IDA PROJECT REPORT

TOPIC 15: REGRESSION ANALYSIS FOR ESTABLISHING A RELATION BETWEEN WEATHER PARAMETERS

SUBMITTED BY: GROUP 15

ARAVIND CHALLA (S20190010169)

SATHWIK RAO (S20190010189)

SAI SATWIK (S20190010092)

LOKESH MORTHA (S20190010167)

PRADYUMNA REDDY (S20190020226)

PROBLEM:

a) Find the relation between each weather parameters. Do proper data pre-processing required before checking for the relation.
b) Report your relation analysis with the following approaches.
    1) Simple linear regression analysis.
    2) Simple non-linear regression analysis.
c) Calculate R^2 values in each of the above-mentioned case and finally conclude your results precisely.

DATASET USED: Climate Data

DATE OF SUBMISSION: 30 NOVEMBER 2021

UNDERSTANDING THE THEORY:

REGRESSION ANALYSIS:

Regression analysis is a statistical method to deal with the formulation of a mathematical model depicting relationships amongst variables, which can be used for the purpose of prediction of the values of the dependent variable, given the values of independent variables.

Regression analysis can be done in many ways like:

1) Linear Regression
2) Logistic Regression
3) Polynomial Regression
4) Lasso Regression
5) Ridge Regression

Here in our project, we mainly deal with Linear Regression and Non-Linear Regression.

LINEAR REGRESSION ANALYSIS:

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the **dependent variable.** The variable you are using to predict the other variable's value is called the **independent variable.**

In linear regression analysis, we assume that the regression function E(Y|X) is a linear function of X: The parameter β is the slope of the true regression line, and can be interpreted as the population mean change in y for a unit change in x.

$$E(Y \mid x)=\alpha +\beta x$$

The parameter β is the slope of the true regression line, and can be interpreted as the population mean change in y for a unit change in x.

The parameter α is the intercept of the true regression line and can be interpreted as the mean value of Y when X is zero.

In Real-Life Scenarios:

In practice, α and β will be unknown parameters, which must be estimated from our data as our observed data will not lie exactly on the true regression line.

So, Instead, we assume that the true regression line is observed with error, i.e., that y_i is given by

$$\mathbf{y_i = \alpha + \beta\, x_i + \varepsilon_i}$$

Here, ε_i is a normally distributed random variable (called the error) with mean 0 and variance σ^2 which does not depend on x. The variance of the error, σ2, measures how close the points are to the true line, in terms of expected squared vertical deviation.

If the true regression line with the error holds, then the values of y will be randomly scattered about the true regression line, and the mean value of y for a given x will be this true regression line, $E(y| x)=\alpha +\beta x$.

LEAST SQUARE ERROR METHOD:

This method uses the concept of residual. A residual is essentially an error in the fit of the model Y = ax + b. Thus i^{th} residual is:

$$\mathbf{e_i = Y_i - \hat{Y}_i} \quad \textbf{where i=1,2,3,4...n}$$

The Residual Sum of Squares also called the sum of squares of the errors(SSE) about the fitted line is denoted as SSE

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i\ )^2$$

We minimize SSE and find the parameters a,b.

For the minimum value of SSE $\frac{\partial (SSE)}{\partial a} = 0$ $\frac{\partial (SSE)}{\partial b} = 0$

$$b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

After solving, we find the values of a and b are $a = \overline{y} - b\overline{x}$ (Eq.1)

NON-LINEAR REGRESSION ANALYSIS:

When the regression equation is in terms of r degree where r>1 we say it as non-linear regression.

Here the relationship between Response Y and Regressor X is

$$Y = aX^n + bX^{n-1} .... + \in$$

MEASURING THE QUALITY OF FIT - R^2

The variability of the fitted model is determined by the quantity R^2, also called as Coefficient of determination.

Sum of squared errors, SSE = $\sum_{i=1}^{n}(y_i - \hat{y})^2$, where $\hat{y}$ is the predicted output, while y is the actual output/label.

The total corrected sum of squares, SST = $\sum_{i=1}^{n}(y_i - \overline{y})^2$, where y is the predicted mean.

SST represents the variation in the response values. The R^2 is

$$R^2 = 1 - \frac{SSE}{SST} \quad \text{(Eq.2)}$$

If the $R^2 \sim 0$, then the model is said to be poorly fit.

$R^2 \sim 1$, then it is said to be a good fit.

IMPLEMENTATION:

DATASET:

We have first loaded the dataset, 'DailyDelhiClimateTrain.csv' file into a variable called "rawData" using the read.csv command.

	date	meantemp	humidity	wind_speed	meanpressure
1	2013-01-01	10.000000	84.50000	0.000000	1015.667
2	2013-01-02	7.400000	92.00000	2.980000	1017.800
3	2013-01-03	7.166667	87.00000	4.633333	1018.667
4	2013-01-04	8.666667	71.33333	1.233333	1017.167
5	2013-01-05	6.000000	86.83333	3.700000	1016.500
6	2013-01-06	7.000000	82.80000	1.480000	1018.000
7	2013-01-07	7.000000	78.60000	6.300000	1020.000
8	2013-01-08	8.857143	63.71429	7.142857	1018.714
9	2013-01-09	14.000000	51.25000	12.500000	1017.000
10	2013-01-10	11.000000	62.00000	7.400000	1015.667

	date	meantemp	humidity	wind_speed	meanpressure
1453	2016-12-23	15.55000	74.70000	1.205000	1014.250
1454	2016-12-24	17.31818	78.63636	5.236364	1011.318
1455	2016-12-25	14.00000	94.30000	9.085000	1014.350
1456	2016-12-26	17.14286	74.85714	8.784211	1016.952
1457	2016-12-27	16.85000	67.55000	8.335000	1017.200
1458	2016-12-28	17.21739	68.04348	3.547826	1015.565
1459	2016-12-29	15.23810	87.85714	6.000000	1016.905
1460	2016-12-30	14.09524	89.66667	6.266667	1017.905
1461	2016-12-31	15.05263	87.00000	7.325000	1016.100
1462	2017-01-01	10.00000	100.00000	0.000000	1016.000

PRE-PROCESSING:

We know that the "date" column will be unique for every record and don't carry any weightage to the model. So we removed this column from the dataset and loaded the processed data into a variable named "trainData".

	meantemp	humidity	wind_speed	meanpressure
1	10.000000	84.50000	0.000000	1015.667
2	7.400000	92.00000	2.980000	1017.800
3	7.166667	87.00000	4.633333	1018.667
4	8.666667	71.33333	1.233333	1017.167
5	6.000000	86.83333	3.700000	1016.500
6	7.000000	82.80000	1.480000	1018.000
7	7.000000	78.60000	6.300000	1020.000
8	8.857143	63.71429	7.142857	1018.714
9	14.000000	51.25000	12.500000	1017.000
10	11.000000	62.00000	7.400000	1015.667

	meantemp	humidity	wind_speed	meanpressure
1453	15.55000	74.70000	1.205000	1014.250
1454	17.31818	78.63636	5.236364	1011.318
1455	14.00000	94.30000	9.085000	1014.350
1456	17.14286	74.85714	8.784211	1016.952
1457	16.85000	67.55000	8.335000	1017.200
1458	17.21739	68.04348	3.547826	1015.565
1459	15.23810	87.85714	6.000000	1016.905
1460	14.09524	89.66667	6.266667	1017.905
1461	15.05263	87.00000	7.325000	1016.100
1462	10.00000	100.00000	0.000000	1016.000

FEATURE SELECTION:

As in our dataset, we have 4 Parameters:-

1) MeanTemp
2) Humidity
3) WindSpeed
4) MeanPressure

We will first use two variables:

1. MeanTemp (Independent Variable/ Control Variable/ Regressor Variable)

2. Humidity (Dependent Variable/ Response Variable)

Similarly, we then consider all the possible pairs of parameters and do regression analysis on them individually for each pair like:

1) MeanTemp vs Humidity (As mentioned above)
2) MeanTemp vs Wind_speed
3) MeanTemp vs MeanPressure
4) Humidity vs Wind_speed
5) Humidity vs MeanPressure
6) Wind_speed vs MeanPressure

SIMPLE LINEAR REGRESSION ANALYSIS:

As we know that we can find a value and b value from Eq.1, i.e.,

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

We implement the same in R.

```
b_value <- function(x,y){
  mean_x= mean(x)
  mean_y = mean(y)
  b_numer = sum((x-mean_x)*(y-mean_y))
  b_denom = sum((x-mean_x)^2)
  b= b_numer/b_denom
  return(b)
}
```

1) As discussed in Eq.1, we write code for finding the value of b in R. We calculate mean_x, mean_y. Then, we calculate the numerator and denominator terms in Eq.1 for the b value and find the value of b.

```
a_value <- function(x,y,b){
  x_mean = mean(x)
  y_mean = mean(y)
  a= y_mean - b*x_mean
  return(a)
}
```

2) Now, From Eq.1, we try to calculate a_value. We first calculate x_mean, y_mean and using them, we will find the value of a.

3) We now have the values of a and b. So we can easily plot our regression line. We plotted our regression line as follows:

```
plotting_regression_line <- function(x,y,a,b){
  print(a)
  print(b)
  y_predict = a + b*x
  return(y_predict)
}
```

4) We now calculate the Coefficient of determination R^2, for the fitted model, to determine the quality of the fit using the function below: (as we know how to calculate R^2 value from Eq.2)

```
R2_value <- function(y,y_predict){
  y_mean= mean(y)
  sst = sum((y-y_mean)^2)
  print(sst)
  sse = sum((y-y_predict)^2)
  print(sse)
  r2 = 1-(sse/sst)
  r2
  return(r2)
}
```

5) Now, we are ready with all the necessary functions to calculate values that are necessary for our linear regression analysis. So, we will use these functions and find the regression line, the intercepts, slope, and the R^2 value. For this, we use:

```
x1=trainData$meantemp
y1=trainData$humidity

b <- b_value(x1,y1)
b
a <- a_value(x1,y1,b)
a
y_predict<-plotting_regression_line(x1,y1,a,b)
R2 <- R2_value(y1,y_predict)
R2
```

First, we consider a pair of parameters(meantemp, humidity) from our given dataset, and perform simple linear regression. We then plotted the regression line using ggplot. The plotting looked as follows:

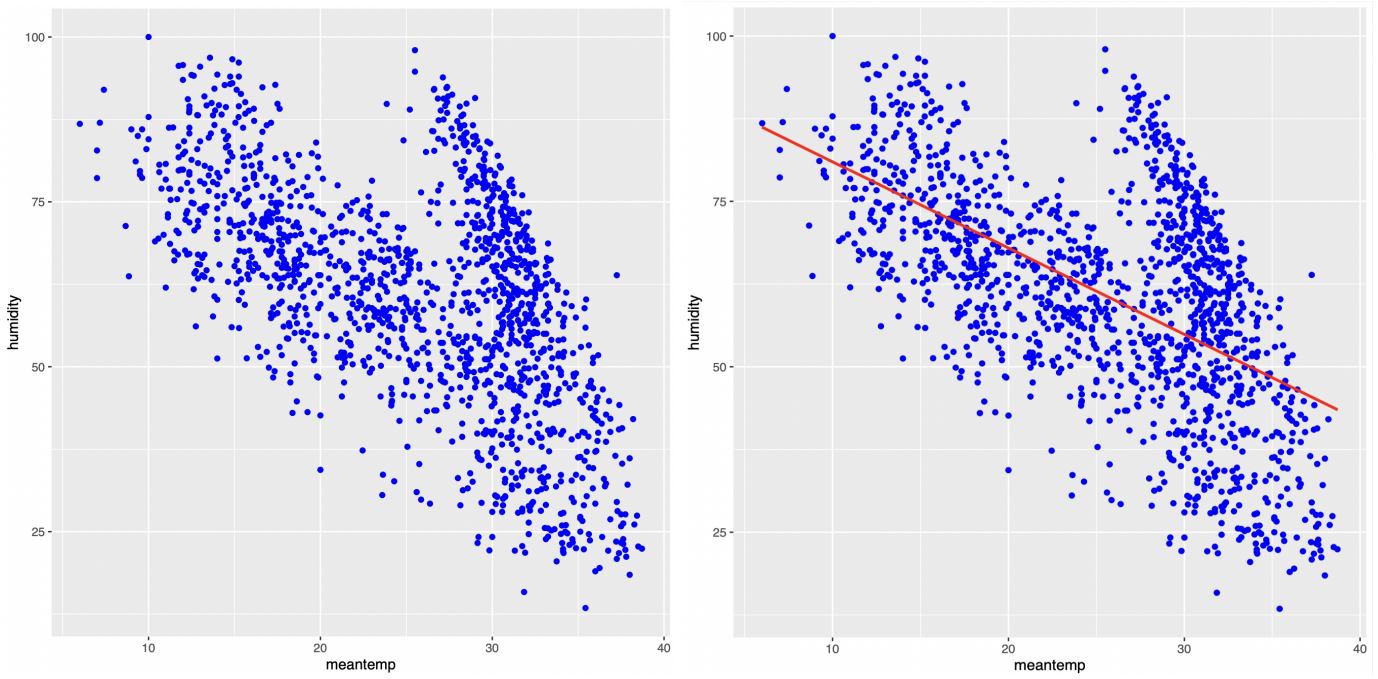


Fig.1 shows the scatter plot of the traindata parameters - meantemp, humidity.

Fig.2 shows the scatter plot of the traindata parameters - meantemp, humidity along with the regression line.

Looks like the regression line is good. But let us discuss more about the fit of the regression line in the RESULTS section.

SIMPLE NON-LINEAR REGRESSION ANALYSIS:

As we know that we can find the values of the coefficients can be calculated from the system of equations:

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i & \cdots & \sum_{i=1}^{n} x_i^k \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \cdots & \sum_{i=1}^{n} x_i^{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_i^k & \sum_{i=1}^{n} x_i^{k+1} & \cdots & \sum_{i=1}^{n} x_i^{2k} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \vdots \\ \sum_{i=1}^{n} x_i^k y_i \end{bmatrix}$$

(Eq.3)

We implemented the same in R.

```
A <- rbind(c(sum((x)^4), sum((x)^3),sum((x)^2)),
           c(sum((x)^3), sum((x)^2), sum(x)),
           c(sum((x)^2), sum(x), length(x)))
B <- c(sum(((x)^2)*y), sum(x*y),sum(y)  )
library(gmp)
q=solve(A,B)
q
y_predict2 <- q[3] + q[2]*x + q[1]*x**2
R2 <- R2_value(y,y_predict2)
R2
```

We input the values as a 3x3 matrix into A, and a 3x1 matrix into B with the values from the equation above.

We then solve this system of equations using solve(an inbuilt function).

This gives us the coefficients a, b, c in $y = a + b*x + c*x^2$. Here, the values q[3], q[2], q[1] is nothing but a, b, c values. Using this, we get the non-linear regression line for degree 2. We test the fit by calculating the R^2 value.

In our project, we did non-linear regression for degrees 2 and 3 without using any inbuilt models.

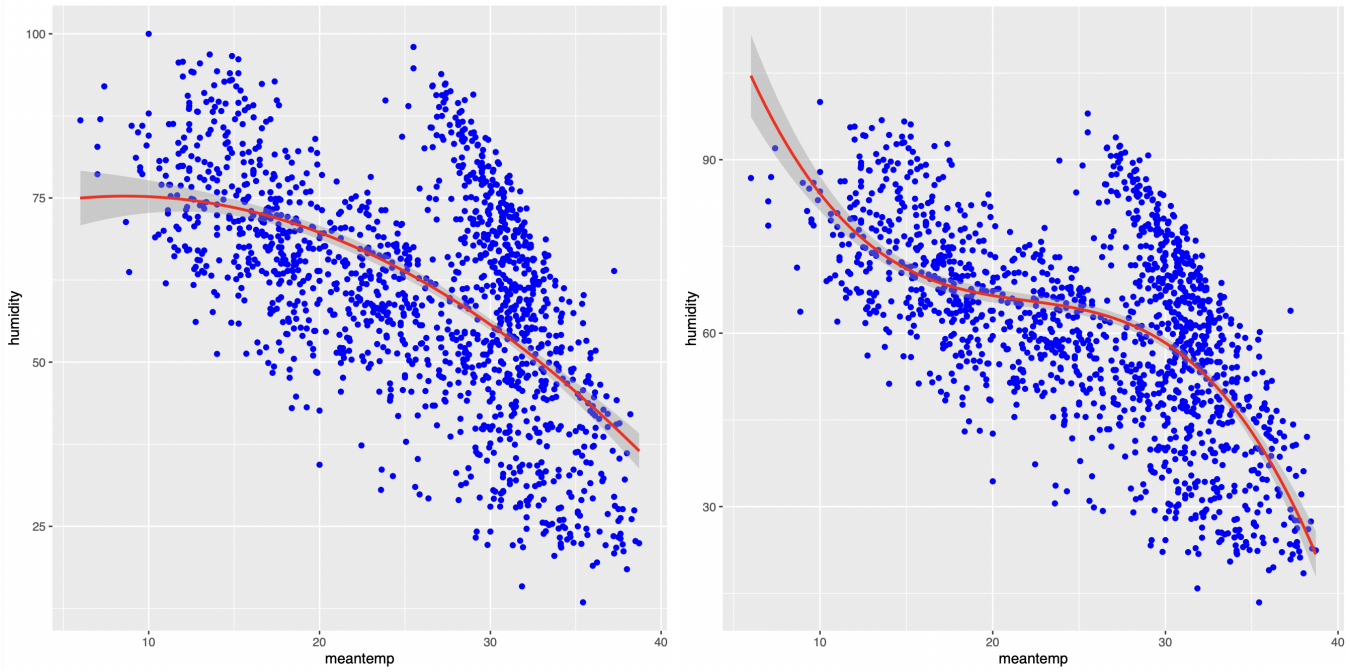
For degree 3: (using the same above Eq.3)

```
C <- rbind(c(sum((x)^6), sum((x)^5), sum((x)^4),sum((x)^3)),
           c(sum((x)^5), sum((x)^4), sum((x)^3),sum((x)^2)),
           c(sum((x)^4), sum((x)^3), sum((x)^2), sum(x)),
           c(sum((x)^3), sum((x)^2), sum(x), length(x)))
D <- c(sum(((x)^3)*y),sum(((x)^2)*y), sum(x*y),sum(y)  )
library(gmp)
p=solve(C,D)
p
y_predict3 <- p[4] + p[3]*x + p[2]*x**2 + p[1]*x**3
R2 <- R2_value(y,y_predict3)
R2
```

We input the values as a 4x4 matrix into c and a 4x1 matrix into D with the values from Eq.3.

This gives us the coefficients a, b, c, d in $y = a + b*x + c*x^2 + d*x^3$. Here, the values p[4], p[3], p[2], p[1] are nothing but a, b, c, d values. Using this, we get the non-linear regression line for degree 3. We test the fit by calculating the R^2 value.

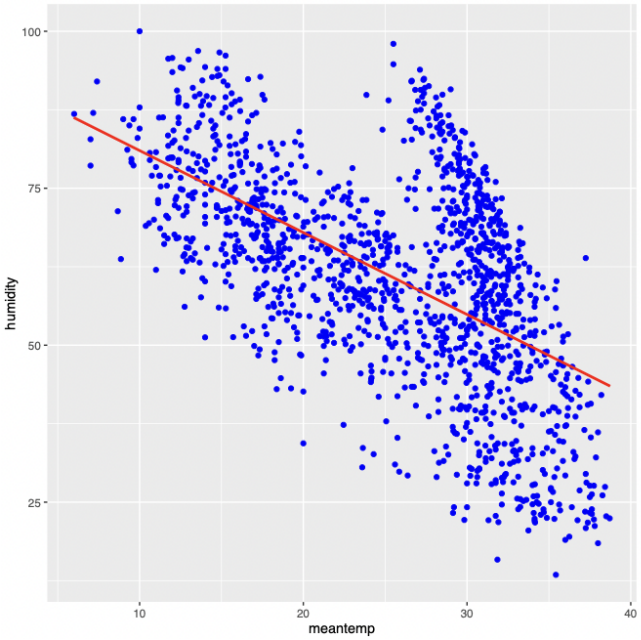
We then plotted the graph for degree 2 and degree 3 using ggplot.



Looks like the regression line is good. But let us discuss more the fit of the regression line in the RESULTS section.

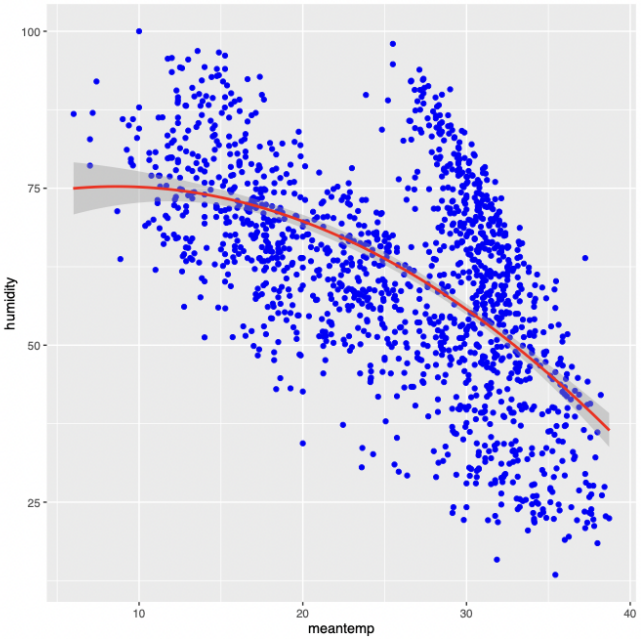
EXPERIMENTAL RESULTS:

For meantemp(x) Vs Humidity(y)



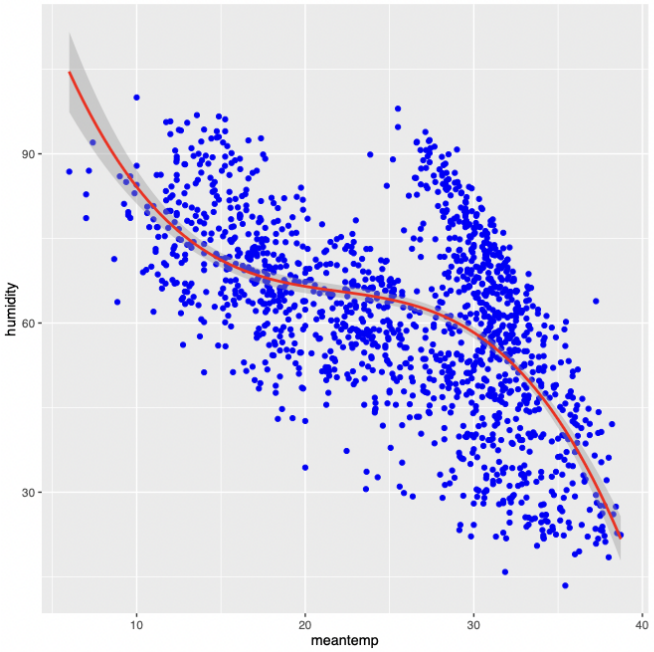
R^2 value = 0.3271276

y = 94.05 - 1.30529*x



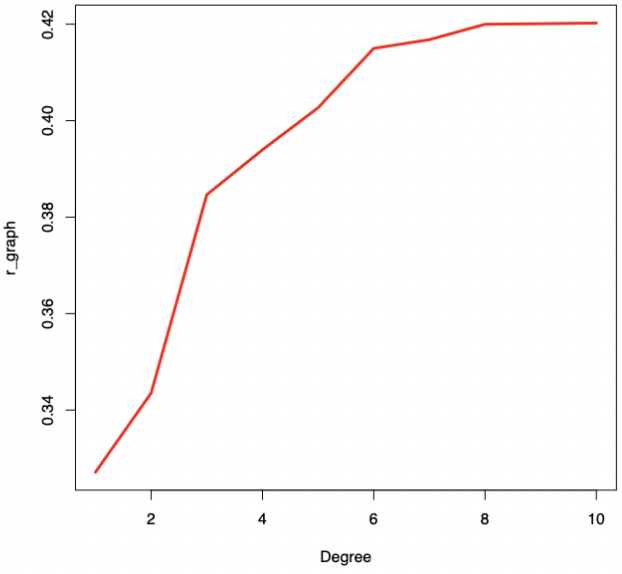
R^2 value = 0.3435334

$y = 72.084 + 0.7396*x - 0.04*x^2$



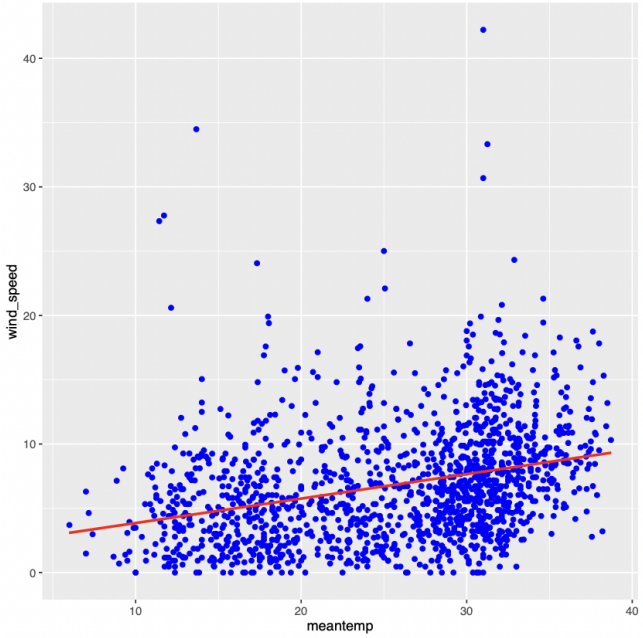
R^2 value = 0.3846457

$y = 158.835 - 11.907*x + 0.52*x^2 - 0.0079*x^3$



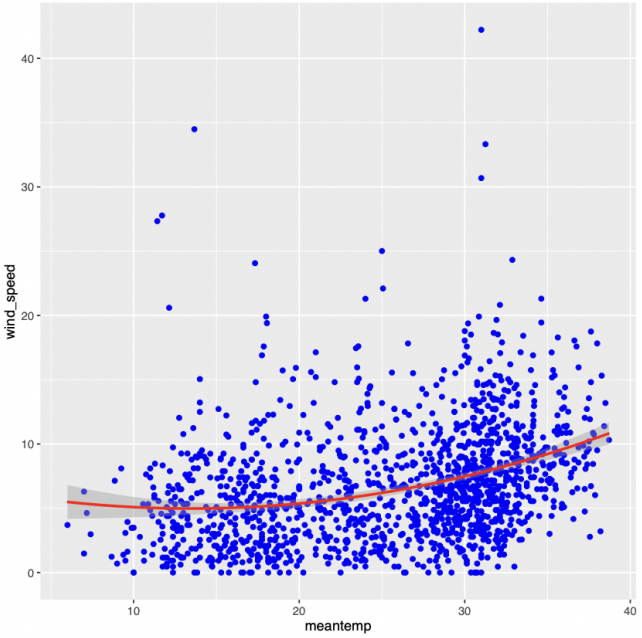
Degree Vs R^2 value

For meantemp(x) Vs Wind_speed(y)



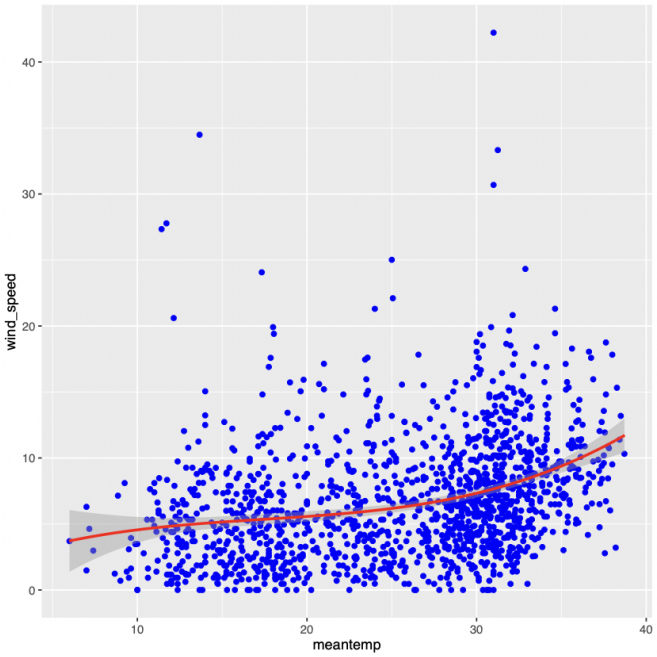
R^2 value = 0.09392246

y = 1.951+ 0.19*x



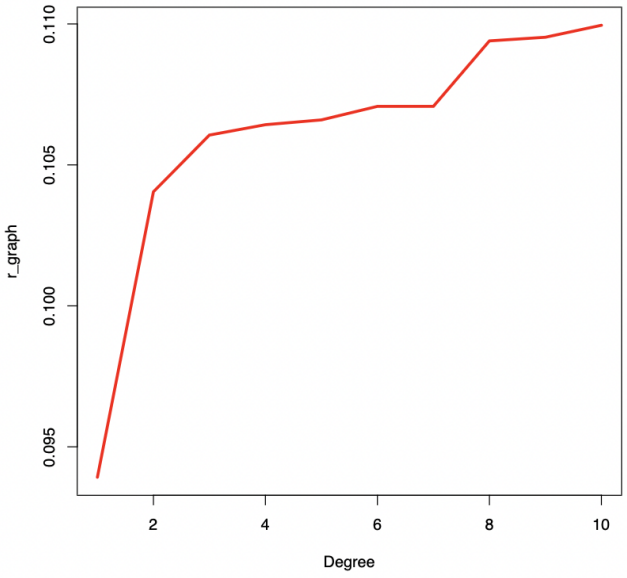
R^2 value = 0.1040446

$y = 6.645 - 0.246*x - 0.009*x^2$



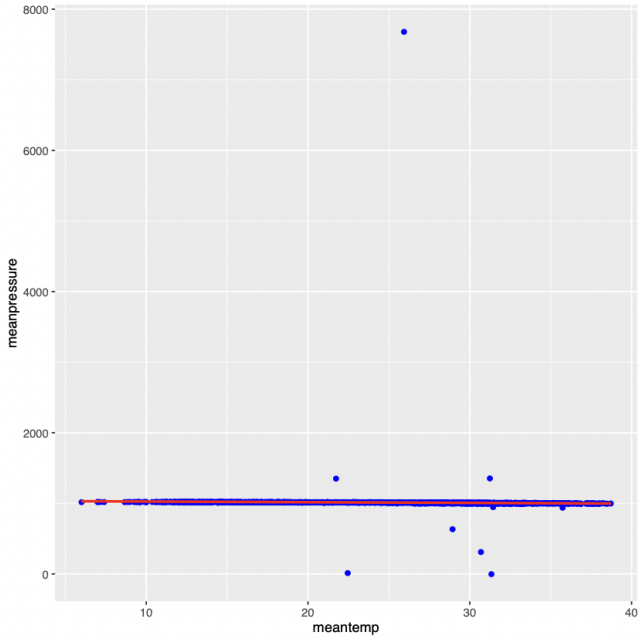
R^2 value = 0.1060581

$y = 1.42 + 0.51*x - 0.02*x^2 + 0.0004*x^3$



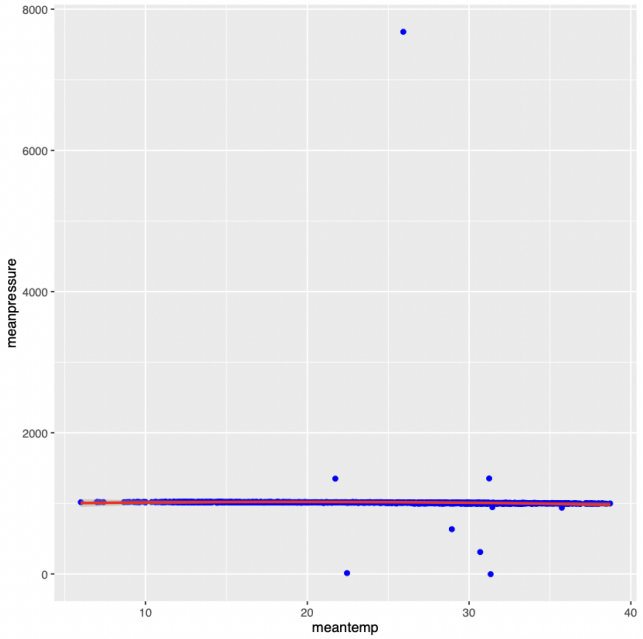
Degree Vs R^2 value

For meantemp(x) Vs meanpressure(y)



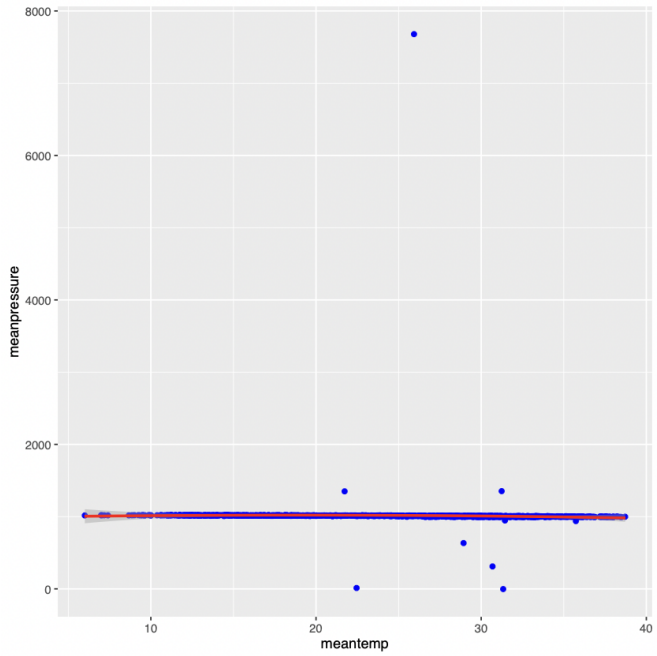
R^2 value = 0.001506851

y = 1035.379 - 0.95*x



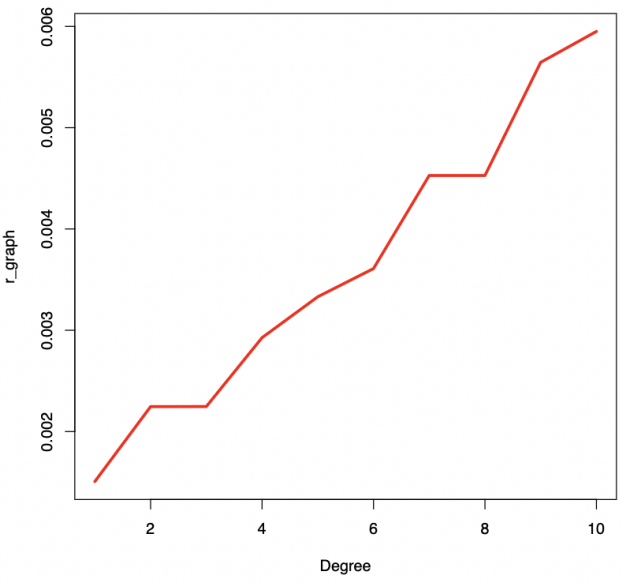
R^2 value = 0.00224618

y = 985.26 + 3.71*x - 0.097*x^2



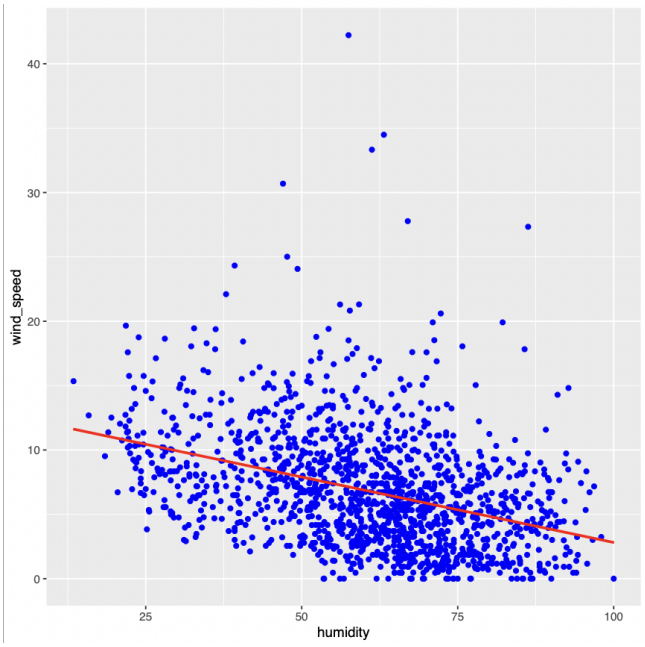
R^2 value = 0.002246901

y = 989.16 +3.14*x - 0.07*x^2 - 0.0003*x^3



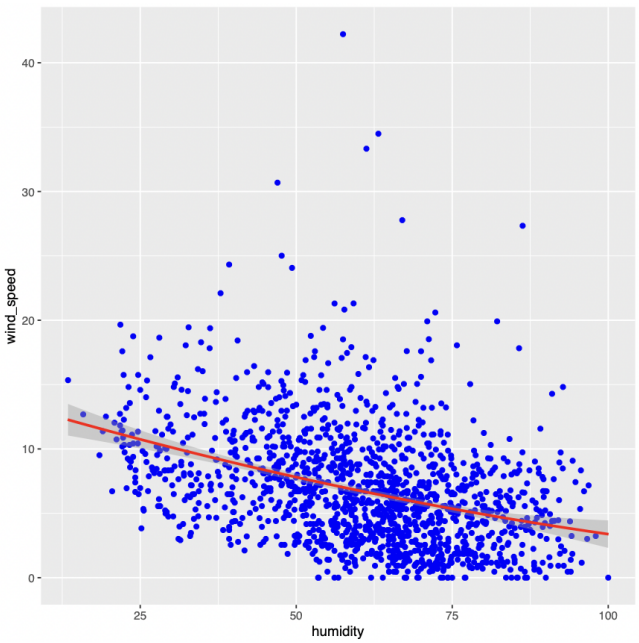
Degree Vs R^2 value

For Humidity(x) Vs Wind_speed(y)



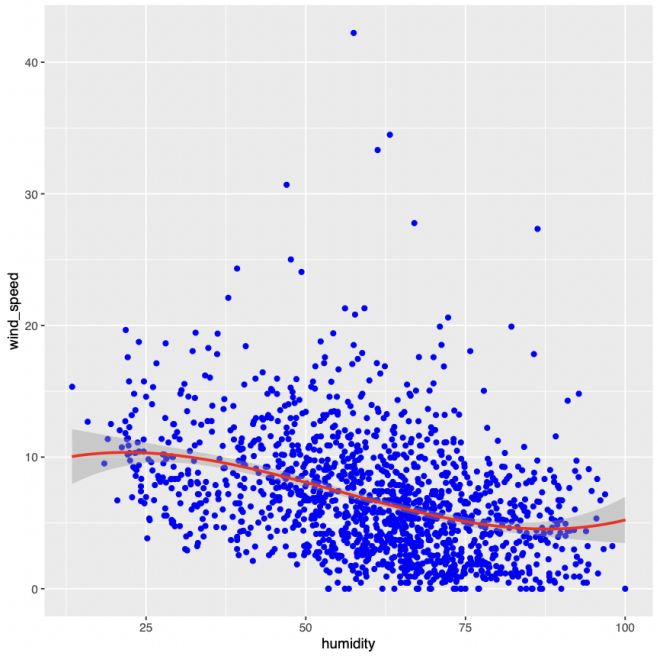
R^2 value =0.1398548

y = 12.984 - 0.101726*x

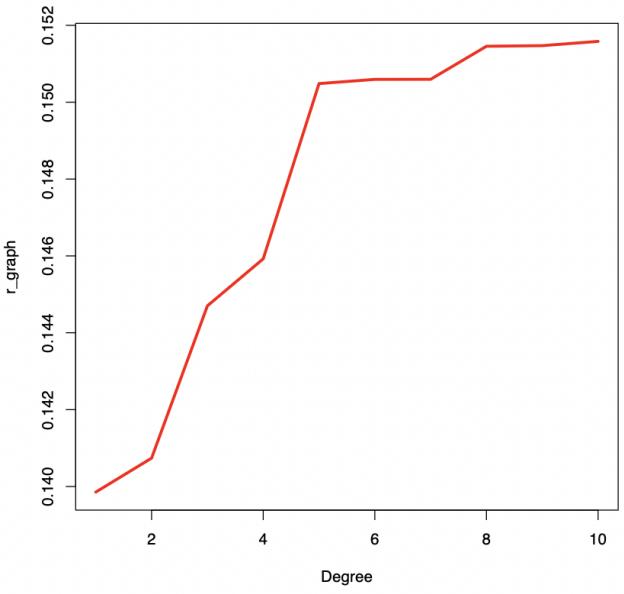


R^2 value = 0.1407384

y = 14.163 - 0.146*x +0.00038*x^2



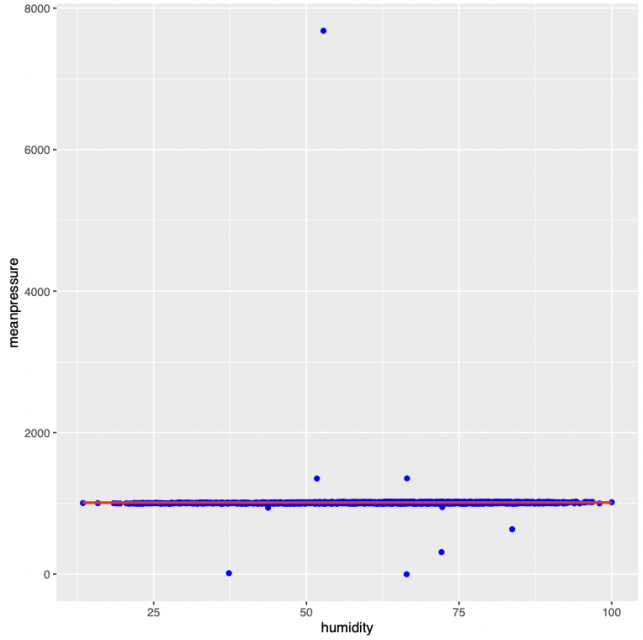
R^2 value = 0.1447004



Degree Vs R^2 value

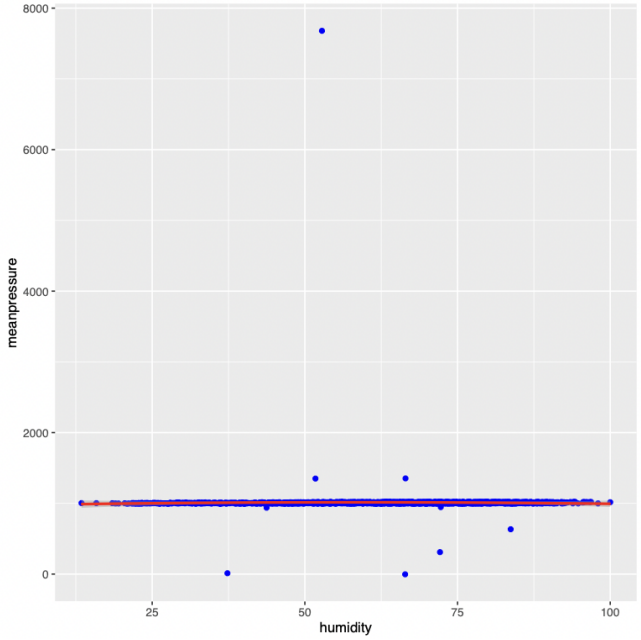
y = 7.992 + 0.235*x - 0.0067*x^2 + 0.00004*x^3

For Humidity(x) Vs meanpressure(y)



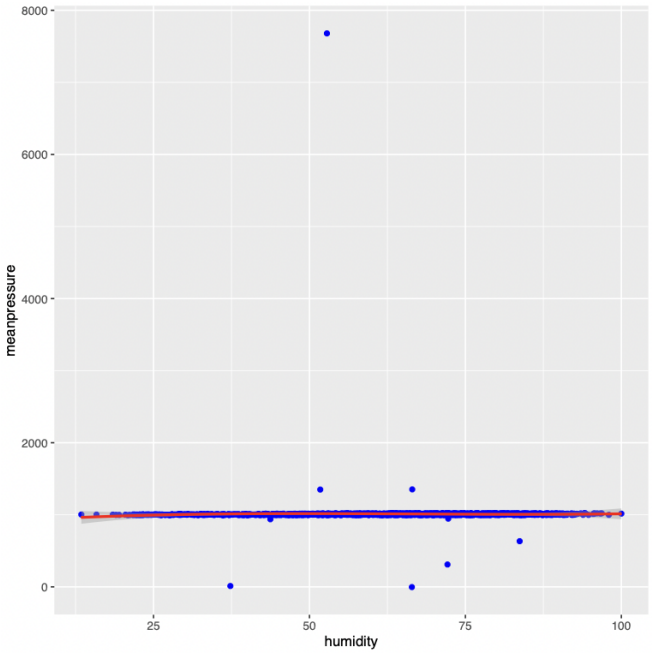
R^2 value = 0.000003006

y = 1009.972+ 0.018*x

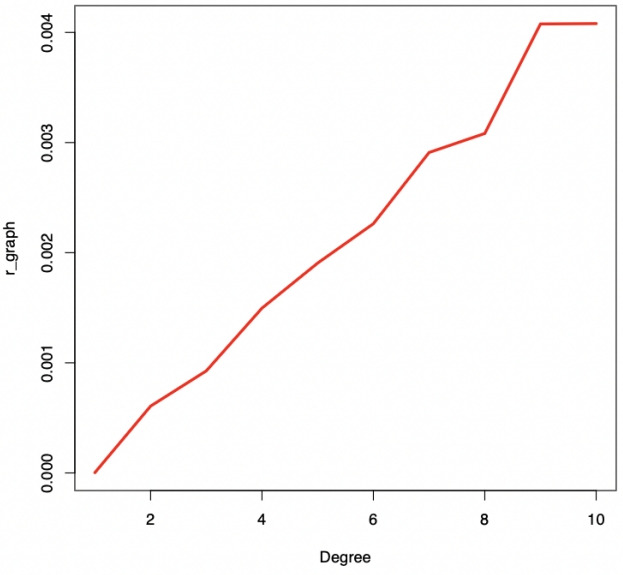


R^2 value = 0.0006066

$y = 971.46 + 1.474*x - 0.01*x^2$



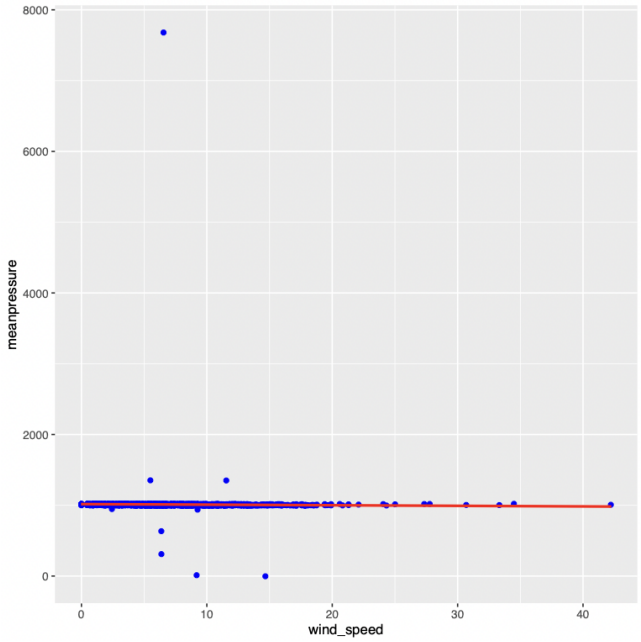
R^2 value = 0.00092650



Degree Vs R^2 value

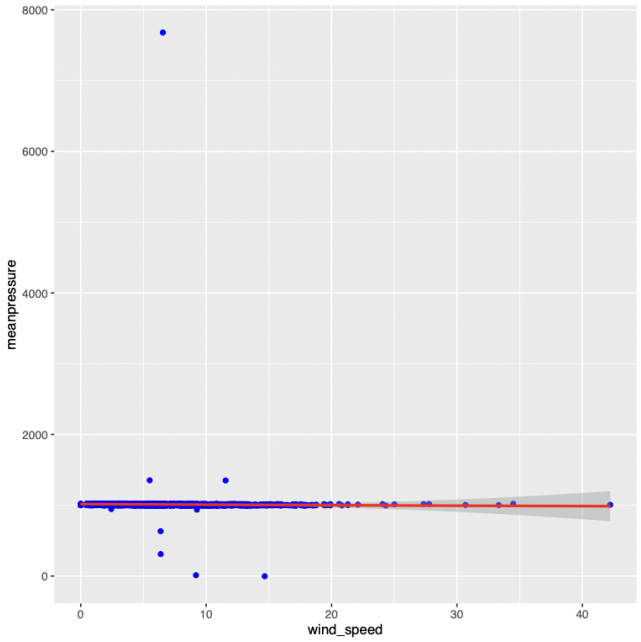
$y = 902.18 + 5.76*x - 0.092*x^2 + 0.00045*x^2$

For wind_speed(x) Vs meanpressure(y)



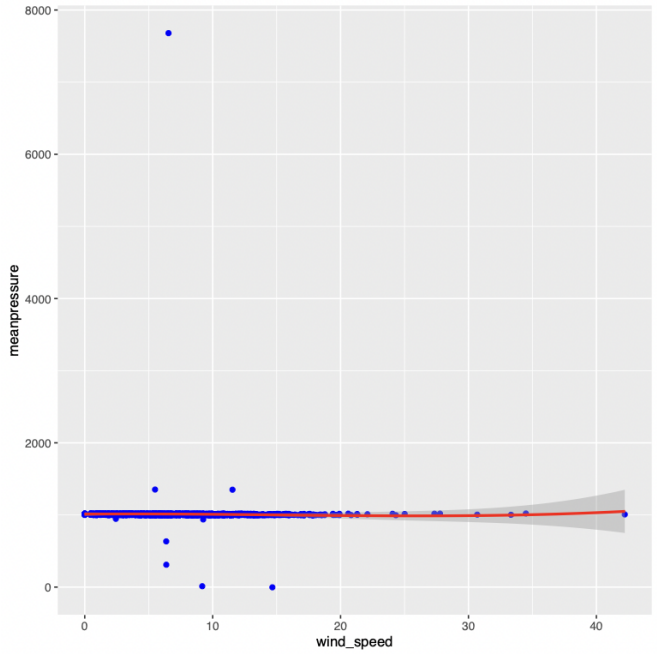
R^2 value = 0.0004272

y = 1016.6 - 0.8166*x

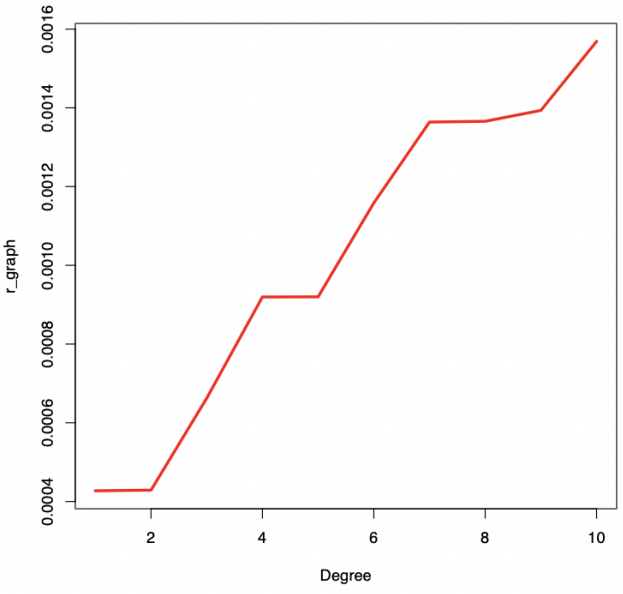


R^2 value = 0.0004289

y = 1017.01 - 0.9182*x + 0.005*x^2



R^2 value = 0.000662



Degree Vs R^2 value

y = 1011.84 +1.374*x - 0.2126*x^2 + 0.004769*x^2

CONCLUSION:

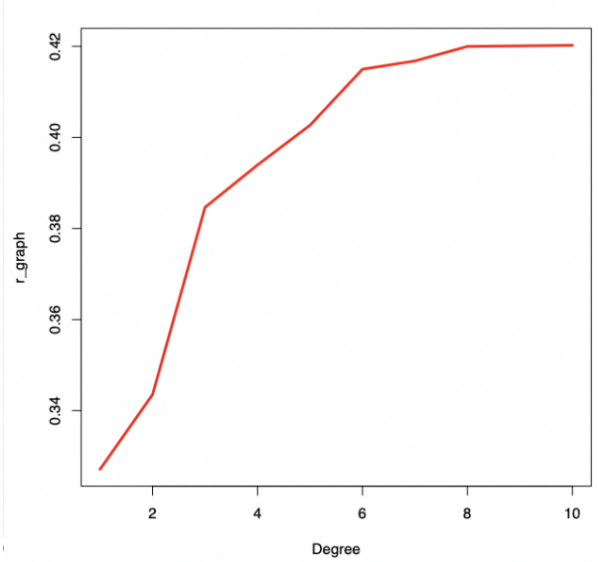
From our experimental results, we can conclude that in all cases, we observe that the R^2 value increases with the increasing degrees.

Let us discuss each case in more detail.

Case 1: MeanTemp vs Humidity

In this case, we can observe from the regression line that, with increasing meantemp, the humidity value decreases.

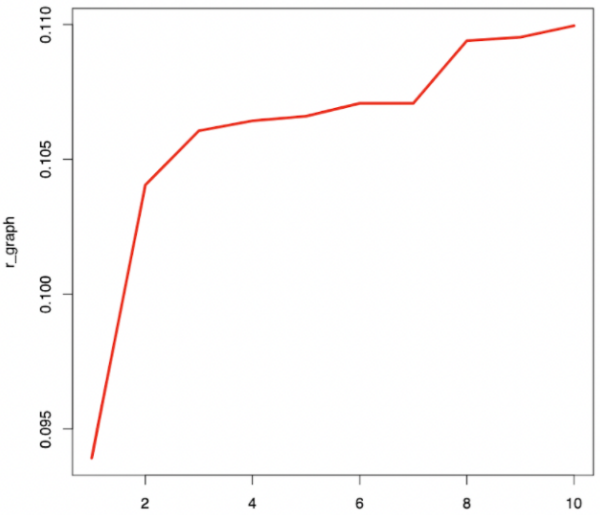
Also, from the R^2 value vs Degree graph, we can conclude that degree 8(non-linear regression line) is the best fit for the data.



Case 2: MeanTemp vs Wind_speed

In this case, we can observe from the regression line that, with increasing meantemp, the windspeed value increases.

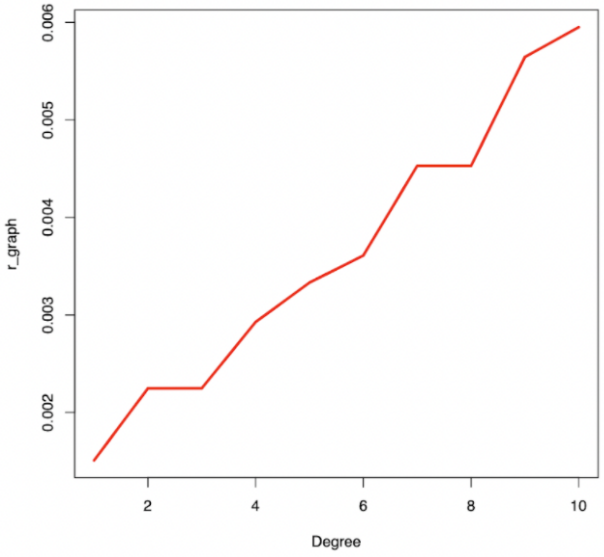
Also, from the R^2 value (which is very low), we can say that our line is not a good fit. But a higher degree gave a higher R^2 value.



Case 3: MeanTemp vs MeanPressure

In this case, we can observe from the regression line that, with increasing meantemp, the humidity value decreases.

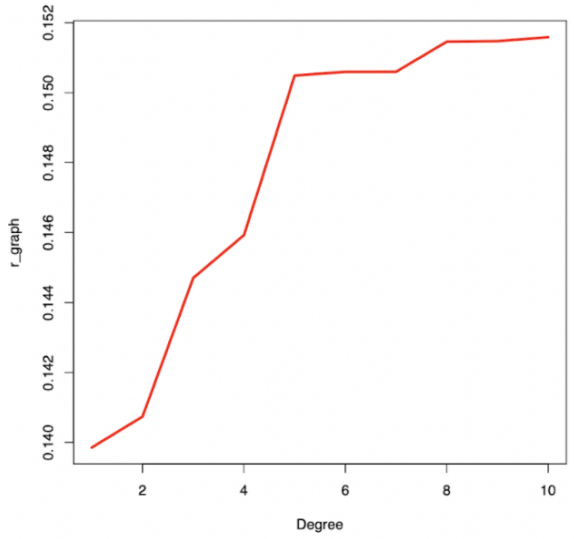
Also, from the R^2 value(which is very low), we conclude that the model is not a good fit.



Case 4: Humidity vs Wind_speed

In this case, we can observe from the regression line that, with increasing meantemp, the humidity value decreases.

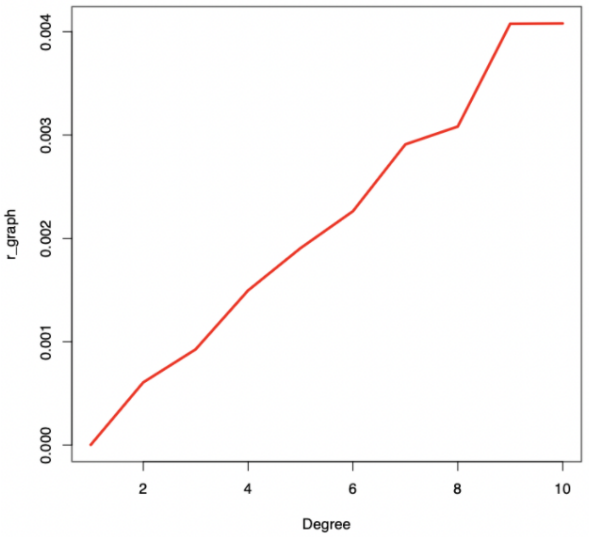
Also, from the R^2 value vs Degree graph, we can conclude that degree 8(non-linear regression line) is the best fit for the data.



Case 5: Humidity vs MeanPressure

In this case, we can observe from the regression line that, with increasing meantemp, the humidity value increases.

Also, from the R^2 value vs Degree graph, we can conclude that degree 9(non-linear regression line) is the best fit for the data. But the R^2 value is very low.



Case 6: Wind_speed vs MeanPressure

In this case, we can observe from the regression line that, with increasing meantemp, the humidity value increases. (But decreases for Linear)

Also, from the R^2 value (which is very low), we can say that our line is not a good fit. But a higher degree gave a higher R^2 value