STATS 3001 Statistical Modelling III
Assignment 2
Due: 4pm Monday 16th April (Week 6), 2018

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.

(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 $\alpha_1 = 0$.

(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 $\alpha_r = -(\alpha_1 + \alpha_2 + \ldots \alpha_{r-1})$.

(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.

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

[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 $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e}$.

(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.]

[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.

(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 can 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)
```

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

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

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

[Total: 25]
