Dhakal_Module4

Kokil Dhakal

2023-11-27

Loading The Data set

```
data.module4 <- read_excel("/Users/kokildhakal/Desktop/STUDY/BU/9.CS555/Module4/module4.data.xlsx")
head(data.module4)
```

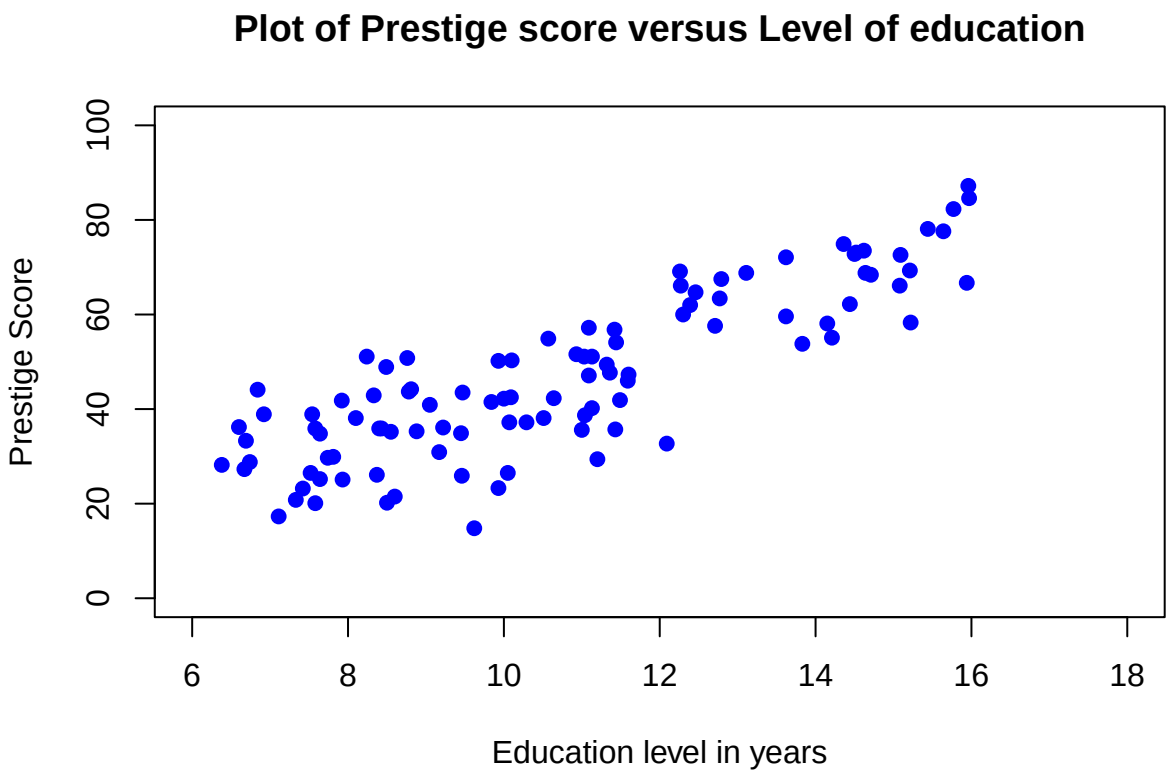
```
## # A tibble: 6 x 5
##   `Occupational Title` `Education Level (years)` `Income ($)`
##   <chr>                                    <dbl>        <dbl>
## 1 GOV_ADMINISTRATORS                        13.1        12351
## 2 GENERAL_MANAGERS                          12.3        25879
## 3 ACCOUNTANTS                               12.8         9271
## 4 PURCHASING_OFFICERS                       11.4         8865
## 5 CHEMISTS                                  14.6         8403
## 6 PHYSICISTS                                15.6        11030
## # i 2 more variables: `Percent of Workforce that are Women` <dbl>,
## #   `Prestige Score` <dbl>
```

#——————————————————————————-

Question 1

To get a sense of the data, generate a scatterplot to examine the association between prestige score and years of education.Briefly describe the form, direction, and strength of the association between the variables.Calculate the correlation coefficient.

```
plot(x=data.module4$`Education Level (years)`,y=data.module4$`Prestige Score`,
     xlab = "Education level in years",
     ylab = "Prestige Score",
     main = "Plot of Prestige score versus Level of education",
     xlim = c(6,18),
     ylim = c(0,100),
     cex=1.1,
     col="blue",
     pch=16)
```




```
#calculation correlation coefficients

cor(data.module4$`Education Level (years)`,data.module4$`Prestige Score`)
```

```
## [1] 0.8501769
```

From above scatter plot,it seems that the form is somewhat linear as data points tend towards a straight line pattern.The direction of the association is positive as we can see that prestige score increase with increase of the Education levels.Also, Strength seems moderate strong as variability of data points are not much and are seems to follow a clear pattern.Also, when I calculate the correlation coefficients between prestige score and years of education,I got about 0.85 which is the indication of positive association as it is positive value as well as there is a strong association between the variables as max positive correlation coefficients is 1.

Correlation between diffret variables(Extra)

Calculating correlation table :

```
cor.table <- data.frame(cor(data.module4[2:5]))
rownames(cor.table) <- c("Education","Income","Women.workforce%","Score")
colnames(cor.table) <- c("Education","Income","Women.workforce%","Score")
kable(cor.table,align = "c",digits = 2)
```

	Education	Income	Women.workforce%	Score
Education	1.00	0.58	0.06	0.85
Income	0.58	1.00	-0.44	0.71
Women.workforce%	0.06	-0.44	1.00	-0.12
Score	0.85	0.71	-0.12	1.00

Question 2

Perform a simple linear regression with prestige score and years of education, and briefly summarize your conclusions (no need to do the 5-step procedure here). Generate a residual plot.Assess whether the model assumptions are met.Are there any outliers or influence points? If so, identify them by ID and comment on the effect of each on the regression.

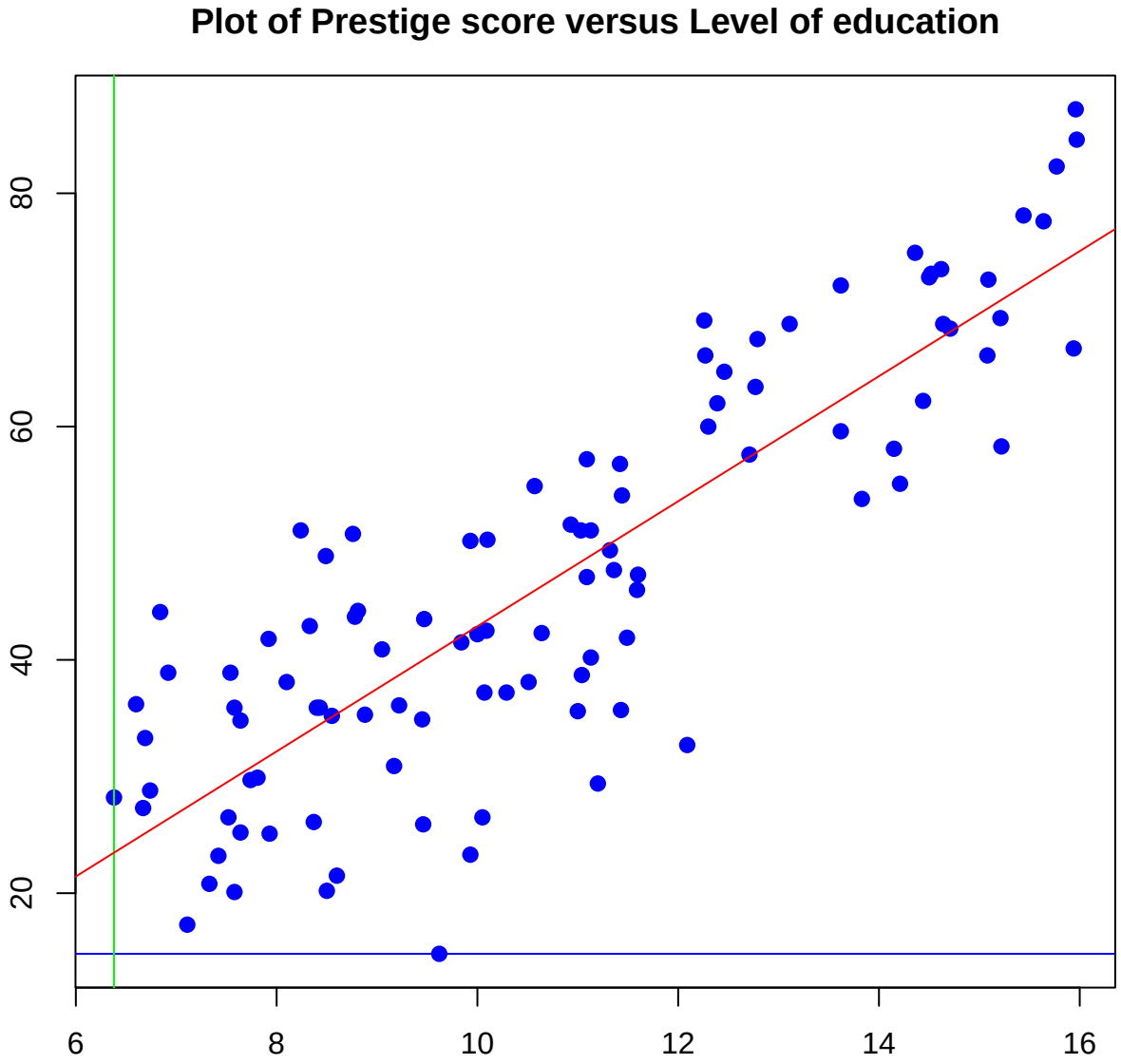
```
set.seed(123)
model.4 <- lm(data = data.module4,`Prestige Score`~ `Education Level (years)`)
summary(model.4)
```

Simple Linear linear regression

```
##
## Call:
## lm(formula = `Prestige Score` ~ `Education Level (years)`, data = data.module4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.0397  -6.5228   0.6611   6.7430  18.1636
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -10.732      3.677  -2.919  0.00434 **
## `Education Level (years)`    5.361      0.332  16.148  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.103 on 100 degrees of freedom
## Multiple R-squared:  0.7228, Adjusted R-squared:   0.72
## F-statistic: 260.8 on 1 and 100 DF,  p-value: < 2.2e-16
```

Adding regression line to the above plot

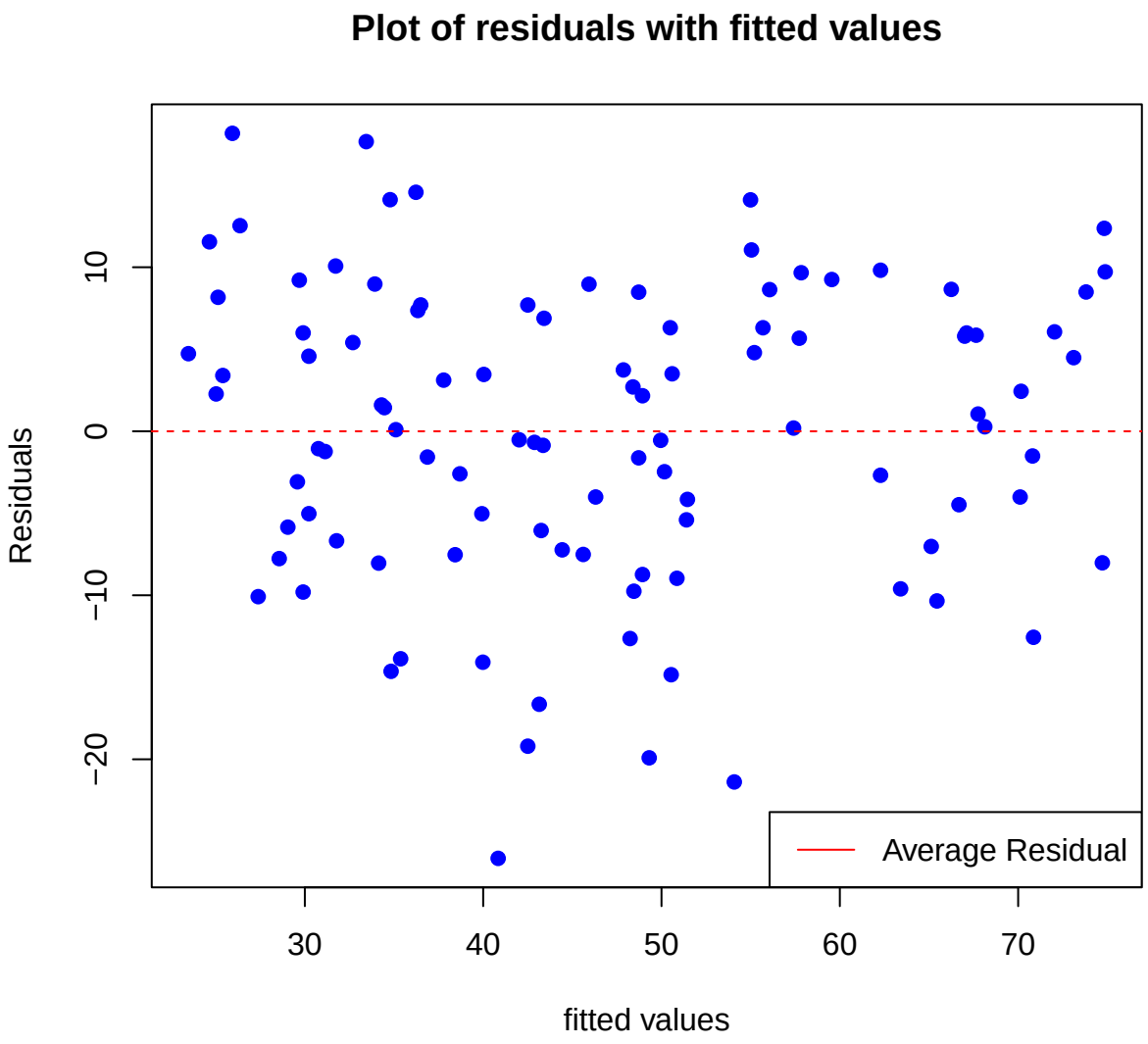
```
par(mar=c(2,2,3,2))
plot(x=data.module4$`Education Level (years)`,y=data.module4$`Prestige Score`,
     xlab = "Education level in years",
     ylab = "Prestige Score",
     main = "Plot of Prestige score versus Level of education",
     cex=1.2,
     col="blue",
     pch=16)
abline(h=min(data.module4$`Prestige Score`),col="blue")
abline(v=min(data.module4$`Education Level (years)`),col="green")
abline(model.4,col="red")
```



After performing simple linear regression and checking summary of the model, it is found that there is strong association between Education level and prestige score as we can see from plot, R-squared value(0.7228) and checking p value which is less than 0.001(i.e. $\beta 1 \neq 0$).

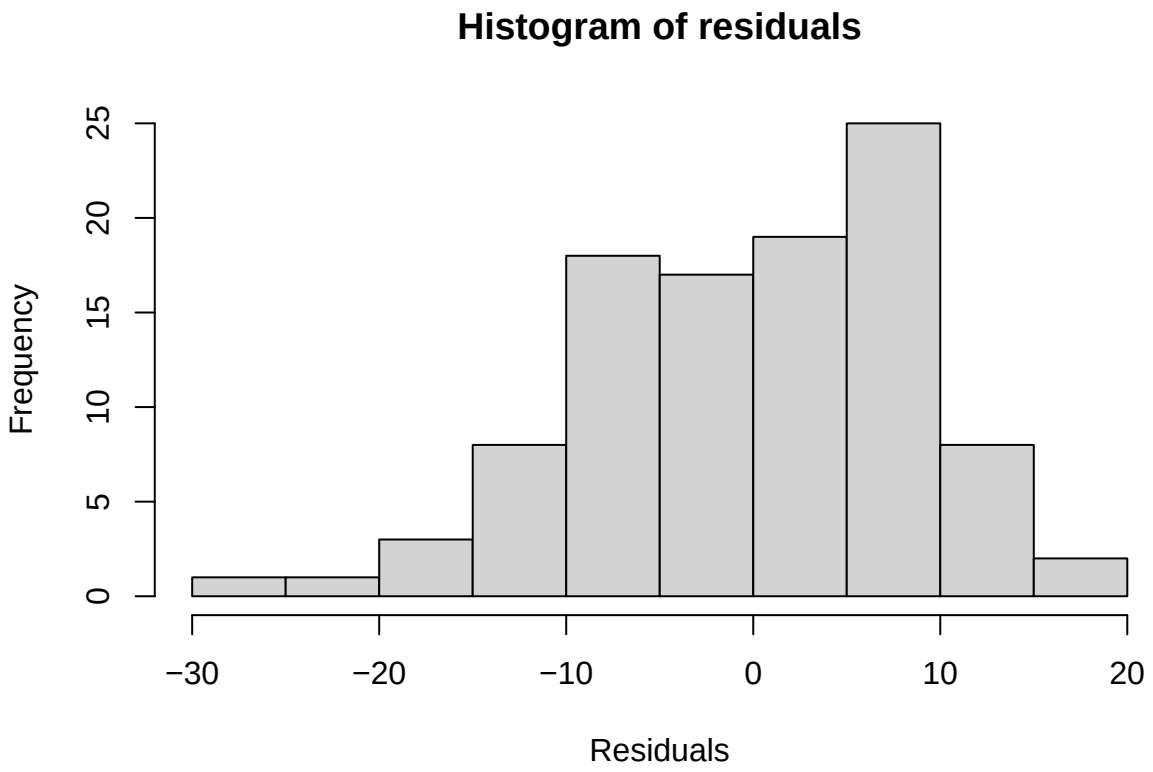
Generating the Residual plot

```
plot(y=model.4$residuals,x=model.4$fitted.values,
     cex=1.1,
     col="blue",
     pch=16,
     xlab ="fitted values",
     ylab = "Residuals",
     main = "Plot of residuals with fitted values")
abline(h=mean(model.4$residuals),col="red",lty=2)
legend("bottomright", legend = ("Average Residual"),
       col = "red", lty = c(1, 1))
```



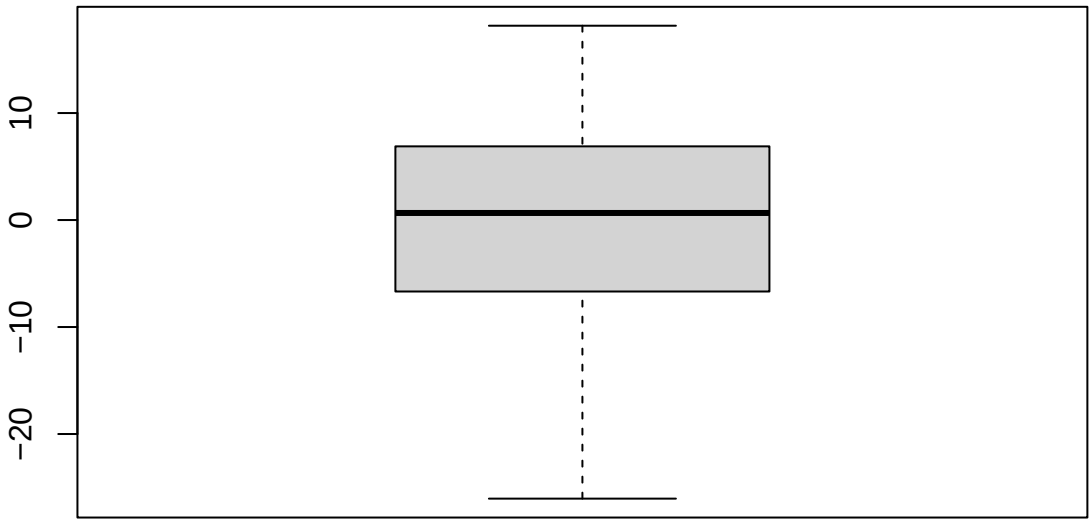
Plotting histogram of Residuals

```
hist(model.4$residuals,main = "Histogram of residuals",xlab = "Residuals",ylab = "Frequency")
```



plotting the boxplot of the residuals.

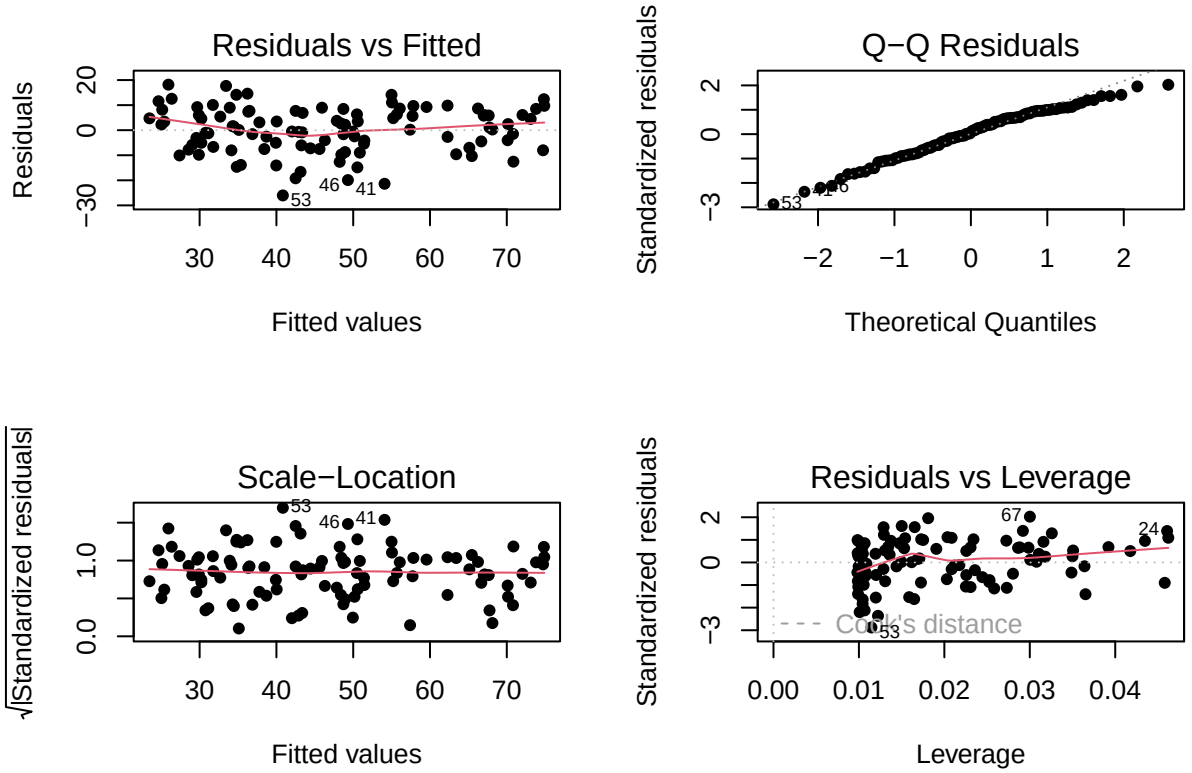
```
boxplot(model.4$residuals)
```



From above plots, it is found that the histogram of the residuals are roughly normally distributed.Also, from the scatter plot between residual and fitted values,no any parttern can be seen which fulfills the model assumptions of The variation of the response variable around the regression line is constant.Plot between education and score shows there is linear association between these variables.Also,we assumed the observations are independent as each observation is for a different occupation.For these reason, we can say that all four model assumptions are met.

Checking model characteristics and influential points.

```r
par(mfrow=c(2,2))
plot(model.4,pch=16,)
```




```r
par(mfrow=c(1,1))
```

The plot above also checks whether the model met the all assumptions.first fig checks whether the association is linear or not, second checks whether residuals are normally distributed and third checks constant variance of the variables.Also, we assume observations are independently taken.

```r
#finding potential influential points using cooks' distance
influential.points <- cooks.distance(model.4)>4/(nrow(data.module4)-2)
cat("Influential points:", which(influential.points), "\n")
```

Finding influential points and checking effect of these points on model

```
## Influential points: 24 53 67
```

```r
# checking influence of each point

set.seed(123)
for (x in which(influential.points)) {
 my.data <- data.module4 #copying main data
 my.data <- my.data[-x,] #removing the row with potential influential point
 #building model after removing the potential influential point
```

```
 my.model <- lm(data = my.data,`Prestige Score`~ `Education Level (years)`)

 #To get the slope of model
 slope.value <- summary(my.model)$coefficients[2,1]
 #to get r-squared from model
 rsquared <- summary(my.model)$r.squared
 cat("After removing row:",x,"Rsquard is ",rsquared,"and slope is",slope.value,"\n")
}
```

```
## After removing row: 24 Rsquard is  0.7123337 and slope is 5.2708
## After removing row: 53 Rsquard is  0.7366151 and slope is 5.321708
## After removing row: 67 Rsquard is  0.7341107 and slope is 5.457958
```

```
cat("while origial r-squared is",summary(model.4)$r.squared,"and slope",summary(model.4)$coefficients[2
```

```
## while origial r-squared is 0.7228007 and slope 5.360878
```

Here, from both plot(residual vs leverage) and calculation, it seems that there are three points that could be potential influential to the model we built. Now, I am going to use loop method to go through each point and for each point it removes that point from the data set and then build the linear regression model. From each model it extracts the slope and R-squared values.Now we will compare each of those values with original slope and r-squared values(without removing those points).From our calculation above, it is found that there is not much change in th slope and R-squared values that is why we can say that these points does not have big individual effect on model.

Question 3

Calculate the least squares regression equation that predicts prestige score from education, income, and percentage of women. Formally test (using the 5-step procedure) whether the set of these predictors are associated with prestige score at the = 0.05 level (Hint: You should be performing the global test)

```
multi.linear.model <- lm(data = data.module4,`Prestige Score` ~`Education Level (years)`+`Income ($)`+`
anova <- anova(multi.linear.model)
summary1 <- summary(multi.linear.model)
anova
```

Setting up the multiple linear regression

```
## Analysis of Variance Table
##
## Response: Prestige Score
##                                   Df  Sum Sq Mean Sq  F value    Pr(>F)
## `Education Level (years)`          1 21608.4 21608.4 350.9741 < 2.2e-16 ***
## `Income ($)`                       1  2248.1  2248.1  36.5153 2.739e-08 ***
## `Percent of Workforce that are Women`  1     5.3     5.3   0.0858    0.7702
```

```
## Residuals                              98  6033.6    61.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary1
```

```
##
## Call:
## lm(formula = `Prestige Score` ~ `Education Level (years)` + `Income ($)` +
##     `Percent of Workforce that are Women`, data = data.module4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.8246  -5.3332  -0.1364   5.1587  17.5045
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        -6.7943342  3.2390886  -2.098   0.0385
## `Education Level (years)`           4.1866373  0.3887013  10.771  < 2e-16
## `Income ($)`                        0.0013136  0.0002778   4.729 7.58e-06
## `Percent of Workforce that are Women` -0.0089052  0.0304071  -0.293   0.7702
##
## (Intercept)                        *
## `Education Level (years)`          ***
## `Income ($)`                       ***
## `Percent of Workforce that are Women`
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.846 on 98 degrees of freedom
## Multiple R-squared:  0.7982, Adjusted R-squared:  0.792
## F-statistic: 129.2 on 3 and 98 DF,  p-value: < 2.2e-16
```

The values from ANOVA and summary Table will be using to test the hypothesis.

Setting up the hypotheses 1.Selecting the alpha level and setting up the hypothesis.

- H0:Beta(education)=Beta(income)=Beta(Woman workforce)=0 (education,income and percentage of women are not significant predictors of prestige score)
- H1:Beta(education) $\neq$ 0 and/or Beta(income) $\neq$ 0 and/or Beta(Woman workforce) $\neq$ 0 (at least one of the slope coefficients is different than 0 education and/or income and/or percentage of women are significant predictors/is a significant predictor of prestige score).
- $\alpha = 0.05$

2.Selecting the appropriate test statistics

- Selecting Global F-stat from summary table with df1=3 and df2=98. alternatively we can calculate the Res SS and Reg SS and calcuate F-value.

3. Stating Decision rule:

-Determining the appropriate value from the F-distribution with 3,102-3-1=98 degrees of freedom and associated with a right hand tail probability of alpha= 0.05(F-distribution has only one-sided curve) -Using software Critical value is:

```
qf(0.05,3,98,lower.tail = FALSE)
```

```
## [1] 2.697423
```

- Decision Rule : Reject H0 if F >=2.679
- Otherwise, do not reject H0

4. Compute the test statistics

```
#Fstatstic from summary table
table <- kable(summary(multi.linear.model)$fstatistic,digits = 1,align = "c")
table
```

	x
value	129.2
numdf	3.0
dendf	98.0

```
# Global Fstatstics can also be calculated manually
totalss <- sum((data.module4$`Prestige Score`- mean(data.module4$`Prestige Score`))^2)
regss <- sum((fitted(multi.linear.model)-mean(data.module4$`Prestige Score`))^2)
resiss <- sum((data.module4$`Prestige Score` - fitted(multi.linear.model))^2)
Fstatistic <- (regss/3)/(resiss/98)
pvalue.f <- pf(Fstatistic,df1 = 3,df2 = 98,lower.tail = FALSE)
Fstatistic
```

```
## [1] 129.1917
```

```
pvalue.f
```

```
## [1] 6.259061e-34
```

From above summary(multi.linear.model), it is found that global F-statistics 129.2 with df1=3 and df2=98 and p_value <0.001.

5.Conclusion:

Since F-Statistics for set of predictors Education,income and woman percentage are greater than critical value 2.679 and similarly p-values of these are way less than alpha value,we reject null hypothesis.This means at least one of the slopes coefficients is different than 0.Also,This can be interpreted at the alpha =0.05 level that education,income and % women workforce when taken together are significant predictors of prestige score.

#——————————————————————————-

Question 4

If the overall model was significant, summarize the information about the contribution of each variable separately at the same significance level as used for the overall model (no need to do a formal 5-step procedure for each one, just comment on the results of the tests). Provide interpretations for any estimates (of the slopes) that are significant.Calculate 95% confidence intervals for any estimates that are significant.

Answer: Checking association of each variable with dependent variable.Here, F-stat and p-values from summary table is obtained using following formula or can be obtained from table itself as follows. From ANOVA table:

- F(edu)= MS Reg(edu)/MS Res =21608.4/61.6 = 350.97 i.e greater than critical value 2.679.Similarly p-value <0.001 -> Significant at level 0.05.

-F(income)= MS Reg(income)/MS Res=2248.1/61.6 = 36,49 i.e greater than critical value 2.679.Similarly p-value <0.001 -> Significant at level 0.05.

- F(women) = MS Reg(women)/MS Res = 5.3/61.6 = 0.085 i.e less than critical value 2.679.Similarly p-value=0.7702 >0.05 -> Not significant at level 0.05.

From above calculation, it can be interpreted as beta(education) and beta(income) are significant at alpha = 0.05. From above summary table beta(education) = 4.1866373 and beta(income)=0.0013136.beta(education)=4.1866373 can be interpreted as for each additional year of education, prestige score is increase by 4.1866373,keeping other variables constant.Similarly, for each additional dollar of income, prestige score is increase by 0.0013136 keeping other variables constant.

Calculating 95% confidence intervals. Since education and income variables are significant at level 0.05,I will be calculating 95% confidence interval for the slopes for those variables.

```
confi.intv <- data.frame(confint(multi.linear.model,level = 0.95))[2:3,]
colnames(confi.intv) <- c("Beta_Lower_value","Beta_Upper_value")
kable(confi.intv,digits = 4,align = "c")
```

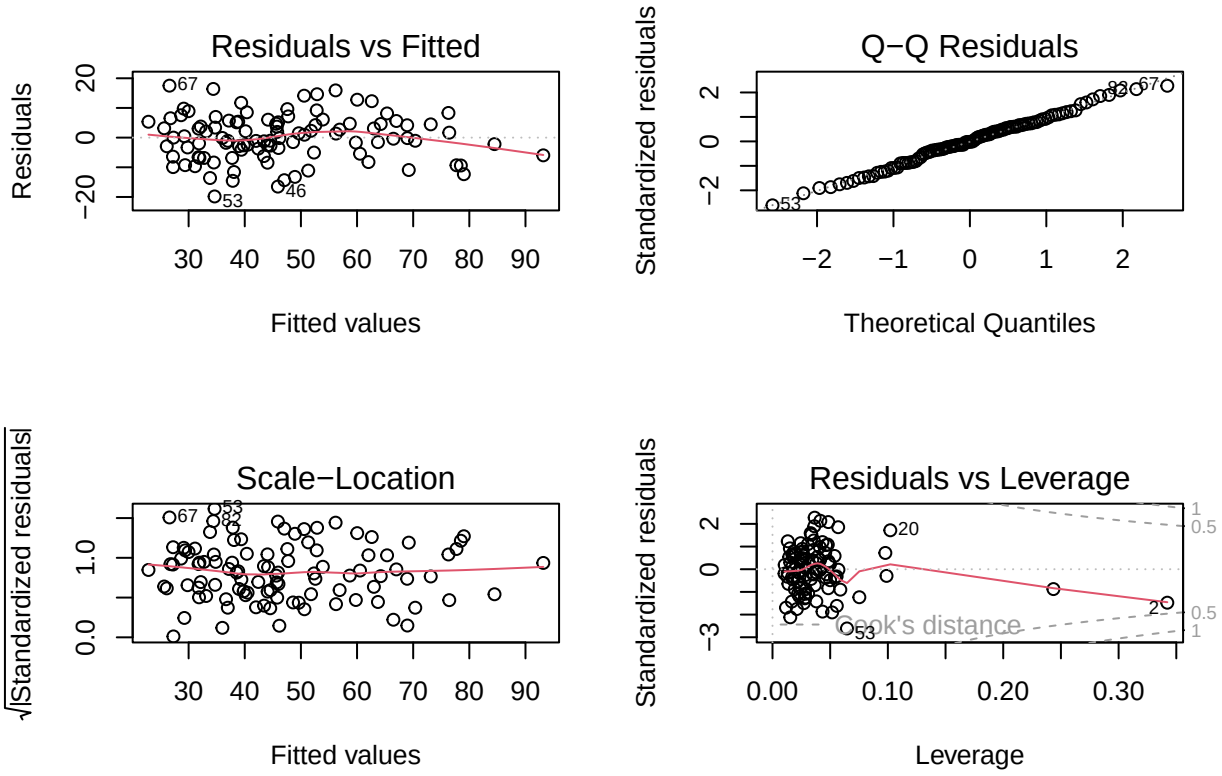
	Beta_Lower_value	Beta_Upper_value
'Education Level (years)'	3.4153	4.9580
'Income ($)'	0.0008	0.0019

Question 5:

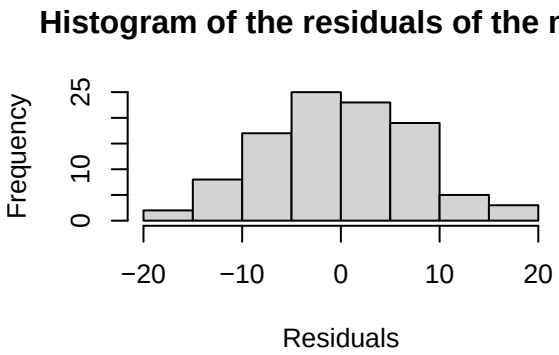
Generate a residual plot showing the fitted values from the regression against the residuals. Is the fit of the model reasonable? Are there any outliers or influence points?

Plot showing the fitted values from the regression against the residuals

```
#Scatter plot of residuals verses fitted values
par(mfrow=c(2,2))
plot(multi.linear.model)
```




```
#histogram of the residuals
hist(multi.linear.model$residuals,main = "Histogram of the residuals of the model",
     xlab = "Residuals")
```



From above plots, it seems that it follows all the assumptions to be reasonably fit to the model. Since plot of the residuals vs fitted values does not follow any pattern which indicate the association of the variables is approximately linear.We do assume that observations are independent. The Q-Q residuals plot is to check whether the residuals are normally distributed. since almost all points fall along a straight line, residuals can be consider approximately normally distributed(also can be seen on hist plot).Scale-Location plot checks constant variance of residuals.Since we can see a horizontal line with points are scattered randomly which indicates there is constant variance of residuals. From these explanations we can say that the fit of the model is reasonable.

Finding influencial points/Outliers I will be using cooks distance method to find any outliers/influence points.

```r
cook.distance.5 <- cooks.distance(multi.linear.model)
# Finding data points that could be the influential points using cooks distance.
# gold standard method of calculating the distance that are higher than 4 divided by number of total da

influencial.points2 <- which(cook.distance.5> 4/(nrow(data.module4)-3-1))
influencial.points2
```

```
##  2 20 24 27 29 53 54 67 82
##  2 20 24 27 29 53 54 67 82
```

```r
# checking influence of each point
set.seed(1234)
for (x in influencial.points2) {
 my.data2 <- data.module4 #copying main data
 my.data2 <- my.data2[-x,] # removing one row of data in each iteration
 my.model2 <- lm(data = my.data2,`Prestige Score`~ `Education Level (years)` + `Income ($)` + `Percent
 rsquared2 <- summary(my.model2)$r.squared
 cat("after removing row",x,"R-squared is",rsquared2,"\n")
}
```

```
## after removing row 2 R-squared is 0.7993014
## after removing row 20 R-squared is 0.7996898
## after removing row 24 R-squared is 0.7880909
## after removing row 27 R-squared is 0.8031303
## after removing row 29 R-squared is 0.8028339
## after removing row 53 R-squared is 0.8054886
## after removing row 54 R-squared is 0.8019492
## after removing row 67 R-squared is 0.8087674
## after removing row 82 R-squared is 0.8074671
```

```r
cat("while origial r-squared is",summary(multi.linear.model)$r.squared)
```

```
## while origial r-squared is 0.7981775
```

When finding the influential point using standard method(cooks' distance), it is found that there are bunch of points that might have high individual effect on model. However after removing each point from dataset and building new model and checking r-squared values, it is found that there is not much difference in R-squared values from original data set,.Among them, point 53 seems to have higher effect on mode than other however not much difference(0.7981775 vs 0.8054886)