ASSIGNMENT 5 – Linear Regression

In this assignment we will implement a linear regression model. The dataset we will be working with will be **countyComplete.csv**. We will be using one variable i.e **poverty** as our response variable and one other variable as our explanatory variable.

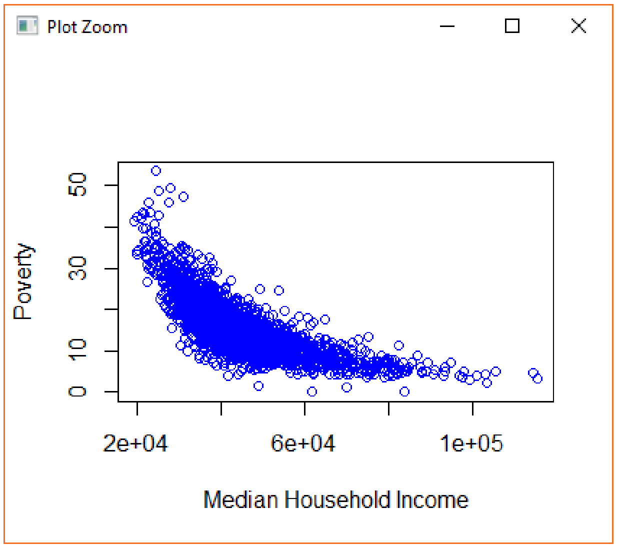
We first read the file by using the read.csv command.

require(mosaic)
county = read.csv("countyComplete.csv")

We will now use the **median population income** variable as our explanatory variable for the poverty estimate.

Before looking at the linear model we first find the trend between the two variables by using a scatterplot. The code to execute it is given below.

plot(county$median_household_income,county$poverty, col = "Blue", type = "p")



The plot clearly shows a negative trend that is if the median household income increases we can observe a corresponding decrease in poverty.

We now calculate the correlation between the variables to find whether our observation from the graph holds or not. We calculate the correlation between the two variables using the below code.

cor(county$median_household_income,county$poverty)

```
[1] -0.7416272
```

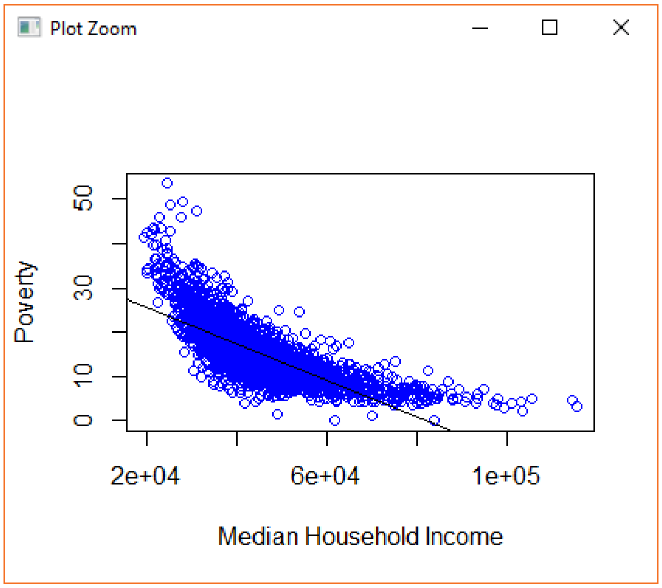
We get the above output after running the code. This shows a negative correlation of just above 74% between the two variables. This means a initial observation hold true.

We can now move forward with implementing the linear model. In R, we use the **lm()** command to generate a linear model for the variables. We generate the linear model in the following way.

model = lm(county$poverty~county$median_household_income)

We now graph the regression line on the scatterplot we created before. To do this we run the code below.

abline(model)



The image above shows the regression line for our model. Certainly, majority of the data point are around the regression line. We now use the **summary()** function to get the information stored in our model.

summary(model)

```
Call:
lm(formula = county$poverty ~ county$median_household_income)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.4094  -2.8108  -0.4418   2.1569  29.8510 

Coefficients:
                                 Estimate Std. Error t value
(Intercept)                     3.365e+01  3.027e-01  111.15
county$median_household_income -4.100e-04  6.617e-06  -61.96
                               Pr(>|t|)    
(Intercept)                      <2e-16 ***
county$median_household_income   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.283 on 3141 degrees of freedom
Multiple R-squared:   0.55,	Adjusted R-squared:  0.5499 
F-statistic:  3839 on 1 and 3141 DF,  p-value: < 2.2e-16
```

The above output shows the information contained in our model.

The first item shows the Call functionality. This only points out to our formula for generating the linear model. The next item are the residuals.
The residuals are the difference between the actual value observed versus the value predicted by the model. If the residuals are symmetrical with the median around 0, it means the model is a good fit. Here we can observe that the that the residuals are not uniformly distributed n the dataset. There is a certain skewness in the values of the residuals. Though the median is close to zero the uniformity of the residual distributed signifies a better fit.

```
Coefficients:
                               Estimate
(Intercept)                   3.365e+01
county$median_household_income -4.100e-04
```

The next section is the coefficient outputs. The first section is the intercept for the model. The intercept in our case is the expected value of the percentage below poverty level by taking the average median household income. Here its comes out to be more than 33%. The second value in the estimate column shows how much change can we observe in the poverty variable given a unit change in the median household income. Here it comes out to be 0.00040% decrease in the percentage variable which is minute.

```
Std. Error
3.365e+01
-4.100e-04
```

The next column in the coefficient is the standard error. A standard error is the estimate of the difference in the coefficient values if we run the model repeatedly. It is a measure of the error in our estimates of the coefficients. We want them to be lower as compared to our coefficients. In this case for the expected poverty percentage it is 0.302 %. Standard errors are also used to calculate confidence interval for hypothesis testing.

```
Multiple R-squared:   0.55,   Adjusted R-squared:  0.5499
```

We now come to the R-squared statistic which provides us a measure of how well the model is fitting the actual data. It takes in the form of how much variance in our response variable is explained by our explanatory variable. It always is in the range of 0 to 1. Here in our case it is 0.55 which is 55%. Our model is able to explain 55% of the variance in the poverty percentage variable.
Recall that the amount of variance explained in our group-based models came out to be 48%. This is a slight improvement over that model. As we increase the variables in the linear model we can get an improvement in the R-square statistic. For that we need to look into the adjusted R-square statistic as it takes into account the increase in the number of variables.

```
F-statistic:  3839 on 1 and 3141 DF
```

Lastly we explain the F-statistic which shows whether there is a relationship between explanatory variable and the response variable. It should be large value in case of many data points to establish that there is a relationship between the two variables. Here we have the value of 3839 which is sufficient to establish that there is a relationship between the poverty and median household income.

Finally, we try to use more than one variable to explain our predictor poverty percentage variable in order to look at multiple linear regression. We will only provide the summary of the model to show the improvement in explaining the variance of our predictor variable. We use home ownership as an additional explanatory variable. The code along with output is below.

model1 = lm(county$poverty~county$median_household_income+county$home_ownership)

summary(model1)

```
Call:
lm(formula = county$poverty ~ county$median_household_income + 
    county$home_ownership)

Residuals:
     Min       1Q   Median       3Q      Max 
-28.7389  -2.3607  -0.4395   1.8679  25.1498 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     4.955e+01  6.939e-01   71.41   <2e-16 ***
county$median_household_income -3.998e-04  6.058e-06  -66.00   <2e-16 ***
county$home_ownership          -2.232e-01  8.932e-03  -24.99   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.912 on 3140 degrees of freedom
Multiple R-squared:  0.6247,	Adjusted R-squared:  0.6244 
F-statistic:  2613 on 2 and 3140 DF,  p-value: < 2.2e-16
```

Here we can see that this model better explains the variance in the poverty percentage variable. The adjusted R-square value is almost 65.5% which is better than the previous 55% we had while using only the median household income.