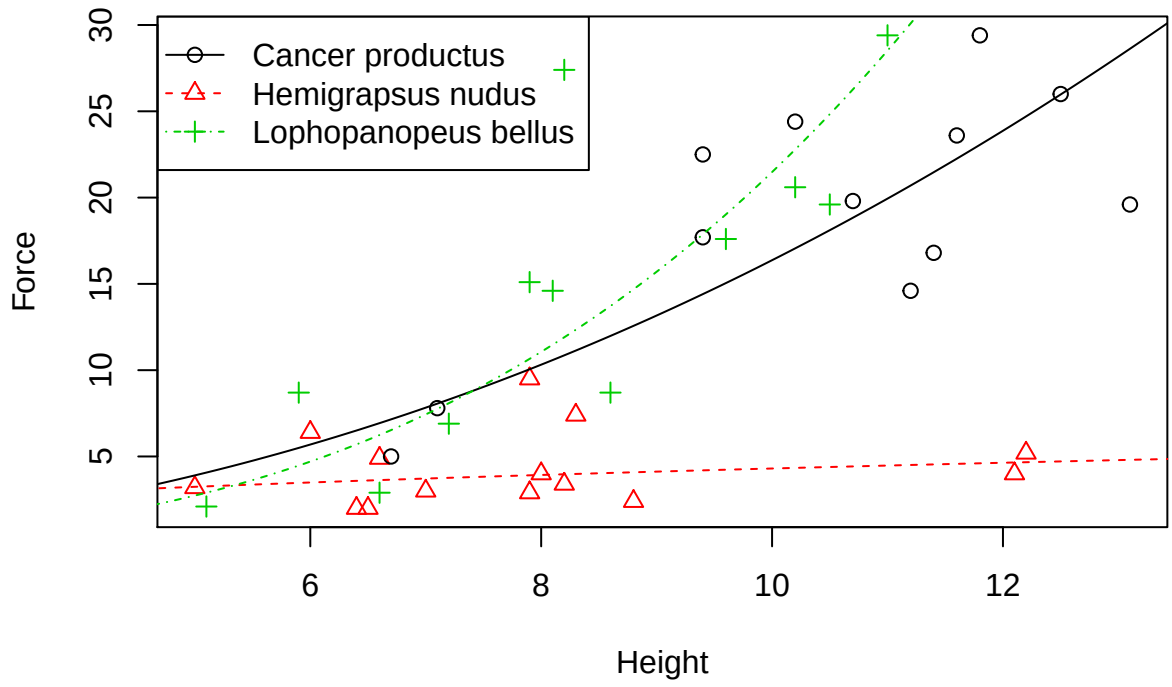
```
summary(m.crab.int)
```

```
##
## Call:
## lm(formula = log10(Force) ~ log10(Height) * Species, data = ex0722)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33300 -0.12380 -0.01001  0.10531  0.38575
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                               -0.8544     0.6293  -1.358
## log10(Height)                              2.0685     0.6208   3.332
## SpeciesHemigrapsus nudus                   1.0798     0.7646   1.412
## SpeciesLophopanopeus bellus               -0.7873     0.8049  -0.978
## log10(Height):SpeciesHemigrapsus nudus    -1.6601     0.7889  -2.104
## log10(Height):SpeciesLophopanopeus bellus  0.9052     0.8302   1.090
##                                          Pr(>|t|)
```

```
## (Intercept)                                 0.18405
## log10(Height)                               0.00219 **
## SpeciesHemigrapsus nudus                    0.16752
## SpeciesLophopanopeus bellus                 0.33536
## log10(Height):SpeciesHemigrapsus nudus      0.04330 *
## log10(Height):SpeciesLophopanopeus bellus   0.28368
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.188 on 32 degrees of freedom
## Multiple R-squared:  0.7945, Adjusted R-squared:  0.7624
## F-statistic: 24.75 on 5 and 32 DF,  p-value: 3.935e-10
```

```
#diagANOVA(m.crab.int)

# Plot the data and model fit
plot(Force ~ Height,data=ex0722,pch=as.numeric(Species),col=as.numeric(Species))
hts = seq(4,14,length=100)
lines(hts,10^(-0.8544 + 2.0685*log10(hts)),col=1,lty=1)                          # C.p.
lines(hts,10^((-0.8544 + 1.0798) + (2.0685 -1.6601)*log10(hts)),col=2,lty=2)  # H.n.
lines(hts,10^((-0.8544 -0.7873) + (2.0685 + 0.9052)*log10(hts)),col=3,lty=4)  # L.b.
legend("topleft",legend=c("Cancer productus","Hemigrapsus nudus",
  "Lophopanopeus bellus"),pch=c(1,2,3),lty=c(1,2,4),col=c(1,2,3))
```



QUESTIONS:

1. Does the change in median Brain size as a function of Gestation depend on the Body size?

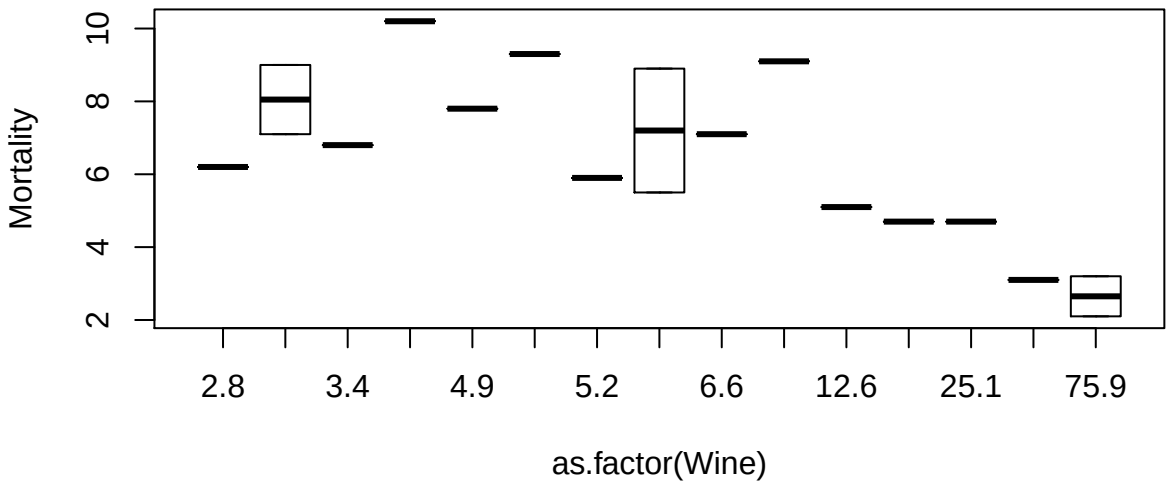
2. Report the R^2 value for this last Model IV fit to the data. How does it compare to the Model III fit to the data?

3. Report and interpret the extra sum of squares test.

MLR words of caution:

- The regression model is (usually) not an exact equation. The models you investigate are tools to help investigate questions of interest by adequately approximating the mean of the response.
- Some statisticians, including the authors of your textbook, talk about the *effect* of a predictor on the response even when an observational study has been performed. This use of *effect* does not imply that the predictor *causes* the response!
- Some statisticians loosely speak of the regression model as describing y as a function of x when in fact the regression models the mean or median of y as a function of x.
- Be careful of **over-fitting** the model to the data. The regression model will not be very helpful if it contains too many explanatory variables or if the model is a complicated function of the predictors. For example:

```
# Example of overfitting
plot(Mortality ~ as.factor(Wine),data=ex0823)
```




```
diagANOVA(lm(log10(Mortality)~ as.factor(Wine),data=ex0823))
```