STATS 3001 Statistical Modelling III
Assignment 2
Due: 4pm Thursday 19th April (Week 6), 2018

```
######################################## solutions ########################################
```

IMPORTANT In keeping with the university policy on plagiarism, you should read the University Policy Statement on Academic Honesty (plagiarism, collusion and related forms of cheating):

http://www.adelaide.edu.au/policies/230.

Assignments must be submitted with a signed Assessment Cover Sheet. These forms are available on MyUni/Canvas under `Modules`→`Assignment cover sheet`. Please note that assignment marks cannot be counted for your assessment unless a signed declaration is received.

Check off the following prior to submitting your assignment:

- ☐ Sufficient working has been provided in each question to satisfactorily demonstrate to the marker that you understand the required concepts and steps in the question.
- ☐ All R output and plots to support your answers are included where necessary.
- ☐ A coversheet is attached to the submission that is completed and signed.
- ☐ Answers are written on their own paper, not on the assignment handout.
- ☐ The submission is neat and provides space for marker's comments.
- ☐ The submission is stapled together.
- ☐ • Submit your assignment into the Statistical Modelling III hand-in box on Level 6 (Ingkarni Wardli building).
  • Late assignments will only be accepted by prior agreement with the Course Co-ordinator and relevant requests should usually be accompanied by a medical certificate.
  .

---

(1) Consider the single factor model,

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

for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, n_i$.

(a) Express this model in the form $\boldsymbol{\eta} = X_0\boldsymbol{\beta}_0$ by writing out the matrix X_0 and vector $\boldsymbol{\beta}_0$, with no constraints on the parameters. Show that the columns of X_0 are linearly dependent.

Solution:

X_0 has $n_1 + n_2 + \cdots + n_r = \sum_{i=1}^{r} n_i$ rows and $r + 1$ columns:

$$X_0 = \begin{pmatrix} 1 & 1 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & & & \vdots \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 0 & 1 & \ldots & 0 \\ 1 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & & & \vdots \\ 1 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & & & \vdots \\ 1 & 0 & 0 & \ldots & 1 \\ 1 & 0 & 0 & \ldots & 1 \\ \vdots & \vdots & & & \vdots \\ 1 & 0 & 0 & \ldots & 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta}_0 = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_r \end{pmatrix}$$

Let $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_r$ denote the columns of X_0. It is easy to check that

$$\boldsymbol{x}_0 - (\boldsymbol{x}_1 + \ldots + \boldsymbol{x}_r) = \mathbf{0}$$

and hence the columns are linearly dependent.

(b) Express the model in the form $\boldsymbol{\eta} = X_1\boldsymbol{\beta}_1$ by writing out the matrix X_1 and the vector $\boldsymbol{\beta}_1$, subject to the constraint $\alpha_1 = 0$.

Solution:

$$X_1 = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & & & \vdots \\ 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & & & \vdots \\ 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & & & \vdots \\ 1 & 0 & 0 & \dots & 1 \\ 1 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & & & \vdots \\ 1 & 0 & 0 & \dots & 1 \end{pmatrix} \text{ and } \boldsymbol{\beta}_1 = \begin{pmatrix} \mu \\ \alpha_2 \\ \vdots \\ \alpha_r \end{pmatrix}$$

(c) Express the model in the form $\boldsymbol{\eta} = X_2\boldsymbol{\beta}_2$ by writing out the matrix X_2 and the vector $\boldsymbol{\beta}_2$, subject to the constraint $\sum_{i=1}^{r} \alpha_i = 0$. **Hint**: Use the fact that $\alpha_r = -(\alpha_1 + \alpha_2 + \dots \alpha_{r-1})$.

Solution:

$$X_2 = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & & & \vdots \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & & & \vdots \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & & & \vdots \\ 1 & 0 & 0 & \dots & 1 \\ 1 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & & & \vdots \\ 1 & 0 & 0 & \dots & 1 \\ 1 & -1 & -1 & \dots & -1 \\ 1 & -1 & -1 & \dots & -1 \\ \vdots & \vdots & & & \vdots \\ 1 & -1 & -1 & \dots & -1 \end{pmatrix} \text{ and } \boldsymbol{\beta}_2 = \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_{r-1} \end{pmatrix}$$

(d) The $r \times r$ matrix A given below is such that $X_1 = X_2A$. Find A^{-1} and hence demonstrate that the two model formulations are equivalent.

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

Solution:

Note This is not part of the solution but note that since $X_1 = X_2 A$, it follows

$$(X_2^T X_2)^{-1} X_2^T X_1 = (X_2^T X_2)^{-1} X_2^T X_2 A = A.$$

A^{-1} can be found by row reduction or by solving $AA^{-1} = I_r$. (Full marks are awarded for enough steps to demonstrate the work was done, or for code if using Matlab or Maple.)

$$A^{-1} = \begin{pmatrix} 1 & 1 & 0 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & & & \ddots & \vdots \\ 0 & -1 & 0 & 0 & \dots & 1 \\ 0 & -2 & -1 & -1 & \dots & -1 \end{pmatrix}$$

Since $X_1 = X_2 A$ with A invertible it follows that

$$\boldsymbol{\eta} = X_1 \boldsymbol{\beta}_1 \Leftrightarrow \boldsymbol{\eta} = X_2 \boldsymbol{\beta}_2$$

where $\boldsymbol{\beta}_2 = A\boldsymbol{\beta}_1$. Hence the two formulations are equivalent.

[Total: 16]

(2) Consider the simple linear regression model

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

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

Solution:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \; X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \; \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \text{ and } \boldsymbol{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

(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.

[**Note**: a concise and complete proof is required for full marks.]

Solution:

Observe that the leverage h_{ii} is defined to be the ith diagonal element of $H = X(X^TX)^{-1}X^T$ which can be written equivalently as

$$h_{ii} = \boldsymbol{x}_i^T (X^TX)^{-1} \boldsymbol{x}_i,$$

where $\boldsymbol{x}_i^T$ is the ith row of X. Next observe that

$$(X^TX) = \begin{pmatrix} n & \sum_j x_j \\ \sum_j x_j & \sum_j x_j^2 \end{pmatrix}$$

and

$$(X^TX)^{-1} = \frac{1}{n\sum_j x_j^2 - (\sum_j x_j)^2} \begin{pmatrix} \sum_j x_j^2 & -\sum_j x_j \\ -\sum_j x_j & n \end{pmatrix} = \frac{1}{nS_{xx}} \begin{pmatrix} \sum_j x_j^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}.$$

Hence,

$$\begin{aligned} h_{ii} &= \boldsymbol{x}_i^T (X^TX)^{-1} \boldsymbol{x}_i \\ &= \frac{1}{nS_{xx}} \{\sum_j x_j^2 - 2nx_i\bar{x} + nx_i^2\} \\ &= \frac{1}{nS_{xx}} \{\sum_j x_j^2 - n\bar{x}^2 + n(\bar{x}^2 - 2x_i\bar{x} + x_i^2)\} \\ &= \frac{1}{nS_{xx}} \{nS_{xx} + n(x_i - \bar{x})^2\} \\ &= \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}. \end{aligned}$$

[Total: 9]

(3) The file `sleep.txt` contains data on brain and body weight, life span, gestation time, time sleeping, and predation and danger indices for 62 species of mammals. Interest in this study is focused on understanding the factors that govern the total amount of sleep and also the proportion of total sleep time that is spent dreaming.

Download the dataset from MyUni and read into `R` in the usual way.

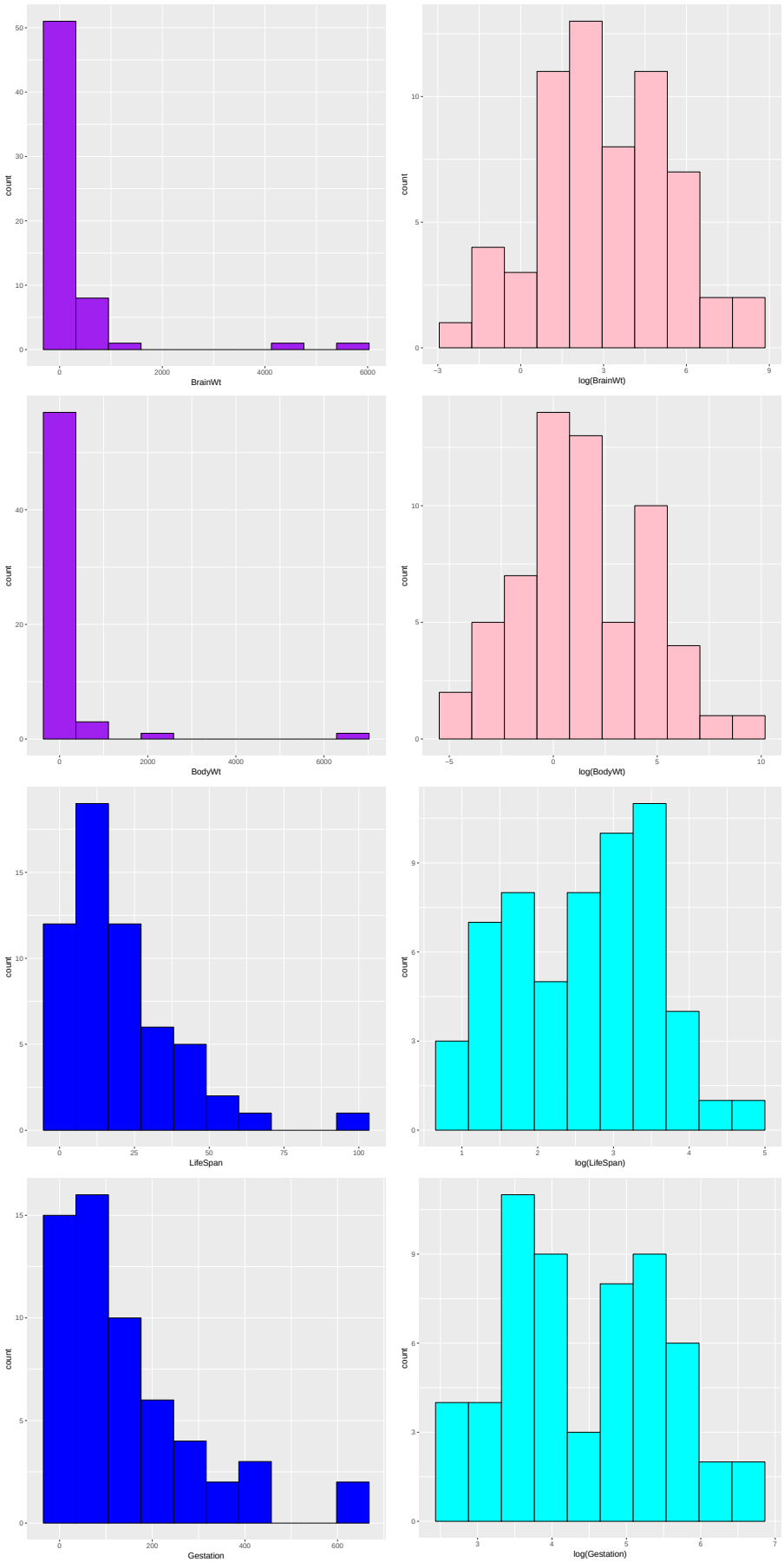
```
sleep <- read.table("sleep.txt", header=T)
```

Variable	Description
BodyWt	body weight (kg)
BrainWt	brain weight (g)
NonDreaming	slow wave ("nondreaming") sleep (hrs/day)
Dreaming	paradoxical ("dreaming") sleep (hrs/day)
TotalSleep	total sleep, sum of slow wave and paradoxical sleep (hrs/day)
LifeSpan	maximum life span (years)
Gestation	gestation time (days)
Predation	predation index (1-5) 1 = minimum (least likely to be preyed upon); 5 = maximum (most likely to be preyed upon)
Exposure	sleep exposure index (1-5) 1 = least exposed (e.g. animal sleeps in a well-protected den); 5 = most exposed
Danger	overall danger index (1-5) (based on the above two indices and other information) 1 = least danger (from other animals); 5 = most danger (from other animals)

(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:

```
attach(sleep)
library(ggplot2)
ggplot(sleep,aes(BrainWt)) + geom_histogram(col="black", fill="purple", bins=10)
ggplot(sleep,aes(log(BrainWt))) + geom_histogram(col="black", fill="pink", bins=10)
ggplot(sleep,aes(BodyWt)) + geom_histogram(col="black", fill="purple", bins=10)
ggplot(sleep,aes(log(BodyWt))) + geom_histogram(col="black", fill="pink", bins=10)
ggplot(sleep,aes(LifeSpan)) + geom_histogram(col="black", fill="blue", bins=10)
ggplot(sleep,aes(log(LifeSpan))) + geom_histogram(col="black", fill="cyan", bins=10)
ggplot(sleep,aes(Gestation)) + geom_histogram(col="black", fill="blue", bins=10)
ggplot(sleep,aes(log(Gestation))) + geom_histogram(col="black", fill="cyan", bins=10)
```


count
BrainWt
log(BrainWt)
BodyWt
log(BodyWt)
LifeSpan
log(LifeSpan)
Gestation
log(Gestation)

Comment: The histograms of the raw data, especially for `BrainWt` and `BodyWt`, are highly right skewed and include outliers that would become points of high leverage in the regression. The histograms of the log transformed data for all four variables are roughly symmetric and do not contain outliers, and hence are less likely to give rise to problems with leverage.

(b) Consider the multiple regression model

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

Fit this model and obtain appropriate diagnostic plots. Comment on whether the model assumptions appear reasonable.
(Since there are missing values in the dataset, you may use the code below to do the plots if you wish.)

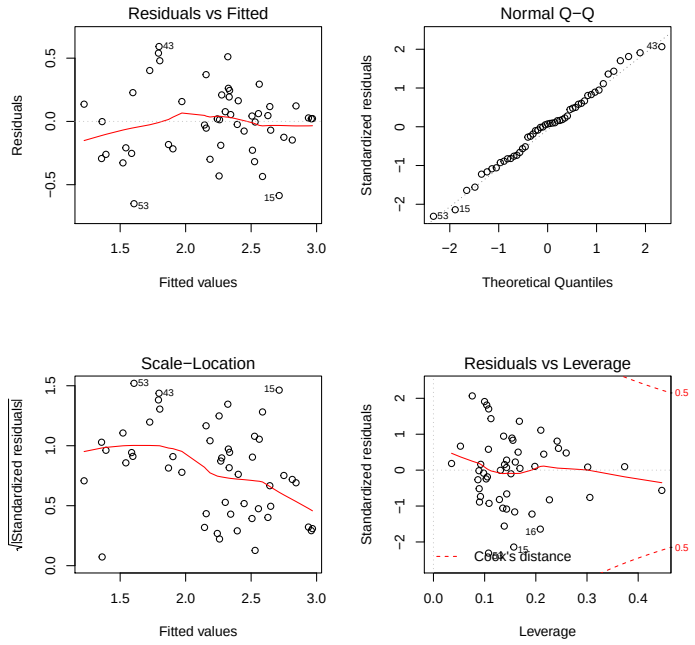
```
lm1=lm(log(TotalSleep)~log(BodyWt)+log(BrainWt)+log(LifeSpan)
       +log(Gestation)+Exposure+Danger+Predation,data=sleep)
par(mfrow=c(2,2))
plot(lm1)
Residuals1=residuals(lm1)
# Because there are some missing values, the vector of residuals is not the
# same length as the original variables.
# First construct an indictor for complete observations (not missing
# any values) required for the regression
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)
```

Solution:

To begin, fit the model and plot the usual set of residual plots from `lm()`:

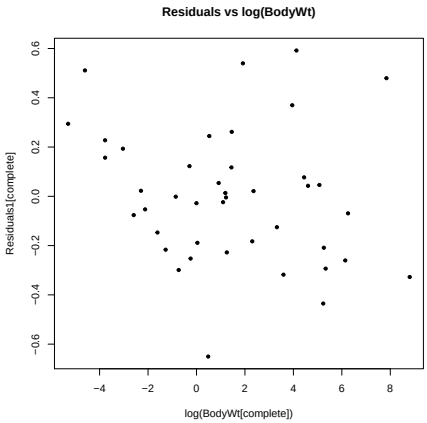
```
lm1=lm(log(TotalSleep)~log(BodyWt)+log(BrainWt)+log(LifeSpan)
       +log(Gestation)+Exposure+Danger+Predation, data=sleep)
summary(lm1)

##
## Call:
## lm(formula = log(TotalSleep) ~ log(BodyWt) + log(BrainWt) + log(LifeSpan) +
##     log(Gestation) + Exposure + Danger + Predation, data = sleep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65033 -0.19874  0.01991  0.16008  0.59184
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.23594    0.31270  10.349 3.02e-13 ***
## log(BodyWt)    -0.05018    0.05131  -0.978  0.33360
## log(BrainWt)   -0.01420    0.07661  -0.185  0.85379
```

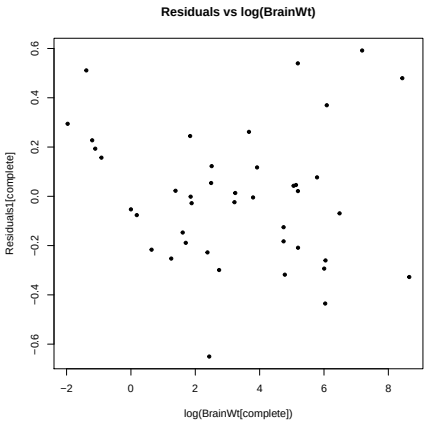
```
## log(LifeSpan)   0.08181    0.08374   0.977  0.33410
## log(Gestation) -0.15333    0.07339  -2.089  0.04264 *
## Exposure        0.03952    0.06316   0.626  0.53483
## Danger         -0.40056    0.12430  -3.223  0.00242 **
## Predation       0.18242    0.10947   1.666  0.10291
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2979 on 43 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.7366,Adjusted R-squared:  0.6937
## F-statistic: 17.18 on 7 and 43 DF,  p-value: 1.324e-10

par(mfrow=c(2,2))
plot(lm1)
```

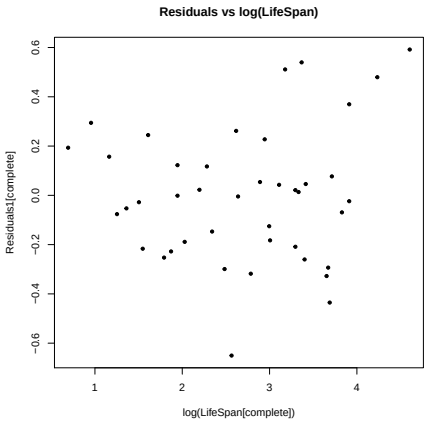


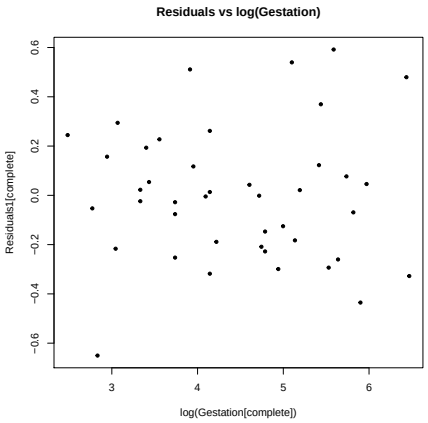
Now plot the residuals versus predictor variable plots:

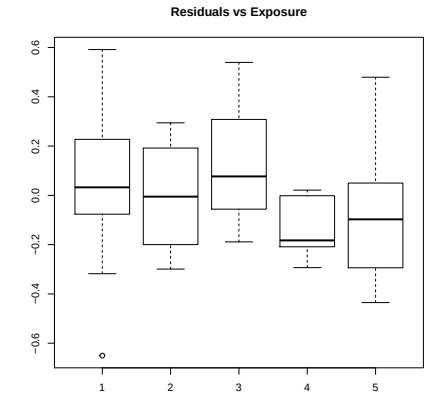
```
Residuals1=residuals(lm1)
# Because there are some missing values, the vector of residuals is not the
# same length as the original variables.
# First construct an indictor for complete observations (not missing
# any values) required for the regression
complete=!is.na(TotalSleep)&!is.na(BodyWt)&!is.na(BrainWt)&!is.na(LifeSpan)&
         !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)
boxplot(Residuals1[complete]~Exposure[complete],main="Residuals vs Exposure")
boxplot(Residuals1[complete]~Danger[complete],main="Residuals vs Danger")
boxplot(Residuals1[complete]~Predation[complete],main="Residuals vs Predation")
```

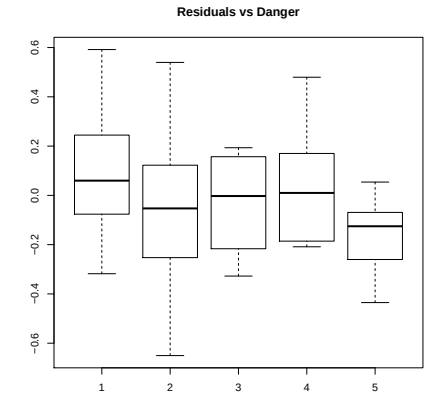

Residuals vs log(BodyWt)
Residuals1[complete]
log(BodyWt[complete])

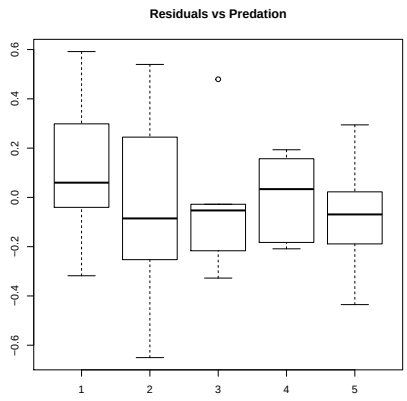

Residuals vs log(BrainWt)
Residuals1[complete]
log(BrainWt[complete])


Residuals vs log(LifeSpan)
Residuals1[complete]
log(LifeSpan[complete])


Residuals vs log(Gestation)
Residuals1[complete]
log(Gestation[complete])


Residuals vs Exposure


Residuals vs Danger


Residuals vs Predation

- The residuals vs fitted values plot shows roughly random scatter with no significant curvature; similarly, the individual plots of the residuals vs predictor variables all display no obvious curvature.
- The Scale-Location plot suggest mildly decreasing variance that could be a consequence of the fact the response variable is bounded; the individual plots of the residuals vs predictor variables also show evidence of mild heteroscedasticity, except possibly for `log(Gestation)` where the scatter looks roughly constant.
- The normal quantile plot shows no departures from normality.
- The residuals vs leverage plot shows no points of high leverage or high influence, with no points outside the Cook's distance contours of 0.5.
- Independence: We are not given information about the mammals, for example, whether they are living in the same geographical areas. If they were, the independence assumption could be violated.

Conclusion: Apart from possible decreasing variance, the assumptions of the multiple regression model appear reasonable.

(c) By successive removal of non-significant terms, find a parsimonious, well-fitting model for the data.

Solution:

```
summary(lm1)

##
## Call:
## lm(formula = log(TotalSleep) ~ log(BodyWt) + log(BrainWt) + log(LifeSpan) +
##     log(Gestation) + Exposure + Danger + Predation, data = sleep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65033 -0.19874  0.01991  0.16008  0.59184
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.23594    0.31270  10.349 3.02e-13 ***
## log(BodyWt)    -0.05018    0.05131  -0.978  0.33360
## log(BrainWt)   -0.01420    0.07661  -0.185  0.85379
## log(LifeSpan)   0.08181    0.08374   0.977  0.33410
## log(Gestation) -0.15333    0.07339  -2.089  0.04264 *
## Exposure        0.03952    0.06316   0.626  0.53483
## Danger         -0.40056    0.12430  -3.223  0.00242 **
## Predation       0.18242    0.10947   1.666  0.10291
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2979 on 43 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.7366,Adjusted R-squared:  0.6937
## F-statistic: 17.18 on 7 and 43 DF,  p-value: 1.324e-10

# Remove log(BrainWt)
lm2=lm(log(TotalSleep)~log(BodyWt)+log(LifeSpan)+log(Gestation)+Exposure+Danger+Predation,
       data=sleep)
summary(lm2)
```

```
##
## Call:
## lm(formula = log(TotalSleep) ~ log(BodyWt) + log(LifeSpan) +
##     log(Gestation) + Exposure + Danger + Predation, data = sleep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65753 -0.20461  0.01117  0.15827  0.58234
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.24402    0.30623  10.593  1.1e-13 ***
## log(BodyWt)    -0.05862    0.02332  -2.514  0.01569 *
## log(LifeSpan)   0.07432    0.07257   1.024  0.31134
## log(Gestation) -0.15900    0.06597  -2.410  0.02020 *
## Exposure        0.04111    0.06188   0.664  0.50996
## Danger         -0.39873    0.12254  -3.254  0.00219 **
## Predation       0.18027    0.10765   1.674  0.10113
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2946 on 44 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.7364,Adjusted R-squared:  0.7005
## F-statistic: 20.49 on 6 and 44 DF,  p-value: 2.812e-11

 # Remove Exposure
lm3=lm(log(TotalSleep)~log(BodyWt)+log(LifeSpan)+log(Gestation)+Danger+Predation,
       data=sleep)
summary(lm3)

##
## Call:
## lm(formula = log(TotalSleep) ~ log(BodyWt) + log(LifeSpan) +
##     log(Gestation) + Danger + Predation, data = sleep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63597 -0.22432  0.02788  0.16916  0.59071
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.19100    0.29381  10.861 3.68e-14 ***
## log(BodyWt)    -0.05376    0.02200  -2.443  0.01856 *
## log(LifeSpan)   0.08583    0.07003   1.226  0.22674
## log(Gestation) -0.14988    0.06413  -2.337  0.02393 *
## Danger         -0.37126    0.11463  -3.239  0.00226 **
## Predation       0.18191    0.10696   1.701  0.09587 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2928 on 45 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.7338,Adjusted R-squared:  0.7042
## F-statistic: 24.81 on 5 and 45 DF,  p-value: 6.554e-12

 # Remove log(LifeSpan)
lm4=lm(log(TotalSleep)~log(BodyWt)+log(Gestation)+Danger+Predation,
       data=sleep)
summary(lm4)

##
```

```
## Call:
## lm(formula = log(TotalSleep) ~ log(BodyWt) + log(Gestation) +
##     Danger + Predation, data = sleep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66037 -0.25134  0.03499  0.18033  0.59161
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.34930    0.26986  12.411   <2e-16 ***
## log(BodyWt)    -0.03854    0.01979  -1.948   0.0572 .
## log(Gestation) -0.14177    0.06302  -2.250   0.0290 *
## Danger         -0.31639    0.10011  -3.161   0.0027 **
## Predation       0.13256    0.09056   1.464   0.1496
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2981 on 49 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.7075,Adjusted R-squared:  0.6836
## F-statistic: 29.63 on 4 and 49 DF,  p-value: 1.528e-12

 # Remove Predation
lm5=lm(log(TotalSleep)~log(BodyWt)+log(Gestation)+Danger, data=sleep)
summary(lm5)

##
## Call:
## lm(formula = log(TotalSleep) ~ log(BodyWt) + log(Gestation) +
##     Danger, data = sleep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6855 -0.2470  0.0346  0.1370  0.5753
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.50847    0.24979  14.046  < 2e-16 ***
## log(BodyWt)    -0.04350    0.01972  -2.206  0.03203 *
## log(Gestation) -0.17275    0.06003  -2.878  0.00588 **
## Danger         -0.17691    0.03106  -5.696 6.47e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3015 on 50 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.6947,Adjusted R-squared:  0.6764
## F-statistic: 37.92 on 3 and 50 DF,  p-value: 6.325e-13
```

`log(BodyWt)`, `log(Gestation)` and `Danger` are all statistically significant so no further simplification is appropriate.

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

Solution:

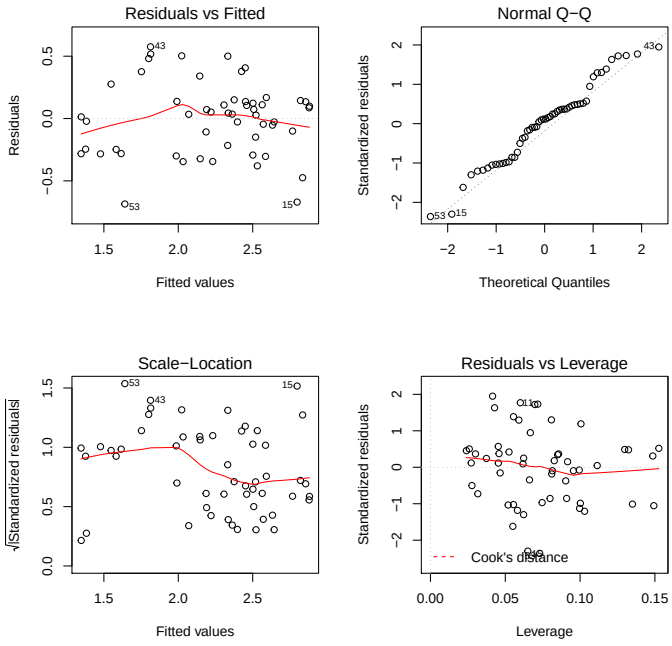
According to the multiple regression, the variables `BodyWt`, `Gestation` and `Danger` have a significant effect on `TotalSleep`.

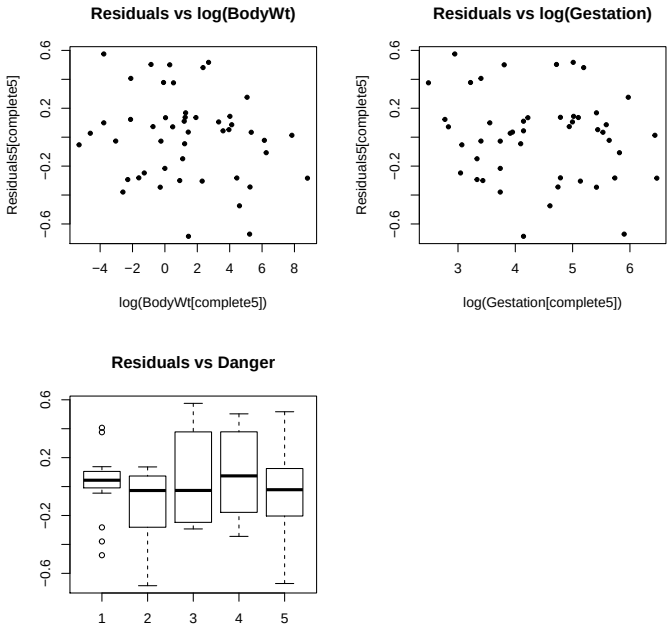
- An increase of one unit in `log(BodyWt)` is estimated to correspond to a 0.0435 unit (log hrs/day) decrease in `log(TotalSleep)` when `Gestation` and `Danger` are held fixed.
- An increase of one unit in `log(Gestation)` is estimated to correspond to 0.17275 a unit (log hours/day) decrease in `log(TotalSleep)` when `BodyWt` and `Danger` are held fixed.
- An increase of one level of `Danger` is estimated to correspond to a 0.17691 unit decrease in `log(TotalSleep)` (log hours/day) when `BodyWt` and `Gestation` are held fixed.

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

Solution:

```
#Generate diagnostic plots as previously
par(mfrow=c(2,2))
plot(lm5)
complete5=!is.na(TotalSleep)&!is.na(BodyWt)&!is.na(Gestation)&!is.na(Danger)
Residuals5=residuals(lm5)
plot(log(BodyWt[complete5]),Residuals5[complete5],main="Residuals vs log(BodyWt)",
     pch=20)
plot(log(Gestation[complete5]),Residuals5[complete5],main="Residuals vs log(Gestation)",
     pch=20)
boxplot(Residuals5[complete5]~Danger[complete5],main="Residuals vs Danger")
```





Comments: The conclusions are very similar to the initial model:

- The residuals vs fitted values plot and the residuals vs predictor variables plots show no significant curvature (that is, random scatter about the zero line).
- The Scale-Location plot suggests mildly decreasing variance that could be a consequence of the response variable being bounded; the `log(BodyWt)` versus residual plot shows very mild heteroscedasticity but the residual plot for `log(Gestation)` does not.
- The normal quantile plot is a little wobbly but overall the normality assumption looks reasonable.
- The residuals vs leverage plot shows no points of high leverage or high influence.
- The independence assumption comments are as before.

Conclusion: Apart from possible decreasing variance, the assumptions of the multiple regression model appear reasonable.

(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:

```
summary(lm5)

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

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6855 -0.2470  0.0346  0.1370  0.5753
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.50847    0.24979  14.046  < 2e-16 ***
## log(BodyWt)    -0.04350    0.01972  -2.206  0.03203 *
## log(Gestation) -0.17275    0.06003  -2.878  0.00588 **
## Danger         -0.17691    0.03106  -5.696 6.47e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3015 on 50 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.6947,Adjusted R-squared:  0.6764
## F-statistic: 37.92 on 3 and 50 DF,  p-value: 6.325e-13
```

```
lm6=lm(log(TotalSleep)~log(BodyWt)+log(Gestation)+Predation, data=sleep)
summary(lm6)
```

```
##
## Call:
## lm(formula = log(TotalSleep) ~ log(BodyWt) + log(Gestation) +
##     Predation, data = sleep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.77547 -0.26674  0.00213  0.15604  0.58586
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.64429    0.27502  13.251  < 2e-16 ***
## log(BodyWt)    -0.04988    0.02114  -2.359  0.02225 *
## log(Gestation) -0.21546    0.06359  -3.388  0.00138 **
## Predation      -0.13985    0.03017  -4.635 2.59e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3238 on 50 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.6478,Adjusted R-squared:  0.6267
## F-statistic: 30.66 on 3 and 50 DF,  p-value: 2.17e-11
```

`Comment:` The model `lm5` obtained on the basis of statistical significance could be taken naively to suggest that `Predation` has no influence on `TotalSleep`, if `Danger` and `Gestation` are held fixed. However, the model `lm6` shows almost as good a fit can be obtained by substituting `Predation` for `Danger`. In other words, statistical significance alone does not identify the factors that affect `TotalSleep`. One explanation is that `Danger` and `Predation` are coding similar information. Since `Danger` is derived from `Predation` and other information, a reasonable conclusion would be that `Predation` is major influence on `Total Sleep` with the other information playing a lesser but nevertheless statistically significant role.

[Total: 25]