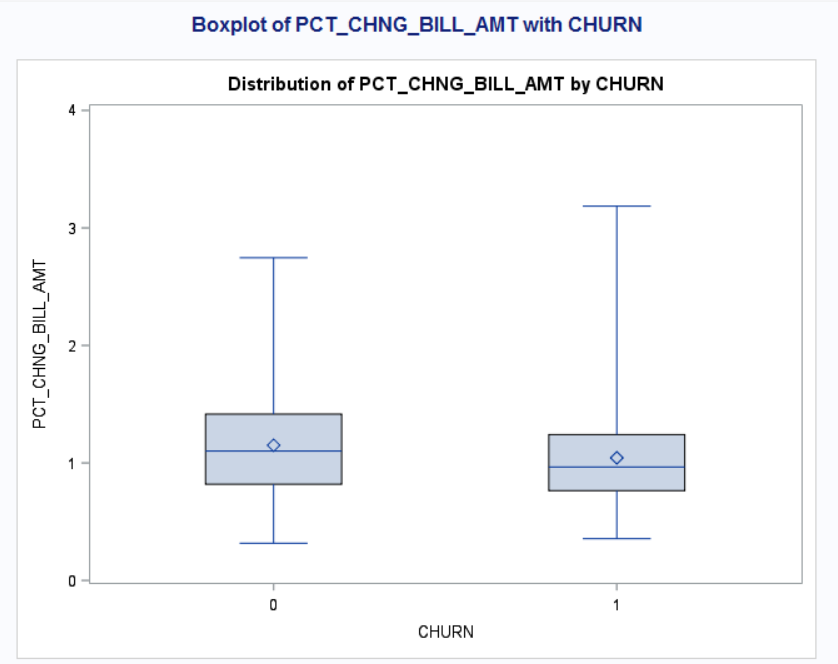
ASSIGNMENT – 6

BY

Sasidhar Mukthinuthalapati

1. Problem -1
   1. The boxplots for PCT\_CHNG\_BILL\_AMT by Churn is as follows:



From the above Boxplot, we can make the following observations:

* The median looks to be the same for both the boxplots related to both the CHURN values and hence we can say that PCT\_CHNG\_BILL\_AMT doesn’t have much effect on the CHURN value.

The code to generate the above plot is as follows:

title "Boxplot of PCT\_CHNG\_BILL\_AMT with CHURN";

**PROC** **sort**;

by CHURN;

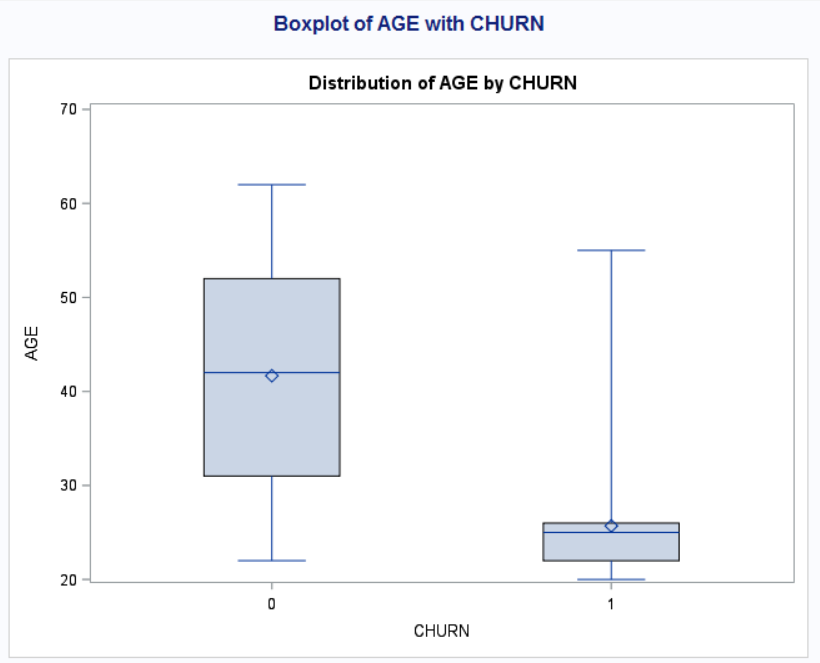
**run**;

**proc** **boxplot**;

plot PCT\_CHNG\_BILL\_AMT\*CHURN;

**run**;

The boxplot for AGE vs CHURN is as follows:



From the above boxplot, we can make the following observations:

* The people who changed network have an average age of around 25 years and hence we can say that **most of the people who have changed networks are under the age of 30**.
* So, an important factor which **influences CHURN value is Age**.

The code to generate the above boxplot is as follows:

title "Boxplot of AGE with CHURN";

**PROC** **sort**;

by CHURN;

**run**;

**proc** **boxplot**;

plot AGE\*CHURN;

**run**;

* 1. To generate the final model, I’ve used the forward selection method and have used the following two options **sls** which sets the criterion for a variable to be included into the next model. This means that the p-value of the variable must be less than 0.05. **sle** is the other option which I’ve used which sets the criteria for entry which means that a variable must have p-value which is less than 0.05 to be included in the model.

The code is as follows:

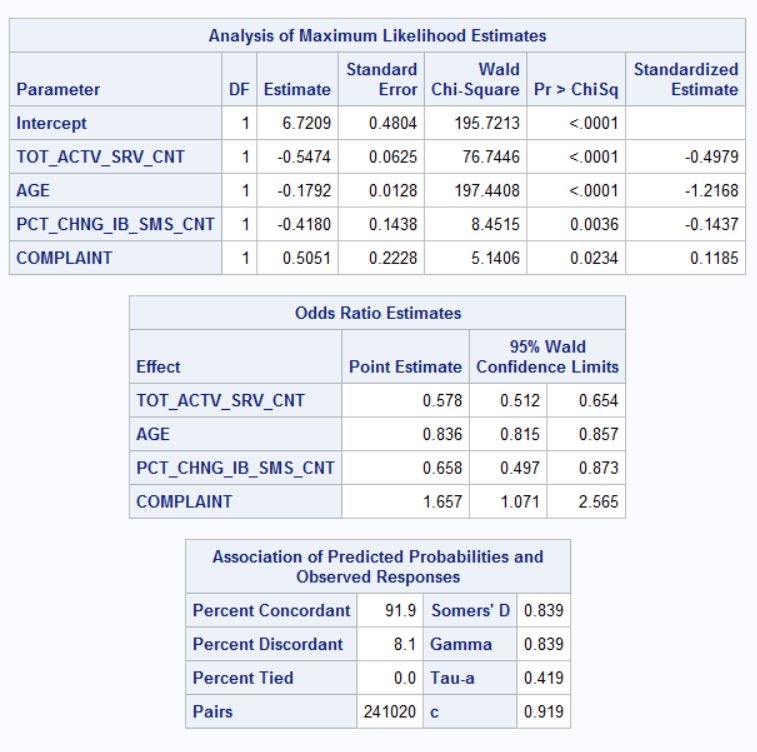
TITLE "FINAL LOGISTIC REGRESSION MODEL";

**PROC** **LOGISTIC**;

MODEL CHURN (EVENT='1') = GENDERNEW ED1 ED2 ED3 ED4 ED5 LAST\_PRICE\_PLAN\_CHNG\_DAY\_CNT TOT\_ACTV\_SRV\_CNT AGE PCT\_CHNG\_IB\_SMS\_CNT PCT\_CHNG\_BILL\_AMT COMPLAINT / stb SELECTION = FORWARD sls = **0.05** sle = **0.05**;

**RUN**;

The above code generates the following final model which is as follows:



From the above Analysis of Maximum Likelihood Estimates table we can see that **TOT\_ACTV\_SRV\_CNT, AGE, PCT\_CHNG\_IB\_SMS\_CNT** and **COMPLAINT** are all significant variables because they have p-value which is less than 0.05.

The Final Equation is as follows:

**LOG(CHURN(EVENT ’1’)) = 6.7209 – 0.5474\*TOT\_ACTV\_SRV\_CNT - 0.1792\*AGE – 0.4180\*PCT\_CHNG\_IB\_SMS\_CNT + 0.5051\*COMPLAINT**

* 1. From the equation generated in the previous bit we can make the following analysis of how a change in each variable will affect the CHURN value.

The Equation is as follows:

**LOG(CHURN(EVENT ’1’)) = 6.7209 – 0.5474\*TOT\_ACTV\_SRV\_CNT - 0.1792\*AGE – 0.4180\*PCT\_CHNG\_IB\_SMS\_CNT + 0.5051\*COMPLAINT**

* When all other variables are kept constant and there is a unit increase in the TOT\_ACTV\_SRV\_CNT, the odds of Churn value (i.e. people shifting to other networks) to **decrease by 42.2%.**
* A unit increase in the AGE would result in the odds of CHURN value (i.e. people shifting to other networks) to **decrease by 16.4%.**
* A unit increase in the PCT\_CHNG\_IB\_SMS\_CNT would result in the odds of CHURN value (i.e. people shifting to other networks) to **decrease by 34.2%.**
* A change of complaints from 0 to 1 would result in the odds of **CHURN** value to **increase by 65.7%**. i.e. the probability of the customer shifting to another network would increase by 65.7%.
  1. Given the following values LAST\_PRICE\_PLAN\_CHNG\_DAY\_CNT=0, TOT\_ACTV\_SRV\_CN=4, PCT\_CHNG\_IB\_SMS\_CNT= 1.04, PCT\_CHNG\_BILL\_AMT= 1.19, and COMPLAINT =1. The predicted CHURN value can be computed using SAS.

To do that we can run the following code inSAS:

TITLE "PREDICTION";

**DATA** CHURN\_test;

INPUT TOT\_ACTV\_SRV\_CNT AGE PCT\_CHNG\_IB\_SMS\_CNT PCT\_CHNG\_BILL\_AMT COMPLAINT CHURN;

DATALINES;

4 43 1.04 1.19 1 .

;

**DATA** pred;

SET CHURN\_test churn\_train;

**RUN**;

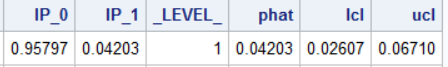
**PROC** **LOGISTIC**;

MODEL CHURN (EVENT='1') = TOT\_ACTV\_SRV\_CNT AGE PCT\_CHNG\_IB\_SMS\_CNT PCT\_CHNG\_BILL\_AMT COMPLAINT ;

OUTPUT OUT=PRED p=phat lower=lcl upper=ucl predprobs=(individual);

**RUN**;

We get the following output:



Here the phat value is the predicted value and we can see that it 0.04203 which is very close to 0 and hence the predicted CHURN value is 0.042 which is equal to 4.203%.

Therefore, the **probability of this person changing the network is 4.203%.**

* 1. The SAS Code for this problem is as follows:

**DATA** churn\_train REPLACE;

INFILE "churn\_train.csv" DELIMITER =',' MISSOVER FIRSTOBS=**2**;

INPUT GENDER $ EDUCATION LAST\_PRICE\_PLAN\_CHNG\_DAY\_CNT TOT\_ACTV\_SRV\_CNT AGE PCT\_CHNG\_IB\_SMS\_CNT PCT\_CHNG\_BILL\_AMT CHURN COMPLAINT ;

IF EDUCATION = '.' THEN EDUCATION = '1';

GENDERNEW = **0**;

ED1 = **0**; ED2 = **0**;

ED3 = **0**;

ED4 = **0**;

ED5 = **0**;

IF GENDER = 'M' THEN GENDERNEW = **1**;

IF EDUCATION = **2** THEN ED1 = **1**;

IF EDUCATION = **3** THEN ED2 = **1**;

IF EDUCATION = **4** THEN ED3 = **1**;

IF EDUCATION = **5** THEN ED4 = **1**;

IF EDUCATION = **6** THEN ED5 = **1**;

**PROC** **PRINT**;

**RUN**;

TITLE "DISTRIBUTION OF EDUCATION";

**PROC** **FREQ**;

TABLES EDUCATION;

**RUN**;

TITLE "REPLACE MISSING VALUES WITH 1 SINCE IT IS THE MOST RECURRING";

**DATA** CHURN;

SET CHURN;

IF EDUCATION = '.' THEN EDUCATION = '1';

**PROC** **PRINT**;

**RUN**;

title "Boxplot of PCT\_CHNG\_BILL\_AMT with CHURN";

**PROC** **sort**;

by CHURN;

**run**;

**proc** **boxplot**;

plot PCT\_CHNG\_BILL\_AMT\*CHURN;

**run**;

title "Boxplot of AGE with CHURN";

**PROC** **sort**;

by CHURN;

**run**;

**proc** **boxplot**;

plot AGE\*CHURN;

**run**;

TITLE "CORRELATION ANALYSIS";

**PROC** **CORR**;

VAR CHURN GENDERNEW ED1 ED2 ED3 ED4 ED5 LAST\_PRICE\_PLAN\_CHNG\_DAY\_CNT TOT\_ACTV\_SRV\_CNT AGE PCT\_CHNG\_IB\_SMS\_CNT PCT\_CHNG\_BILL\_AMT COMPLAINT;

**RUN**;

TITLE "FINAL LOGISTIC REGRESSION MODEL";

**PROC** **LOGISTIC**;

MODEL CHURN (EVENT='1') = GENDERNEW ED1 ED2 ED3 ED4 ED5 LAST\_PRICE\_PLAN\_CHNG\_DAY\_CNT TOT\_ACTV\_SRV\_CNT AGE PCT\_CHNG\_IB\_SMS\_CNT PCT\_CHNG\_BILL\_AMT COMPLAINT / stb SELECTION = FORWARD sls = **0.05** sle = **0.05**;

**RUN**;

TITLE "PREDICTION";

**DATA** CHURN\_test;

INPUT TOT\_ACTV\_SRV\_CNT AGE PCT\_CHNG\_IB\_SMS\_CNT PCT\_CHNG\_BILL\_AMT COMPLAINT CHURN;

DATALINES;

4 43 1.04 1.19 1 .

;

**DATA** pred;

SET CHURN\_test churn\_train;

**RUN**;

**PROC** **LOGISTIC**;

MODEL CHURN (EVENT='1') = TOT\_ACTV\_SRV\_CNT AGE PCT\_CHNG\_IB\_SMS\_CNT PCT\_CHNG\_BILL\_AMT COMPLAINT ;

OUTPUT OUT=PRED p=phat lower=lcl upper=ucl predprobs=(individual);

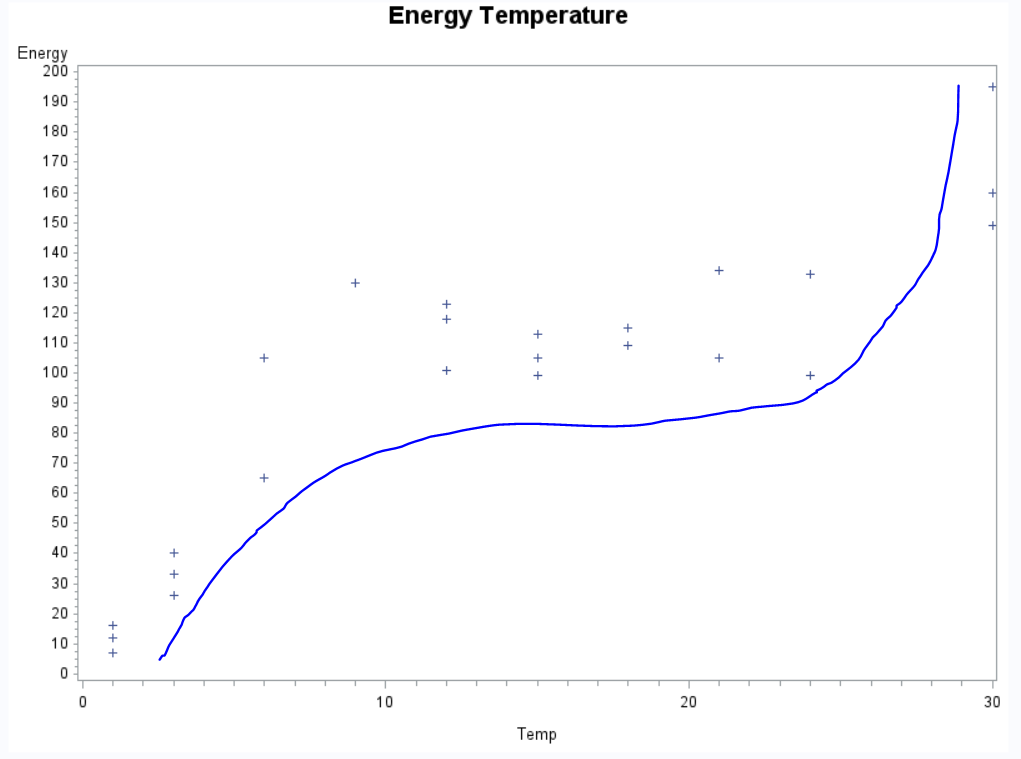
**RUN**;

**proc** **print** data = churn\_test;

**run**;

PROBLEM – 2 – EnergyTemp

1. Scatterplot between Energy and Temperature is as follows:



To generate the above plot, I ran the following code in SAS:  
**proc** **gplot**;

plot Energy\*temp;

**run**;

We can make the following observations from the above scatterplot plot:

* The relation between the Energy and Temperature isn’t linear and hence we may need to add in few higher order terms (Cubic) in order to establish a relation between Energy and Temperature.

1. To add in Squared and Cubic terms into the model we have to run the following code in SAS:

title "Energy Temperature";

**data** energyTemp;

infile "energytemp.txt" delimiter =' ' missover firstobs=**2**;

input Energy Temp;

temp2 = temp\*\***2**;

temp3 = temp\*\***3**;

**proc** **print**;

**run**;

The above code will add 2 terms of which one is the squared of temperature and the other is Cube of Temperature.

In the previous question, we had observed that cubic terms may be required in order to establish some kind of relation between Energy and Temperature.

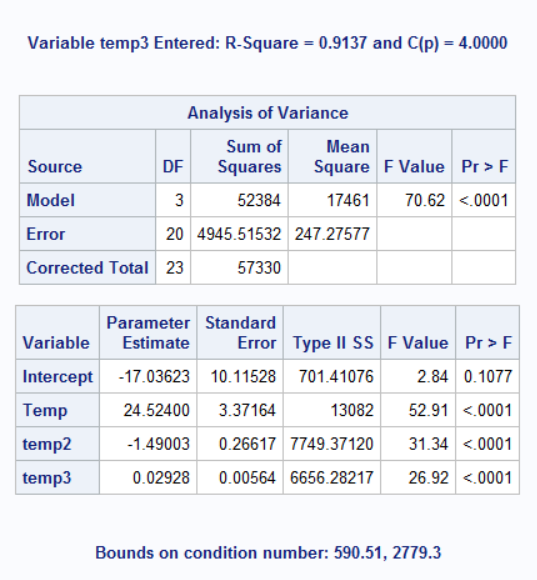
The Final Cubic model can be generated by running the following code in SAS:

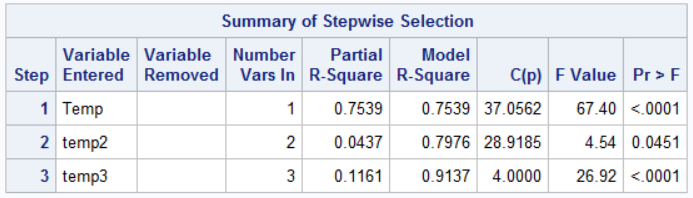
**proc** **reg**;

model energy = temp temp2 temp3 / selection = stepwise;

**run**;

The following output is generated:



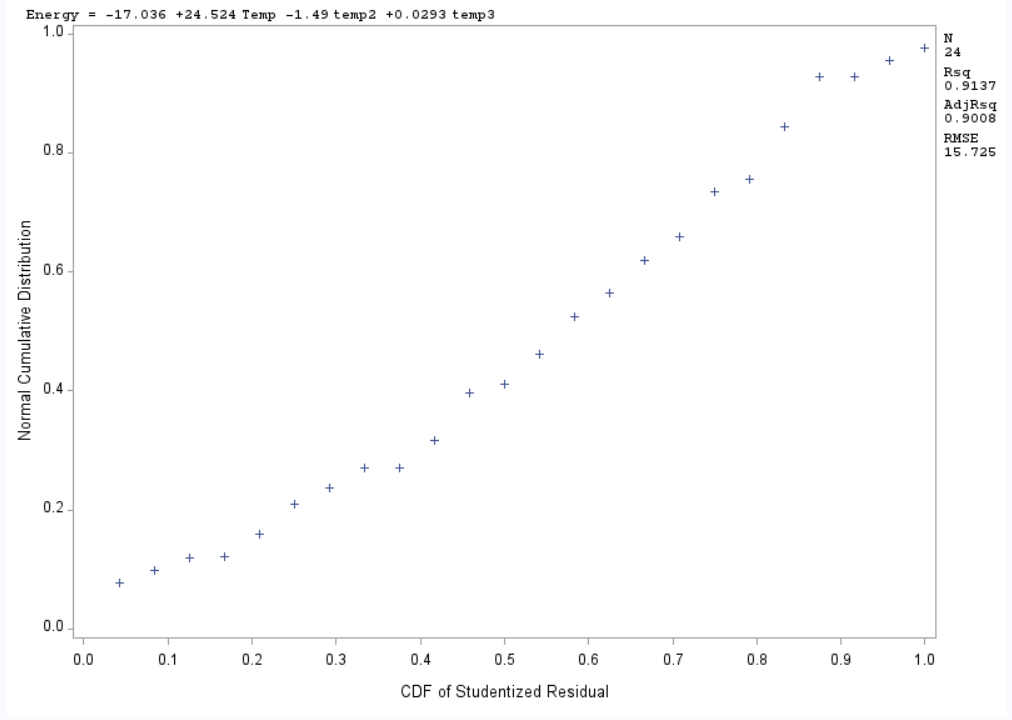


From the above tables, we can come to the following conclusion for the Final Model Equation (Cubic):

**ENERGY = -17.036 + 24.524\*TEMP – 1.490\*TEMP2 + 0.02928\*TEMP3**

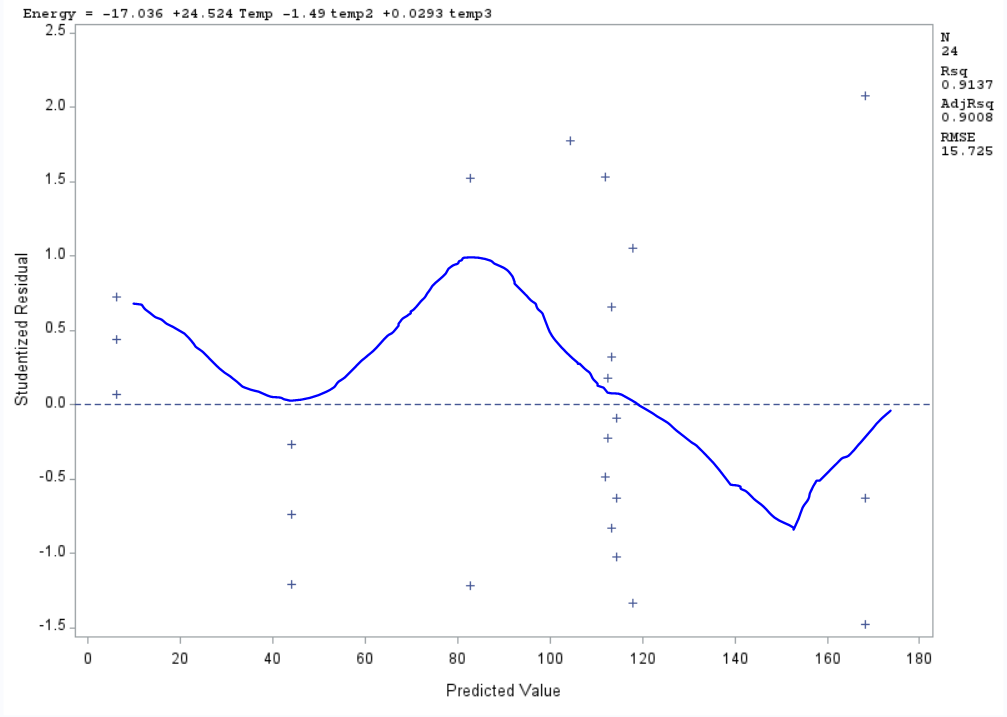
Here, Temp2 is the Temp\*\*2 and Temp3 is Temp\*\*3. This **final model can explain up to 90.08% variance in Energy**. Also, the 2nd and 3rd order terms are significant and hence they have to be included in the model.

1. Yes, based on the above-mentioned output all the variables mentioned in the above model are significant.
2. The Normality plot is as follows:



From the above plot we can see that the linearity is present and there aren’t any significant outliers which are visible form the above normality plot.

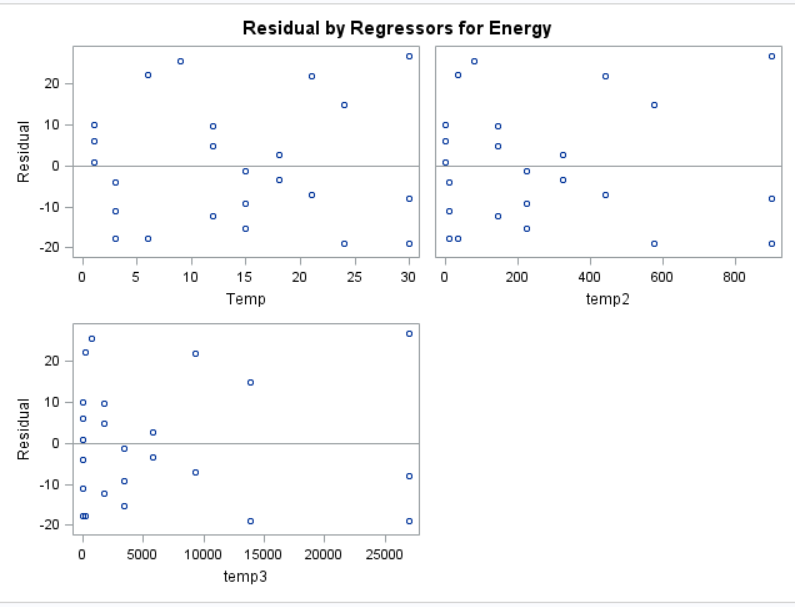
The Studentized Residual plot vs the Predicted Values is as follows:



We can see that the points are somewhat randomly distributed across the plot. There is a significant pattern which can be observed i.e. a somewhat sinusoidal wave which means that there isn’t a constant variance and that the points are not completely independent of each other. There are no outliers present in this dataset.

The Residual plots wrt each regressor is as follows and we can make the following observations from the plots:

* There isn’t a significant pattern which can be observed in the plots.
* The points seem to be randomly scattered over the plot and hence we can say that somewhat constant variance can be observed for the temp and temp2 plots. We can also say that the points are mostly independent of each other.
* But in the case of temp3 we can say that the points are mainly concentrated in the 0-5000 range which means that the points don’t have a constant variance and the points are not completely independent of each other.



The code to generate all the plots is as follows:

title "Normality Plot";

**proc** **reg**;

model energy = temp temp2 temp3;

plot npp.\*student.;

**run**;

title "Residual and Influential point analysis";

**proc** **reg**;

model energy = temp temp2 temp3/r;

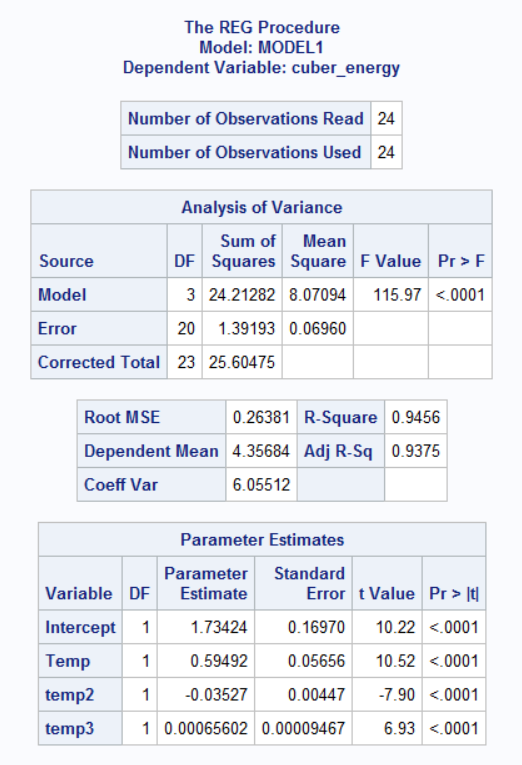
plot student.\*predicted.;

**run**;

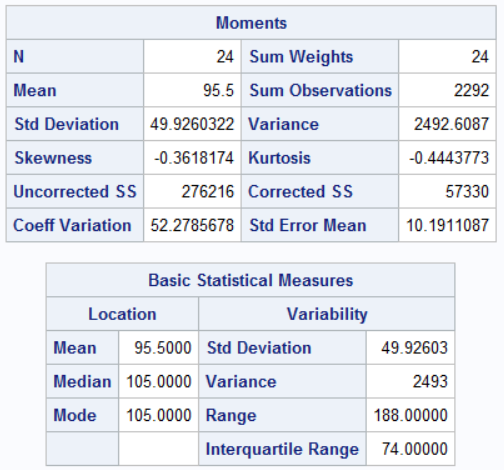
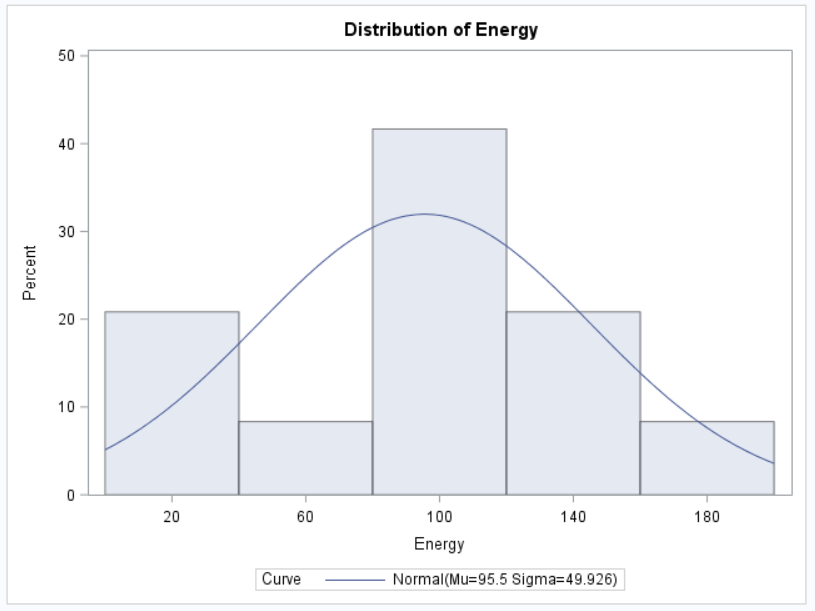
NOTE: The individual plots were taken from the output in which I had generated the studentized vs Predicted Residual values.

1. I am not satisfied with the model and hence decided to experiment with transformations for Energy Variable. I had tried Log, Inverse, Squared, Square root and Cube root transformation for Energy variable and observed that the final model which resulted by applying Cube root transformation to energy variable had an Adjusted R-Squared value of **0.9375 i.e. 93.75% variance in energy could be explained by applying cube root transformation on energy.**

The output is as follows:

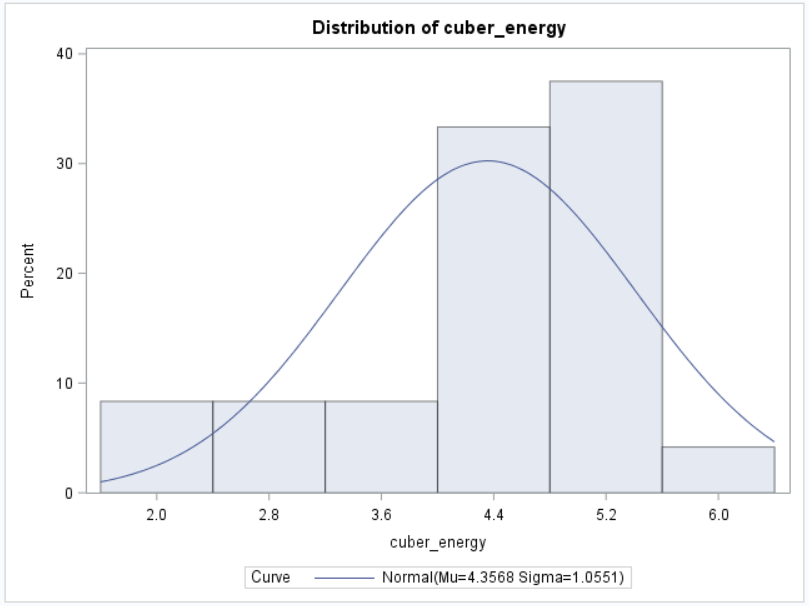


**The histogram of energy initially was as follows:**



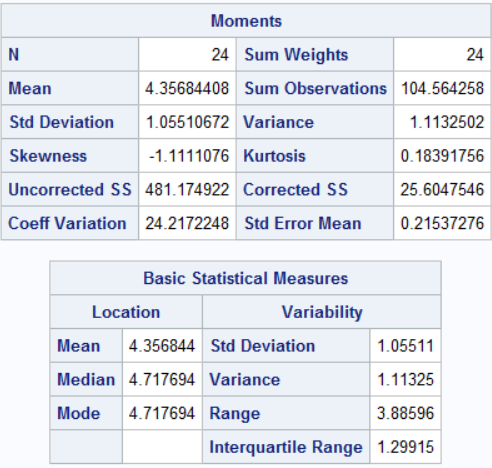
From the basic statistical Measures table, we can see that the difference between mean and Median is somewhat large and from the histogram we can see that nearly 22% of the data are at the starting and hence I felt a need to apply transformation on Energy.

**After applying Cube root, Transformation on Energy** I got the following Histogram:



Here we can see that the values have been centered.

The other relevant output are as follows:



To apply the transformation and to generate the histogram I’ve used the following code:

**data** Energytemp;

set Energytemp;

cuber\_energy = energy\*\*(**1**/**3**);

**run**;

**proc** **univariate** normal;

var cuber\_energy;

histogram/normal (mu=est sigma =est);

**run**;

**proc** **reg**;

model cuber\_energy = temp temp2 temp3 ;

**run**;

title "Normality Plot";

**proc** **reg**;

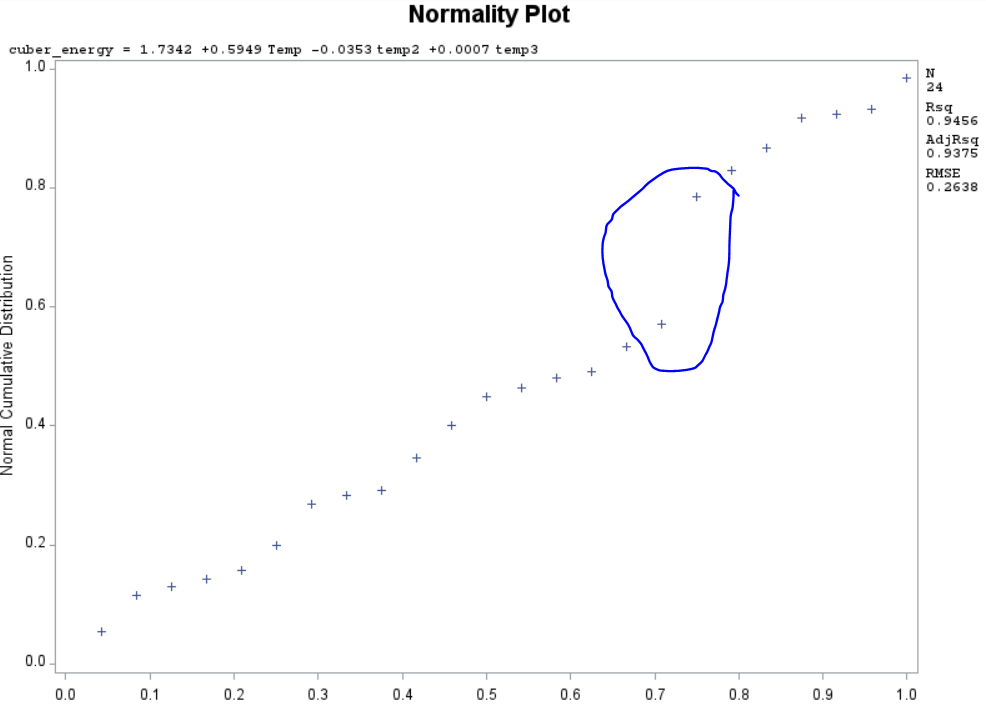
model cuber\_energy = temp temp2 temp3;

plot npp.\*student.;

**run**;

The output for the transformed Energy regression model was mentioned at the start of the question.

The normality plot for this model is as follows:



Here we can see that there is a small jump which has been circled and apart from that the normality plot seems to be a straight line.

The final model equation after applying the transformation is as follows:

**CUBER\_ENERGY = 1.7342 + 0.5949\*TEMP – 0.0353\*TEMP2 + 0.0007\*TEMP3**

1. For the prediction, I’ve used the model which I generated initially i.e. the model in which I **didn’t apply transformation to the energy variable**.

To perform the prediction in SAS I ran the following code:

TITLE "PREDICTION";

**DATA** pred;

INPUT Energy temp;

temp2 = temp\*\***2**;

temp3 = temp\*\***3**;

DATALINES;

. 10

;

**proc** **print**;

**run**;

**DATA** new;

SET pred energy;

temp2 = temp\*\***2**;

temp3 = temp\*\***3**;

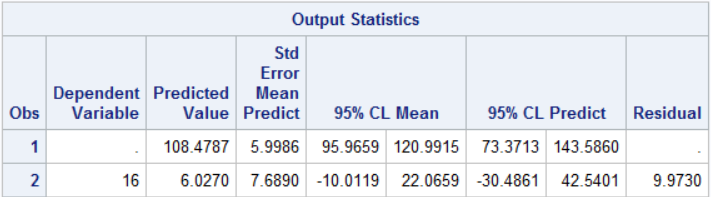
**RUN**;

**PROC** **reg** data=new;

model energy = temp temp2 temp3 /p cli clm;

**RUN**;

we get the following output i.e. the predicted value is as follows:



And hence the predicted Energy value is **108.4787 with 95% confidence interval (73.37, 143.5860) with a prediction interval of (73.3713, 143.5860).**

1. The whole SAS code for this problem is as follows:

title "Energy Temperature dataset";

**data** energyTemp;

infile "energytemp.txt" delimiter =' ' missover firstobs=**2**;

input Energy Temp;

temp2 = temp\*\***2**;

temp3 = temp\*\***3**;

**proc** **print**;

**run**;

title "Scatterplot Energy vs Temperature";

**proc** **gplot**;

plot Energy\*temp;

**run**;

title "Final Model";

**proc** **reg**;

model energy = temp temp2 temp3 / selection = stepwise;

**run**;

title "Normality Plot";

**proc** **reg**;

model energy = temp temp2 temp3;

plot npp.\*student.;

**run**;

title "Residual and Influential point analysis";

**proc** **reg**;

model energy = temp temp2 temp3/r;

plot student.\*predicted.;

**run**;

title "Transformation on Energy";

**data** Energytemp;

set Energytemp;

cuber\_energy = energy\*\*(**1**/**3**);

**run**;

title "Histogram";

**proc** **univariate** normal;

var cuber\_energy;

histogram/normal (mu=est sigma =est);

**run**;

title "Regression Model";

**proc** **reg**;

model cuber\_energy = temp temp2 temp3 ;

**run**;

title "Normality Plot";

**proc** **reg**;

model cuber\_energy = temp temp2 temp3;

plot npp.\*student.;

**run**;

TITLE "PREDICTION";

**DATA** pred;

INPUT Energy temp;

temp2 = temp\*\***2**;

temp3 = temp\*\***3**;

DATALINES;

. 10

;

**proc** **print**;

**run**;

**DATA** new;

SET pred energy;

temp2 = temp\*\***2**;

temp3 = temp\*\***3**;

**RUN**;

**PROC** **reg** data=new;

model energy = temp temp2 temp3 /p cli clm;

**RUN**;