THE UNIVERSITY OF MELBOURNE

Student number

Semester 1 Assessment, 2021

School of Mathematics and Statistics

MAST30025 Linear Statistical Models Assignment 2

Submission deadline: **Friday April 30, 5pm**

This assignment consists of 13 pages (including this page)

Instructions to Students

Writing

- There are 5 questions with marks as shown. The total number of marks available is 40.
- This assignment is worth 7% 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, including the `lm` function unless specified. 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 (4 marks)

Prove Theorem 4.8: show that the maximum likelihood estimator of the error variance σ^2 is

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n}.$$

The log-likelihood is given in the lecture notes as

$$\begin{aligned}
\ln L(\boldsymbol{\beta}, \sigma^2) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta}) \\
\frac{\partial}{\partial\sigma^2}\ln L(\boldsymbol{\beta}, \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta}) = 0 \\
\sigma^2 &= \frac{1}{n}(\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta})
\end{aligned}$$

which gives the required formula on the substitution of the ML estimator $\mathbf{b}$ for $\boldsymbol{\beta}$.

Question 2 (11 marks)

We wish to predict the price of apartments in Melbourne using some of their features. Let y be the apartment price per square metre, x_1 be the apartment age (in years), x_2 be the distance (in metres) to the nearest train station, and x_3 be the number of convenience stores nearby. The following data is collected:

x_1	x_2	x_3	y ($\times 10^3$)
32	84.9	10	37.9
19.5	306.6	9	42.2
13.3	562.0	5	47.3
13.3	562.0	5	43.1
5	390.6	5	54.8
7.1	2175.0	3	47.1
34.5	623.5	7	40.3

For this question, you may NOT use the `lm` function in R.

(a) Fit a linear model to the data and estimate the parameters and variance.

(b) Find a 90% confidence interval for the expected price per square metre of a 10 year old apartment that is 100 meters away from the train station and has 6 convenience stores nearby.

(c) Find the standard error of $\beta_1 - \beta_3$.

(d) Test the hypothesis that the price per square metre falls by \$1000 for every year that the apartment ages, at the 5% significance level.

(e) Test for model relevance using a corrected sum of squares.

(a)

```
> n <- 7
> p <- 4
> X <- matrix(c(rep(1,n),32, 19.5, 13.3, 13.3, 5, 7.1, 34.5,
+         84.9, 306.6, 562.0, 562.0, 390.6, 2175.0, 623.5,
+         10, 9, 5, 5, 5, 3, 7),n,p)
> y <- c(37.9, 42.2, 47.3, 43.1, 54.8, 47.1, 40.3)
> (b <- solve(t(X)%*%X,t(X)%*%y))

             [,1]
[1,] 58.369312708
[2,] -0.346291960
[3,] -0.002900359
[4,] -0.887671692

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

[1] 13.06871
```

(b)

```
> xst <- as.vector(c(1,10,100,6))
> xst %*% b + c(-1,1) * qt(0.95,df=n-p)*
+         sqrt(s2 * t(xst) %*% solve(t(X)%*%X) %*% xst)

[1] 43.27252 55.30814
```

A 90% confidence interval is $(43.27, 55.31)$.

(c)

```
> tt <- c(0,1,0,-1)
> sqrt(s2 * t(tt) %*% solve(t(X)%*%X) %*% tt)

         [,1]
[1,] 1.388968
```

The standard error of $\beta_1 - \beta_3$ is 1.389.

(d) This is a general linear hypothesis with $C = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix}$ and $\boldsymbol{\delta}^* = -1$.

```
> C <- matrix(c(0,1,0,0),1,4)
> dst <- -1
> Fstat <- (t(C%*%b-dst)%*%solve(C%*%solve(t(X)%*%X)%*%t(C))%*%
+         (C%*%b-dst)/1)/s2
> pf(Fstat, 1, n-p, lower=F)

           [,1]
[1,] 0.04945829
```

We reject the null hypothesis at the 5% significance level.

(e)

```
> SSReg <- t(y) %*% X %*% b - sum(y)^2 / n
> SSRes <- s2*(n-p)
> ( Fstat <- (SSReg/(p-1))/(SSRes/(n-p)) )

      [,1]
[1,] 3.819

> pf(Fstat, p-1, n-p, lower.tail = FALSE)

          [,1]
[1,] 0.1500833
```

We do not reject the null hypothesis of model irrelevance.

Question 3 (5 marks)

Consider two full rank linear models $\mathbf{y} = X_1\boldsymbol{\gamma}_1 + \boldsymbol{\varepsilon}_1$ and $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}_2$, where all predictors in the first model ($\boldsymbol{\gamma}_1$) are also contained in the second model ($\boldsymbol{\beta}$). Show that the SS_{Res} for the first model is at least the SS_{Res} for the second model.

Let $\hat{\boldsymbol{\gamma}}_1$ be the least squares estimates for $\boldsymbol{\gamma}_1$ in the first model. Then $\begin{bmatrix} \hat{\boldsymbol{\gamma}}_1 \\ \mathbf{0} \end{bmatrix}$ is a (not necessarily optimal) estimate for $\boldsymbol{\beta}$ in the second model, with residual sum of squares

$$\left(\mathbf{y} - X\begin{bmatrix} \hat{\boldsymbol{\gamma}}_1 \\ \mathbf{0} \end{bmatrix}\right)^T \left(\mathbf{y} - X\begin{bmatrix} \hat{\boldsymbol{\gamma}}_1 \\ \mathbf{0} \end{bmatrix}\right) = (\mathbf{y} - X_1\hat{\boldsymbol{\gamma}}_1)^T(\mathbf{y} - X_1\hat{\boldsymbol{\gamma}}_1)$$
$$= SS_{Res} \text{ (first model)}.$$

But the least squares estimates $\mathbf{b}$ of $\boldsymbol{\beta}$ minimise the residual sum of squares for the second model, so we get

$$SS_{Res} \text{ (second model)} \leq SS_{Res} \text{ (first model)}.$$

Question 4 (10 marks)

In this question, we study the mtcars dataset. This dataset contains data published by the US magazine *Motor Trends* in 1974, on fuel consumption of cars for 32 different models. It includes the variables:

- mpg: miles/(US) gallon
- disp: displacement (cu. in.)
- hp: gross horsepower
- drat: rear axle ratio
- wt: weight (1000 lbs)
- qsec: 1/4 mile time

The dataset is distributed with R. Open it, select the appropriate variables, and take a logarithmic transformation of the data with the following commands:

```
> data(mtcars)
> mtcars = log(mtcars[, c(1,3:7)])
```

We wish to use a linear model to model mpg in terms of the other variables.

(a) Plot the data and comment.

(b) Perform model selection using forward selection.

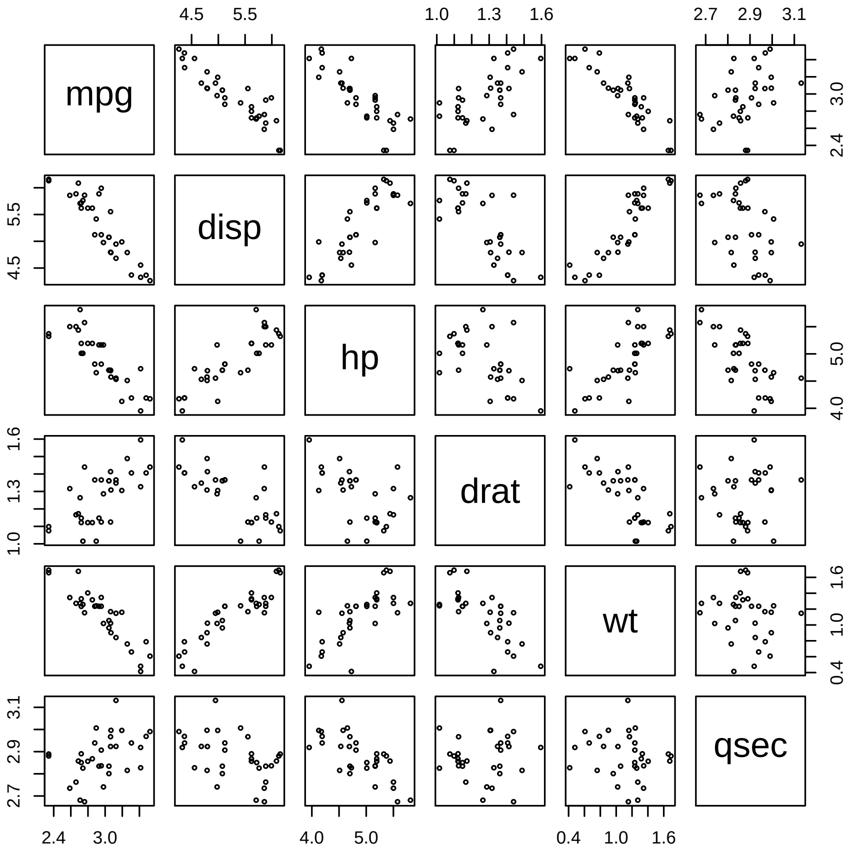
(c) Starting from the full model, perform model selection using stepwise selection with AIC.

(d) Write down the final fitted model from stepwise selection.

(e) Produce diagnostic plots for your final model from stepwise selection and comment.

(a) Looking at `mpg` against the other variables, a linear assumption seems to be reasonable.

```
> pairs(mtcars,cex=0.5)
```



(b)

```
> model0 <- lm(mpg ~ 1, data=mtcars)
> add1(model0, scope= ~ . + disp + hp + drat + wt + qsec, test="F")

Single term additions

Model:
mpg ~ 1
       Df Sum of Sq     RSS      AIC  F value    Pr(>F)
<none>              2.74874  -76.547
disp    1   2.25596 0.49277 -129.550 137.3427 1.006e-12 ***
hp      1   1.96733 0.78140 -114.797  75.5310 1.080e-09 ***
drat    1   1.23131 1.51742  -93.559  24.3435 2.807e-05 ***
wt      1   2.21452 0.53422 -126.966 124.3596 3.406e-12 ***
qsec    1   0.47755 2.27119  -80.654   6.3079   0.01763 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> model1 <- lm(mpg ~ disp, data=mtcars)
> add1(model1, scope= ~ . + hp + drat + wt + qsec, test="F")

Single term additions

Model:
mpg ~ disp
       Df Sum of Sq     RSS     AIC F value  Pr(>F)
<none>              0.49277 -129.55
hp      1  0.045531 0.44724 -130.65  2.9523 0.09641 .
drat    1  0.001383 0.49139 -127.64  0.0816 0.77711
wt      1  0.098796 0.39398 -134.71  7.2722 0.01154 *
qsec    1  0.000308 0.49247 -127.57  0.0181 0.89382
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> model2 <- lm(mpg ~ disp + wt, data=mtcars)
> add1(model2, scope= ~ . + hp + drat + qsec, test="F")

Single term additions

Model:
mpg ~ disp + wt
       Df Sum of Sq     RSS     AIC F value  Pr(>F)
<none>              0.39398 -134.71
hp      1  0.078605 0.31537 -139.83  6.9789 0.01334 *
drat    1  0.007358 0.38662 -133.31  0.5329 0.47146
qsec    1  0.057788 0.33619 -137.79  4.8130 0.03671 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> model3 <- lm(mpg ~ disp + wt + hp, data=mtcars)
> add1(model3, scope= ~ .  + drat + qsec, test="F")

Single term additions

Model:
mpg ~ disp + wt + hp
       Df Sum of Sq     RSS     AIC F value Pr(>F)
<none>              0.31537 -139.83
drat    1 0.0000095 0.31536 -137.83  0.0008 0.9774
qsec    1 0.0033067 0.31206 -138.17  0.2861 0.5971
```

The final variables are `disp`, `wt` and `hp`.

(c)
```
> fullmodel <- lm(mpg ~ ., data = mtcars)
> model <- step(fullmodel, scope = ~ .)

Start:  AIC=-136.21
mpg ~ disp + hp + drat + wt + qsec

       Df Sum of Sq     RSS     AIC
- drat  1  0.000402 0.31207 -138.17
- disp  1  0.002104 0.31377 -138.00
- qsec  1  0.003699 0.31536 -137.83
<none>              0.31166 -136.21
- hp    1  0.023697 0.33536 -135.87
- wt    1  0.103076 0.41474 -129.07

Step:  AIC=-138.17
mpg ~ disp + hp + wt + qsec

       Df Sum of Sq     RSS     AIC
- qsec  1  0.003307 0.31537 -139.83
- disp  1  0.004372 0.31644 -139.72
<none>              0.31207 -138.17
- hp    1  0.024123 0.33619 -137.79
+ drat  1  0.000402 0.31166 -136.21
- wt    1  0.103779 0.41584 -130.98

Step:  AIC=-139.83
mpg ~ disp + hp + wt

       Df Sum of Sq     RSS     AIC
- disp  1  0.006635 0.32201 -141.16
<none>              0.31537 -139.83
+ qsec  1  0.003307 0.31207 -138.17
+ drat  1  0.000010 0.31536 -137.83
- hp    1  0.078605 0.39398 -134.71
- wt    1  0.131870 0.44724 -130.65

Step:  AIC=-141.17
mpg ~ hp + wt

       Df Sum of Sq     RSS     AIC
<none>              0.32201 -141.16
+ disp  1   0.00664 0.31537 -139.83
+ qsec  1   0.00557 0.31644 -139.72
+ drat  1   0.00112 0.32089 -139.28
- hp    1   0.21221 0.53422 -126.97
- wt    1   0.45939 0.78140 -114.80
```

The final variables are `hp` and `wt`.

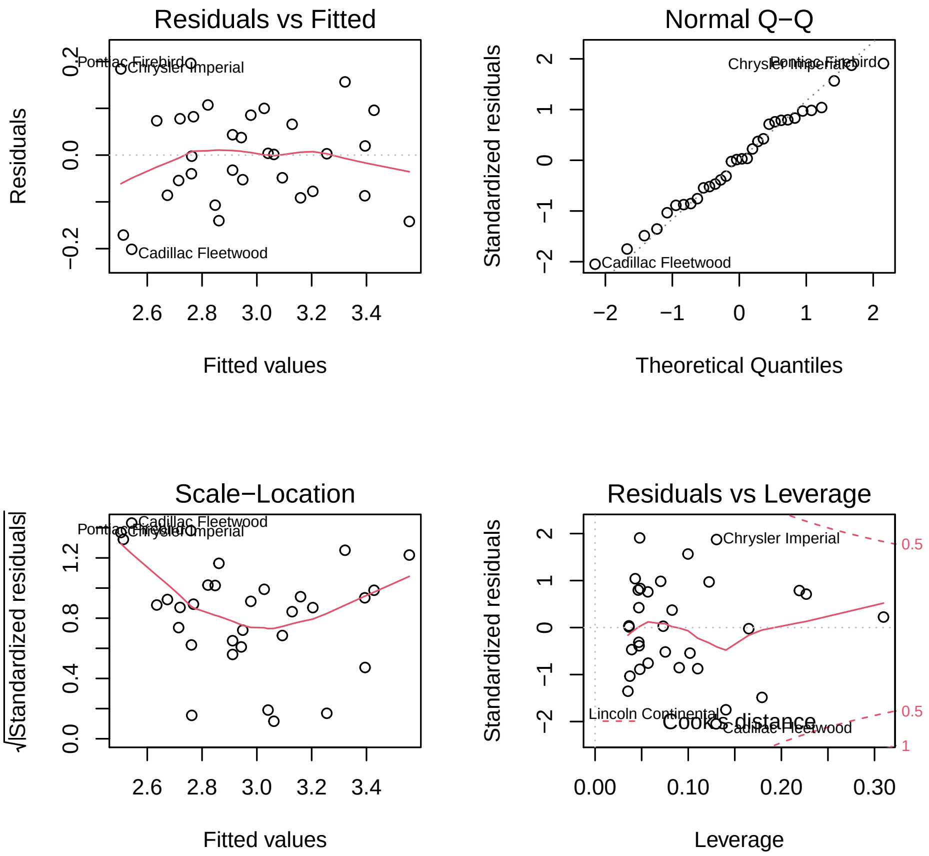
(d)
```
> model$coef

(Intercept)          hp          wt
  4.8346929  -0.2553185  -0.5622822
```

The final model (accounting for the transformation) is

$$\texttt{mpg} = 125.8 \cdot \texttt{hp}^{-0.255} \cdot \texttt{wt}^{-0.562}.$$

(e)

```
> opar <- par(mfrow=c(2,2))
> plot(model, which=1)
> plot(model, which=2)
> plot(model, which=3)
> plot(model, which=5)
> par <- opar
```



Diagnostic plots show a reasonable fit to linear model assumptions. There appears to be some pattern in the residuals when the fitted values are low or high, but on closer inspection it appears to be driven more by a small number of (not too) extreme points than a general trend.

Question 5 (10 marks)

For ridge regression, we choose parameter estimators $\mathbf{b}$ which minimise

$$\sum_{i=1}^{n} e_i^2 + \lambda \sum_{j=0}^{k} b_j^2,$$

where λ is a constant penalty parameter.

(a) Show that these estimators are given by

$$\mathbf{b} = (X^T X + \lambda I)^{-1} X^T \mathbf{y}.$$

(b) Show that $\mathbf{b}$ is biased if $\lambda \neq 0$.

(c) One way to calculate the optimal value for the penalty parameter is to minimise the AIC. Since the number of parameters p does not change, we use a slightly modified version:

$$AIC = n \ln \frac{SS_{Res}}{n} + 2\,df,$$

where df is the "effective degrees of freedom" defined by

$$df = tr(H) = tr(X(X^T X + \lambda I)^{-1} X^T).$$

We will use the data from Q2. In order to avoid penalising some parameters unfairly, we must first standardise the variables; this also means an intercept parameter is not used. You can do this with `scale`:

```
> X <- scale(X[,-1],center=T,scale=T)
> y <- scale(y,center=T,scale=T)
> p <- 3
```

Construct a plot of λ against AIC. Thereby find the optimal value for λ.

(a) We have

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{b}}\left[\sum_{i=1}^{n} e_i^2 + \lambda \sum_{j=0}^{k} b_j^2\right] &= \frac{\partial}{\partial \mathbf{b}}\left[(\mathbf{y} - X\mathbf{b})^T(\mathbf{y} - X\mathbf{b}) + \lambda \mathbf{b}^T\mathbf{b}\right] \\
&= \frac{\partial}{\partial \mathbf{b}}\left[\mathbf{y}^T\mathbf{y} - \mathbf{y}^T X\mathbf{b} + \mathbf{b}^T X^T X\mathbf{b} + \lambda \mathbf{b}^T\mathbf{b}\right] \\
&= -2X^T\mathbf{y} + 2(X^TX + \lambda I)\mathbf{b} = 0 \\
(X^TX + \lambda I)\mathbf{b} &= X^T\mathbf{y} \\
\mathbf{b} &= (X^TX + \lambda I)^{-1}X^T\mathbf{y}.
\end{aligned}$$

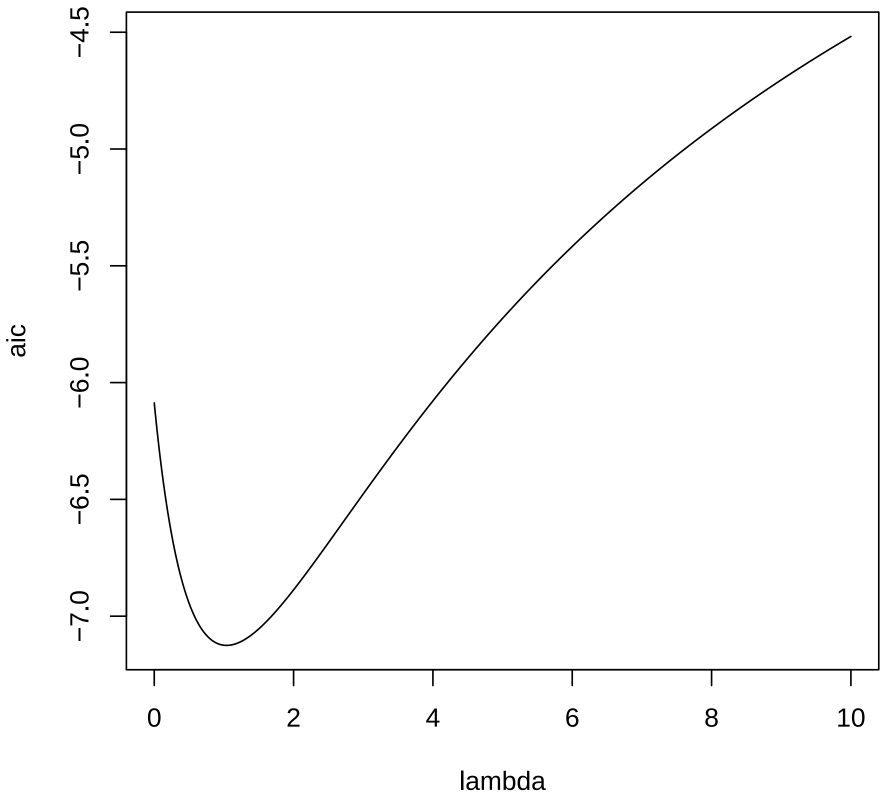
(b) We have

$$\begin{aligned}
E(\mathbf{b}) &= E((X^TX + \lambda I)^{-1}X^T\mathbf{y}) \\
&= (X^TX + \lambda I)^{-1}X^T E(\mathbf{y}) \\
&= (X^TX + \lambda I)^{-1}X^T X\boldsymbol{\beta} \\
&= (X^TX + \lambda I)^{-1}(X^TX + \lambda I - \lambda I)\boldsymbol{\beta} \\
&= \boldsymbol{\beta} - \lambda(X^TX + \lambda I)^{-1}\boldsymbol{\beta}.
\end{aligned}$$

Therefore, $\mathbf{b}$ is biased if $\lambda \neq 0$.

(c)

```
> lambda <- seq(0,10,0.01)
> aic <- c()
> for (l in lambda) {
+         b <- solve(t(X)%*%X + l*diag(p),t(X)%*%y)
+         ssres <- sum((y-X%*%b)^2)
+         H <- X %*% solve(t(X)%*%X + l*diag(p)) %*% t(X)
+         aic <- c(aic, n*log(ssres/n) + 2*sum(diag(H)))
+ }
> plot(lambda,aic,type='l')
> lambda[which.min(aic)]

[1] 1.04
```



End of Assignment — Total Available Marks = 40