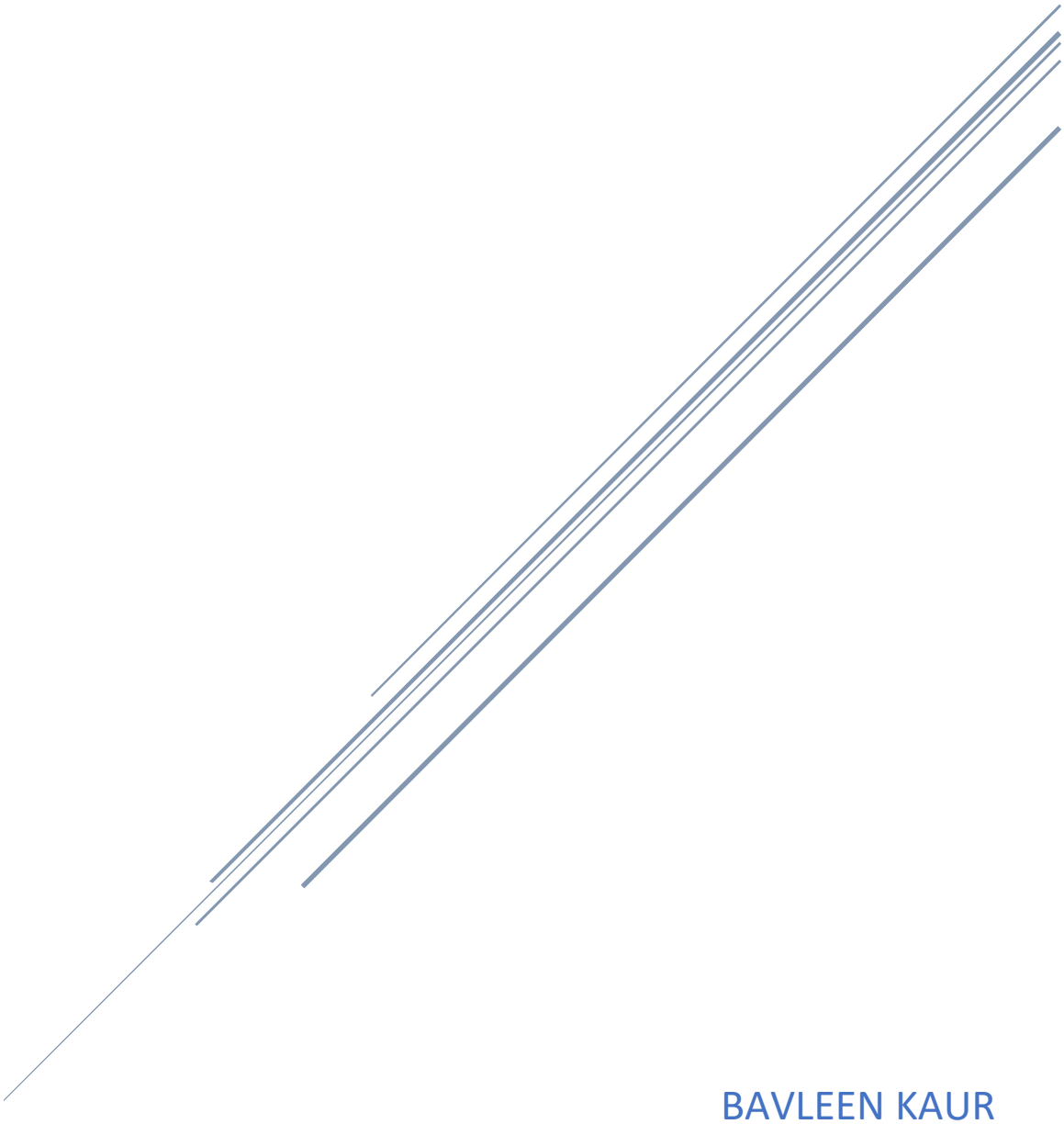
ASSIGNMENT 3

APPLIED REGRESSION ANALYSIS

STAT 3101-001 F2021



BAVLEEN KAUR
#3122587

Question 1 A hospital administrator is studying the relation between patient satisfaction (Y, anindex) and patient's age (X1, in years), severity of illness (X2, an index), and anxiety level (X3, an index). Data are reported for 23 randomly selected patients. Forall index variables, higher values indicate more (satisfaction, severity, anxiety). Data(patient.txt) can be found on NEXUS Assignment 3 folder.

1a. Fit a regression model for Y using the three predictor variables (usingX1,X2,X3 to predict Y) and state the regression parameters and their standard errors. Comment on the overall results.

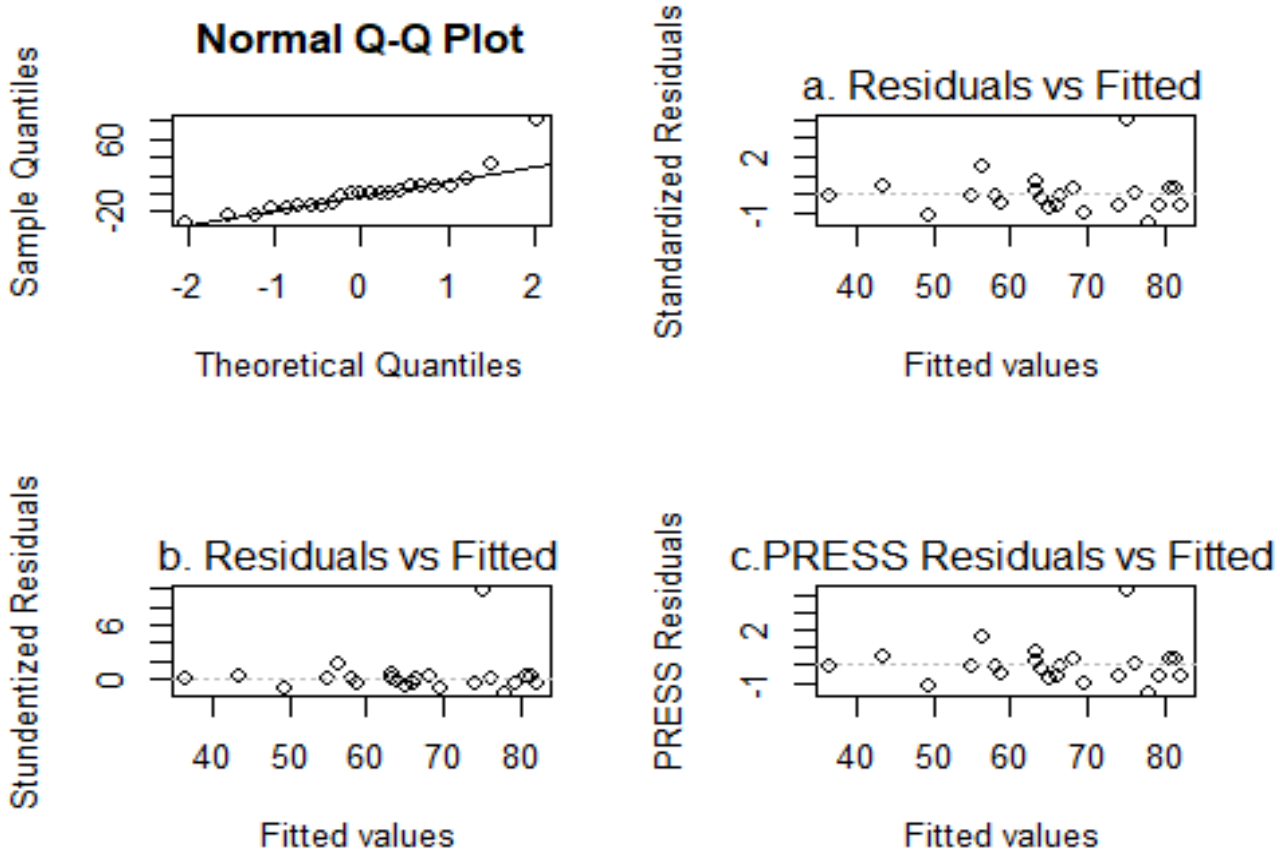
```
fit <- lm(formula = y~x1+x2+x3)
summary(fit)

##
## Call:
## lm(formula = y ~ x1 + x2 + x3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.999 -13.878  -0.924   7.317  81.882
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 114.06659   61.67693   1.849   0.0800 .
## x1           -0.08374    0.72132  -0.116   0.9088
## x2            1.62492    1.96451   0.827   0.4184
## x3          -55.57100   29.29132  -1.897   0.0731 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.62 on 19 degrees of freedom
## Multiple R-squared:  0.2239, Adjusted R-squared:  0.1014
## F-statistic: 1.827 on 3 and 19 DF,  p-value: 0.1764
```

$$y = 114.06659 - 0.08374x_1 + 1.62492x_2 - 55.571x_3$$

It seems like none of the predictors are significant.

1b. Use the appropriate diagnostics to check (standardized, studentized, and PRESS residuals vs fitted values plots and also three Q-Q plots) if any of the model assumptions is violated and comment on your findings. Identify, if there is any suspected outlier.

So, we can see the normal qq plot, seems a little right skewed. The next three are the standardized, studentized, and PRESS residuals plots. There are suspected outliers.



1.c. Obtain the leverages, Cook's distances, and DFBETAs for each of the 23 observations. Do not paste all the values into your report; just use them to identify outliers and influential cases. Carefully examine all three measures for 23 individuals and find out there is any problematic observation?

```
#Studentized residual

b <- data1[stdres(fit)>4,]
b

##    id   y x1 x2  x3
## 15 15 157 53 54 2.2

lev = hat(model.matrix(fit))
data1[which(lev >= 0.34),]

## [1] id y  x1 x2 x3
## <0 rows> (or 0-length row.names)
```

 # k =4, n = 23 , cutoff = 2k/n = 8/23 = 0.35

No leverage points.

```r
#Cook's distance
data1[which(cooks.distance(fit) >1 ), ]
```

```
##    id   y x1 x2  x3
## 15 15 157 53 54 2.2
```

```r
#DFBeta
dfbeta = dfbeta(fit)[,-1]
data1[which(dfbeta > 0.417),]
```

```
##      id   y x1 x2  x3
## 15   15 157 53 54 2.2
```

Observation 15 is the problematic one.

1.d. Remove the most problematic data point(s) (if there is any) in turn and re-fit the model. Explain the effect of removing problematic data point(s) on the estimates of the model parameters, sum of squared errors, and R2.

We will remove observation 15.

```r
full = lm(y~x1+x2+x3, data = data1)

anova(full)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## x1         1   725.4  725.42  1.1967 0.28766
## x2         1   415.9  415.91  0.6861 0.41778
## x3         1  2181.9 2181.87  3.5993 0.07311 .
## Residuals 19 11517.7  606.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
out1 = round(c(full$coef, anova(full)$"Mean sq"[3], summary(full)$adj.r.squared),3)


data0 = data1[-15,]
fit0 = lm(y~x1+x2+x3, data = data0)
anova(fit0)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x1         1 3966.3  3966.3 40.5918 5.317e-06 ***
## x2         1  400.3   400.3  4.0968   0.05807 .
```

```
## x3         1    0.1     0.1  0.0009   0.97618
## Residuals 18 1758.8    97.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
out2 = round(c(fit0$coef, anova(fit0)$"Mean Sq"[3], summary(fit0)$adj.r.squar
ed),3)

result = data.frame(rbind(out1, out2))

names(result) = c("beta0", "beta1", "beta2", "beta3", "MS residual", "Adj-Rsq
uare")
result
```

```
##         beta0  beta1  beta2   beta3 MS residual Adj-Rsquare
## out1 114.067 -0.084  1.625 -55.571       0.101     114.067
## out2 172.238 -1.426 -1.105   0.394       0.090       0.665
```

We can observe a big change in the Mean Squared error and The Adjusted R^2, after removal of observation 15.

Question 2 Fit the regression model based on each of following selection procedures and summarize your findings. Finally, you will recommend the best regression model for this data

Backward selection

```
#model fitting all predictors
g1 = lm(income~., data = final)
summary(g1)
```

```
##
## Call:
## lm(formula = income ~ ., data = final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3945 -0.3895  0.1085  0.4874  3.1321
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.009924   0.177836  39.418  < 2e-16 ***
## race         0.096710   0.079469   1.217 0.224017
## gender       0.249001   0.065159   3.821 0.000144 ***
```

```
## citizen     -0.237987   0.146824  -1.621 0.105467
## lang         0.097116   0.090030   1.079 0.281069
## married      0.212273   0.068147   3.115 0.001911 **
## COLHS        0.373362   0.071607   5.214 2.41e-07 ***
## GradHS       0.900629   0.099726   9.031  < 2e-16 ***
## hoursWorked  0.049302   0.002667  18.485  < 2e-16 ***
## age          0.018419   0.002289   8.046 3.44e-15 ***
## timetoWork   0.002638   0.001395   1.892 0.058915 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8385 on 734 degrees of freedom
## Multiple R-squared:  0.525,  Adjusted R-squared:  0.5186
## F-statistic: 81.14 on 10 and 734 DF,  p-value: < 2.2e-16
```

deleting lang based on high p-value

```
g2 = update(g1, income~.-lang)
summary(g2)
```

```
##
## Call:
## lm(formula = income ~ race + gender + citizen + married + COLHS +
##     GradHS + hoursWorked + age + timetoWork, data = final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3722 -0.3824  0.1012  0.4947  3.1435
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.017153   0.177730  39.482  < 2e-16 ***
## race         0.114552   0.077738   1.474 0.141023
## gender       0.252374   0.065091   3.877 0.000115 ***
## citizen     -0.174681   0.134600  -1.298 0.194770
## married      0.208454   0.068063   3.063 0.002274 **
## COLHS        0.377513   0.071512   5.279 1.71e-07 ***
## GradHS       0.899345   0.099730   9.018  < 2e-16 ***
## hoursWorked  0.049034   0.002656  18.463  < 2e-16 ***
## age          0.018605   0.002283   8.149 1.57e-15 ***
## timetoWork   0.002623   0.001395   1.881 0.060377 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8386 on 735 degrees of freedom
## Multiple R-squared:  0.5243, Adjusted R-squared:  0.5185
## F-statistic:    90 on 9 and 735 DF,  p-value: < 2.2e-16
```

```r
#deleting citizen, based on highest p value 0.194770
g3 = update(g2, income~.-citizen)
summary(g3)
```

```
##
## Call:
## lm(formula = income ~ race + gender + married + COLHS + GradHS +
##     hoursWorked + age + timetoWork, data = final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3950 -0.3758  0.1059  0.4870  3.1355
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.881302   0.143698  47.887  < 2e-16 ***
## race        0.091996   0.075805   1.214 0.225295
## gender      0.254285   0.065105   3.906 0.000103 ***
## married     0.205861   0.068065   3.024 0.002577 **
## COLHS       0.370115   0.071317   5.190 2.73e-07 ***
## GradHS      0.901826   0.099758   9.040  < 2e-16 ***
## hoursWorked 0.049053   0.002657  18.462  < 2e-16 ***
## age         0.018303   0.002272   8.055 3.19e-15 ***
## timetoWork  0.002768   0.001391   1.991 0.046902 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.839 on 736 degrees of freedom
## Multiple R-squared:  0.5232, Adjusted R-squared:  0.518
## F-statistic:   101 on 8 and 736 DF,  p-value: < 2.2e-16
```

```r
#deleting race based on highest p value, 0.225295
g4 = update(g3, income~.-race)
summary(g4)
```

```
##
## Call:
## lm(formula = income ~ gender + married + COLHS + GradHS + hoursWorked +
##     age + timetoWork, data = final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4720 -0.3734  0.1142  0.4977  3.1516
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.945183   0.133753  51.925  < 2e-16 ***
## gender      0.255414   0.065119   3.922 9.59e-05 ***
## married     0.213915   0.067762   3.157  0.00166 **
## COLHS       0.373918   0.071271   5.246 2.03e-07 ***
```

```
## GradHS      0.896946   0.099709   8.996  < 2e-16 ***
## hoursWorked 0.049073   0.002658  18.463  < 2e-16 ***
## age         0.018401   0.002271   8.101 2.26e-15 ***
## timetoWork  0.002682   0.001389   1.930  0.05393 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8392 on 737 degrees of freedom
## Multiple R-squared:  0.5222, Adjusted R-squared:  0.5177
## F-statistic: 115.1 on 7 and 737 DF,  p-value: < 2.2e-16
```

```
#deleting timetowork based on highest p value, 0.0539(close call)
g5 = update(g4, income~.-timetoWork)
summary(g5)
```

```
##
## Call:
## lm(formula = income ~ gender + married + COLHS + GradHS + hoursWorked +
##     age, data = final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4829 -0.3943  0.1014  0.4828  3.1142
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.985219   0.132379  52.767  < 2e-16 ***
## gender      0.259082   0.065211   3.973 7.79e-05 ***
## married     0.218528   0.067845   3.221  0.00133 **
## COLHS       0.390229   0.070899   5.504 5.13e-08 ***
## GradHS      0.897948   0.099892   8.989  < 2e-16 ***
## hoursWorked 0.049565   0.002650  18.701  < 2e-16 ***
## age         0.018438   0.002276   8.102 2.23e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8408 on 738 degrees of freedom
## Multiple R-squared:  0.5198, Adjusted R-squared:  0.5159
## F-statistic: 133.2 on 6 and 738 DF,  p-value: < 2.2e-16
```

The process stops here, as all the left predictors are significant. We see our R-squared value changed from 0.525 to 0.5918 after removal of the predictors.

Forward selection using AIC

```
summary(FORW1)
```

```
##
## Call:
```

```
## lm(formula = income ~ hoursWorked + age + GradHS + COLHS + gender +
##     married + timetoWork, data = final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4720 -0.3734  0.1142  0.4977  3.1516
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.945183   0.133753  51.925  < 2e-16 ***
## hoursWorked 0.049073   0.002658  18.463  < 2e-16 ***
## age         0.018401   0.002271   8.101 2.26e-15 ***
## GradHS      0.896946   0.099709   8.996  < 2e-16 ***
## COLHS       0.373918   0.071271   5.246 2.03e-07 ***
## gender      0.255414   0.065119   3.922 9.59e-05 ***
## married     0.213915   0.067762   3.157  0.00166 **
## timetoWork  0.002682   0.001389   1.930  0.05393 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8392 on 737 degrees of freedom
## Multiple R-squared:  0.5222, Adjusted R-squared:  0.5177
## F-statistic: 115.1 on 7 and 737 DF,  p-value: < 2.2e-16
```

It shows that number of hours worked, age, education, gender and marital status are all significant predictors.

Stepwise selection with AIC Criterion

```
summary(STEP)
```

```
##
## Call:
## lm(formula = income ~ hoursWorked + age + GradHS + COLHS + gender +
##     married + timetoWork, data = final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4720 -0.3734  0.1142  0.4977  3.1516
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.945183   0.133753  51.925  < 2e-16 ***
## hoursWorked 0.049073   0.002658  18.463  < 2e-16 ***
## age         0.018401   0.002271   8.101 2.26e-15 ***
## GradHS      0.896946   0.099709   8.996  < 2e-16 ***
## COLHS       0.373918   0.071271   5.246 2.03e-07 ***
## gender      0.255414   0.065119   3.922 9.59e-05 ***
```

```
## married     0.213915   0.067762   3.157  0.00166 **
## timetoWork  0.002682   0.001389   1.930  0.05393 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8392 on 737 degrees of freedom
## Multiple R-squared:  0.5222, Adjusted R-squared:  0.5177
## F-statistic: 115.1 on 7 and 737 DF,  p-value: < 2.2e-16
```

It shows that number of hours worked, age, education, gender, and marital status are all significant predictors.

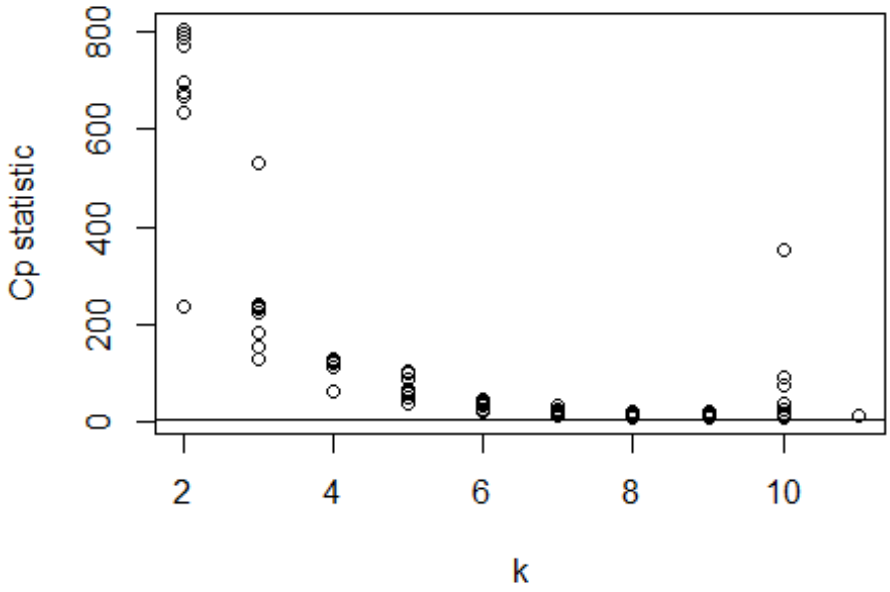
Adjusted R^2

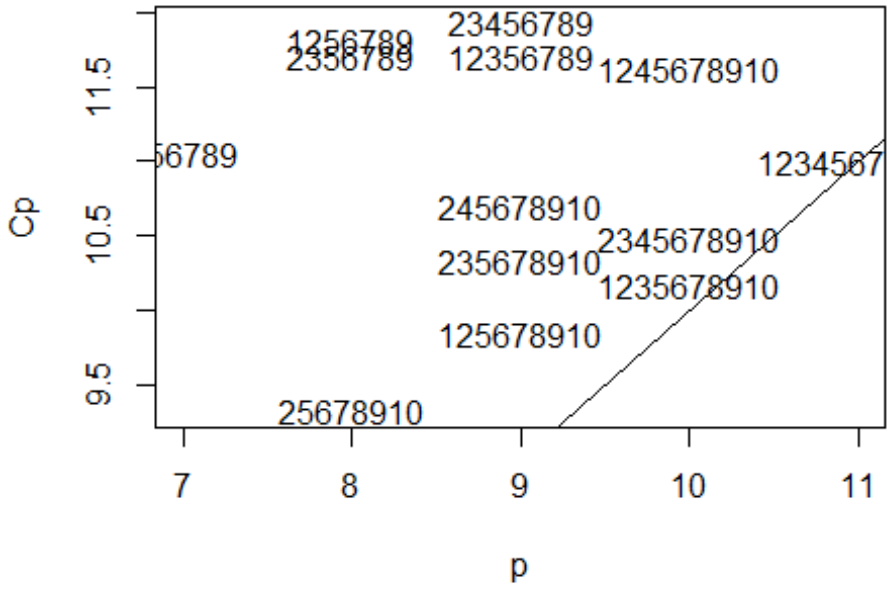
```
maxadjr(ADJR2,11)
```

```
## 1,2,3,4,5,6,7,8,9,10   1,2,3,5,6,7,8,9,10   2,3,4,5,6,7,8,9,10
##                0.519                0.518                0.518
##    1,2,5,6,7,8,9,10      2,5,6,7,8,9,10     2,3,5,6,7,8,9,10
##                0.518                0.518                0.518
##   1,2,4,5,6,7,8,9,10    2,4,5,6,7,8,9,10    1,2,3,4,5,6,7,8,9
##                0.517                0.517                0.517
##     1,2,3,5,6,7,8,9      2,3,4,5,6,7,8,9
##                0.517                0.517
```

We see that the model with all the predictors has the largest adjusted R2

Mellow's Cp





Mean Squared Error

```r
MSE = round(SSres/(13-p),2)
MSE
```

```
##      1      1      1      1      2      2      2      2      3      3
3
##  62.55  87.87  90.12  90.55  60.98  62.85  64.71  67.69  62.54  66.20  66.
28
##      3      4      4      4      4      5      5      5      5      6
6
##  66.43  67.87  68.87  69.06  69.67  75.58  76.12  77.08  77.33  86.95  87.
68
##      6      6      7      7      7      7      8      8      8      8
##  87.94  88.06 103.82 104.15 104.17 104.28 129.51 129.60 129.66 129.84
```

Appendix

R code

```
setwd("C:/Users/12048/Desktop/Regression Analysis/Assignment 3")
data <- read.csv("patient.csv")
attach(data)
head(data)

id=seq(1,23)
y <- data[,1]

x1 <- data[,2]
x2 <- data[,3]
x3 <- data[,4]
data1 = data.frame(id,y,x1,x2,x3)
#1. regression
fit <- lm(formula = y~x1+x2+x3)
summary(fit)



pr = residuals(fit)/(1-lm.influence(fit)$hat)
PRESS <- sum(pr^2)
VAR = var(residuals(fit))/(1-lm.influence(fit)$hat)
STDVAR = pr/sqrt(VAR)

 library(MASS)

par(mfrow=c(2,2))

qqnorm(fit$res)
```

```
qqline(fit$res)

plot(fit$fitted, stdres(fit), xlab="Fitted values", ylab="Standardized Residuals")
mtext("a. Residuals vs Fitted", line = .25)
abline(h=0, lty=3, col="grey")


plot(fit$fitted, studres(fit), xlab="Fitted values", ylab="Stundentized Residuals")
mtext("b. Residuals vs Fitted", line = .25)
abline(h=0, lty=3, col="grey")


plot(fit$fitted,STDVAR, xlab="Fitted values", ylab="PRESS Residuals")
mtext("c.PRESS Residuals vs Fitted", line = .25)
abline(h=0, lty=3, col="grey")

stdres(fit)
sort(stdres(fit))
a <- data1[stdres(fit)>4,]


studres(fit)
sort(studres(fit))
b <- data1[stdres(fit)>4,]
b

lev= hat(model.matrix(fit))
plot(lev)
c <- data1[lev>0.24,]
```

```
c


sort(STDVAR)
d <- data1[STDVAR>4,]


rstudent(fit)
sort(rstudent(fit))


lev = hat(model.matrix(fit))
data1[which(lev >= 0.34),]


# k =4, n = 23 , cutoff = 2k/n = 8/23 = 0.35
# no leverage point


#Cook's distance
data1[which(cooks.distance(fit) >1 ), ]


#DFBeta
dfbeta = dfbeta(fit)[,-1]
data1[which(dfbeta > 0.417),]


full = lm(y~x1+x2+x3, data = data1)


anova(full)
 anova(full)$"Mean sq"
out1 = round(c(full$coef, anova(full)$"Mean sq"[3], summary(full)$adj.r.squared),3)



data0 = data1[-15,]
```

```
fit0 = lm(y~x1+x2+x3, data = data0)
anova(fit0)

out2 = round(c(fit0$coef, anova(fit0)$"Mean Sq"[3], summary(fit0)$adj.r.squared),3)

result = data.frame(rbind(out1, out2))
names(result) = c("beta0", "beta1", "beta2", "beta3", "MS residual", "Adj-Rsquare")
result

setwd("C:/Users/12048/Desktop/Regression Analysis/Assignment 3")
myData <- read.csv("ACS.csv")
attach(myData)
head(myData)

#ERD on raw data
#library(graphics)

par(mfrow = c (2,2))
income = myData[,2]
hrswork = myData[,4]
age = myData[,6]
timetowork = myData[,9]
data1 = na.omit(myData)
newdata = data1[data1$income !=0 ,]
head(newdata)
newdata <- newdata[,-3]
head(newdata)

income = log(newdata$income)
```

```
race = ifelse(newdata$race == "white" , 1, 0 )

gender = ifelse(newdata$gender == "male" , 1, 0 )

citizen = ifelse(newdata$citizen == "yes" , 1, 0 )

lang = ifelse(newdata$lang == "english" , 1, 0 )

married = ifelse(newdata$married == "yes" , 1, 0 )

COLHS = ifelse(newdata$edu == "college", 1, 0)

GradHS = ifelse(newdata$edu == "grad", 1, 0)

hoursWorked = newdata$hrs_work

age = newdata$age

timetoWork = newdata$time_to_work




final = data.frame(income, race, gender, citizen, lang, married, COLHS, GradHS, hoursWorked, age,
timetoWork)

head(final)


#models fitting all predictors

g1 = lm(income~., data = final)

summary(g1)


#deleting lang based on high p-value

g2 = update(g1, income~.-lang)

summary(g2)


#deleting citizen, based on highest p value 0.194770

g3 = update(g2, income~.-citizen)

summary(g3)
```

```
#deleting race based on highest p value, 0.225295
g4 = update(g3, income~.-race)
summary(g4)

#deleting timetowork based on highest p value, 0.0539(close call)
g5 = update(g4, income~.-timetoWork)
summary(g5)

#The process stops here as all oher are significant.

#R^2 0.525 ---> 0.5198

#install.packages("faraway")
library(faraway)
library(stats)
library(MASS)
#install.packages("olsrr")
library(olsrr)

#Fitting model with intercept only
null = lm(income~1, data = final)
#fitting model with all predictors
full = lm(income~., data = final)

FORW1 = stepAIC(null, scope=list(lower=null, upper=full), data=final, direction="forward")

summary(FORW1)
FORW1$anova
```

```
STEP = stepAIC(null, scope = list(upper=full),data = final, direction="both")
summary(STEP)
STEP$anova

library(leaps)
full = lm(income~.,data = final)
x= model.matrix(full)[,-1]
y=final$income
ADJR2 = leaps(x,y,method  = "adjr2")

#install.packages("faraway")
library(faraway)

maxadjr(ADJR2,11)

full = lm(income~.,data = final)
x= model.matrix(full)[,-1]
y=final$income
CP = leaps(x,y,method = c("Cp"))

plot(CP$size, CP$Cp, xlab = "k" , ylab = "Cp statistic")
abline( h =5)
Cpplot(CP)

all = regsubsets(x,y,  all.best = FALSE, nbest = 4)
SSres = round(summary(all)$rss, 2)
Mat = round(summary(all)$which ,2)
p = apply(Mat , 1, sum)
MSE = round(SSres/(13-p),2)
```

MSE