THE UNIVERSITY OF MELBOURNE

Student number

Semester 1 Assessment, 2023

School of Mathematics and Statistics

MAST30025 Linear Statistical Models Assignment 1

Submission deadline: **Friday March 24, 5pm**

This assignment consists of 9 pages (including this page) with 5 questions and 33 total marks

Instructions to Students

Writing

- This assignment is worth 6% of your total mark.
- You may choose to either typeset your assignment in LaTeX, or handwrite and scan it to produce an electronic version.
- You may use R for this assignment, but for matrix calculations only (you may not use the `lm` function). If you do, include your R commands and output.
- Write your answers on A4 paper. Page 1 should only have your student number, the subject code and the subject name. Write on one side of each sheet only. Each question should be on a new page. The question number must be written at the top of each page.

Scanning and Submitting

- Put the pages in question order and all the same way up. Use a scanning app to scan all pages to PDF. Scan directly from above. Crop pages to A4.
- Submit your scanned assignment as a single PDF file and carefully review the submission in Gradescope. Scan again and resubmit if necessary.

Question 1 (4 marks)

Prove that if a symmetric matrix A has eigenvalues which are all either 0 or 1, it is idempotent.

Diagonalise A and write it as

$$A = PDP^T.$$

Now D has only 0 or 1 on the diagonal, and hence $D^2 = D$. So

$$\begin{aligned} A^2 &= PDP^T PDP^T \\ &= PD^2P^T \\ &= PDP^T = A. \end{aligned}$$

Question 2 (6 marks)

We wish to prove (without using Theorem 2.5) that if A, B, and $A+B$ are $n \times n$ idempotent matrices, then $AB = BA = 0$.

(a) Show that $AB + BA = 0$.

(b) By Theorem 2.2, there exists a matrix P which diagonalises A:

$$P^T A P = D = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix},$$

where $r = r(A)$.

Write

$$P^T B P = \Lambda = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix},$$

where the block matrices are of the same dimensions as above. Show that $\Lambda_{11} = 0$, $\Lambda_{12} = 0$ and $\Lambda_{21} = 0$.

(c) Show that $AB = BA = 0$.

(a) By the idempotence of A, B, and $A+B$,

$$A + B = (A+B)^2 = A^2 + B^2 + AB + BA = A + B + AB + BA,$$

and therefore

$$AB + BA = 0.$$

(b)

$$\begin{aligned} 0 &= PDP^T P\Lambda P^T + P\Lambda P^T PDP^T \\ &= PD\Lambda P^T + P\Lambda D P^T \\ 0 &= D\Lambda + \Lambda D = \begin{bmatrix} 2\Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & 0 \end{bmatrix}. \end{aligned}$$

Therefore, $\Lambda_{11} = 0$, $\Lambda_{12} = 0$, and $\Lambda_{21} = 0$.

(c)

$$\begin{aligned} D\Lambda &= \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \Lambda_{21} \end{bmatrix} = 0 \\ AB &= PD\Lambda P^T = 0, \end{aligned}$$

and likewise $BA = 0$.

Question 3 (4 marks)

Show directly that for any random vector $\mathbf{y}$ and compatible matrix A, we have $\text{var } A\mathbf{y} = A(\text{var } \mathbf{y})A^T$.

Let $\boldsymbol{\mu} = E[\mathbf{y}]$. From the definition,

$$\begin{aligned}
\text{var } A\mathbf{y} &= E[(A\mathbf{y} - A\boldsymbol{\mu})(A\mathbf{y} - A\boldsymbol{\mu})^T] \\
&= E[A(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^T A^T] \\
&= A\, E[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^T]\, A^T \\
&= A(\text{var } \mathbf{y})A^T.
\end{aligned}$$

Question 4 (7 marks)

Let $\mathbf{y} = (y_1, y_2, y_3)^T$ be a 3-dimensional multivariate normal random vector with mean and variance

$$\boldsymbol{\mu} = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}, \quad V = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{bmatrix}.$$

(a) Find the eigenvalues and eigenvectors of V.

(b) Find the distribution of $z_1 = 3y_1 + 2y_2 + y_3$.

(c) Find the distribution of $z_2 = y_1^2 + \left(\frac{y_2+y_3}{2}\right)^2 + \left(\frac{y_2-y_3}{2}\right)^2$.

(a) The eigenvalues are $\lambda_1 = 2$, $\lambda_2 = 1$ and $\lambda_3 = 0$, and corresponding eigenvectors are $\mathbf{x}_1 = (0, 1, -1)^T$, $\mathbf{x}_2 = (1, 0, 0)^T$, and $\mathbf{x}_3 = (0, 1, 1)^T$.

(b) Write $z_1 = \mathbf{b}^T\mathbf{y}$, where $\mathbf{b} = (3, 2, 1)^T$. This has a normal distribution with mean $\mathbf{b}^T\boldsymbol{\mu} = 11$ and variance $\mathbf{b}^T V\mathbf{b} = 10$.

```
> mu <- c(1,2,4)
> V <- matrix(c(1,0,0,0,1,-1,0,-1,1),3,3)
> b <- c(3,2,1)
> t(b)%*%mu

     [,1]
[1,]   11

> t(b)%*%V%*%b

     [,1]
[1,]   10
```

(c) From the mean and variance of $\mathbf{y}$, it can be seen that $y_3 = 6 - y_2$, so

$$\left(\frac{y_1 + y_3}{2}\right)^2 = 9.$$

Likewise, the vector $(y_1, \frac{y_2-y_3}{2})^T$ has a multivariate normal distribution with mean $(1, -1)^T$ and variance I_2. So the sum of squares of its elements has a non-central χ^2 distribution with 2 degrees of freedom and non-centrality parameter

$$\frac{1}{2}\begin{bmatrix} 1 & -1 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \end{bmatrix} = 1.$$

So z_2 has a distribution of 9 plus a $\chi^2_{2,1}$ distribution.

Note: The original intended solution was as follows:

Write $z_2 = \mathbf{y}^T A\mathbf{y}$, where $A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix}$. Since AV is an idempotent matrix with rank $= 2$, z_2 has a non-central χ^2 distribution with 2 degrees of freedom and noncentrality parameter 5.5.

```
> A <- matrix(c(1,0,0,0,1/2,0,0,0,1/2),3,3)
> round(A %*% V - A %*% V %*% A %*% V, 5)

     [,1] [,2] [,3]
[1,]    0    0    0
[2,]    0    0    0
[3,]    0    0    0

> t(mu)%*%A%*%mu/2

     [,1]
[1,]  5.5
```

It turns out that Theorem 3.8 doesn't work for a singular variance matrix — but we gave full marks for that anyway.

Question 5 (12 marks)

A secondary school teacher wants to know if the marks of students in Specialist Mathematics can be predicted from their marks in General Mathematics. A linear model is assumed, and the following data is obtained from nine students:

ID	General Mathematics	Specialist Mathematics
1	86	85
2	85	97
3	89	76
4	82	79
5	84	76
6	86	99
7	84	49
8	78	72
9	92	83

(a) Write down the linear model as a matrix equation, writing out the matrices in full.

(b) Calculate the least squares estimate of the parameters.

(c) Calculate the sample variance s^2.

(d) Calculate the standardised residuals for all students.

(e) Calculate the Cook's distances for all students.

(f) Predict (using a point estimate) the mark of Specialist Mathematics for a student whose mark for General Mathematics is 90.

(a) $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where

$$\mathbf{y} = \begin{bmatrix} 85 \\ 97 \\ 76 \\ 79 \\ 76 \\ 99 \\ 49 \\ 72 \\ 83 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 86 \\ 1 & 85 \\ 1 & 89 \\ 1 & 82 \\ 1 & 84 \\ 1 & 86 \\ 1 & 84 \\ 1 & 78 \\ 1 & 92 \end{bmatrix},$$

and $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ are obvious.

(b)
```
> n <- 9
> p <- 2
> y <- c(85, 97, 76, 79, 76, 99, 49, 72, 83)
> X <- cbind(1, c(86, 85, 89, 82, 84, 86, 84, 78, 92))
> (b <- solve(t(X)%*%X,t(X)%*%y))

           [,1]
[1,] -3.2451839
[2,]  0.9728546
```

(c)
```
> e <- y - X%*%b
> SSRes <- sum(e^2)
> (s2 <- SSRes/(n-p))

[1] 231.447
```

(d)
```
> H <- X %*% solve(t(X)%*%X) %*% t(X)
> (z <- e / sqrt(s2 * (1 - diag(H))))

             [,1]
 [1,]  0.32041485
 [2,]  1.22381070
 [3,] -0.54984795
 [4,]  0.18018732
 [5,] -0.17347884
 [6,]  1.29991629
 [7,] -2.06627662
 [8,] -0.05983821
 [9,] -0.29839771
```

(e)
```
> (D <- 1/p * z^2 * diag(H) / (1-diag(H)))
            [,1]
 [1,] 0.006824008
 [2,] 0.093699263
 [3,] 0.045229533
 [4,] 0.003743555
 [5,] 0.002068276
 [6,] 0.112316883
 [7,] 0.293421750
 [8,] 0.001860639
 [9,] 0.041946276
```

(f)
```
> c(1, 90)%*%b
         [,1]
[1,] 84.31173
```

End of Assignment — Total Available Marks = 33