Project Report

SYED AMEER.S

data analysis program.

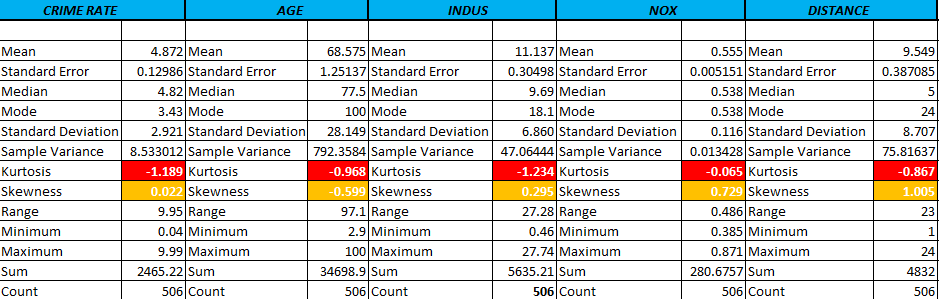
2023

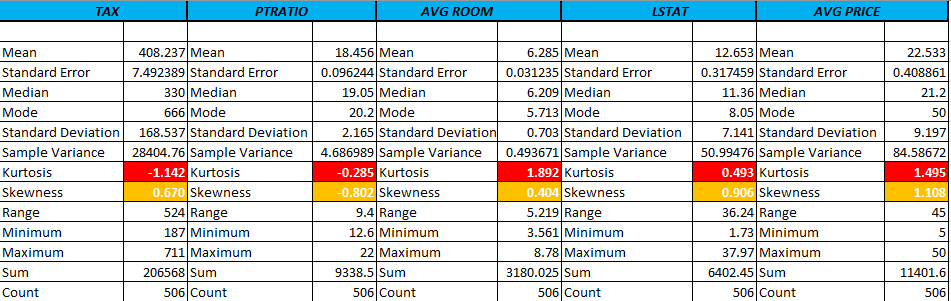
**OBJECTIVES**

**AND**

**INTERPRETATIONS**

**1)Generate the summary statistics for each variable in the table. Write down your observation. (Use Data Analysis tool pack).**

****

****

From the summary statistics, I am using the skewness and kurtosis to find out how the data distributed actually.

* 1. **What is Skewness?**

Skewness is used to find in which direction frequency distribution has a departure from symmetry. It is the measure of asymmetry that occurs when our data deviates from the norm. Skewness is basically of two types.

* Positive skewed (Right tailed)
* Negative skewed (Left tailed)

Rules:

* (-0.5 to +0.5) - nearly symmetrical
* (-0.5 to -1 (or) 0.5 to 1) - negative skewed (or) positive skewed
* (< -1 (or) > +1) – extremely negative (or) positive skewed

**1.2. What is Kurtosis?**

Kurtosis is defined as a parameter of the relative sharpness of the curvature of possible distribution curve. It is used to indicate the flatness or height of a distribution curve and usually measures the tails with or without distribution. Kurtosis basically of three types.

* Leptokurtic (kurtosis more than normal distribution) (kurtosis > 3)
* Mesokurtic (kurtosis same as the normal distribution) (kurtosis = 3)
* Platykurtic (kurtosis less than normal distribution) (kurtosis < 3).

**Skewness Interpretation:**

1.For Crime rate (0.022), the skewness is distributed positively. The distributed value is nearly symmetrical.

2.For Age (-0.599), The distributed value is in between the -0.5 to -1, so we considered as negatively skewed.

3.For Indus (0.295), the skewness is distributed positively. The distributed value is nearly symmetrical

4. For NOX (0.729), The distributed value is in between the 0.5 to 1, so we considered as positively skewed.

5. For Distance (1.005), The distributed value is greater than 1, so we considered as extremely positive skewed.

6.For Tax (0.670), The distributed value is in between the 0.5 to 1, so we considered as positively skewed.

7.For P-Ratio (-0.802), The distributed value is in between the -0.5 to -1 , so we considered as negatively skewed.

8.For AverageRoom (0.404), the skewness is distributed positively. The distributed value is nearly symmetrical.

9.For LSTAT (0.906), The distributed value is in between the 0.5 to 1, so we considered as positively skewed.

10.For Average Price (1.108), The distributed value greater than 1, so we considered as extremely positive skewed.

**Kurtosis Interpretation:**

1. CrimeRate - (-)1.189

2. Age - (-)0.968

3. Indus - (-)1.234

4. Nox - (-)0.065

5. Distance - (-)0.867

6. Tax - (-)1.142

7. P-Ratio - (-)0.285

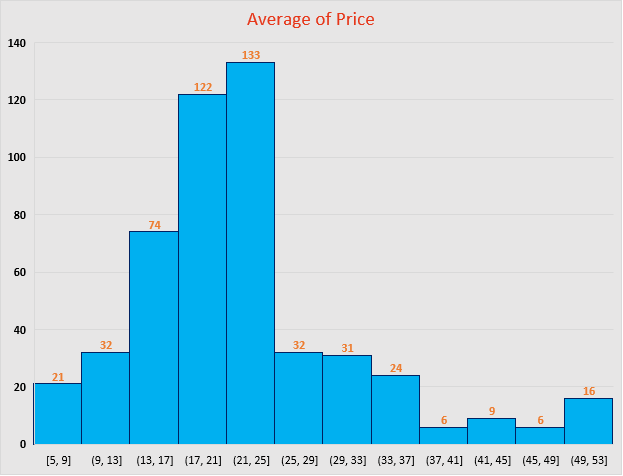
8. AvgRoom - 1.892

9. LSTAT - 0.493

10.AvgPrice - 1.495

The values of kurtosis are most of them negative in our data and the range also in between the -1.2 to +1.9, which is equal to zero, but not exactly equals to zero. So, from the above values we considered as mesokurtic curve approximately, also represents the curve are distributed normally.

2**) Plot a histogram of the AvgPrice variable. What do you infer?**



**Fig.2.1.**

**Interpretation:**

Histogram describes that most of the houses in Boston having an average price (in $1000’s) between the 17k to 25k range which is shown in the fig.2.1. The distribution of the this data is skewed positively. It means that the structure is extended to the upper right side, means a good inclination. Sometimes the results would be slightly change based upon the bin size. Our bin size is 4.

**3) Compute the covariance matrix. Share your observations.**

**3.1. What is Covariance?**

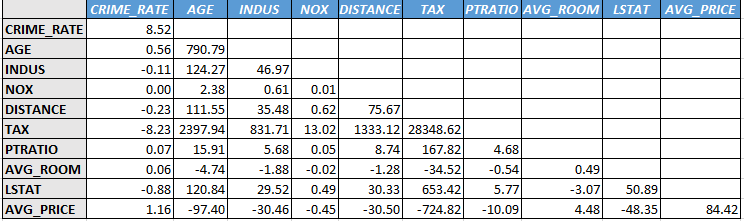
Covariance in statistics is used to evaluate the relationship between two variables. It is actually a measure of the variance among two variables. Covariance is estimated in units and is determined by multiplying the units of the two given variables.

* Positive covariance:

If the covariance for any two variables is positive, that means, both the variables move in same direction. Here the variables show similar behaviour. That means, if the variables (greater or lesser) of one variable corresponds to the value of another variable, then they are said to be in positive covariance.

* Negative covariance:

If the covariance for any two variables is negative, that means, both the variables move in the opposite direction. It is the opposite case of positive covariance, where greater values of one variable correspond to lesser values of another variable and vice-versa.

****

**Interpretation:**

It’s statistically used to find the direction of variables. But correlation is the best way to interpret the behaviour of two variables. Because, the accuracy of interpretation is given by the correlation are better than compared to the covariance and also values of covariance change (or) differ based on the measure of the data and scale of the data . In our data the most of the data covariance between the two variables moves in the same direction.

**4) Create a correlation matrix of all the variables (Use Data analysis tool pack).**

**4.1. What is Correlation?**

Correlation coefficient is a statistical concept, which helps in establishing a relation between predicted and actual values obtained in a statistical experiment. The calculated value of the correlation coefficient explains the exactness between the predicted and actual values. Types of correlation as given below.

* Positive Correlation:( +1)

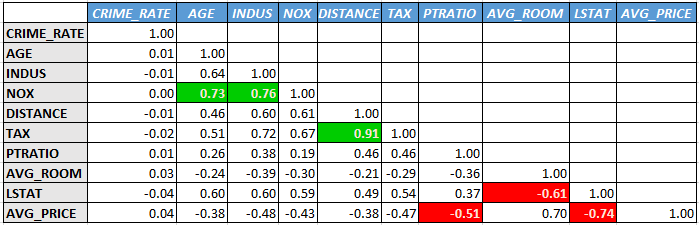
The value of one variable increases linearly with increase in another variable. This indicates a similar relation between the both variables. So its correlation coefficient would be positive or 1 in this case.

* Negative Correlation:(-1)

When there is a decrease in values of one variable with increase in values of another variable. In that case, correlation coefficient would be negative.

* Zero Correlation:(close to 0 or 0)

There is one more situation when there is no specific relation between two variables.

****

**Interpretation:**

1. **Which are the top 3 positively correlated pairs:**

1. Distance vs Tax (0.91) - a perfect positive correlation.

2. Indus vs Nox (0.76) - a high positive correlation.

3. Age vs Nox (0.73) - a high positive correlation.

1. **Which are the top 3 negatively correlated pairs:**

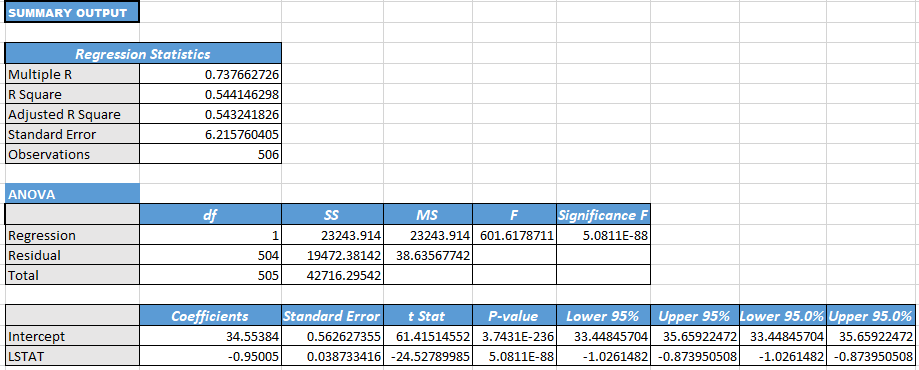
1. LSTAT vs AvgPrice (-0.74) – a high negative correlation.

2. AvgRoom vs LSAT (-0.61) – a high negative correlation.

3. P-Ratio vs AvgPrice (-0.51) – a moderate negative correlation.

**5) Build an initial regression model with AVG\_PRICE as ‘y’ (Dependent variable) and LSTAT variable as Independent Variable. Generate the residual plot.**

**a)** **What do you infer from the Regression Summary output in terms of variance explained, coefficient value, Intercept, and the Residual plot?**

****

**Interpretation (by coefficient value, Intercept**):

Regression Equation,

y = 34.554 – 0.950\*LSTAT

**β1** = It means that, if the 1unit(percentage) of LSTAT increases then it is affects the AvgPrice by decreasing -0.905(in $1000’s).

**β0 =** The value for the intercept term in this model is 34.554. This would mean the average of AvgPrice ($1000’s) when their x variables (predictor variables) are equal to zero. The intercept just doesn’t have any meaningful for this model.

**Interpretation (by residual plot**):

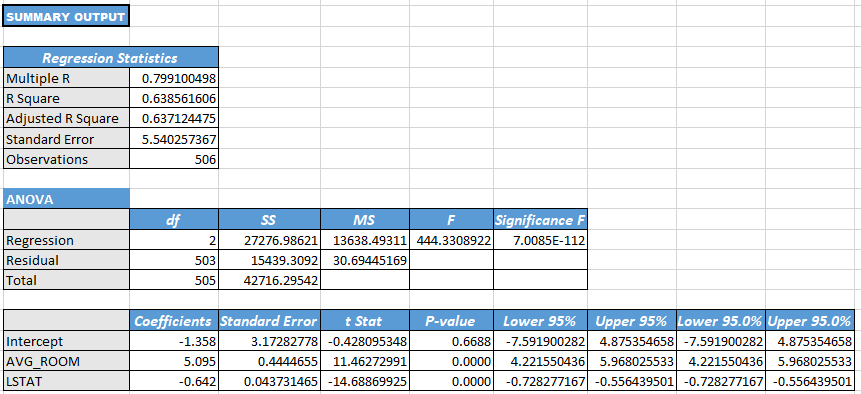
**Fig.5.1**

In this residual plot, there is a pattern that can be described. The data points are above the residuals=0 line near [1.73,17.08] which is shown in fig5.1. Then the points are scattered randomly around the residual = 0 line. We can conclude that the linear model is appropriate for modelling this data.

**b) Is LSTAT variable significant for the analysis based on your model?**

A p-value for the LSTAT is 5.0811E-88 which is equal to 0.00, means, the p - value is less than the alpha level 0.05, so we reject the null hypothesis and accept the alternative hypothesis. It means that there is some relationship between the LSTAT and AvgPrice. It is statistically significant for the analysis.

**6)** **Build a new Regression model including LSTAT and AVG\_ROOM together as independent variables and AVG\_PRICE as dependent variable.**

****

**a)** **Write the Regression equation. If a new house in this locality has 7 rooms (on an average) and has a value of 20 for L-STAT, then what will be the value of AVG\_PRICE? How does it compare to the company quoting a value of 30000 USD for this locality? Is the company Overcharging/ Undercharging?**

Regression Equation,

y = -1.358 + 5.095\*AvgRoom – 0.642\*LSTAT

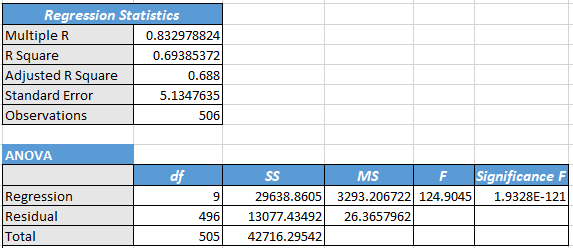
y= Avg price, β1= 5.095, β2= (-)0.642, X1= AvgRoom= 7, X2= LSTAT= 20

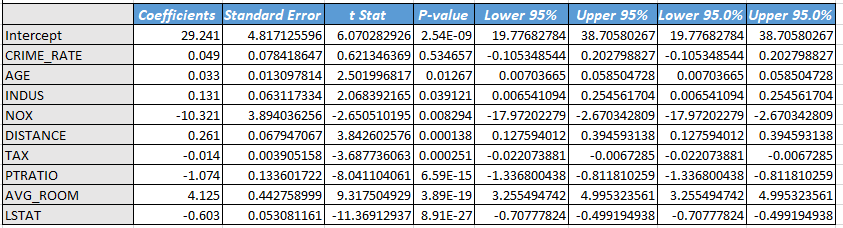
y = -1.358 + 5.095\*7 – 0.642\*20 **y = 21.46**

The AvgPrice value while after given the values for AvgRoom as 7 and LSTAT as 20 is $21460. Comparing this value with the company quoting value of $30000 for this locality clearly shows the company is overcharging for houses.

**b) Is the performance of this model better than the previous model you built in Question 5? Compare in terms of adjusted R-square and explain.**

In the first model (Question 5) were we simply had % lower status of the population (LSTAT) as our only predictor, our adjusted R squared is 54.32%. However, for the second model (Question 6) were we have the two variables [% lower status of the population (LSTAT) and average number of rooms per house (AVG\_ROOM)], and our adjusted R squared has increased to 63.71%. When any variable is added to the model, the R squared increases. However, if the added variable doesn't really provide any new information or is completely unrelated, the adjusted R squared does not increase. First, adjusted R squared applies a penalty for the number of predictors included in the model. And second, we choose models with higher adjusted R squared over others. So, in this case the second model (Question 6) has the higher adjusted R squared value. So, we can conclude that the performance of this model (Question 6) better than the previous model (Question 5).

**7)** **Build another Regression model with all variables where AVG\_PRICE alone be the Dependent Variable and all the other variables are independent. Interpret the output in terms of adjusted R-square, coefficient and Intercept values. Explain the significance of each independent variable with respect to AVG\_PRICE.**



**Interpretation (by adjusted R-square):**

The adjusted R square is similar to R square it is used to measure the goodness of fit. In our data the adjusted R square value is 0.688, this indicates that, the relation between Y variable (response variable) the X variables (predictor variables) accounts for 69% of the variation. It is moderately good fit.

**Interpretation (by coefficient and Intercept values):**

Regression Equation,

y = 29.241 + 0.049\*CrimeRate + 0.033\*Age + 0.131\*Indus - 10.321\*Nox + 0.261\*Distance

-0.014\*Tax - 1.074\*Pratio + 4.125\*AvgRoom - 0.603\*LSTAT

**β1 =** It means that, if the 1 unit (per capita) of CrimeRate increases, then it is affects the AvgPrice by increasing 0.049(in $1000’s).

**β2 =** It means that, if the 1 unit (in %) of Age increases, then it is affects the AvgPrice by increasing 0.033(in $1000’s).

**β3 =** It means that, if the 1 unit (in %) of Indus increases, then it is affects the AvgPrice by increasing 0.131(in $1000’s).

**β4 =** It means that, if the 1 unit (per 10 million) of Nox increases, then it is affects the AvgPrice by decreasing -10.321(in $1000’s).

**β5 =** It means that, if the 1 unit (in miles) of Distance increases, then it is affects the AvgPrice by increasing 0.261(in $1000’s).

**β6 =** It means that, if the 1 unit (per $10,000) of Tax increases, then it is affects the AvgPrice by decreasing -0.014(in $1000’s).

**β7 =** It means that, if the 1 unit of Pratio increases, then it is affects the AvgPrice by decreasing -1.074(in $1000’s).

**β8 =** It means that, if the 1 unit of AvgRoom increases, then it is affects the AvgPrice by increasing 4.125 (in $1000’s).

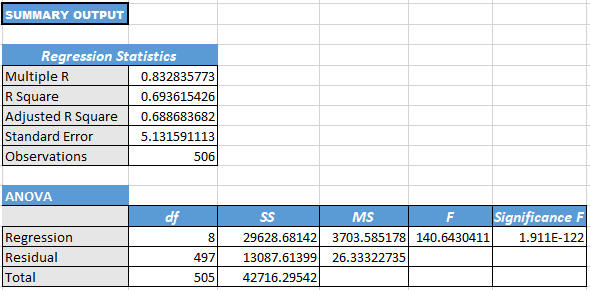
**β9 =** It means that, if the 1 unit (in %) of LSTAT increases, then it is affects the AvgPrice by decreasing -0.603 (in $1000’s).

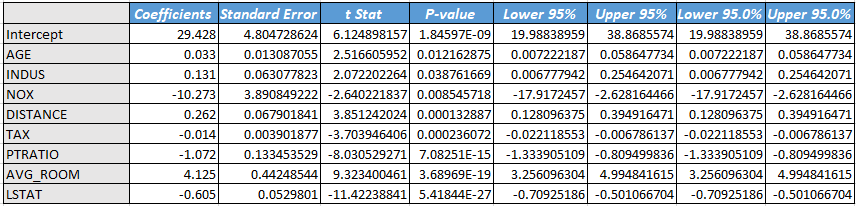
**β0 =** The value for the intercept term in this model is 29.241. This would mean the average of AvgPrice ($1000’s) when their x variables (predictor variables) are equal to zero. The intercept just doesn’t have any meaningful for this model.

**Interpretation (by significance value):**

* A p-value for the CrimeRate is 0.534 means, the p - value is greater than the alpha level 0.05, so we accept the null hypothesis. It means that there is no relationship between the CrimeRate and AvgPrice. It is statistically not significant.
* A p-value for the Age is 0.012 which is equal to 0.00, means, the p - value is less than the alpha level 0.05, so we reject the null hypothesis and accept the alternative hypothesis. It means that there is some relationship between the Age and AvgPrice. It is statistically significant.
* A p-value for the Indus is 0.039 which is equal to 0.00, means, the p - value is less than the alpha level 0.05, so we reject the null hypothesis and accept the alternative hypothesis. It means that there is some relationship between the Indus and AvgPrice. It is statistically significant.
* A p-value for the Nox is 0.008 which is equal to 0.00, means, the p - value is less than the alpha level 0.05, so we reject the null hypothesis and accept the alternative hypothesis. It means that there is some relationship between the Nox and AvgPrice. It is statistically significant.
* A p-value for the Distance is 0.000 which is equal to 0.00, means, the p - value is less than the alpha level 0.05, so we reject the null hypothesis and accept the alternative hypothesis. It means that there is some relationship between the Distance and AvgPrice. It is statistically significant.
* A p-value for the Tax is 0.000 which is equal to 0.00, means, the p - value is less than the alpha level 0.05, so we reject the null hypothesis and accept the alternative hypothesis. It means that there is some relationship between the Tax and AvgPrice. It is statistically significant.
* A p-value for the Pratio is 6.59E-15 which is equal to 0.00, means, the p - value is less than the alpha level 0.05, so we reject the null hypothesis and accept the alternative hypothesis. It means that there is some relationship between the and Pratio and AvgPrice. It is statistically significant.
* A p-value for the AvgRoom is 3.89E-19 which is equal to 0.00, means, the p - value is less than the alpha level 0.05, so we reject the null hypothesis and accept the alternative hypothesis. It means that there is some relationship between the AvgRoom and AvgPrice. It is statistically significant.
* A p-value for the LSTAT is 8.91E-17 which is equal to 0.00, means, the p - value is less than the alpha level 0.05, so we reject the null hypothesis and accept the alternative hypothesis. It means that there is some relationship between the LSTAT and AvgPrice. It is statistically significant.

**8)** **Pick out only the significant variables from the previous question. Make another instance of the Regression model using only the significant variables you just picked and answer the questions below:**

****

****

**a)** **Interpret the output of this model:**

**Interpretation (by R square):**

The R square is used to measure the goodness of fit. In our data the R square value is **0.693**, this indicates that, the relation between Y variable (response variable) the X variables (predictor variables) accounts for 69% of the variation. It is moderately good fit.

**b)** **Compare the adjusted R-square value of this model with the model in the previous question, which model performs better according to the value of adjusted R-square?**

In the first model (Question 7) were we had (CRIME RATE, AGE, INDUS, NOX, DISTANCE, TAX, PTRATIO, AVGROOM, LSTAT) as our predictor variables, and our adjusted R squared is 68.83%. However, for the second model (Question 8) were we removed CRIME RATE variable because it is not significant variable and the remaining predictor variables are significant variables, and our adjusted R squared has increased to 68.87%. In the concept of adjusted R square, we choose models with higher adjusted R squared over others. So in this case the second model (Question 8) has the higher adjusted R squared value. So we can conclude that the performance of this model (Question 8) better than the previous model (Question 7).

**c)** **Sort the values of the Coefficients in ascending order. What will happen to the average price if the value of NOX is more in a locality in this town?**

The coefficient value of NOX is -10.321, after arranging in ascending order the value of AvgPrice is decreased. It means that, if the 1 unit of NOX (per 10 million) increases the AvgPrice will be decreased.

**d)** **Write the regression equation from this model.**

Regression Equation,

y = 29.928 + 0.033\*Age +0.131\*Indus – 10.273\*Nox + 0.262\*Distance -0.014\*Tax - 1.072\*Pratio + 4.125\*AvgRoom – 0.605\*LSAT.