MAST30025: Linear Statistical Models

Assignment 3, 2019 Solutions

Total marks: 50

Due: 5pm Friday, May 31 (week 12)

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

   (a) Show that $r(A^c A) = r(A)$.
   Solution [2 marks]:
   $$r(A) = r(AA^c A) \leq r(A^c A) \leq r(A).$$

   (b) Show that $I - A(A^T A)^c A^T$ is idempotent.
   Solution [2 marks]: This follows from the idempotency of $A(A^T A)^c A^T$ (proved in lectures) and the fact that if H is idempotent, then $I - H$ is also idempotent.

   (c) Show that $r(I - A(A^T A)^c A^T) = n - r(A)$.
   Solution [3 marks]:
   $$\begin{aligned} r(I - A(A^T A)^c A^T) &= tr(I - A(A^T A)^c A^T) \\ &= tr(I_n) - tr(A(A^T A)^c A^T) \\ &= n - r(A(A^T A)^c A^T) \\ &= n - r(A). \end{aligned}$$

2. We are interested in examining the yield of tomato plants that have been grown with certain types of fertiliser. A study is conducted and the following data obtained:

Fertiliser		
1	2	3
43	33	56
45	37	54
47	38	57
46	35	
48		

   We fit the model
   $$y_{ij} = \mu + \tau_i + \varepsilon_{ij},$$
   where μ is the overall mean and τ_i is the effect of using the ith fertiliser.

   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.
   Solution [2 marks]:

```
> X <- matrix(c(rep(1,17),rep(0,12),rep(1,4),rep(0,12),rep(1,3)),12,4)
> y <- as.vector(c(43,45,47,46,48,33,37,38,35,56,54,57))
> t(X) %*% X
     [,1] [,2] [,3] [,4]
[1,]   12    5    4    3
[2,]    5    5    0    0
[3,]    4    0    4    0
[4,]    3    0    0    3
> 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 0.00 0.0000000
[2,]    0  0.2 0.00 0.0000000
[3,]    0  0.0 0.25 0.0000000
[4,]    0  0.0 0.00 0.3333333
```

(b) Characterise all solutions to the normal equations.

```
> (b <- XtXc %*% t(X) %*% y)
```

```
         [,1]
[1,]  0.00000
[2,] 45.80000
[3,] 35.75000
[4,] 55.66667
```

```
> diag(rep(1,4)) - XtXc%*%t(X)%*%X
```

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

Solution [3 marks]: All solutions to the normal equations are of the form $\mathbf{b} + \begin{bmatrix} 1 \\ -1 \\ -1 \\ -1 \end{bmatrix} z$ for arbitrary z.

(c) Is $4\mu + 2\tau_1 + \tau_2 + \tau_3$ estimable?

Solution [2 marks]: Yes, as it is a linear combination of $\mu + \tau_1$, $\mu + \tau_2$ and $\mu + \tau_3$, which are all elements of $X\boldsymbol{\beta}$ and therefore estimable.

(d) Find a 95% prediction interval for the yield of a tomato plant grown on fertiliser 1.

Solution [3 marks]:

```
> n <- 12
> r <- 3
> tt <- as.vector(c(1,1,0,0))
> s2 <- sum((y - X%*%b)^2)/(n-r)
> hw <- qt(0.975,df=n-r)*sqrt(s2)*sqrt(1 + t(tt) %*% XtXc %*% tt)
> tt%*%b + c(-1,1)*hw
```

```
[1] 40.96818 50.63182
```

(e) Test the hypothesis that fertilisers 2 and 3 have no difference in yield.

Solution [2 marks]: This hypothesis can be written as $H_0 : C\boldsymbol{\beta} = \mathbf{0}$, where

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

We reject the null hypothesis firmly.

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

```
         [,1]
[1,] 178.8633
```

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

```
             [,1]
[1,] 3.042802e-07
```

3. Consider a linear model with only categorical predictors, written in matrix form as $\mathbf{y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1$. Now 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

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

   Solution [5 marks]:

   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

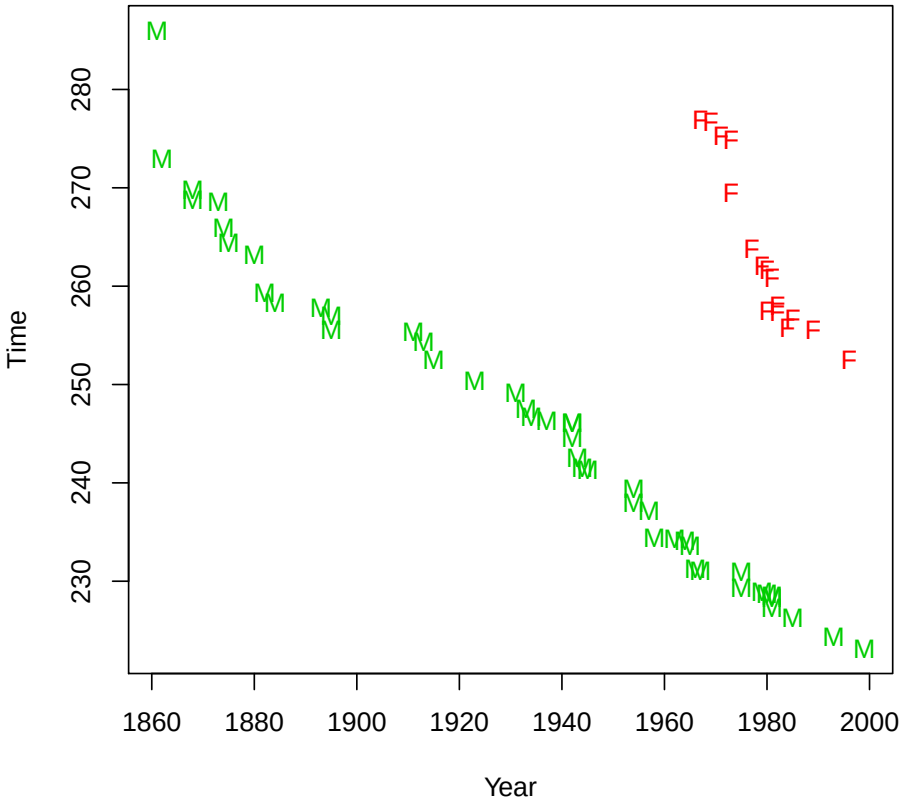
$$\begin{aligned} r\left(\left[X^TX\middle|\mathbf{t}\right]\right) &= r\left(\left[\begin{array}{cc|c} X_1^TX_1 & X_1^TX_2 & \mathbf{t}_1 \\ X_2^TX_1 & X_2^TX_2 & \mathbf{t}_2 \end{array}\right]\right) \\ &= r\left(\left[\begin{array}{cc|c} X_1^TX_1 & X_1^TX_2 & X_1^TX_1\mathbf{z}_1 \\ X_2^TX_1 & X_2^TX_2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

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

4. 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`.

   (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?

```
> 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))
```



Solution [3 marks]: 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) Test the hypothesis that there is no interaction between the two predictor variables. Interpret the result in the context of the study.

```
> 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
```

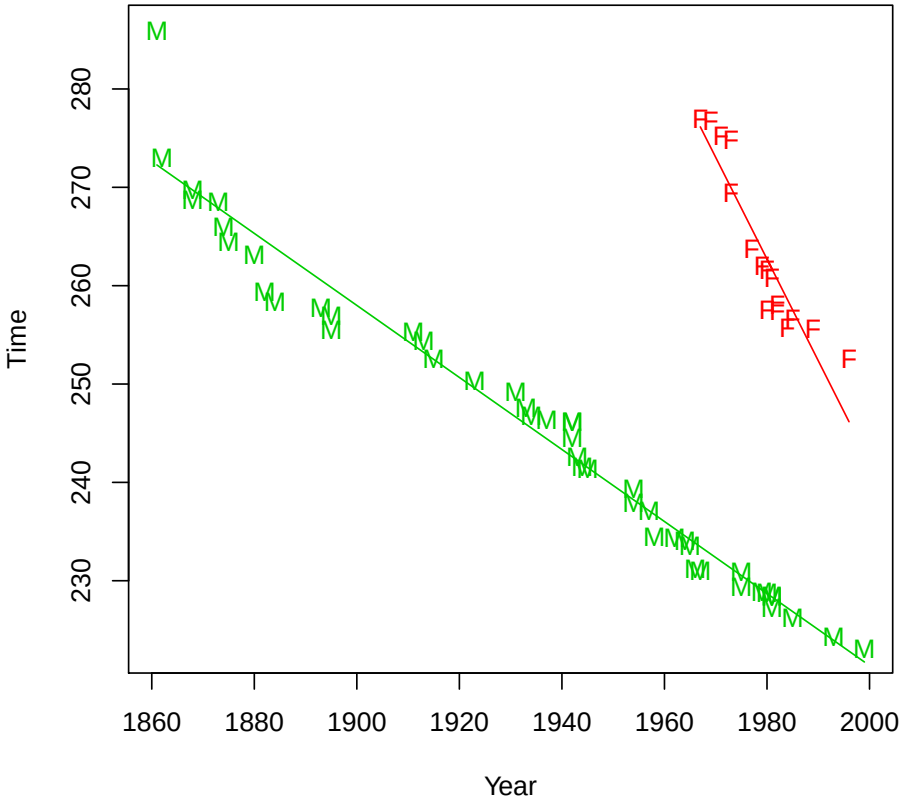
Solution [3 marks]: We conclude that there is significant interaction, i.e., there is a significant difference between the rate of improvements in the male and female records. We should use the model with interaction.

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

```
> 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)) }
```



Solution [3 marks]: For females,

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

For males,

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

(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?

```
> -imodel$coef[3]/imodel$coef[4]
```

```
GenderMale
   2030.95
```

Solution [3 marks]: 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) Is the year when the female world record will equal the male world record an estimable quantity? Is your answer consistent with part (d)?

Solution [2 marks]: 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) Calculate a 95% confidence interval for the amount by which the gap between the male and female world records narrow every year.

Solution [2 marks]:

```
> confint(imodel)[4,]
    2.5 %    97.5 %
0.4620087 0.8730100
```

(g) Test the hypothesis that the male world record decreases by 0.3 seconds each year.

```
> linearHypothesis(imodel, c(0,1,0,1), -0.3)
Linear hypothesis test

Hypothesis:
Year  + Year:GenderMale = - 0.3

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

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     59 850.63
2     58 518.03  1     332.6 37.238 9.236e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Solution [2 marks]: We reject this hypothesis; the record decreases faster than this rate.

5. 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.

Solution [5 marks]: 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) Perform the random allocation. You must use R for randomisation and include your R commands and output.

Solution [2 marks]:

```
> x <- sample(50, 50)
> x[1:9]
[1] 40  6  4 19 46 42 12 11 16
> x[10:23]
 [1] 17 14 38 36 43  3 27 15 24 34 25 13 31 37
> x[24:50]
 [1] 30 20 29 23 39  9 49 48 22  1 10 44 28 33 45 18  7 32  8  2 35 26 47  5 21
[26] 50 41
```