Midterm Question 1: 2019

Question 1: http://www.statsci.org/data/general/uscrime.html

Criminologists are interested in the effect of punishment regimes on crime rates. This has been studied using aggregate data on 47 states of the USA for 1960. The data set **USCrime.csv** contains the following columns:

Variable Description M percentage of males aged 14–24 in total state population So indicator variable for a southern state
Ed mean years of schooling of the population aged 25 years or over
Po1 per capita expenditure on police protection in 1960
Po2 per capita expenditure on police protection in 1959 LF labour force participation rate of civilian urban males in the age-group 14-24 M.F number of males per 100 females Pop state population in 1960 in hundred thousands NW percentage of nonwhites in the population U1 unemployment rate of urban males 14???24 U2 unemployment rate of urban males 35???39 Wealth wealth: median value of transferable assets or family income Ineq income inequality: percentage of families earning below half the median income Prob probability of imprisonment: ratio of number of commitments to number of offenses Time average time in months served by offenders in state prisons before their first release Crime crime rate: number of offenses per 100,000 population in 1960

Our goal in this study is to find which factors affect crime rate using this dataset.

(a) Read the data into the dataframe *UScrime*. Which state has the lowest and which state has the highest crime rates respectively in 1960? Ans. New Hampshire and Nevada

```
UScrime <-read.csv("UScrime.csv")
UScrime[which.min(UScrime$Crime),]$State
```

```
## [1] "NH"
```

```
UScrime[which.max(UScrime$Crime),]$State
```

```
## [1] "NV"
```

(b) Find the average crime rate (*Crime*) in states where per capita expenditure on police protection in 1960 was above US$ 8. Also find the average crime rate for states where per capita expenditure on police protection in 1960 was less than or equal to US$ 8.

Ans. 1137.826 and 682.0417

```
mean(UScrime[UScrime$Po1>8,]$Crime)
```

```
## [1] 1137.826
```

```
mean(UScrime[UScrime$Po1<=8,]$Crime)
```

```
## [1] 682.0417
```

(c) Run a two-sample t-test to verify if there is difference in crime rates for states spending more than 8 dollars per month and states spending less than or equal to 8 dollars per month on police protection (1960). Write down the null hypothesis for your test.

Ans.

$$H_0 : \beta_1 = \beta_2$$

vs.

$$H_1 : \beta_1 \neq \beta_2$$

(d) For the test conducted in part (c) write down the p-value and your conclusion.

Ans. P-value = 2.74471e-05. We reject null.

```
 ttest<-t.test(UScrime[UScrime$Po1>8,]$Crime,UScrime[UScrime$Po1<=8,]$Crime)
ttest$p.value
```

```
## [1] 2.74471e-05
```

(e) In this question, you will develop a simple linear regression model. Use the data on the first 42 states to fit the model. Use the subset function and remove the "States" variable by using argument select=-States. Let this data set be UStrain.

Create a simple linear regression model to fit the Crime rates of a state against their per capita expenditure on police protection ().

What is the R-squared for your model? Ans. 0.4985

```
UStrain<-subset(UScrime[1:42,], select=-States)
UStest<-subset(UScrime[43:47,], select=-States)
simple1<-lm(Crime~Po1,data=UStrain)
summary(simple1)
```

```
##
## Call:
## lm(formula = Crime ~ Po1, data = UStrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -622.13 -182.44   33.76  131.25  547.29
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   143.68     131.57   1.092    0.281
## Po1            91.65      14.54   6.305 1.76e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 285 on 40 degrees of freedom
## Multiple R-squared:  0.4985, Adjusted R-squared:  0.4859
## F-statistic: 39.76 on 1 and 40 DF,  p-value: 1.76e-07
```

(f) Create a dataset for testing the model by choosing the 43rd to 47th states (VA, WA, WV, WI, WY). Call this set UStest (remove the States variable as before). What is the 99% confidence interval for the predicted crime rate for Washington (WA)? Is the actual value within this interval?

Ans. Interval = (889.5083, 1139.268). Actual = 1030. It is inside.

```
pred1<-predict(simple1,newdata=UStest,interval=c("confidence"),level=.99)
pred1
```

```
##          fit      lwr       upr
## 43  831.0810 705.4374  956.7246
## 44 1014.3881 889.5083 1139.2679
## 45  565.2856 370.2857  760.2856
## 46 1115.2070 971.1244 1259.2897
## 47  968.5613 848.2131 1088.9096
```

```
pred1[2,2]
```

```
## [1] 889.5083
```

```
pred1[2,3]
```

```
## [1] 1139.268
```

```
UStest$Crime[2]
```

```
## [1] 1030
```

```
UScrime$Crime[which(UScrime$State=="WA")]
```

```
## [1] 1030
```

(g) What is the 99% confidence interval for the predicted crime rate for Wisconsin (WI)? Is the actual value within this interval?

Ans. Interval = (971.1244, 1259.29). Actual = 508. It is NOT in the interval.

```
pred1[4,2]
```

```
## [1] 971.1244
```

```
pred1[4,3]
```

```
## [1] 1259.29
```

```
UStest$Crime[4]
```

```
## [1] 508
```

(h) Now develop a multiple linear regression model using data on the first 45 states (created earlier) to predict Crime rates using all available predictor variables (and intercept). Which variables are significant at 0.05 level and what is the R-squared for this model?

Ans. M, Ed, Ineq. Rsquared= 0.8434

```
lmtotal<-lm(Crime~.,data=UStrain)
summary(lmtotal)
```

```
##
## Call:
## lm(formula = Crime ~ ., data = UStrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -359.11  -84.76  -43.88  112.53  445.01
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7263.9590  1947.6843  -3.730 0.000943 ***
## M              91.5784    40.7479   2.247 0.033320 *
## So             18.9650   150.1473   0.126 0.900459
## Ed            163.5351    65.5230   2.496 0.019238 *
## Po1           174.7967   130.3789   1.341 0.191621
## Po2           -99.8353   142.9119  -0.699 0.491017
## LF          -1297.2067  1566.9559  -0.828 0.415289
## M.F            32.7246    21.8503   1.498 0.146262
## Pop            -0.4918     1.2910  -0.381 0.706318
## NW              2.4013     6.9922   0.343 0.734041
## U1          -7042.4358  4397.4651  -1.601 0.121353
## U2            165.7345    83.3891   1.987 0.057498 .
## Wealth          0.1712     0.1053   1.626 0.115906
## Ineq           77.6886    23.1187   3.360 0.002414 **
## Prob        -4447.4192  2252.2974  -1.975 0.059022 .
## Time           -2.7099     8.2524  -0.328 0.745259
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 197.5 on 26 degrees of freedom
## Multiple R-squared:  0.8434, Adjusted R-squared:  0.753
## F-statistic: 9.333 on 15 and 26 DF,  p-value: 5.718e-07
```

(i) Using only the variables that you found significant at 0.05 level (p-value less than 0.05) in the previous model in part (h), create a new linear regression model. What is the R-squared for the new model? Ans. 0.1025

```
lmselect<-lm(Crime~M+Ed+Ineq,data=UStrain)
summary(lmselect)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Ineq, data = UStrain)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -560.92 -247.73  -38.29  164.01  893.85
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1098.32    1499.80  -0.732   0.4685
## M              11.18      64.03   0.175   0.8623
## Ed            153.12      85.43   1.792   0.0811 .
## Ineq           13.05      25.31   0.515   0.6092
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 391.1 on 38 degrees of freedom
## Multiple R-squared:  0.1025, Adjusted R-squared:  0.03165
## F-statistic: 1.447 on 3 and 38 DF,  p-value: 0.2444
```

(j) Among the models you created in parts (e), (h) and (i) which one would you prefer? What comparison criteria did you use to come to your conclusion? Ans. Adjusted R2. Model (h). Otherwise if we use test SSE:

```
predsim<-predict(simple1,newdata = UStest)
SSEsim= mean((predsim-UStest$Crime)^2)
SSEsim
```

```
## [1] 79093.44
```

```
predlmt<-predict(lmtotal,newdata = UStest)
SSElmt= mean((predlmt-UStest$Crime)^2)
SSElmt
```

```
## [1] 92697.84
```

```
predlmsel<-predict(lmselect,newdata = UStest)
SSElmsel= mean((predlmsel-UStest$Crime)^2)
SSElmsel
```

```
## [1] 55491.31
```

(k) We now use the **regsubsets** function in the **leaps** package for a best subset selection. How many variables are included in the model if you use adjusted R-squared to pick your model? Ans. 10

```
library(leaps)
modsubex<-regsubsets(Crime~.,data=UStrain,nvmax=15)
summary(modsubex)$adjr2
```

```
##  [1] 0.4859287 0.5947319 0.6694974 0.7018156 0.7251878 0.7451737 0.7628699
##  [8] 0.7765279 0.7856578 0.7860374 0.7813537 0.7760594 0.7693077 0.7620143
## [15] 0.7530126
```

```
idx <- which.max(summary(modsubex)$adjr2)
length(coef(modsubex, idx)) - 1 # -1 for intercept
```

```
## [1] 10
```

```
coef(modsubex, idx)
```

```
##   (Intercept)             M            Ed           Po1            LF
## -8214.9652466   101.2849955   164.7081572    82.5132186 -1202.9337062
##           M.F            U1            U2        Wealth          Ineq
##    38.9493880 -7493.3390043   183.8828703     0.1630412    80.0172862
##          Prob
## -3486.7652338
```

(l) How many variables are selected if you use BIC to pick your model with best subset selection? Ans. 8

```
summary(modsubex)$bic
```

```
##  [1] -21.50827 -28.82210 -34.74064 -36.44494 -37.28624 -37.90299 -38.40567
##  [8] -38.41336 -37.72004 -35.39026 -32.12030 -28.80162 -25.29022 -21.77272
## [15] -18.06081
```

```
idx <- which.min(summary(modsubex)$bic)
length(coef(modsubex, idx)) - 1
```

```
## [1] 8
```

```
coef(modsubex, idx)
```

```
## (Intercept)           M          Ed         Po1         M.F          U1
## -7281.20993    90.57845   164.76357   100.71144    35.05074 -7650.63856
##          U2        Ineq        Prob
##   207.72305    57.22673 -3607.86374
```

(m) How many variables are selected if you use BIC to pick your model with forward stepwise subset selection?

Ans. 7

```
modsubforward<-regsubsets(Crime~.,data=UStrain,method="forward")
summary(modsubforward)$bic
```

```
## [1] -21.50827 -28.82210 -34.74064 -36.44494 -35.84528 -36.95185 -37.80652
## [8] -36.97251
```

```
idx <- which.min(summary(modsubforward)$bic)
length(coef(modsubex, idx)) - 1
```

```
## [1] 7
```

```
coef(modsubex, idx)
```

```
##   (Intercept)             M            Ed           Po1            U2
## -6302.8238362   120.9305845   172.9962400    91.9826351    91.3947647
##        Wealth          Ineq          Prob
##     0.1974631    90.0693833 -3085.3996117
```

(n) Now using UStest set for your out-of-sample validation, you are asked to choose one of the models from parts (k), (l), and (m). Create separate multiple linear regression models using the predictors found in parts (k), (l) and (m). Use these models to find the sum-of- squared errors (SSE) for the UStest data set.

What is the test set SSE for the model in part (k)?

Ans. SSEK = 507749.7

```
ModK<-lm(Crime~M+Ed+Po1+LF+M.F+U1+U2+Wealth+Ineq+Prob,data=UStrain)
PredK<-predict(ModK,newdata = UStest)
SSEK= sum((PredK-UStest$Crime)^2)

ModL<-lm(Crime~M+Ed+Po1+M.F+U1+U2+Ineq+Prob,data=UStrain)
PredL<-predict(ModL,newdata = UStest)
SSEL= sum((PredL-UStest$Crime)^2)

ModM<-lm(Crime~M+Ed+Po1+M.F+Wealth+Ineq+Prob,data=UStrain)
PredM<-predict(ModM,newdata = UStest)
SSEM= sum((PredM-UStest$Crime)^2)

SSEK
```

```
## [1] 507749.7
```

```
SSEL
```

```
## [1] 358031.2
```

```
SSEM
```

```
## [1] 436948
```

(o) What are the test set SSE for the models in parts (l) and (m)?

Ans. SSEL= 358031.2, SSEM = 436948

(p) Which model do you prefer among the ones created in parts (k), (l) and (m) based on your findings above?

Ans. The lowest SSE is for model (l), which we choose.

END