THE UNIVERSITY OF MELBOURNE

Student ID

Semester 1 Assessment, 2018

School of Mathematics and Statistics

MAST30025 Linear Statistical Models

Writing time: 3 hours

Reading time: 15 minutes

This is NOT an open book exam

This paper consists of 9 pages (including this page)

Authorised materials:

- The only permitted scientific calculator is the Casio FX82.
- Two A4 double-sided handwritten sheets of notes.

Instructions to Students

- You must NOT remove this question paper at the conclusion of the examination.
- You should attempt all questions. Marks for individual questions are shown.
- The total number of marks available is 90.

Instructions to Invigilators

- Students must NOT remove this question paper at the conclusion of the examination.

This paper may be held in the Baillieu Library

This paper must not be removed from the examination room

Question 1 (9 marks)

(a) Let $A_1, \ldots, A_m$ be a set of symmetric idempotent matrices with $A_i A_j = 0$ for $i \neq j$. Show directly that $\sum_{i=1}^{m} A_i$ is idempotent.

(b) Let $\mathbf{y}$ be a random vector with $E[\mathbf{y}] = \boldsymbol{\mu}$ and var $\mathbf{y} = V$. Show directly that var $\mathbf{c}^T\mathbf{y} = \mathbf{c}^T V \mathbf{c}$.

(c) Find a conditional inverse for the matrix

$$A = \begin{bmatrix} 5 & 5 & 0 & 6 \\ 9 & 13 & 2 & 14 \\ -4 & -8 & -2 & -8 \end{bmatrix}.$$

Question 2 (12 marks)

(a) Let $X \sim \chi^2_{k,\lambda}$. Show that $E[X] = k + 2\lambda$.

(b) Let $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim MVN(\boldsymbol{\mu}, V)$, where

$$\boldsymbol{\mu} = \begin{bmatrix} 4 \\ -1 \end{bmatrix}, \qquad V = \begin{bmatrix} 3 & -1 \\ -1 & 2 \end{bmatrix}.$$

Describe the distribution of $\begin{bmatrix} y_1 + y_2 \\ y_1 - y_2 \end{bmatrix}$.

(c) Describe the distribution of $2y_1^2 + 3y_2^2 + 2y_1y_2$.

(d) Using your answers to (a) and (c), find $E[2y_1^2 + 3y_2^2 + 2y_1y_2]$.

Question 3 (17 marks) In this question, we study a dataset of 50 US states. This dataset contains the variables:

- `Life.Exp`: life expectancy in years (1969–71)
- `Murder`: murder and non-negligent manslaughter rate per 100,000 population (1976)
- `HS.Grad`: percentage of high-school graduates (1970)

We wish to model life expectancy in terms of the other variables. The data are stored in the `statedata` data frame and the following R calculations performed:

```
> X <- cbind(rep(1,50), statedata$Murder, statedata$HS.Grad)
> y <- statedata$Life.Exp
> t(X) %*% X

       [,1]     [,2]      [,3]
[1,]   50.0   368.90   2655.40
[2,]  368.9  3389.49  18878.61
[3,] 2655.4 18878.61 144219.64

> solve(t(X)%*%X)

            [,1]          [,2]          [,3]
[1,]  1.62861066 -0.0377838755 -0.0250403183
[2,] -0.03778388  0.0019656212  0.0004383807
[3,] -0.02504032  0.0004383807  0.0004105962

> t(X)%*%y

          [,1]
[1,]   3543.93
[2,]  25957.51
[3,] 188520.36

> t(y)%*%y

         [,1]
[1,] 251277.1

> qt(0.975,45:50)

[1] 2.014103 2.012896 2.011741 2.010635 2.009575 2.008559

> qf(0.95,1,45:50)

[1] 4.056612 4.051749 4.047100 4.042652 4.038393 4.034310

> qf(0.95,2,45:50)

[1] 3.204317 3.199582 3.195056 3.190727 3.186582 3.182610
```

(a) Calculate the least squares estimates of $\boldsymbol{\beta}$, the parameters of the model.

(b) Calculate the sample variance s^2.

(c) Calculate a 95% confidence interval for the parameter corresponding to `Murder`.

(d) Calculate a 95% prediction interval for the life expectancy in a state with murder rate 7 per 100,000 population and 50% of high-school graduates.

(e) Test the relevance of the `HS.Grad` variable at the 5% level.

Question 4 (14 marks) Consider the full rank linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with p parameters.

(a) Calculate the variance of the residuals $\mathbf{e} = \mathbf{y} - X\mathbf{b}$.

(b) Ridge regression estimates the parameters $\boldsymbol{\beta}$ by minimising $\mathbf{e}^T\mathbf{e} + \lambda\mathbf{b}^T\mathbf{b}$. Derive an expression for the resulting estimators $\mathbf{b}$.

(c) Calculate the expected value of SS_{Reg}, the regression sum of squares.

(d) Consider two nested models $\mathbf{y} = X_2\boldsymbol{\gamma}_2 + \boldsymbol{\varepsilon}_2$ and $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \end{bmatrix}$ and $\boldsymbol{\gamma}_i$ contains p_i parameters ($i = 1, 2$). It can be shown that the second model has a larger AIC if and only if

$$\frac{SS_{Res}(\boldsymbol{\gamma}_2)}{SS_{Res}(\boldsymbol{\beta})} \leq c,$$

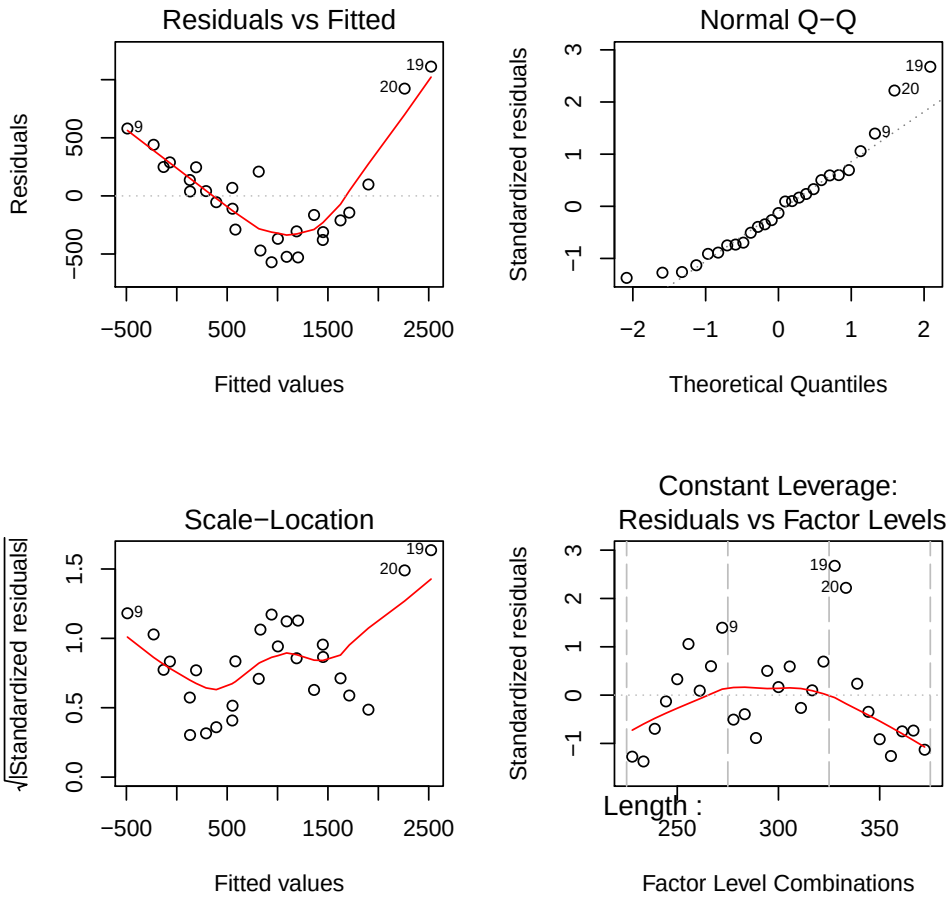
for some constant c. Find c.

Question 5 (16 marks) An experiment was conducted to understand the strength of wool as a function of three factors. The variables measured are:

- `Length`: Length of test specimen (200, 300, 350 mm)
- `Amplitude`: Amplitude of loading cycle (8, 9, 10 mm)
- `Load`: Load put on the specimen (40, 45, 50 g)
- `Cycles`: Number of cycles until the specimen fails

One sample was measured for each combination of `Length`, `Amplitude` and `Load` (i.e., 27 samples in total). The data is analysed below.

```
> wool <- read.csv('wool.csv', header=T)
> wool$Length <- factor(wool$Length)
> wool$Amplitude <- factor(wool$Amplitude)
> wool$Load <- factor(wool$Load)
> model1 <- lm(Cycles ~ ., data = wool)
> par(mfrow=c(2,2))
> plot(model1)
```



```
> model2 <- lm(log(Cycles) ~ ., data = wool)
> summary(model2)

Call:
lm(formula = log(Cycles) ~ ., data = wool)

Residuals:
     Min       1Q   Median       3Q      Max
-0.36860 -0.13002  0.00902  0.10129  0.30469

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.48287    0.09644  67.225  < 2e-16 ***
Length300    0.91833    0.08928  10.286 1.97e-09 ***
Length350    1.66477    0.08928  18.646 4.10e-14 ***
Amplitude9  -0.65521    0.08928  -7.339 4.31e-07 ***
Amplitude10 -1.26173    0.08928 -14.132 7.19e-12 ***
Load45      -0.32529    0.08928  -3.643  0.00162 **
Load50      -0.78524    0.08928  -8.795 2.62e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1894 on 20 degrees of freedom
Multiple R-squared:  0.9691,        Adjusted R-squared:  0.9598
F-statistic: 104.5 on 6 and 20 DF,  p-value: 4.979e-14

> anova(model2)

Analysis of Variance Table

Response: log(Cycles)
          Df  Sum Sq Mean Sq F value    Pr(>F)
Length     2 12.5159  6.2579 174.456 2.193e-13 ***
Amplitude  2  7.1674  3.5837  99.905 3.889e-11 ***
Load       2  2.8019  1.4010  39.055 1.239e-07 ***
Residuals 20  0.7174  0.0359
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> model3 <- lm(log(Cycles) ~ .^2, data = wool)
> anova(model2, model3)

Analysis of Variance Table

Model 1: log(Cycles) ~ Length + Amplitude + Load
Model 2: log(Cycles) ~ (Length + Amplitude + Load)^2
  Res.Df     RSS Df Sum of Sq     F Pr(>F)
1     20 0.71742
2      8 0.16591 12   0.55151 2.216 0.1325
```

```
> model4 <- step(model3)

Start:  AIC=-99.49
log(Cycles) ~ (Length + Amplitude + Load)^2

                   Df Sum of Sq     RSS      AIC
- Amplitude:Load    4   0.01460 0.18051 -105.211
<none>                          0.16591  -99.487
- Length:Load       4   0.13575 0.30167  -91.345
- Length:Amplitude  4   0.40116 0.56707  -74.304

Step:  AIC=-105.21
log(Cycles) ~ Length + Amplitude + Load + Length:Amplitude +
    Length:Load

                   Df Sum of Sq     RSS      AIC
<none>                          0.18051 -105.211
- Length:Load       4   0.13575 0.31626  -98.069
- Length:Amplitude  4   0.40116 0.58167  -81.618

> summary(model4)

Call:
lm(formula = log(Cycles) ~ Length + Amplitude + Load + Length:Amplitude +
    Length:Load, data = wool)

Residuals:
      Min        1Q    Median        3Q       Max
-0.153728 -0.055232 -0.008017  0.067786  0.175706

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            6.384806   0.091416  69.843  < 2e-16 ***
Length300              0.913780   0.129282   7.068 1.30e-05 ***
Length350              1.963516   0.129282  15.188 3.37e-09 ***
Amplitude9            -0.449946   0.100142  -4.493 0.000735 ***
Amplitude10           -1.232398   0.100142 -12.307 3.65e-08 ***
Load45                -0.401464   0.100142  -4.009 0.001734 **
Load50                -0.649468   0.100142  -6.485 3.00e-05 ***
Length300:Amplitude9  -0.001114   0.141622  -0.008 0.993851
Length350:Amplitude9  -0.614678   0.141622  -4.340 0.000961 ***
Length300:Amplitude10  0.064964   0.141622   0.459 0.654638
Length350:Amplitude10 -0.152966   0.141622  -1.080 0.301328
Length300:Load45       0.083463   0.141622   0.589 0.566565
Length350:Load45       0.145059   0.141622   1.024 0.325914
Length300:Load50      -0.133655   0.141622  -0.944 0.363913
Length350:Load50      -0.273658   0.141622  -1.932 0.077269 .
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1226 on 12 degrees of freedom
Multiple R-squared:  0.9922,        Adjusted R-squared:  0.9831
F-statistic: 109.3 on 14 and 12 DF,  p-value: 1.968e-10

> qt(0.975,20:27)

[1] 2.085963 2.079614 2.073873 2.068658 2.063899 2.059539 2.055529 2.051831
```

(a) Identify the features in the diagnostic plots which support the use of the logarithmic transformation on the `Cycles` variable.

(b) From the additive model, calculate a 95% confidence interval for the average ratio of the number of cycles to failure for 50g loads against 40g loads. (*Hint: The logarithm of the ratio is the difference of the logarithms.*)

(c) From the additive model, test whether length has an effect on wool strength, at the 5% significance level.

(d) Calculate the change in AIC if the amplitude variable was removed from the additive model.

(e) Test for the presence of 2-way interaction between the factors.

(f) Is your answer above consistent with the results of the variable selection? Why or why not?

(g) Using the model resulting from variable selection, calculate a point estimate for the average number of cycles to failure for a wool specimen of length 350mm, loading cycle amplitude 8mm, with 45g load.

Question 6 (14 marks) Consider the general linear model, $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. This model may be of full or less than full rank.

(a) Explain the difference between an error and a residual.

(b) Define and explain the purpose of the standardised residual of a point.

(c) When is a model with fewer explanatory variables more desirable than a model with more explanatory variables? When is it less desirable?

(d) State the general linear hypothesis and explain how it is tested.

(e) Define a treatment contrast and explain its usage.

(f) Explain what randomisation is and its use in experimental design.

(g) Explain what a Latin square is and its use in experimental design.

Question 7 (8 marks) An experiment compares four different mixtures of the components of a rocket propellant; the mixtures contain different proportions of oxidizer, fuel, and binder. To compare the mixtures, five different samples of propellant are prepared for each mixture. Each of five investigators is randomly assigned one sample of each of the four mixtures and is asked to measure the propellant thrust. The data is given below:

Mixture	Investigator					Mixture
	1	2	3	4	5	Total
A	2340	2355	2362	2350	2348	11755
B	2658	2650	2665	2640	2653	13266
C	2449	2458	2432	2437	2445	12221
D	2403	2410	2418	2397	2405	12033
Investigator Total	9850	9873	9877	9824	9851	

(a) What type of experimental design is described above?

(b) Which are the treatment and blocking variables in this experiment?

(c) Is it better to analyse this data as a complete block design or completely randomised design? Justify your answer.

(d) A larger experiment is planned with the goal of testing whether mixture D, a newly developed formula, is more effective than industry standard mixtures A and B. This experiment has resources to prepare 100 samples of propellant. Calculate the best number of samples for each mixture. (*Hint: In a completely randomised design with treatment effects* τ_i*, we have* $var\ \tau_i = \frac{\sigma^2}{n_i}$*. To minimise a function* $f(\mathbf{x})$ *under the constraint* $g(\mathbf{x}) = c$*, minimise* $f(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda(g(\mathbf{x}) - c)$.)

End of Exam—Total Available Marks = 90.