STATS 2107
Statistical Modelling and Inference II
Assignment 4

Jono Tuke

Semester 2 2017

CHECKLIST

- ☐: Have you shown all of your working, including probability notation where necessary?
- ☐: Have you given all numbers to 3 decimal places?
- ☐: Have you included all R output and plots to support your answers where necessary?
- ☐: Have you included all of your R code?
- ☐: Have you made sure that all plots and tables each have a caption?
- ☐: If before the deadline, have you submitted your assignment via the online submission on MyUni?
- ☐: Is your submission a single pdf file - correctly orientated, easy to read? If not, penalties apply.
- ☐: Penalties for more than one document - 10% of final mark for each extra document. Note that you may resubmit and your final version is marked, but the final document should be a single file.
- ☐: Penalties for late submission - within 24 hours 40% of final mark. After 24 hours, assignment is not marked and you get zero.
- ☐: Assignments emailed instead of submitted by the online submission on MyUni will not be marked and will recieve zero.
- ☐: Have you checked that the assignment submitted is the correct one, as we cannot accept other submissions after the due date?

Due date: Friday 6th October 2017 (Week 9), 5pm.

Q1 Writing design matrices

A survey, conducted in a city by a local internet provider, collected data for the following variables from 35 randomly selected households.

- Y: number of hours of Internet used in a week
- X_1: number of children in the household
- X_2: net household income (in dollars) in the previous week
- X_3: number of years of formal education of the highest income earner in the household

The values for the first 4 households are shown below.

Household	1	2	3	4
Y	12.8	9.4	14	15.6
X_1	1.0	0.0	2	3.0
X_2	1590.0	968.0	732	780.0
X_3	15.0	11.0	12	13.0

Consider the multiple regression model

$$M: \; \boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $Y_i \sim N(\eta_i, \sigma^2)$ independently for $i = 1, 2, \ldots, n$ and $\boldsymbol{\eta} = X\boldsymbol{\beta}$.

(a) Write down the dimensions of $\boldsymbol{Y}$, X, $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$.

[4 marks]

(b) Write down $\boldsymbol{\beta}$ in full, and the first four rows of X and $\boldsymbol{y}$ (the vector of observed values).

[3 marks]

[Question total: 7]

Q2 Linear transformations of the design matrix

Suppose X is an $n \times p$ matrix with linearly independent columns.

Let $X^* = XA$, where A is an invertible $p \times p$ matrix.

(a) Show that the columns of X^* are also linearly independent.
[*Hint*: Prove by contradiction, *i.e.*, start by assuming the columns of X^* are **not** linearly independent.]

[3 marks]

(b) Show that $X^*(X^{*T}X^*)^{-1}X^{*T} = X(X^TX)^{-1}X^T$.

[3 marks]

(c) Consider two alternative models

$$M: \ \boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{and} \quad M^*: \ \boldsymbol{Y} = X^*\boldsymbol{\beta^*} + \boldsymbol{\epsilon}.$$

Show that $\hat{\boldsymbol{\eta}}^* = \hat{\boldsymbol{\eta}}$, *i.e.*, the vector of fitted values is the same, whatever the form of the design matrix X.

[3 marks]

[Question total: 9]

Q3 Matrix calculations in R

For this question, you may use R to perform the matrix calculations. Include your code for full marks.

An experiment was conducted to investigate the simultaneous influence of three predictor variables X_1, X_2, X_3 on a response variable Y. Data were recorded for a sample of seven subjects, as shown in the table below.

Subject	1	2	3	4	5	6	7
y	1	0	0	1	2	3	3
x_1	-3	-2	-1	0	1	2	3
x_2	5	0	-3	-4	-3	0	5
x_3	-1	1	1	0	-1	-1	1

Suppose the data satisfy the assumptions of the multiple regression model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$ independently for $i = 1, 2, \ldots, 7$.

Consider the matrix formulation of this model:

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

(a) Write down the design matrix, X, and the vector of observed values, $\boldsymbol{y}$, and enter them into R.

[3 marks]

(b) Use direct matrix calculations in R to find the least squares estimates given by

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}.$$

Recall, that `t(X)` gives the transpose of a matrix; `%*%` multiplies two matrices; and `solve()` can be used to find the inverse of a matrix

[1 mark]

(c) Continuing to use R for your calculations, find the predicted value of Y when $x_1 = 1$, $x_2 = -3$, $x_3 = -1$.

[1 mark]

(d) Test the null hypothesis that X_3 has no effect on Y, *i.e.*, test $H_0 : \beta_3 = 0$, as follows.

i. The test statistic takes the form:

$$T = \frac{\boldsymbol{\lambda}^T \hat{\boldsymbol{\beta}} - 0}{s_e \sqrt{\boldsymbol{\lambda}^T (X^T X)^{-1} \boldsymbol{\lambda}}} \quad \text{where } T \sim t_{n-p} \text{ if } H_0 \text{ is true.}$$

In this case, write down $\boldsymbol{\lambda}$, n, and p.

[3 marks]

ii. Calculate the observed value of the test statistic for this sample.

[*Hint*: Recall $s_e^2 = \frac{1}{n-p} \|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2 = \frac{1}{n-p} (\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^T (\boldsymbol{y} - X\hat{\boldsymbol{\beta}})$.]

[2 marks]

iii. Calculate the P-value, and hence state whether you reject or retain H_0 at significance level $\alpha = 0.05$.

[2 marks]

iv. Find a 95% confidence interval for the expected value of Y given $x_1 = 1$, $x_2 = -3$, $x_3 = -1$. *i.e.*, a 95% confidence interval for

$$\boldsymbol{\lambda}^T \boldsymbol{\beta} = \beta_0 + \beta_1 - 3\beta_2 - \beta_3,$$

where $\boldsymbol{\lambda}^T = (1, 1, -3, -1)$.

[3 marks]

[Question total: 15]

Q4 Rats versus cats question

For full marks this question must be typed up in latex, word or Rmarkdown.

For full marks, please include your code, your output and remember to caption all tables and figures.

In this question, you will perform a two-sample t-test to assess if Carnivora or Rodentia sleep on average longer.

(a) Read in the data (`msleep`).

[1 mark]

(b) Produce and include side-by-side boxplots of the sleep total for Carnivora and Rodentia. Describe the distributions.

[5 marks]

(c) Decide if a pooled two-sample t-test can be used. Give reasons.

[3 marks]

(d) Perform a two-sample t-test. For full marks include:
  (i) the null and alternative hypotheses,
  (ii) the value of the test statistic,
  (iii) the distribution of the test statistic if the null hypothesis is true,
  (iv) P-value,
  (v) and your conclusion.

[6 marks]

(e) Check the assumptions, including appropriate plots if neccesary.

[3 marks]

[Question total: 18]

[[Assignment total: 49]]