STATS 3001 Statistical Modelling III
Assignment 1
Due: 4pm Monday 19th March (Week 4), 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) Let A and B be matrices as given in Tutorial 1, Question 4. Furthermore, assume C is a constant $n \times n$ matrix. Prove that

$$\text{tr}(ABC) = \text{tr}(BCA)$$

[**Hint**: use the result from Q4, T1.]

[Total: 4]

(2) Let $\boldsymbol{Y}$ be a random $n \times 1$ vector with mean $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$, and let $\boldsymbol{A}$ be a constant $n \times n$ matrix. Show that

$$E(\boldsymbol{Y}^T \boldsymbol{A} \boldsymbol{Y}) = \text{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T \boldsymbol{A} \boldsymbol{\mu}.$$

[Total: 8]

(3) Consider the multiple linear regression model

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\mathcal{E}}.$$

Prove that $X^T X$ is invertible if and only if the columns of X are linearly independent.

[Total: 5]

(4) The data file `fev.txt` contains data from an early study on the deleterous effects of smoking in children and young adults. Measures of lung function (FEV, forced expiratory volume in litres per second) were made in 654 healthy children seen for a routine check up in a paediatric clinic. The children participating in this study were asked whether they were current smokers. A higher FEV is usually associated with better respiratory function and it is well known that prolonged smoking diminishes FEV in adults.

The following variables were recorded: 654 subjects aged 3-19.

Variable	Description
`ID`	ID number
`Age`	years
`FEV`	litres
`Height`	inches
`Sex`	`Male` or `Female`
`Smoker`	`Non` = nonsmoker, `Current` = current smoker

The purpose of your analysis is to relate lung capacity (as measured by `FEV`) to the predictor variables: `Sex`, `Height`, `Smoker` and `Age` using multiple linear regression. We can do this by fitting a model of the form:

$$f\left(\texttt{FEV}_i\right) = \beta_0 + \beta_1 \times \texttt{Sex}_i + \beta_2 \times \texttt{Height}_i + \beta_3 \times \texttt{Smoker}_i + \beta_4 \times \texttt{Age}_i + \mathcal{E}_i,$$

where f is some suitable transformation and $\mathcal{E}_i \overset{iid}{\sim} N(0, \sigma^2)$; $i = 1, 2, \ldots, 654$.

(a) In R, create an appropriate design matrix X using the function `model.matrix()`. Call this matrix `X` and provide the first 10 rows.

(b) In your matrix `X`, what are the values for the *categorical variables*? What levels of the variables do these values correspond to?

(c)
- Using the `pairs` command in R, create a scatterplot matrix for all the variables.
- Comment on the relationship between `FEV` and the continuous predictor variables.
- Repeat the scatterplot matrix but now using `log(FEV)` as the response variable. Comment again on the relationship between `FEV` and the continuous predictor variables.

(d) Obtain a set of diagnostic plots (and neatly include them in your answers) for each of the models

$$\texttt{log(FEV)} \sim X + 0$$

and

$$\texttt{FEV} \sim X + 0.$$

(Note the +0 in the formula ensures `R` does not automatically include an intercept term; the intercept is already accounted for in your design matrix.)

Which assumptions look better for the model fit on the log-transformed data compared to the raw data? In each case, justify your answer.

(e) Provide a careful interpretation of the coefficient for the predictor variable of `Sex` on the *orginal scale.*

(f) (i) Explain why it is possible to obtain a $100(1-\alpha)\%$ prediction interval for `log(FEV)`. *In your answer, give the form of the interval on each of the transformed and untransformed scales.*

(ii) Can you perform a similar calculation for a confidence interval for `log(FEV)`? Justify your answer.

[Total: 25]