



Semester 1 Assessment, 2021

School of Mathematics and Statistics

MAST30025 Linear Statistical Models Assignment 1

Submission deadline: **Friday March 26, 5pm**

This assignment consists of 10 pages (including this page)

Instructions to Students

- If you have a printer, print the assignment template.

Writing

- There are 5 questions with marks as shown. The total number of marks available is 40.
- 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 the page.

Scanning

- 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. Check PDF is readable.

Submitting

- Go to the Gradescope window. Choose the Canvas assignment for this assignment. Submit your file as a single PDF document only. Get Gradescope confirmation on email.
- It is your responsibility to ensure that your assignments are submitted correctly and on time, and problems with online submissions are not a valid excuse for submitting a late or incorrect version of an assignment.

Question 1 (7 marks)

Which of the following statements are true, and which are false? For the ones which are true, give a proof. For the ones which are false, give a counterexample.

- (a) For a symmetric matrix A , if $A^2 = A^3$, A is idempotent.
 (b) For a symmetric matrix A , if $A = A^3$, A is idempotent.

(a) True. Let λ be an eigenvalue of A with eigenvector \mathbf{x} . Then

$$\begin{aligned}\lambda^2 \mathbf{x} &= A^2 \mathbf{x} \\ &= A^3 \mathbf{x} \\ &= \lambda^3 \mathbf{x}.\end{aligned}$$

Hence $\lambda^2(1 - \lambda) = 0$, so $\lambda = 0$ or 1 . Therefore all eigenvalues of A are 0 or 1 . Now diagonalise A and write it as

$$A = PDP^T.$$

Then

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

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

- (b) False. Let $A = [-1]$. We have $A^3 = [-1] = A$, but $A^2 = [1] \neq A$, so A is not idempotent.

Question 2 (4 marks)

Let A_1, A_2, \dots, A_m be a set of symmetric $k \times k$ matrices. Suppose that there exists an orthogonal matrix P such that $P^T A_i P$ is diagonal for all i . Show that $A_i A_j = A_j A_i$ for every pair $i, j = 1, 2, \dots, m$.

$$\begin{aligned} P^T A_i A_j P &= (P^T A_i P) (P^T A_j P) \\ &= (P^T A_j P) (P^T A_i P) \\ &= P^T A_j A_i P. \end{aligned}$$

Premultiply by P and postmultiply by P^T to get $A_i A_j = A_j A_i$.

Question 3 (4 marks)

Let A be an $n \times p$ matrix with rank p . Show that $A^T A$ is a positive definite matrix.

First, for any vector \mathbf{x} , $\mathbf{x}^T A^T A \mathbf{x} = \|A\mathbf{x}\|^2 \geq 0$, so $A^T A$ is positive semi-definite. Now $r(A^T A) = r(A) = p$, and $A^T A$ is a $p \times p$ matrix. Thus it is invertible, and therefore has no 0 eigenvalues. So it must be positive definite.

Question 4 (10 marks)

Let $x_1, x_2, x_3 \sim N(\mu, \sigma^2)$ be a sequence of independent normal random variables, $\bar{x} = \frac{x_1+x_2+x_3}{3}$, $\mathbf{x} = (x_1, x_2, x_3)$, and $\mathbf{y} = (x_1 - \bar{x}, x_2 - \bar{x}, x_3 - \bar{x})^T$.

- (a) Let $\mathbf{y} = A\mathbf{x}$. Find A .
- (b) Find the rank of A .
- (c) Find $E(\mathbf{y}^T \mathbf{y})$.
- (d) Using Theorem 3.5, find the distribution of $\mathbf{y}^T \mathbf{y} / \sigma^2$.

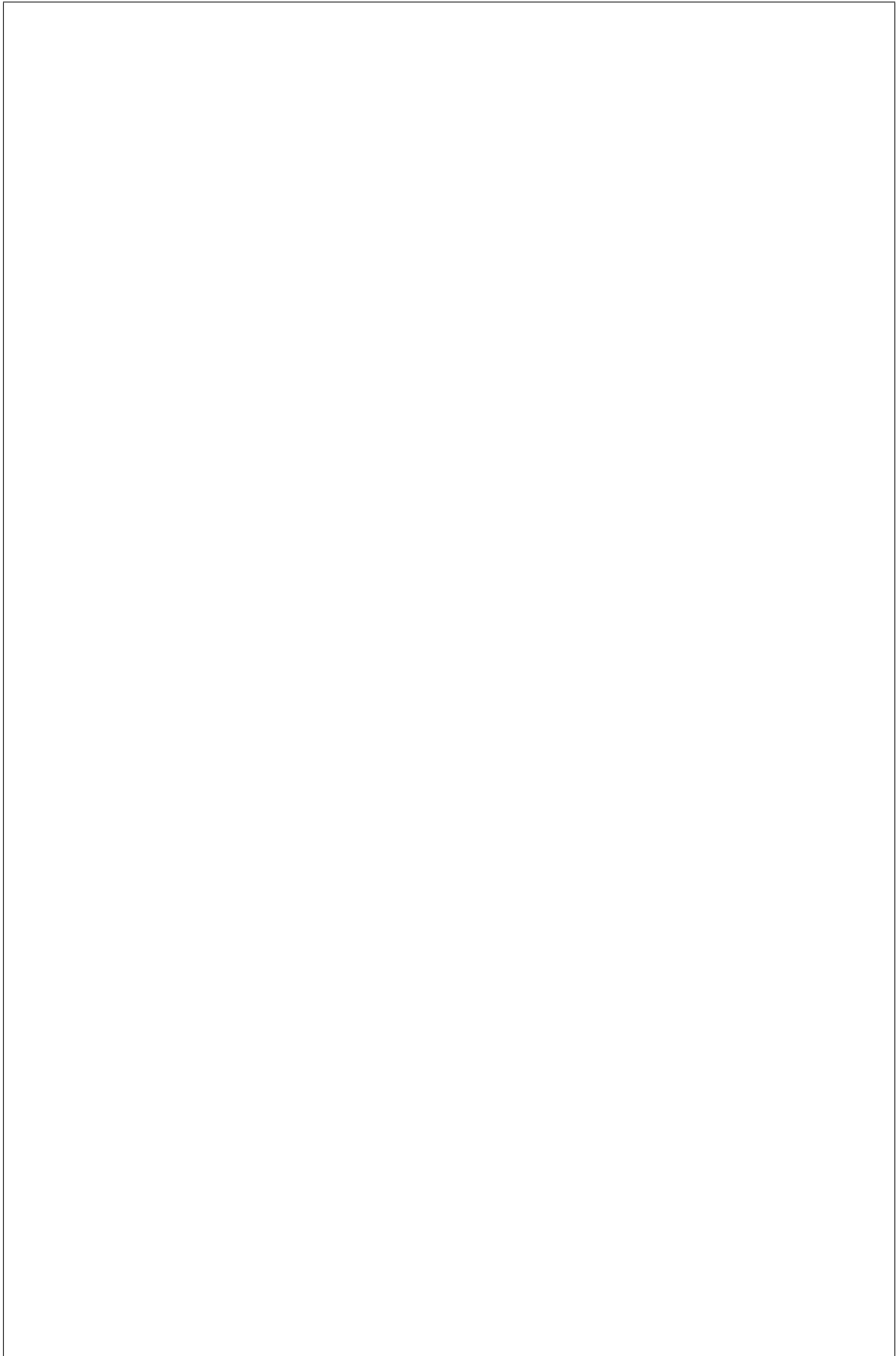
$$(a) \quad A = \begin{bmatrix} 2/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & -1/3 \\ -1/3 & -1/3 & 2/3 \end{bmatrix}.$$

(b) It is clear that $r(A) \geq 2$. However, $\det(A) = 0$, so $r(A) \leq 2$. Therefore, $r(A) = 2$.

(c) Note that $\mathbf{x} \sim MVN(\mu \mathbf{1}, \sigma^2 I)$ and $A^T A = A$. From Theorem 3.2,

$$E(\mathbf{y}^T \mathbf{y}) = E(\mathbf{x}^T A \mathbf{x}) = \sigma^2 \text{tr}(A) + \mu \mathbf{1}^T A \mathbf{1} \mu = 2\sigma^2.$$

(d) Let $\mathbf{z} = \mathbf{x} / \sigma \sim MVN(\boldsymbol{\mu}, I)$. Then we have $\mathbf{y}^T \mathbf{y} / \sigma^2 = \mathbf{z}^T A \mathbf{z}$. Since A is idempotent with rank 2, from Theorem 3.5, $\mathbf{y}^T \mathbf{y} / \sigma^2$ has a noncentral χ^2 distribution with $df = 2$ and noncentrality parameter $\lambda = 0$. That is, $\mathbf{y}^T \mathbf{y} / \sigma^2$ follows a χ^2 distribution with $df = 2$.



Question 5 (15 marks)

The table below shows prices in US cents per pound received by fishermen and vessel owners for various species of fish and shellfish in 1970 and 1980. (Taken from Moore & McCabe, Introduction to the Practice of Statistics, 1989.)

Type of fish	Price (1970)	Price (1980)
Cod	13.1	27.3
Flounder	15.3	42.4
Haddock	25.8	38.7
Menhaden	1.8	4.5
Ocean perch	4.9	23.0
Salmon, chinook	55.4	166.3
Salmon, coho	39.3	109.7
Tuna, albacore	26.7	80.1
Clams, soft-shelled	47.5	150.7
Clams, blue hard-shelled	6.6	20.3
Lobsters, american	94.7	189.7
Oysters, eastern	61.1	131.3
Sea scallops	135.6	404.2
Shrimp	47.6	149.0

We will model the 1980 price of fish, based on the 1970 price.

- The linear model is of the form $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Write down the matrices and vectors involved in this equation.
- Find the least squares estimators of the parameters.
- Calculate the sample variance s^2 .
- A fisherman sold ocean trout for 28c/pound in 1970. Predict the price for ocean trout in 1980.
- Calculate the standardised residual for sea scallops.
- Calculate the Cook's distance for sea scallops.
- Does sea scallops fit the linear model? Justify your argument.

(a)

$$y = \begin{bmatrix} 27.3 \\ 42.4 \\ 38.7 \\ 4.5 \\ 23.0 \\ 166.3 \\ 109.7 \\ 80.1 \\ 150.7 \\ 20.3 \\ 189.7 \\ 131.3 \\ 404.2 \\ 149.0 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 13.1 \\ 1 & 15.3 \\ 1 & 25.8 \\ 1 & 1.8 \\ 1 & 4.9 \\ 1 & 55.4 \\ 1 & 39.3 \\ 1 & 26.7 \\ 1 & 47.5 \\ 1 & 6.6 \\ 1 & 94.7 \\ 1 & 61.1 \\ 1 & 135.6 \\ 1 & 47.6 \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \\ \epsilon_{10} \\ \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \end{bmatrix}.$$

```
(b) > n <- 14
> p <- 2
> y <- c(27.3,42.4,38.7,4.5,23.0,166.3,109.7,80.1,150.7,
+        20.3,189.7,131.3,404.2,149.0)
> X <- matrix(c(rep(1,n),13.1,15.3,25.8,1.8,4.9,55.4,
+               39.3,26.7,47.5,6.6,94.7,61.1,135.6,47.6),n,p)
> ( b <- solve(t(X)%*%X, t(X)%*%y) )
```

```
      [,1]
[1,] -1.233836
[2,]  2.701553
```

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

[1] 777.1528
```



```
(d) > c(1,28)%*%b
```

```
      [,1]  
[1,] 74.40965
```

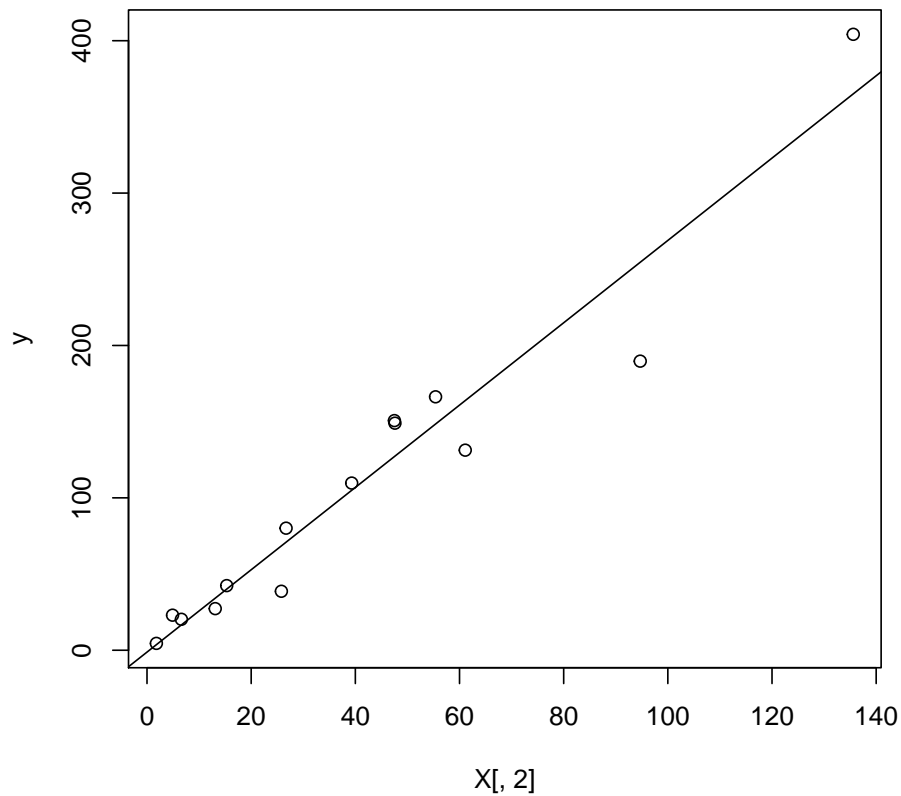
```
(e) > H <- X %*% solve(t(X)%*%X) %*% t(X)  
> z <- e / sqrt(s2 * (1 - diag(H)))  
> z[13]
```

```
[1] 2.104999
```

```
(f) > D <- 1/p * z^2 * diag(H) / (1-diag(H))  
> D[13]
```

```
[1] 2.774008
```

```
(g) > plot(X[,2],y)  
> abline(b[1],b[2])
```



The Cook's distance certainly indicates it should be of some concern; however looking at the plot, it seems that the fit is actually okay. There is considerable evidence for heteroskedasticity — the variance increases with x (the design variable). Sea scallops has (by far) the largest x and so may be prone to a larger variance than the remaining points. The high Cook's distance therefore comes primarily from a very high leverage, rather than a bad fit to the model.

End of Assignment — Total Available Marks = 40