final test

Mehrab Atighi

6/20/2021

set.seed(1)  
library("ISLR")

## Warning: package 'ISLR' was built under R version 4.0.5

data("swiss")  
attach(swiss)  
head(swiss)

## Fertility Agriculture Examination Education Catholic  
## Courtelary 80.2 17.0 15 12 9.96  
## Delemont 83.1 45.1 6 9 84.84  
## Franches-Mnt 92.5 39.7 5 5 93.40  
## Moutier 85.8 36.5 12 7 33.77  
## Neuveville 76.9 43.5 17 15 5.16  
## Porrentruy 76.1 35.3 9 7 90.57  
## Infant.Mortality  
## Courtelary 22.2  
## Delemont 22.2  
## Franches-Mnt 20.2  
## Moutier 20.3  
## Neuveville 20.6  
## Porrentruy 26.6

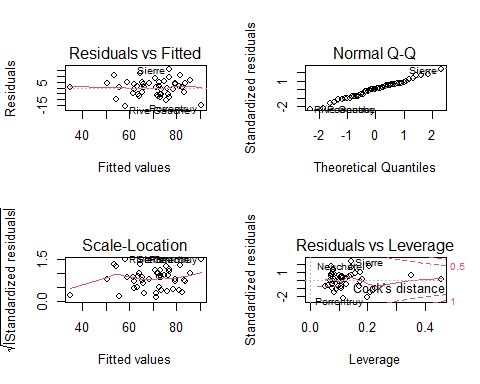
# a)

## for make Full model and plot we have:

fit<-lm(Fertility~. , data=swiss)  
summary(fit)

##   
## Call:  
## lm(formula = Fertility ~ ., data = swiss)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.2743 -5.2617 0.5032 4.1198 15.3213   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 66.91518 10.70604 6.250 1.91e-07 \*\*\*  
## Agriculture -0.17211 0.07030 -2.448 0.01873 \*   
## Examination -0.25801 0.25388 -1.016 0.31546   
## Education -0.87094 0.18303 -4.758 2.43e-05 \*\*\*  
## Catholic 0.10412 0.03526 2.953 0.00519 \*\*   
## Infant.Mortality 1.07705 0.38172 2.822 0.00734 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.165 on 41 degrees of freedom  
## Multiple R-squared: 0.7067, Adjusted R-squared: 0.671   
## F-statistic: 19.76 on 5 and 41 DF, p-value: 5.594e-10

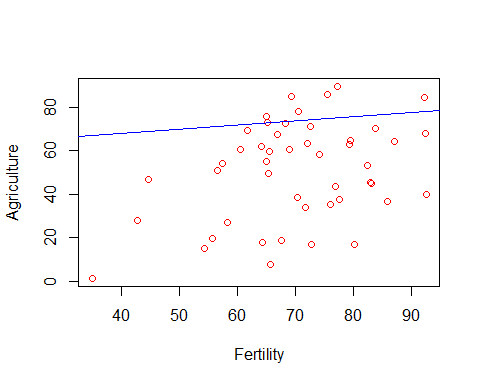
par(mfrow=c(2,2))  
plot(fit)



we can see that just Examination and Agriculture variables (predictors) are not signifact and anothers are becuse they have a p-values less than 0.05(alpha). the adjust R sqrue is 0.67 and R squre is 0.7 and our standard error is 7.165 and Agriculture , Examination , Education predictors have negative relationships with our response.

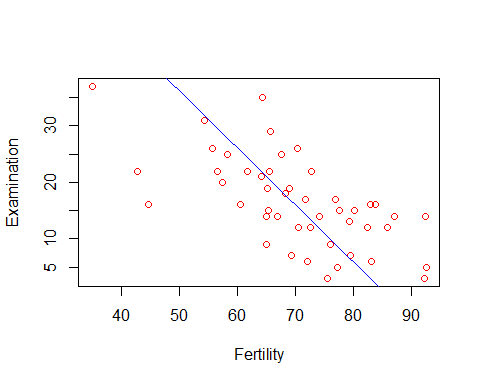
## for make plot of each varibale with response we have:

par(mfrow=c(1,1))  
plot(Fertility,Agriculture , col="red")  
abline(lm(Fertility~Agriculture), col="Blue")



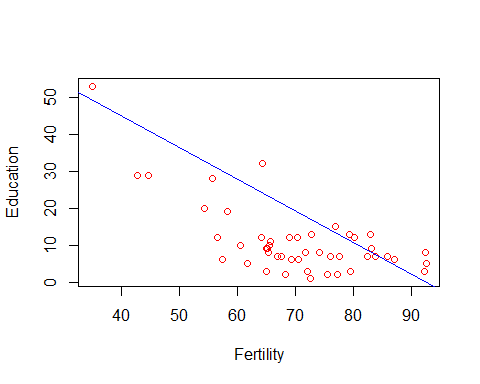
According to this plot we see the approxmiatly posetive realationships and we have 1 High leverage point (the left and down side of plot).

plot(Fertility,Examination , col="red")  
abline(lm(Fertility~Examination), col="Blue")



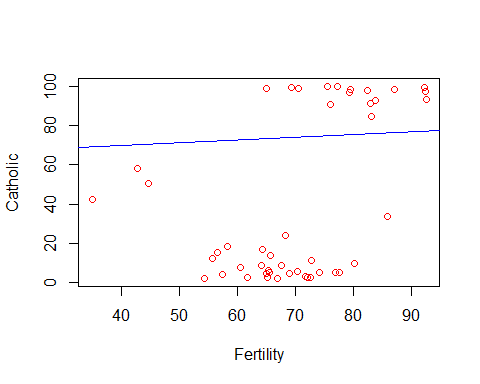
now we can see the negative realationships between response and this predictor and we have 3 High leverage points (the left side).

plot(Fertility,Education , col="red")  
abline(lm(Fertility~Education), col="Blue")



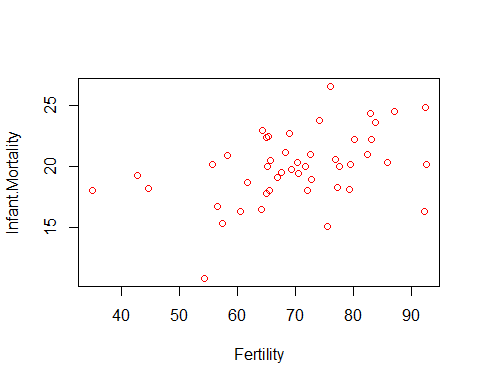
now we can see the negative realationships between response and this predictor and we have 3 High leverage points (the left side).

plot(Fertility,Catholic , col="red")  
abline(lm(Fertility~Catholic), col="Blue")



we can see the posetive realationships and 3High leverage points (the left side).

plot(Fertility,Infant.Mortality , col="red")  
abline(lm(Fertility~Infant.Mortality), col="Blue")



here we can see the posetive realationships with one outlires and high leverage point.

## for make covariance matrix of variables we have:

cov(swiss)

## Fertility Agriculture Examination Education Catholic  
## Fertility 156.04250 100.169149 -64.366929 -79.729510 241.56320  
## Agriculture 100.16915 515.799417 -124.392831 -139.657401 379.90438  
## Examination -64.36693 -124.392831 63.646623 53.575856 -190.56061  
## Education -79.72951 -139.657401 53.575856 92.456059 -61.69883  
## Catholic 241.56320 379.904376 -190.560611 -61.698830 1739.29454  
## Infant.Mortality 15.15619 -4.025851 -2.649537 -2.781684 21.31812  
## Infant.Mortality  
## Fertility 15.156193  
## Agriculture -4.025851  
## Examination -2.649537  
## Education -2.781684  
## Catholic 21.318116  
## Infant.Mortality 8.483802

according to this out put we see all of the covariance between each variables.

# b)

## best subset selection:

at the first we need to library this package:

library(leaps)

## Warning: package 'leaps' was built under R version 4.0.5

Now we make our model from swiss data:

best.subset.fit1<-regsubsets(Fertility~. , data = swiss , nvmax = 19)  
summary(best.subset.fit1)

## Subset selection object  
## Call: regsubsets.formula(Fertility ~ ., data = swiss, nvmax = 19)  
## 5 Variables (and intercept)  
## Forced in Forced out  
## Agriculture FALSE FALSE  
## Examination FALSE FALSE  
## Education FALSE FALSE  
## Catholic FALSE FALSE  
## Infant.Mortality FALSE FALSE  
## 1 subsets of each size up to 5  
## Selection Algorithm: exhaustive  
## Agriculture Examination Education Catholic Infant.Mortality  
## 1 ( 1 ) " " " " "\*" " " " "   
## 2 ( 1 ) " " " " "\*" "\*" " "   
## 3 ( 1 ) " " " " "\*" "\*" "\*"   
## 4 ( 1 ) "\*" " " "\*" "\*" "\*"   
## 5 ( 1 ) "\*" "\*" "\*" "\*" "\*"

now here we can see that in the first step we have select M1 model or a model with one predictors from all of the models with one predictors and we can see that this method choose the Education predictors from these models. like this for the model with 2 predictors we choose model with M1 + Catholic=M2 predictors from all of them. like this for the model with 3 predictors we choose model with M2+Infant.Mortality =M3 predictors from all of them. like this for the model with 4 preditos we choose model with M3+Agriculture =M4 and at the last for model with all of predictors we have just 1 model or full model with all of the predictors M4 +Examinarion=M5

# backward subset selection:

backward.fit1<-regsubsets(Fertility~.,data=swiss ,nvmax= 19 , method = "backward")  
summary(backward.fit1)

## Subset selection object  
## Call: regsubsets.formula(Fertility ~ ., data = swiss, nvmax = 19, method = "backward")  
## 5 Variables (and intercept)  
## Forced in Forced out  
## Agriculture FALSE FALSE  
## Examination FALSE FALSE  
## Education FALSE FALSE  
## Catholic FALSE FALSE  
## Infant.Mortality FALSE FALSE  
## 1 subsets of each size up to 5  
## Selection Algorithm: backward  
## Agriculture Examination Education Catholic Infant.Mortality  
## 1 ( 1 ) " " " " "\*" " " " "   
## 2 ( 1 ) " " " " "\*" "\*" " "   
## 3 ( 1 ) " " " " "\*" "\*" "\*"   
## 4 ( 1 ) "\*" " " "\*" "\*" "\*"   
## 5 ( 1 ) "\*" "\*" "\*" "\*" "\*"

here at the first we have compelte model(M0) with all of the predictors in the M1 we just delete the Education preditor from our M0 it means that M1=M0-Education in the next step we have: M2= M1-Catholic the next step we have: M3= M2-Infant.mortality the next step we have: M4 = M3 - Agriculture the M5 is the model with 0 predictor. M5= M4-Examination

# c)

Now we want to make a training and test data with probabilty (0.7 , 0.3 ) and again chek models with bic:

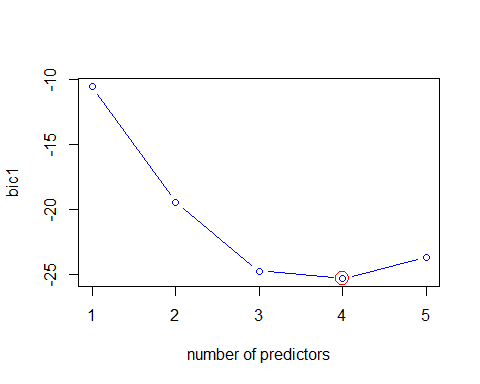
sample<-sample(c(TRUE , FALSE ) , nrow(swiss) , replace = T , prob=c(0.7,0.3))  
train<-swiss[sample,]  
test<-swiss[!sample,]

Now again we make model with best subset selection and backward method here:

best.subset.fit2<-regsubsets(Fertility~. , data = train , nvmax = 19)  
backward.fit2<-regsubsets(Fertility~.,data= train ,nvmax= 19 , method = "backward")

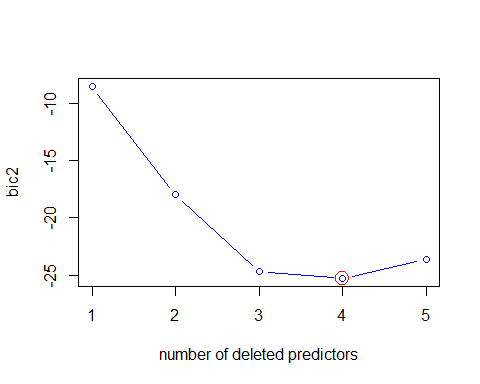
now we want to calculate and see the bic of each method here:

bic1<-summary(best.subset.fit2)$bic  
plot(bic1 , type="b" , col="Blue" , xlab="number of predictors")  
points(4,bic1[which.min(bic1)] ,cex=2 , col="red")



according to this plot we choose the model with 4 predictors that have the minimum of the BIC.

bic2<-summary(backward.fit2)$bic  
plot(bic2 , type="b" , col="Blue",xlab="number of deleted predictors")  
points(4,bic2[which.min(bic1)] ,cex=2 , col="red")



according to this plot we choose the model with 1 predictor that have the minimum of the BIC.attention:backward method.

End. by: Mehrab Atighi