

CSC 423: Data Analysis & Regression

(Spring 2017)

Weather Data Analysis of Szeged



**Guide**: Prof. Nandhini Gulasingham

**Team Members:**

1. Jaymin, Parekh
2. Keertika, Rai
3. N K Vidya Shanmukha Maruthi Sasidhar, Mukthinuthalapati
4. Vishakh, Lakshmikanth

**Abstract**

This project aims to analyze and predict the temperature (dependent variable) based on the values of various parameters (independent variables) in the given geographic location.

To perform this analysis, we would be using the linear regression methodologies and principles which we learned during the CSC423 course. For this, we collected weather data for the duration between January till December 2016 of Szeged, Hungary.

During this project path, we split the whole dataset into four individual quarters based on the months. This was then split into training and test sets (80:20). Using the training set we generated the various models thus enabling us to identify the underlying relation between temperature and other parameters. Based on the backdrops of each model, we made changes by applying transformations and the inclusion of interaction terms.

Then we used the final significant model to predict the temperature on the given test data set and computed the performance metrics.

**Introduction**

The weather data for Szeged, Hungary was collected by using DarkSky.Net API which makes use of various weather-related sensors to read the values of various influential variables such as humidity, air pressure, wind speed and direction, visibility and cloud cover. The data is recorded on an hourly basis everyday throughout the year.

This data was then portioned among the team members based on months (quarterly). Each member then divided their respective dataset into two parts based on the 80:20 split ratio in which the larger portion was the training set and the remaining part was considered as test set. Then later everyone performed regression analysis on their respective quarters and came up with two unique models and then tested those models on the test set.

The outcomes and observations of this analysis can be used by the citizens to improve their understanding of local weather. These outcomes can further be used by the tourists and tourism department of Hungary for better tourist experience.

**Methodology**

**Methodology I:**

1. **Data cleaning and analysis:**

I have chosen to work on months from January to March 2016. This data set is extracted using proc sql command from the formatted date. Formatted date has date and time together so we have decided to separate out to get the months. After equally splitting the data, I come know that average temperature in 3 months were 4.5 C. Apart from this, the frequency of snow and rain is approximately 468 to 1715 which shows that there is no other value present in my data set for recip type.

Final variable list used after model 1 construction:

Temperature: this is dependent variable tells about the temperature

Humidity: the amount of moisture present in the air

Wind Speed (Km/h): the wind flow in velocity caused due to the difference in pressure

Visibility (km): distance through which person can identify certain object

Pressure millibars: measure of air pressure

Time numeric = separated time from formatted date

Snow:

Clr (clear)

MC (Mostly Cloudy)

OC (overcast)

1. **Data splitting:**

The data splitting is done in two parts, 80 % of training and 20% of testing set. It has 1747 and 436 observation respectively. As per my team mate discussion, the model which we have developed is to predict the temperature by using our course knowledge of regression and analysis. The training set is defined as weather\_Train1 and for testing part I have extracted the records from the Nwether\_123 (total set 100% for 3 months) dataset.

1. **Exploratory Data Analysis**

Histogram and Boxplot Analysis:

* This is the step where we need to analyze histogram and residual analysis in order to signify the variable for our individual model. I have run the histogram for temperature and I come to know that it is almost symmetric in nature (Fig. J-1).Box plot analysis has given the general overview of temperature distribution in the specified country.

Residual Analysis:

* Next part would always be the residual analysis for the training set. Based on the observation, I got the normality plot to be little bit bend in the end Hence, I have decided to apply transformation (y2 ) to the training dataset (Fig. J-2). What I have found is that it has decreased the adj R2 and R2 of my model. As a consequence of this step, I have planned to apply different selection method to check the residual plots for further exploration.

The results of correlation is fascinating because I didn’t get any multicollinearity issue for my part. However I have excluded the apparent temperature in my model because what I have learned so far is that when we have two variables with similar connotation it can cause the multicollinearity. I am planning to explore that variable later for my future enhancement of this project. As far as scatterplot is concerned, visually it difficult to address the issue related to the strong association for my model so I have decided to check the P value and VIF values for filtering out the variables.

The results of two selection methods for training set are below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | Adj R2 | R2 | RMSE | No of predictors |
| **Stepwise** | 0.6825 | 0.6841 | 9.25 | 9 |
| **Backward** | 0.6825 | 0.6841 | 9.25 | 9 |

Table P-1: comparison of two methods

Selected Method for Training set

As you can see that, I got similar results for my method selection. After that I have decided to go with stepwise selection method (Training Model 1) for further analysis.

Variable selection and Outlier Analysis

* After checking the null hypothesis test and evaluating p (must be α<0.05) value of my model I have come to know that all variables which I have gathered from selected model are significant. List of fixed variable for model building is below:

Snow, Humidity, OC (over cast), time numeric, Wind Speed (km), MC (mostly cloudy), Visibility (km), Pressure (millibars), Clr (clear).

* AS per the Fig. J-4, I have detected less than 5 variables as Influential and outliers. Out of 1747 records few of them are detected as outliers. In my analysis, I have decided to keep those for model building process.

Since the values of adj R2 and R2 are similar I was in the situation where I couldn’t take effective decision for model selection. However, the Interaction terms is another way to boost up the model performance so I have used the following Interaction terms for model construction. One of the research study also have extensive iteration process to choose the variable it doesn’t have similar scenario but it has helped me to understand the process of linear regression1

hwp = Humidity\*Wind Speed (km/h)\*Pressure (millibars)

hs = Humidity\*Snow

ps = Pressure(millibars)\*Snow

tmc  = time numeric\*MC

vtmc = Visibility(km)\*time numeric\*MC

ptmc = Pressure(millibars)\*time(numeric)\*MC

vsmc = Visibility(km)\*Snow\*MC

hoc = Humidity\*OC

htoc = Humidity\*time numeric\*OC

soc = Snow\*OC

Model 2 – Interaction Model

I got the Adj R- square improved by 3% after adding interaction terms. It is my improved model for the training set. I have checked the linearity for the model but it has slightly different at the end side apart from this the distribution is considerable compare to model1.

Equation of Model 1:

Temperature\_c = -11.93606(Humidity) + 0.06109 (Wind\_Speed\_km\_h\_) -0.07868(Visibility\_km\_)-0.00300 (Pressure\_\_millibars\_)+ 0.00001752(time\_numeric -) + 9.00278(Snow) -0.86333(Clr) 0.91424(MC) + 1.62742(OC);

Equation of Model2:

Temperature\_c = 13.64725 -11.27678(Humidity) +0.28465(Wind\_Speed\_km\_h\_)-0.0002(hwp) +15.6 (hs) -0.020 (ps) +0.00084(tmc) -0.00000233(vtmc) -7.77154E-7(ptmc) -0.12001(vsmc) +2.25115(hoc) +0.00002376htoc -3.35399(soc)

Prediction for selected model (Model2 of training Model):

TITLE ‘prediction model 2 training set 95% confidence interval';

**data** pred;

input Temperature\_c Humidity Wind\_Speed\_km\_h\_ hwp hs ps tmc vtmc ptmc vsmc hoc htoc soc;

datalines;

. 1.00 11.1090 1 0 0 0 1 1 0 1 1 1

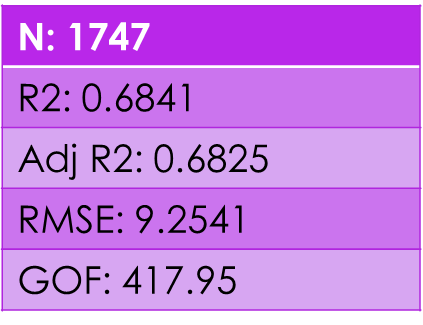
;

The result of the prediction is 4.42 C for the 11.10 wind speed and 1.00 humidity (Fig. J-6)

1. **Validation test**

* In order to test the model performance we need to check validation test for 30% observation of the current data set. What are the criteria to check those values? We can compare our testing set on RMSE, MAE and R-Square with the Training set. For this procedure I have generated phat vale for the testing set and subtracted with my original temperature variable. It can simplify the prediction values for the testing data set which is also utilized for unseen data for prediction.
* Next step would be evaluating RMSE, MAE value by differentiating the original Temperature to predicted Yhat value. It will give us the performance for model 1 and 2 for testing dataset (Fig.7 and 8).
* At last we can check K fold cross validation to check error progression and it will stop when Error start to increase.
* ASE plots is one of the way in which we can check the error progression for training and testing set which I have attached in the appendix (Fig.J-8)

**Performance Matrix for both Model 1 and 2 -Training set Testing set**

****

****

****

In comparison we can say that my model 2 has considerable performance in terms of lower RMSE value compare to training model 1

However, I can also check the model performance by including subject matter expert because he/she has better domain knowledge of weather prediction. As per the testing performance both model has lover CV R Square (must be less than 0.3) which means both can perform better.

As an individual I can include the apparent temperature to see if the prediction for visiting the country is going to improve or not.

References

1. Debasis Kundu, G.Murli. Model selection in Linear Regression. Computational Statistics & Analysis 22, 1996. Page 461-469.

**Screenshots:**

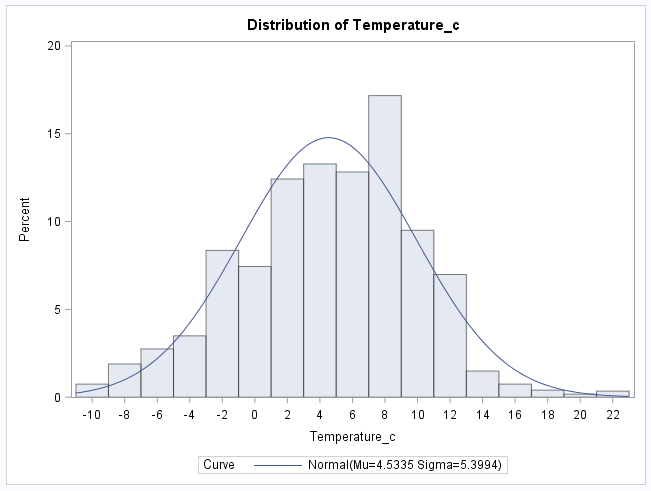
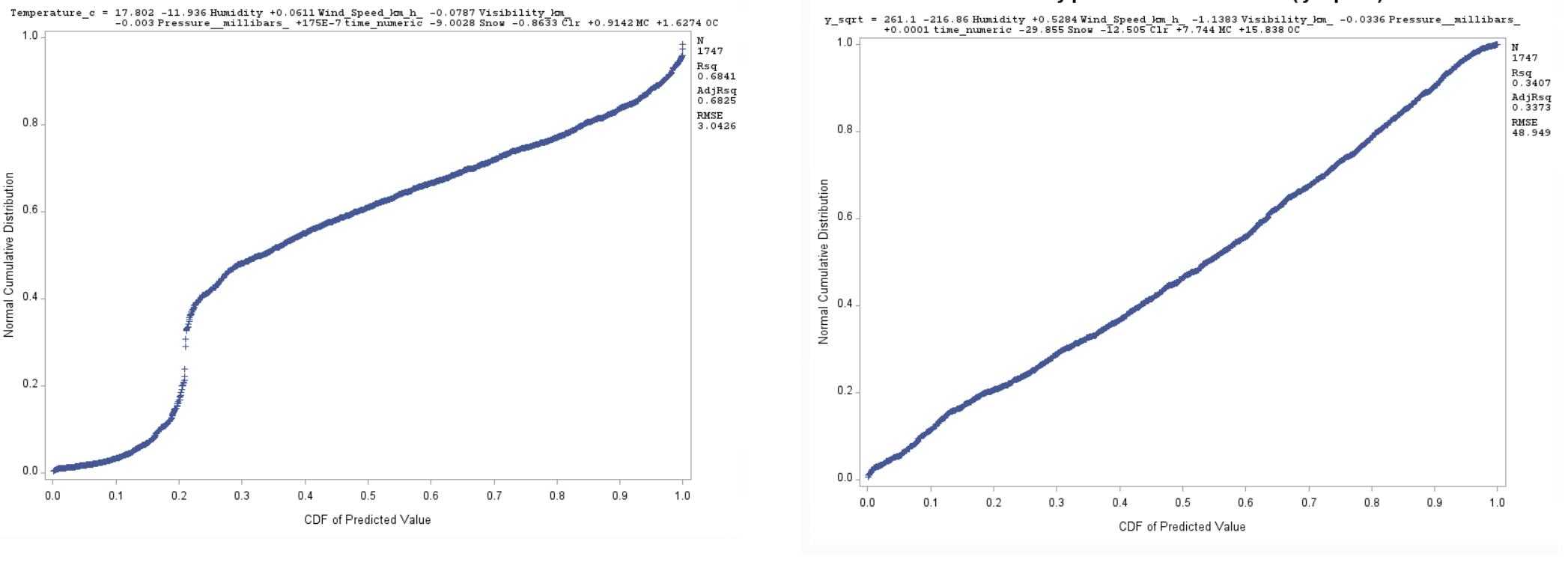


Figure: J-1 Histogram of Training set [Temperature in C]



Transformation

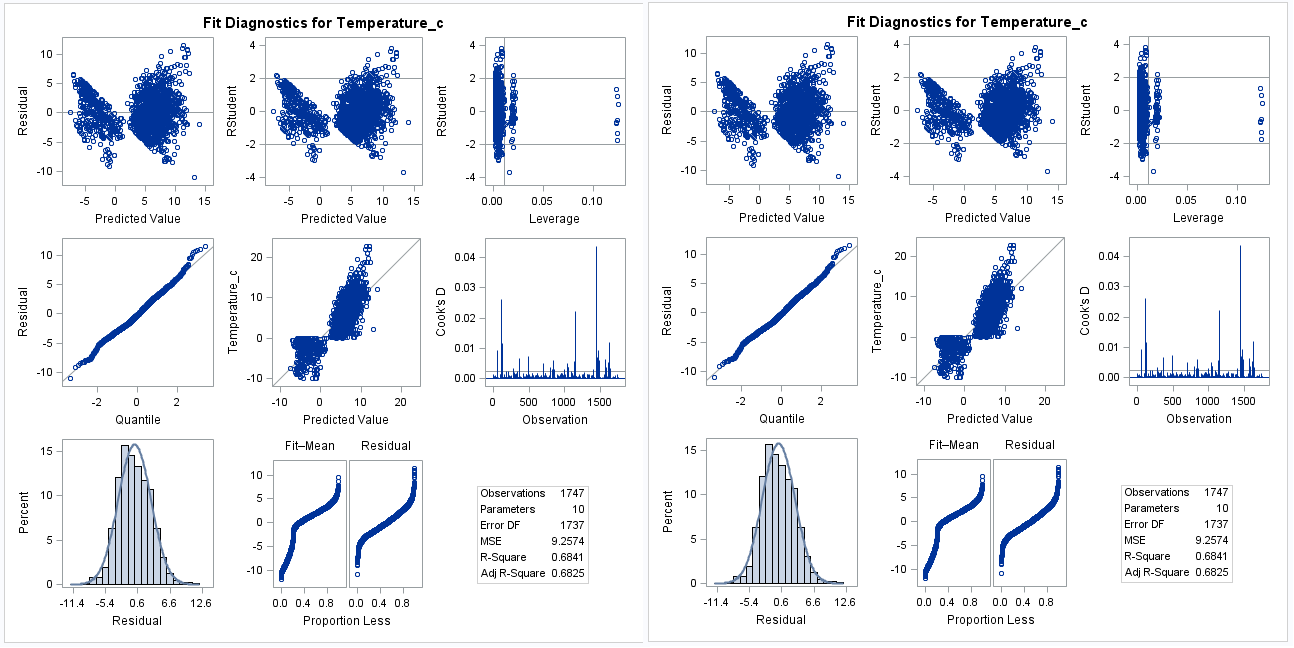
N:1747

R2:0.3407

Adj R2: 0.3373

RMSE: 48.949

Figure: J-2 applied transformation [Residual Analysis]



Applied Backward – training set

Applied stepwise – training set

Figure: J-3 Comparison of two methods [Model selection]

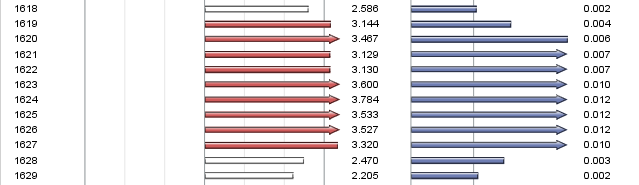
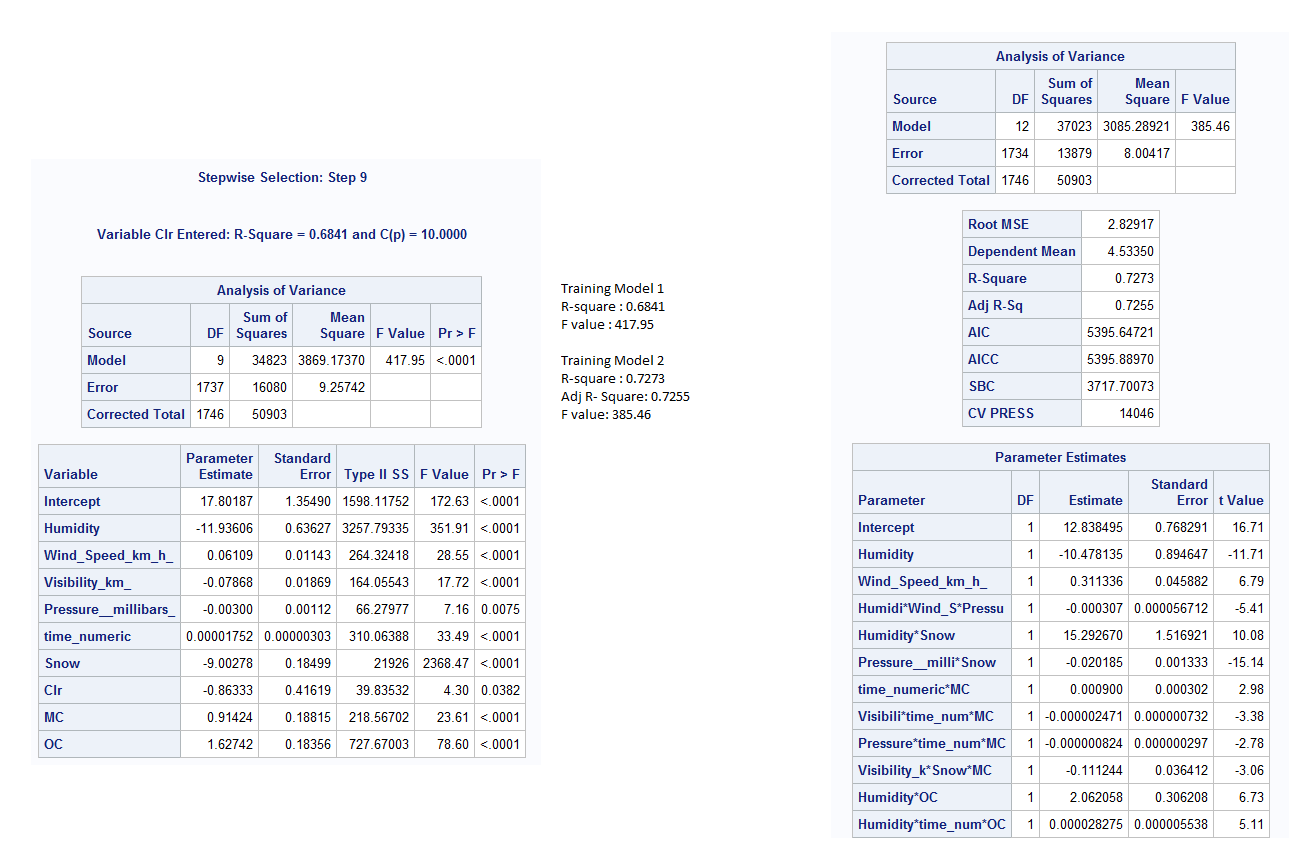


Figure: J-4 Influential points and outliers



Training Model 1 - Stepwise

Training Model 2 – Interaction Model

Figure: J-5 internal model comparison

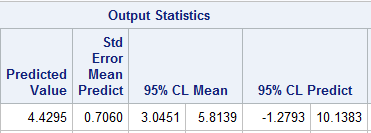


Figure: J-6 Training Model Prediction

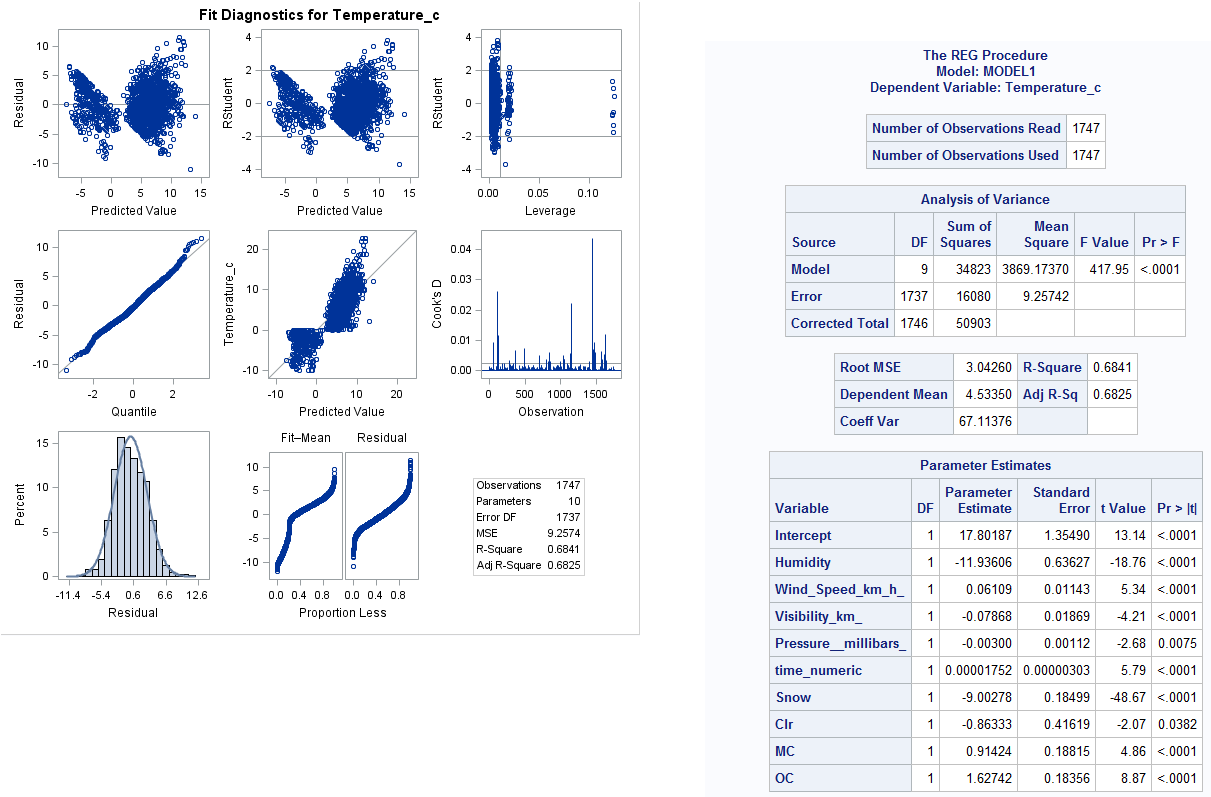


Figure J-7: Comparison of training and testing model1

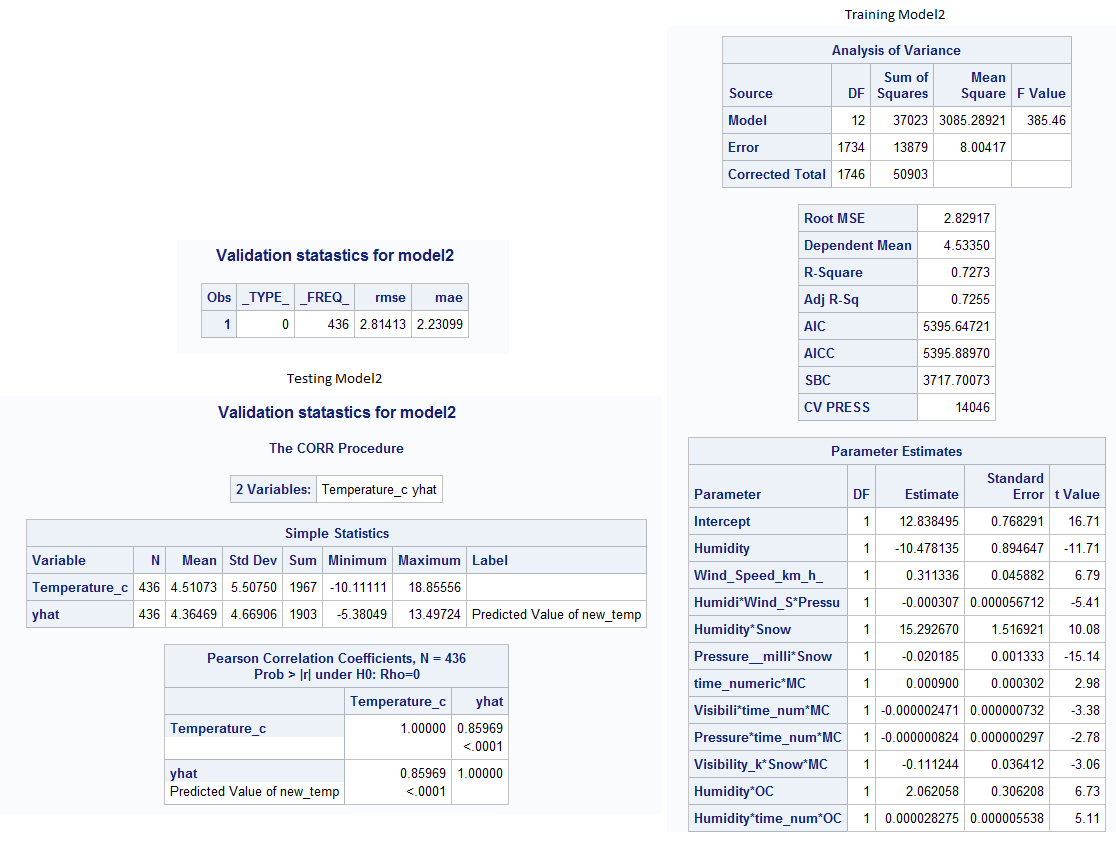


Figure J-8: Comparison of training and testing model2

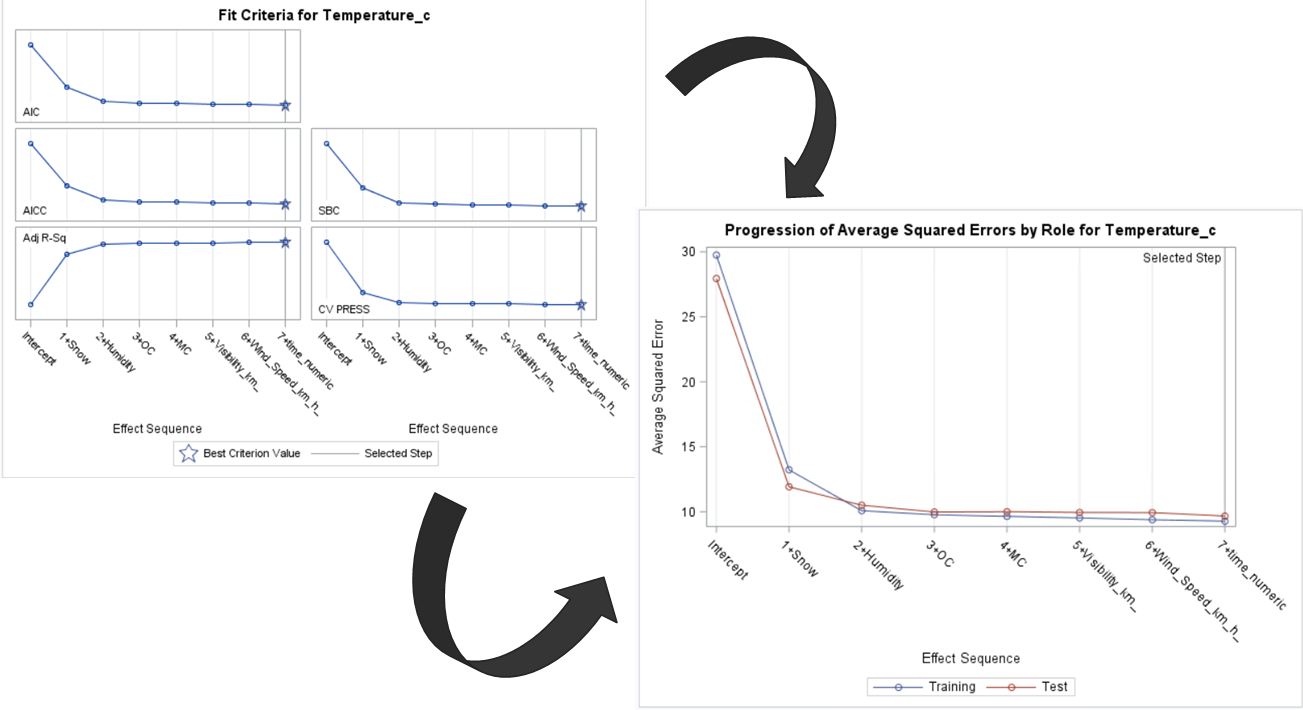


Figure J-9: ASE plots for Training and Testing

**Methodology II:**

1. **DataSet Description:**

In this dataset, we have the following variables which are qualitative and hence would require us to create dummy variables, the qualitative variables are the following:

* Precipitation type: This takes two values ‘Rain’ and ‘Snow’ and hence we would be required to create a single dummy variable whose base case is ‘Rain’ and for it the dummy variable will have a value of 0. For ‘Snow’ the dummy variable will have a value of 1.
* Summary: This column has 14 different types of values and would require us to create 13 variables. The dummy variables have been named as B1 till B13. For each value of Summary, the associated dummy variable will be set to 1 and all other will be set to 0. For this Base case has been chosen as Mostly Cloudy since it had the most number of occurrences. For the Base case, all the dummy variables will be set to 0.
  + If Summary = Breezy then Breeze=1 and rest are 0’s.
  + If Summary = Breezy and Mostly Cloudy then B\_MC = 1.
  + If Summary = Breezy and Overcast then B\_OC = 1.
  + If Summary = Breezy and Partly Cloudy then B\_PC = 1.
  + If Summary = Clear then clr = 1.
  + If Summary = Drizzle then Drizzle = 1.
  + If Summary = Foggy then Fog = 1.
  + If Summary = Humid and Overcast then H\_OC = 1.
  + If Summary = Humid and Mostly Cloudy then H\_MC = 1.
  + If Summary = Light Rain then Lgt\_rain = 1.
  + If Summary = Overcast then OC = 1.
  + If Summary = Partly Cloudy then PCld = 1.
  + If Summary = Rain then RN = 1.
* Interaction Terms: For my model analysis, I have generated and included the following interaction terms;
  + IV1 = Interaction term that combines the effect of Apparent Temperature and Wind Speed.
  + IV2 = Interaction term that combines the effect of Humidity and Wind Speed.
  + IV3 = Interaction term that combines the effect of Apparent Temperature, Wind Speed and Humidity.
  + IV4 = Interaction term that is been obtained by the interaction between Apparent temperature, time and Most Cloudy atmosphere.
  + IV5 = Interaction term that is been generated due to the interaction between Humidity, time and Most Cloudy atmosphere.
  + IV6 = Interaction term that is created due to the interaction between Pressure, time and Most Cloudy atmosphere.

1. **Data Cleaning:**

In my dataset, I had noticed many null values for Precip\_Type variable and according to the principles of Data Analysis & Regression, It is not favorable to remove all the null observations so I performed a background research of weather in Szeged thus overcoming the null values issue in two steps that includes, understanding the real-time geographical weather and analyzing the frequency of each Precip\_Type in the dataset. Hence after the analysis I concluded by replacing all the null values with Rain value.

1. **Data Split/Partition:**

The dataset has been split into two parts Training set which contains 1748 records and which constitutes up to 80% of the data. The Test set contains 20% containing 436 records.

The training set will be used to come up with the best models which explain the maximum variance for the Temperature variable. The models which we came up using the training set will be used to come up with the predicted value for temperature using the Test set.

1. **Data Analysis:**

In this data analysis phase, I generated the histogram to verify the distribution of the records along the dataset and Pearson Correlation to infer the association among the predictors. Histogram also provides an insight for the presence of outliers.

From the Histogram, we can infer that the distribution is quite symmetric and possess the normal distribution for Temperature variable which is our prediction variable. The histogram does not show any presence of notable outlier and thus the below statistics of Histogram [Fig: M1 (a)]:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Mean | Median | Mode | Kurtosis | Q1 | Q2 | Q3 | Q4 | IQR |
| 16.97 | 16.72 | 16.08 | -0.4 | 12.64 | 16.71 | 21.42 | 34.55 | 8.78 |

Thus, the kurtosis value we can observe that the histogram possesses flat tail distribution.

From the box plots [Fig: M1 (b)] of Temperature v/s Wmonth, we can observe that temperature in gradually increasing from the month of April to June and we can conclude that Szeged is having its highest temperature in the month of June (acc. to my dataset). Through the analysis of another box plot of temperature v/s time numeric, we can infer that the temperature is cooler from 12am to 7 am and later it gradually increases and at the end of the day from 7pm the day starts to get cooler (The highest temperature of the day can be noticed at 2pm). To extend the interest and better understanding the influence of other variables on temperature regarding the geographic location I finally generated the boxplot of Humidity v/s time [Fig: M1 (c)] thus observing that the humidity will be at its highest peak during the time period of 12am to 9am and later gradually decrease till 5pm later starts to increase when the temperature starts to decrease.

As a part of Data Exploration, I generated the Pearson Correlation plot/ table [Fig: M1 (q)] in which the highest positive association/correlation was found between (Temperature, Apparent Temperature) with an association value of 0.99.

Apparent temperature can alone define the 99.3% of correlation with Temperature variable. The scatter plot is not much of informative when compared to Pearson Correlation Tabulation since it contains all the predictors.

1. **Model Construction:**

**MODEL 1: (Training Set)**

From my dataset, I have built the model that consists of base variables and dummy variables thus making my Model 1.

Model Statement for Fitted Model (unitedly consists of significant variables as a Training Set) [Fig: M1 (e)]

Temperature\_c = 2.1061 + 0.9041\*(Apparent\_temperature\_\_C\_) – 0.80367\*(Humidity) + 0.0160\*(Wind\_Speed\_km\_h\_) + 1.317\*(B\_OC) + 0.160\*(MC) + 0.9682\*(OC) + er

To achieve the final regression equation/model, I utilized the Forward, Backward, Stepwise, AdjRsrq selection methods, out of which AdjRsq and Stepwise gave the same efficiency value but I selected with the Stepwise method since it had the least number of predictors when compared to AdjRsrq selection method with an AdjR2 = 99.08% and RMSE =59.97%

Then later I removed the remaining insignificant predictors based on the Pr > |t| value which has the value greater than 0.05

Later, I executed my final model with ‘stb’ option which generated the standardized estimates for each final significant predictor thus showcasing that the apparent temperature variable had the highest influence on the Y-variable (Temperature) and the least influential variable is OverCast.

As a part of data exploration and verification, I ran the model with ‘vif’ & ‘tol’ option that generates the vif value used to verify the multicollinearity and tolerance value. According to my data model, the vif value of all the predictors was below the threshold value (vif<10) hence proving that my model didn’t have any multicollinearity among variables.

Then later I verified the presence of outliers, influential points in the fitted model and later I investigated the residual plots to verify the model assumptions.

While verifying the outliers & influential points, there were many observations that were both outlier and an influential point so, I removed those observations that were both outlier & influential point.

After investigating the residual plots, we can infer that Studentized\_residual v/s predicted [Fig: M1 (g)]and studentized\_residual v/s apparent temperature[Fig: M1 (h)] does not have constant variance, independence, linearity or normality. Studentized\_residual v/s Humidity[Fig: M1 (j)] acknowledges the principles of Constant variance and independence.

But, in the normality plot[Fig: M1 (l)] we can clearly identify the presence of ‘s’ shape at the top (beginning) and at the left bottom thus mentioning that the distribution of errors in the model is uneven. The normality probability plot does not comply with the principles or Normality and linearity.

**MODEL 1: (Testing Set)**

Due to an extensive analysis from the above-mentioned training set, I learnt the characteristics of the dataset and I have incorporated the techniques and model understandings into the test set thus enabling me to predict the desired predictor. In my case I am aimed at predicting the Temperature variable and I ran the prediction code block thus generated the predicted Temperature value of 12.39 which gets falls within the confidence interval range of (11.29, 13.49) with 95% confidence Interval. [Fig: M1 (p)]

**Validation Testing:** [Fig: M1 (r)]

From my validation Testing on Model 2, I can infer that the model 2 is highly stable and efficient so as it can be used for accurate prediction. The validation test numeric results can be found as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| CV R2 | R2 | MAE | Yhat |
| 0.00048 | 0.9908 | 0.4165 | 0.9956 |

I was successful in deriving the CV-R2 value by squaring the Yhat value and subtracting the R2 value of training set. From the above tabulation, we can observe that the CV-R2 value is <0.3 thus making this model efficient.

**MODEL 2: (Training Set)**

From my dataset, I have built the model that consists of base variables, dummy variables and newly calculated Interaction variables/terms thus making my Model 2.

I have used the ‘GlmSelect’ procedure with stepwise selection and cv as stop option to generate the 3rd degree (Cubic) interaction terms.

Model Statement for Fitted Model (unitedly consists of significant variables as a Training Set): [Fig: M2 (b)]

Temperature\_c = 0.539 + 0.9799\*(Apparent\_temperature\_\_C\_) – 0.4527\*(Humidity) – 0.05872\*(Wind\_Speed\_km\_h\_) + 0.7342\*(B\_MC) – 0.4673\*(B\_OC) + 0.12608\*(Clr) + 0.17861\*(MC) + 0.08683\*(PCld) + 0.00474\*(IV1) + 0.2787\*(IV2) – 0.1806\*(IV3) + 7.551891E-7 \*(IV4) + 0.00001373\*(IV5) - -2.25483E-8 \*(IV6) + e

To achieve the final regression equation/model, I utilized the Forward, Backward, Stepwise, AdjRsrq selection methods, out of which Backward and Stepwise gave the same efficiency value but I selected with the backward method since it had the least number of predictors when compared to stepwise selection method with an AdjR2 = 99.48% and RMSE =45.124%

Then later I removed the remaining insignificant predictors based on the Pr > |t| value which has the value greater than 0.05

Later, I executed my final model with ‘stb’ option which generated the standardized estimates for each final significant predictor thus showcasing that the apparent temperature variable had the highest influence on the Y-variable (Temperature) and the least influential variable is Clr.

As a part of data exploration and verification, I ran the model with ‘vif’ & ‘tol’ option that generates the vif value used to verify the multicollinearity and tolerance value. According to my data model, the vif value for Wind speed increased drastically due to the addition of interaction variables. Then later I tried the technique to center the variable by removing the mean of the Wind\_Speed and then again, I calculated the interaction variables but it was unfortunate that the vif value was decreased by 2% which was not efficient and mainly after centering the variable my AdjR2 decreased drastically which was unfavorable. So, I decided to retain the Wind\_Spped variable along my final model. The tolerance value for all the variable was within the desirable limits.

Then later I verified the presence of outliers, influential points in the fitted model and later I investigated the residual plots to verify the model assumptions. [Fig: M2 (d)]

While verifying the outliers & influential points, there were many observations that were both outlier and an influential point so, I removed those observations that were both outlier & influential point. [Fig: M2 (c)]

After investigating the residual plots, we can infer that Studentized\_residual v/sHumidity [Fig: M2 (e)] and studentized\_residual v/s Wind\_Speed [Fig: M2 (f)] have constant variance and independence. Studentized\_residual v/s Predicted [Fig: M2 (g)] and studentized\_residual v/s apparent temperature [Fig: M2 (h)] does not have constant variance and independence since majority of the observations are not distributed evenly and has a ‘u’ shape in the residual plot.

Fortunately, due to the addition of interaction terms the ‘s’ shape which was found in the normality plot of model1 is enhanced and re-corrected thus making to a straight 45o line in npp plot of model 2. The normality probability plot of model 2 comply with the principles of Normality and Linearity [Fig: M2 (l)].

**MODEL 2: (Testing Set)**

Due to an extensive analysis from the above-mentioned training set, I learnt the characteristics of the dataset and I have incorporated the techniques and model understandings into the test set thus enabling me to predict the desired predictor. In my case I am aimed at predicting the Temperature variable and I ran the prediction code block thus generated the predicted Temperature value of 12.35 which gets falls within the confidence interval range of (11.46, 13.24) with 95% confidence Interval [Fig: M2 (q)].

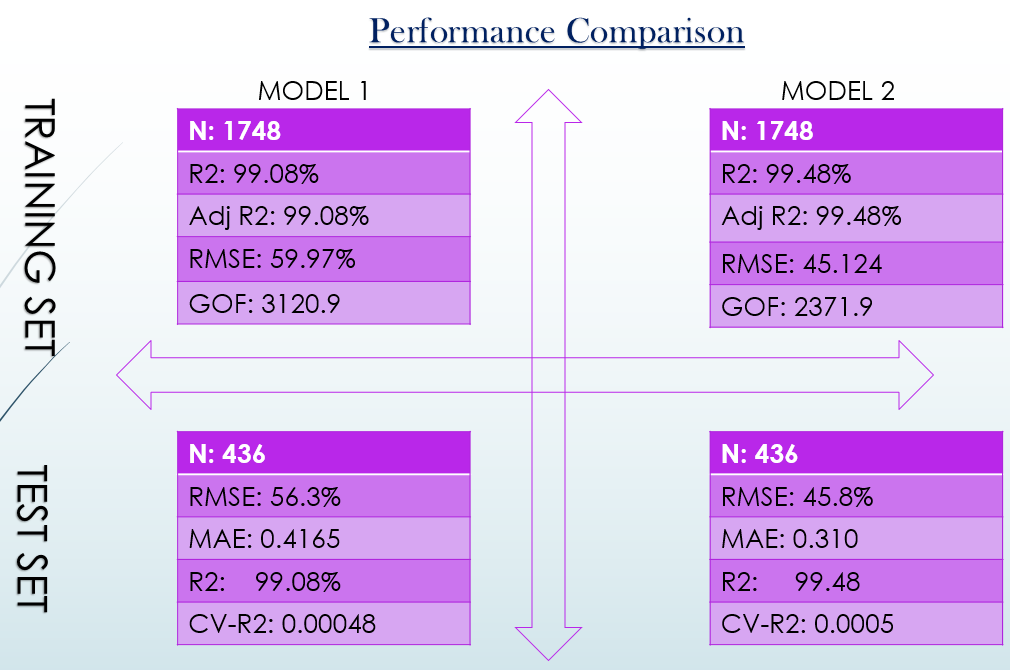
**Validation Testing:** [Fig: M2 (r)]

From my validation Testing on Model 2, I can infer that the model 2 is highly stable and efficient so as it can be used for accurate prediction. The validation test numeric results can be found as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| CV-R2 | R2 | MAE | Yhat |
| 0.00055 | 0.9948 | 0.310 | 0.99712 |

I was successful in deriving the CV-R2 value by squaring the Yhat value and subtracting the R2 value of training set. From the above tabulation, we can observe that the CV-R2 value is <0.3 thus making this model efficient.

**Performance Metrics:**



1. Comparison between the Training Sets:

As per the performance metrics, we can conclude that Model 2 is better since the AdjR2 is higher.

1. Comparison between Testing Sets:

As per the performance metrics, we can observe that both the test models are good models since the CV-R2 is less than 0.3 thus making it stable for prediction.

1. Comparison between Model 1 Test & Training Sets:

From the above metrics we can observe that Model 1 Test set is efficient since the RMSE value is less when comapred to training set.

1. Comparision between Model 2 Test & Training Sets:

From the above metrics, we can observe that Model 2 Training set is better since the RMSE value is less when compared to test set.

SnapShots:

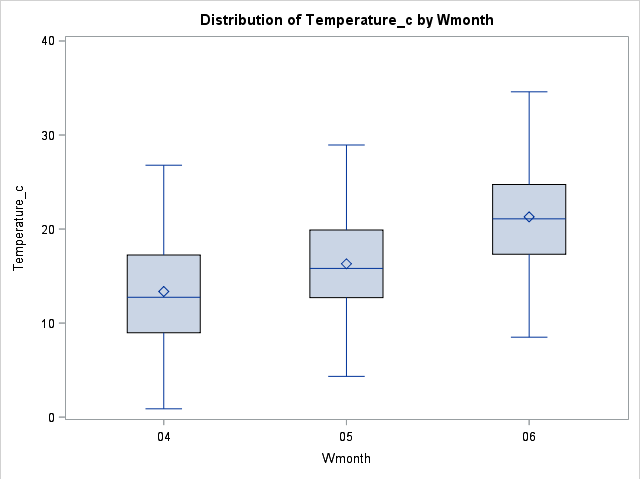
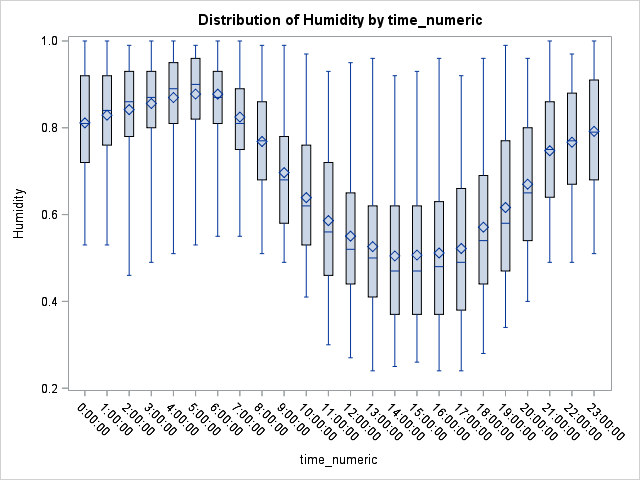
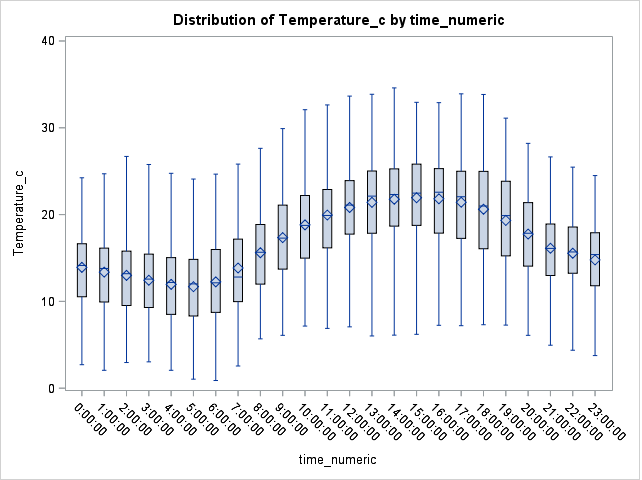
 

Fig: M1 (a) Fig: M1 (b) Fig: M1 (c)

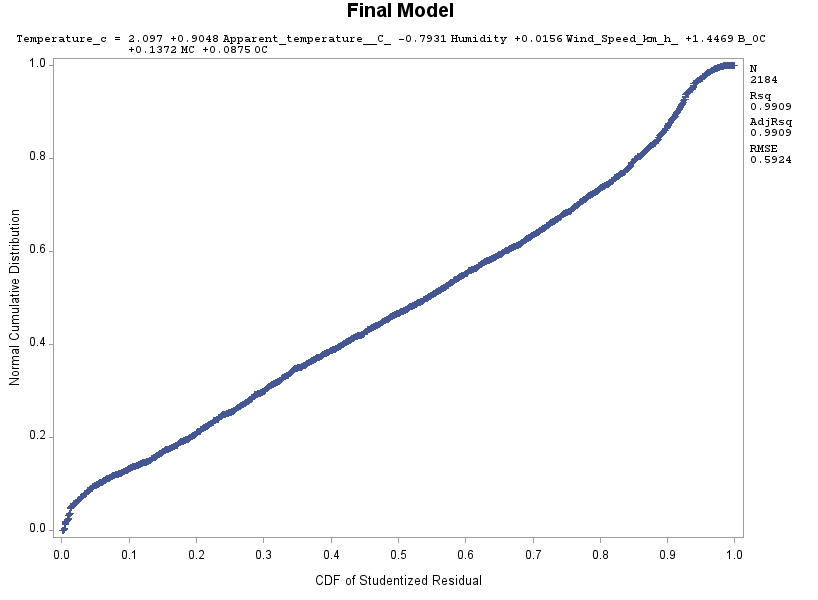
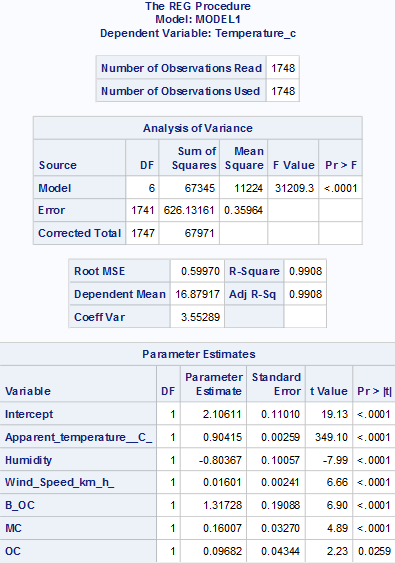
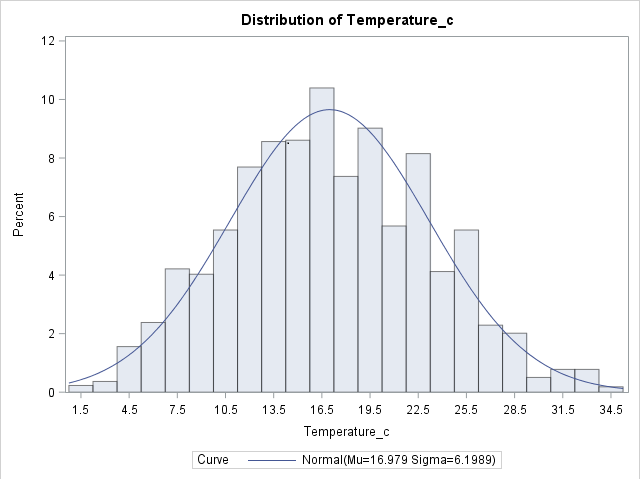


Fig: M1 (d) Fig: M1 (e) Fig: M1 (f)

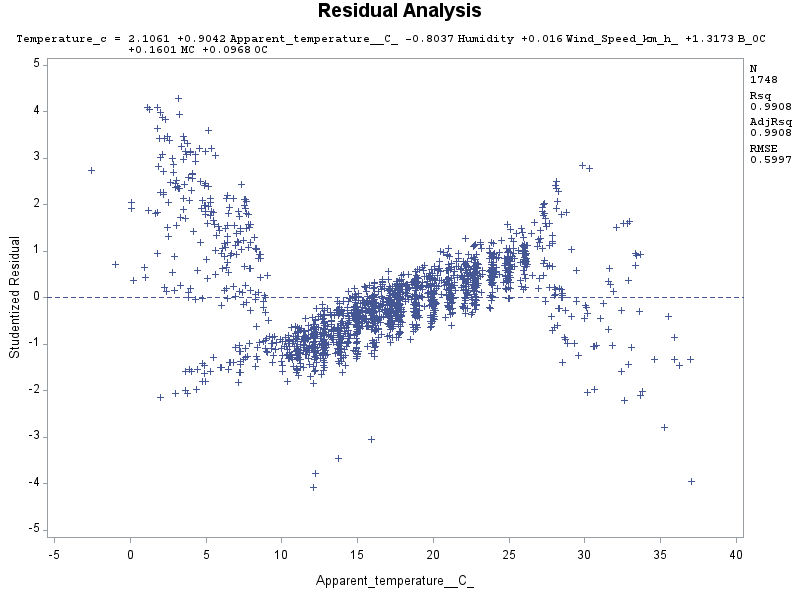
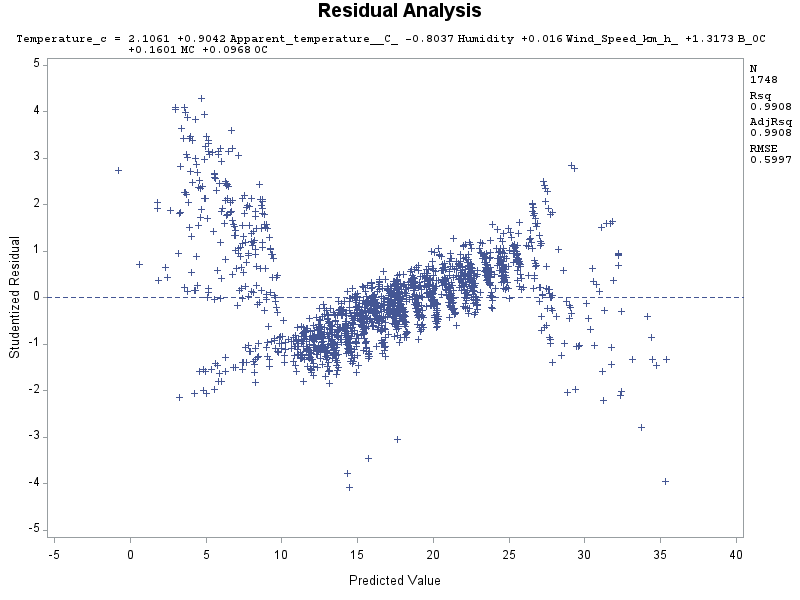


Fig: M1 (g) Fig: M1 (h) Fig: M1 (i)

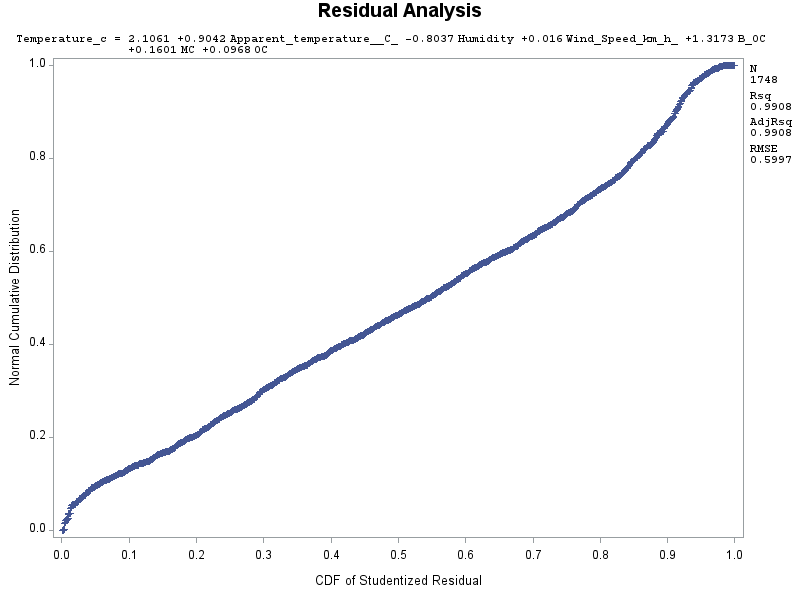
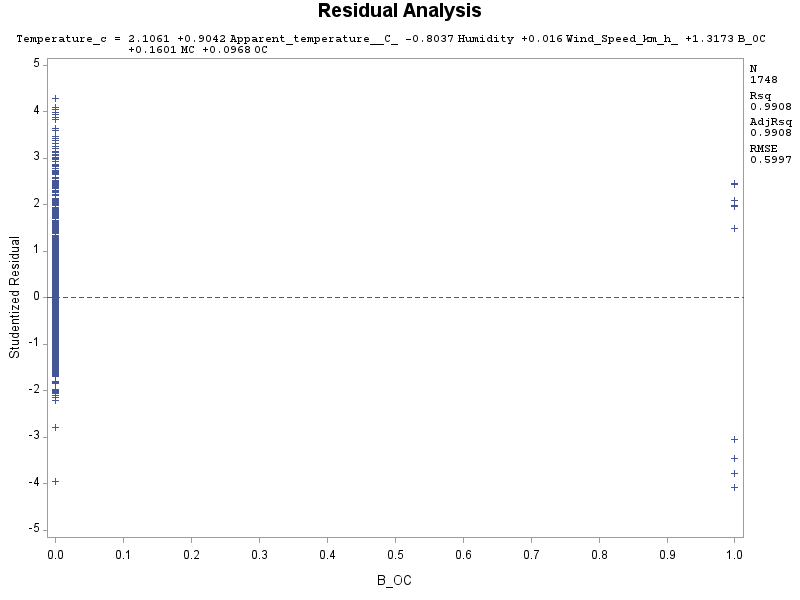
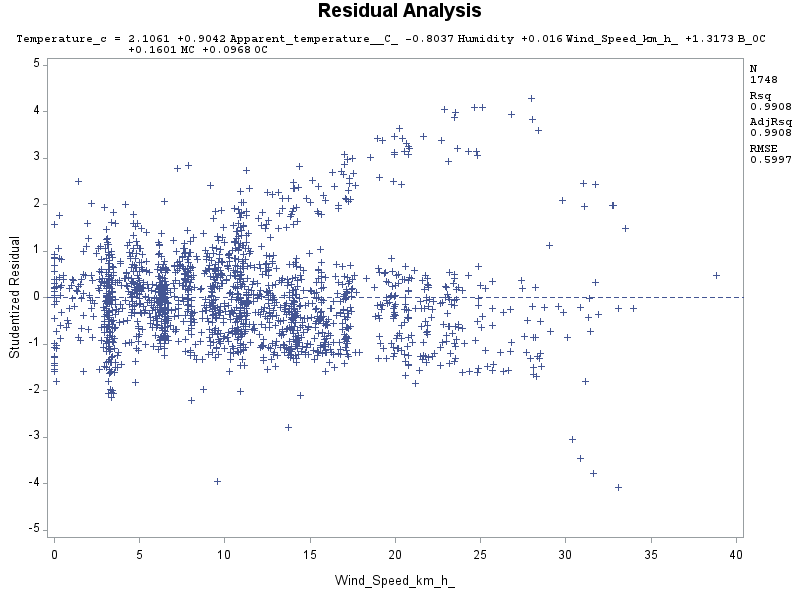


Fig: M1 (j) Fig: M1 (k) Fig: M1 (l)

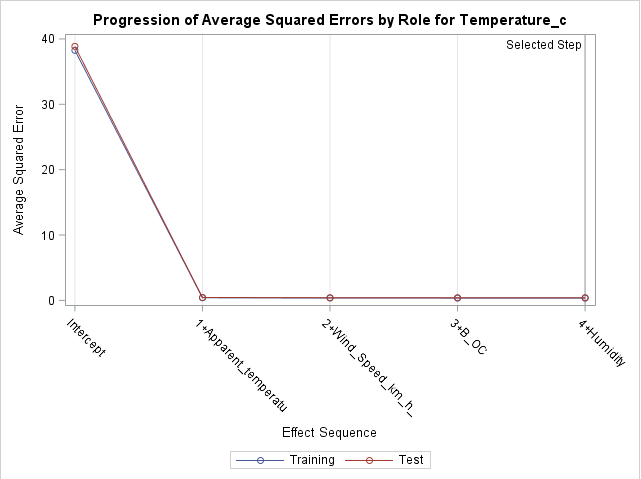
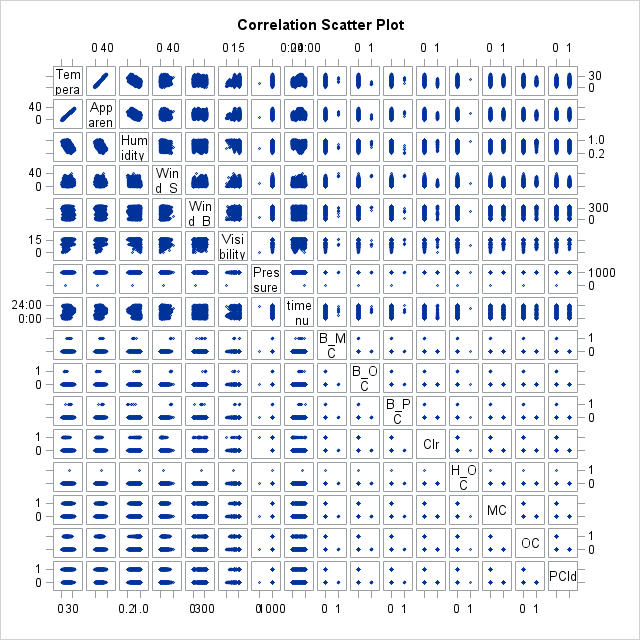


Fig: M1 (m) Fig: M1 (n)

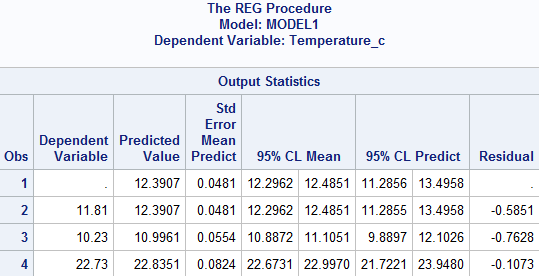
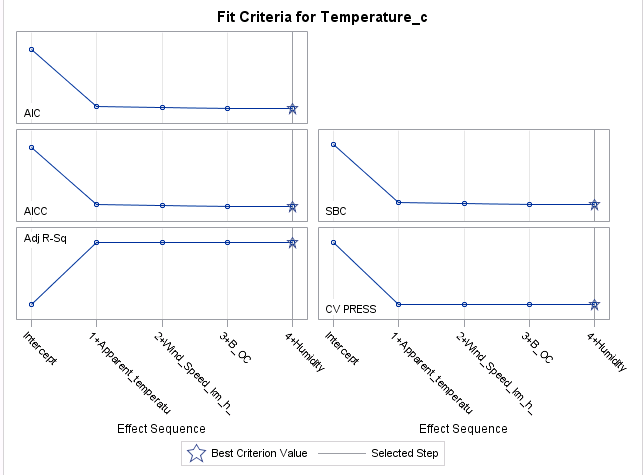


Fig: M1 (o) Fig: M1 (p)

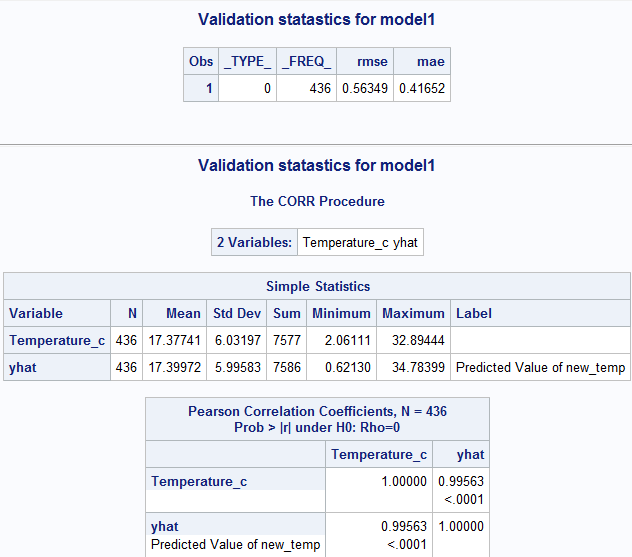
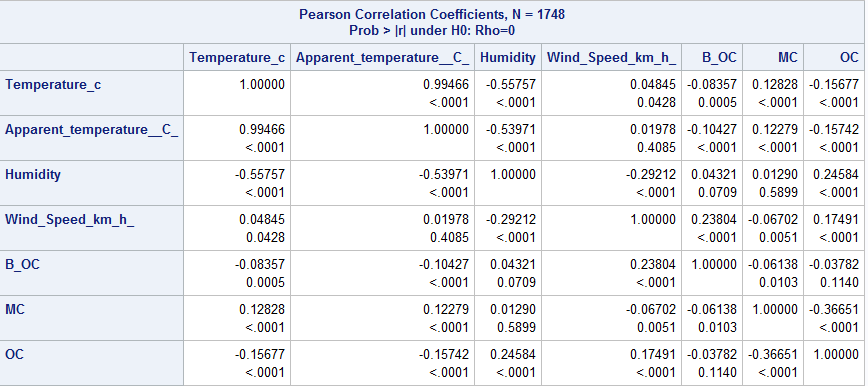


Fig: M1 (q) Fig: M1 (r)

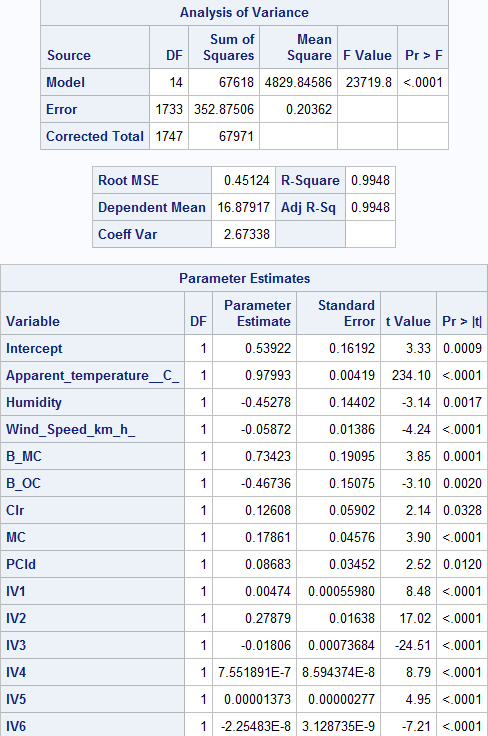
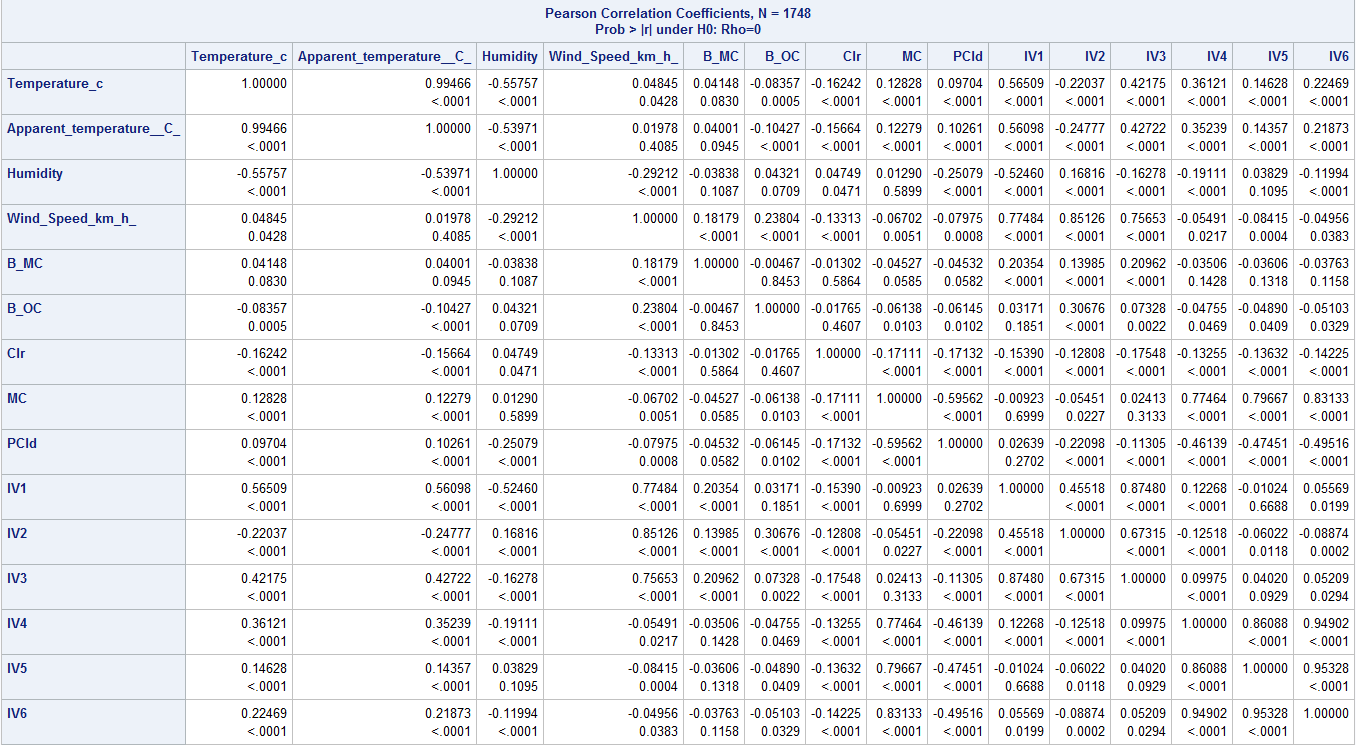


Fig: M2 (a) Fig: M2 (b)

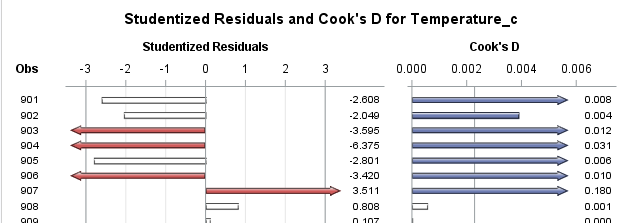


Fig: M2 (c) Fig: M2 (d)

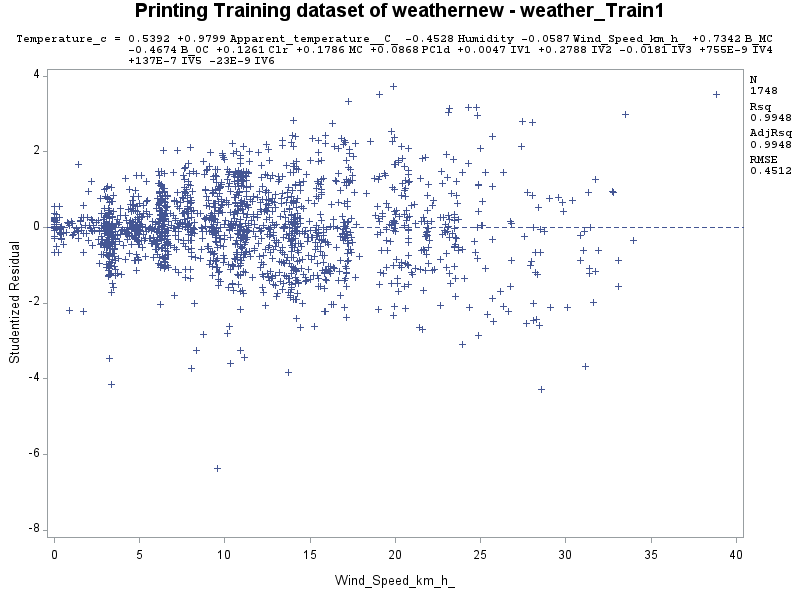
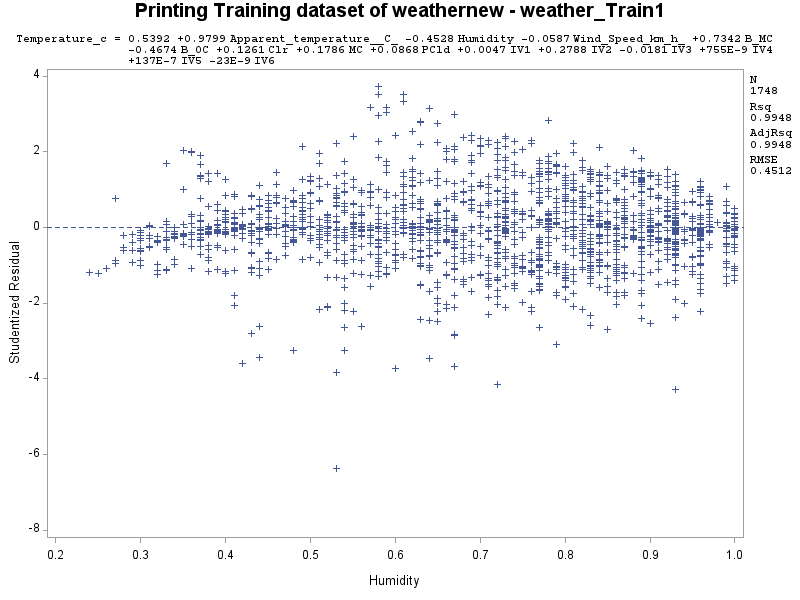
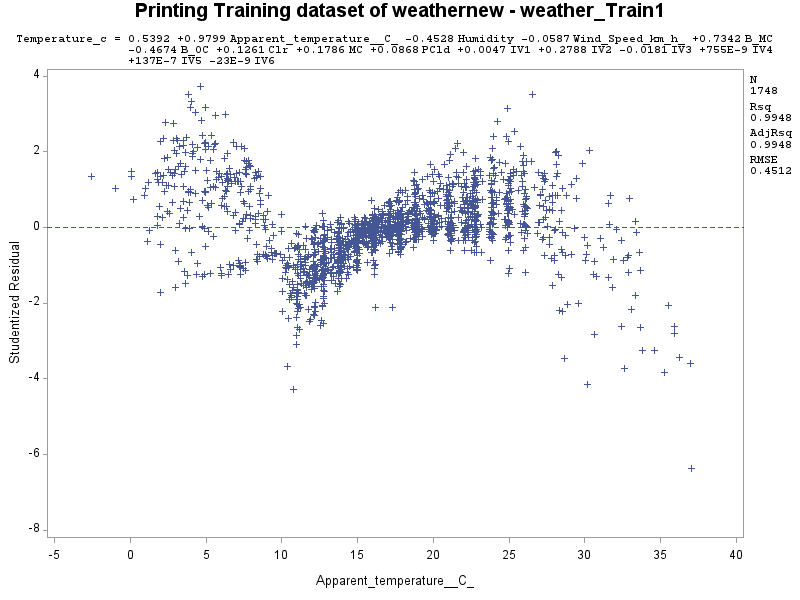


Fig: M2 (e) Fig: M2 (f) Fig: M2 (g)

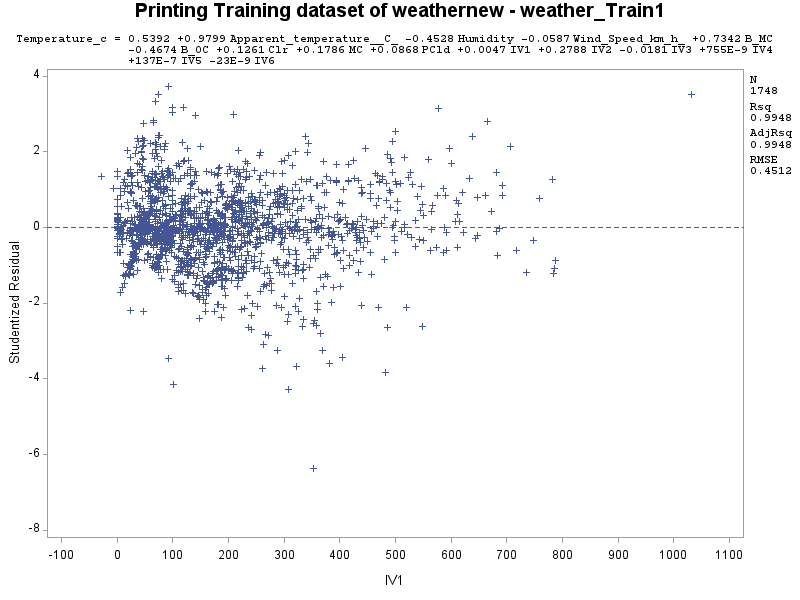
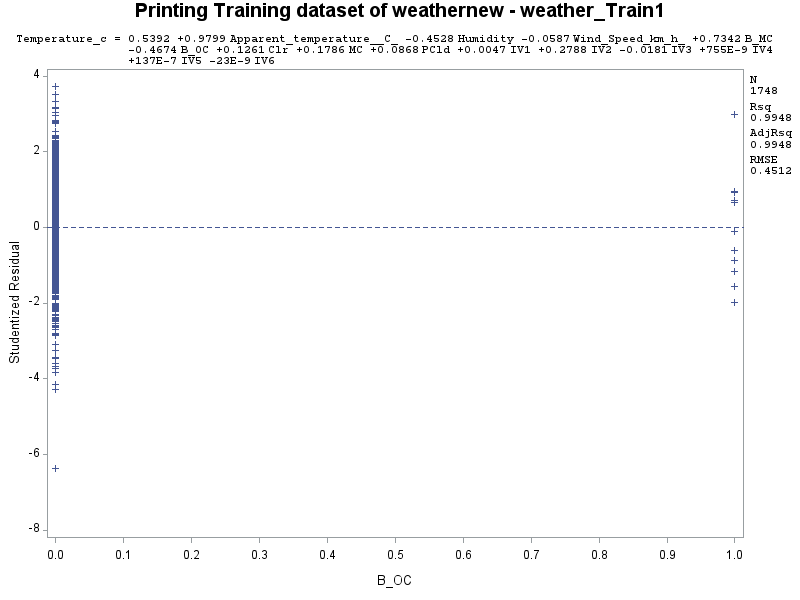
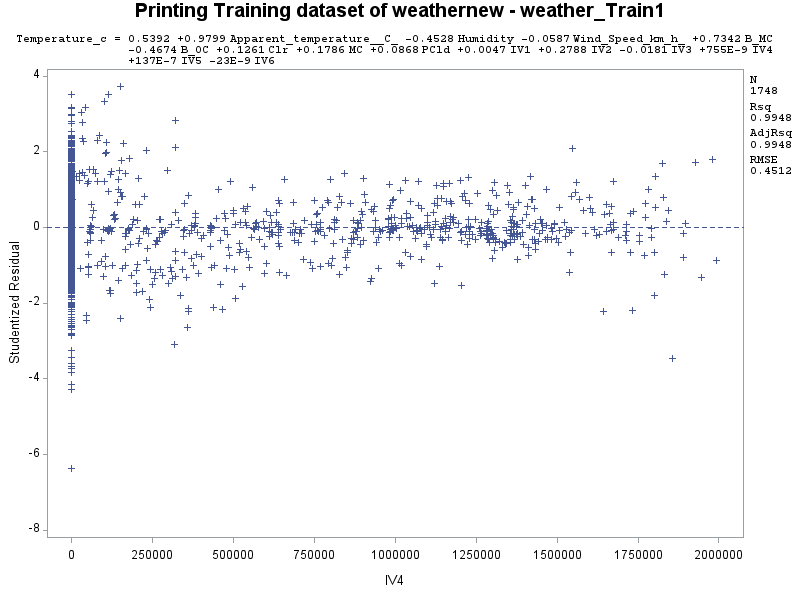
 

Fig: M2 (h) Fig: M2 (i) Fig: M2 (j)

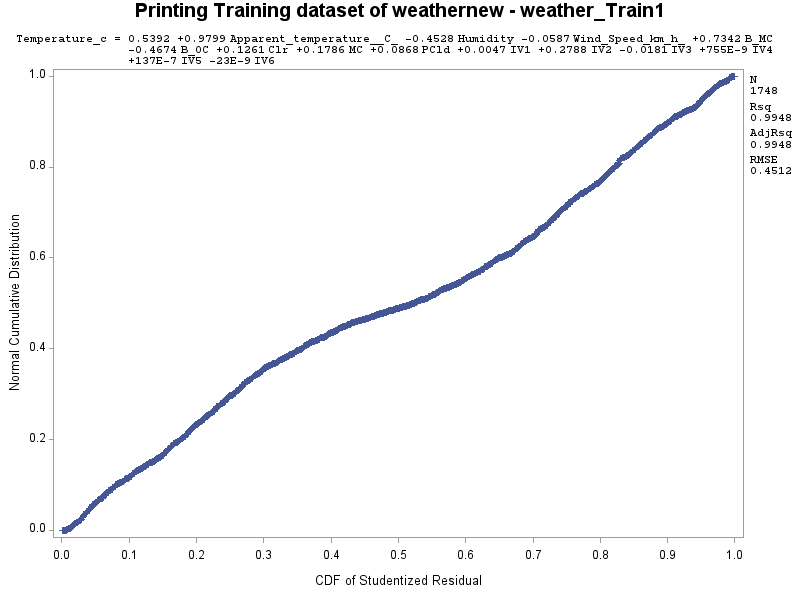
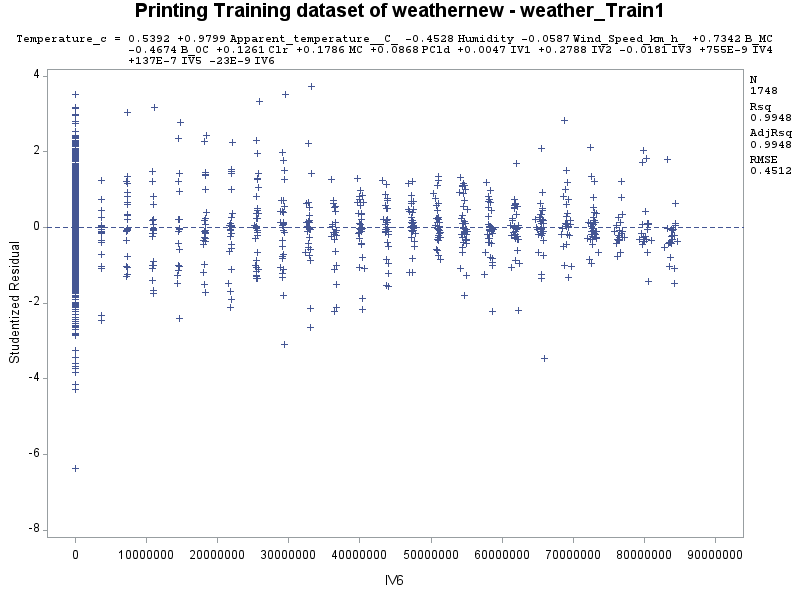


Fig: M2 (k) Fig: M2 (l)

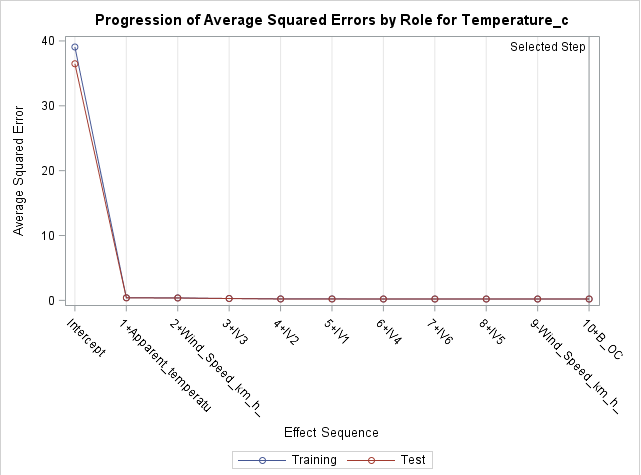
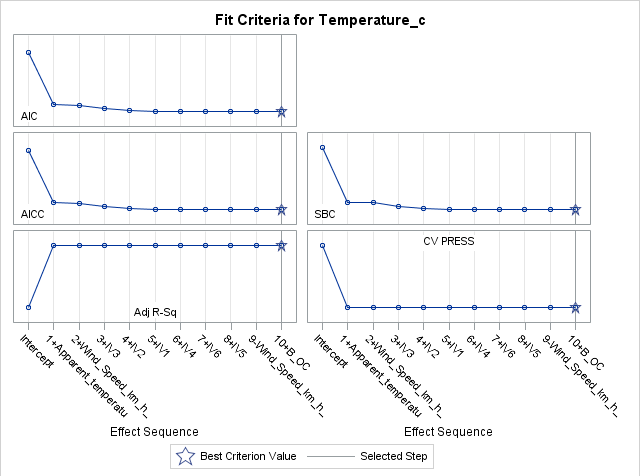


Fig: M2 (m) Fig: M2 (n) Fig: M2 (o)

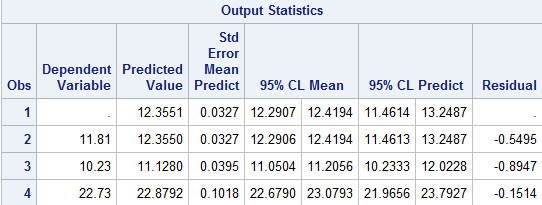
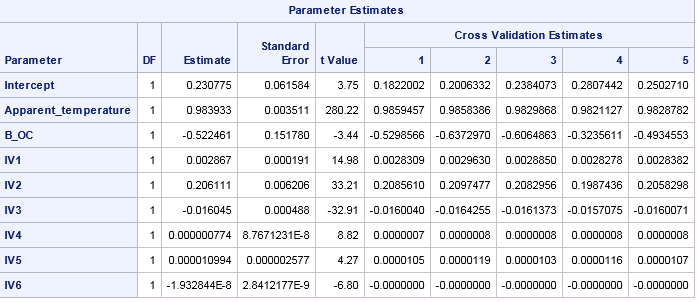


Fig: M2 (p) Fig: M2 (q)

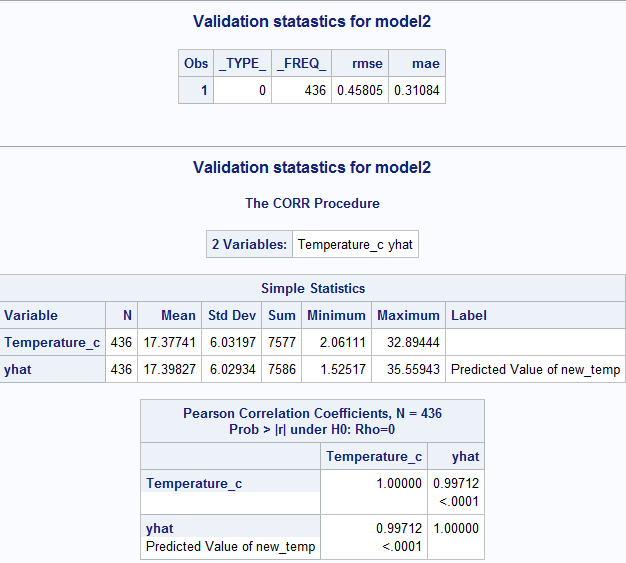


Fig: M2 (r)

**Methodology III:**

1. **Data Cleaning:**

In the original dataset there were 96,000 observations which was very difficult to use for project. Hence, we have decided to break the dataset and took dataset only for the year of 2016 which has 8785 observations. After that again we have split the dataset into four parts among ourselves and I took the dataset for the month of July, August and September.

We found that there were few null values for variable ‘Precipitation Type’ when we were analysis the dataset. But for my part (Dataset for July, August & September) of the Dataset there were no missing values for Precipitation Type. Hence, Data cleaning was not a big task for me.

I have created dummy variables for two variables ‘Precipitation Type’ and ‘Summary’ as below.

* DPrecip = ( Precip\_type = "snow")
* DBreezy = (Summary = "Breezy")
* DB\_M\_Cloudy = (Summary = "Breezy and Mostly Cloudy")
* DB\_Cast = (Summary = "Breezy and Overcast ")
* DB\_P\_Cloudy = (Summary = "Breezy and Partly Cloudy")
* DClr = (Summary = "Clear")
* DDrizzle = (Summary = "Drizzle")
* DFroggy = (Summary = " Froggy")
* DHum\_Cast = (Summary = "Humid and Overcast")
* DL\_Rain = (Summary = "Light Rain")
* DM\_Cloudy = (Summary = "Mostly Cloudy")
* DCast = (Summary = "Overcast")
* DP\_Cloudy = (Summary = "Partly Cloudy")
* DRain = (Summary = "Rain")

I took ‘Mostly Cloudy’ Summary as base case to create dummy variables. There are 13 dummy variables for ‘Summary’ variables and 1 dummy variable for ‘Precipitation Type’ variable. I have used only four dummy variables DClr, DM\_Cloudy, DCast, DP\_Cloudy because only these are part of my dataset.

1. **Dataset Splitting:**

In this step, we have split the dataset into training and testing set. Total number of observations for my part of the dataset are 2209. Dataset has been split into training and testing set by 80% and 20% respectively. Number of observations for training set are 1768 and for testing set its 441.

We have used training dataset to compare two different models and come up with the best model which can explain the variance of our y variables which is ‘Temperature’.

1. **Exploratory Data Analysis:**

In this step, I have done the Histogram, Boxplot and Residual analysis. Histogram analysis will tell us whether the data is normally distributed or not. I have created two boxplots, between time and temperature and between months and temperature. These Boxplots will tell us about temperature variations during different months and time. I have computed the Correlation values and standardized estimates as well to find the strongest x-variable which we can use for Temperature prediction.

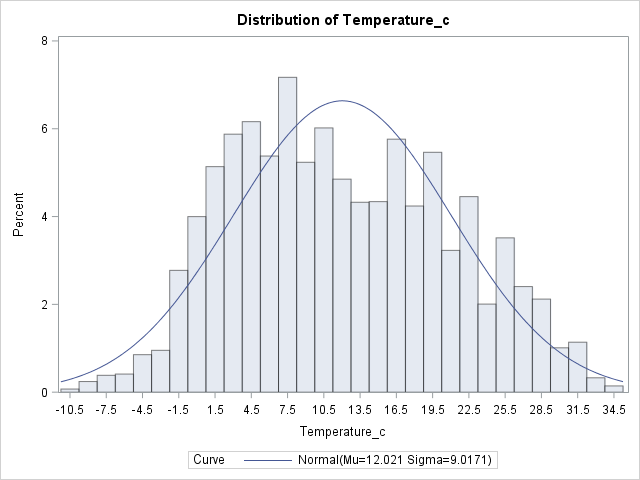
***Histogram:***

Mean = 20.88, Median = 20.88 i.e. Mean=Median which means Histogram is symmetric.

There are no outliers.

Kurtosis = -0.61 which is <3 which means negative kurtosis and curve is flatter with heavy tails.

We can say that temperature would be between 15-27 degrees Celsius for maximum number of time for the Month of July, August and September.

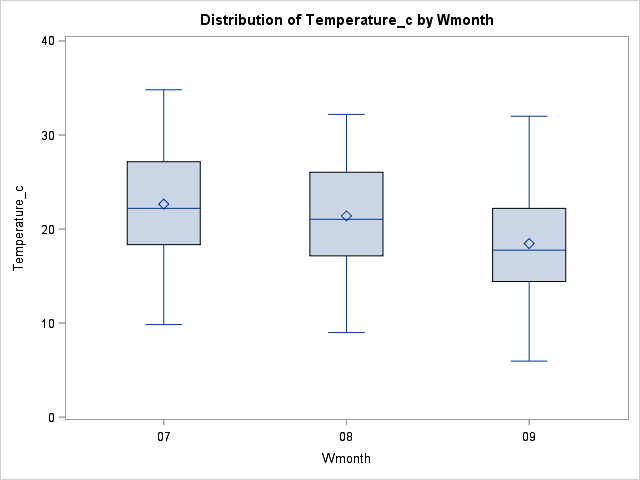


***Boxplots:***



As per above boxplot between Temperature and Months we can say that minimum temperature would start decreasing for the month of august and September.

And maximum temperature would be higher for the month of July.



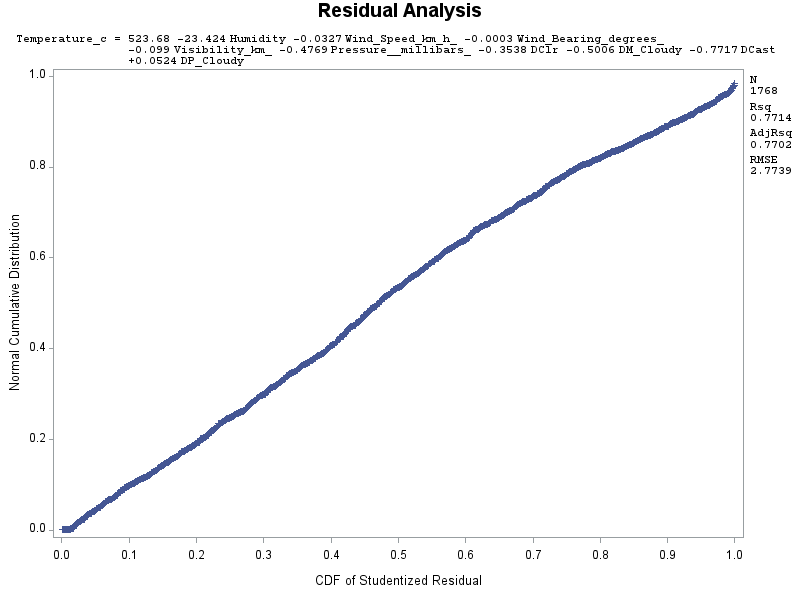
As per above boxplot between Temperature and Months we can say that minimum temperature would start decreasing for the month of august and September.

And maximum temperature would be higher for the month of July.

***Residual Analysis:***

I have performed residual analysis to detect any possible problems in the regression.

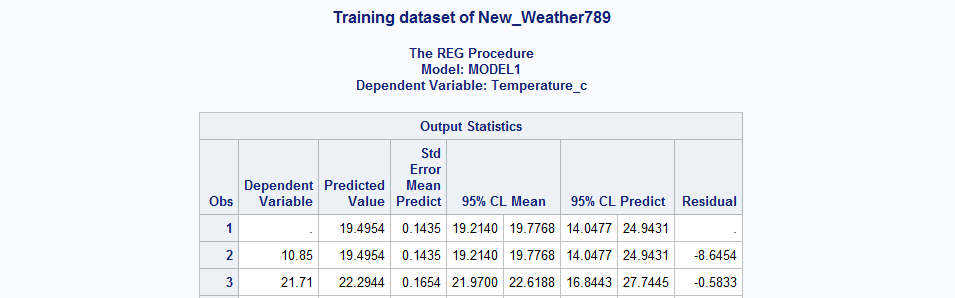
Normal Probability plot is showing almost 45 degree angel so we can say that it linear and normally distributed (No Need of transformation).



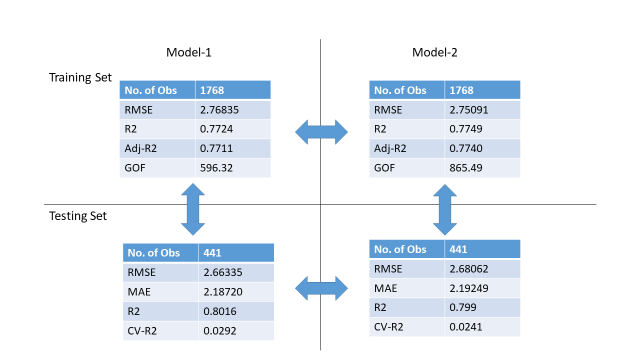
***Outliers and Influential Points:*** There were only one outlier and 3-4 influential points in my model. I kept them as it is because number of observations for my dataset was not much.

***Strongest Predictor:* Humidity** is the strongest predictor for my dataset.

***Prediction Interval and Predicted values:***



I have done the prediction for Model 1 (training set). I have used data from my dataset only and Predicted value is equal to observed value. Hence we can say our model is predicting the values correctly.

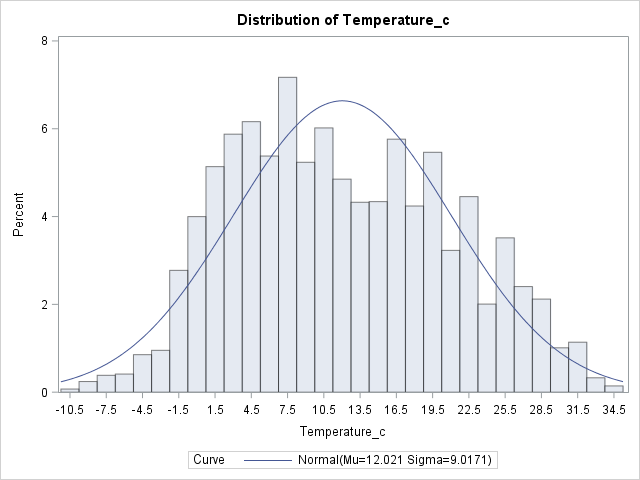


As per ***Performance Metrics***, we can say that Model 2 is better as Adj-R2 is higher in Model 2 for Training set. For Testing Set, Model 2 is better as RMSE is decreasing and for it.

* **Model 1 Training Vs Testing:** Testing set is better that Training set as RMSE is decreasing for Testing set and R2 is increasing.
* **Model 2 Training Vs Testing:** Testing set is better that Training set as RMSE is decreasing for Testing set and R2 is increasing.

**Outputs:**

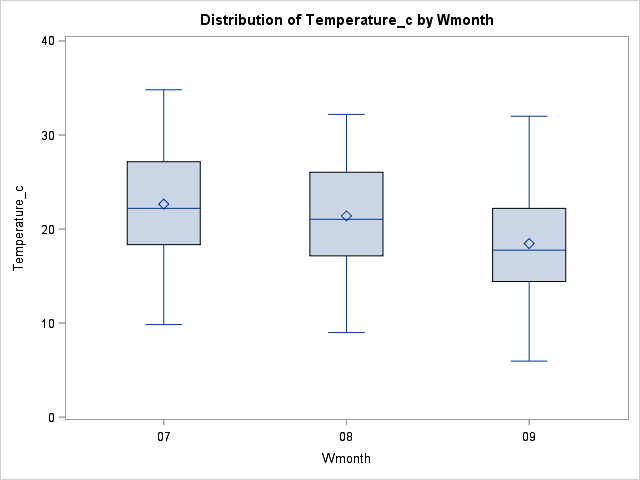
Histogram:



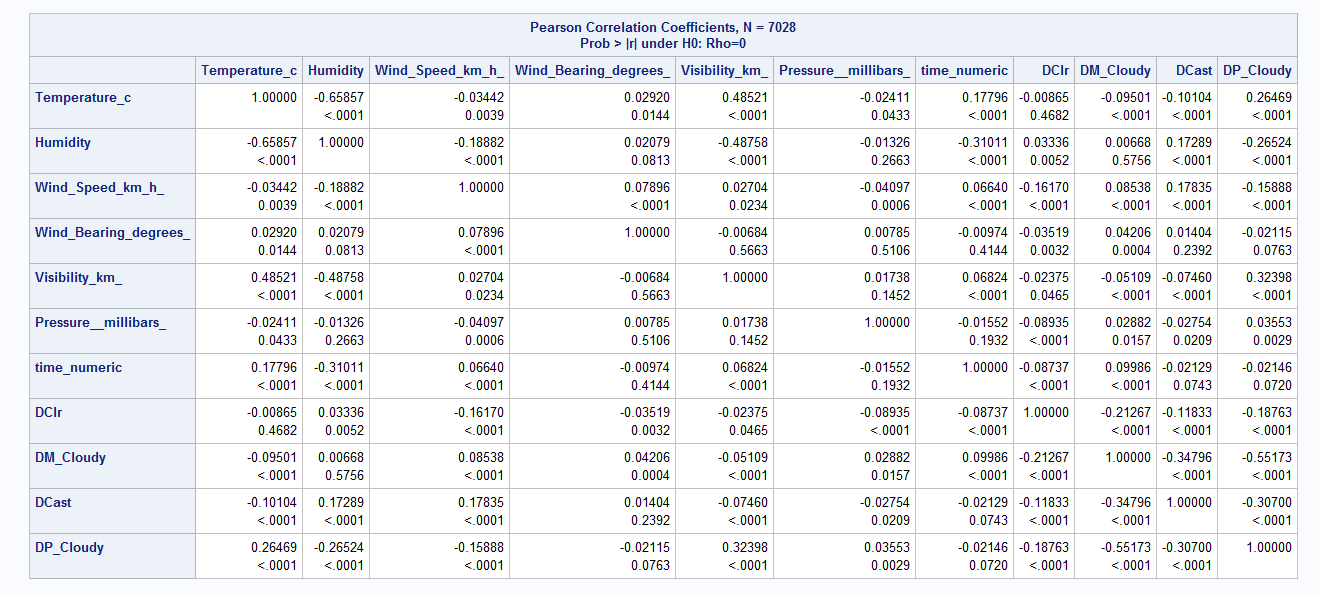


Boxplots:

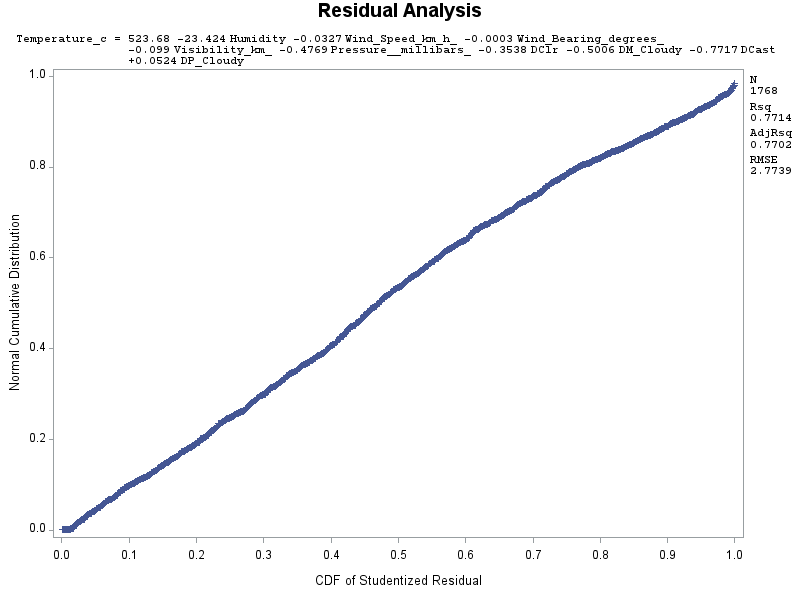




Correlation Values:



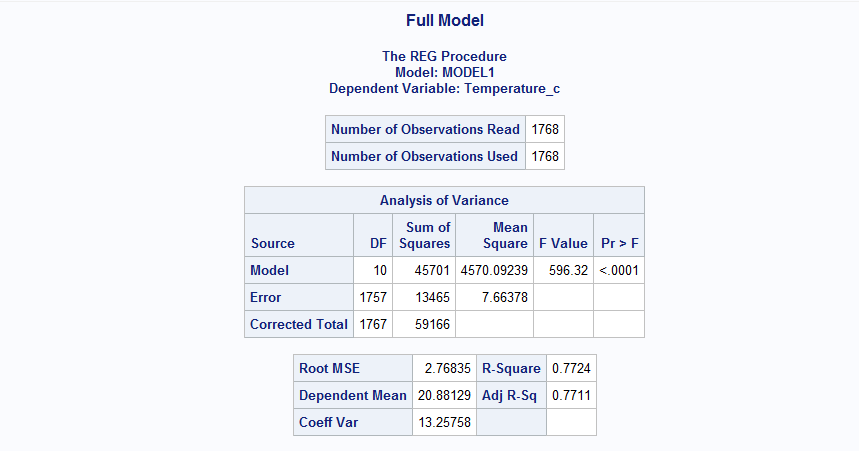
Normality Plot:

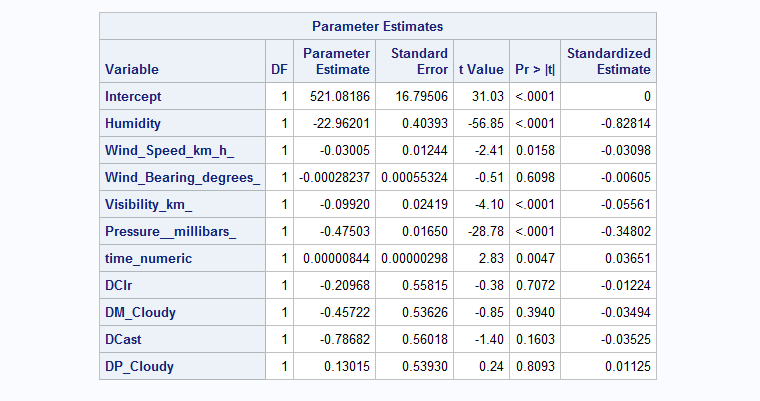


Predicted value plot:



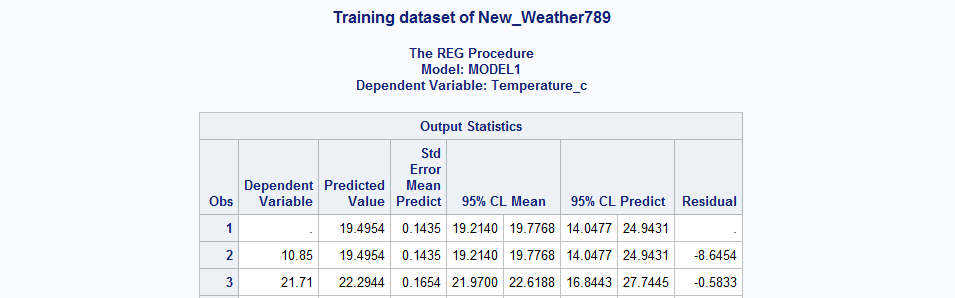
Full-Model:





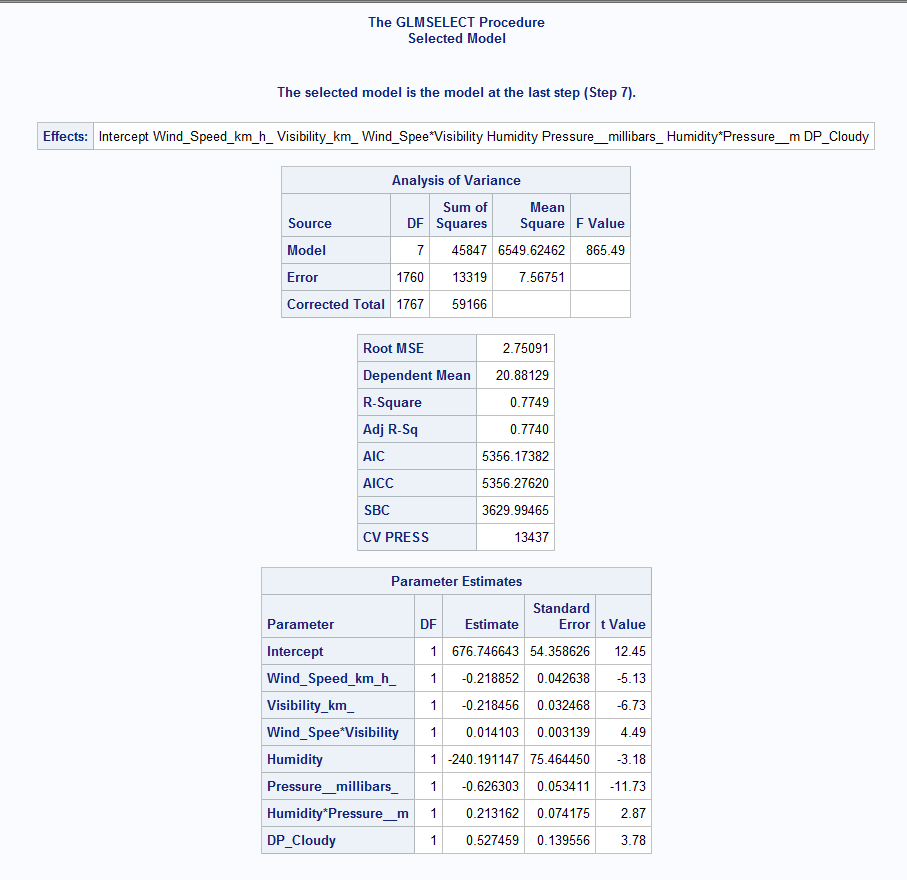
Prediction of Model 1- Training set:



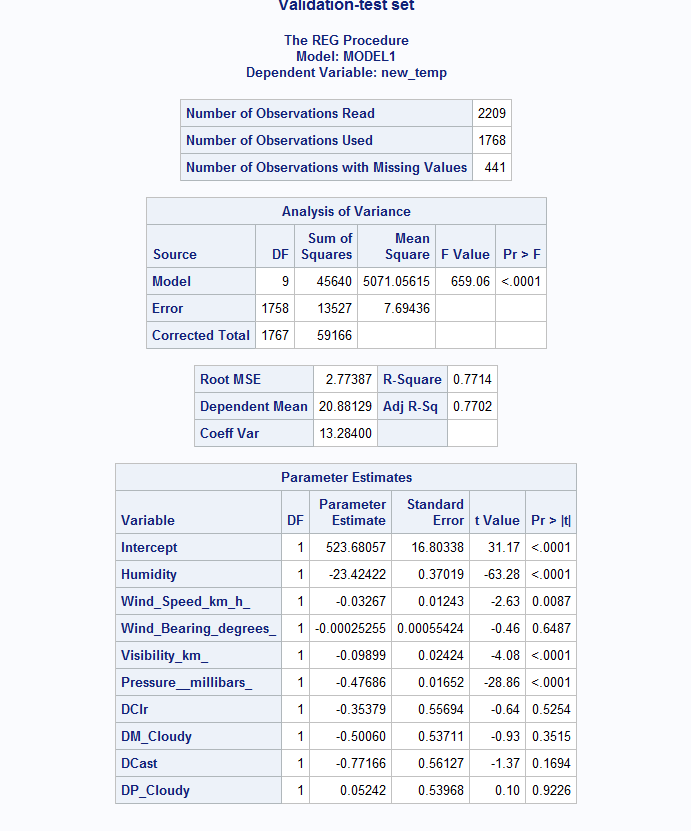


Predicted value is equal to observed value. Hence we can say our model is predicting the values correctly.

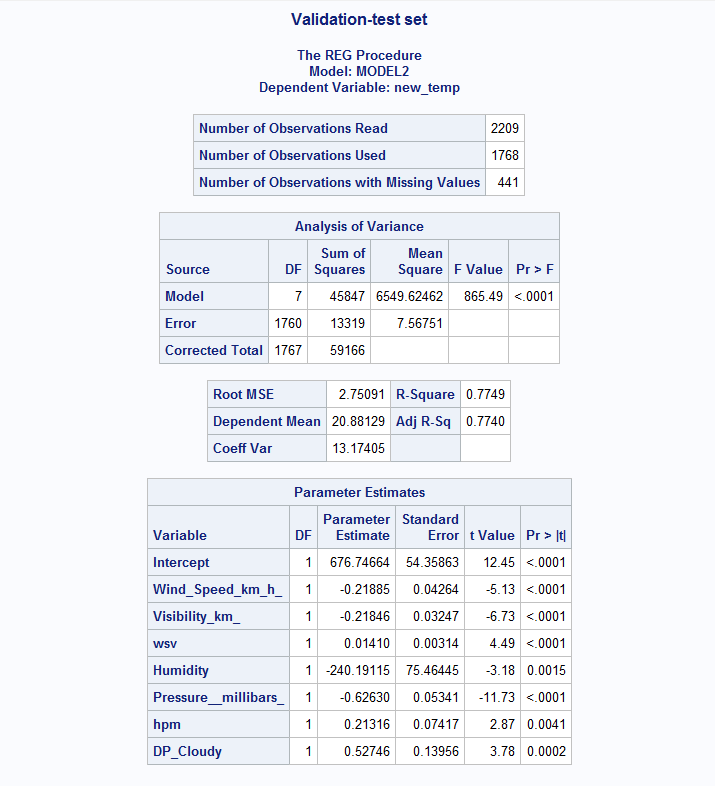
Interaction Variables (Stepwise- Model 2):



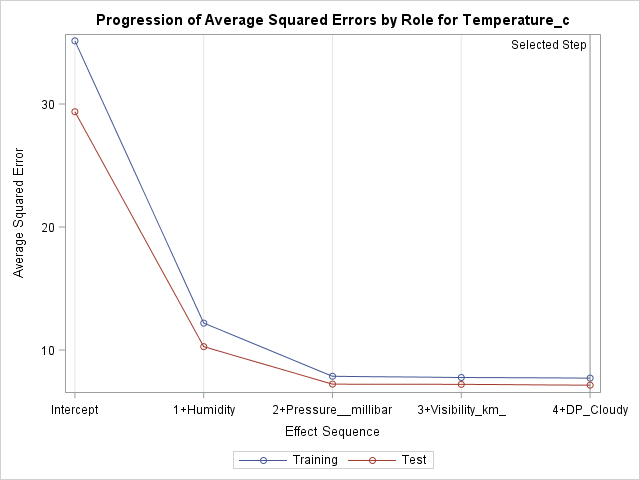
Validation test set: (Model 1)

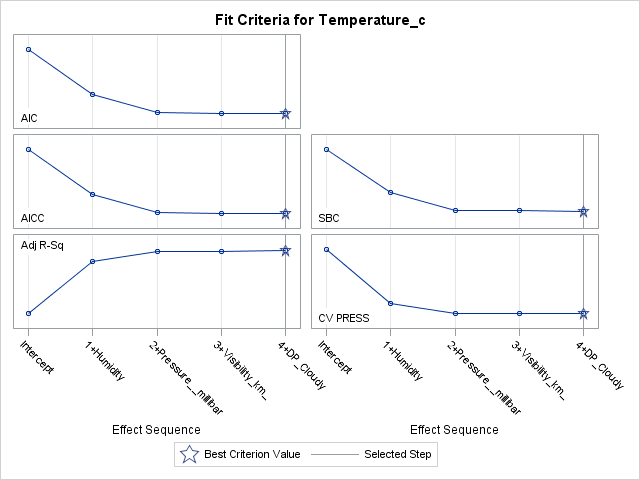


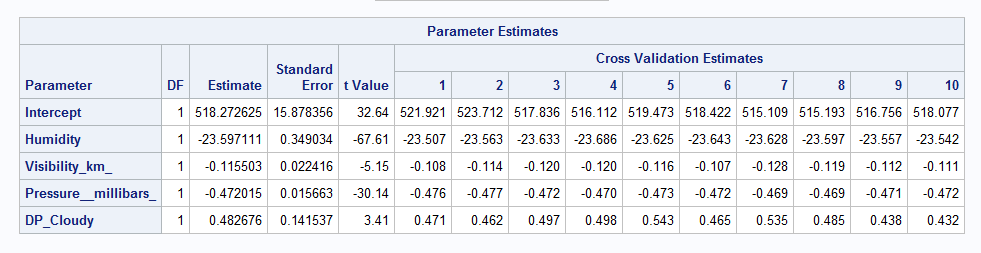
Validation test sets: (Model 2)



5-Fold: Stepwise







**Methodology IV:**

In this dataset, I have the following variables which are qualitative and hence required creating dummy variables, the qualitative variables are the following:

* **Precipitation type:** This takes two values ‘Rain’ and ‘Snow’ and hence we would be required to create a single dummy variable whose base case is ‘Rain’ and for it the dummy variable will have a value of 0. For ‘Snow’ the dummy variable will have a value of 1.
* **Summary:** This column has **14 different types** of values and would require us to create **13 variables**. The dummy variables have been named as B1 till B13. For each value of Summary, the associated dummy variable will be set to 1 and all other will be set to 0. For this Base case has been chosen as Mostly Cloudy since it had the most number of occurrences. For the Base case, all the dummy variables will be set to 0.
  + If Summary = Breezy then B1=1 and rest are 0’s.
  + If Summary = Breezy and Mostly Cloudy then B2 = 1.
  + If Summary = Breezy and Overcast then B3 = 1.
  + If Summary = Breezy and Partly Cloudy then B4 = 1.
  + If Summary = Clear then B5 = 1.
  + If Summary = Drizzle then B6 = 1.
  + If Summary = Foggy then B7 = 1.
  + If Summary = Humid and Overcast then B8 = 1.
  + If Summary = Humid and Mostly Cloudy then B9 = 1.
  + If Summary = Light Rain then B10 = 1.
  + If Summary = Overcast then B11 = 1.
  + If Summary = Partly Cloudy then B12 = 1.
  + If Summary = Rain then B13 = 1.
* Time\_numeric is being treated as qualitative variable with each hour as its own group. Though binning might be required while making changes in the future.

1. **Data Cleaning:**

In my part of the dataset there were missing values present for Precipitation Type. After doing a bit of research about the weather at Szeged, Hungary I have concluded that snow is usually present during the month of December and hence the missing values during the **month of December have been replaced by 1 for the PTN dummy variable**. For the **month of October and November I have replaced the missing values for the PTN dummy variable with 0** since there is higher probability of Rain during this period.

1. **Dataset Splitting:**

The dataset has been split into two parts Training set which contains 1767 records and which constitutes up to 80% of the data. The Test set contains 20% of the data which is about 442 records.

The training set will be used to come up with the best models which explain the maximum variance for the Temperature variable and will be used in testing later.

1. **Exploratory Data Analysis:**

**Figure 1,** The histogram of the Temperature variable shows skewed distribution but I decided not to go ahead with the transformation initially. After running the basic model and then looking at the normality plot I would decide whether to go ahead with the transformation or not.

**Figure 2**, shows the quantitative variables histograms and as we can see fee of them skewed but in Linear Regression no assumption is made for the independent variables and hence no transformation for these variables is required.

**Figure 3**, Various Boxplots gives us the following information:

* Temperature and Time: The **best time to go out and visit places during this time would be 11:00 AM and 7:PM during October to December**
* Pressure and Time: Pressure is usually **constant throughout the day**.
* Temperature vs Month: The average **temperature keeps decreasing form October through December**.
* Visibility vs Time: **Visibility during this duration is low(5km) during the morning time and is prettty much clear the rest of the day**.
* Wind Speed vs Time: Overall there is a **gentle breeze with speeds reaching a maximum of 14 km/hr** but it isn’t as windy as Chicago.

**Figure 4**, shows the **Pearson’s Correlation Values** and we can see that the Apparent Temperature is highly correlated with temperature and it alone can explain **99.29%** variance in the Temperature. I have excluded this variable from my analysis in order to come up with a model that explains nearly as much of the variance in the temperature as possible and also to see the effect of the transformations and interaction terms.

After examining of all the other values, I can conclude that there isn’t any Multicollinearity present in this dataset because no Independent variables were highly correlated with other independent variables.

**Figure 5**, shows the scatterplot matrix of the variables I have excluded the binary variables in this plot to see the strength of the correlation between dependent and independent variables. Based on Pearson and Scatterplot we can see that the Humidity is strongly but negatively correlated with temperature. The other scatterplots show that there exists correlation between temperature and other independent variables but it isn’t as strong as humidity.

**Figure 6**, Frequency of distribution of the weather conditions over this duration of the analysis shows that few of the dummy variables are absent which means those conditions weren’t present so frequently during this duration.

1. **Model Building:**

The Formatted\_Date column in this dataset contained the Date and Time together and hence we had to run a small code to extract the time and month. The time column is named as **time\_numeric** and month column as **Wmonth.**

**Figure 7,** After running the basic model and checking the normality plot I saw there were curves at the top and bottom of the normality plot and hence I decided to apply transformation and rerun the model but the adjusted R-Squared decreased by nearly 10%. Hence, I scrapped the idea of transformation and decided to go ahead by adding interaction terms.

In the GLMSELECT I have used order 3 since the normality plot was showing two curves and hence the interaction terms were as follows:

* Humidity Time and Pressure had interaction between them and hence these were | separated in the procedure.
* Wind Speed, wind Direction and Visibility was the other interaction term.

**Model – 1: This model will remove all the insignificant variables and wont contain any interaction or higher order terms and neither any kind of transformation is applied to the Temperature\_C.**

Figure 8, The Parameter estimates and the output for this model have been shown and as we can see that this model can explain 84.09% of variance in Temperature. In this model, I’ve used the following options: for the PROC REG procedure in SAS:

* **Selection**: In this I used Stepwise method for selection of variables into the model since it combines both the forward and backward selection methods.
* **Sls**: Criteria for staying in the model.
* **Sle**: Criteria for entering in the model.

The Residual analysis shows the presence of outliers and that the model cannot deal with outliers properly and this might be the reason for the curves at the top and bottom of the normality plot.

The final Equation for this model is:

**Temperature\_C = 241.01 – 34.401\*Humidity + 1.00\*B5 + 5.2611\*B6 +3.279\*B10 +2.77\*B11 – 3.15\*PTN - 0.1034\*WIND\_SPEED\_KM\_H + 0.0056\*WIND\_BEARING\_DEGREES\_ - 407E-8\*TIME\_NUMERIC – 0.2017\*PRESSURE\_\_MILLIBARS\_**

**Model – 2: GLMSELECT model with 3rd order Interaction Terms.**

In this model, since by adding interaction terms multi collinearity will increase and hence I decided to center the variables but that lead to a huge drop in Adjusted R-Squared value and hence didn’t use the centered variables in this model.

**Figure 9**, Shows the parameter estimates of the interaction terms and the other variables which were returned after the final step. We can see that by using higher order terms we are able to increase the amount of variance which can be explained. For the stopping condition I used Cross Validation CV.

The Residual analysis and the normality plot shows that most of the outliers from the previous model were handled but still there were few. I haven’t removed the outliers because then I would’ve been left with lesser observations for learning. The residuals in which time was included as part of interaction terms had some pattern which I feel can be handled by creating bins for time variable which is also a part of the future scope for this dataset. The normlaity plot still has slight curves and in future this can be handled by binning time variable.

**Performance Matix** is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| TRAINING |  | MODEL – 1 | MODEL – 2 |
| RMSE | 1.7344 | 1.7619 |
| R-SQUARED | 0.8418 | 0.8366 |
| ADJUSTED R-SQ | 0.8409 | 0.8356 |
| GOF | 934.34 | 898.77 |
|  | | | |
| TESTING | R-SQUARED | 0.233 | 0.20498 |
| CV-RSQUARED | 0.60764 | 0.6317 |
| MAE | 3.36606 | 3.61193 |
| RMSE | 3.9332 | 4.25995 |

The Equation for the second model including the interaction terms is as follows:

**Temperature\_C = 152.86 + 5.288\*B6 -3.2816\*PTN + 603E-7\*WSWBV -0.0331\*HUMP - 0.1146\*PRESSURE\_\_MILLIBARS\_ + 0.0016\*WIND\_BEARING\_DEGREES\_ - 0.2445\*WIND\_SPEED\_KN\_H**

From the above table we can see that the two models aren’t performing well on the test data even after applying interactions terms and hence getting more varibles into the model might be helpful and/or by creating bins for time might make it significant and have a positive impact on the models.

**Figure 10** and **Figure 11** show the rmse mae and yhat values when Model -1 and Model -2 were ran on the test set. Using these values I had computed the above mentioned Performance Matrix.

After Adding the Apparent Temperature in the final models I got the following performance matrix for the training and test set.

|  |  |  |  |
| --- | --- | --- | --- |
| TRAINING |  | MODEL – 1 | MODEL – 2 |
| RMSE | 0.32219 | 0.3155 |
| R-SQUARED | 0.9945 | 0.9948 |
| ADJUSTED R-SQ | 0.9945 | 0.9947 |
| GOF | 26627.2 | 30289.5 |
|  | | | |
| TESTING | R-SQUARED | 0.96324 | 0.96198 |
| CV-RSQUARED | 0.03125 | 0.03281 |
| MAE | 0.63299 | 0.62876 |
| RMSE | 0.79246 | 0.78537 |

Now we can say that both the models are performing very nicely and hence it would be up to the clients to select the model which they would prefer.

**Figure 13** and **Figure 14** show the SAS output of running these models on test cases.

The Figures are as follows:

Figure 1: Histogram of Temperature without transformation.

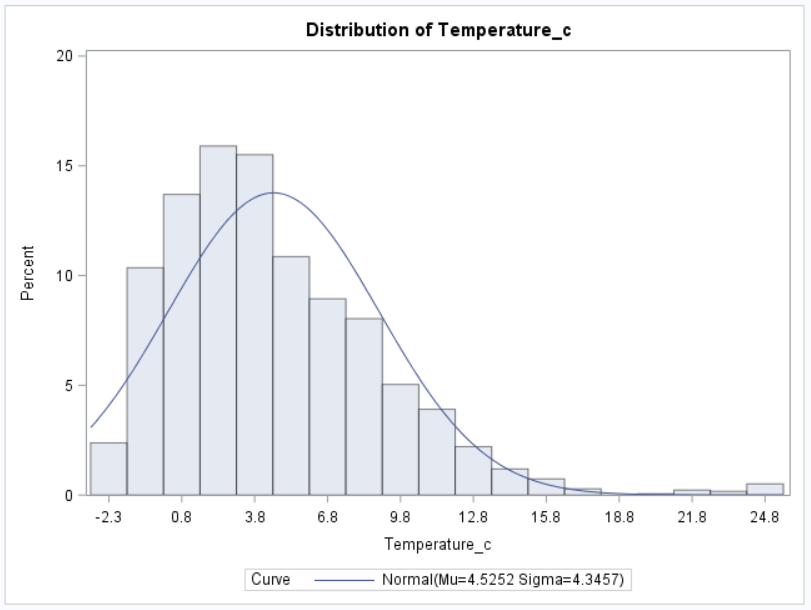


Figure 2: Histogram of independent variables

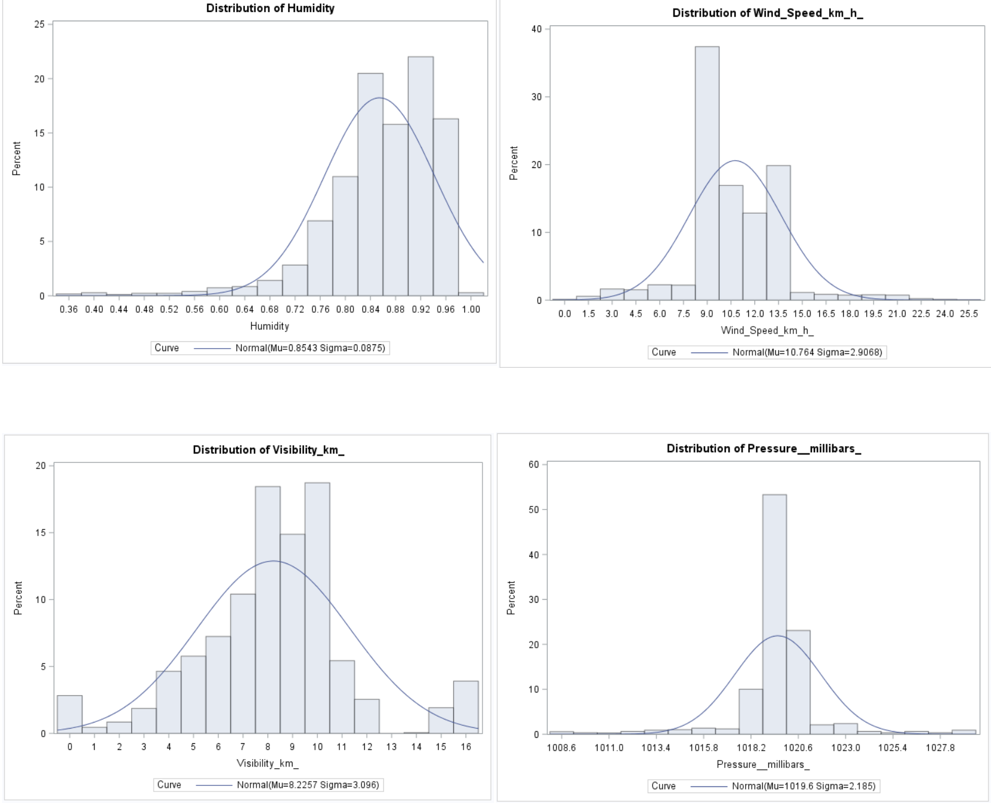


Figure 3: Boxplots

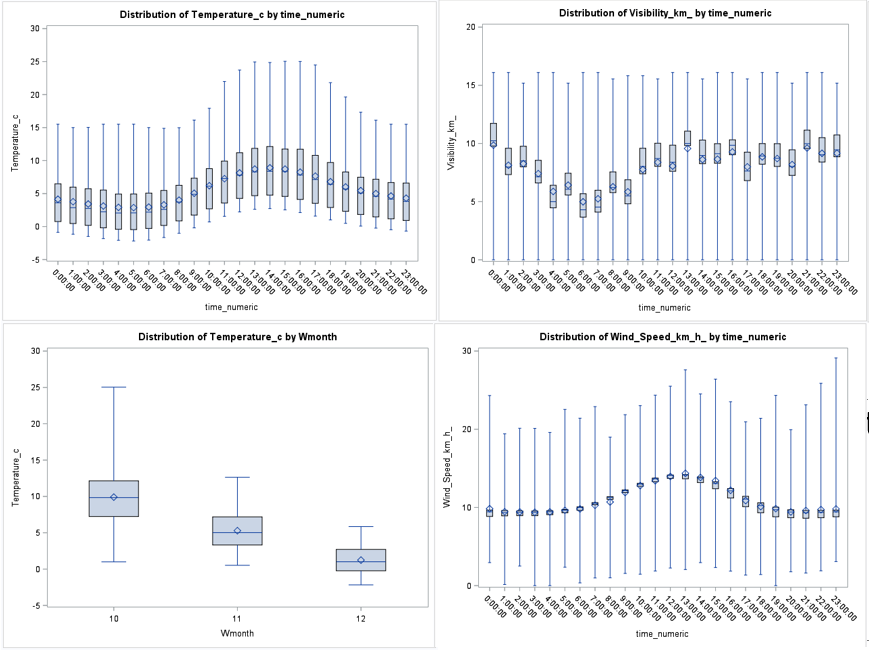


Figure 4: Pearson Correlation values:

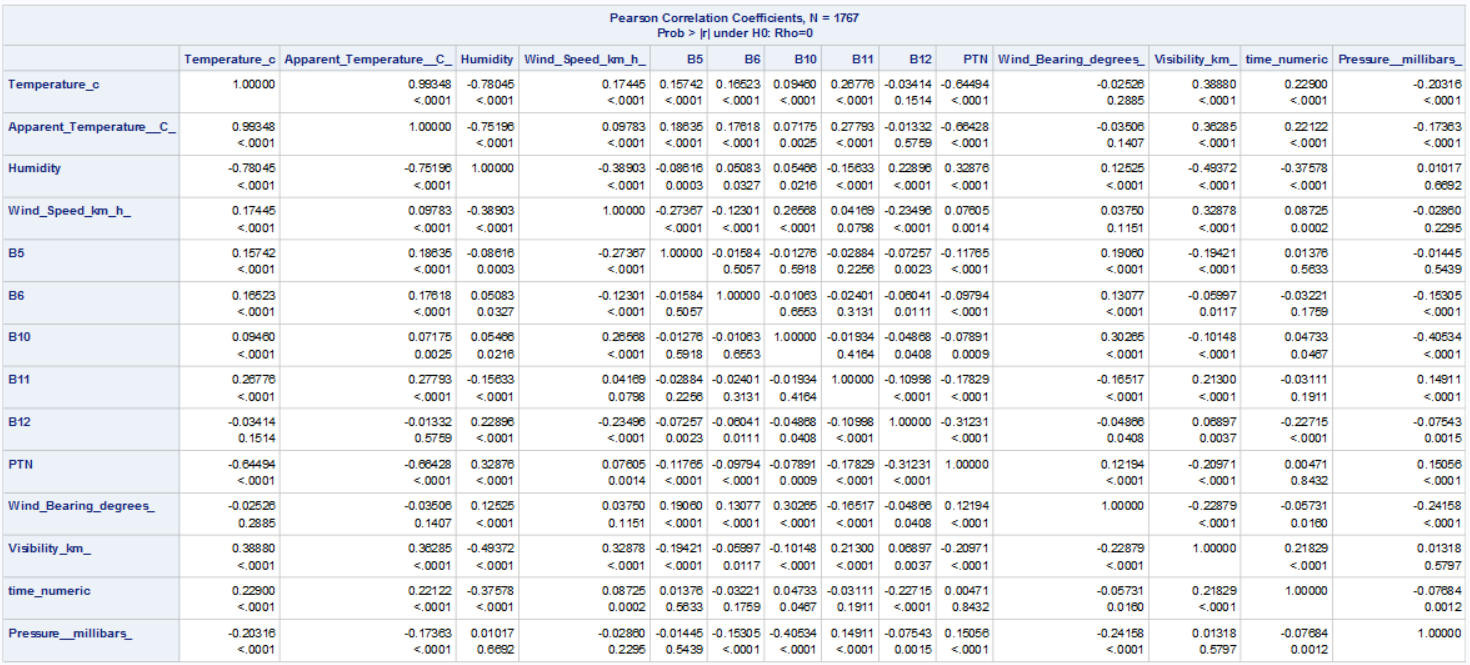


Figure 5: Scatterplots

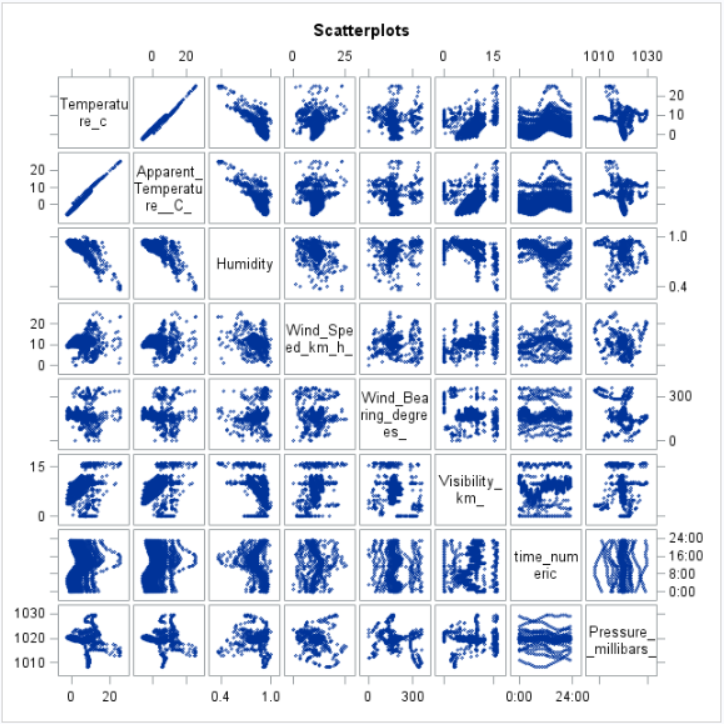


Figure 6: Frequency of Summary

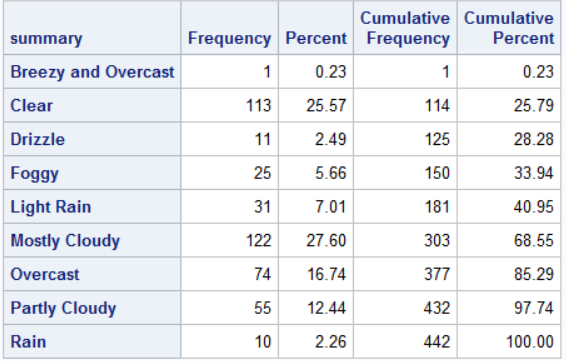


Figure 7: Normality plot after applying ¼ root transformation on Temperature

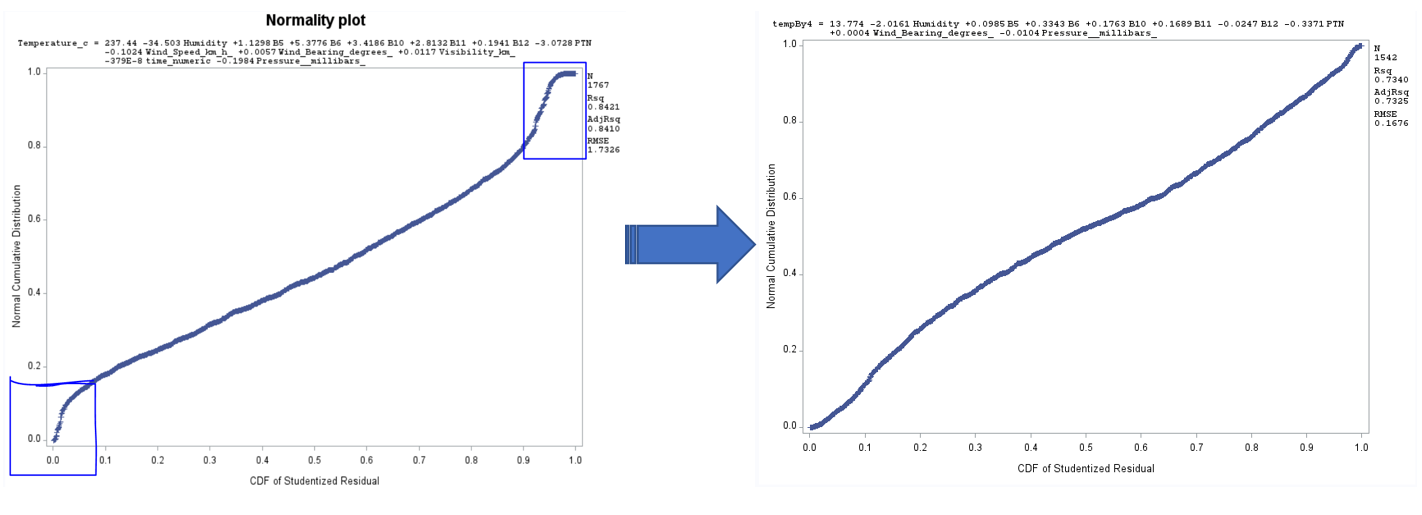
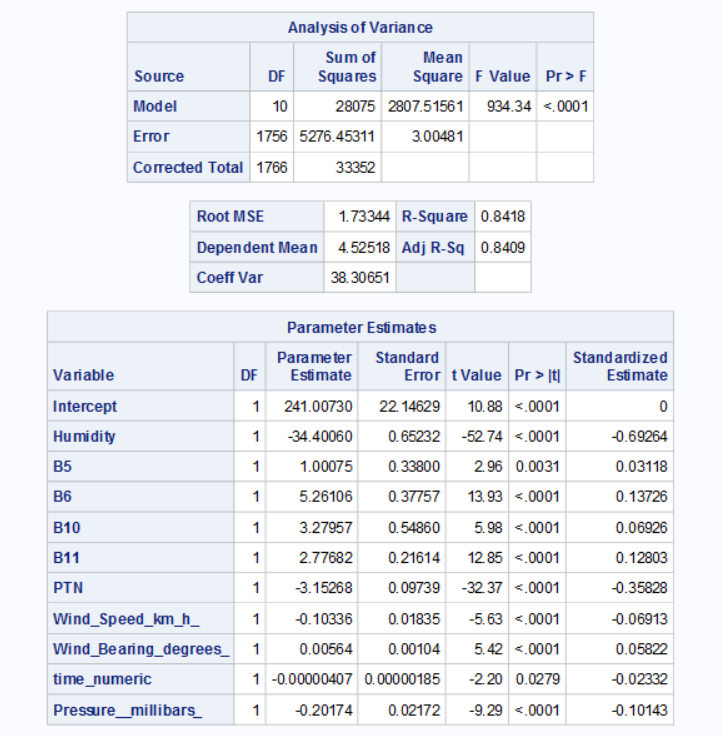


Figure – 8 Parameter estimates, Normality plot and output for Model – 1.



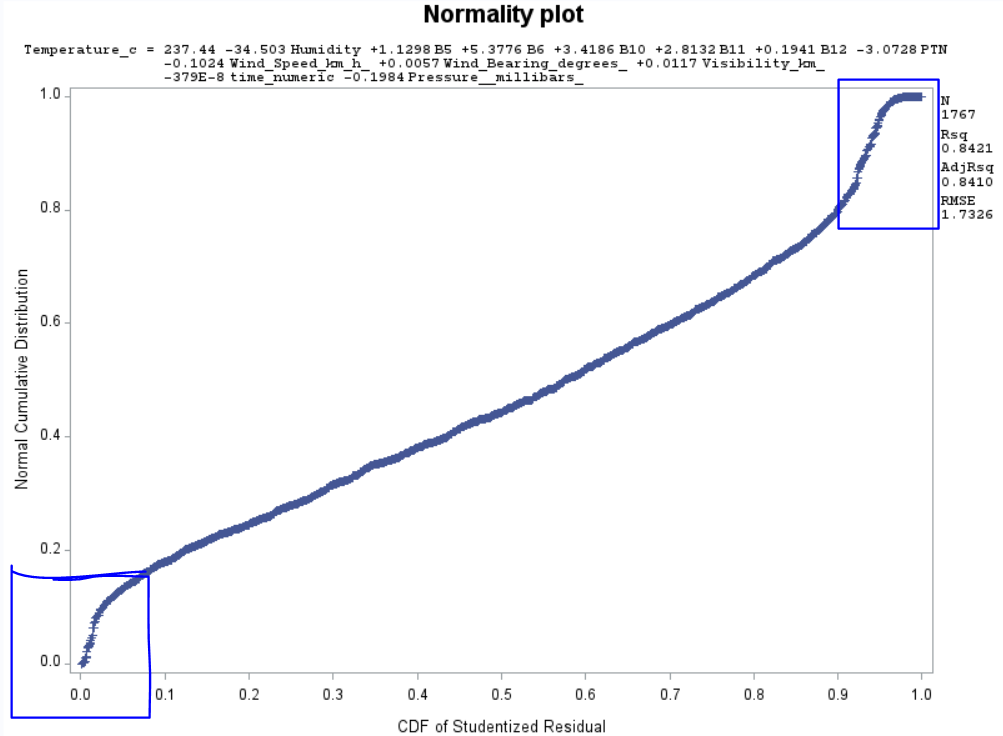
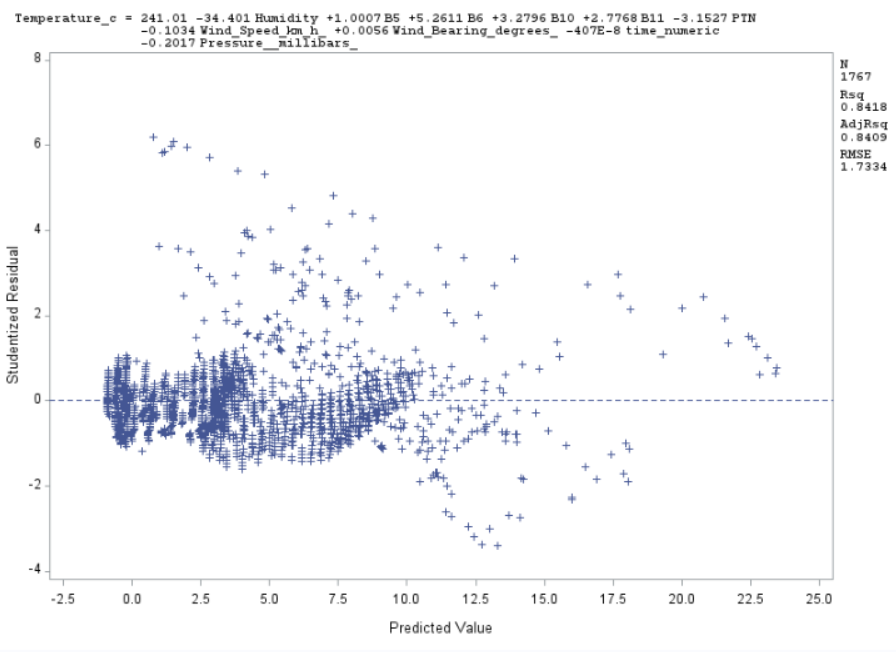
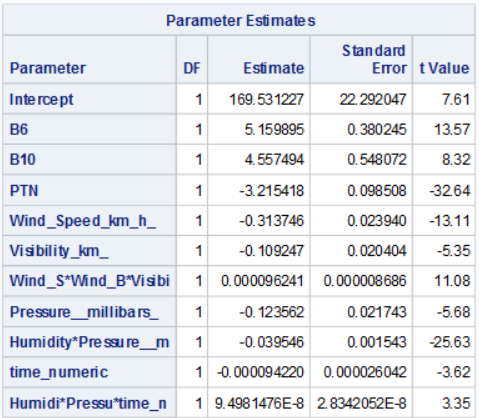
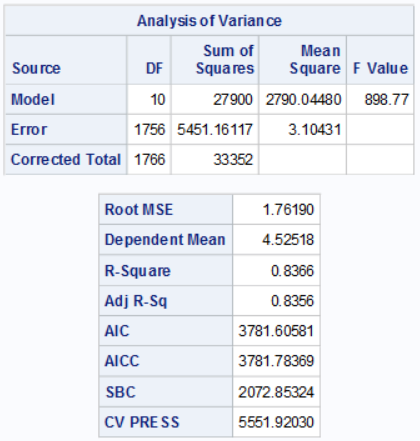
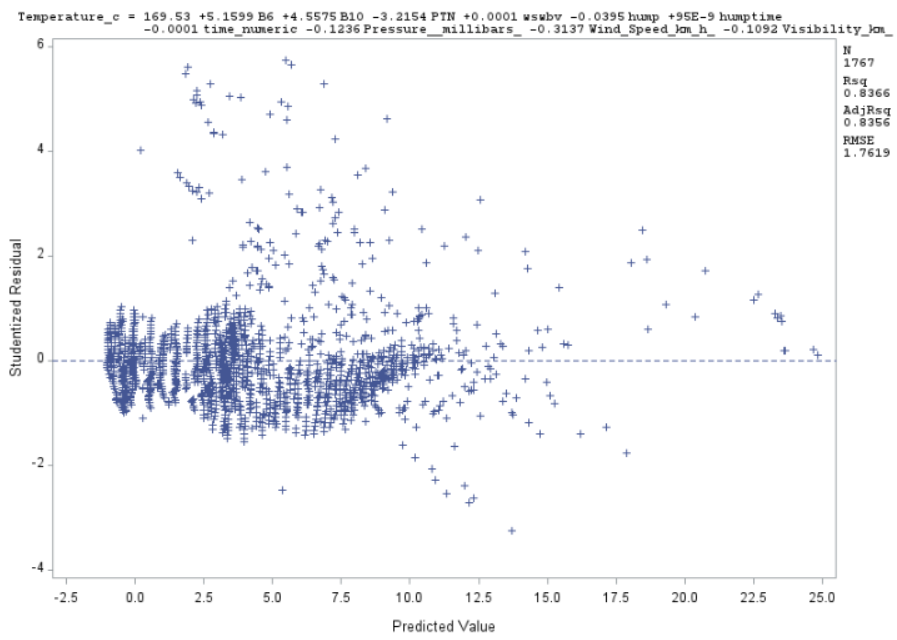
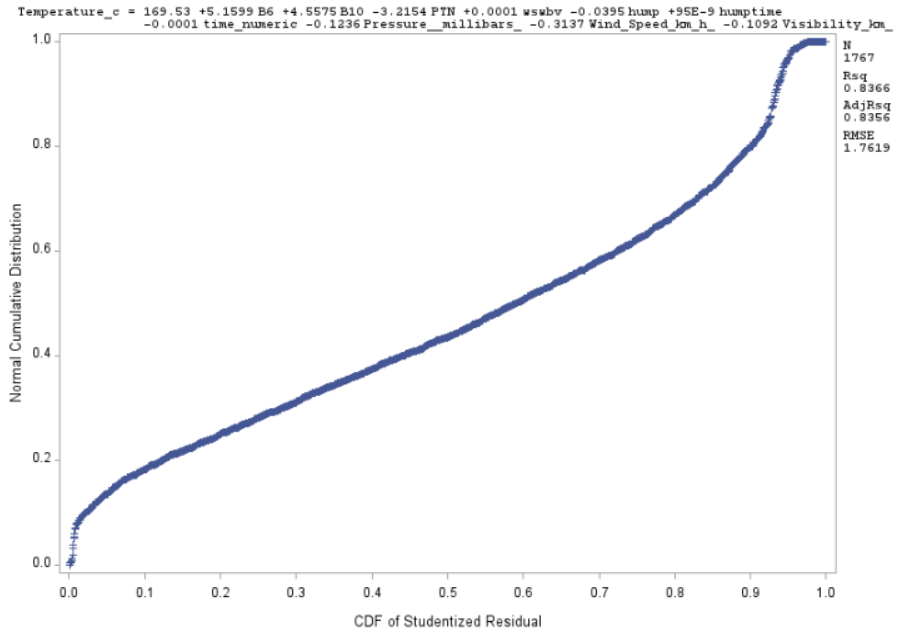
 

Figure – 9: Parameter Estimates for GLMSELECT model which includes Interaction terms Normality and Predicted values plots.





SFigure – 11: Running model-1 on test set

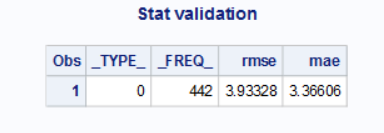
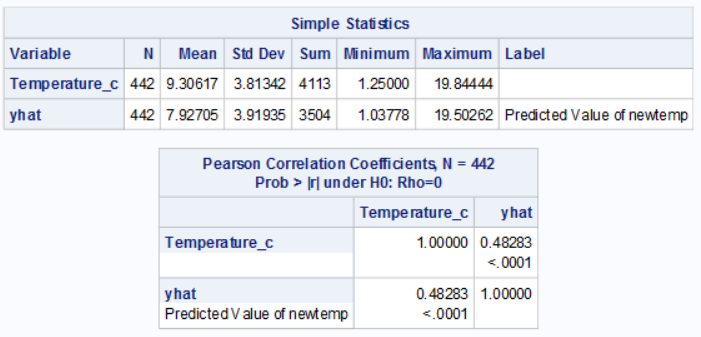
 

Figure – 12: Running model-2 on test set.

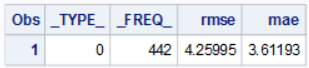
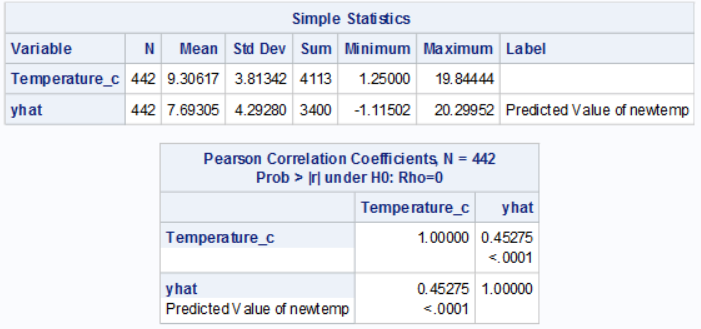
 

Figure – 13: Running model – 1 on test set after including Apparent Temperature.

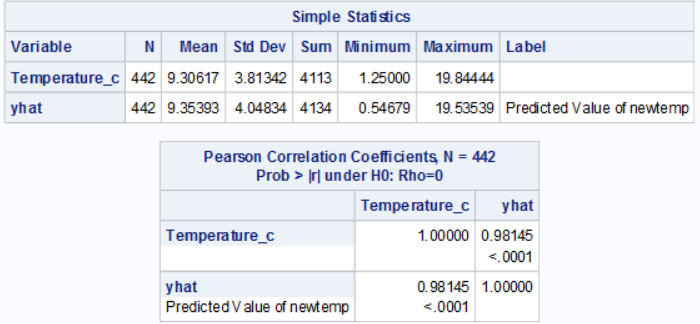
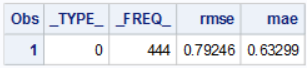
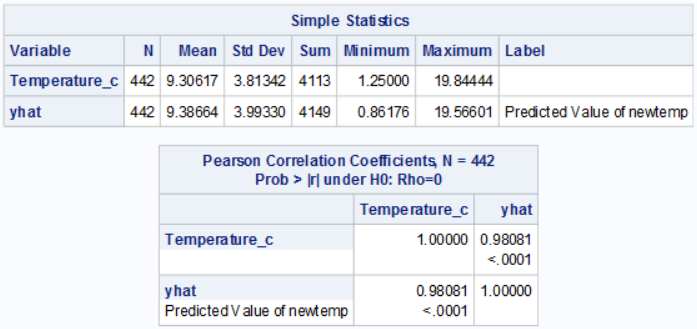
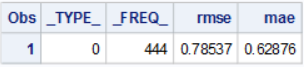


Figure – 14: Running model – 2 on test set after including Apparent Temperature without making any changes to other variables.



**Methodology I SAS Code:**

\*Creation of dummy variabels;

\*Jaymin parekh CSC 423 Data analysis and regression;

**DATA** weather;

length Formatted\_Date $30.;

length summary $30.;

length daily\_Summary $40.;

INFILE "weather.csv" DELIMITER =',' MISSOVER FIRSTOBS=**2**;

INPUT Formatted\_Date Summary $ Precip\_type $ Temperature\_c Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Loud\_Cover Pressure\_\_millibars\_ Daily\_Summary $;

Wmonth=substr(Formatted\_Date,**6**,**2**);

format time\_numeric time8.;

time\_numeric = input(substr(Formatted\_Date,**12**,**2**) !! ':' !! substr(Formatted\_Date,**15**,**2**) !! ':' !! substr(Formatted\_Date,**18**,**2**),time8.);

Snow = ( Precip\_type = "snow");

Breeze = ( Summary = "Breezy");

B\_MC = ( Summary = "Breezy and Mostly Cloudy");

B\_OC = ( Summary = "Breezy and Overcast ");

B\_PC = ( Summary = "Breezy and Partly Cloudy");

Clr = ( Summary = "Clear");

Drizzle = (Summary = "Drizzle");

Fog = ( Summary = " Foggy");

HMC = (Summary = " Humid and Mostly Cloudy");

H\_OC = (Summary = "Humid and Overcast");

Lgt\_rain = ( Summary = "Light Rain");

MC = (Summary = "Mostly Cloudy");

OC = (Summary = "Overcast");

PCld = (Summary = "Partly Cloudy");

RN = (Summary = "Rain");

**run**;

**PROC** **PRINT**;

**RUN**;

TITLE "Select only the required months [month-JAN, FEB, MAR] ";

**proc** **sql**;

delete from weather

where wmonth not IN ('01','02','03');

**QUIT**;

**PROC** **PRINT**;

**RUN**;

TITLE'Sorts the data Formatted\_Date in Ascending Order';

**PROC** **SORT** DATA= weather OUT=Sort\_Weather;

By Formatted\_Date;

**PROC** **PRINT**;

**RUN**;

TITLE'Histogram of Temperature\_c';

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

var Temperature\_c;

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

**run**;

\*Verifying full model(with dummy variables);

TITLE "Full Model";

**proc** **reg**;

model Temperature\_c = Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Loud\_Cover Pressure\_\_millibars\_ time\_numeric Snow Breeze B\_MC B\_OC B\_PC Clr Drizzle Fog HMC H\_OC Lgt\_rain MC OC PCld RN;

**run**;

**QUIT**;

TITLE'taining and testing data set';

**proc** **surveyselect**

data = weather out = Nwether\_123 seed = **231201**

samprate = **0.80** outall;

**proc** **print**;

**run**;

TITLE'Check the frequency ';

**proc** **freq** data = Nwether\_123;

**run**;

TITLE'print Nweather for month 1,2,3';

**proc** **print** data = Nwether\_123;

**run**;

TITLE'Printing Training dataset of weathernew - weather\_Train1';

**data** weather\_Train1

(where = ( selected = **1** ));

set Nwether\_123;

**run**;

**proc** **print** data = weather\_Train1;

**run**;

TITLE "Histogram for Temperature - training data set";

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

var Temperature\_c;

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

**run**;

TITLE"Boxplot of temp vs months in Jan feb and March ";

**PROC** **BOXPLOT**;

PLOT Temperature\_c\*wmonth;

**RUN**;

TITLE "Correlation values for tarining set";

**proc** **corr**;

var Temperature\_c Humidity Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Pressure\_\_millibars\_ time\_numeric Snow Breeze B\_MC B\_OC B\_PC Clr Drizzle Fog HMC H\_OC Lgt\_rain MC OC PCld RN;

**run**;

TITLE "Full Model for training set";

**proc** **reg** data = weather\_Train1;

model Temperature\_c = Humidity Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Pressure\_\_millibars\_ time\_numeric Snow Breeze B\_MC B\_OC B\_PC Clr Drizzle Fog HMC H\_OC Lgt\_rain MC OC PCld RN;

**run**;

TITLE "training set -step wise ";

\*Model 1 step wise ;

**proc** **reg** data = weather\_Train1;

model Temperature\_c = Humidity Wind\_Speed\_km\_h\_ Visibility\_km\_ Pressure\_\_millibars\_ time\_numeric Snow Breeze Clr Drizzle Fog H\_OC Lgt\_rain MC OC RN/selection = stepwise;

**run**;

TITLE " training set -backward ";

\*Model 1 backword ;

\* this method giving the same resuts as stepwise;

**proc** **reg** data = weather\_Train1;

model Temperature\_c = Humidity Wind\_Speed\_km\_h\_ Visibility\_km\_ Pressure\_\_millibars\_ time\_numeric Snow Breeze Clr Drizzle Fog H\_OC Lgt\_rain MC OC RN/selection = backward;

**run**;

TITLE ' Refit model(M1) - getting influential points and outliers ';

**proc** **reg** data = weather\_Train1;

model Temperature\_c = Humidity Wind\_Speed\_km\_h\_ Visibility\_km\_ Pressure\_\_millibars\_ time\_numeric Snow Clr MC OC / influence r ;

**run**;

\* resudual Analysis model 1;

TITLE'Studentised Resudual plots for model1';

**proc** **reg**;

model Temperature\_c = Humidity Wind\_Speed\_km\_h\_ Visibility\_km\_ Pressure\_\_millibars\_ time\_numeric Snow Clr MC OC;

plot student.\*(Humidity Wind\_Speed\_km\_h\_ Visibility\_km\_ Pressure\_\_millibars\_ time\_numeric Snow Clr MC OC) ;

**run**;

TITLE'Studentised vs predicted values plots';

**proc** **reg**;

model Temperature\_c = Humidity Wind\_Speed\_km\_h\_ Visibility\_km\_ Pressure\_\_millibars\_ time\_numeric Snow Clr MC OC;

plot student.\*predicted.;

**run**;

TITLE'Normality plot for model1';

**proc** **reg**;

model Temperature\_c = Humidity Wind\_Speed\_km\_h\_ Visibility\_km\_ Pressure\_\_millibars\_ time\_numeric Snow Clr MC OC;

plot npp.\*predicted.;

**run**;

title 'transformation y square';

**data** weather\_Train1;

set weather\_Train1;

y\_sqrt = Temperature\_c \*\***2**;

**run**;

TITLE'Normality plots after transformation( y square)';

\* r square and adj r2 decreased for the model 1 ;

**proc** **reg**;

model y\_sqrt = Humidity Wind\_Speed\_km\_h\_ Visibility\_km\_ Pressure\_\_millibars\_ time\_numeric Snow Clr MC OC;

plot npp.\*predicted.;

**run**;

title 'transformation via log';

**data** weather\_Train1;

set weather\_Train1;

Log\_Y = log(Temperature\_c);

**run**;

TITLE'Normality plots after transformation( log y)';

**proc** **reg**;

model Log\_Y = Humidity Wind\_Speed\_km\_h\_ Visibility\_km\_ Pressure\_\_millibars\_ time\_numeric Snow Clr MC OC;

plot npp.\*predicted.;

**run**;

\* model prediction for stepwise M1;

\* i have removed insignificant variables to use gmselect of model 2;

TITLE'prediction model 1 training set';

**data** pred;

input Temperature\_c Humidity Wind\_Speed\_km\_h\_ Visibility\_km\_ Pressure\_\_millibars\_ time\_numeric Snow Clr MC OC;

datalines;

. 1.00 11.1090 6.1985 1017.78 0 0 0 0 1

;

\*this prediction shows me that if the overcast is there and visibility is good then there wont be a rain and temprature should be closer to 10;

\* this observation is same as observation 1 in training data set;

**data** new;

set pred weather\_Train1;

**run**;

**proc** **reg** data = new;

model Temperature\_c = Humidity Wind\_Speed\_km\_h\_ Visibility\_km\_ Pressure\_\_millibars\_ time\_numeric Snow Clr MC OC/p clm cli alpha=**0.05**;

**run**;

TITLE ' MODEL 2 - Interaction terms';

**proc** **glmselect** data = weather\_Train1;

model Temperature\_c = Humidity|Wind\_Speed\_km\_h\_|Visibility\_km\_|Pressure\_\_millibars\_|time\_numeric|Snow|Clr|MC|OC @**3** / selection = stepwise(stop=cv);

**run**;

\*defining variables name ;

**data** Nwether\_123;

set Nwether\_123;

hwp = Humidity\*Wind\_Speed\_km\_h\_\*Pressure\_\_millibars\_;

hs = Humidity\*Snow;

ps = Pressure\_\_millibars\_\*Snow;

tmc = time\_numeric\*MC;

vtmc = Visibility\_km\_\*time\_numeric\*MC;

ptmc = Pressure\_\_millibars\_\*time\_numeric\*MC;

vsmc = Visibility\_km\_\*Snow\*MC;

hoc = Humidity\*OC;

htoc = Humidity\*time\_numeric\*OC;

soc = Snow\*OC;

**run**;

TITLE'prediction model 2 training set';

**data** pred;

input Temperature\_c Humidity Wind\_Speed\_km\_h\_ hwp hs ps tmc vtmc ptmc vsmc hoc htoc soc;

datalines;

. 1.00 11.1090 1 0 0 0 1 1 0 1 1 1

;

**data** new;

set pred Nwether\_123;

**run**;

**proc** **reg** data = new;

model Temperature\_c = Humidity Wind\_Speed\_km\_h\_ hwp hs ps tmc vtmc ptmc vsmc hoc htoc soc/p clm cli alpha=**0.05**;

**run**;

Title"new y(temp) ";

**data** Nwether\_123;

set Nwether\_123;

if selected then new\_temp = Temperature\_c;

**run**;

title "Evaluating test set";

**proc** **reg** data=Nwether\_123;

model new\_temp = Humidity Wind\_Speed\_km\_h\_ hwp hs ps tmc vtmc ptmc vsmc hoc htoc soc;

output out = outm2(where=(new\_temp=**.**)) p=yhat;

model new\_temp = Humidity Wind\_Speed\_km\_h\_ Visibility\_km\_ Pressure\_\_millibars\_ time\_numeric Snow Clr MC OC;;

output out = outm1(where=(new\_temp=**.**)) p=yhat;

**run**;

TITLE'Testing data set p hat values';

**proc** **print** data = outm1;

\* outm1 deifines data set containg model 1 predicted values for the test set;

**run**;

**proc** **print** data = outm2;

\* outm2 deifines data set containg model 2 predicted values for the test set;

**run**;

\* next task is to get difference for two models in test set;

\* this value will gives me the performance test stasts RMSE MAE R2 for TEST SET;

\* this is model 1 of testing data set N=436 to get RMSE r2 MAE values;

Title'difference between observed and predicted val in test set';

**data** outm1\_sum;

set outm1;

d = Temperature\_c - yhat;

absd = abs(d);

**run**;

**proc** **summary** data = outm1\_sum;

var d absd;

output out = outm1\_stats std(d) = rmse mean(absd) = mae;

**run**;

**proc** **print** data = outm1\_stats;

TITLE'Validation statastics for model1';

**run**;

**proc** **corr** data = outm1;

var Temperature\_c yhat;

**run**;

\* this is model 2 of testing data set N=436 to get RMSE r2 MAE values;

Title'difference between observed and predicted val in test set';

**data** outm2\_sum;

set outm2;

d = Temperature\_c - yhat;

absd = abs(d);

**run**;

**proc** **summary** data = outm2\_sum;

var d absd;

output out = outm2\_stats std(d) = rmse mean(absd) = mae;

**run**;

**proc** **print** data = outm2\_stats;

TITLE'Validation statastics for model2';

**run**;

**proc** **corr** data = outm2;

var Temperature\_c yhat;

**run**;

\* performing on test data set;

TITLE'5 - fold cross validation for model selection - step wise';

**proc** **glmselect** data = Nwether\_123

plots = (aseplot Criteria);

\* generate ASE plots;

partition fraction ( test = **0.20**);

model Temperature\_c = Humidity Wind\_Speed\_km\_h\_ Visibility\_km\_ Pressure\_\_millibars\_ time\_numeric Snow Clr MC OC /selection = stepwise( stop = cv) cvMethod = split(**5**) cvDetails = all;

**run**;

**Methodology II SAS Code:**

**\*MODEL 1;**

\*Import of data and creation of dummy variables;

**DATA** weather;

length Formatted\_Date $30.;

length summary $30.;

length daily\_Summary $40.;

INFILE "weather.csv" DELIMITER =',' MISSOVER FIRSTOBS=**2**;

INPUT Formatted\_Date Summary $ Precip\_type $ Temperature\_c Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Loud\_Cover Pressure\_\_millibars\_ Daily\_Summary $;

Wmonth=substr(Formatted\_Date,**6**,**2**);

format time\_numeric time8.;

time\_numeric = input(substr(Formatted\_Date,**12**,**2**) !! ':' !! substr(Formatted\_Date,**15**,**2**) !! ':' !! substr(Formatted\_Date,**18**,**2**),time8.);

Snow = ( Precip\_type = "snow");

Breeze = ( Summary = "Breezy");

B\_MC = ( Summary = "Breezy and Mostly Cloudy");

B\_OC = ( Summary = "Breezy and Overcast ");

B\_PC = ( Summary = "Breezy and Partly Cloudy");

Clr = ( Summary = "Clear");

Drizzle = (Summary = "Drizzle");

Fog = ( Summary = " Foggy");

HMC = (Summary = " Humid and Mostly Cloudy");

H\_OC = (Summary = "Humid and Overcast");

Lgt\_rain = ( Summary = "Light Rain");

MC = (Summary = "Mostly Cloudy");

OC = (Summary = "Overcast");

PCld = (Summary = "Partly Cloudy");

RN = (Summary = "Rain");

**run**;

**PROC** **PRINT**;

**RUN**;

TITLE "Select only the required months [month-JAN, FEB, MAR] ";

**proc** **sql**;

delete from weather

where wmonth not IN ('04','05','06');

**QUIT**;

**PROC** **PRINT**;

**RUN**;

TITLE'Sorts the data Formatted\_Date in Ascending Order';

**PROC** **SORT** DATA= weather OUT=Sort\_Weather;

By Formatted\_Date;

**PROC** **PRINT**;

**RUN**;

TITLE'Creation of Histogram (with dummy variables)';

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

var Temperature\_c;

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

**run**;

TITLE "Boxplots - Temp vs wmonth";

**PROC** **SORT**;

BY Wmonth;

**RUN**;

**PROC** **BOXPLOT**;

PLOT Temperature\_c\*Wmonth ;

**RUN**;

TITLE "Boxplots - Temp vs time\_numeric";

**PROC** **SORT**;

BY time\_numeric;

**RUN**;

**PROC** **BOXPLOT**;

PLOT Temperature\_c\*time\_numeric ;

**RUN**;

\*Splitting the Data into Test and Training;

TITLE'taining and testing data set';

**proc** **surveyselect**

data = weather out = Nwether\_456 seed = **231201**

samprate = **0.80** outall;

**proc** **print**;

**run**;

TITLE'print Nweather for month 4,5,6';

**proc** **print** data = Nwether\_456;

**run**;

TITLE'Extraction of Training Set Data Observations';

**data** weather\_Train1

(where = ( selected = **1** ));

set Nwether\_456;

**run**;

**proc** **print** data = weather\_Train1;

**run**;

\*Histogram Generation;

TITLE "Histogram for Temperature - training data set";

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

var Temperature\_c;

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

**run**;

\*Verifying full model(for Training dataset);

TITLE "Full Model";

**proc** **reg**;

model Temperature\_c = Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_

Wind\_Bearing\_degrees\_ Visibility\_km\_ Pressure\_\_millibars\_ time\_numeric

B\_MC B\_OC B\_PC Clr H\_OC MC OC PCld ;

**run**;

**QUIT**;

\*Model Selection Methods;

TITLE "Full Model";

**proc** **reg**;

model Temperature\_c = Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_

Wind\_Bearing\_degrees\_ Visibility\_km\_ Pressure\_\_millibars\_ time\_numeric

B\_MC B\_OC B\_PC Clr H\_OC MC OC PCld /selection=STEPWISE;

**run**;

**QUIT**;

\*Fitting the final model with significant predictors;

TITLE "Final Model";

**proc** **reg**;

model Temperature\_c = Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_

B\_OC MC OC;

**run**;

**QUIT**;

\*5-Fold Cross Validation;

TITLE "5-FOld Cross Validation";

**PROC** **glmselect** data=weather

plots=(asePlot Criteria);

partition fraction (test=**0.25**);

model Temperature\_c = Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_

B\_OC MC OC; / selection=stepwise(stop=cv)

cvMethod=split(**5**) cvDetails=all;

**RUN**;

\*Final Model (Provides Correlation Values);

TITLE "Correlation values";

**proc** **corr**;

Var Temperature\_c Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_

B\_OC MC OC;

**run**;

\*Verifcation of Standardized Estimates;

**proc** **reg**;

model Temperature\_c = Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_

B\_OC MC OC /stb;

**run**;

**QUIT**;

\*Verification of Tolerance Values;

**proc** **reg**;

model Temperature\_c = Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_

B\_OC MC OC /tol;

**run**;

**QUIT**;

\*Verification of Outliers,influential points,multicollinearity;

**PROC** **reg** data=weather;

model Temperature\_c = Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_

B\_OC MC OC / influence r vif;

plot student.\*( Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_

B\_OC MC OC predicted.);

plot npp.\*student.;

**run**;

\*Enables to remove the outlier/influential points;

**DATA** NewWeather;

SET weather;

if \_n\_=**1573** then delete;

**run**;

TITLE "Residual Analysis";

**PROC** **REG**;

model Temperature\_c = Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_

B\_OC MC OC;

plot student.\* predicted.;

plot student.\*(Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_

B\_OC MC OC);

plot npp.\*student.;

**run**;

\*TESTING DATASET PROCEDURE;

\*Procedure for extraction of Test set data observations;

TITLE'Test Set Data Observations - weather\_Test1';

**data** weather\_Test1

(where = ( selected = **0** ));

set Nwether\_456;

**run**;

**proc** **print** data = weather\_Test1;

**run**;

**data** pred;

input Temperature\_c Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ B\_OC MC OC;

datalines;

. 11.8056 0.71 10.9963 0 0 0

;

**data** new;

set pred weather\_Test1;

**run**;

**proc** **reg**;

model Temperature\_c = Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ B\_OC MC OC /p clm cli alpha=**0.05**;

**run**;

\*VALIDATION TESTING;

Title"new y(temp) ";

**data** Nwether\_456;

set Nwether\_456;

if selected then new\_temp = Temperature\_c;

**run**;

**PROC** **PRINT**;

**RUN**;

\*MODEL 1;

title "Evaluating test set";

**proc** **reg** data=Nwether\_456;

model new\_temp = Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ B\_OC MC OC;

output out = outm2(where=(new\_temp=**.**)) p=yhat;

Title'difference between observed and predicted val in test set';

**data** outm2\_sum;

set outm2;

d = Temperature\_c - yhat;

absd = abs(d);

**run**;

**proc** **summary** data = outm2\_sum;

var d absd;

output out = outm2\_stats std(d) = rmse mean(absd) = mae;

**run**;

**proc** **print** data = outm2\_stats;

TITLE'Validation statastics for model1';

**run**;

**proc** **corr** data = outm2;

var Temperature\_c yhat;

**run**;

**\*MODEL 2;**

TITLE "DataSet Import";

**DATA** weather;

length Formatted\_Date $30.;

length summary $30.;

length daily\_Summary $40.;

INFILE "weather.csv" DELIMITER =',' MISSOVER FIRSTOBS=**2**;

INPUT Formatted\_Date Summary $ Precip\_type $ Temperature\_c Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Loud\_Cover Pressure\_\_millibars\_ Daily\_Summary $;

Wmonth=substr(Formatted\_Date,**6**,**2**);

format time\_numeric time8.;

time\_numeric = input(substr(Formatted\_Date,**12**,**2**) !! ':' !! substr(Formatted\_Date,**15**,**2**) !! ':' !! substr(Formatted\_Date,**18**,**2**),time8.);

Snow = ( Precip\_type = "snow");

Breeze = ( Summary = "Breezy");

B\_MC = ( Summary = "Breezy and Mostly Cloudy");

B\_OC = ( Summary = "Breezy and Overcast ");

B\_PC = ( Summary = "Breezy and Partly Cloudy");

Clr = ( Summary = "Clear");

Drizzle = (Summary = "Drizzle");

Fog = ( Summary = " Foggy");

H\_OC = (Summary = "Humid and Overcast");

Lgt\_rain = ( Summary = "Light Rain");

MC = (Summary = "Mostly Cloudy");

OC = (Summary = "Overcast");

PCld = (Summary = "Partly Cloudy");

RN = (Summary = "Rain");

**run**;

**PROC** **PRINT**;

**RUN**;

title "Selection of required months [month-April, May, June] ";

**proc** **sql**;

delete from weather

where wmonth not IN ('04','05','06');

**QUIT**;

**PROC** **PRINT**;

**RUN**;

\*Sorts the data Formatted\_Date in Ascending Order;

**PROC** **SORT** DATA= weather OUT=Sort\_Weather;

By Formatted\_Date;

**PROC** **PRINT**;

**RUN**;

\*Generates Interaction variables;

TITLE "Interaction Variables";

**proc** **glmselect**;

model Temperature\_c = Apparent\_temperature\_\_C\_| Humidity|Wind\_Speed\_km\_h\_|Wind\_Bearing\_degrees\_|Visibility\_km\_|Pressure\_\_millibars\_|time\_numeric| B\_MC| B\_OC| B\_PC| Clr| Drizzle| Fog| H\_OC | Lgt\_rain| MC| OC| PCld| RN @**3** / selection = stepwise(stop=cv);

**run**;

\*Adding the Interaction terms for the existing dataset;

**data** weather;

set weather;

IV1 = Apparent\_temperature\_\_C\_\*Wind\_Speed\_km\_h\_;

IV2 = Humidity\*Wind\_Speed\_km\_h\_;

IV3 = Apparent\_temperature\_\_C\_\*Humidity\*Wind\_Speed\_km\_h\_;

IV4 = Apparent\_temperature\_\_C\_\*time\_numeric\*MC;

IV5 = Humidity\*time\_numeric\*MC;

IV6 =Pressure\_\_millibars\_\*time\_numeric\*MC;

**run**;

**PROC** **PRINT**;

**RUN**;

\*Creation of Histogram (with dummy & Interaction variables);

TITLE "Histogram for Temperature";

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

var Temperature\_c;

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

**run**;

\*Splitting the Data into Test and Training;

TITLE'taining and testing data set';

**proc** **surveyselect**

data = weather out = Nwether\_456 seed = **231201**

samprate = **0.80** outall;

**proc** **print**;

**run**;

TITLE'Extraction of Training Set from the split operation';

**data** weather\_Train2

(where = ( selected = **1** ));

set Nwether\_456;

**run**;

**proc** **print** data = weather\_Train2;

**run**;

\*Full Model;

TITLE "Full Model";

**proc** **reg** data=weather\_Train2;

model Temperature\_c = Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_

Visibility\_km\_ Loud\_Cover Pressure\_\_millibars\_ time\_numeric Snow Breeze

B\_MC B\_OC B\_PC Clr Drizzle Fog H\_OC Lgt\_rain MC OC PCld RN IV1 IV2 IV3 IV4 IV5 IV6;

**run**;

**QUIT**;

\*SELECTION METHODS;

TITLE "Full Model";

**proc** **reg**;

model Temperature\_c = Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_

Visibility\_km\_ Loud\_Cover Pressure\_\_millibars\_ time\_numeric Snow Breeze

B\_MC B\_OC B\_PC Clr Drizzle Fog H\_OC Lgt\_rain MC OC PCld RN IV1 IV2 IV3 IV4 IV5 IV6/selection=backward;

**run**;

**QUIT**;

\*Final Fitted Model;

TITLE "Fit Model";

**proc** **reg** data=weather\_Train2;

model Temperature\_c = Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ B\_MC B\_OC Clr MC PCld IV1 IV2 IV3 IV4 IV5 IV6;

**run**;

**QUIT**;

\*5-Fold Cross Validation;

TITLE "5-FOld Cross Validation";

**PROC** **glmselect** data=weather

plots=(asePlot Criteria);

partition fraction (test=**0.25**);

model Temperature\_c = Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ B\_MC B\_OC Clr MC PCld IV1 IV2 IV3 IV4 IV5 IV6 / selection=stepwise(stop=cv)

cvMethod=split(**5**) cvDetails=all;

**RUN**;

\*Verification of Residual Analysis Plots;

TITLE "Residual Analysis";

**proc** **reg**;

model Temperature\_c = Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ B\_MC B\_OC Clr MC PCld IV1 IV2 IV3 IV4 IV5 IV6;

plot Student.\* predicted.;

plot Student.\*(Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ B\_MC B\_OC Clr MC PCld IV1 IV2 IV3 IV4 IV5 IV6);

plot npp.\* Student.;

**run**;

\*Verification of Association/Correlation;

**PROC** **corr**;

Var Temperature\_c Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ B\_MC B\_OC Clr MC PCld IV1 IV2 IV3 IV4 IV5 IV6;

**RUN**;

**Quit**;

**PROC** **gplot**;

plot Temperature\_c\*(Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ B\_MC B\_OC Clr MC PCld IV1 IV2 IV3 IV4 IV5 IV6);

**run**;

**PROC** **sgscatter**;

matrix Temperature\_c Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ B\_MC B\_OC Clr MC PCld IV1 IV2 IV3 IV4 IV5 IV6;

**run**;

\*Verification of Outliers, influential points, multicollinearity;

**PROC** **reg** data=weather\_Train2;

model Temperature\_c = Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ B\_MC B\_OC Clr MC PCld IV1 IV2 IV3 IV4 IV5 IV6 / influence r tol stb vif;

plot student.\*(Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ B\_MC B\_OC Clr MC PCld IV1 IV2 IV3 IV4 IV5 IV6 predicted.);

plot npp.\*student.;

**run**;

\*Enables to remove the outlier/influential points;

**DATA** NewWeather;

SET weather;

if \_n\_=**1573** then delete;

**run**;

\*TESTING PHASE;

\*Prediction Procedure;

TITLE'Printing Training dataset of weathernew - weather\_Train1';

**data** weather\_Test2

(where = ( selected = **0** ));

set Nwether\_456;

**run**;

**proc** **print** data = weather\_Test2;

**run**;

\*Confidence Intervals, Prediction Intervals, Prediction Value;

**data** pred;

input Temperature\_c Apparent\_temperature\_\_C\_ B\_MC B\_OC B\_PC IV1 IV2 IV3 IV4 IV5 IV6;

datalines;

. 11.8056 0 0 0 129.817 7.8074 92.170 0.00 0 0

;

**data** new;

set pred weather\_Test2;

**run**;

**proc** **reg**;

model Temperature\_c = Apparent\_temperature\_\_C\_ B\_MC B\_OC B\_PC IV1 IV2 IV3 IV4 IV5 IV6/p clm cli alpha=**0.05**;

**run**;

\*VALIDATION TESTING;

Title"new y(temp) ";

**data** Nwether\_456;

set Nwether\_456;

if selected then new\_temp = Temperature\_c;

**run**;

**PROC** **PRINT**;

**RUN**;

title "Evaluating test set";

**proc** **reg** data=Nwether\_456;

model new\_temp = Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ B\_MC B\_OC Clr MC PCld IV1 IV2 IV3 IV4 IV5 IV6;

output out = outm2(where=(new\_temp=**.**)) p=yhat;

Title'difference between observed and predicted val in test set';

**data** outm2\_sum;

set outm2;

d = Temperature\_c - yhat;

absd = abs(d);

**run**;

**proc** **summary** data = outm2\_sum;

var d absd;

output out = outm2\_stats std(d) = rmse mean(absd) = mae;

**run**;

**proc** **print** data = outm2\_stats;

TITLE'Validation statastics for model2';

**run**;

**proc** **corr** data = outm2;

var Temperature\_c yhat;

**run**;

**Methodology III SAS Code:**

title "Separate date and time";

**DATA** weather;

length Formatted\_Date $30.;

length summary $30.;

length daily\_Summary $40.;

INFILE "weather.csv" DELIMITER =',' MISSOVER FIRSTOBS=**2**;

INPUT Formatted\_Date Summary $ Precip\_type $ Temperature\_c Apparent\_temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Loud\_Cover Pressure\_\_millibars\_ Daily\_Summary $;

Wmonth=substr(Formatted\_Date,**6**,**2**);

format time\_numeric time8.;

time\_numeric = input(substr(Formatted\_Date,**12**,**2**) !! ':' !! substr(Formatted\_Date,**15**,**2**) !! ':' !! substr(Formatted\_Date,**18**,**2**),time8.);

DPrecip = ( Precip\_type = "snow");

DBreezy = ( Summary = "Breezy");

DB\_M\_Cloudy = ( Summary = "Breezy and Mostly Cloudy");

DB\_Cast = ( Summary = "Breezy and Overcast ");

DB\_P\_Cloudy = ( Summary = "Breezy and Partly Cloudy");

DClr = ( Summary = "Clear");

DDrizzle = (Summary = "Drizzle");

DFroggy = ( Summary = " Froggy");

DHum\_Cast = (Summary = "Humid and Overcast");

DL\_Rain = ( Summary = "Light Rain");

DM\_Cloudy = (Summary = "Mostly Cloudy");

DCast = (Summary = "Overcast");

DP\_Cloudy = (Summary = "Partly Cloudy");

DRain = (Summary = "Rain");

**run**;

**proc** **print**;

**run**;

title "Select only the required months ";

**proc** **sql**;

create table weathernew as

select \* from weather

where wmonth IN ('07','08','09');

**QUIT**;

**proc** **print**;

**run**;

TITLE'Spliting dataset into Training and Testing dataset';

**proc** **surveyselect** data = weathernew out = New\_Weather789 seed = **415217** samprate = **0.80** outall;

**proc** **print**;

**run**;

TITLE'Training dataset of New\_Weather789';

**data** weathernew\_Train1

(where = ( selected = **1** ));

set New\_Weather789;

**run**;

**proc** **print** data = weathernew\_Train1;

**run**;

TITLE "Histogram for Temperature";

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

var Temperature\_c;

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

**run**;

TITLE "Boxplot - time\_numeric and Temperature\_c";

**proc** **sort**;

by time\_numeric;

**run**;

**proc** **boxplot**;

plot Temperature\_c\*time\_numeric;

**run**;

TITLE "Boxplot - WMonth and Temperature\_c";

**proc** **sort**;

by WMonth;

**run**;

**proc** **boxplot**;

plot Temperature\_c\*WMonth;

**run**;

TITLE "Correlation values";

**proc** **corr**;

var Temperature\_c Humidity Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Pressure\_\_millibars\_ time\_numeric DClr DM\_Cloudy DCast DP\_Cloudy;

**run**;

TITLE "Full Model";

**proc** **reg**;

model Temperature\_c = Humidity Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Pressure\_\_millibars\_ time\_numeric DClr DM\_Cloudy DCast DP\_Cloudy/stb;

**run**;

TITLE "Prediction Model 1 Training Set";

**data** pred;

input Temperature\_c Humidity Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Pressure\_\_millibars\_ DClr DM\_Cloudy DCast DP\_Cloudy;

datalines;

. 0.70 3.6708 10 15.5526 1019.55 0 0 0 1

;

**data** new;

set pred weathernew\_Train1;

**run**;

**proc** **reg**;

model Temperature\_c= Humidity Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Pressure\_\_millibars\_ DClr DM\_Cloudy DCast DP\_Cloudy/p clm cli alpha=**0.05**;

**run**;

TITLE "Residual Analysis";

**proc** **reg**;

model Temperature\_c = Humidity Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Pressure\_\_millibars\_ DClr DM\_Cloudy DCast DP\_Cloudy;

\*Residual Plot: residuals vs predicted values;

plot student.\*predicted.;

\*Residual Plot: residuals vs x variables;

\*Linearity, Independence, Constant Variance Assumptions;

plot student.\*(Humidity Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Pressure\_\_millibars\_ DClr DM\_Cloudy DCast DP\_Cloudy);

\*Normal Probability Plot or QQ Plot;

\*Normality Assumption;

plot npp.\*student.;

**run**;

TITLE "Interaction Variables";

**proc** **glmselect**;

model Temperature\_c = Wind\_Speed\_km\_h\_|Wind\_Bearing\_degrees\_|Visibility\_km\_ Humidity|Pressure\_\_millibars\_ DClr DM\_Cloudy DCast DP\_Cloudy @**3** / selection = stepwise(stop=cv);

**run**;

TITLE "Prediction Model 2 Training Set";

**data** pred;

input Temperature\_c Wind\_Speed\_km\_h\_ Visibility\_km\_ wsv Humidity Pressure\_\_millibars\_ hpm DP\_Cloudy;

datalines;

. 10.9963 16.1 0.71 1013.95 0

;

**data** new;

set pred weathernew\_Train1;

**run**;

**proc** **reg**;

model Temperature\_c= Temperature\_c Wind\_Speed\_km\_h\_ Visibility\_km\_ wsv Humidity Pressure\_\_millibars\_ hpm DP\_Cloudy/p clm cli alpha=**0.05**;

**run**;

TITLE"Create new\_temp";

**data** New\_Weather789;

set New\_Weather789;

if selected then new\_temp=Temperature\_c;

**run**;

**proc** **print** data= New\_Weather789;

**run**;

**data** New\_Weather789;

set New\_Weather789;

hws = Humidity\*Wind\_Speed\_km\_h\_;

wsv = Wind\_Speed\_km\_h\_\*Visibility\_km\_;

hpm = Humidity\*Pressure\_\_millibars\_;

**run**;

TITLE "Validation-test set";

**proc** **reg** data=New\_Weather789;

\*Model 1;

model new\_temp = Humidity Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Pressure\_\_millibars\_ DClr DM\_Cloudy DCast DP\_Cloudy;

output out = outm1(where=(new\_temp=**.**)) p=yhat;

\*Model 2;

model new\_temp = Wind\_Speed\_km\_h\_ Visibility\_km\_ wsv Humidity Pressure\_\_millibars\_ hpm DP\_Cloudy;

output out = outm2(where=(new\_temp=**.**)) p=yhat;

**run**;

TITLE'Difference between Observed and Predicted in Test Set- Model1';

**data** outm1\_sum;

set outm1;

\*d is the difference between observed and predicted values in test set;

d = Temperature\_c - yhat;

absd = abs(d);

**run**;

**proc** **summary** data = outm1\_sum;

var d absd;

output out = outm1\_stats std(d) = rmse mean(adsd) = mae;

**run**;

**proc** **print** data = outm1\_stats;

TITLE'Validation statistics for model';

**run**;

**proc** **corr** data = outm1;

var Temperature\_c yhat;

**run**;

TITLE' Difference between Observed and Predicted in Test Set- Model2';

**data** outm2\_sum;

set outm2

d = Temperature\_c - yhat;

absd = abs(d);

**run**;

**proc** **summary** data = outm2\_sum;

var d absd;

output out = outm2\_stats std(d) = rmse mean(adsd) = mae;

**run**;

**proc** **print** data = outm2\_stats;

TITLE'Validation statistics for model';

**run**;

**proc** **corr** data = outm2;

var Temperature\_c yhat;

**run**;

TITLE'5 - fold cross validation for model selection - step wise';

**proc** **glmselect** data = New\_Weather789

plots = (aseplot Criteria);

\* generate ASE plots;

partition fraction ( test = **0.20**);

model Temperature\_c = Humidity Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Pressure\_\_millibars\_ DClr DM\_Cloudy DCast DP\_Cloudy /

selection = stepwise( stop = cv) cvMethod = split(**10**) cvDetails = all;

**run**;

TITLE'5 - fold cross validation for model selection - backward';

**proc** **glmselect** data = New\_Weather789

plots = (aseplot Criteria);

\* generate ASE plots;

partition fraction ( test = **0.20**);

model Temperature\_c = Humidity Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Pressure\_\_millibars\_ DClr DM\_Cloudy DCast DP\_Cloudy /

selection = backward( stop = cv) cvMethod = split(**10**) cvDetails = all;

**run**;

**Methodology IV SAS Code:**

TITLE "Separate date and time";

**DATA** weather;

length Formatted\_Date $30.;

length summary $30.;

INFILE "weather.csv" DELIMITER =',' MISSOVER FIRSTOBS=**2** ;

INPUT Formatted\_Date Summary $ Precip\_type $ Temperature\_c Apparent\_Temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Loud\_Cover Pressure\_\_millibars\_ ;

Wmonth=substr(Formatted\_Date,**6**,**2**);

PTN = ( Precip\_type = "snow");

B1 = ( Summary = "Breezy");

B2 = ( Summary = "Breezy and Mostly Cloudy");

B3 = ( Summary = "Breezy and Overcast ");

B4 = ( Summary = "Breezy and Partly Cloudy");

B5 = ( Summary = "Clear");

B6 = (Summary = "Drizzle");

B7 = ( Summary = " Foggy");

B8 = (Summary = "Humid and Overcast");

B9 = (Summary = "Humid and Mostly Cloudy");

B10 = ( Summary = "Light Rain");

B11 = (Summary = "Overcast");

B12 = (Summary = "Partly Cloudy");

B13 = (Summary = "Rain");

PROC PRINT;

**RUN**;

TITLE "Select only the required months";

**PROC** **SQL**;

DELETE FROM Weather

WHERE Wmonth NOT IN ('10','11','12');

**QUIT**;

PROC PRINT;

**RUN**;

TITLE "EXTRACT TIME AND REPLACE MISSING VALUES";

**DATA** Weather;

SET WEATHER;

FORMAT time\_numeric TIME8.;

time\_numeric = input(substr(Formatted\_Date,**12**,**2**) !! ':' !! substr(Formatted\_Date,**15**,**2**) !! ':' !! substr(Formatted\_Date,**18**,**2**),TIME8.);

IF Wmonth = '10' THEN PTN = **0**;

IF Wmonth = '11' THEN PTN = **0**;

IF Wmonth = '12' THEN PTN = **1**;

**RUN**;

TITLE "SPLITing 80-20";

**DATA** Weather;

SET Weather;

selected = **0**;

n=RANUNI(**7777**);

**DATA** weather;

SET Weather nobs=nobs;

IF \_n\_<=**.8**\*nobs THEN selected=**1**;

**RUN**;

TITLE"Create interaction terms";

**data** weather;

set weather;

wswbv = Wind\_Bearing\_degrees\_\*Wind\_Speed\_km\_h\_\*Visibility\_km\_;

hump = Humidity\*Pressure\_\_millibars\_;

humptime = Humidity\*Pressure\_\_millibars\_\*Time\_numeric;

**run**;

TITLE "creating a dataset calles as weather\_training";

**proc** **sql**;

create table weather\_training as

select \* from weather where selected = **1**;

**quit**;

**run**;

TITLE "Histogram analysis of Temperature\_c";

**PROC** **Univariate** normal data=weather;

var Temperature\_c;

HISTOGRAM/NORMAL (mu = est sigma = est);

**RUN**;

TITLE "Histogram analysis of Humidity";

**PROC** **Univariate** normal data=weather;

var Humidity;

HISTOGRAM/NORMAL (mu = est sigma = est);

**RUN**;

TITLE "Histogram analysis of Wind\_Speed\_km\_h\_";

**PROC** **Univariate** normal data=weather;

var Wind\_Speed\_km\_h\_;

HISTOGRAM/NORMAL (mu = est sigma = est);

**RUN**;

TITLE "Histogram analysis of Visibility\_km\_";

**PROC** **Univariate** normal data=weather;

var Visibility\_km\_;

HISTOGRAM/NORMAL (mu = est sigma = est);

**RUN**;

TITLE "Histogram analysis of Pressure\_\_millibars\_";

**PROC** **Univariate** normal data=weather;

var Pressure\_\_millibars\_;

HISTOGRAM/NORMAL (mu = est sigma = est);

**RUN**;

TITLE "Histogram analysis of Time\_numeric";

**PROC** **Univariate** normal data=weather;

var Time\_numeric;

HISTOGRAM/NORMAL (mu = est sigma = est);

**RUN**;

\*In Regression all the assumptions are made for the dependent variabel and hence the distribution of the independent variable is not required;

TITLE "BOXPLOT OF TIME AND TEMPERATURE";

**PROC** **SORT** DATA = weather;

BY time\_numeric;

**RUN**;

**PROC** **BOXPLOT**;

PLOT Temperature\_c\*time\_numeric;

**RUN**;

TITLE "BOXPLOT OF MONTH AND TEMPERATURE";

**PROC** **SORT** DATA = weather;

BY Wmonth;

**RUN**;

**PROC** **BOXPLOT**;

PLOT Temperature\_c\*wmonth;

**RUN**;

TITLE "BOXPLOT OF Pressure AND TEMPERATURE";

**PROC** **SORT** DATA = weather;

BY time\_numeric;

**RUN**;

**PROC** **BOXPLOT**;

PLOT Pressure\_\_millibars\_\*time\_numeric;

**RUN**;

TITLE "BOXPLOT OF VISIBILITY AND TEMPERATURE";

**PROC** **SORT** DATA = weather;

BY time\_numeric;

**RUN**;

**PROC** **BOXPLOT**;

PLOT Visibility\_km\_\*time\_numeric;

**RUN**;

TITLE "BOXPLOT OF Wind Speed AND TEMPERATURE";

**PROC** **SORT** DATA = weather;

BY time\_numeric;

**RUN**;

**PROC** **BOXPLOT**;

PLOT Wind\_Speed\_km\_h\_\*time\_numeric;

**RUN**;

TITLE "Distribution of conditions";

**proc** **freq**;

tables summary;

**run**;

TITLE "Correlation with dummy variables";

**PROC** **CORR**;

VAR Temperature\_c Apparent\_Temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ B1-B13 PTN Wind\_Bearing\_degrees\_ Visibility\_km\_ Time\_numeric Pressure\_\_millibars\_ ;

**RUN**;

TITLE "Correlation with dummy variables only significant";

**PROC** **CORR**;

VAR Temperature\_c Apparent\_Temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ B5 B6 B10-B12 PTN Wind\_Bearing\_degrees\_ Visibility\_km\_ Time\_numeric Pressure\_\_millibars\_ ;

**RUN**;

TITLE "Correlation";

**PROC** **CORR**;

VAR Temperature\_c Apparent\_Temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Time\_numeric Pressure\_\_millibars\_ ;

**RUN**;

TITLE "Scatterplots";

**PROC** **sgscatter**;

MATRIX Temperature\_c Apparent\_Temperature\_\_C\_ Humidity Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Time\_numeric Pressure\_\_millibars\_;

**RUN**;

TITLE "Basic model:1 with Apparent Temperature";

**PROC** **REG** DATA=weather\_training;

model Temperature\_c = Apparent\_Temperature\_\_C\_ Humidity B1-B13 PTN Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Time\_numeric Pressure\_\_millibars\_ / stb;

**run**;

TITLE "Normality plot";

**PROC** **REG** DATA=weather\_training;

model Temperature\_c = Apparent\_Temperature\_\_C\_ Humidity B1-B13 PTN Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Time\_numeric Pressure\_\_millibars\_;

plot npp.\*student.;

**run**;

TITLE "Basic model:2 without Apparent Temperature";

**PROC** **REG** DATA=weather\_training;

model Temperature\_c = Humidity B1-B13 PTN Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Time\_numeric Pressure\_\_millibars\_/ stb;

**run**;

TITLE "Normality plot";

**PROC** **REG** DATA=weather\_training;

model Temperature\_c = Humidity B1-B13 PTN Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Time\_numeric Pressure\_\_millibars\_;

**run**;

TITLE "Model:3 without insignificant terms";

**PROC** **REG** DATA=weather\_training;

model Temperature\_c = Humidity B1-B13 PTN Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Time\_numeric Pressure\_\_millibars\_ / selection = stepwise stb sls = **0.05** sle = **0.05**;

**run**;

TITLE "Normality plot for this model";

**PROC** **REG** DATA=weather\_training;

model Temperature\_c = Humidity B1-B13 PTN Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Time\_numeric Pressure\_\_millibars\_ / selection = stepwise stb sls = **0.05** sle = **0.05**;

plot npp.\*student.;

plot student.\*predicted.;

plot student.\*(Humidity B5 B6 B10 B11 PTN Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Time\_numeric Pressure\_\_millibars\_);

**run**;

TITLE "Model 4: Apply sqrt transformation";

**data** weather\_training;

set weather\_training;

sqrt\_temp = sqrt(Temperature\_c);

**run**;

TITLE "Histogram for the sqrt";

**PROC** **Univariate** normal data=weather\_training;

var sqrt\_temp;

HISTOGRAM/NORMAL (mu = est sigma = est);

**RUN**;

TITLE "sqrt\_temp model";

**PROC** **REG** DATA=weather\_training;

model sqrt\_temp = Humidity B1-B13 PTN Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Time\_numeric Pressure\_\_millibars\_/ selection = stepwise stb sls = **0.05** sle = **0.05**;

**run**;

TITLE "Model 5: Apply 1/4th root transformation";

**data** weather\_training;

set weather\_training;

tempBy4 = temperature\_C\*\*(**1**/**4**);

**run**;

TITLE "Histogram for the sqrt";

**PROC** **Univariate** normal data=weather\_training;

var tempBy4;

HISTOGRAM/NORMAL (mu = est sigma = est);

**RUN**;

TITLE "1/4 model";

**PROC** **REG** DATA=weather\_training;

model tempBy4 = Humidity B1-B13 PTN Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Time\_numeric Pressure\_\_millibars\_/ selection = stepwise stb sls = **0.05** sle = **0.05**;

**run**;

**proc** **reg** data=weather\_training;

model tempBy4 = Humidity B1-B13 PTN Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Time\_numeric Pressure\_\_millibars\_/ selection = stepwise stb sls = **0.05** sle = **0.05**;

plot npp.\*student.;

**run**;

TITLE "Model 6:GLMSELECT";

ods graphics on;

**PROC** **glmselect** DATA = weather\_training plots =criteria;

model Temperature\_c = B6 B9 B10 B12 PTN Wind\_Speed\_km\_h\_|Wind\_Bearing\_degrees\_|Visibility\_km\_ Humidity|Pressure\_\_millibars\_|Time\_numeric @**3**/selection=stepwise(stop=CV);

**run**;

ods graphics off;

ods graphics on;

TITLE "Normality plot after adding interaction terms";

**proc** **reg** data=weather\_training;

model temperature\_c = B6 B10 PTN wswbv hump humptime time\_numeric Pressure\_\_millibars\_ Wind\_Speed\_km\_h\_ Visibility\_km\_;

plot npp.\*student.;

plot student.\*predicted.;

plot student.\*(B6 B10 PTN wswbv hump humptime time\_numeric Pressure\_\_millibars\_ Wind\_Speed\_km\_h\_ Visibility\_km\_);

**run**;

ods graphics off;

TITLE "setting the newtemp for training set";

**data** weather;

set weather;

if selected then newtemp=temperature\_c;

**run**;

TITLE "Prediction on test set using model 1";

**proc** **reg** data = weather;

model newtemp = Humidity B5 B6 B10 B11 PTN Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Time\_numeric Pressure\_\_millibars\_ ;

output out = outmodel1(where=(newtemp=**.**)) p=yhat;

**run**;

TITLE "Prediction on test set using model 2";

**proc** **reg** data = weather;

model newtemp = B6 B10 PTN wswbv hump humptime time\_numeric Pressure\_\_millibars\_ Wind\_Speed\_km\_h\_ Visibility\_km\_;

output out = outmodel2(where=(newtemp=**.**)) p=yhat;

**run**;

TITLE "Difference between observed and Predicted for model 1";

**data** outmodel1\_stat;

set outmodel1;

d = Temperature\_c - yhat;

absd = abs(d);

**run**;

**proc** **summary** data = outmodel1\_stat;

var d absd;

output out=outm1\_stats std(d) = rmse mean(absd)=mae;

**run**;

**proc** **print** data = outm1\_stats;

TITLE"Stat validation";

**run**;

**proc** **corr** data=outmodel1;

var temperature\_c yhat;

**run**;

TITLE "Difference between observed and Predicted for model 2";

**data** outmodel2\_stat;

set outmodel2;

d = Temperature\_c - yhat;

absd = abs(d);

**run**;

**proc** **summary** data = outmodel2\_stat;

var d absd;

output out=outm2\_stats std(d) = rmse mean(absd)=mae;

**run**;

**proc** **print** data = outm2\_stats;

TITLE"Stat validation";

**run**;

**proc** **corr** data=outmodel2;

var temperature\_c yhat;

**run**;

TITLE "Prediction on test set using model 1 including Apparent temperature";

**proc** **reg** data = weather;

model newtemp = Apparent\_Temperature\_\_C\_ Humidity B5 B6 B10 B11 PTN Wind\_Speed\_km\_h\_ Wind\_Bearing\_degrees\_ Visibility\_km\_ Time\_numeric Pressure\_\_millibars\_ ;

output out = outmodel11(where=(newtemp=**.**)) p=yhat;

**run**;

TITLE "Prediction on test set using model 2 including Apparent Temperature";

**proc** **reg** data = weather;

model newtemp = B6 B10 PTN wswbv hump humptime time\_numeric Pressure\_\_millibars\_ Wind\_Speed\_km\_h\_ Visibility\_km\_ Apparent\_Temperature\_\_C\_;

output out = outmodel21(where=(newtemp=**.**)) p=yhat;

**run**;

TITLE "Difference between observed and Predicted for model 1 after adding apparent temperature";

**data** outmodel11\_stat;

set outmodel11;

d = Temperature\_c - yhat;

absd = abs(d);

**run**;

**proc** **summary** data = outmodel11\_stat;

var d absd;

output out=outm11\_stats std(d) = rmse mean(absd)=mae;

**run**;

**proc** **print** data = outm11\_stats;

TITLE"Stat validation";

**run**;

**proc** **corr** data=outmodel11;

var temperature\_c yhat;

**run**;

TITLE "Difference between observed and Predicted for model 2 after adding apparent temperature";

**data** outmodel21\_stat;

set outmodel21;

d = Temperature\_c - yhat;

absd = abs(d);

**run**;

**proc** **summary** data = outmodel21\_stat;

var d absd;

output out=outm21\_stats std(d) = rmse mean(absd)=mae;

**run**;

**proc** **print** data = outm21\_stats;

TITLE"Stat validation";

**run**;

**proc** **corr** data=outmodel21;

var temperature\_c yhat;

**run**;

**References**

1.Implementation of GOSSTANDART technique for verifying and validating Goodness of Fit and maximal test power:

B.Yu.Lemeshko, S.N.Postovalov, E.V.Chimitova, Rules of Application of Goodness- of – fit in simple and Composite Hypothesis Testing, pg. 126 – 132.

2. Implementation of Vector Regression Modeling technique for better/accurate prediction:

Kavitha S, Varuna S, Ramya R, A Comparative Analysis on Linear Regression and Support Vector Regression, 2016 Online International Conference on Green Engineering and Technologies (IC-GET)

3. Proposes MRDDV (Mutiple linear Regression with Dependent Dummy Variable) that combines quantitative and qualitative variables to estimate the behavior of other predictors

N.J.Park, K.M. George-N.Park, A Multiple Regression Model for Trend Change Prediction, Proceedings of the *ITI 2010 32nd Int. Conf. on Information Technology Interfaces,* June 2010.

**Future Work**

* Combining different datasets to come up with better model equation.
* We can enhance this model by adding more influential variable such as greenhouse gas indicators, pollutants, forest cover and population indicators.

**Limitations**

* + More variables and observations would be good for learning.
  + The frequency of recording the observation is done on hourly basis and it’d be better if it was done 4 times in an hour.
  + It’s hard to explain sudden changes using Linear Regression models.