STATS 3001 Statistical Modelling III
Assignment 3
Due: 4pm Friday 4$^{\text{th}}$ May (Week 8), 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) For $y > 0$ prove that

$$\lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} = \log y.$$

[Total: 3]

(2) Suppose

$$\boldsymbol{Y} = \boldsymbol{\eta} + \boldsymbol{\mathcal{E}},$$

where $\mathcal{E}_i \sim N(0, \sigma^2)$, $i = 1, 2, \ldots, n$, independently and let X be a $n \times p$ matrix with linearly independent columns. Show that

$$E\left(\|\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}\|^2\right) = (n-p)\sigma^2 + \|(I - P)\boldsymbol{\eta}\|^2,$$

where

$$P = X(X^T X)^{-1} X^T.$$

Under what condition is

$$E\left(\|\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}\|^2\right) = (n-p)\sigma^2?$$

[Total: 8]

(3) Use the result of Question 2 to justify the claim that Mallow's C_p will satisfy

$$C_p \approx p$$

for a correct model.

[Total: 5]

(4) The data in the file `prostate-a3.csv` are from a study by Stamey (1989) who examined the correlation between prostate specific antigen (PSA) and a number of clinical measures in 97 men who were about to undergo a radical prostatectomy. The aim of the analysis is to predict PSA from the clinical measures, which are described in the following table:

Variable	Description
`psa`	prostate specific antigen (*ng/ml*)
`lcavol`	log cancer volume (*cc*)
`lweight`	log prostate weight (*cm*)
`age`	in years
`lbph`	log of benign prostatic hyperplasia amount (cm^2)
`svi`	seminal vesicle invasion (1, 0 otherwise)
`lcp`	log of capsular penetration (*cm*)
`gleason`	Gleason scores $6, 7, 8$ or 9
`pgg45`	percent of Gleason scores 4 or 5

(a) Read the data into R.

(i) Plot the data in a pairwise scatterplot matrix.

(ii) Create a correlation matrix of all the predictor variables.

(iii) Comment on any observed relationships between `psa` and the predictor variables, and

(iv) comment on any observed relationships amongst the predictor variables.

(b) Use the Box-Cox method to find a suitable transformation of the response variable `psa` in the context of the following model:

```
lm(psa~lcavol+lweight+age+lbph+svi+lcp+gleason+pgg45)
```

State, with justification, your chosen value of λ.

(c) Re-fit the linear model in part (b) using the Box-Cox transformed response.

(d) Use Mallow's C_p and stepwise model selection to obtain the most appropriate reduced model for the transformed data.

(e) Obtain appropriate diagnostic plots for your selected model and present them neatly in your answers. Comment on whether your model is appropriate for the transformed data.

(f) Obtain a scatter plot of `psa` versus `lcavol` showing the back-transformed 95% prediction bands. Comment on whether or not they appear appropriate.

[Total: 28]