MAST30025: Linear Statistical Models

Akira Wang, Student ID: 913391
Assignment 3, 2019

May 31, 2019

All R output is provided at the end of the assignment. Written using LaTeXon Overleaf.

1. (a) We have $r(A^cA) \leq r(A)$. Then,

$$\begin{aligned} r(AA^cA) &\leq r(A^cA) \leq r(A) \\ r(A) &\leq r(A^cA) \leq r(A) \end{aligned}$$

Therefore $r(A^cA) = r(A)$.

(b) Matrix is idempotent iff $A^2 = A$.

$$\begin{aligned} [I - A(A^TA)^cA^T]^2 &= I^2 - 2A(A^TA)^cA^T + A(A^TA)^cA^TA(A^TA)^cA^T \\ &= I - 2A(A^TA)^cA^T + A(A^TA)^cA^T \\ &= I - A(A^TA)^cA^T. \end{aligned}$$

Therefore $I - A(A^TA)^cA^T$ is idempotent.

(c) We are given that A is $n \times p$, so I is $n \times n$.
First we show that $r(A(A^TA)^cA^T) = r(A)$.

$$\begin{aligned} r(A(A^TA)^cA^TA) &\leq r(A(A^TA)^cA^T) \leq r(A) \\ r(A) &\leq r(A(A^TA)^cA^T) \leq r(A) \\ r(A(A^TA)^cA^T) &= r(A). \end{aligned}$$

Hence,

$$\begin{aligned} r(I_n - A(A^TA)^cA^T) &= r(I_n) - r(A(A^TA)^cA^T) \\ &= n - r(A). \end{aligned}$$

2. (a) The rank is $p = r(X^TX) = 4$, so we take the 4×4 bottom right principle minor of X^TX. Therefore,

$$(X^TX)^c = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0.25 & 0 \\ 0 & 0 & 0 & 0.33 \end{bmatrix} \tag{1}$$

(b) Any solution to the normal equations can be expressed in the form

$$\mathbf{b} = (X^TX)^cX^T\mathbf{y} + [I_p - (X^TX)^cX^TX]\mathbf{z}, \tag{2}$$

where $\mathbf{z}$ is $p \times 1$.
Using the conditional matrix found from part (a),

$$\begin{aligned}
\mathbf{b} &= \begin{bmatrix} 0 \\ 45.8 \\ 35.75 \\ 55.667 \end{bmatrix} + \left(\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \right) \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} \\
&= \begin{bmatrix} 0 \\ 45.8 \\ 35.75 \\ 55.667 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} \\
&= \begin{bmatrix} 0 \\ 45.8 \\ 35.75 \\ 55.667 \end{bmatrix} + \begin{bmatrix} z_1 \\ -z_1 \\ -z_1 \\ -z_1 \end{bmatrix} \\
&= \begin{bmatrix} z_1 \\ 45.8 - z_1 \\ 35.75 - z_1 \\ 55.667 - z_1 \end{bmatrix}, \text{ for an arbitrary } z_1.
\end{aligned}$$

(c) $4\mu + 2\tau_1 + \tau_2 + \tau_3$ is estimable.

(d) Let $\mathbf{t} = \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix}^T$, the mean yield $(\mu + \tau_1)$ of tomato's grown on fertiliser 1.
The 95% Prediction Interval is given as

$$\mathbf{t}^T\mathbf{b} \pm t_{\alpha/2}s\sqrt{1 + \mathbf{t}^T(X^TX)^c\mathbf{t}} = [40.96818, 50.63182]. \tag{3}$$

(e) We test for the null hypothesis $H_0 : \tau_2 = \tau_3$.
Using the General Linear Hypothesis, we let

$$C = \begin{bmatrix} 0 & 0 & 1 & -1 \end{bmatrix}, \tag{4}$$

where $m = r(C) = 1$.
Next, we find our sample variance

$$\begin{aligned}
s^2 &= \frac{SS_{Res}}{n - r} \\
&= 3.801852.
\end{aligned}$$

Therefore our F statistic is

$$\begin{aligned}
F_{\text{stat}} &= \frac{(C\mathbf{b})^T[C(X^TX)^cC^T]^{-1}C\mathbf{b}/m}{s^2} \\
&= 178.8633 \geq F_{m,n-r} = 7.570882.
\end{aligned}$$

At the significance level of $\alpha = 0.05$, we reject the null that there is no difference in yield between fertilisers 2 and 3.

3. Let $X = \left[X_1 | X_2\right]$, and $\mathbf{z} = \begin{bmatrix}\mathbf{z}_1 \\ \mathbf{z}_2\end{bmatrix}$.

Then by Theorem 6.9,

$$\begin{aligned} E[(X\mathbf{z})^T\mathbf{y}] &= E(\mathbf{z}^T X^T \mathbf{y}) \\ &= \mathbf{z}^T X^T E(\mathbf{y}) \\ &= \mathbf{z}^T X^T X\beta \\ &= \mathbf{t}^T\beta. \end{aligned}$$

Therefore $\mathbf{t}^T\beta$ is estimable.

Now, we show that Theorem 6.3 holds and that our system is consistent.

We write our equation $X^TX|\mathbf{t}$ as an augmented matrix,

$$X^TX|\mathbf{t} = \left[\left[X_1|X_2\right] \left[\begin{array}{c} X_1^T \\ \hline X_2^T \end{array}\right] \left| \left[\begin{array}{c} \mathbf{t}_1 \\ \hline \mathbf{t}_2 \end{array}\right]\right.\right]. \tag{5}$$

Expanding the augmented matrix gives us:

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

By definition, $r(\left[X_1|X_2\right]) = r(X_1) + r(X_2)$.

Hence,

$$r\left(\begin{bmatrix} X_1^T & 0 \\ 0 & X_2^T \end{bmatrix} \left[\begin{array}{cc|c} X_1 & X_2 & X_1\mathbf{z}_1 \\ X_1 & X_2 & X_2\mathbf{z}_2 \end{array}\right]\right) \leq r\left(\begin{bmatrix} X_1^T & 0 \\ 0 & X_2^T \end{bmatrix}\right) \tag{6}$$

$$= r(X_1) + r(X_2) \tag{7}$$

$$= r(X), \tag{8}$$

and

$$r(X^TX|\mathbf{t}) \geq r(X^TX) = r(X). \tag{9}$$

Therefore

$$r(X) = r(X^TX) = r(X^TX|\mathbf{t}), \tag{10}$$

and we have shown that the system $X^TX\mathbf{z} = \mathbf{t}$ is consistent.

Thus if $\mathbf{t}_1^T\beta$ is estimable in the first model, $\mathbf{t}^T\beta$ is estimable in the second model.

4. (a) Although the plot suggests it is linear given the data, we cannot assume the data is independent. This is because a new record can only be faster than the previous record, implying that at one point the record will reach unrealistic times (such as negative time). As such, the assumptions of a linear model are violated.

   (b) When testing for interaction, we reject the null hypothesis at $\alpha = 0.05$ for interaction between the two predictor variables. Taking the context of the study into account, the interaction term tells us that there is a difference between male and female records (slope of female model is much steeper than male model).

(c) For m = male, f = female, the fitted models are:

$$\text{Time}_m = 953.7469611 - 0.3661867 \times \text{Year} \tag{11}$$

$$\text{Time}_f = 2309.424748 - 1.033696 \times \text{Year} \tag{12}$$

(d) The point estimate is 2030.95. Therefore, the suggest year for when the female record equals the male record is 2031. We expect this estimate to be inaccurate estimate since it is outside the scope of the data set, where there are simply too many other factors to consider (i.e. natural body limitations between genders).

(e) The year when the female world record will equal the male world record is estimable and therefore consistent with part (d)

(f) Let $\mathbf{t} = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}^T$, the gap between the male and female records for Year.
Then, the 95% Confidence Interval is given as

$$\mathbf{t}^T\mathbf{b} \pm t_{\alpha/2} s \sqrt{\mathbf{t}^T (X^T X)^c \mathbf{t}} = [0.4620087, 0.8730100] \tag{13}$$

(g) At $\alpha = 0.05$, we reject the null hypothesis that the male world record decreases by 0.3 seconds each year.

5. (a) We want to minimise our function

$$f(n_1, n_2, n_3, \lambda) = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{2}{n_3} \right) + \lambda(5n_1 + 2n_2 + n_3 - 100). \tag{14}$$

We differentiate with respect to n_1, n_2, n_3,

$$\begin{aligned}
\frac{\partial f}{\partial n_1} &= -\frac{\sigma^2}{n_1^2} + \lambda = 0 \\
\Rightarrow n_1^2 &= \frac{\sigma^2}{5\lambda} \\
\frac{\partial f}{\partial n_2} &= \frac{-\sigma^2}{n_2^2} + 2\lambda = 0 \\
\Rightarrow n_2^2 &= \frac{\sigma^2}{2\lambda} \\
\frac{\partial f}{\partial n_3} &= \frac{-\sigma^2}{n_3^2} + \lambda = 0 \\
\Rightarrow n_3^2 &= \frac{2\sigma^2}{\lambda}
\end{aligned}$$

Rearrange terms to get

$$\begin{aligned}
n_1 &= \frac{1}{\sqrt{10}} n_3 \\
n_2 &= \frac{1}{2} n_3.
\end{aligned}$$

Therefore,

$$\begin{aligned} 5n_1 + 2n_2 + n_3 &= 100 \\ \frac{5}{\sqrt{10}} + n_3 + n_3 &= 100 \\ \Rightarrow n_3 &= 27.92408 \approx 27 \\ \Rightarrow n_2 &= 13.96204 \approx 14 \\ \Rightarrow n_1 &= 8.830369 \approx 9, \end{aligned}$$

integer rounding to minimise the variance.

(b) *Refer to R output below.*

R

Question 2 (setup)

```
library(Matrix)
```

```
## Warning: package 'Matrix' was built under R version 3.5.2
```

```
y = c(43,45,47,46,48,33,37,38,35,56,54,57)
X = matrix(c(rep(1,12),rep(1,5),rep(0,7),rep(0,5),rep(1,4),rep(0,3),rep(0,9),rep(1,3)),12,4)
n = dim(X)[1]
r = rankMatrix(X)[1]
```

Question 2. a)

```
xtx = t(X)%*%X
xtxc = diag(c(0,1/5,1/4,1/3))

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
```

Question 2. b) (helper)

```
b = xtxc%*%t(X)%*%y

b
```

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

Question 2. c)

```
tt = c(4,2,1,1)

tt == round(tt%*%xtxc%*%xtx,3)
```

```
##      [,1] [,2] [,3] [,4]
## [1,] TRUE TRUE TRUE TRUE
```

Question 2. d)

```
tt1 = c(1,1,0,0)
e = y - X%*%b
s2 = sum(e^2)/(n-r)
ta = qt(0.975, df=(n-r))
CI = c(tt1%*%b) + c(-1,1)*c(ta*sqrt(s2)*sqrt(1+t(tt1)%*%xtxc%*%tt1))

CI
```

```
## [1] 40.96818 50.63182
```

Question 2. e)

```
C = matrix(c(0,0,1,-1),1,4)
m = rankMatrix(C)[1]
SS = t(C%*%b)%*%solve(C%*%xtxc%*%t(C))%*%C%*%b
Fstat = (SS/m)/s2

pf(Fstat, m, n-r, lower=F) < 0.05
```

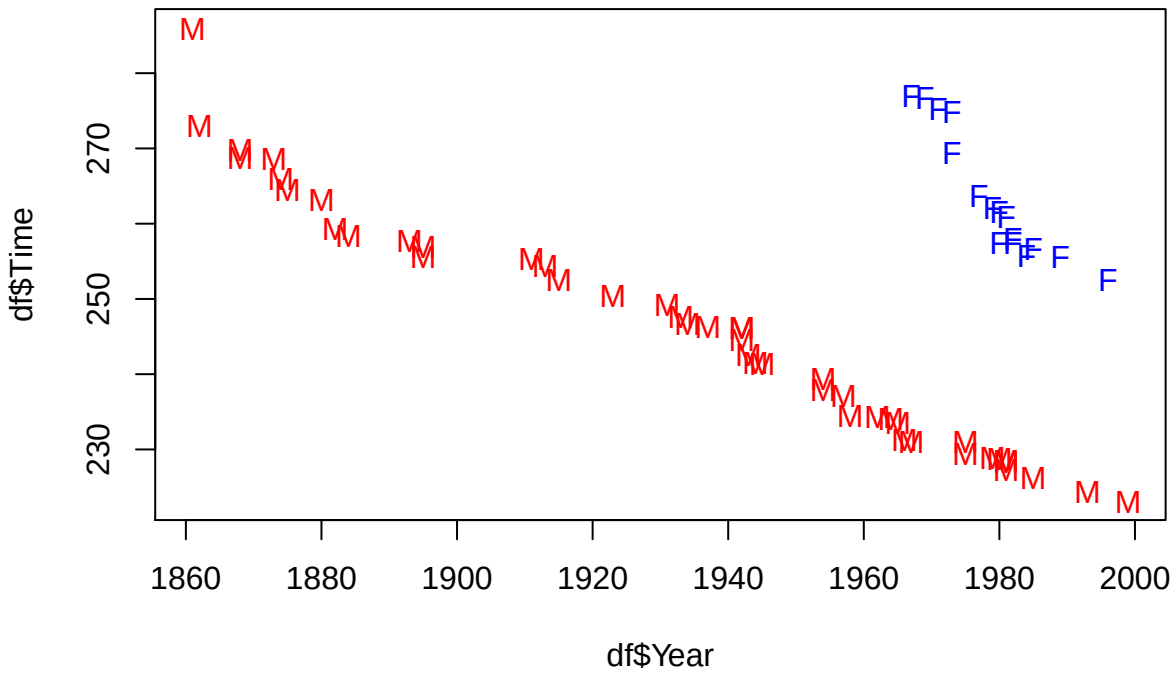
```
##      [,1]
## [1,] TRUE
```

Question 4 (setup)

```
setwd("C:\\Users\\akira\\Dropbox\\University\\Linear Statistical Models\\Lab Data")
df = read.csv("mile.csv")
```

Question 4. a)

```
palette(c("blue","red"))
plot(df$Time~df$Year, pch=array(df$Gender), col=df$Gender)
```



Question 4. b)

```
amodel = lm(Time ~ Gender+Year, df)
imodel = lm(Time ~ Gender*Year+Gender+Year, df)

anova(amodel, imodel)
```

```
## Analysis of Variance Table
##
## Model 1: Time ~ Gender + Year
## Model 2: Time ~ Gender * Year + Gender + Year
##   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
```

Question 4. c)

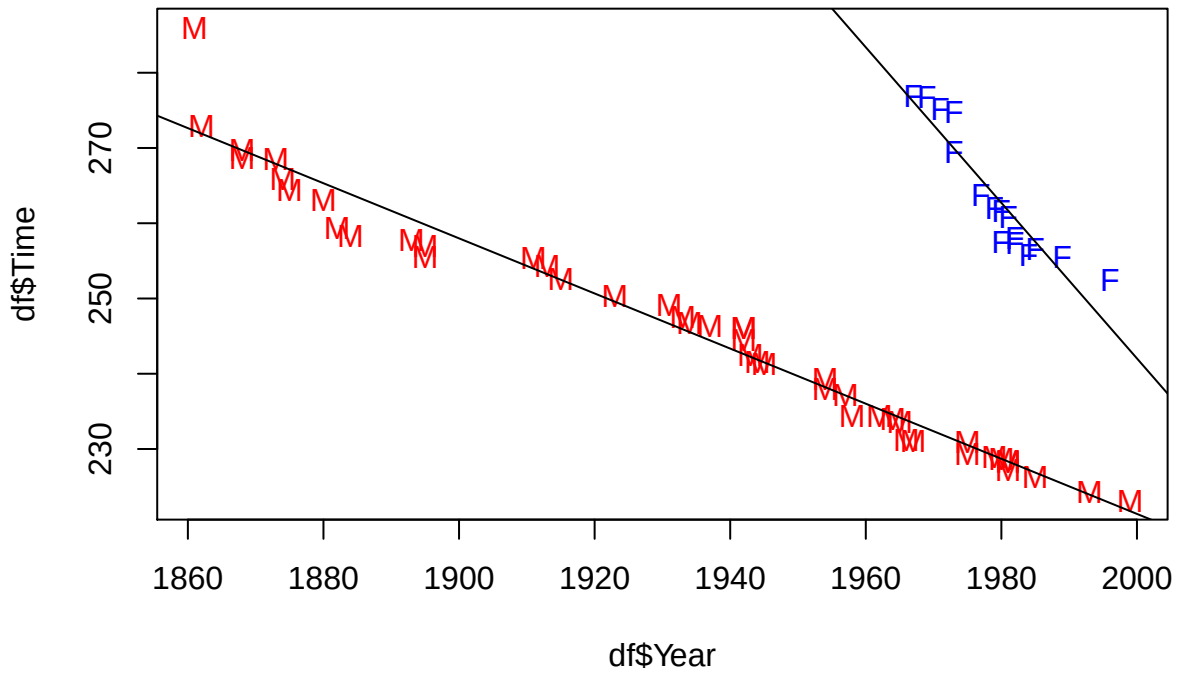
```
summary(imodel)
```

```
##
## Call:
## lm(formula = Time ~ Gender * Year + Gender + Year, data = df)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4512 -1.6160 -0.1137  1.1784 13.7265
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2309.4247   202.0583  11.429  < 2e-16 ***
## GenderMale      -1355.6778   203.1441  -6.673 1.03e-08 ***
## Year               -1.0337     0.1021 -10.126 1.95e-14 ***
## GenderMale:Year     0.6675     0.1027   6.502 2.00e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.989 on 58 degrees of freedom
## Multiple R-squared:  0.9663, Adjusted R-squared:  0.9645
## F-statistic: 553.8 on 3 and 58 DF,  p-value: < 2.2e-16
```

```r
male = c(imodel$coefficients[1] + imodel$coefficients[2], imodel$coefficients[3] + imodel$coefficients[4])
female = c(imodel$coefficients[1], imodel$coefficients[3])

plot(df$Time~df$Year, pch=array(df$Gender), col=df$Gender)
abline(male)
abline(female)
```



Question 4. d)

```
point_estimate = -imodel$coefficients[2]/imodel$coefficients[4]

point_estimate
```

```
## GenderMale
##    2030.95
```

Question 4. e)

```
tt = c(0,1,-1, 0, 2031, -2031)
n = nrow(df)
p = length(df)
X = matrix(0, n, p)
y = df$Time

X[,1] = 1
mapper = unlist(Map({function(i) if (i=="Male") 1 else 2}, df$Gender))
X[cbind(1:n, mapper+1)] = 1
X[,4] = df$Year
X[cbind(1:n, mapper+4)] = df$Year

xtx = t(X)%*%X
xtxc = matrix(0, dim(X)[2], dim(X)[2])
xtxc[c(2:3,5:6),c(2:3,5:6)] = t(solve(xtx[c(2:3,5:6),c(2:3,5:6)]))
A = t(xtxc)%*%xtx

tt == round(tt%*%A)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,] TRUE TRUE TRUE TRUE TRUE TRUE
```

Question 4. f)

```
ci = gmodels::estimable(imodel, c(0,0,0,1), conf.int=0.95)
c(ci$Lower, ci$Upper)
```

```
## [1] 0.4620087 0.8730100
```

Question 4. g)

```
car::linearHypothesis(imodel, c(0,0,1,1), -0.3)
```

```
## Linear hypothesis test
##
## Hypothesis:
## Year  + GenderMale:Year = - 0.3
##
## Model 1: restricted model
## Model 2: Time ~ Gender * Year + Gender + Year
```

```
##
##   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
```

Question 5. a)

```
n3 = 100/(5*sqrt(2/5) + 2*sqrt(1/2) + 1)
n2 = round(n3*sqrt(1/2))
n1 = round(n3*sqrt(2/5))
n3 = round(n3)
```

Question 5. b)

```
n = c(n1,n2,n3)
nsum = sum(n)
x = sample(nsum, nsum)
j1 = x[1:n[1]]
j2 = x[(n[1]+1):(n[1]+n[2])]
j3 = x[(n[1]+n[2]+1):nsum]

print("Treatment 1 Patients - $5000")
```

```
## [1] "Treatment 1 Patients - $5000"
```

```
(j1)
```

```
##  [1] 23 37 16  4 29  6 36 22 27 24 18
```

```
print("Treatment 2 Patients - $2000")
```

```
## [1] "Treatment 2 Patients - $2000"
```

```
(j2)
```

```
##  [1]  9 41 21  2 40 11 17 35  7 30 26  5 34
```

```
print("Treatment 3 Patients - $1000")
```

```
## [1] "Treatment 3 Patients - $1000"
```

```
(j3)
```

```
##  [1] 13 14 20 10 12 42 28 15  3  8 38 25 19 39 33 31  1 32
```