9.4-9.6: More MLR models (III and IV and more!) with examples

November 8, 2017

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.

MLR words of caution:

- Do not think of the regression model as some exact, discoverable 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:

```r
library(Sleuth3)
source("diagANOVA.r")
diagANOVA(lm(log10(Mortality)~ as.factor(Wine),data=ex0823))
```
In this sample of size n=18, correlation of the residuals in the qq-plot is r=0.891592
In the following table, if r < critical.r, then the qq-plot suggests the residuals are not normal:

<table>
<thead>
<tr>
<th>n</th>
<th>critical.r</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.832</td>
</tr>
<tr>
<td>2</td>
<td>0.880</td>
</tr>
<tr>
<td>3</td>
<td>0.911</td>
</tr>
<tr>
<td>4</td>
<td>0.929</td>
</tr>
<tr>
<td>5</td>
<td>0.941</td>
</tr>
<tr>
<td>6</td>
<td>0.949</td>
</tr>
<tr>
<td>7</td>
<td>0.960</td>
</tr>
<tr>
<td>8</td>
<td>0.966</td>
</tr>
<tr>
<td>9</td>
<td>0.971</td>
</tr>
<tr>
<td>10</td>
<td>0.976</td>
</tr>
</tbody>
</table>

plot(Mortality ~ as.factor(Wine), data=ex0823)
An example of MLR Model III

Model III (from section 9.2.1): A model of the mean response $Y$ with two predictors (i.e., quantitative variables $X_1$ and $X_2$):

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

This is an equal slopes model, because for any fixed value of $X_2$, there is a linear relationship between the mean of $Y$ and $X_1$ with the same slope $\beta_1$. These lines have different $y$-intercepts $= \beta_0 + \beta_2 X_2$. The relationship between $X_1$ and $X_2$ and $Y$ over all values forms a plane in 3D that may be hard to visualize (but we will try below).

The R-code to fit this model to data is

```
lm(Y ~ X_1 + X_2).
```
```r
# Get the data
summary(case0902)
```

```r
## 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
##
```

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

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

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

```r
## 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 explanatory 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:

```r
# 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 log10-transform Brain size only then use another matrix plot (leaving out Species this time) to assess:

```r
pairs(log10(Brain) ~ Gestation + Litter + Body, data=case0902)
```
This last matrix plot indicates a non-linear relationship between the mean log-Brain size and each of Gestation and Body size. Consider log-transforms of these explanatory variables next:

```r
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:

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

Let’s fit the MLR in equation (2) and assess residual plots:

```r
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:

$$
\mu \{ \log_{10}(\text{Brain})|\text{Gestation, Body} \} = \beta_0 + \beta_1 \log_{10}(\text{Gestation}) + \beta_2 \log_{10}(\text{Body})
$$  \hspace{1cm} (3)

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:

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

The $y$-intercept = $\beta_0 + \beta_2 \log_{10}(\text{Body})$ but the slope is a constant $\beta_1$ regardless of Gestation.

**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 (3) to the data:

```r
# 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
```

```r
anova(lm(log10(Brain) ~ 1, data = case0902), m3)
```

```
## Analysis of Variance Table
##
## Model 1: log10(Brain) ~ 1
## Model 2: log10(Brain) ~ log10(Gestation) + log10(Body)
## Res.Df RSS Df Sum of Sq F Pr(>F)
## 1    95 84.463
## 2    93  4.215 2   80.247 885.25 < 2.2e-16 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```
## Assess model fit

diagANOVA(m3)

---

### Density Plot

- Raw Residuals

---

### Boxplot

- Raw Residuals

---

### Normal Plot

- Theoretical Quantiles

---

### Res Vs. Fits

- fitted(m)

---

---

---

### Code Snippet

```r
# Now 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
```
The parallelism between the lines in the visualization 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 (3) 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 lack-of-fit test. Is the MLR better than the single mean null model?

5. Give an estimate of $\sigma$, the constant SD

6. Give the value of $R^2$ as a quantitative measure of the model’s fit.

An example of Model IV - including an interaction

We can add an interaction term to Model III to arrive at Model IV (just as we added an interaction to Model I to get Model II):

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

This is a separate slopes model, because for any fixed value of $X_2$ you get a linear relationship between the mean of $Y$ and $X_1$ that has a slope that depends on $X_2$: $\beta_1 + \beta_3 X_2$; and a $y$-intercept that depends on $X_2$: $\beta_0 + \beta_2 X_2$.

In R:

```r
lm(Y ~ X1 * X2).
```

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:

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

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

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

QUESTION: Which parameter is of interest to answer the question: 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:
```r
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:  "." 0.1  "" 0.05  *** 0.01  ** 0.001  ***

## 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 a lack-of fit 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 0.0798 .
## ---
## Signif. codes:  "." 0.1  "" 0.05  *** 0.01  ** 0.001  ***

# This visualization looks basically the same as for Model III without interaction
```
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 lack of fit 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 investigates their simultaneous affect on Brain size. Our graphical assessment using matrix plots suggests using Litter directly in the model:

$$
\mu\{\log_{10}(\text{Brain})|\text{Gestation, Body}\} = \beta_0 + \beta_1 \log_{10}(\text{Gestation}) + \beta_2 \log_{10}(\text{Body}) + \beta_3 \text{Litter}. \quad (4)
$$

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

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

or
\[ \mu_{\log_{10}(\text{Brain}) | \text{Gestation, Body}} = \beta_0 + \beta_1 \log_{10}(\text{Gestation}) + \beta_2 \log_{10}(\text{Body}) + \beta_3 \text{Litter} + \beta_4 \log_{10}(\text{Gestation}) \times \text{Litter} \]

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:

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

<table>
<thead>
<tr>
<th>Litter</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>2.0</td>
<td>2.5</td>
<td>3.0</td>
<td>3.5</td>
<td>4.0</td>
<td>4.5</td>
<td>5.0</td>
<td>5.5</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Litter</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>2.0</td>
<td>2.5</td>
<td>3.0</td>
<td>3.5</td>
<td>4.0</td>
<td>4.5</td>
<td>5.0</td>
<td>5.5</td>
</tr>
</tbody>
</table>
Let’s fit the simpler model with no interactions in equation (4) as well as the most complicated model that includes all interactions:

```r
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
## -0.4078 -0.1213 -0.0040  0.1244  0.6938
##
## 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)

## [1] "In this sample of size n=96, correlation of the residuals in the qq-plot is r=0.990605"
## [1] "In the following table, if r < critical.r, then the qq-plot suggests the residuals are not normal:
## n  critical.r
##  1   0.832
##  2   0.880
##  3   0.911
##  4   0.929
##  5   0.941
##  6   0.949
##  7   0.960
##  8   0.966
##  9   0.971
## 10   0.976
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 lack-of-fit 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 (that we learned about in the first part of Chapter 9) 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 researchers 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.

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

<table>
<thead>
<tr>
<th></th>
<th>Force</th>
<th>Height</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>##</td>
<td>Min.</td>
<td>Min.</td>
<td>Cancer productus</td>
</tr>
<tr>
<td>##</td>
<td>1st Qu.</td>
<td>1st Qu.</td>
<td>Hemigrapsus nudus</td>
</tr>
<tr>
<td>##</td>
<td>Median</td>
<td>Median</td>
<td>Lophopanopeus bellus</td>
</tr>
<tr>
<td>##</td>
<td>Mean</td>
<td>Mean</td>
<td></td>
</tr>
<tr>
<td>##</td>
<td>12.13</td>
<td>8.813</td>
<td></td>
</tr>
<tr>
<td>##</td>
<td>3rd Qu.</td>
<td>3rd Qu.</td>
<td></td>
</tr>
<tr>
<td>##</td>
<td>19.60</td>
<td>10.650</td>
<td></td>
</tr>
<tr>
<td>##</td>
<td>Max.</td>
<td>Max.</td>
<td></td>
</tr>
<tr>
<td>##</td>
<td>29.40</td>
<td>13.100</td>
<td></td>
</tr>
</tbody>
</table>

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:

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

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:

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

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.

```r
# 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

<table>
<thead>
<tr>
<th>#</th>
<th>Model 1: log10(Force) ~ log10(Height) + Species</th>
<th>Model 2: log10(Force) ~ log10(Height) * Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Res.Df</td>
<td>RSS</td>
<td>Sum of Sq</td>
</tr>
<tr>
<td>1</td>
<td>34</td>
<td>1.5809</td>
</tr>
<tr>
<td>2</td>
<td>32</td>
<td>1.1311</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Signif. codes:</td>
<td>0 '<em><strong>' 0.001 '</strong>' 0.01 '</em>' 0.05 '.' 0.1 ' ' 1</td>
<td></td>
</tr>
</tbody>
</table>

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

21
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 lack of fit test.