Student
number

Semester 1 Assessment, 2021

School of Mathematics and Statistics

MAST30025 Linear Statistical Models Assignment 3

Submission deadline: **Friday May 28, 5pm**

This assignment consists of 12 pages (including this page)

Instructions to Students

Writing

- There are 5 questions with marks as shown. The total number of marks available is 48.
- 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 (7 marks)

Let A be an $n \times p$ matrix with $n \geq p$.

(a) Show directly that $r(A^cA) = r(A)$.

(b) Show directly that A^cA is idempotent.

(c) Show directly that $A(A^TA)^cA^T$ is unique (invariant to the choice of conditional inverse).

(a)

$$r(A) = r(AA^cA) \leq r(A^cA) \leq r(A).$$

(b)

$$(A^cA)(A^cA) = A^c(AA^cA) = A^cA.$$

(c)

$$\begin{aligned} A(A^TA)_1^cA^T &= \left[A(A^TA)_2^cA^TA\right](A^TA)_1^cA^T \\ &= A(A^TA)_2^c\left[A(A^TA)_1^cA^TA\right]^T \\ &= A(A^TA)_2^cA^T. \end{aligned}$$

Question 2 (11 marks)

We are interested in examining the lifetime of three different types of bulb (in 1000 hours). A study is conducted and the following data obtained:

Bulb type		
1	2	3
22	16	28
23	18	27
24	19	29
22		29
26		

We fit the model

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij},$$

where μ is the overall mean and τ_i is the effect of the ith type of bulb.

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

(a) Find a conditional inverse for $X^T X$, using the algorithm given in Theorem 6.2.

(b) Find s^2.

(c) Is $\mu + 2\tau_1 + \tau_2$ estimable?

(d) Find a 90% confidence interval for the lifetime of the 2nd type of bulb.

(e) Test the hypothesis that there is no difference in lifetime between 1st and 3rd types of bulb, at the 5% significance level.

(a)
```
> X <- matrix(c(rep(1,17),rep(0,12),rep(1,3),rep(0,12),
+                           rep(1,4)),12,4)
> y <- as.vector(c(22,23,24,22,26,16,18,19,28,27,29,29))
> XtXc <- matrix(0,4,4)
> XtXc[2:4,2:4] <- solve((t(X)%*%X)[2:4,2:4])
> XtXc

     [,1] [,2]      [,3] [,4]
[1,]    0  0.0 0.0000000 0.00
[2,]    0  0.2 0.0000000 0.00
[3,]    0  0.0 0.3333333 0.00
[4,]    0  0.0 0.0000000 0.25
```

(b)

```
> n <- 12
> r <- 3
> b <- XtXc %*% t(X) %*% y
> (s2 <- sum((y - X%*%b)^2)/(n-r))

[1] 2.068519
```

(c)

```
> tt <- as.vector(c(1,2,1,0))
> tt %*% XtXc %*% t(X) %*% X

     [,1] [,2] [,3] [,4]
[1,]    3    2    1    0
```

No, since $\mathbf{t}^T(X^TX)^cX^TX \neq \mathbf{t}^T$.

(d)

```
> tt <- as.vector(c(1,0,1,0))
> hw <- qt(0.95,df=n-r)*sqrt(s2)*sqrt(t(tt) %*% XtXc %*% tt)
> tt%*%b + c(-1,1)*hw

[1] 16.14451 19.18882
```

(e) This hypothesis can be written as $H_0 : C\boldsymbol{\beta} = \mathbf{0}$, where

$$C = \begin{bmatrix} 0 & 1 & 0 & -1 \end{bmatrix}.$$

```
> C <- matrix(c(0,1,0,-1),1,4)
> (Fstat <- t(C%*%b) %*% solve(C%*%XtXc%*%t(C)) %*% C%*%b / s2)

         [,1]
[1,] 25.27037

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

             [,1]
[1,] 0.0007123037
```

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

Question 3 (5 marks)

Consider a linear model with only categorical predictors, written in matrix form as $\mathbf{y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1$. Suppose we add some continuous predictors, resulting in an expanded model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Now consider a quantity $\mathbf{t}^T\boldsymbol{\beta}$, where $\mathbf{t}^T = [\mathbf{t}_1^T|\mathbf{t}_2^T]$ is partitioned according to the categorical and continuous predictors. Show that if $\mathbf{t}_1^T\boldsymbol{\beta}_1$ is estimable in the first model, then $\mathbf{t}^T\boldsymbol{\beta}$ is estimable in the second model.
If you write $X = [X_1|X_2]$, you may assume that $r(X) = r(X_1) + r(X_2)$.

Hint: Use Theorems 6.9 and 6.3. For any vectors $\mathbf{z}_1$ *and* $\mathbf{z}_2$*, you can write*

$$\left[\begin{array}{cc|c} X_1^TX_1 & X_1^TX_2 & X_1^TX_1\mathbf{z}_1 \\ X_2^TX_1 & X_2^TX_2 & X_2^TX_2\mathbf{z}_2 \end{array}\right] = \left[\begin{array}{cc} X_1^T & \mathbf{0} \\ \mathbf{0} & X_2^T \end{array}\right] \left[\begin{array}{cc|c} X_1 & X_2 & X_1\mathbf{z}_1 \\ X_1 & X_2 & X_2\mathbf{z}_2 \end{array}\right].$$

We have

$$X^TX = \left[\begin{array}{c|c} X_1^TX_1 & X_1^TX_2 \\ \hline X_2^TX_1 & X_2^TX_2 \end{array}\right].$$

We need to show that $X^TX\mathbf{z} = \mathbf{t}$ has a solution for $\mathbf{z}$. To do this we show that

$$r\left(\left[X^TX\middle|\mathbf{t}\right]\right) = r(X^TX) = r(X) = r(X_1) + r(X_2).$$

We first observe that since $\mathbf{t}_1^T\boldsymbol{\beta}_1$ is estimable in the first model, there exists a solution $\mathbf{z}_1$ to the system $X_1^TX_1\mathbf{z} = \mathbf{t}_1$. Likewise, there exists a solution $\mathbf{z}_2$ to the system $X_2^TX_2\mathbf{z} = \mathbf{t}_2$, since X_2 can be assumed to be of full rank.
Now, it is immediate that

$$r\left(\left[X^TX\middle|\mathbf{t}\right]\right) \geq r(X^TX).$$

To show the reverse inequality, we have

$$\begin{aligned} r\left(\left[X^TX\middle|\mathbf{t}\right]\right) &= r\left(\left[\begin{array}{cc|c} X_1^TX_1 & X_1^TX_2 & \mathbf{t}_1 \\ X_2^TX_1 & X_2^TX_2 & \mathbf{t}_2 \end{array}\right]\right) \\ &= r\left(\left[\begin{array}{cc|c} X_1^TX_1 & X_1^TX_2 & X_1^TX_1\mathbf{z}_1 \\ X_2^TX_1 & X_2^TX_2 & X_2^TX_2\mathbf{z}_2 \end{array}\right]\right) \\ &= r\left(\left[\begin{array}{cc} X_1^T & \mathbf{0} \\ \mathbf{0} & X_2^T \end{array}\right]\left[\begin{array}{cc|c} X_1 & X_2 & X_1\mathbf{z}_1 \\ X_1 & X_2 & X_2\mathbf{z}_2 \end{array}\right]\right) \\ &\leq r\left(\left[\begin{array}{cc} X_1^T & \mathbf{0} \\ \mathbf{0} & X_2^T \end{array}\right]\right) \\ &= r(X_1) + r(X_2). \end{aligned}$$

Thus the equality is proved and $\mathbf{t}^T\boldsymbol{\beta}$ is estimable in the full model.

Question 4 (18 marks)

Data was collected on the world record times (in seconds) for the one-mile run. For males, the records are from the period 1861–1999, and for females, from the period 1967–1996. The data is given in the file `mile.csv`, available on the LMS.

(a) Plot the data, using different colours and/or symbols for male and female records. Without drawing diagnostic plots, do you think that this data satisfies the assumptions of the linear model? Why or why not?

(b) Test the hypothesis that there is no interaction between the two predictor variables. Interpret the result in the context of the study.

(c) Write down the final fitted models for the male and female records. Add lines corresponding to these models to your plot from part (a).

(d) Calculate a point estimate for the year when the female world record will equal the male world record. Do you expect this estimate to be accurate? Why or why not?

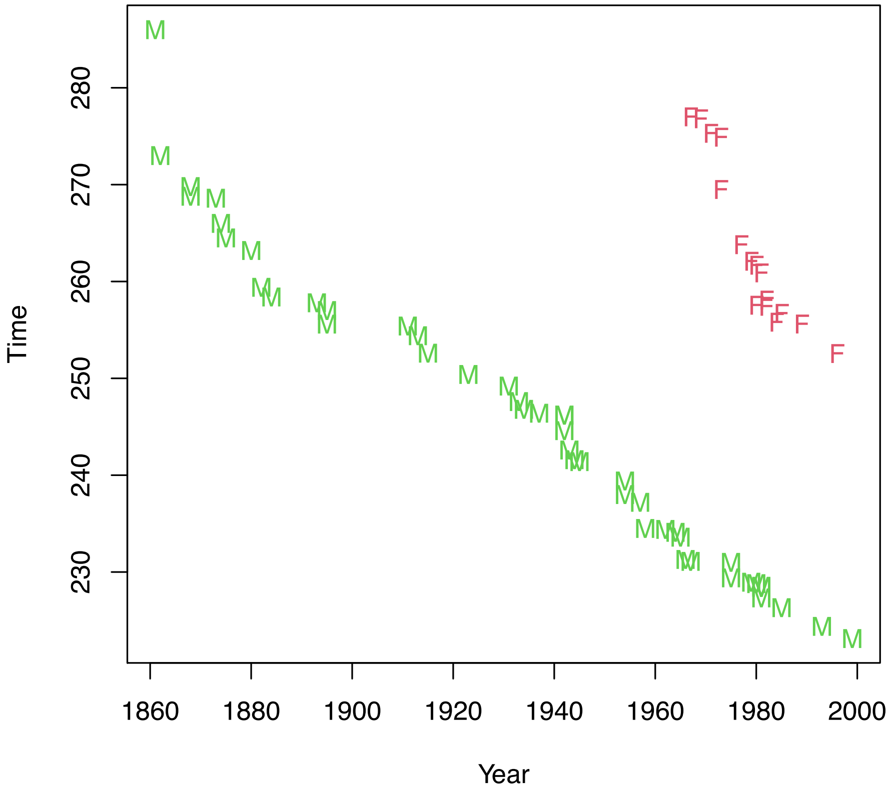
(e) Is the year when the female world record will equal the male world record an estimable quantity? Is your answer consistent with part (d)?

(f) Calculate a 95% confidence interval for the amount by which the gap between the male and female world records narrow every year.

(g) Test the hypothesis that the male world record decreases by 0.4 seconds each year.

(a)

```
> mile <- read.csv('../../data/mile.csv', header=T)
> mile$Gender <- factor(mile$Gender)
> plot(Time ~ Year, data = mile, col=as.numeric(Gender)+1,
+          pch=as.character(Gender))
```



While the data look quite linear, there is a big problem: the record can only decrease. So the data are not independent and cannot satisfy the linear model assumptions.

(b)

```
> imodel <- lm(Time ~ (Year + Gender)^2, data = mile)
> amodel <- lm(Time ~ Year + Gender, data = mile)
> anova(amodel, imodel)

Analysis of Variance Table

Model 1: Time ~ Year + Gender
Model 2: Time ~ (Year + Gender)^2
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     59 895.62
2     58 518.03  1    377.59 42.276 2.001e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

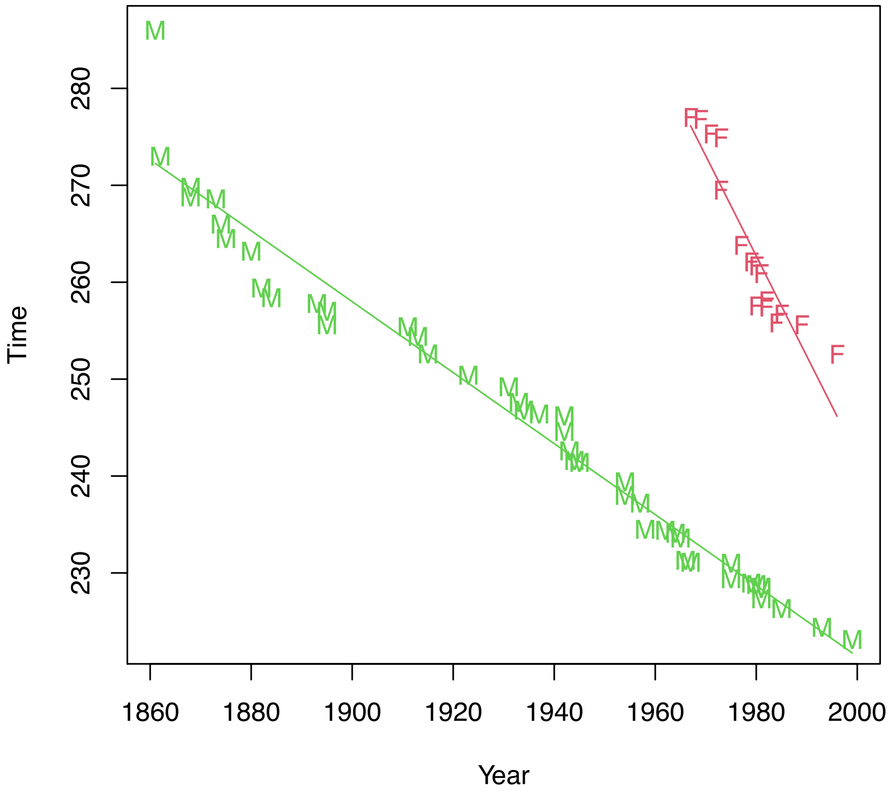
We conclude that there is significant interaction, i.e., there is a significant difference between the rate of improvements in the male and female records. We should use the model with interaction.

(c)

```
> imodel$coef[c(1,2)] + imodel$coef[c(3,4)]

(Intercept)        Year
953.7469611  -0.3661867

> plot(Time ~ Year, data = mile, col=as.numeric(Gender)+1
+          , pch=as.character(Gender))
> for (i in 1:2) { with(mile, lines(Year[as.numeric(Gender)==i],
+              fitted(imodel)[as.numeric(Gender)==i], col=i+1)) }
```



For females,

$$\text{Time} = 2309 - 1.034\,\text{Year}.$$

For males,

$$\text{Time} = 954 - 0.366\,\text{Year}.$$

(d)
```
> -imodel$coef[3]/imodel$coef[4]

GenderMale
   2030.95
```

We expect that the world records will be equal around the year 2031. However this is unlikely to be an accurate estimate, as we are extrapolating well beyond the range of the data.

(e) Strangely, this is a not an estimable quantity; it is not even expressible as a linear function of the parameters. This is consistent with part (d), because "estimable" really means *linearly* estimable. So we can estimate a value for this quantity even though it is not estimable.

(f)
```
> confint(imodel)[4,]

    2.5 %    97.5 %
0.4620087 0.8730100
```

(g)
```
> linearHypothesis(imodel, c(0,1,0,1), -0.4)

Linear hypothesis test

Hypothesis:
Year  + Year:GenderMale = - 0.4

Model 1: restricted model
Model 2: Time ~ (Year + Gender)^2

  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     59 604.84
2     58 518.03  1    86.806 9.7191 0.002837 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We reject this hypothesis; the record decreases slower than this rate.

Question 5 (7 marks)

You wish to perform a study to compare 2 medical treatments (and a placebo) for a disease. Treatment 1 is an experimental new treatment, and costs $5000 per person. Treatment 2 is a standard treatment, and costs $2000 per person. Treatment 3 is a placebo, and costs $1000 per person. You are given $100,000 to complete the study. You wish to test if the treatments are effective, i.e., $H_0 : \tau_1 = \tau_2 = \tau_3$.

(a) Determine the optimal allocation of the number of units to assign to each treatment.

(b) Perform the random allocation. You must use R for randomisation and include your R commands and output.

(a) The best contrasts to use here are contrasts against treatment 3, as it is the cheapest. As in the lecture notes, we have

$$\text{var}\ \widehat{\tau_i - \tau_3} = \sigma^2 \left(\frac{1}{n_i} + \frac{1}{n_3} \right),$$

and so we want to minimise

$$f(n_1, n_2, n_3, \lambda) = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{2}{n_3} \right) + \lambda \left(5n_1 + 2n_2 + n_3 - 100\right).$$

This gives

$$\begin{aligned}
\frac{\partial f}{\partial n_1} &= -\frac{\sigma^2}{n_1^2} + 5\lambda = 0 \\
\frac{\partial f}{\partial n_2} &= -\frac{\sigma^2}{n_2^2} + 2\lambda = 0 \\
\frac{\partial f}{\partial n_3} &= -2\frac{\sigma^2}{n_3^2} + \lambda = 0 \\
n_1^2 &= \frac{\sigma^2}{5\lambda} = \frac{2}{5} n_2^2 = \frac{1}{10} n_3^2 \\
n_1 &= \frac{1}{\sqrt{10}} n_3, \qquad n_2 = \frac{1}{2} n_3 \\
\frac{5}{\sqrt{10}} n_3 + \frac{2}{2} n_3 + n_3 &= 100
\end{aligned}$$

This gives $n_3 = 27.92$ (rounded to 27 to fit budget constraints), $n_1 = 9, n_2 = 14$.

(b)

```
> x <- sample(50, 50)
> x[1:9]

[1] 37 42 32  7  2 39 31 11 33

> x[10:23]

 [1]  1 28 38  6 19 48 18 45 35 14 49 41 12 30

> x[24:50]

 [1] 20 23 34 44 43  8 25 16  4 22  9 27 40 15 46  3 21 47 26 10 13
[26] 29 24
```

End of Assignment — Total Available Marks = 48