Assignment 2, Statistical Modelling III

Andrew Martin

April 19, 2018

1. Consider the single factor model,

$$\eta_{ij} = \mu + \alpha_i$$

for $i = 1, \ldots, r$ and $j = 1, \ldots, n_i$

(a) Express this model in the form $\eta = X_0\beta_0$ by writing out the matrix X_0 and vector β_0, with no constraints on the parameters. Show that the columns of X_0 are linearly dependent

Solution

$$\eta = \begin{pmatrix} 1 & 1 & 0 & \ldots & 0 & 0 \\ 1 & 1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 1 & 0 & 1 & \ldots & 0 & 0 \\ 1 & 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 1 & 0 & 0 & \ldots & 0 & 1 \\ 1 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_r \end{pmatrix}$$

This is linearly dependent as the first column can be written using the other columns, i.e. $c_1 = c_1 + c_2 + \ldots + c_r$ Which means the first column is linearly dependent **As Required**

(b) Express the model in the form $\eta = X_1\beta_1$ by writing out the matrix X_1 and the vector β_1, subject to the constraint $\alpha_1 = 0$

Solution

$$\eta = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 & 0 \\ 1 & 0 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 1 & 1 & 0 & \ldots & 0 & 0 \\ 1 & 1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 1 & 0 & 1 & \ldots & 0 & 0 \\ 1 & 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 1 & 0 & 0 & \ldots & 0 & 1 \\ 1 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_2 \\ \vdots \\ \alpha_r \end{pmatrix}$$

As Required

(c) Express the model in the form $\eta = X_2\beta_2$ by writing out the matrix X_2 and the vector β_2, subject to the constraint $\sum_{i=1}^{r} \alpha_i = 0$ **Hint:** use the fact that $\alpha_r = -(\alpha_1 + \ldots + \alpha_{r-1})$

Solution

$$\eta = \begin{pmatrix} 1 & 1 & 0 & \ldots & 0 & 0 \\ 1 & 1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 1 & 0 & 1 & \ldots & 0 & 0 \\ 1 & 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 1 & 0 & 0 & \ldots & 0 & 1 \\ 1 & 0 & 0 & \ldots & 0 & 1 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 1 & -1 & -1 & \ldots & -1 & -1 \\ 1 & -1 & -1 & \ldots & -1 & -1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{r-1} \end{pmatrix}$$

As Required

(d) The $r \times r$ matrix A given is such that $X_1 = X_2 A$. Find A^{-1} and hence demonstrate that the two model formulations are equivalent.

$$A = \begin{pmatrix} 1 & \frac{1}{r} & \frac{1}{r} & \dots & \frac{1}{r} & \frac{1}{r} \\ 0 & -\frac{1}{r} & -\frac{1}{r} & \dots & -\frac{1}{r} & -\frac{1}{r} \\ 0 & \frac{r-1}{r} & -\frac{1}{r} & \dots & -\frac{1}{r} & -\frac{1}{r} \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ 0 & -\frac{1}{r} & -\frac{1}{r} & \dots & \frac{r-1}{r} & -\frac{1}{r} \end{pmatrix}$$

Solution

$$\begin{aligned} X_1\beta_1 &= X_2 A\beta_1 \\ &= X_2(A\beta_1) \\ &= X_2\beta_2 \end{aligned}$$

As Required

2. Consider the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i \text{ for } i = 1, 2, \dots, n$$

(a) Express the model in the form $\mathbf{y} = X\beta + e$

Solution

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

As Required

(b) Show that the leverage h_{ii} is given by:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ for simple linear regression

Solution The H matrix is

$$H = X(X^T X)^{-1} X^T$$

$$\begin{aligned} (X^T X)^{-1} &= \begin{pmatrix} \sum_{j=1}^{n} 1 & \sum_{j=1}^{n} x_j \\ \sum_{j=1}^{n} x_j & \sum_{j=1}^{n} x_j^2 \end{pmatrix}^{-1} \\ &= \frac{1}{\left(n \times \sum_{j=1}^{n} x_j^2\right) - \left(\sum_{j=1}^{n} x_j \times \sum_{j=1}^{n} x_j\right)} \begin{pmatrix} \sum_{j=1}^{n} x_j^2 & -\sum_{j=1}^{n} x_j \\ -\sum_{j=1}^{n} x_j & n \end{pmatrix} \\ X(X^T X)^{-1} &= \frac{1}{\left(n \sum_{j=1}^{n} x_j^2\right) - \left(\sum_{j=1}^{n} x_j \times \sum_{j=1}^{n} x_j\right)} \begin{pmatrix} \sum_{j=1}^{n} x_j^2 - x_1 \sum_{j=1}^{n} x_j & -\sum_{j=1}^{n} x_j + nx_1 \\ \sum_{j=1}^{n} x_j^2 - x_2 \sum_{j=1}^{n} x_j & -\sum_{j=1}^{n} x_j + nx_2 \\ \vdots & \vdots \\ \sum_{j=1}^{n} x_j^2 - x_n \sum_{j=1}^{n} x_j & -\sum_{j=1}^{n} x_j + nX_n \end{pmatrix} \end{aligned}$$

for simplicity only writing the diagonal elements as they are all the ones of concern

$$\begin{aligned} X(X^T X)^{-1} X^T = &\frac{1}{\left(n \sum_{j=1}^{n} x_j^2\right) - \left(\sum_{j=1}^{n} x_j \times \sum_{j=1}^{n} x_j\right)} \\ &\times \begin{pmatrix} \sum_{j=1}^{n} x_j^2 - x_1 \sum_{j=1}^{n} x_j - x_1 \left(\sum_{j=1}^{n} x_j + nx_1\right) & & \\ & \ddots & \\ & & \sum_{j=1}^{n} x_j^2 - x_n \sum_{j=1}^{n} x_j - x_n \left(\sum_{j=1}^{n} x_j + nx_n\right) \end{pmatrix} \end{aligned}$$

h_{ii} is the i^{th} diagonal term of this matrix. I.e.

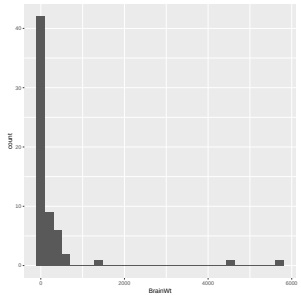
$$
\begin{aligned}
h_{ii} &= \frac{1}{\left(n\sum_{j=1}^{n} x_j^2\right) - \left(\sum_{j=1}^{n} x_j \times \sum_{j=1}^{n} x_j\right)} \times \left(\sum_{j=1}^{n} x_j^2 - x_i \sum_{j=1}^{n} x_j - x_i \left(\sum_{j=1}^{n} x_j + n x_i\right)\right) \\
&= \frac{1/n}{\left(\sum_{j=1}^{n} x_j^2\right) - \frac{1}{n}\left(\sum_{j=1}^{n} x_j \times \sum_{j=1}^{n} x_j\right)} \times \left(\sum_{j=1}^{n} x_j^2 - 2x_i \sum_{j=1}^{n} x_j - n x_i^2\right) \\
&= \frac{1/n}{\sum_{j=1}^{n}\left(x_j^2\right) - n\bar{x}^2} \left(\sum_{j=1}^{n} x_j^2 - 2x_i\bar{x} - n x_i^2\right) \\
&= \frac{1/n}{\sum_{j=1}^{n}\left(x_j^2\right) + n\bar{x}^2 - n\bar{x}^2 - n\bar{x}^2} \left(\sum_{j=1}^{n} x_j^2 - n\bar{x}^2 + n\bar{x}^2 - 2nx_i\bar{x} - n x_i^2\right) \\
&= \frac{1/n}{\sum_{j=1}^{n}\left(x_j^2\right) + n\bar{x}^2 - 2n\bar{x}^2} \left(\sum_{j=1}^{n} (x_j - \bar{x})^2 + n(x_i - \bar{x})^2\right) \\
&= \frac{1/n}{\sum_{j=1}^{n}\left(x_j^2\right) + n\bar{x}^2 - 2n\bar{x}^2} \left(\sum_{j=1}^{n} (x_j - \bar{x})^2 + n(x_i - \bar{x})^2\right) \\
&= \frac{1/n}{\sum_{j=1}^{n}\left(x_j^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2\right)} \left(\sum_{j=1}^{n} (x_j - \bar{x})^2 + n(x_i - \bar{x})^2\right) \\
&= \frac{1/n}{\sum_{j=1}^{n}\left(x_j^2 - 2x_j\bar{x} + \bar{x}^2\right)} \left(\sum_{j=1}^{n} (x_j - \bar{x})^2 + n(x_i - \bar{x})^2\right) \\
&= \frac{1/n}{\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2} \left(\sum_{j=1}^{n} (x_j - \bar{x})^2 + n(x_i - \bar{x})^2\right) \\
&= \frac{1}{n} \frac{\sum_{j=1}^{n} (x_j - \bar{x})^2 + n(x_i - \bar{x})^2}{\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2} \\
&= \frac{1}{n}\left(1 + \frac{n(x_i - \bar{x})^2}{\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2}\right) \\
&= \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}
\end{aligned}
$$

As Required

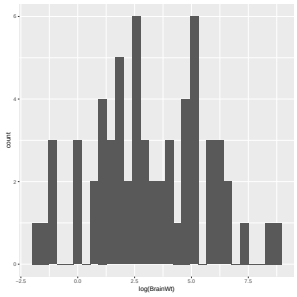
3. `sleep.txt` contains data for 62 species of mammals. Interested in factors that govern total amount of sleep, and the proportion spent dreaming.

   (a) Obtain histograms for each of `BrainWt`, `BodyWt`, `LifeSpan` and `Gestation` on the original scale, and also after taking the log transformation. Explain, making reference to *leverage*, why it may be useful to consider the transformed variables

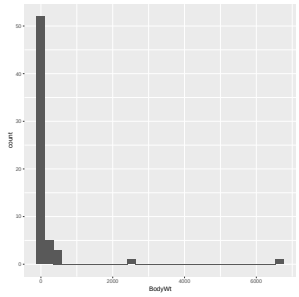
   Solution



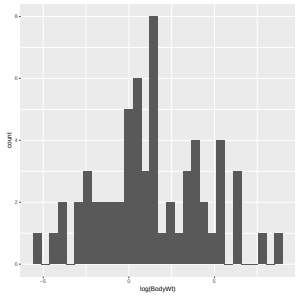
(a) Histogram of Brain Weight against sleep



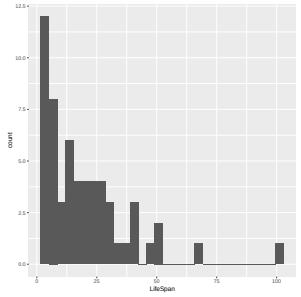
(b) Histogram of log(Brain Weight) against sleep



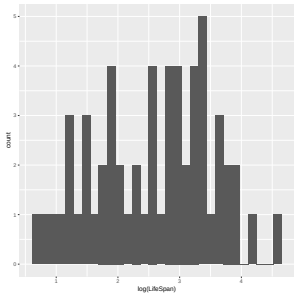
(c) Histogram of BodyWeight against sleep



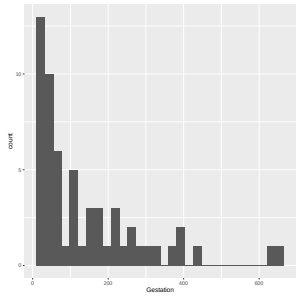
(d) Histogram of log(BodyWeight) against sleep



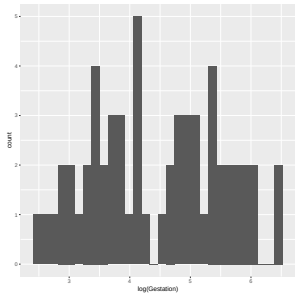
(e) Histogram of LifeSpan against sleep



(f) Histogram of log(LifeSpan) against sleep



(g) Histogram of Gestation against sleep



(h) Histogram of log(Gestation) against sleep

The histograms on the following page are those for the variables on each scale. By taking the logarithm, the distance between the points is decreased, thus decreasing the leverage. With lower leverage there is less room for the outliers to mask the correct data E.g. Looking at plots (e) and (g), there is a clear outlier on the far right of the scale. Looking at (f) and (h) these outliers are brought closer to the data and their leverage is effectively removed. This is why it would be useful to consider the log transformation instead of the original scale.

As Required

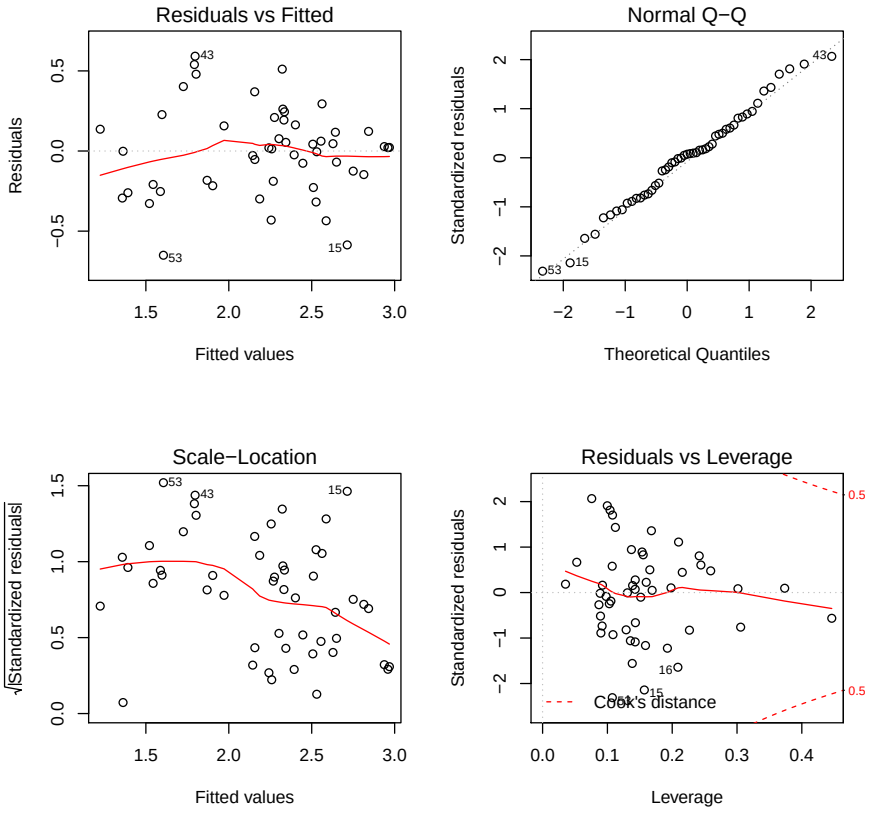
(b) Consider the multiple regression model

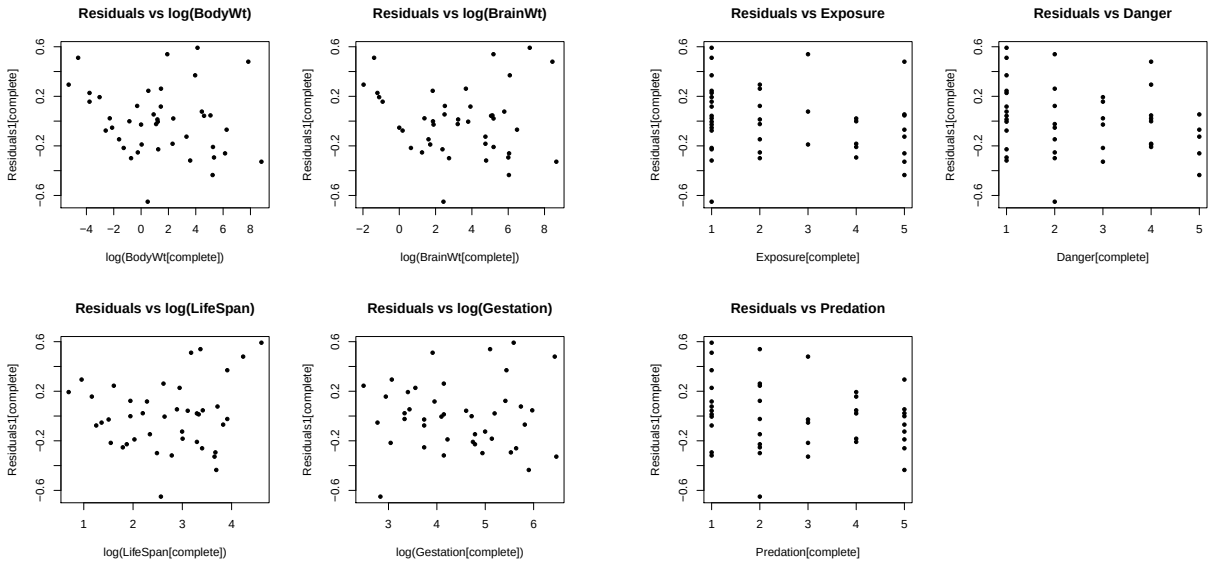
```
log(TotalSleep) ~ log(BodyWt)+log(BrainWt)+log(LifeSpan)+log(Gestation)
+ Exposure + Danger + Predation
```

Fit this model and obtain diagnostic plots. Comment on whether the model assumptions appear reasonable (use the code given to generate the plots given the missing values)

Solution

Figure 1: Diagnostic plots for the given model





(a) Residual plots of sleep against (clockwise): log(bodywt),log(brainwt),log(lifespan),log(gestation)

(b) Residual plots of sleep against (clockwise): Exposure, Danger, Predation

First check for heteroscedasticity: the assumption appears more-or-less reasonable from the spread of the points in the residuals vs individual predictor plots For each of the plots, the variation seems reasonably uniform.
The fit of the logarithm scaled data seems reasonable, as there are no high leverage, outlying points which could affect the fit (indicated on the residuals vs leverage plot)
Linearity is indicated in the residuals vs fitted plot, where the red line shows a reasonable linear trend.

The normality assumption is indicated in the Normal Q-Q plot, where most of the data fits on the line.
As Required

(c) By removing non-significant terms, find a parsimonious well-fitting model for the data

Solution Using $\alpha = 0.05$ for significance, using the output from `summary(newmodel)` gives: Note that this was done using only statistical significance, and neglecting the marginality principle.

```
Call:
lm(formula = log(TotalSleep) ~ log(BodyWt) + log(Gestation) +
    Danger, data = sleep[complete, ])

Residuals:
     Min       1Q   Median       3Q      Max
-0.69280 -0.22225  0.02824  0.13927  0.58412

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.47472    0.25282  13.744  < 2e-16 ***
log(BodyWt)    -0.04783    0.01955  -2.446   0.0182 *
log(Gestation) -0.15797    0.06052  -2.610   0.0121 *
Danger         -0.18716    0.03102  -6.034 2.39e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.2972 on 47 degrees of freedom
Multiple R-squared:  0.7136,Adjusted R-squared:  0.6953
F-statistic: 39.04 on 3 and 47 DF,  p-value: 8.193e-13
```

From this, the model ends up including an intercept, the `log(BodyWt)` term, `log(Gestation)` and `Danger` **As Required**

(d) Provide an interpretation of the coefficient estimates from your final model

Solution

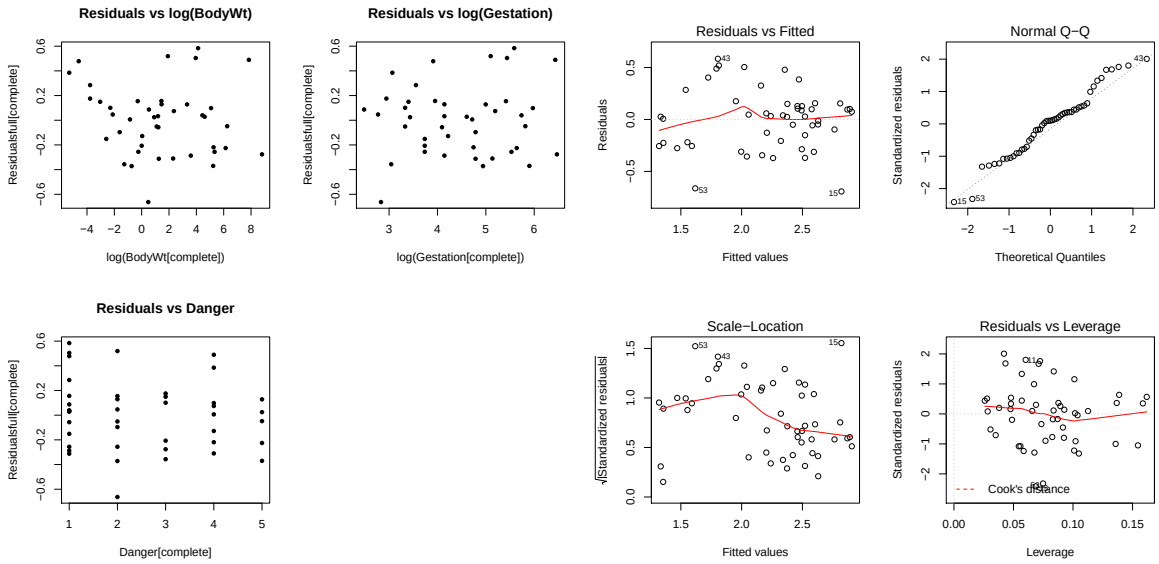
Intercept has value 3.47472. The default value in this model, with other variables = 0, will be the intercept. This is the default sleep time for a creature with 0 body weight, no gestation time, and a 0 danger index (which is not possible but this is due to danger being an integer but ignore this flaw)

log(BodyWt) has coefficient -0.04783 which equates to BodyWt having a coefficient of 0.95330. Since BodyWt is in kg, this coefficient shows that the sleep time for a mammal will be increased by roughly 95% of its weight. This would be related to larger creatures needing more time to rest.

log(Gestation) has coefficient -0.15797 which equates to gestation having a coefficient of 0.85388. This means that the number of days the mammal spends in the womb relates to an increase of sleep time by 0.85388 per day spent in the womb. Creatures who spend more time in the womb require more rest.

Danger's coefficient is -0.18716, this will multiplied by (1-5) depending on the mammal's Danger index, meaning that the more Danger the mammal is potentially in, the less sleep it will get. Mammals who are at greater risk will sleep less possibly due to fear and possibly to prevent the danger (giving opportunity to fight/flee) **As Required**

(e) Obtain diagnostic plots for this model and comment on whether the assumptions appear reasonable.
Solution



(c) Residual plots for the new model

(d) Diagnostic plots for the new model

The assumptions in this case follow from the first model:

The assumption of homoscedasticity is observed in the residuals vs variable plots as the residuals have no trend as the variables change.
Fit - using the residuals vs leverage plot it is clear that there are no high leverage outliers indicated on the plot (the boundaries to check are not even shown) so there are no high leverage values affecting the fit of the data.
Linearity is shown in the red line in the residuals vs fitted plot. This line is mostly horizontal about 0 for the residuals, showing that the linearity assumption is valid
Normality is lost slightly when reducing to this model, but the data as shown on the normal q-q plot still follows the trend. I.e. this is a valid assumption
As Required

(f) The final model you obtained should have included the term `Danger`. Consider the same model with `Danger` replaced by `Predation`. Fit this model and hence comment on whether it is possible to simply use statistical significance to identify the factors that affect `TotalSleep`. Explain your reasoning clearly
Solution

Danger is essentially an interaction term - to include it you have to include predation and exposure (marginality) as it depends on those two. Danger could potentially be a term with form: $Danger = Predation \times Exposure \times othervariables$. This means it still has some dependence on the other two variables, implying they must both be included in the model if we wish to include danger.

"Whenever an interaction term is included in the model, all implied lower order interactions and main effects must also be included. Even if they are not statistically significant." Based on statistical significance alone, they have about the same significance in each model (less than an order of 10 for very small numbers).
While it is possible to do so, the resulting model would be invalid.
As Required

Code Used

```
library(tidyverse)
library(ggplot2)
library(gridExtra)
setwd("~/Uni/2018/Statistical Modelling")
sleep = read.table("sleep.txt", header=T)

#Put it all in a tidy pdf

pdf(file="DiagPlotsAssignment2.pdf")

###a
#histograms of BrainWt, BodyWt, LifeSpan, and  Gestation on each scale
ggplot(aes(BrainWt),data = sleep) + geom_histogram()
ggplot(aes(log(BrainWt)),data = sleep) + geom_histogram()
```

```
ggplot(aes(BodyWt),data = sleep) + geom_histogram()
ggplot(aes(log(BodyWt)),data = sleep) + geom_histogram()
ggplot(aes(LifeSpan),data = sleep) + geom_histogram()
ggplot(aes(log(LifeSpan)),data = sleep) + geom_histogram()
ggplot(aes(Gestation),data = sleep) + geom_histogram()
ggplot(aes(log(Gestation)),data = sleep) + geom_histogram()
###b
#fit model and obtain diagnostic plot
lm1=lm(log(TotalSleep)~log(BodyWt)+log(BrainWt)+log(LifeSpan)
       +log(Gestation)+Exposure+Danger+Predation,data=sleep)

## Given plot code ##
par(mfrow=c(2,2))
plot(lm1)
Residuals1=residuals(lm1)
# Because there are some missing values, the vector of residuals is not the
# same length as the original variables.
# First construct an indictor for complete observations (not missing
# any values) required for the regression

##Code so the given code works...
TotalSleep = sleep$TotalSleep
BodyWt = sleep$BodyWt
BrainWt = sleep$BrainWt
LifeSpan = sleep$LifeSpan
Gestation = sleep$Gestation
Exposure = sleep$Exposure
Danger = sleep$Danger
Predation = sleep$Predation

complete=!is.na(TotalSleep)&!is.na(BodyWt)&!is.na(BrainWt)&!is.na(LifeSpan)&
          !is.na(Gestation)&!is.na(Exposure)&!is.na(Danger)&!is.na(Predation)

# Plot the residuals vs individual predictors for the complete data subset.
plot(log(BodyWt[complete]),Residuals1[complete],main="Residuals vs log(BodyWt)",pch=20)
plot(log(BrainWt[complete]),Residuals1[complete],main="Residuals vs log(BrainWt)",pch=20)
plot(log(LifeSpan[complete]),Residuals1[complete],main="Residuals vs log(LifeSpan)",pch=20)
plot(log(Gestation[complete]),Residuals1[complete],main="Residuals vs log(Gestation)",pch=20)
plot(Exposure[complete],Residuals1[complete],main="Residuals vs Exposure",pch=20)
plot(Danger[complete],Residuals1[complete],main="Residuals vs Danger",pch=20)
plot(Predation[complete],Residuals1[complete],main="Residuals vs Predation",pch=20)
plot.new() #this is just to fill a spot for formatting purposes
###c
#remove non-sig terms and find good model
summary(lm1)
#this is just lm1 but using sleep[complete,]
lm2=lm(log(TotalSleep)~log(BodyWt)+log(BrainWt)+log(LifeSpan)
       +log(Gestation)+Exposure+Danger+Predation,data=sleep[complete,])
#step function makes this step trivial
newmodel = step(lm2,direction = "both")
summary(newmodel)
#this didn't omit predation (it notices its an interaction term) omit for this problem though
newmodel = lm(log(TotalSleep)~log(BodyWt)+log(Gestation)+Danger ,data=sleep[complete,])


###e
#Diagnostic plots for the new model
Residualsfull = residuals(newmodel)
plot(log(BodyWt[complete]),Residualsfull[complete],main="Residuals vs log(BodyWt)",pch=20)
plot(log(Gestation[complete]),Residualsfull[complete],main="Residuals vs log(Gestation)",pch=20)
plot(Danger[complete],Residualsfull[complete],main="Residuals vs Danger",pch=20)
plot.new()
plot(newmodel)


###f
#Replace danger with predation
```

```
#fit model and check significance
newmodelpred =  lm(log(TotalSleep)~log(BodyWt)+log(Gestation)+Predation ,data=sleep[complete,])


dev.off()
```