MAST30025 Linear Statistical Models ——Assignment 3

Name: Lei Chen
Student ID : 1013025
Tutor's name: Yao-ban Chan
Tutorial time: Fri 10-11am

1. a). Since $r(XY) \le r(X), r(Y)$.

$\Rightarrow r(A^cA) \geqslant r(A^cAA) = r(A) \geqslant r(A^cA)$

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

b). $(I - A(A^TA)^cA^T)^2$

$= (I - A(A^TA)^cA^T)(I - A(A^TA)^cA^T)$

$= I - A(A^TA)^cA^T - A(A^TA)^cA^T + A\underline{(A^TA)^cA^TA}(A^TA)^cA^T$

$= I - A(A^TA)^cA^T - A(A^TA)^cA^T + A(A^TA)^cA^T$

$= I - A(A^TA)^cA^T$.

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

c). $r(I - A(A^TA)^cA^T) = r(I) - r(A(A^TA)^cA^T)$

$= n - r(A(A^TA)^cA^T)$.

since $r(XY) \le r(X), r(Y)$.

$\Rightarrow r(A(A^TA)^cA^T) \geqslant r(A(A^TA)^cA^TA) = r(A)$

$r(A) \geqslant r(A(A^TA)^cA^T)$.

Hence. $r(A(A^TA)^cA^T) = r(A)$.

Therefore, $r(I - A(A^TA)^cA^T) = n - r(A)$.

Question 2

```
> y = matrix(c(43,45,47,46,48,33,37,38,35,56,54,57))
> X = matrix(c(rep(1,17),rep(0,12),rep(1,4),rep(0,12),rep(1,3)),12,4)

> # 2(a)
> A = t(X) %*% X
> library(Matrix)
> rankMatrix(A)[1]
[1] 3
> # r(A) = 2
> M = A[2:4,2:4]
> Ac = matrix(0,4,4)
> Ac[2:4,2:4] = solve(M)
> Ac
     [,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
```

Hence, a conditional inverse is XtXc = Ac.

```
> # 2(b)
> XtXc = Ac
> b = XtXc %*% t(X) %*% y
> b
         [,1]
[1,]  0.00000
[2,] 45.80000
[3,] 35.75000
[4,] 55.66667
> I = diag(c(rep(1,4)))
> I - 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
```

One solution to the normal equation is b.
Another solution to the normal equation is b2 = b + (I - XtXc %*% t(X) %*% X) %*% z,
Where z is an arbitrary 4*1 vector.

```
> # 2(c)
> tt = c(4,2,1,1)
> t(tt) %*% XtXc %*% t(X) %*% X
     [,1] [,2] [,3] [,4]
[1,]    4    2    1    1
```

So, yes, it is estimable.

```
> # 2(d)
> n = length(y)
> library(Matrix)
> r = rankMatrix(X)[1]
> e = y - X %*% b
> (s2 = sum(e^2)/(n-r))
[1] 3.801852
> tt = c(1,1,0,0)
> t(tt) %*% XtXc %*% t(X) %*% X
     [,1] [,2] [,3] [,4]
[1,]    1    1    0    0        #so estimable
> hw = qt(0.975, df=n-r) * sqrt(s2) * sqrt(1 + t(tt) %*% XtXc %*% tt)
> c(tt %*% b - hw, tt %*% b + hw)
[1] 40.96818 50.63182
```

95% prediction interval for the yield of a tomato plant grown on fertiliser 1 is (40.96818, 50.63182).

```
> # 2(e)
> # H0:  τ2 − τ3 = 0
> 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 = F)
             [,1]
[1,] 3.042802e-07
```

So, we can reject the null hypothesis at a 5% level:
fertilisers 2 and 3 have no difference in yield.

3. Since $t_1^T\beta_1$ is estimable. Apply Theorem 6.9 (In the general linear model $y = X\beta + \varepsilon$, $t^T\beta$ is estimable if and only if there is a solution to the linear system $X^TXz = t$.)

$\Rightarrow$ There is a solution to the linear system $X_1^TX_1z_1 = t_1$.

Since X_2 is full rank, $X_2^TX_2$ has inverse. Hence, there is ~~an~~ solution to the linear system $X_2^TX_2z = t_2$.

$\Rightarrow$ ~~$z_2 = $~~ $t_2 = X_2^TX_2z_2$.

Thus, we now have $X^TXz = t = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$.

Now, Apply Theorem 6.3. (The system $Ax = g$ is consistent if and only if the rank of $(A|g)$ is equal to the rank of A).

$\Rightarrow$ To prove $t^T\beta$ is estimable in the second model, we need to prove the rank of $[X^TX | \begin{smallmatrix} t_1 \\ t_2 \end{smallmatrix}]$ is equal to the rank of $[X^TX]$.

~~rank~~ $$r([X^TX|t]) = r\left(\left[\begin{matrix} X_1^TX_1 & X_1^TX_2 \\ X_2^TX_1 & X_2^TX_2 \end{matrix} \middle| \begin{matrix} X_1^TX_1z_1 \\ X_2^TX_2z_2 \end{matrix}\right]\right)$$
$$= r\left(\begin{bmatrix} X_1^T & 0 \\ 0 & X_2^T \end{bmatrix}\left[\begin{matrix} X_1 & X_2 \\ X_1 & X_2 \end{matrix} \middle| \begin{matrix} X_1z_1 \\ X_2z_2 \end{matrix}\right]\right).$$
$$\leq r\left(\begin{bmatrix} X_1^T & 0 \\ 0 & X_2^T \end{bmatrix}\right)$$
$$= r(X_1^T) + r(X_2^T).$$
$$= r(X_1) + r(X_2)$$
$$= r(X).$$

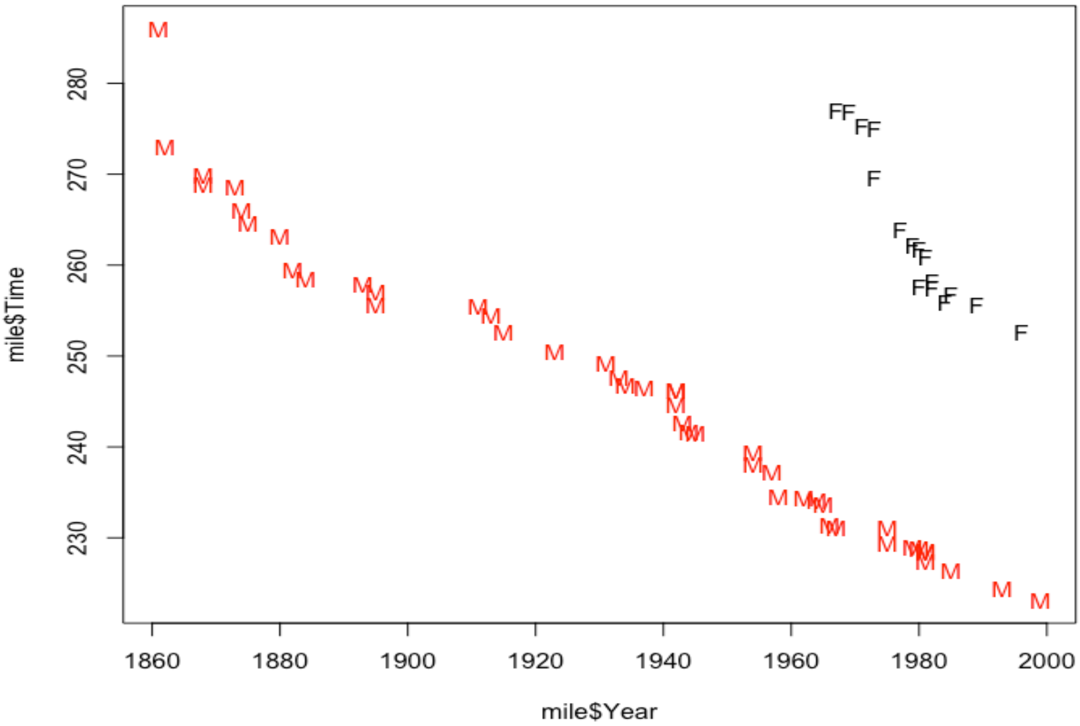
$$r([X^TX|t]) \geq r(X^TX) = r(X).$$

Therefore, we have $r(X^TX|t) = r(X) = r(X^TX)$.

So we proved that $t^T\beta$ is estimable in the second model.

Question 4

```
> # 4(a)
> mile = read.csv('mile.csv')
> mile$Gender.f = factor(mile$Gender)
> plot(mile$Year, mile$Time, pch=array(mile$Gender.f), col=mile$Gender)
```



The data looks linear, but the record can only decrease.
Hence, the data are not independent and cannot satisfy the linear model assumptions.

```
> # 4(b)
> imodel = lm(Time ~ Year * Gender, data = mile)
> summary(imodel)
```

```
Call:
lm(formula = Time ~ Year * Gender, data = mile)

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 ***
Year            -1.0337     0.1021 -10.126 1.95e-14 ***
GenderMale     -1355.6778   203.1441  -6.673 1.03e-08 ***
Year:GenderMale     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

> amodel = lm(Time ~ Year + Gender, data = mile)
> summary(amodel)
Call:
lm(formula = Time ~ Year + Gender, data = mile)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9071 -2.0988 -0.1141  1.2002 13.1863

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1003.00334   27.84691   36.02   <2e-16 ***
Year          -0.37364    0.01406  -26.57   <2e-16 ***
GenderMale   -34.85078    1.30099  -26.79   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.896 on 59 degrees of freedom
Multiple R-squared:  0.9417,  Adjusted R-squared:  0.9397
F-statistic: 476.3 on 2 and 59 DF,  p-value: < 2.2e-16

> anova(imodel, amodel)
Analysis of Variance Table

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

Since 2.001e-08 << 0.05,we reject H0 and conclude that there is a significant interaction between two predict variables.

```
> # 4(c)
lm(formula = Time ~ Year * Gender, data = mile)

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 ***
Year              -1.0337     0.1021 -10.126 1.95e-14 ***
GenderMale     -1355.6778   203.1441  -6.673 1.03e-08 ***
Year:GenderMale    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
```

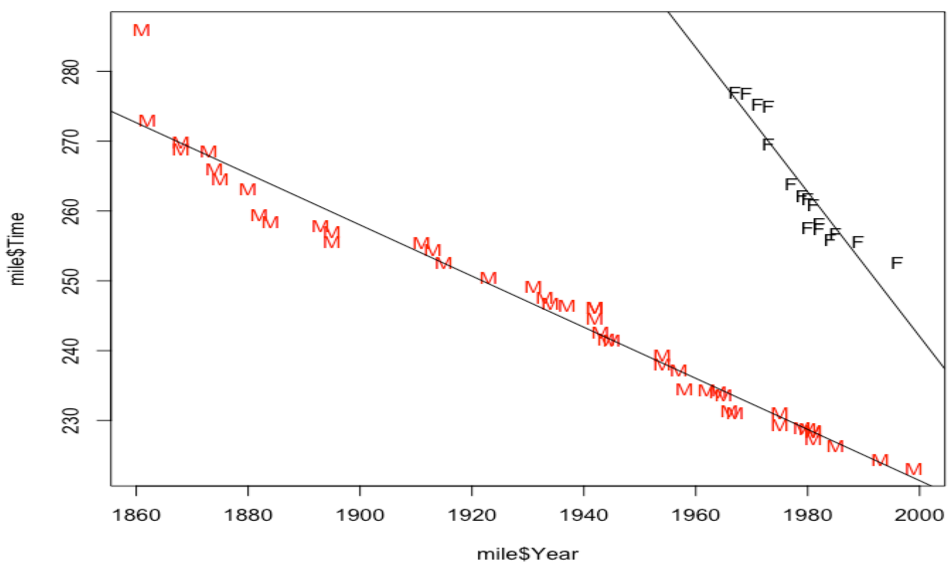
We should use the model with interaction, so the final fitted models:

For females: Time = 2309.4247 - 1.0337 * Year

For males: Time = (2309.4247-1355.6778) + (-1.0337+0.6675)* Year

=953.747 - 0.3662 * Year

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



```
> # 4(d)
> -imodel$coef[3]/imodel$coef[4]
GenderMale
   2030.95
```

We expect that the female wrold record will equal the male world record around the year 2031. However this is unlikely to be an accurate estimate since this is beyond the range of the data.

4.(e) The model is : $y_{ij} = \mu + \tau_i + \beta x_{ij} + \xi_i x_{ij} + \varepsilon_{ij}$.

the quantity is : $\frac{-(\tau_2 - \tau_1)}{\xi_2 - \xi_1} = $ Year.

since $(\tau_2 - \tau_1)$ and $(\xi_2 - \xi_1)$ are contrasts, so the answer is consistent with part (d).

Thus, this is esitmable.

```
> # 4(f)
> se <- coef(summary(imodel))[,2]
> imodel$coef[4] + c(-1,1)*qt(0.975,df=58)*se[4]
[1] 0.4620087 0.8730100
```

Hence, 95% confidence interval for the amount by which the gap between the male and female world records narrow every year is (0.4620087,0.87301).

```
> # 4(g)
> library(car)
> 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

  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
```

Since 9.236e-08 << 0.05, so we can reject the H0.
We conclude that the male world record is not decreases by 0.3 seconds each year.

5. a) Since we wish to test if the treatments are effective: $H_0: \tau_1 = \tau_2 = \tau_3$. So we want to study the treatments contrasts $\tau_2 - \tau_1$, $\tau_3 - \tau_1$. We have $var\ \widehat{\tau_i - \tau_1} = \sigma^2 (\frac{1}{n_1} + \frac{1}{n_i})$.

and we need: $5000 n_1 + 2000 n_2 + 1000 n_3 \le 100000$

$$\Rightarrow 5n_1 + 2n_2 + n_3 \le 100.$$

So we minimise:

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

we get: $\frac{\partial f}{\partial n_1} = -2\frac{\sigma^2}{n_1^2} + 5\lambda = 0$

$$n_1^2 = \frac{2\sigma^2}{5\lambda}$$

$$\frac{\partial f}{\partial n_2} = -\frac{\sigma^2}{n_2^2} + 2\lambda = 0$$

$$n_2^2 = \frac{\sigma^2}{2\lambda}$$

$$\frac{\partial f}{\partial n_3} = -\frac{\sigma^2}{n_3^2} + \lambda = 0$$

$$n_3^2 = \frac{\sigma^2}{\lambda}$$

$$\frac{\partial f}{\partial \lambda} = 5n_1 + 2n_2 + n_3 - 100 = 0.$$

Therefore: $n_1^2 = \frac{4}{5} n_2^2 = \frac{2}{5} n_3^2 \quad \Rightarrow n_1 = \frac{2}{\sqrt{5}} n_2 = \sqrt{\frac{2}{5}} n_3$

$$\Rightarrow 5n_1 + 2 \cdot \frac{\sqrt{5}}{2} n_1 + \sqrt{\frac{5}{2}} n_1 = 100$$

$$n_1 (5 + \sqrt{5} + \sqrt{\tfrac{5}{2}}) = 100$$

$$n_1 \approx 11.3415$$

$n_2 = \frac{\sqrt{5}}{2} n_1 \approx 12.68$, $n_3 = n_1 / \sqrt{\frac{2}{5}} = 17.932$.

~~$n_3 = 100 - 55 - 2\times 13 = 19$~~ (we choose $n_1 = 11$, $n_2 = 13$, $n_3 = 19$: $11 \times 5000 + 13 \times 2000 + 19 \times 1000 = 100000$)

Therefore, we choose 11 treatment 1, 13 treatment 2 and 19 treatment 3. (n1 = 11, n2 = 13, n3 = 19)

```
> # 5(b)
> n = c(11, 13, 19)
> nsum = sum(n)
> x = sample(nsum, nsum)

> (j1 = x[1:n[1]])     # treatment 1
 [1] 33  4 21 41 32 36 23 40 10 29 18
> (j2 = x[(n[1] + 1):(n[1] + n[2])])     # treatment 2
 [1] 24 34 25 39 19 31  7 20 22 38 35 43  8
> (j3 = x[(n[1] + n[2] + 1): nsum])     # treatment 3
 [1] 16  2 13 11  1 17  5  6 30 15 12 14 42 37  3 27  9 28 26
```