**THE UNIVERSITY OF TEXAS AT ARLINGTON**



**SIMPLE LINEAR REGRESSION PROJECT ON UTILITY COSTS**

**APPLIED REGRESSION ANALYSIS**

**IE 5318- 005**

**Under the Guidance of & Submitted to: Dr. Aera K Leboulluec**

**SUBMITTED BY:**

**Achuth Chandrasekhar Venkata Subramaniam- 1001487479**

**Anutham Eashwar Padmanabhan- 1001371738**

**Nobel Victor Bombacha - 1001583388**

**Salvius Stephen Kinny- 1001552262**

**Tejas Sutar Nandkumar- 1001558442**

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S. No.** | **CONTENT** | **PAGE NUMBER** |
| 1 | PROJECT PROPOSAL | 3 |
| 2 | PROBLEM STATEMENT AND DATA COLLECTION PROCESS | 3 |
| 3 | SCATTER PLOT INTERPRETATION | 6 |
| 4 | MEANINGFULNESS OF THE MODEL | 7 |
| 5 | SIMPLE LINEAR REGRESSION MODEL | 7 |
| 6 | INFERENCES ON THE PARAMETERS | 8 |
| 7 | MODEL ASSUMPTIONS | 10 |
| 8 | INFERENCE ON THE TRUE LINE AND PREDICTION | 11 |
| 9 | NORMALITY TEST | 15 |
| 10 | FINAL DISCUSSION | 16 |

**TABLE OF TABLES**

|  |  |  |
| --- | --- | --- |
| **TABLE NUMBER** | **TABLE NAME** | **PAGE NUMBER** |
| 1 | PROJECT DATA SET | 3 |
| 2 | ANOVA, SIMPLE STATISTICS AND PARAMETER ESTIMATES TABLE | 8 |
| 3 | CONFIDENCE BAND LIMIT FOR EVERY X | 14 |
| 4 | MODIFIED LEVENE TEST- SAS OUTPUT | 15 |

**TABLE OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **FIGURE NUMBER** | **FIGURE NAME** | **PAGE NUMBER** |
| 1 | SCATTER PLOTS FOR EACH OF 4 REGRESSOR AND RESPONSE VARIABLE | 6 |
| 2 | SCATTER PLOT FOR SLR MODEL | 7 |
| 3 | RESIDUAL Vs PREDICTED VALUES | 10 |
| 4 | PLOT OF NORMALITY FOR RESIDUALS | 11 |
| 5 | INTERVAL PLOT | 12 |
| 6 | CONFIDENCE BAND WITH FITTED LINE | 14 |
| 7 | BOX PLOT | 15 |

1. **PROJECT PROPOSAL**

**PROBLEM STATEMENT:**

For the project, data of the utility bills of a residence in the United States was collected from the time- period 1999 to 2010.The data set obtained consists of Response (Y) & Predictor Variables (X1, X2, X3, X4) which is listed below:

Y- Total Utilities Cost (in dollars)

X1- Temperature (°F)

X2- Kilowatt hour - kWh

X3- Centum Cubic Feet -CCF

X4- Therms per day - thm

The data set was collected from the following link:

<https://raw.github.com/vincentarelbundock/Rdatasets/master/csv/mosaicData/Utilities.csv>

The primary objective of this project is to plot and analyze the effects of the various predictor variables X1, X2, X3, X4 against the response variable Y. For the data set chosen we have obtained the relation between cost of utilities and the factors that affect it. Modeling this data would be useful because it will give us an idea as to which variables/factors are to be optimized to reduce the overall utility cost. Additionally, we can also predict the total utility cost for budgeting/saving purposes and energy conservation.

**SCATTER PLOT INTERPRETATION:**

**DISCUSSION OF SCATTER PLOTS:**

* The scatter plot of Temperature (X1) vs Total Utilities Cost (Y) that shows a downward linear trend with a R2 value of 0.6766, which depicts that the observations are close to the trend line, there are a few outliers. The regression model fits the data by 67.66%.
* The scatter plot for kWh (X2) vs Total Utilities Cost (Y) shows an upward linear trend with a R2 value of 0.2069, which depicts that the observations are scattered around the trend line with a few outliers. The regression model fits the data only by 20.69%.
* The scatter plot for CCF (X3) vs Total Utilities Cost (Y) shows an **upward linear trend** with a R2 value of 0.7049, which depicts that the observations are close to the trend line and there are a few outliers. The regression model fits the data by **70.49%.**
* The scatter plot for Therms per day (X4) vs Total Utilities Cost (Y) shows an upward linear trend with a R2 value of 0.6906, which depicts that most of the observations are close to the trend line and there are a few outliers. The regression model fits the data by 69.06%.

**MEANINGFULNESS**

The decided predictor variable (X) will have the strongest correlation with the dependent variable (Y). From the above analysis through scatter plots, we observe that amongst all the predictor variables Temperature (X1), kWh(X2), CCF (X3), Therms per day (X4) we infer that CFF (X3) has a **high correlation** with Y. This is seen as the highest R2 value on the scatter plot created: 0.7049. Thus, of the 4 predictor variables, we take **CCF**- **X3** into consideration for further analysis in relation to the dependent variable- Y (**Total Utilities Cost)**

**SIMPLE LINEAR REGRESSION MODEL**

A quantitative relationship between the two variables X (CCF in our case- Predictor) and Y (Total Utilities Cost- Response) is developed by the Simple Linear Regression Model. From the above regression analysis, the regression model shows the equation:

**ŷ =92.693+0.7796x**

where,

b0 🡪 Point estimator of **β0** is 92.693

b1 🡪 Point estimator of **β1** is 0.7796

The slope **β1** = 0.7796 indicates that the increase in each value of CCF leads to an increase in the mean of probability distribution of Y of ***$0.7796.*** The intercept **β0** = ***92.693*** indicates the value of the regression function at X = 0. The above equation states that ŷ is the response variable i.e. Utility Bill and X is the predictor variable i.e. CCF. The regression model shows a positive slope with upward linear trend which states that as the CCF increases, so will the Utility Bill. The given relationship has a **good fit**.

**MODEL FORM**

**Figure 2: Scatter plot of CCF(X) vs Total Utility Cost (Y) [SLR MODEL]**

A simple linear regression model goes by the equation: **Yi = β0 + β1 + Ɛi**.

**Y**i 🡪 Value of the response variable of the ith trial of the Total Utilities Cost.

**Xi**🡪A constant/ Value of Predictor variable of the ith trial of the CCF.

**Ɛ** 🡪Residual/ Random Error/ Vertical variation

**β0, β1** 🡪Regression coefficients. **β0** 🡪 Y intercept of regression line & **β1** 🡪 Slope of the regression line.

**ŷ** 🡪 Estimated mean CCF.

The estimated regression function for this SLR model is **ŷ = 0.7796x + 92.693** as shown in the scatter plot of CCF vs. Utilities Bill. The obtained function is of the form ŷ = b0+b1 (X where b0 is the point estimator of the y-intercept and b1 is point estimator of the slope. Here, ŷ denotes the estimated mean of CCF levels.

The ANOVA output table and the estimates of the parameters are shown below. The output has been developed on SAS. From the parameter estimates of the given output, we can obtain the value of y-intercept and the slope of the true line.

**ANOVA TABLE WITH PARAMETER ESTIMATES**

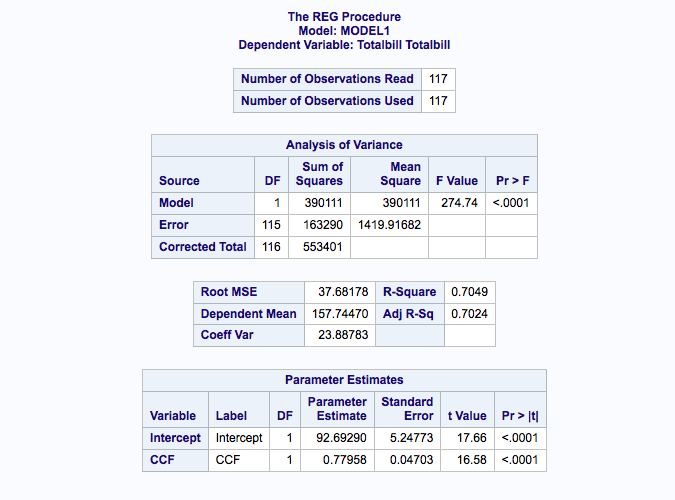
**SSTO (TOTAL SUM OF SQUARES):**

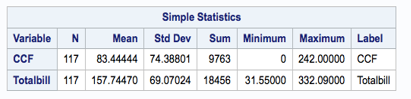
A value of ***553,401*** is obtained as SSTO defined as the sum of the overall observations of the squared difference between dependent variable and its mean. It is to be noted that the larger the variation among the Yi observations, the larger is the SSTO, hence we state that it is **a measure of the uncertainty** related to the total bill amount corresponding to the CCF, when the CCF is not considered.

* **SSE (ERROR SUM OF SQUARES):**

A value of ***163,290*** is obtained as SSE, which is a measure of sum of squares of deviations of the predicted values from the mean value of the response variable. The SSE is used as **a measure of the optimal condition** while selecting the parameter and the model.

**SSE 🡪 Variation of the Yi observations.**





**Table 2: ANOVA Table, Parameter Estimates and Simple Statistics**

**SSR (REGRESSION SUM OF SQUARES):**

A value of ***390,111*** is obtained as SSR and, is a measure of that part of the variability of Yi which is associated with the regression line. The larger SSR is in relation to SSTO, the greater is the effect of the regression relation accounted for the total variation in the Yi observations.

SSR = SSTO – SSE = 553401 – 163290 = ***390,111***

**MSR (REGRESSION MEAN SQUARES):**

A value of ***390,111*** is obtained as MSR**,** which is **the ratio of the Sum of Squares of the Regression to its Degrees of Freedom**. [DOF (for SLR) = 1]

**MSE (ERROR MEAN SQUARES):**

A value of ***1,419.91682*** is obtained as MSE, which is SSE/(N-2) = 163290/115 = ***1,419.917***

F\*=MSR/MSE = 390111/1419.917 = ***274.712***

Let **H0:** β1 = 0 🡪 A linear relationship **DOES NOT** exist.

**H1:** β1 ≠ 0 🡪 A linear relationship **EXISTS**.

The level of significance is assumed to be, α= 0.05, since n= 117, we require F (0.95,1,115)

Since F\* = 274.712 > 3.9201 (F-table value), we accept that hypothesis H1: β1 ≠ 0. In other words, there is a linear association between the Utility Bill and CCF. Thus, it is concluded that the **model is significant** and the data **fits well** in the regression model.

**RMSE (ROOT ERROR MEAN SQUARES):**

A value of ***37.68178*** which represents the sample standard deviation of the differences between predictor values and observed values.

**DEPENDENT MEAN:**

It is the mean the dependent variable with the value of ***157.74470***

**R-SQUARED:**

The coefficient of determination is ***0.7049*** that is the proportionate reduction of total variation **associated** with the use of the predictor variable X (CCF), so the larger R-Squared is, the higher is the total variation of Y will reduce due to the introduction of the predictor variable X (CCF).

**ADJUSTED R-SQUARED:**

The adjusted R-squared attempts to provide **more reliable estimate of R-squared** for the population and is found to be 0.7024.

**COEFFICIENT OF VARIATION:**

This is the ratio of the standard deviation to the mean = ***157.74470*** which indicates the extent of variability in relation to the mean of the population.

**STANDARD ERROR OF b0 & b1:**

These are the standard errors associated with the coefficient with values of ***5.24773*** & ***0.04703*** respectively.

**MODEL ASSUMPTIONS:**

To verify the SLR model assumptions, we need to conduct the residual analysis for the original data. We need to verify the following SLR model assumptions:

* A linear model is reasonable 🡪 Verified by comparing the plot for ei vs ŷi for curvature.
* The residuals have constant variance 🡪 Verified by comparing the plot for ei vs ŷi for funnel shape.
* The residuals follow a normal distribution 🡪 Verified by comparing the normal probability plot.
* **Assumption 1**: ***A Linear relationship exists between Total Cost and CCF.***

From the figure 2, we notice that there isn’t any existence of the curvature in the graph, we can conclude that the linear model is reasonable i.e. **the first assumption of SLR model is satisfied**.

* **Assumption 2**: ***The residuals have constant variance***

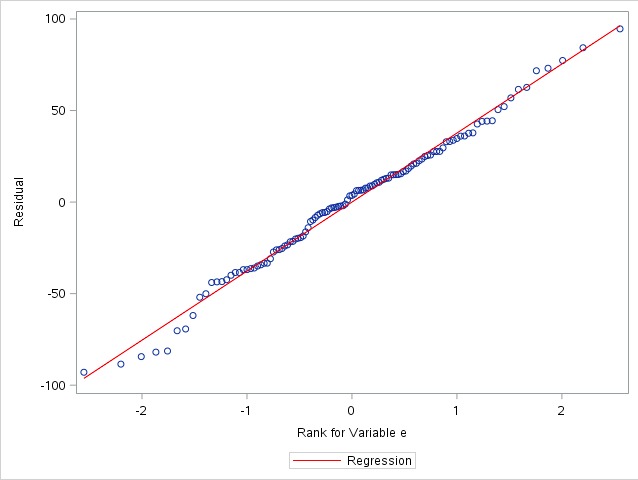
From the figure 3 in the plot below, we can clearly notice a funnel shape variation of the data. Hence, we can conclude that the residuals **do not have constant variance** i.e. the second assumption of SLR model is violated. Hence a **transformation would be required** in this case to bring out a constant variance in the model.

* **Assumption 3**: ***The residuals follow a normal distribution***

From the figure 4 of the plot provided below, the data almost has a perfect straight line. Hence, we can conclude that the residuals follow a normal distribution i.e. the **third assumption of SLR model is satisfied**.

|  |
| --- |
|  |

**Figure 3: Plot of Residual vs. Predicted Value of Y**

 **Figure 4: Plot of Normality for Residuals**

**Confidence Interval for Slope**:

To test whether the linear relationship exists between the Total Cost and CCF, we are going to perform hypothesis test for slope b1.

The regression equation is given by: **ŷ = 92.693 + 0.7796x**

* **Hypothesis Test**:

H0: β1 = 0 (No linear relationship exists)

H1: β1 ≠ 0 (linear relationship exists)

Decision Rule: If 0 is not in the confidence interval we will reject H0.

**Two-sided test**: **b1±t (1-α/2; n-2) s{b1}**

standard error of b1: s{b1} = √ (MSE/ Sxx)

MSE = 1419.91682 (From ANOVA in table)

SSR = (b12 \* Sxx)

390111= (92.693) ^2 \* Sxx

**Sxx** =390111/ [(92.693) ^2]

**Sxx** = 641,866.2423

🡺 s{b1} =√1419.91682 / 641,866.2423

**s{b1} = 0.0470337** (Close to the table value obtained in the ANOVA table)

From the t-distribution table, t (0.975, 115) = 1.98168

🡺 b1 ± t (1-α/2; n-2) s{b1}

🡺 0.7796± (1.98168) \* (0.047033704)

🡺 (0.0264, 0.8728)

Since the interval does not include 0, we reject Ho, and conclude that a linear relationship exists between the 2 quantitative variables (Temperature and CCF).

We are **95% confident that the mean Utility Bill lies between** (0.0264, 0.8728) for an increase in CCF.

* **INFERENCES ON TRUE LINE AND PREDICTION:**

As part of Hand Calculation, we take a particular value of x (CCF) to calculate the mean amount of Total Utilities Cost to get a predicted value that corresponds to the x-value.

ŷh = b0 + b1x = 92.693 + 0.7796x

ŷ│x=90 = 92.693 + 0.7796 \* (139)

ŷ│x=90 = 201.0574

* **Confidence Interval for the mean response**

The confidence level for this hand calculation is assumed to be 95% or α= 0.05 to determine the Confident Interval.

* **2- Sided Confidence Interval is calculated as**:

ŷ │x= 90 ±t (1-α/2; n-2)] \* [s {ŷ h}]

ŷ │x=90 = 201.0574

From the ANOVA, we know that **MSE**= 1419.91682

From previous calculation, **Sxx** = 641,866.2423

Number of observations, n = 117

From Sample Statistics table brought out in the ANOVA table, ∑xi / n = 82.44444

We use the above values and substitute into the follow equation,

S {ŷ h} = √MSE {1/n + [(Xh−X̄̄) ^2/Sxx]}

=√1419.91682{1/88 + [(88– 82.4444) ^2/ 641,866.2423]}

S {ŷ h} = 4.0253

From the t-distribution table we get; t (0.975, 88) = 1.9906

Thus, the Confidence Interval for the mean response is calculated as:

🡺 ŷ│x=90 ± t (1-α/2; n-2) S {ŷh}

🡺 201.0574 ± (1.9906\*4.0253)

🡺 (193.044,209.070)

We are therefore, **95% confident** that the Total Utilities Cost for a CCF value of 182 is between the range **193.044 and 209.070** for a CCF value of 182.

* **Prediction Interval for the mean response**

The confidence level for this hand calculation to determine Prediction Interval is assumed to be 95% or α= 0.05.

* **2- Sided Prediction Interval is calculated as**:

🡺 ŷ│x=90 ± t (1-α/2; n-2) s{pred.}

🡺 MSE = 1419.91682

{S (ŷh)}2 = (4.0253) 2 = 16.20304

by substituting the above value in the formula below, we get

s{pred} = √S2 (ŷh) + MSE

= √16.20304+ 1419.91682

s{pred} = 37.8961721

🡺 t (0.975;88) = 1.9906, is obtained from the t-Distribution table.

🡺 Ŷ│x=90 = 201.0574

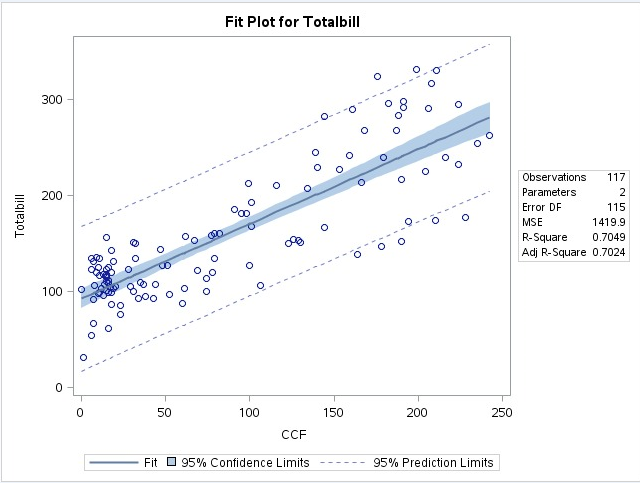
By utilizing the calculated values, we obtain the Prediction Interval:

201.0574 ± (1.9906\*37.8961721)

🡺 (125.62, 276.494)

From the above calculation, we are 95% confident that our Prediction Interval lies between (125.62, 276.494) for the CCF value of 182. Our calculation is right because the Prediction Interval is wider than the Confidence Interval and is within the 95% Confidence limits.

**INTERVAL PLOT**



**Figure 4: Interval Plot of Total Bill vs. CCF**

* **Two-Sided Confidence Band**:

ŷ │x=90 ± [√2F (1 − α, 2, n − 2)] S{Ŷh}

ŷ │x=90 = 201.0574

S {ŷh} = 4.0253

We obtain F (0.95, 2, 115) = 3.0784 from the F-distribution table.

We calculate the Confidence Band with the above values:

116.81851 ± √2 ∗ 3.2746 \*(2.731972356)

116.81851 ± 6.991572217

(109.83,123.81)

We come to a conclusion that we are 95% confident that Total Utilities Cost for CCF of 182 will lie within a range of 109.83 and 123.81.

* **INTERPRETATION OF CONFIDENCE BANDS ON EXCEL**

The term **s.e. {ŷ| x = x0}** is computed using MS Excel for all the observations to compute the Confidence Band. Here, Width of Confidence Band < Width of Confidence Interval < Width of Prediction Interval. The confidence band table is given below:

**Table 3: Confidence Band Calculation Table**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **CCF (X)** | **Predicted Bill (ŷ)** | **S.E {ŷ}** | **LB** | **UB** |
| 95 | 166.755 | 3.525801 | 159.7711 | 173.7389 |
| 123 | 188.5838 | 3.949298 | 180.761 | 196.4066 |
| 126 | 190.9226 | 4.017693 | 182.9644 | 198.8808 |
| 129 | 193.2614 | 4.089813 | 185.1603 | 201.3625 |
| 139 | 201.0574 | 4.354675 | 192.4317 | 209.6831 |
| 144 | 204.9554 | 4.499712 | 196.0424 | 213.8684 |
| 144 | 204.9554 | 4.499712 | 196.0424 | 213.8684 |
| 166 | 222.1066 | 5.216503 | 211.7737 | 232.4395 |
| 178 | 231.4618 | 5.64918 | 220.2719 | 242.6517 |
| 194 | 243.9354 | 6.258808 | 231.538 | 256.3328 |

The values of the lower bound and the upper bound are plotted: CCF (x-axis) vs. Total Utilities Bill (y-axis).

Below is the Confidence Band Figure 5 produced from the above table on MS Excel. The Total Utilities Cost is the Blue dots, Upper bound is seen as the Green Dots and the Lower Bound is seen as the Red Dots. The Confidence Band is seen as the width between the Upper and the Lower Bound. This proves that the Regression Line is a **good fit**.

**Figure 5: Confidence Band from MS Excel**

The Mean, Median, Quartiles, Minimum and Maximum values of the model is shown in the above image obtained from SAS. Q1 is the First Quartile, Q2 is the Second Quartile or Median, and Q3 is the third Quartile. The two horizontal lines that arise on either side of the box are called whiskers. The rhombus inside the box is the mean and the line inside the box is the mean. We see a Bell curve, hence we conclude that the model is perfect as this shows a Normal Distribution or Symmetry of the data values.

* **NORMALITY TEST**

Hypothesis:

Hypothesis for the normality test is as follows:

H0: Normality is OK

H1: Normality is violated.

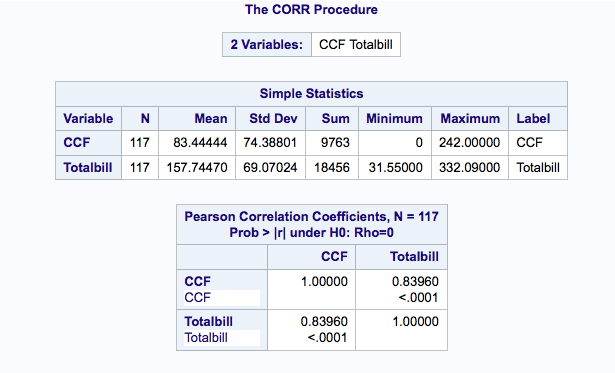
From the above SAS output, we can note that the value of e(Residual) = 0.8396

Also, the cut-off value c (α = 0.10, n=117) from the table is 1.290

Since, e = 0.8396 < 0.993, **the decision is to reject H0**.

Hence, **Normality is violated** for all values of α.

**Table 4: SAS output of Non- Constant Variance table**



* **MODIFIED LEVENE TEST** (Test for Non-Constant Variance):

H0: Means of the absolute deviation have equal variance

H1: Means of the absolute deviation do not have equal variance

We assume α = 0.05

From the above SAS output, the p value of F\* statistics is 0.003

Since, p value < α, the **decision is to reject H0**

Hence, we conclude that the means of the absolute deviation are not equal i.e. **error variance of the model is not equal.**

* **DISCUSSION ABOUT MODEL ASSUMPTIONS**
* From the test of Normality, we found that the normality is violated. But, from the boxplot and the Normality figure we consider the Normality to be OK.
* There is no constant variance in the model.
* **NEED FOR TRANSFORMATION**

We notice a funnel shape in the figure of residuals (Residual vs. Total Utilities Cost plot), thus a **transformation is required for this model**. For a non-constant variance and non-normality, we would need a “**Variance stabilizing transformation**”.

* **FINAL DISCUSSION**
* The objective of a simple linear regression analysis is to model the relationship between 2 quantitative variables CCF and Total Utilities Bill. It was concluded, after plotting the scatter plots, a positive linear trend exists in the model, in other words, there would be a rise in the Total Utilities Bill if the CCF increases.
* The correlation of these two variables is **0.7049**, and the sum of the residuals is equal to zero.
* All the output tables and figures were plotted on SAS and MS Excel.
* The Confidence Interval for the model is **193.044 and 209.070** at 95% confidence for a CCF- 182.
* The Confidence Band for the model is **109.83 and 123.81** at 95% confidence for a CCF- 182.
* The Prediction Interval for the model is **125.62, 276.494** at 95% confidence for a CCF- 182.
* As a test of variance, we performed the Normality test and the Modified Levene Test. It was the concluded that the Normality was violated and there existed a constant variance.
* A Funnel shape was found in the plot of Residual Vs. Total Utilities Cost hence, a **Variance-Stabilizing transformation** would be required.
* If we need additional knowledge about the response variable (Total Utilities Bill) we need to perform Multi Linear Regression. This is because, the introduction of additional variables would make the model more appropriate in terms of being fit and providing additional information.
* **REFERENCES**
* Applied Linear Statistical Models | Fifth Edition | Michael H. Kutner, Christopher J. Nachtsheim, John Neter, William Li
* <https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_reg_sect003.htm>
* <https://raw.github.com/vincentarelbundock/Rdatasets/master/csv/mosaicData/Utilities.csv>