Regression Models

Introduction: In this Assignment, We are going to use Regression models on our selected data. I have selected the USArrests data with the columns "Murder", "Assault", "UrbanPop" and "Rape".

Objective The main objective of this assignment to use and interpret the regression model on the selected data. I will do exploratory data analysis including classical univariate and bivariate analysis.

```
#importing all the libraries
library(dplyr)
library(tidyverse)library(ggplot2)
library(reshape2)
library(MASS)
library(Information)
library(gridExtra)
library(stringr)
library(caret)
library(car)
library(ggcorrplot)
library(hrbrthemes)


library(help="datasets")
USArrests

##                Murder Assault UrbanPop Rape
## Alabama          13.2     236       58 21.2
## Alaska           10.0     263       48 44.5
## Arizona           8.1     294       80 31.0
## Arkansas          8.8     190       50 19.5
## California        9.0     276       91 40.6
## Colorado          7.9     204       78 38.7
## Connecticut       3.3     110       77 11.1
## Delaware          5.9     238       72 15.8
## Florida          15.4     335       80 31.9
## Georgia          17.4     211       60 25.8
## Hawaii            5.3      46       83 20.2
## Idaho             2.6     120       54 14.2
## Illinois         10.4     249       83 24.0
## Indiana           7.2     113       65 21.0
## Iowa              2.2      56       57 11.3
## Kansas            6.0     115       66 18.0
## Kentucky          9.7     109       52 16.3
## Louisiana        15.4     249       66 22.2
## Maine             2.1      83       51  7.8
## Maryland         11.3     300       67 27.8
## Massachusetts     4.4     149       85 16.3
## Michigan         12.1     255       74 35.1
```

```
## Minnesota          2.7      72       66 14.9
## Mississippi       16.1     259       44 17.1
## Missouri           9.0     178       70 28.2
## Montana            6.0     109       53 16.4
## Nebraska           4.3     102       62 16.5
## Nevada            12.2     252       81 46.0
## New Hampshire      2.1      57       56  9.5
## New Jersey         7.4     159       89 18.8
## New Mexico        11.4     285       70 32.1
## New York          11.1     254       86 26.1
## North Carolina    13.0     337       45 16.1
## North Dakota       0.8      45       44  7.3
## Ohio               7.3     120       75 21.4
## Oklahoma           6.6     151       68 20.0
## Oregon             4.9     159       67 29.3
## Pennsylvania       6.3     106       72 14.9
## Rhode Island       3.4     174       87  8.3
## South Carolina    14.4     279       48 22.5
## South Dakota       3.8      86       45 12.8
## Tennessee         13.2     188       59 26.9
## Texas             12.7     201       80 25.5
## Utah               3.2     120       80 22.9
## Vermont            2.2      48       32 11.2
## Virginia           8.5     156       63 20.7
## Washington         4.0     145       73 26.2
## West Virginia      5.7      81       39  9.3
## Wisconsin          2.6      53       66 10.8
## Wyoming            6.8     161       60 15.6
```

```
View(USArrests)
```

Understanding the selected data: Before we begin with our models, It’s best to understand and analyze the variables. Now, I can see that there are no NA values and missed values in my data. Our data contains the number of Murder, Assault, and Rape cases for each of the states in the USA in 1973. It also contains the percentage of people living in urban areas.

```
str(USArrests)
```

```
## 'data.frame':    50 obs. of  4 variables:
##  $ Murder  : num  13.2 10 8.1 8.8 9 7.9 3.3 5.9 15.4 17.4 ...
##  $ Assault : int  236 263 294 190 276 204 110 238 335 211 ...
##  $ UrbanPop: int  58 48 80 50 91 78 77 72 80 60 ...
##  $ Rape    : num  21.2 44.5 31 19.5 40.6 38.7 11.1 15.8 31.9 25.8 ...
```

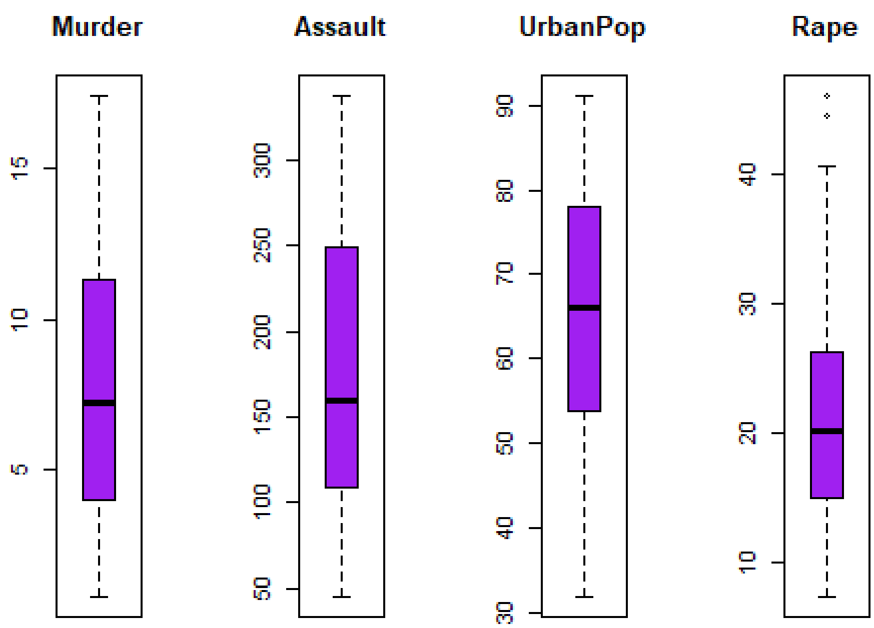
Now, We can see 50 observations refer to the USA states including 4 Variables i.e. Murder, Assault, UrbanPop, and Rape.

```
summary(USArrests)
```

```
##      Murder          Assault         UrbanPop          Rape      
##  Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30
```

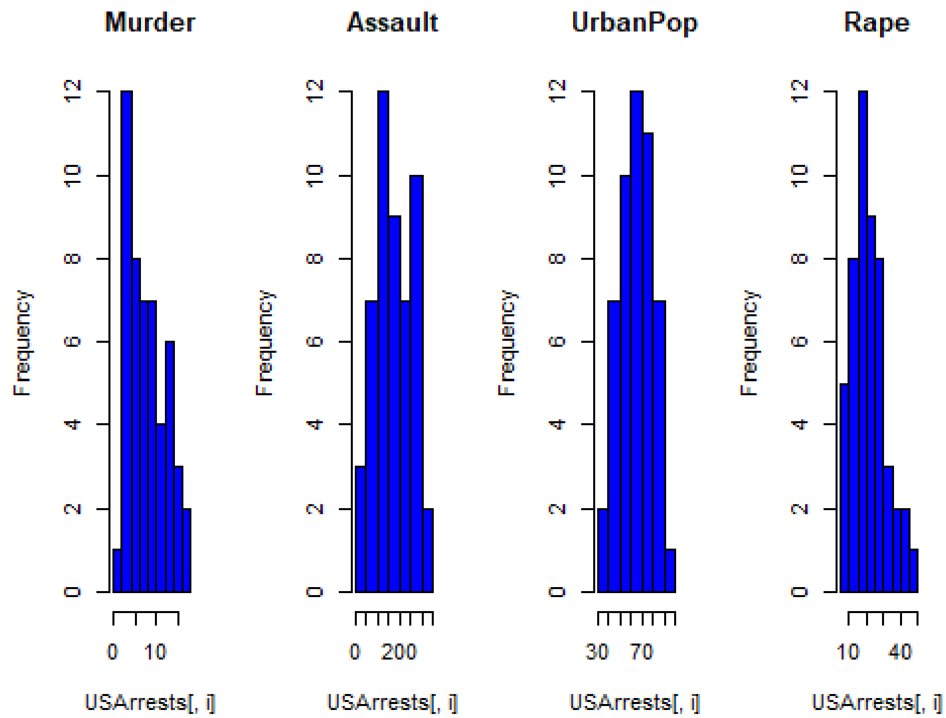
```
##   1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07
##   Median : 7.250   Median :159.0   Median :66.00   Median :20.10
##   Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23
##   3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18
##   Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00
```

```
#Box Plot
par(mfrow=c(1,4))
for(i in 1:4) {
    boxplot(USArrests[,i], main=names(USArrests)[i],
   col = c("purple"))
}
```



BOX PLOT: Box plot displays the distribution of data on a five-number summary(“min”,Q1, median, Q3, and "maximum) It also shows the outliers. The two hinges are the version of the first and third quartile.

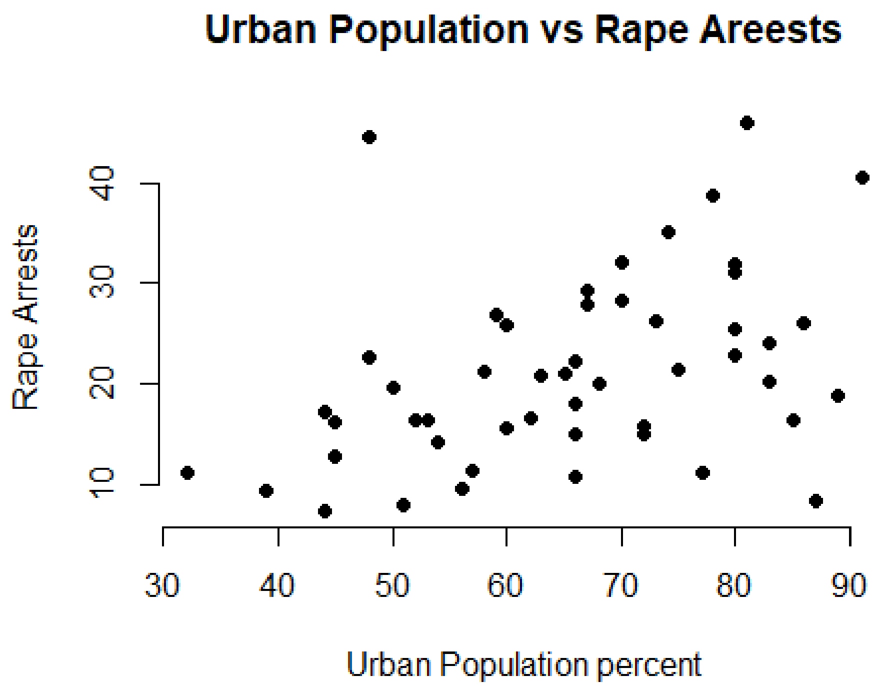
```
#Histogram
par(mfrow=c(1,4))
for(i in 1:4) {
    hist(USArrests[,i], main=names(USArrests)[i],
   col = c("Blue"))
  }
```



Histogram: We can see the histogram of Rape Arrests in the graphs. As we can see that rape arrests are highest under the category of 15-25 whereas lowest in 42-50.

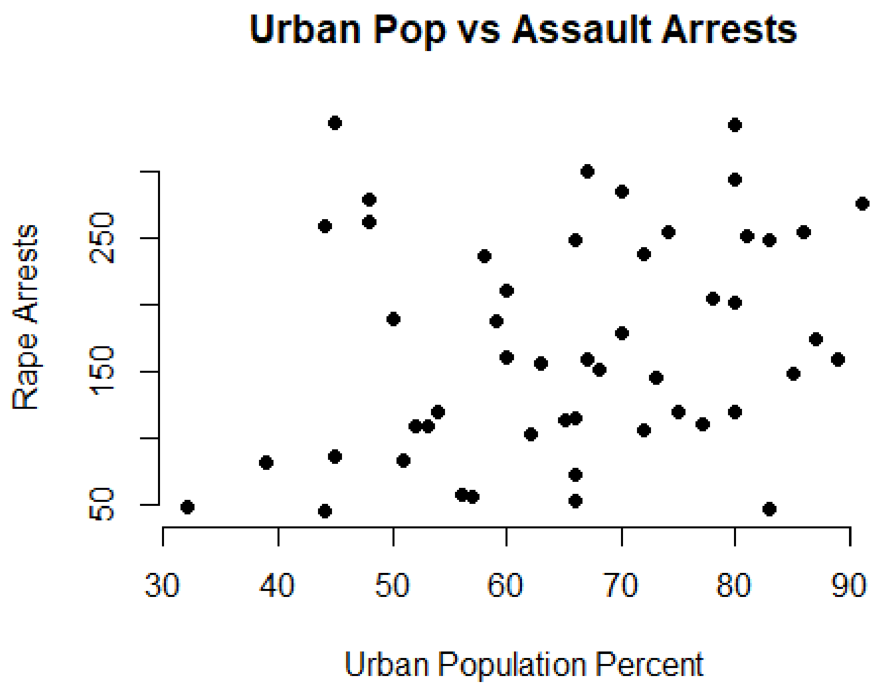
```
#scatter plot 1

x <- USArrests$UrbanPop
y <- USArrests$Rape
# Plot with main and axis titles
# Change point shape (pch = 19) and remove frame.
plot(x, y, main = "Urban Population vs Rape Areests",
     xlab = "Urban Population percent", ylab = "Rape Arrests",
     pch = 19, frame = FALSE)
```



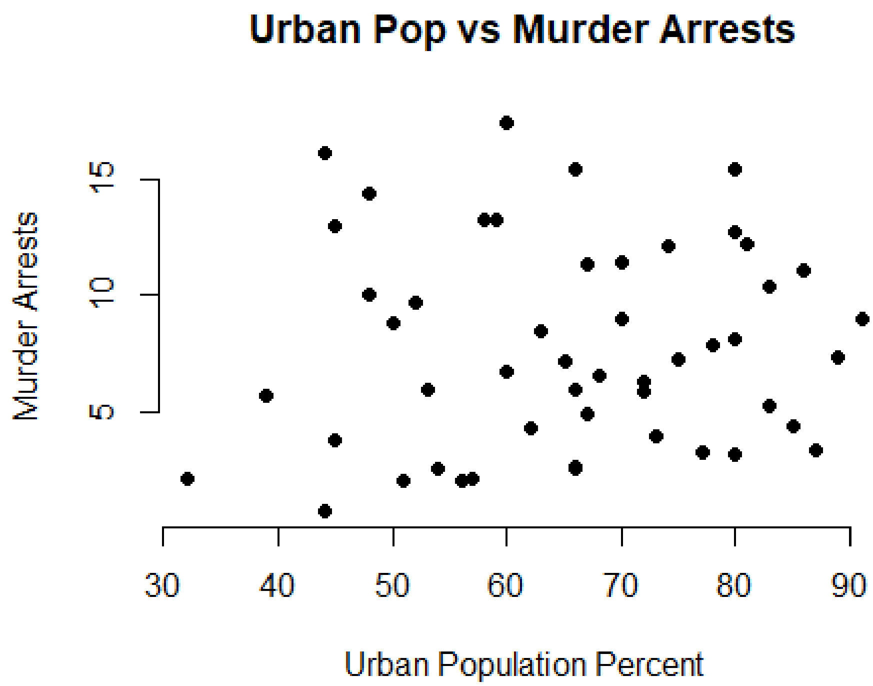

```
#scatter plot 2

x <- USArrests$UrbanPop
y <- USArrests$Assault
# Plot with main and axis titles
# Change point shape (pch = 19) and remove frame.
plot(x, y, main = "Urban Pop vs Assault Arrests",
     xlab = "Urban Population Percent", ylab = "Rape Arrests",
     pch = 19, frame = FALSE)
```



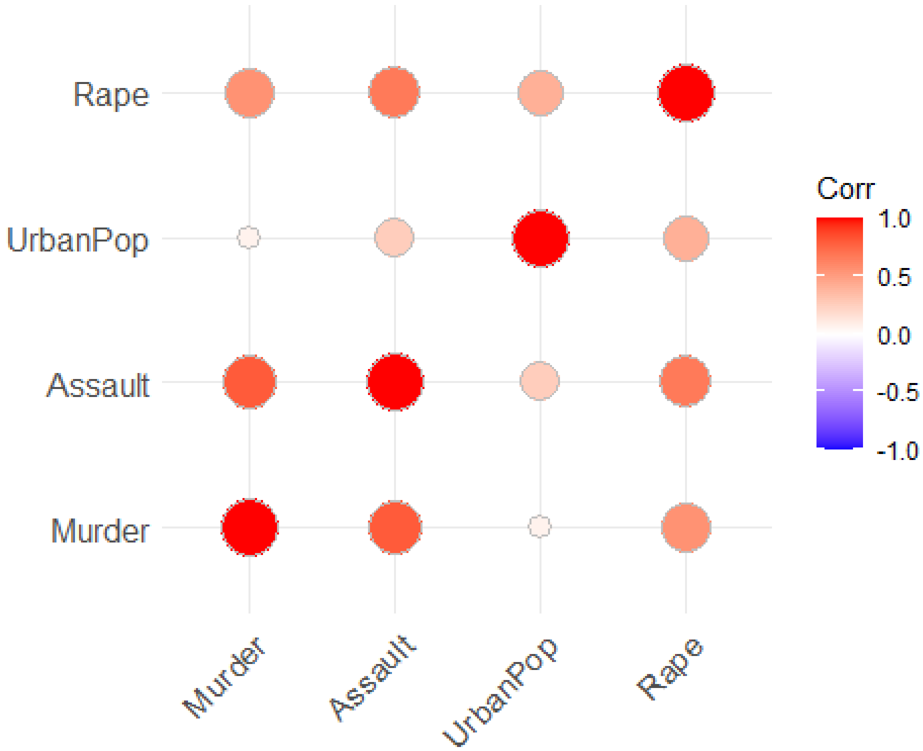

```r
#scatter plot 3

x <- USArrests$UrbanPop
y <- USArrests$Murder
# Plot with main and axis titles
# Change point shape (pch = 19) and remove frame.
plot(x, y, main = "Urban Pop vs Murder Arrests",
     xlab = "Urban Population Percent", ylab = "Murder Arrests",
     pch = 19, frame = FALSE)
```



Scatter plots show the relationship between two variables. I have plotted three scatter plots using UrbanPop as an explanatory variable and others as response variables. According to the selected data, we can see if the Assault/Murder/Rape variable increases/decreases in a straight line as UrbanPop variable increases/decreases that can provide us the evidence of the relationship between two variable. In our scatter plot 1, 2 & 3, It is difficult to observe if there is any linear relationship between the data.

```
#correlation of data
correlation_data <- cor(USArrests[,1:4])
ggcorrplot(correlation_data, method = "circle")
```



Finding the correlation between the features or predictors is important in the model because we can use the correlation to make the predictions. Correlation takes values between 1 to -1. We can see in the screenshot above that 3 variables are highly correlated i.e. Murder, Rape , and Assault whereas UrbanPop is less correlated. According to our data if murder arrests increase/decrease then the Assault and Rape will also increase/decrease that are showing high correlation with each other. UrbanPop is closer to 0 which shows the weak relationship with variables.

```
#SIMPLE LINEAR REGRESSION
# build linear regression model on full data
linearMod <- lm(Rape ~ UrbanPop, data=USArrests)
print(linearMod)

##
## Call:
## lm(formula = Rape ~ UrbanPop, data = USArrests)
##
## Coefficients:
## (Intercept)     UrbanPop
##      3.7871       0.2662
```

Now, We have intercept and slope, we can say that every single %age increase in the Urban state population, the number of Rape cases increases by 0.266. Our simple regression model is y=β0 + β1x or y=3.7871 + 0.2662x

```
summary(linearMod)
```

```
##
## Call:
## lm(formula = Rape ~ UrbanPop, data = USArrests)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.644  -5.476  -1.216   5.885  27.937
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.78707    5.71128   0.663    0.510
## UrbanPop     0.26617    0.08513   3.127    0.003 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.626 on 48 degrees of freedom
## Multiple R-squared:  0.1692, Adjusted R-squared:  0.1519
## F-statistic: 9.776 on 1 and 48 DF,  p-value: 0.003001
```

Now, We have found the p-value, F-statistic, Residual standard error, Adjusted R square. P-value is very important in analysis, we can check if our linear model is statistically significant or not that generally when the p-value is less than 0.05. F-statistic is basically on the ratio of mean squares. The more the value, the better the model. If we check the Mean square error in the model that is large 8.626 means the regression line is not precise to the data sets. R2 represents the correlation between the variables that is 0.1692. This agrees with the result of MSE as well. The higher the R2, the better the model.

```
#MULTIPLE LINEAR REGRESSION
# build linear regression model on full data
linearMod1 <- lm(UrbanPop ~ Murder+ Rape+ Assault, data=USArrests)
print(linearMod1)
```

```
##
## Call:
## lm(formula = UrbanPop ~ Murder + Rape + Assault, data = USArrests)
##
## Coefficients:
## (Intercept)       Murder         Rape      Assault
##     52.8419      -1.4115       0.6984       0.0519
```

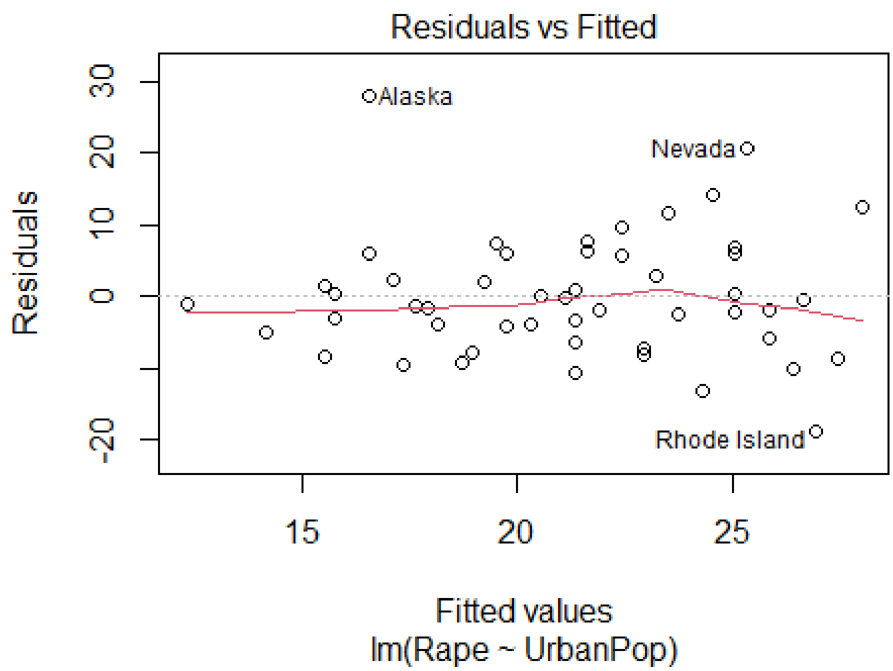
```
summary(linearMod1)
```

```
##
## Call:
## lm(formula = UrbanPop ~ Murder + Rape + Assault, data = USArrests)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.456  -6.950   0.077   7.770  25.221
##
```

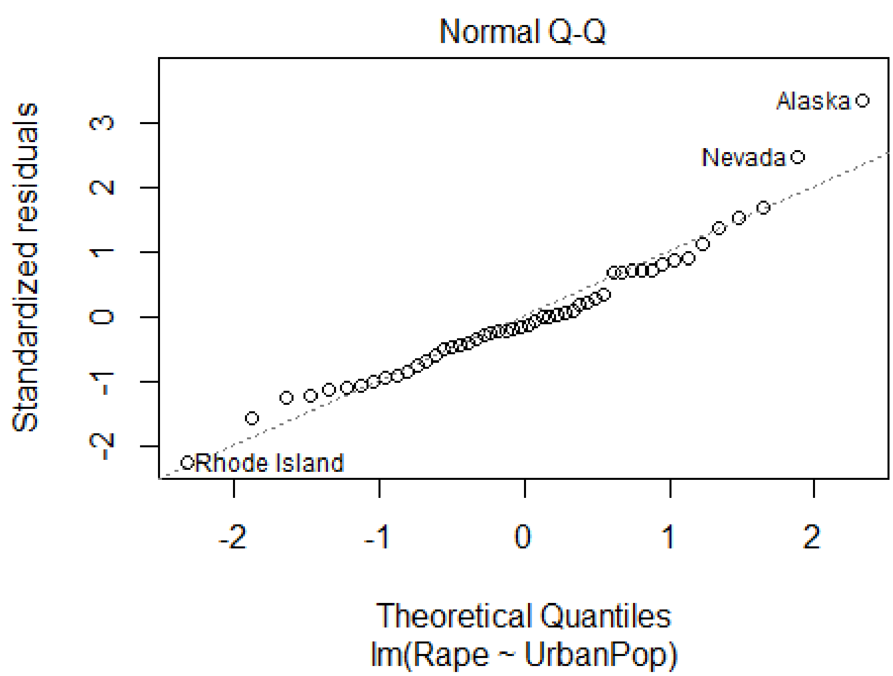
```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 52.84187    4.82483  10.952 2.09e-14 ***
## Murder      -1.41154    0.71954  -1.962   0.0559 .
## Rape         0.69841    0.26776   2.608   0.0122 *
## Assault      0.05190    0.04161   1.247   0.2186
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.08 on 46 degrees of freedom
## Multiple R-squared:  0.2337, Adjusted R-squared:  0.1837
## F-statistic: 4.676 on 3 and 46 DF,  p-value: 0.006208
```

Now, We can compare both the models using R2 and MSE and can check which model is better. If we check the Residual standard error, It is more in multiple regression and less in Simple regression. Both the models have p value less than 0.05.

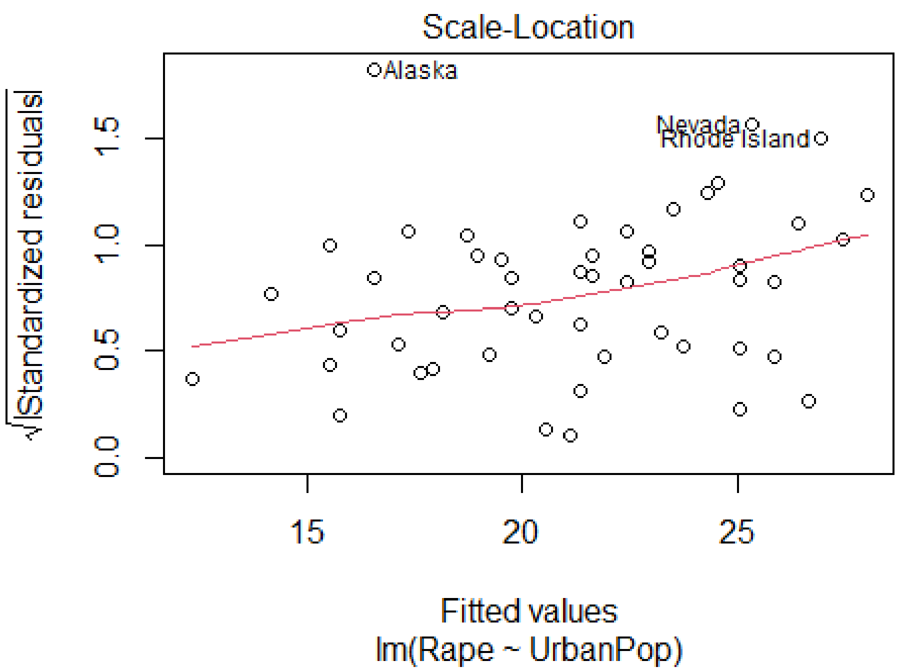
```
#Plotiing Q-Q,.....
plot(linearMod)
```

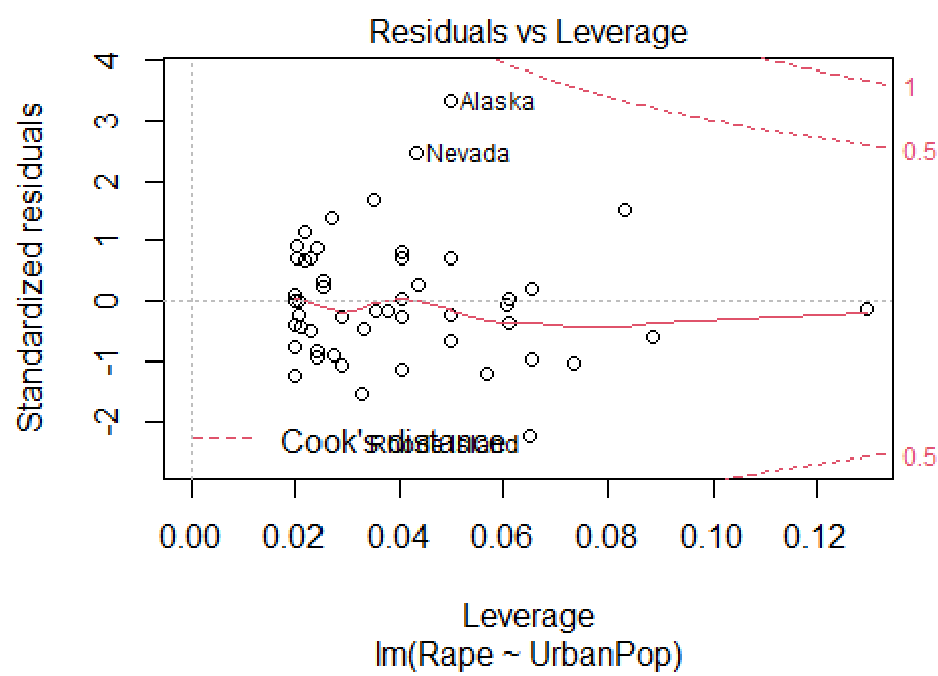


Residuals vs Fitted
Alaska
Nevada
Rhode Island
Residuals
30
20
10
0
-20
15
20
25
Fitted values
lm(Rape ~ UrbanPop)



Normal Q-Q
Alaska
Nevada
Rhode Island
Standardized residuals
3
2
1
0
-1
-2
-2
-1
0
1
2
Theoretical Quantiles
lm(Rape ~ UrbanPop)



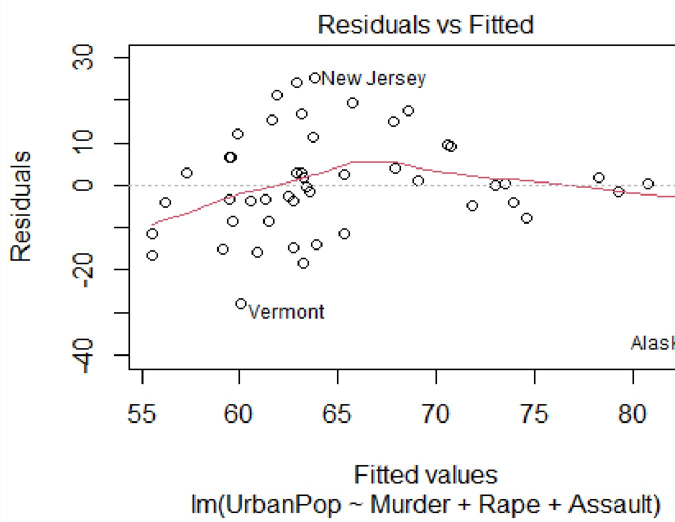


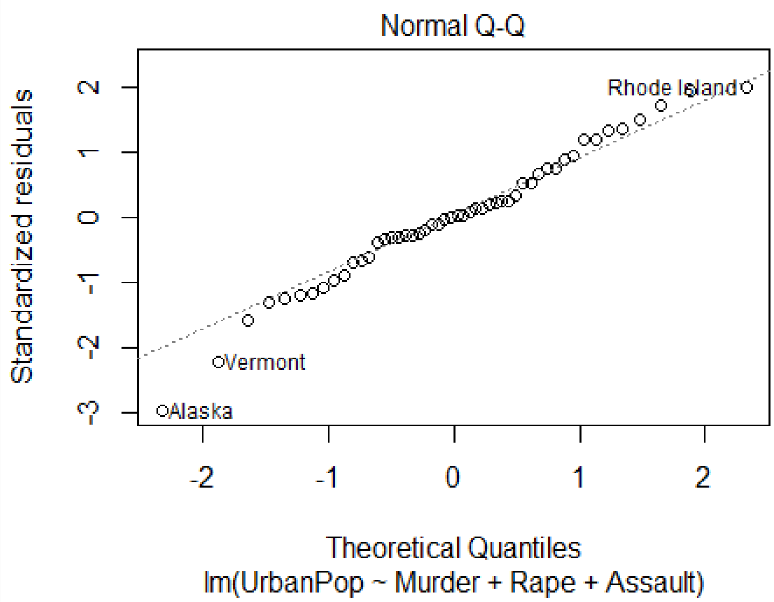
Visualizing model fit is a very important aspect to see how the regression works in R. Main feature to recognize is when residuals are normally distributed and the mean of residuals is zero.

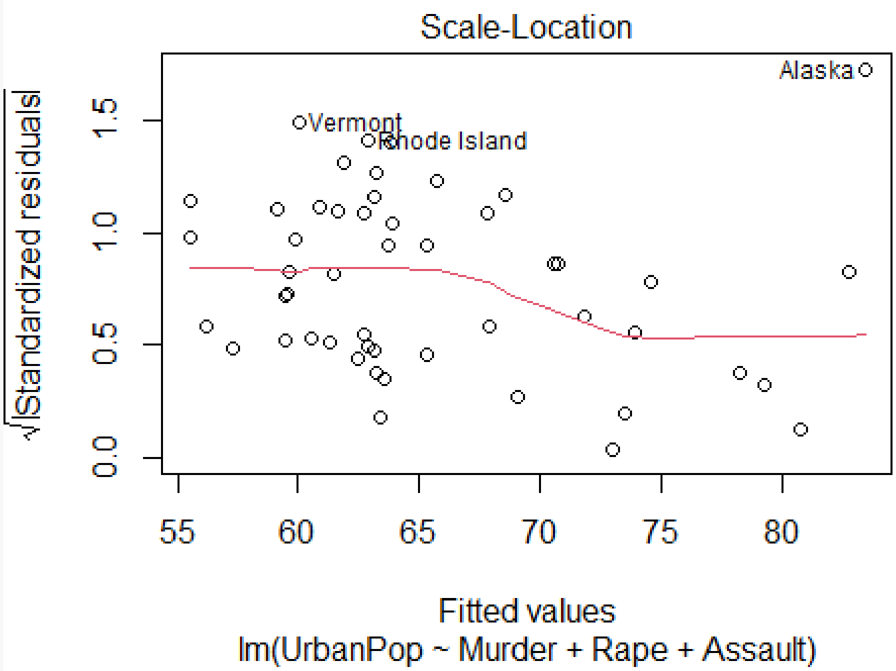
From the residuals vs fitted graph, it is a scatter plot of residuals on the y-axis and fitted values (estimated responses) on the x-axis. we observe that the plot detects linearity while checking the percentage of the UrbanPop population with the rape category. The residual variance is constant across urban pop percent, so we can say the residuals are normally distributed.

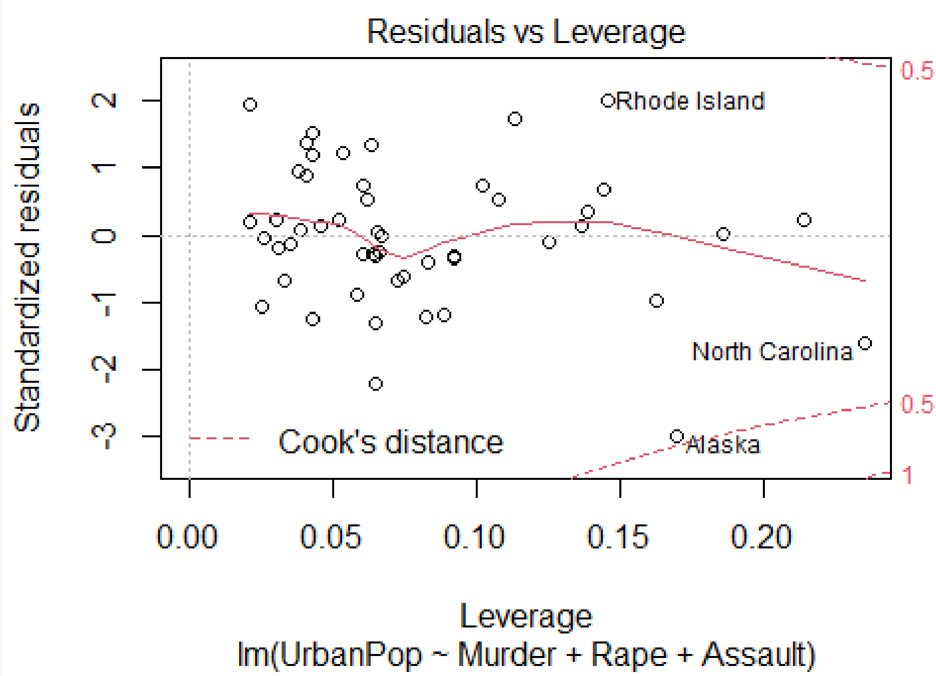
In the Q-Q plot, Residuals are following the straight path that shows residuals are normally distributed.

```
plot(linearMod1)
```










```
#Predictions and Confidence Interval
newdata <- data.frame(UrbanPop=70)
confy <- predict(linearMod, newdata, interval="confidence", level = .95)
confy
```

```
##        fit      lwr      upr
## 1 22.41913 19.85036 24.98789
```

We have found the confidence interval for Rape Arrests for a state that has 70 percent urban population. This shows us that 95% of the samples will create mean of rape arrests between 19.85 to 24.98 for a state which has the 70% urban population.

```
predy <- predict(linearMod, newdata, interval="predict", level=.95)
predy

##        fit      lwr      upr
## 1 22.41913 4.886679 39.95158
```

A prediction interval is wider than the confidence because it must account for both the uncertainty in estimating the population means, plus the random variation of the individual values. We have again used the 70% urban Population.

Now, Both the intervals need to be centered at the same point. We will check the accuracy.

```
confy[1] == predy[1]

## [1] TRUE
```

Hence, both intervals are centered at one point.

Conclusion: In this Assignment, We built the Linear regression model on the selected dataset USArrests that includes the data of the number of rape, murder, and Assault arrests in the 50 states of the USA. We have seen that UrbanPop shows a weak relationship with other variables.