THE UNIVERSITY OF MELBOURNE

Student number	

Semester 1 Assessment, 2021

School of Mathematics and Statistics

MAST30025 Linear Statistical Models

Reading time: 30 minutes — Writing time: 3 hours — Upload time: 30 minutes

This exam consists of 23 pages (including this page)

Permitted Materials

- This exam and/or an offline electronic PDF reader, one or more copies of the masked exam template made available earlier, blank loose-leaf paper and a Casio FX-82 calculator.
- One double sided A4 page of notes (handwritten only).

Instructions to Students

- If you have a printer, print the exam one-sided. If you cannot print, download the exam to a second device and disconnect that device from the internet.
- Ask the supervisor if you want to use the device running Zoom.
- Check your scanned PDF before submitting.
- You must submit while in the Zoom room and no submissions will be accepted after you have left the Zoom room.

Writing

- There are 7 questions with marks as shown. The total number of marks available is 90.
- Write your answers in the boxes provided on the exam that you have printed or the masked exam template that has been previously made available. If you need more space, you can use blank paper. Note this in the answer box, so the marker knows. The extra pages can be added to the end of the exam to scan.
- If you have been unable to print the exam and do not have the masked template write your answers on A4 paper. The first page should contain only your student number, the subject code and the subject name. Write on one side of each sheet only. Start each question on a new page and include the question number at the top of each page.

Scanning

- Put the pages in number order and the correct way up. Add any extra pages to the end. Use a scanning app to scan all pages to PDF. Scan directly from above. Crop pages to A4. Check PDF is readable.

Submitting

- **You must submit while in the Zoom room.** No submissions will be accepted after you have left the Zoom room.
- Go to the Gradescope window. Choose the Canvas assignment for this exam. Submit your file as a single PDF document only. Get Gradescope confirmation on email. Tell your supervisor it is submitted.

Question 1 (10 marks)

Let $A = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}$ and $B = \begin{bmatrix} 5 & 4 \\ 10 & 8 \end{bmatrix}$.

(a) [**3 marks**] Show that A is positive definite.

(b) [**3 marks**] Find all possible values of c such that $c(A - I)$ is idempotent.

(c) **[2 marks]** Find a conditional inverse for B.

(d) **[2 marks]** Let $\mathbf{y} = (y_1, y_2)^T$. Find $\frac{\partial \mathbf{y}^T B \mathbf{y}}{\partial \mathbf{y}}$ in terms of y_1 and y_2.

Question 2 (13 marks)

Let

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \sim MVN\left(\begin{bmatrix} 4 \\ 2 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} \right),$$

and $\mathbf{z} = (y_1, y_2)^T \sim MVN(\boldsymbol{\mu}, V)$.

(a) [**2 marks**] Find $\boldsymbol{\mu}$ and V.

(b) [**3 marks**] Describe the distribution of $\begin{bmatrix} y_1 + y_2 \\ y_1 - 2y_2 \end{bmatrix}$.

Page 5 of 23 — add any extra pages after page 23 — Page 5 of 23

(c) **[4 marks]** Describe the distribution of $\frac{2}{3}(y_1^2 + y_2^2 + y_1 y_2)$.

(d) **[4 marks]** Describe the distribution of $\frac{2}{3}(y_1^2 + y_2^2 + y_1 y_2) + \frac{1}{2}y_3^2$.

Question 3 (17 marks)

In this question, we study the number of sales of hot dogs in a stadium over 20 days. This dataset contains the variables:

- `sale` (y): the number of hot dogs sold on the day
- `temp` (x_1): the maximum temperature on the day
- `nppl` (x_2): the number of people in the stadium on the day (in thousands).

The data are stored in the `saledata` data frame and the following R calculations performed:

```
> X = cbind(rep(1,n), saledata$temp, saledata$nppl)
> y = saledata$sales
>
> t(X)%*%X

     [,1]   [,2] [,3]
[1,] 20.0   -4.8 33.8
[2,] -4.8 1104.8  6.6
[3,] 33.8    6.6 70.0

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

       [,1]     [,2]    [,3]
[1,]  0.278  0.00202 -0.1345
[2,]  0.002  0.00092 -0.0011
[3,] -0.134 -0.00106  0.0794

> t(X)%*%y

     [,1]
[1,]  319
[2,] 2381
[3,]  661

> t(y)%*%y

      [,1]
[1,] 11271

> qt(0.975,15:20)
[1] 2.131 2.120 2.110 2.101 2.093 2.086
> qf(0.95,1,15:20)
[1] 4.543 4.494 4.451 4.414 4.381 4.351
> qf(0.95,2,15:20)
[1] 3.682 3.634 3.592 3.555 3.522 3.493
> qf(0.95,3,15:20)
[1] 3.287 3.239 3.197 3.160 3.127 3.098
```

(a) **[2 marks]** Calculate the least squares estimates of the parameters.

(b) **[2 marks]** Calculate the sample variance s^2.

(c) **[2 marks]** Calculate a 95% confidence interval for the parameter corresponding to `temp`.

(d) **[3 marks]** Test for model relevance at the 5% significance level.

(e) **[4 marks]** Consider the linear model without an intercept,

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon.$$

Calculate the least squares estimates of the parameters.

(f) **[4 marks]** Test the hypothesis $H_0 : \beta_0 = 0$ in the full model (including the intercept), using an F-test at the 5% significance level.

Question 4 (8 marks)

Consider the general full rank linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. For given inputs $\mathbf{x}^*$, we wish to predict the difference of two responses with these inputs,

$$y_1^* = (\mathbf{x}^*)^T\boldsymbol{\beta} + \varepsilon_1^*,$$
$$y_2^* = (\mathbf{x}^*)^T\boldsymbol{\beta} + \varepsilon_2^*,$$

where $\varepsilon_1^* \sim N(0, \sigma^2)$, $\varepsilon_2^* \sim N(0, \sigma^2)$ and is independent of ε_1^*. Let $\mathbf{b}$ be the least squares estimator of $\boldsymbol{\beta}$ and s^2 be the sample variance.

(a) **[1 mark]** Write down the predictor of $y_1^* - y_2^*$.

(b) **[2 marks]** Find the distribution of the prediction error.

(c) [**3 marks**] Formulate a t-distributed random variable based on the prediction error, and specify its degrees of freedom.

(d) [**2 marks**] Based on (c), construct a $100(1-\alpha)\%$ prediction interval for $y_1^* - y_2^*$.

Question 5 (16 marks)

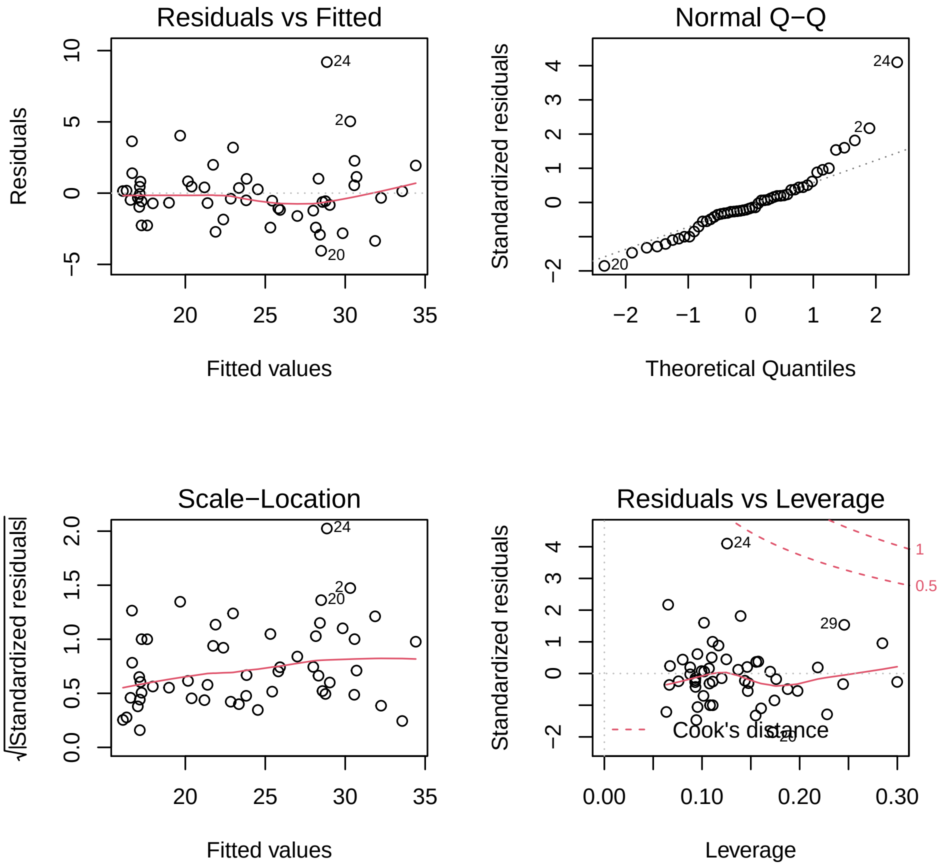
A small university collected data on the salary of its faculty members in the early 1980s. The dataset contains the following variables:

- `degree`: Highest qualification (Masters or PhD)
- `rank`: Faculty position (Asst, Assoc, or Prof)
- `sex`: Gender (Male or Female)
- `year`: Years in current rank
- `ysdeg`: Years since highest degree
- `salary`: Current salary ($k/yr)

We wish to model the salary of the faculty members in terms of the other variables. The following R calculations are produced:

```
> model <- lm(salary ~ ., data=salary)
> model2 <- lm(salary ~ . * sex, data=salary)
> anova(model, model2)

Analysis of Variance Table

Model 1: salary ~ degree + rank + sex + year + ysdeg
Model 2: salary ~ (degree + rank + sex + year + ysdeg) * sex
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     45 258.86
2     40 241.45  5    17.406 0.5767 0.7174

> summary(model)

Call:
lm(formula = salary ~ ., data = salary)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0452 -1.0947 -0.3615  0.8132  9.1931

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.74605    0.80018  19.678  < 2e-16 ***
degreePhD    1.38861    1.01875   1.363    0.180
rankAssoc    5.29236    1.14540   4.621 3.22e-05 ***
rankProf    11.11876    1.35177   8.225 1.62e-10 ***
sexFemale    1.16637    0.92557   1.260    0.214
year         0.47631    0.09491   5.018 8.65e-06 ***
ysdeg       -0.12457    0.07749  -1.608    0.115
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.398 on 45 degrees of freedom
Multiple R-squared:  0.855,       Adjusted R-squared:  0.8357
F-statistic: 44.24 on 6 and 45 DF,  p-value: < 2.2e-16
```

```
> par(mfrow=c(2,2))
> plot(model)
```



```
> model3 <- step(model)

Start:  AIC=97.46
salary ~ degree + rank + sex + year + ysdeg

         Df Sum of Sq    RSS     AIC
- sex     1      9.13 267.99  97.265
<none>                258.86  97.462
- degree  1     10.69 269.55  97.566
- ysdeg   1     14.87 273.73  98.366
- year    1    144.87 403.73 118.574
- rank    2    399.79 658.65 142.025

Step:  AIC=97.27
salary ~ degree + rank + year + ysdeg

         Df Sum of Sq    RSS     AIC
- degree  1      6.68 274.68  96.547
- ysdeg   1      7.87 275.87  96.771
<none>                267.99  97.265
- year    1    147.64 415.64 118.085
- rank    2    404.11 672.10 141.077

Step:  AIC=96.55
salary ~ rank + year + ysdeg

        Df Sum of Sq    RSS     AIC
- ysdeg  1      2.31 276.99  94.983
<none>               274.68  96.547
- year   1    141.11 415.78 116.104
- rank   2    478.54 753.22 145.002

Step:  AIC=94.98
salary ~ rank + year

       Df Sum of Sq    RSS     AIC
<none>              276.99  94.983
- year  1    161.95 438.95 116.923
- rank  2    632.06 909.05 152.780
```

```
> summary(model3)

Call:
lm(formula = salary ~ rank + year, data = salary)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4620 -1.3028 -0.2992  0.7835  9.3816

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.20327    0.63868  25.370  < 2e-16 ***
rankAssoc    4.26228    0.88289   4.828 1.45e-05 ***
rankProf     9.45452    0.90583  10.437 6.12e-14 ***
year         0.37570    0.07092   5.298 2.90e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.402 on 48 degrees of freedom
Multiple R-squared:  0.8449,       Adjusted R-squared:  0.8352
F-statistic: 87.15 on 3 and 48 DF,  p-value: < 2.2e-16

> linearHypothesis(model3, c(0,-1,1,-10), 0)

Linear hypothesis test

Hypothesis:
- rankAssoc  + rankProf - 10 year = 0

Model 1: restricted model
Model 2: salary ~ rank + year

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     49 284.38
2     48 276.99  1    7.3901 1.2806 0.2634

> person <- data.frame(degree='PhD', rank='Asst', sex='Female', year=4, ysdeg=10)
> predict(model3, person, interval='confidence', level=0.95)

       fit      lwr      upr
1 17.70605 16.56736 18.84474

> qt(0.975,45:50)

[1] 2.014103 2.012896 2.011741 2.010635 2.009575 2.008559

> qf(0.95,2,45:50)

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

(a) **[4 marks]** Based on the diagnostic plots, comment on the fit of the linear model. Suggest a possible transformation for consideration that might improve the fit.

(b) **[2 marks]** Interpret the output of the `anova` function in the context of the study.

(c) **[2 marks]** Calculate a confidence interval for the average difference in salary between females and males.

(d) **[2 marks]** Interpret the output of the `step` function in the context of the study.

(e) [**2 marks**] Interpret the output of the `linearHypothesis` function in the context of the study.

(f) [**4 marks**] Based on `model3`, calculate a 95% *prediction* interval for the salary of a female assistant professor with a PhD for 10 years, who has been in her position for 4 years.

Question 6 (12 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) **[2 marks]** State the distributional assumption used in fitting a linear model and explain how it may be verified.

(b) **[2 marks]** Define the hat matrix and explain its importance in the theory of linear models.

(c) **[2 marks]** State the correction factor for a corrected sum of squares, and explain its use in testing for model relevance.

(d) **[2 marks]** Outline the similarities and differences between Akaike's information criterion and the Bayesian information criterion.

(e) **[2 marks]** State two differences between a conditional inverse and a regular inverse.

(f) **[2 marks]** Define interaction between two categorical predictors, and explain how to model it.

Question 7 (14 marks)

In this question, we consider the balanced incomplete block design. We use the standard notation that I_n is the $n \times n$ identity matrix and J_n is the $n \times n$ matrix with all elements equal to 1.

(a) **[2 marks]** Under what circumstances should one prefer a balanced incomplete block design to a complete block design, and vice versa?

(b) **[2 marks]** Write down a design matrix and parameter vector for a balanced incomplete block design with 3 treatments and 3 blocks, each of size 2. Ensure that subjects in the same block are placed consecutively.

(c) [**3 marks**] Calculate the reduced design matrix $X_{2|1}$ for your design. You are given that the hat matrix corresponding to the nuisance parameters has the block structure

$$H_1 = X_1(X_1^T X)^c X_1^T = \frac{1}{2}\begin{bmatrix} J_2 & 0 & 0 \\ 0 & J_2 & 0 \\ 0 & 0 & J_2 \end{bmatrix}.$$

(d) **[4 marks]** It can be shown that the reduced design matrix $X_{2|1}$ for this model satisfies

$$X_{2|1}^T X_{2|1} = \frac{3}{2}\left[I_3 - \frac{1}{3}J_3\right].$$

Show that

$$(X_{2|1}^T X_{2|1})^c = cI$$

for some constant c, and find the value of c.

(e) **[3 marks]** Using your answers above, show directly that the reduced normal equations for this design does not have the same solution as the normal equations for a completely randomised design of 6 experimental units over 3 treatments.

End of Exam — Total Available Marks = 90