MAST30025 Assignment 3 2019 R Code

#Question 2 #Part a: Find a conditional inverse t(X)%\*%X, using the algorrithm given in Theorem 6.2

#Creating a data frame first

yield = c(43,45,47,46,48,33,37,38,35,56,54,57)  
fertiliser = c(1,1,1,1,1,2,2,2,2,3,3,3)  
typefert = factor(fertiliser)

tomato = data.frame(yield,typefert)  
tomato

## yield typefert  
## 1 43 1  
## 2 45 1  
## 3 47 1  
## 4 46 1  
## 5 48 1  
## 6 33 2  
## 7 37 2  
## 8 38 2  
## 9 35 2  
## 10 56 3  
## 11 54 3  
## 12 57 3

#Response variable  
y = tomato$yield  
#Our Design Matrix  
X = matrix(0,12,4)  
X[,1]=1  
for (i in 1:3){X[typefert == i, i +1] = 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 1 0  
## [10,] 1 0 0 1  
## [11,] 1 0 0 1  
## [12,] 1 0 0 1

#Finding the Condition Inverse according from the lectures!  
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

#Part b: Characterise all solutions to the normal equations.

#First from the tutorials we find another Conditional inverse using the ginv function  
library(MASS)  
library(Matrix)  
XtXc2 = ginv(t(X)%\*%X)  
XtXc2

## [,1] [,2] [,3] [,4]  
## [1,] 0.048958333 0.001041667 0.01354167 0.034375  
## [2,] 0.001041667 0.148958333 -0.06354167 -0.084375  
## [3,] 0.013541667 -0.063541667 0.17395833 -0.096875  
## [4,] 0.034375000 -0.084375000 -0.09687500 0.215625

b = XtXc%\*%t(X)%\*%y  
b

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

b2 = XtXc2%\*%t(X)%\*%y  
b2

## [,1]  
## [1,] 34.304167  
## [2,] 11.495833  
## [3,] 1.445833  
## [4,] 21.362500

#More to add below #According to Theorem 6.7

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.

#Part c: Is 4μ + 2τ1 + τ2 + τ3 estimable? #Attempt 1

library(MASS)  
library(Matrix)  
tt = c(4,2,1,1)  
tt%\*%ginv(t(X)%\*%X)%\*%t(X)%\*%X

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

#It is estimable which satisfies Theorem 6.10 # Tutor Comment: Yes, as it is a linear combination of μ + τ1, μ + τ2 and μ + τ3, which are all elements of Xβ and therefore estimable

#Part d: Find a 95% prediction interval for the yield of a tomato plant grown on fertiliser 1.

library(Matrix)  
n = 12  
SSRes = sum((y-X%\*%b)^2)  
s2 = SSRes/(n-rankMatrix(X)[1])  
#Attempt 1  
tt = c(1,1,0,0) #yield from fertiliser 1  
ta = qt(0.975,n-rankMatrix(X)[1])  
s = sqrt(s2)  
  
#Creating our Prediction Interval  
halfwidth = ta\*s\*sqrt(1+t(tt)%\*%ginv(t(X)%\*%X)%\*%tt)  
c(tt%\*%b-halfwidth,tt%\*%b+halfwidth)

## [1] 40.96818 50.63182

#Actual Solution (Correct from above)

#Part e: Test the hypothesis that fertilisers 2 and 3 have no difference in yield.

#Attempt 1  
library(Matrix)  
C = matrix(c(1,0,-1,0,1,0,0,-1),2,4)  
m = rankMatrix(C)[1]  
num = t(C%\*%b)%\*%solve(C%\*%ginv(t(X)%\*%X)%\*%t(C))%\*%C%\*%b  
Fstat = (num/m)/s2  
Fstat

## [,1]  
## [1,] 1952.131

pf(Fstat,m,n-rankMatrix(X)[1], lower = F)

## [,1]  
## [1,] 1.341734e-12

#We can reject the null, conclude there is no differenece in yield between fertilisers 2 and 3.

#Actual Solution  
library(Matrix)  
C = matrix(c(0,0,1,-1),1,4)  
m = rankMatrix(C)[1]  
num = t(C%\*%b)%\*%solve(C%\*%ginv(t(X)%\*%X)%\*%t(C))%\*%C%\*%b  
Fstat = (num/m)/s2  
Fstat

## [,1]  
## [1,] 178.8633

pf(Fstat,m,n-rankMatrix(X)[1], lower = F)

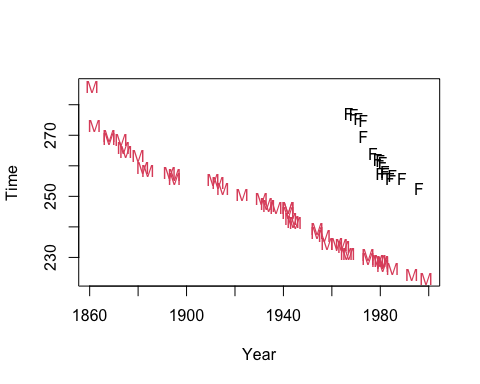
## [,1]  
## [1,] 3.042802e-07

#Question 4 #Part 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?

#Setting the directory  
setwd("~/Desktop/UNIMELB 2021 Material/UNIMELB S1 2021 (Currently)/MAST30025/Tutorials /Tutorials/Rfile/data")  
mile = read.csv("mile.csv")  
str(mile)

## 'data.frame': 62 obs. of 6 variables:  
## $ Year : int 1861 1862 1868 1868 1873 1874 1875 1880 1882 1884 ...  
## $ Time : num 286 273 270 269 269 ...  
## $ Name : chr "N.S.,Greene" "George,Farran" "Walter,Chinnery" "William,Gibbs" ...  
## $ Country: chr "IRL" "IRL" "GBR" "GBR" ...  
## $ Place : chr NA NA NA NA ...  
## $ Gender : chr "Male" "Male" "Male" "Male" ...

#Ensure you convert gender into your Factor Type.   
genderfactor = factor(mile$Gender)  
plot(Time ~ Year, pch=as.character(genderfactor), col=genderfactor, data=mile)



#The data looks linear, but its decreasing as the years gone by. #They are not independant and does not satisfy the linear model assumptions.

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

#Attempt 1

#Remember we still dealing with the additive model with interaction!  
library(MASS)  
library(Matrix)  
y = mile$Time  
n = length(mile$Time)  
X = matrix(c(rep(1,n),rep(0,n\*5)),n,6)  
X[cbind(1:n,as.numeric(genderfactor)+1)]=1  
X[cbind(1:n,as.numeric(genderfactor)+3)]=1  
#Trickest part since Year in this case we do not account them as integers  
X

## [,1] [,2] [,3] [,4] [,5] [,6]  
## [1,] 1 0 1 0 1 0  
## [2,] 1 0 1 0 1 0  
## [3,] 1 0 1 0 1 0  
## [4,] 1 0 1 0 1 0  
## [5,] 1 0 1 0 1 0  
## [6,] 1 0 1 0 1 0  
## [7,] 1 0 1 0 1 0  
## [8,] 1 0 1 0 1 0  
## [9,] 1 0 1 0 1 0  
## [10,] 1 0 1 0 1 0  
## [11,] 1 0 1 0 1 0  
## [12,] 1 0 1 0 1 0  
## [13,] 1 0 1 0 1 0  
## [14,] 1 0 1 0 1 0  
## [15,] 1 0 1 0 1 0  
## [16,] 1 0 1 0 1 0  
## [17,] 1 0 1 0 1 0  
## [18,] 1 0 1 0 1 0  
## [19,] 1 0 1 0 1 0  
## [20,] 1 0 1 0 1 0  
## [21,] 1 0 1 0 1 0  
## [22,] 1 0 1 0 1 0  
## [23,] 1 0 1 0 1 0  
## [24,] 1 0 1 0 1 0  
## [25,] 1 0 1 0 1 0  
## [26,] 1 0 1 0 1 0  
## [27,] 1 0 1 0 1 0  
## [28,] 1 0 1 0 1 0  
## [29,] 1 0 1 0 1 0  
## [30,] 1 0 1 0 1 0  
## [31,] 1 0 1 0 1 0  
## [32,] 1 0 1 0 1 0  
## [33,] 1 0 1 0 1 0  
## [34,] 1 0 1 0 1 0  
## [35,] 1 0 1 0 1 0  
## [36,] 1 0 1 0 1 0  
## [37,] 1 0 1 0 1 0  
## [38,] 1 0 1 0 1 0  
## [39,] 1 0 1 0 1 0  
## [40,] 1 0 1 0 1 0  
## [41,] 1 0 1 0 1 0  
## [42,] 1 0 1 0 1 0  
## [43,] 1 0 1 0 1 0  
## [44,] 1 0 1 0 1 0  
## [45,] 1 0 1 0 1 0  
## [46,] 1 0 1 0 1 0  
## [47,] 1 1 0 1 0 0  
## [48,] 1 1 0 1 0 0  
## [49,] 1 1 0 1 0 0  
## [50,] 1 1 0 1 0 0  
## [51,] 1 1 0 1 0 0  
## [52,] 1 1 0 1 0 0  
## [53,] 1 1 0 1 0 0  
## [54,] 1 1 0 1 0 0  
## [55,] 1 1 0 1 0 0  
## [56,] 1 1 0 1 0 0  
## [57,] 1 1 0 1 0 0  
## [58,] 1 1 0 1 0 0  
## [59,] 1 1 0 1 0 0  
## [60,] 1 1 0 1 0 0  
## [61,] 1 1 0 1 0 0  
## [62,] 1 1 0 1 0 0

r = rankMatrix(X)[1]  
XtXc = ginv(t(X)%\*%X)  
b = XtXc%\*%t(X)%\*%y  
b

## [,1]  
## [1,] 127.30172  
## [2,] 68.09008  
## [3,] 59.21164  
## [4,] 68.09008  
## [5,] 59.21164  
## [6,] 0.00000

s2 = sum((y-X%\*%b)^2)/(n-r)  
C = matrix(c(0,0,0,1,-1,0,0,0,0,1,0,-1),2,6,byrow = T)  
Fstat = t(b)%\*%t(C)%\*%solve(C%\*%XtXc%\*%t(C))%\*%C%\*%b/2/s2  
Fstat

## [,1]  
## [1,] 10043.3

pf(Fstat,2,n-r,lower=F)

## [,1]  
## [1,] 1.653802e-76

#Using the lm

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

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

#We can reject the H0 and conclude that there is a significant interaction between two predict variables.

#Part 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 = lm(Time~Year\*Gender, data = mile)  
imodel$coefficients

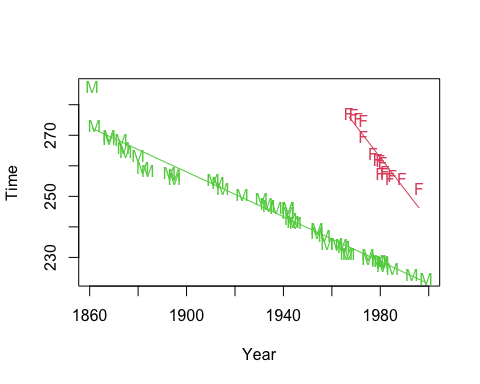
## (Intercept) Year GenderMale Year:GenderMale   
## 2309.4247477 -1.0336960 -1355.6777866 0.6675093

#Actual solution

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

## (Intercept) Year   
## 953.7469611 -0.3661867

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



#One for Females #Time = 2309 -1.034\*Year

#One for Males #Time = 954 - 0.366\*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

#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.

#Part 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)?

#This is not an estimable quantity; it is not 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.

#Part 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)[4,]

## 2.5 % 97.5 %   
## 0.4620087 0.8730100

#Part g. Test the hypothesis that the male world record decreases by 0.3 seconds each year.

library(car)

## Loading required package: carData

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

#Question 5 #Part b: Perform the random allocation. You must use R for randomisation and include your R com- mands and output.

x = sample(50,50)  
x[1:9]

## [1] 50 11 32 41 44 10 9 23 40

x[10:23]

## [1] 18 21 12 2 42 15 13 45 16 36 25 19 1 46

x[24:50]

## [1] 49 24 3 7 14 20 26 6 43 48 31 4 27 22 30 5 39 47 17 29 33 28 35 38 8  
## [26] 34 37