THE UNIVERSITY OF MELBOURNE

Student number 998008

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 3 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.

Linear Statistical Models Assignment 3

Kim Seang CHY

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

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

Let $\mathbf{A} = \left[\begin{array}{c|c} \mathbf{M} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right]$ where $\mathbf{M}$ is a square matrix with $r(\mathbf{A}) \times r(\mathbf{A})$ and $r(\mathbf{A}) = a \leq p$.
By definition $\mathbf{M}$ is a matrix with $r(\mathbf{M}) = r(\mathbf{A})$

Using theorem 6.2, we can find $\mathbf{A}^c$ such that it has the following partition:
$\mathbf{A}^c = \left[\begin{array}{c|c} \mathbf{M}^{-1} & 0 \\ \hline 0 & 0 \end{array}\right]$ where $\mathbf{M}^{-1}$ is the inverse of $\mathbf{M}$.

Hence $\mathbf{A}^c\mathbf{A} = \left[\begin{array}{c|c} I_a & \mathbf{M}^{-1}A_{12} \\ \hline 0 & 0 \end{array}\right]$

All column vectors in $\mathbf{M}^{-1}A_{1,2}$ is a linear combination of the independent column vectors in the identity matrix I_a. Thus $r(\mathbf{A}^c\mathbf{A}) = r(I_a) = r(\mathbf{M}) = a$. Since $r(\mathbf{A}) = r(\mathbf{M})$, this implied $r(\mathbf{A}) = r(\mathbf{A}^c\mathbf{A})$.

b. Show directly that A^cA is idempotent.

Since A^c is a conditional inverse for A then AA^cA=A. Thus,

$$\begin{aligned}(A^cA)^2 &= A^cAA^cA \\ &= A^cA\end{aligned}$$

Hence, A^cA is idempotent

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

Consider conditional inverse $(A^TA)_1^c$ and an arbitrary conditional inverse $(A^TA)_i^c$ where $i \neq 1$. Now using the properties $A = A(A^TA)_i^c(A^TA)$ and $A^T = (A^TA)_1^c(A^TA)A^T$., we get the following:

$$\begin{aligned}A(A^TA)_1^cA^T &= A(A^TA)_i^c(A^TA)(A^TA)_1^cA^T \\ &= A(A^TA)_i^cA^T\end{aligned}$$

Since,$A(A^TA)_1^cA^T = A(A^TA)_i^cA^T$, this implied it is unique and invariant to the choice of conditional inverse.

Question 2:

Pre-question

```
#Setting up Matrix Y
Y <- matrix(c(22,23,24,22,26,16,18,19,28,27,29,29),12,1)
Y
```

```
##       [,1]
##  [1,]   22
##  [2,]   23
##  [3,]   24
##  [4,]   22
##  [5,]   26
##  [6,]   16
##  [7,]   18
##  [8,]   19
##  [9,]   28
## [10,]   27
## [11,]   29
## [12,]   29
```

```
#Setting up matrix X
X <- matrix(rep(0,36),12,4)
X[,1] = 1 ;X[1:5,2]=1; X[6:8,3]=1; X[9:12,4]=1
X
```

```
##       [,1] [,2] [,3] [,4]
##  [1,]    1    1    0    0
##  [2,]    1    1    0    0
##  [3,]    1    1    0    0
##  [4,]    1    1    0    0
##  [5,]    1    1    0    0
##  [6,]    1    0    1    0
##  [7,]    1    0    1    0
##  [8,]    1    0    1    0
##  [9,]    1    0    0    1
## [10,]    1    0    0    1
## [11,]    1    0    0    1
## [12,]    1    0    0    1
```

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

```
#Computing X^tX
XtX <- t(X)%*%(X)
XtXc <- matrix(rep(0,36),4,4)
#Computing one of the conditional inverse
XtXc[2:4,2:4] <-solve(XtX[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.Find s^2.

```
#Computing s2
df <- 12-3 #n=12 and reading X we can see r(X)=3
I <- diag(12)
SSres <- t(Y)%*%(I-X%*%XtXc%*%t(X))%*%Y
SSres
```

```
##           [,1]
## [1,] 18.61667
```

```
s2 <- SSres/df
```

c.Is $\mu + 2\tau_1 + \tau_2$ estimable?

```
#Testing for estimablility using Thm 6.10
t1 <- matrix(c(1,2,1,0),4,1)
t(t1)%*%XtXc%*%XtX
```

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

Since $\mathbf{t}_1^T(X^TX)^C(X^TX) \neq \mathbf{t}_1^T$, by theorem 6.10, $\mu + 2\tau_1 + \tau_2$ is not estimable.

c.Find a 90% confidence interval for the lifetime of the 2nd type of bulb.

```
#Computing Beta
b <- XtXc%*%t(X)%*%Y
b
```

```
##           [,1]
## [1,]  0.00000
## [2,] 23.40000
## [3,] 17.66667
## [4,] 28.25000
```

```
#Computing 90% CI interval for 2nd type of bulb
t2 <- matrix(c(0,0,1,0),4,1)
tstat_halfalpha <- qt(0.95,df)
t2.std <- tstat_halfalpha*sqrt(s2*(t(t2)%*%XtXc%*%t2))
CI.90 <- c(t(t2)%*%b - t2.std,t(t2)%*%b + t2.std)
CI.90
```

```
## [1] 16.14451 19.18882
```

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

Testing for $H_0 : \tau_1 = \tau_3$ vs $H_1 : \tau_1 \neq \tau_3$

```
#Testing for H_0:tau.1=tau.3 and H_1:tau.1!=tau.3
C <- matrix(c(0,1,0,-1),1,4)
#C has a rank of 2
dst <- 0
numerator <- (t(C%*%b-dst)%*%solve(C%*%XtXc%*%t(C))%*%C%*%b-dst)
```

```r
Fstat <- (numerator/2)/s2
pf(Fstat,2,df,lower=F)
```

```
##              [,1]
## [1,] 0.002437569
```

We can firmly reject the null at a 5% statistical significant, thus there are difference in lifetime between 1st and and 3rd types of bulb.

Question 3:

Let $t = \left[\begin{array}{c} t_1 \\ \hline t_2 \end{array}\right] = \left[\begin{array}{c} X_1^T X_1 z_1 \\ \hline X_2^T X_2 z_2 \end{array}\right]$ and $X = \left[\begin{array}{c|c} X_1 & X_2 \end{array}\right]$

$$X^T X\mathbf{a} = \left[\begin{array}{c} X_1^T \\ \hline X_2^T \end{array}\right] \left[\begin{array}{c|c} X_1 & X_2 \end{array}\right] \mathbf{a} = \left[\begin{array}{c|c} X_1^T X_1 & X_1^T X_2 \\ \hline X_2^T X_1 & X_2^T X_2 \end{array}\right] \mathbf{a}$$

We can to rewrite the system of linear equation for $X^T X\mathbf{a} = t$, as an augmented matrix form as $[X^T X|t]$.

$$[X^T X|t] = \left[\begin{array}{cc|c} X_1^T X_1 & X_1^T X_2 & X_1^T X_1 z_1 \\ X_2^T X_1 & X_2^T X_2 & X_2^T X_2 z_2 \end{array}\right] = \left[\begin{array}{cc} X_1^T & 0 \\ 0 & X_2^T \end{array}\right] \left[\begin{array}{cc|c} X_1 & X_2 & X_1 z_1 \\ X_1 & X_2 & X_2 z_2 \end{array}\right]$$

Since, X_2 is continuous factor we can inferred that X_2 is column full rank and since X_1 less then full column rank we can inferred that X_1 can be written as a linear combination X_2.

Thus by theorem 6.2 that $r\left(\left[\begin{array}{cc|c} X_1 & X_2 & X_1 z_1 \\ X_1 & X_2 & X_2 z_2 \end{array}\right]\right) = r(X^T X)$ if and only if $\left[\begin{array}{cc|c} X_1 & X_2 & X_1 z_1 \\ X_1 & X_2 & X_2 z_2 \end{array}\right]$ is a consistent system.

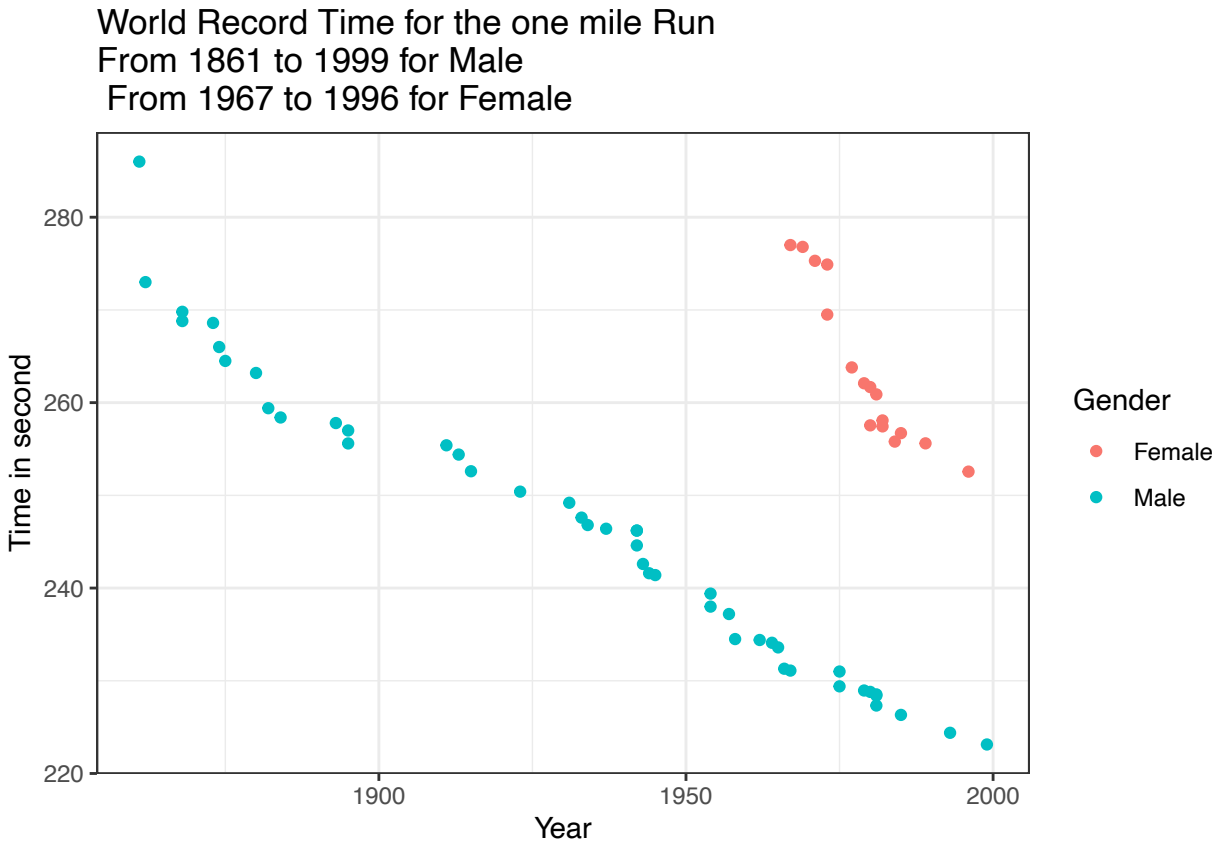
Thus by the fact that X_2 is a linear combination of $X_1$and then $\left[\begin{array}{cc} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{array}\right]$ are linear combination of $\left[\begin{array}{c} t_1 \\ \hline t_2 \end{array}\right]$ if there exist a z_1 such that $X_1^T X_1 z_1 = t_1$ is a consistent system.

Since, $t_1^T \beta_1$ is estimable then $X_1^T X_1 z_1 = t_1$ is a consistent system hence $X^T X z = t$ is also a consistent system. Thus, if $t_1^T \beta_1$ is estimable then $t^T \beta$ where $\beta^T = [\beta_1^T | \beta_2^T]$.

Question 4:

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?

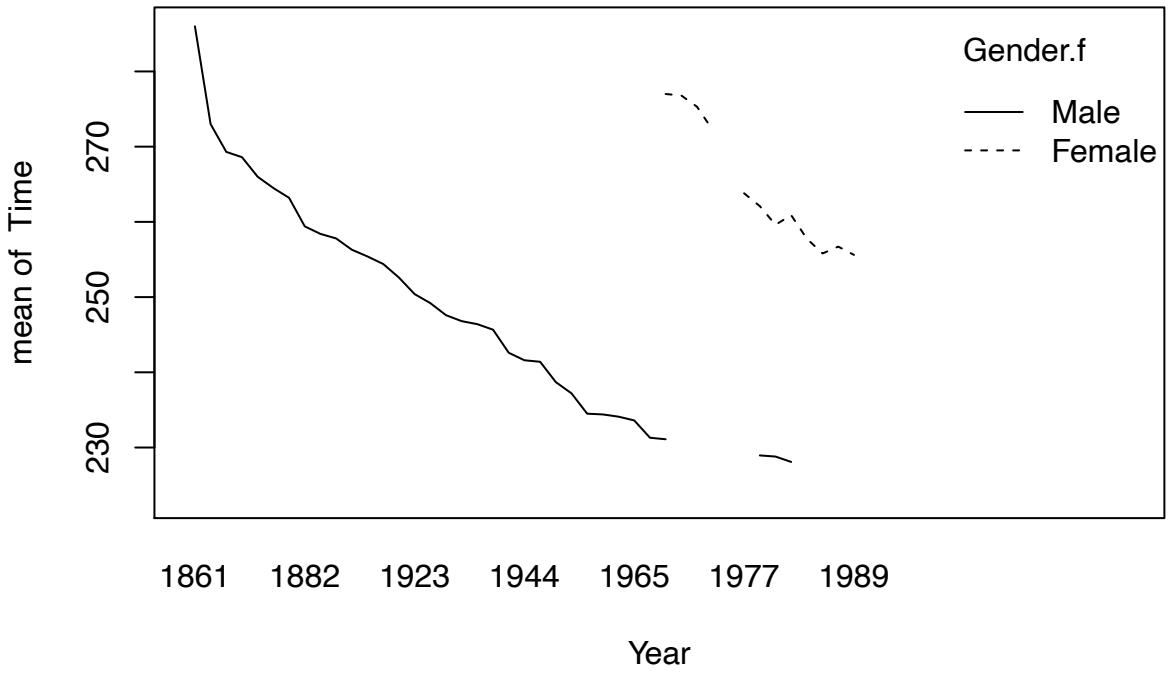
```
library(ggplot2)
mile2 <- read.csv(file="mile2.csv")
mile2$Gender.f <- factor(mile2$Gender)
g1 <- ggplot(mile2, aes(x=Year, y=Time,colour=Gender)) + geom_point()
g2 <- g1 + labs(x="Year", y="Time in second",
         title="World Record Time for the one mile Run
From 1861 to 1999 for Male \n From 1967 to 1996 for Female")
g2 + theme_bw()
```



From the graph, we can see a linear relationship for Male. Homoscedasticity seem to also be met as beside the big outlier of 1861, we can the point seem be randomly distributed around the regression.For the big outlier we will need to do further analysis to see if the point has a high crook's distance. For female beside the lack of a big outlier, the same thing could be said about the linear relationship and homoscedasticity assumption.

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

```
#Determine if there are interaction using R
with(mile2, interaction.plot(Year,Gender.f, Time))
```




```
model <- lm(Time~Gender.f+Year, mile2)
imodel <- lm(Time~Gender.f*Year, mile2)
anova(model,imodel)
```

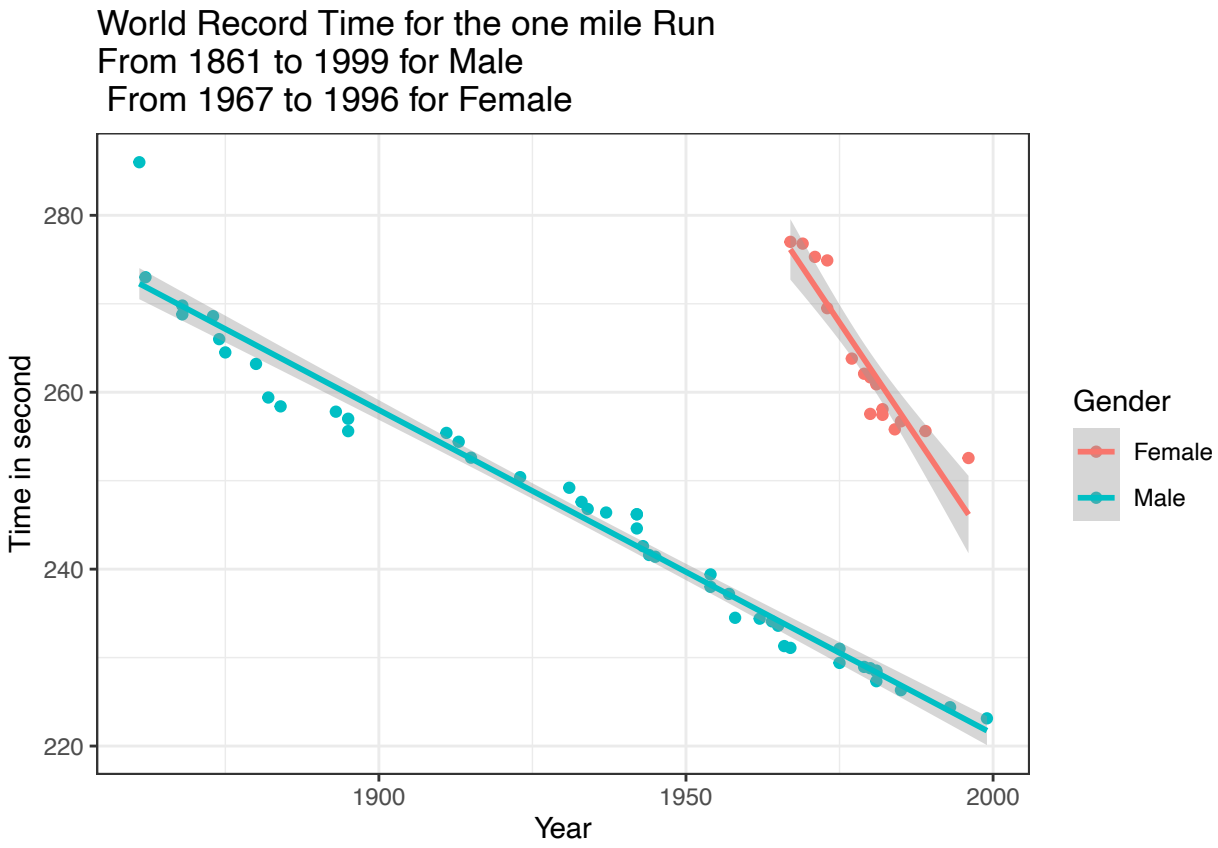
```
## Analysis of Variance Table
##
## Model 1: Time ~ Gender.f + Year
## Model 2: Time ~ Gender.f * Year
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     59 895.62
## 2     58 518.03  1    377.59 42.276 2.001e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1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

Looking at the interaction we can see that there an interaction between year and gender as the line is not parallel to one another. Furthermore, this is confirmed by our statistical test where we reject the null at 5%. Thus indicating

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).

```
library(MASS)
library(Matrix)
g3 <- g2 + geom_smooth(method = lm)
g3+theme_bw()
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
#Computing Beta
Y <- mile2$Time
n <- length(Y)
X <- matrix(0,n,6)
X[,1] <- 1
X[cbind(1:n,as.numeric(mile2$Gender.f)+1)] <-1
X[,4] <- mile2$Year
X[cbind(1:n,as.numeric(mile2$Gender.f)+4)] <- mile2$Year
XtX <- t(X)%*%X
r <- rankMatrix(X)[1]
XtXc <- matrix(0,6,6)
M <- XtX[c(2:3,5:6),c(2:3,5:6)]
XtXc[c(2:3,5:6),c(2:3,5:6)] <-  solve(M)
b <- XtXc%*%t(X)%*%Y
b
```

```
##              [,1]
## [1,]    0.0000000
## [2,] 2309.4247477
```

```
## [3,]  953.7469611
## [4,]    0.0000000
## [5,]   -1.0336960
## [6,]   -0.3661867
```

The model is given by:

$$\text{Time} = 0.0321 - 2309.4247 \text{ Female} - 953.747 \text{ Male} + 0.0322 - 1.0337 \text{ Female:Year} - 0.3662 \text{ Male:Year}$$

Where Female is an indicator taking the value of 1 for female and 0 for male. Similarly, Male is also an indicator with value of 1 for male and 0 for female.

Hence the respective subpopulation model for female is given by:

$$\text{Time}_{Female} = 2309.4247 - 1.0337 \text{ Year}$$

For male is given by:

$$\text{Time}_{Male} = 953.747 - 0.3662 \text{ Year}$$

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?

```
(1355.6777866)/(0.6675093)
```

```
## [1] 2030.95
```

```
#Computing time taken in 2030.95 for female
t1 <- matrix(c(1,1,0,1,(1355.6777866)/(0.6675093),0),6,1)
t(t1)%*%b
```

```
##          [,1]
## [1,] 210.0401
```

```
#Computing time taken in 2030.95 for female
t2 <- matrix(c(1,0,1,1,0,(1355.6777866)/(0.6675093)),6,1)
t(t2)%*%b
```

```
##          [,1]
## [1,] 210.0402
```

Letting $\text{Time}_{Female} = \text{Time}_{Male}$ or $2309.4247 - 1.0337 \text{ Year} = 953.747 - 0.3662 \text{ Year}$, we get the estimate year where male and female will be equal to each other near the end of the year 2030 with a time of 210.1045 second.
First of all such a point would not be estimable on our model.Secondly, our model may not have the power to predict what happen in the future because the model assumed perpuity improvement instead of taking into account of the physical limitation gap between Male and Female.

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)?

```
t(t1)%*%XtXc%*%XtX
```

```
##      [,1] [,2] [,3]    [,4]    [,5] [,6]
## [1,]    1    1    0 2030.95 2030.95    0
```

```
t(t2)%*%XtXc%*%XtX
```

```
##      [,1] [,2] [,3]    [,4] [,5]    [,6]
## [1,]    1    0    1 2030.95    0 2030.95
```

Using theorem 6.10 we can see that both $\mathbf{t}_1^T(X^TX)^CX^TX) \neq \mathbf{t}_1$. Similarly, $\mathbf{t}_2^T(X^TX)^CX^TX) \neq \mathbf{t}_2$. Thus, such a point estimate is not estimatable which was consistent with the answer given in part c.

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

```
confint(imodel)
```

```
##                           2.5 %       97.5 %
## (Intercept)        1904.9610773 2713.8884180
## Gender.fMale      -1762.3149667 -949.0406065
## Year                 -1.2380466   -0.8293454
## Gender.fMale:Year     0.4620087    0.8730100
```

The Gender.fMale:Year is the differences between female and male record each year. Reading from the above we get the 95% CI for amount of gap by which female and male record narrow every year of $(0.462, 0.873)$ second every year.

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

We want to test $H_0 : \beta_2 = -0.4$

```
C <- matrix(c(0,0,0,0,0,1),1,6)
#C has a rank of 1
dst <- -0.4
numerator <- t(C%*%b-dst)%*%solve(C%*%XtXc%*%t(C))%*%(C%*%b-dst)
Fstat <- (numerator)/s2
Fstat
```

```
##          [,1]
## [1,] 41.96548
```

```
pf(Fstat,1,df,lower=F)
```

```
##              [,1]
## [1,] 0.0001143321
```

We can reject the null that the male world record will decreases by -0.4 second at 5% statistical significant.

Question 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.

We are testing for $H_0 : \tau_1 = \tau_2 = \tau_3$ or equivalently we are testing the contrast differences $\tau_1 - \tau_3$ and $\tau_2 - \tau_3$. By doing this I put a heavy emphasis on placebo group relative to other two groups.

The first constraint we have is $5n_1 + 2n_2 + n_3 = 100$

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

$$f(n_1, n_2, n_3, \lambda) = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{2}{n_3} \right) + \lambda \left(\sum_{i=1}^{3} n_i - n \right)$$

Solve $\frac{df}{dn_i} = 0$, we get $n_3 = \sqrt{2}n_2 = \sqrt{2}n_2$. Subbing this back into our first constraint of $5n_1 + 2n_2 + n_3 = 100$.

```
#Solving for n3
100/(1+(7/sqrt(2)))
```

```
## [1] 16.80744
```

```
#Solving for n1=n2
16.80744/sqrt(2)
```

```
## [1] 11.88465
```

We get $n_1 = n_2 = 12$ and $n_3 = 16$ after rounding out the samples.

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

```
#Performing random selection
n <- c(12,12,16)
nsum <- sum(n)
x <- sample(nsum, nsum)
n1 <- x[1:n[1]]
n2 <- x[n[1]+1:n[2]]
n3 <- x[n[2]+1:n[3]]
```

Thus our assignment to treatment 1 for y_i will be:

```
n1
```

```
##  [1] 14 25 35 11  6 23 30  1 29 20 36  7
```

Our assignment to treatment 2 for y_i will be:

```
n2
```

```
##  [1]  3 24 12 33 19 10 40 28 17 13 21  5
```

And our assignment to placeobo for y_i will be:

```
n3
```

```
##  [1]  3 24 12 33 19 10 40 28 17 13 21  5  4 15 18  8
```