MAST30025 Assignment 2 2019 R Questions

#Question 2. An experiment is conducted to estimate the annual demand for cars, based on their cost, the current unemployment rate, and the current interest rate. A survey is conducted and the following measurements obtained.

#Part a) Fit a linear model to the data and estimate the parameters and variance

y = c(5.5,5.9,6.5,5.9,8,9,10,10.8)  
X = matrix(c(rep(1,8),7.2,10,9,5.5,9,9.8,14.5,8,8.7,9.4,10,9,12,11,12,13.7,5.5,4.4,4,7,5,6.2,5.8,3.9),8,4)  
b = solve(t(X)%\*%X,t(X)%\*%y)  
b #estimates of the parameters

## [,1]  
## [1,] -7.4044796  
## [2,] 0.1207646  
## [3,] 1.1174846  
## [4,] 0.3861206

#beta parameters

n=8  
p=4  
e = y - X%\*%b  
SSRes = sum(e^2)  
s2 = SSRes/(n-p)  
s2 #Sample variance

## [1] 0.3955368

#Part b) Which two of the parameters have the highest (in magnitude) covariance in their estimators?

#The unemployment rate and the Interest rate.   
#Actual answer!!  
#The intercept term and the Interest rate!(Forgot to account the intercept)  
C = solve(t(X)%\*%X)  
C

## [,1] [,2] [,3] [,4]  
## [1,] 13.49743324 -0.054817613 -0.69854293 -1.029731987  
## [2,] -0.05481761 0.024498395 -0.01478859 -0.001937333  
## [3,] -0.69854293 -0.014788594 0.06226378 0.031714790  
## [4,] -1.02973199 -0.001937333 0.03171479 0.135362495

#Part c) Find a 99% confidence interval for the avarage number of $8,000 cars sold in a year which has unemployment rate 9% and interest rate 5%

xst = c(1,8,9,5)  
s = sqrt(s2)  
s

## [1] 0.6289172

xst%\*%b

## [,1]  
## [1,] 5.549602

#99% t quantile

ta = qt(0.995,df=8-4)  
ta

## [1] 4.604095

#Lower Bound

xst%\*%b - ta\*s\*sqrt(t(xst)%\*%solve(t(X)%\*%X)%\*%xst)

## [,1]  
## [1,] 3.926075

#Upper Bound

xst%\*%b + ta\*s\*sqrt(t(xst)%\*%solve(t(X)%\*%X)%\*%xst)

## [,1]  
## [1,] 7.173129

#Part d) A prediction interval for the number of cars sold in such a year is calculated to be (4012,7087). Find the confidence level used!

LB = 4.012  
UB = 7.087  
xst%\*%b

## [,1]  
## [1,] 5.549602

s\*sqrt(1 + t(xst)%\*%solve(t(X)%\*%X)%\*%xst)

## [,1]  
## [1,] 0.7210287

#To help us solve for ta

ta = 2.1325

#guessing the confidence interval

ta = qt(0.95,df=8-4)  
ta

## [1] 2.131847

#Can do for the upper bound gives us the same result  
#alpha level is 0.10   
#It is a 90% Confidence interval!! QED!!

#Part e) Test for model relevance using a corrected sum of squares.

SSTotal = sum(y^2)  
SSReg = SSTotal - SSRes  
SSReg

## [1] 502.1779

SSRes

## [1] 1.582147

#There is a strong linear signal in the data.

MSReg = SSReg/4  
MSRes = SSRes/(8-4)  
Fstat = MSReg/MSRes  
Fstat

## [1] 317.4027

qf(0.95,8,4)

## [1] 6.041044

pf(Fstat,4,4,lower.tail = FALSE)

## [1] 2.952959e-05

#We test for model relevance when H0: beta = 0 (We reject the null hypothesis for model irrelvance)!!

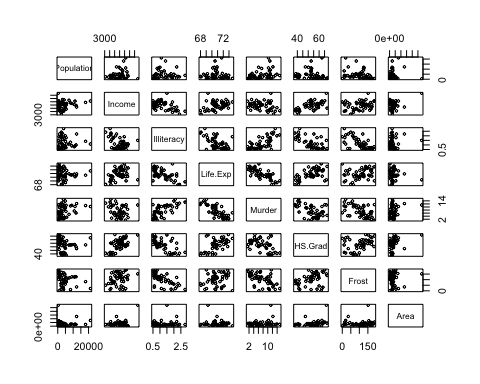
#Question 4. In this question, we study a dataset of 50 US states. This dataset contains the variables: Population: population estimate as of July 1, 1975� Income: per capita income (1974) Illiteracy: illiteracy (1970, percent of population) Life.Exp: life expectancy in years (1969–71) Murder: murder and non-negligent manslaughter rate per 100,000 population (1976) HS.Grad: percentage of high-school graduates (1970) Frost: mean number of days with minimum temperature below freezing (1931–1960) in capital or large city Area: land area in square miles #The dataset is distributed with R. Open it with the following commands:

data(state)  
statedata = data.frame(state.x77, row.names=state.abb, check.names=TRUE)

#We wish to use a linear model to model the murder rate in terms of the other variables.

#Part a) Plot the data and comment. Should we consider any variable transformations?

pairs(statedata,cex = 0.5)



#Actual  
statedata$logPopulation <- log(statedata$Population)  
statedata$logArea <- log(statedata$Area)

#Looking at murder rate against the other variables, there is evidence of a linear relationship with income, illiteracy, life expectancy, high school grad and frost. There is no obvious relationship with population and area. #Population and area both have distributions heavily skewed to the right. log(population) and log(area) would be less skewed and might fit better with the other variables. #There is potential heteroskedasticity in high school grad, and non-linearity in illiteracy, but neither enough for immediate concern.

NOTE: Key things to understand this part is to observe any non linear relationships between the two variables, and check for any heteroskedasticity!

#Part b) Perform model selection using forward selection, using all the variable transformations which may be relevant.

basemodel = lm(statedata$Murder~1, data = statedata)  
add1(basemodel, scope = ~ .+statedata$Population + statedata$Income+statedata$Illiteracy+statedata$Life.Exp+statedata$HS.Grad+statedata$Frost+statedata$Area, test = "F")

## Single term additions  
##   
## Model:  
## statedata$Murder ~ 1  
## Df Sum of Sq RSS AIC F value Pr(>F)   
## <none> 667.75 131.594   
## statedata$Population 1 78.85 588.89 127.311 6.4273 0.0145504 \*   
## statedata$Income 1 35.35 632.40 130.875 2.6829 0.1079683   
## statedata$Illiteracy 1 329.98 337.76 99.516 46.8943 1.258e-08 \*\*\*  
## statedata$Life.Exp 1 407.14 260.61 86.550 74.9887 2.260e-11 \*\*\*  
## statedata$HS.Grad 1 159.00 508.75 119.996 15.0017 0.0003248 \*\*\*  
## statedata$Frost 1 193.91 473.84 116.442 19.6433 5.405e-05 \*\*\*  
## statedata$Area 1 34.83 632.91 130.916 2.6416 0.1106495   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#The highest F value is Life.Exp

model2 = lm(statedata$Murder~statedata$Life.Exp, data = statedata)  
add1(model2, scope = ~ .+statedata$Population + statedata$Income+statedata$Illiteracy+statedata$HS.Grad+statedata$Frost+statedata$Area, test = "F")

## Single term additions  
##   
## Model:  
## statedata$Murder ~ statedata$Life.Exp  
## Df Sum of Sq RSS AIC F value Pr(>F)   
## <none> 260.61 86.550   
## statedata$Population 1 56.615 203.99 76.303 13.0442 0.0007374 \*\*\*  
## statedata$Income 1 0.958 259.65 88.366 0.1733 0.6790605   
## statedata$Illiteracy 1 60.549 200.06 75.329 14.2249 0.0004533 \*\*\*  
## statedata$HS.Grad 1 1.124 259.48 88.334 0.2035 0.6539823   
## statedata$Frost 1 80.104 180.50 70.187 20.8575 3.576e-05 \*\*\*  
## statedata$Area 1 14.121 246.49 85.764 2.6926 0.1074933   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Now Frost has the highest F value

model3 = lm(statedata$Murder~statedata$Life.Exp + statedata$Frost, data = statedata)  
add1(model3, scope = ~ .+statedata$Population + statedata$Income+statedata$Illiteracy+statedata$HS.Grad+statedata$Area, test = "F")

## Single term additions  
##   
## Model:  
## statedata$Murder ~ statedata$Life.Exp + statedata$Frost  
## Df Sum of Sq RSS AIC F value Pr(>F)   
## <none> 180.50 70.187   
## statedata$Population 1 23.7098 156.79 65.146 6.9559 0.01136 \*  
## statedata$Income 1 5.5598 174.94 70.622 1.4619 0.23281   
## statedata$Illiteracy 1 6.0663 174.44 70.477 1.5997 0.21231   
## statedata$HS.Grad 1 2.0679 178.44 71.610 0.5331 0.46901   
## statedata$Area 1 21.0840 159.42 65.976 6.0837 0.01743 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Population

model4 = lm(statedata$Murder~statedata$Life.Exp + statedata$Frost+statedata$Population, data = statedata)  
add1(model4, scope = ~ .+statedata$Income+statedata$Illiteracy+statedata$HS.Grad+statedata$Area, test = "F")

## Single term additions  
##   
## Model:  
## statedata$Murder ~ statedata$Life.Exp + statedata$Frost + statedata$Population  
## Df Sum of Sq RSS AIC F value Pr(>F)   
## <none> 156.79 65.146   
## statedata$Income 1 0.7393 156.06 66.909 0.2132 0.64650   
## statedata$Illiteracy 1 11.8262 144.97 63.225 3.6710 0.06173 .  
## statedata$HS.Grad 1 1.8215 154.97 66.561 0.5289 0.47083   
## statedata$Area 1 19.0402 137.75 60.672 6.2198 0.01637 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Area

model5 = lm(statedata$Murder~statedata$Life.Exp + statedata$Frost+statedata$Population+statedata$Area, data = statedata)  
add1(model5, scope = ~ .+statedata$Income+statedata$Illiteracy+statedata$HS.Grad, test = "F")

## Single term additions  
##   
## Model:  
## statedata$Murder ~ statedata$Life.Exp + statedata$Frost + statedata$Population +   
## statedata$Area  
## Df Sum of Sq RSS AIC F value Pr(>F)   
## <none> 137.75 60.672   
## statedata$Income 1 1.2408 136.51 62.220 0.3999 0.53040   
## statedata$Illiteracy 1 8.7227 129.03 59.402 2.9745 0.09161 .  
## statedata$HS.Grad 1 0.7708 136.98 62.392 0.2476 0.62126   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Our final model will be we use Population,Life Expectancy,Frost and Area #murder = beta0 + beta1*Population + beta4*Life Expectancy + beta6*Frost+beta7*Area

#Actual Answer!!! #Forgot to add in log(Population) & log(Area) you need to add in from part (a)!! Ensure you observe for any non linearities relationships requires any transformation. To ensure it is the ’BEST FIT FOR THE MODEL".

basemodel = lm(statedata$Murder~1, data = statedata)  
add1(basemodel, scope = ~ .+statedata$Population + statedata$Income+statedata$Illiteracy+statedata$Life.Exp+statedata$HS.Grad+statedata$Frost+statedata$Area+statedata$logArea+statedata$logPopulation, test = "F")

## Single term additions  
##   
## Model:  
## statedata$Murder ~ 1  
## Df Sum of Sq RSS AIC F value Pr(>F)   
## <none> 667.75 131.594   
## statedata$Population 1 78.85 588.89 127.311 6.4273 0.0145504 \*   
## statedata$Income 1 35.35 632.40 130.875 2.6829 0.1079683   
## statedata$Illiteracy 1 329.98 337.76 99.516 46.8943 1.258e-08 \*\*\*  
## statedata$Life.Exp 1 407.14 260.61 86.550 74.9887 2.260e-11 \*\*\*  
## statedata$HS.Grad 1 159.00 508.75 119.996 15.0017 0.0003248 \*\*\*  
## statedata$Frost 1 193.91 473.84 116.442 19.6433 5.405e-05 \*\*\*  
## statedata$Area 1 34.83 632.91 130.916 2.6416 0.1106495   
## statedata$logArea 1 58.63 609.12 128.999 4.6201 0.0366687 \*   
## statedata$logPopulation 1 86.37 581.37 126.668 7.1313 0.0103090 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Wait wouldn’t just give us the same result (just to add in the other design variables to create a BEST FIT?!) –>getting more marks?! #Our final model will be we use Population,Life Expectancy,Frost, log(Area) and Illiteracy

#Part c) Starting from the full model, perform model selection using stepwise selection with the AIC.

fullmodel = lm(statedata$Murder~1, data = statedata)  
model2 = step(fullmodel,scope = ~ .+statedata$Population + statedata$Income+statedata$Illiteracy+statedata$Life.Exp+statedata$HS.Grad+statedata$Frost+statedata$Area+statedata$logArea+statedata$logPopulation)

## Start: AIC=131.59  
## statedata$Murder ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + statedata$Life.Exp 1 407.14 260.61 86.550  
## + statedata$Illiteracy 1 329.98 337.76 99.516  
## + statedata$Frost 1 193.91 473.84 116.442  
## + statedata$HS.Grad 1 159.00 508.75 119.996  
## + statedata$logPopulation 1 86.37 581.37 126.668  
## + statedata$Population 1 78.85 588.89 127.311  
## + statedata$logArea 1 58.63 609.12 128.999  
## + statedata$Income 1 35.35 632.40 130.875  
## + statedata$Area 1 34.83 632.91 130.916  
## <none> 667.75 131.594  
##   
## Step: AIC=86.55  
## statedata$Murder ~ statedata$Life.Exp  
##   
## Df Sum of Sq RSS AIC  
## + statedata$Frost 1 80.10 180.50 70.187  
## + statedata$Illiteracy 1 60.55 200.06 75.329  
## + statedata$Population 1 56.62 203.99 76.303  
## + statedata$logPopulation 1 50.86 209.75 77.694  
## + statedata$logArea 1 30.22 230.39 82.386  
## + statedata$Area 1 14.12 246.49 85.764  
## <none> 260.61 86.550  
## + statedata$HS.Grad 1 1.12 259.48 88.334  
## + statedata$Income 1 0.96 259.65 88.366  
## - statedata$Life.Exp 1 407.14 667.75 131.594  
##   
## Step: AIC=70.19  
## statedata$Murder ~ statedata$Life.Exp + statedata$Frost  
##   
## Df Sum of Sq RSS AIC  
## + statedata$logArea 1 30.973 149.53 62.774  
## + statedata$Population 1 23.710 156.79 65.146  
## + statedata$Area 1 21.084 159.42 65.976  
## + statedata$logPopulation 1 12.213 168.29 68.684  
## <none> 180.50 70.187  
## + statedata$Illiteracy 1 6.066 174.44 70.477  
## + statedata$Income 1 5.560 174.94 70.622  
## + statedata$HS.Grad 1 2.068 178.44 71.610  
## - statedata$Frost 1 80.104 260.61 86.550  
## - statedata$Life.Exp 1 293.331 473.84 116.442  
##   
## Step: AIC=62.77  
## statedata$Murder ~ statedata$Life.Exp + statedata$Frost + statedata$logArea  
##   
## Df Sum of Sq RSS AIC  
## + statedata$Population 1 16.347 133.18 58.985  
## + statedata$logPopulation 1 9.131 140.40 61.623  
## + statedata$Illiteracy 1 8.737 140.79 61.764  
## <none> 149.53 62.774  
## + statedata$Income 1 4.786 144.74 63.147  
## + statedata$Area 1 1.239 148.29 64.358  
## + statedata$HS.Grad 1 0.190 149.34 64.710  
## - statedata$logArea 1 30.973 180.50 70.187  
## - statedata$Frost 1 80.854 230.39 82.386  
## - statedata$Life.Exp 1 270.087 419.62 112.366  
##   
## Step: AIC=58.99  
## statedata$Murder ~ statedata$Life.Exp + statedata$Frost + statedata$logArea +   
## statedata$Population  
##   
## Df Sum of Sq RSS AIC  
## + statedata$Illiteracy 1 13.919 119.26 55.466  
## <none> 133.18 58.985  
## + statedata$Area 1 2.091 131.09 60.194  
## + statedata$Income 1 0.920 132.26 60.639  
## + statedata$logPopulation 1 0.523 132.66 60.789  
## + statedata$HS.Grad 1 0.083 133.10 60.954  
## - statedata$Population 1 16.347 149.53 62.774  
## - statedata$logArea 1 23.611 156.79 65.146  
## - statedata$Frost 1 51.061 184.24 73.212  
## - statedata$Life.Exp 1 274.770 407.95 112.957  
##   
## Step: AIC=55.47  
## statedata$Murder ~ statedata$Life.Exp + statedata$Frost + statedata$logArea +   
## statedata$Population + statedata$Illiteracy  
##   
## Df Sum of Sq RSS AIC  
## <none> 119.26 55.466  
## + statedata$Income 1 3.724 115.54 55.880  
## - statedata$Frost 1 7.639 126.90 56.570  
## + statedata$HS.Grad 1 2.022 117.24 56.611  
## + statedata$logPopulation 1 0.463 118.80 57.272  
## + statedata$Area 1 0.446 118.82 57.279  
## - statedata$Illiteracy 1 13.919 133.18 58.985  
## - statedata$Population 1 21.529 140.79 61.764  
## - statedata$logArea 1 25.704 144.97 63.225  
## - statedata$Life.Exp 1 127.359 246.62 89.792

#Actual Answer:The model is the same from the forward selection!

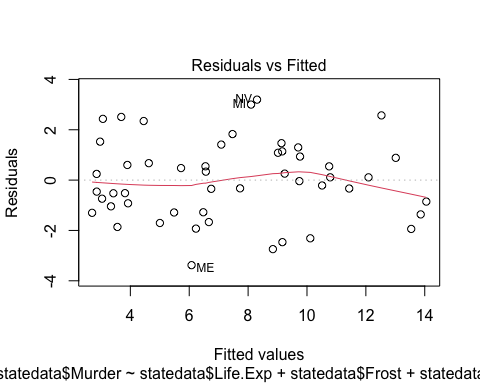
lm(formula = statedata$Murder ~ statedata$Frost + statedata$Illiteracy + statedata$Population+statedata$logArea +statedata$Life.Exp, data = statedata)

##   
## Call:  
## lm(formula = statedata$Murder ~ statedata$Frost + statedata$Illiteracy +   
## statedata$Population + statedata$logArea + statedata$Life.Exp,   
## data = statedata)  
##   
## Coefficients:  
## (Intercept) statedata$Frost statedata$Illiteracy   
## 108.713249 -0.011293 1.474305   
## statedata$Population statedata$logArea statedata$Life.Exp   
## 0.000162 0.632740 -1.542284

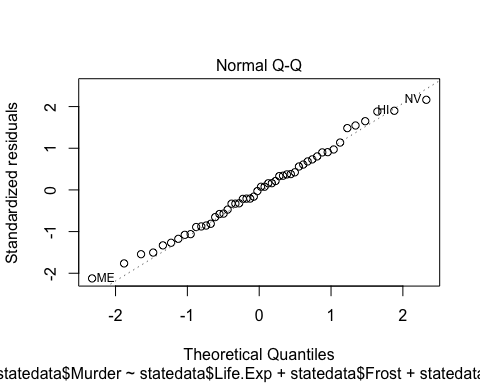
#Part d) Write down your final fitted model (including any variable transformations used). #murder = beta0 + beta1*Population + beta3*Illiteracy + beta4*Life Expectancy + beta6*Frost+beta8*log(Area) #murder = 108.713249 + 0.000162*Population + 1.474305 *Illiteracy -1.542284*Life Expectancy -0.011293*Frost + 0.632740*log(Area)

#Part e) Produce diagnostic plots for your final model and comment.

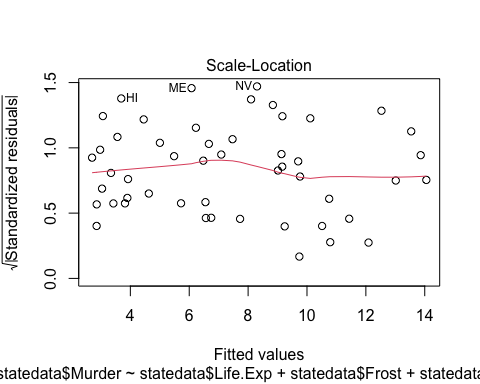
plot(model2, which = 1)



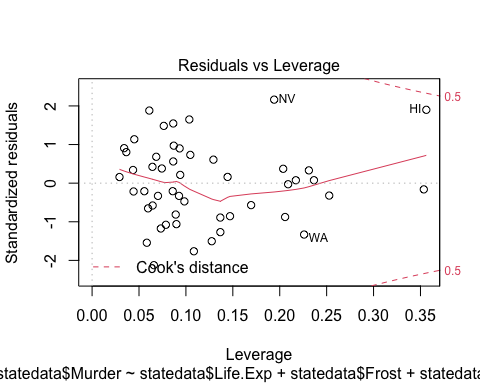
plot(model2, which = 2)



plot(model2, which = 3)



plot(model2, which = 5)



#Demonstator comment: It looks like a linear normal fit (or best fit)! There is a slight negative trend for higher fitted values and moderate leverages.

#Question 5b) Calculate the ridge regression estimates for the data from Q2 with penalty parameter λ = 0.5. In order to avoid penalising some parameters unfairly, we must first scale every predictor variable so that it is standardised (mean 0, variance 1), and centre the response variable (mean 0), in which case an intercept parameter is not used. (Hint: This can be done with the scale function).

Xs <- scale(X[,-1],center=T,scale=T)  
ys <- scale(y,center=T,scale=F)  
p = 4  
p <- p-1  
solve(t(Xs)%\*%Xs + diag(rep(0.5,p)),t(Xs)%\*%ys)

## [,1]  
## [1,] 0.3494789  
## [2,] 1.7899861  
## [3,] 0.3432961

#Question 5c) One way to calculate the optimal value for the penalty parameter is to minimise the AIC. Since the number of parameters p does not change, we use a slightly modified version:

AIC=n\*ln(SSRes/n) +2df,

#where df is the “effective degrees of freedom” defined by

df = tr(H) = tr(X(XT X + λI)^−1\*XT).

#For the data from Q2, construct a plot of λ against AIC. Thereby find the optimal value for λ.

n = 8  
lambda <- seq(0,1,0.001)  
aic <- c()  
for (l in lambda) {  
 b <- solve(t(Xs)%\*%Xs + diag(rep(l,p)),t(Xs)%\*%ys)  
 ssres <- sum((ys-Xs%\*%b)^2)  
 H <- Xs %\*% solve(t(Xs)%\*%Xs + diag(rep(l,p))) %\*% t(Xs)  
 aic <- c(aic, n\*log(ssres/n) + 2\*sum(diag(H)))  
}  
lambda[which.min(aic)]

## [1] 0.136

plot(lambda,aic,type='l')

