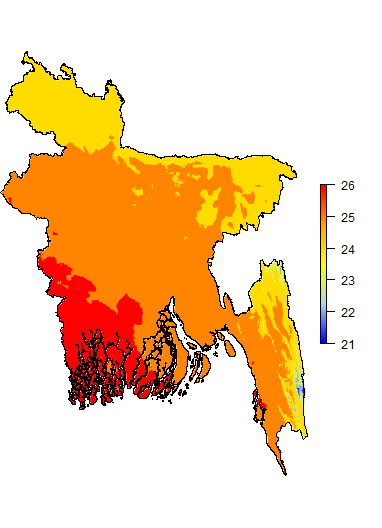


****

**Temperature Prediction of Bangladesh Using Linear Regression**



**Report on**

**Temperature Prediction of Bangladesh Using**

**Linear Regression**

**Course: CSE 829**

| **Submitted by**  **Md. Mehedi Hasan (BSSE 1037)**  **Fahim Ahmed (BSSE 1039)**  **Sudip Kumar Sah (BSSE 1040)**  **S.M. Shahriyar (BSSE 1044)** | **Supervised by**  **Dr. B M Mainul Hossain   Professor**  **IIT, University of Dhaka** |
| --- | --- |



**Institute of Information Technology**

**University of Dhaka**

**Date: 09/11/2022**

**Table of Contents**

[**Abstract**](#_bwshsserb1fr) **1**

[**Data Source**](#_2ajyap2wpcps) **2**

[**Methodology:**](#_px0iqkr8mttv) **3**

[Data preprocessing:](#_pfccwb2thlw2) 3

[**Assumptions of Linear Regression**](#_fsfuqxp3vxcg) **6**

[Check linear relationship among feature and target variables](#_2n0c42om418z) 6

[Normality of errors](#_t8pv9thh7u9p) 7

[No heteroscedasticity (common variance)](#_warzy532w2gc) 8

[No autocorrelation](#_1ob65qnjcz6b) 9

[No multicollinearity](#_11eqfged7d08) 10

[Selected Features:](#_638wr7beb619) 12

[**Transformations**](#_uwuec6v297jd) **13**

[Result without any transformation:](#_e9czkb203wui) 13

[Log transformation on “CO2\_world” feature:](#_hzg86nbmq5bn) 14

[Log transformation on “Solar\_Radiation” feature:](#_i4zmjcd7s5ai) 15

[Log transformation on “Rainfall” feature:](#_wdf8uiez7wgk) 16

[Exponential transformation on “Rainfall” feature:](#_sdhyu1gkgkd5) 17

[Square-root transformation on “Rainfall” feature:](#_g7ojawm7ynw2) 18

[Cube-root transformation on “Rainfall” feature:](#_iuejajm9nf7e) 19

[Inverse(1/x) transformation on “Rainfall” feature:](#_ynvu4yazo0o) 20

[Inverse(1x3) transformation on “Rainfall” feature:](#_j3ozieqhbdtu) 21

[**Results and Discussion:**](#_bvbw0ko6uq5d) **22**

[**Final Equation**](#_ufxefkr3bffi) **23**

[**Conclusion and Future Work**](#_agj13tq4yjt4) **23**

# 

# Abstract

As a result of global warming, there is a significant change in our climate. There are several reasons for global warming and as a result of global warming there are changes in different aspects of nature. In this finding, we considered CO2 emission as a potential reason for global warming, we considered solar radiation and rainfall as potential reasons for climate change and we considered temperature as a candidate of climate change indicator. In this case, the scope can be very large. So, we considered the temperature in Dhaka - which we wanted to predict. For predicting the temperature (target variable) of Bangladesh from feature variables - Bangladesh CO2 emission, global CO2 emission, solar radiation and rainfall data of Bangladesh.

# Data Source

For this prediction, we did not collect data by ourselves. Therefore, we relied on secondary data sources. But all secondary sources are not reliable. So we relied on websites of the Bangladesh Government and International data source which is trusted and used by : Harvard, Stanford university; The Guardian, The New York Times, CNN, BBC and other news media.

**Temperature and solar radiation dataset:**

<http://barcapps.gov.bd/climate/>

**CO2 emission dataset:**

<https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions>

**Rainfall Data:**

<https://oasishub.co/dataset/bangladesh-historical-daily-rainfall-record-1948-2014-bangladesh-meteorological-department>





Figure: Data Sources

# Methodology:

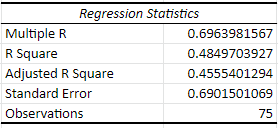
## 

## Data preprocessing:

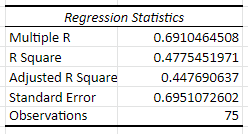
We did not find a ready made dataset for our analysis. We had to merge some datasets from different sources to calculate our statistical analysis. We considered the following steps for our data pre-processing.

1. For rainfall data, we did not consider the following months - December, January, February, March - as there was little to no rain data in these months.
2. We extracted the solar radiation data and temperature data of Bangladesh from a large unstructured dataset.
3. We had daily rainfall data, so we calculated the **mean rainfall value** for each month.
4. For each month, we considered rainfall and temperature data and calculated the r-squared value.

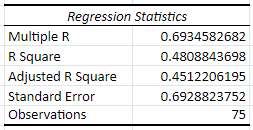
April:



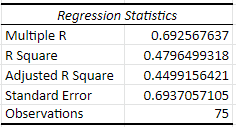
May:



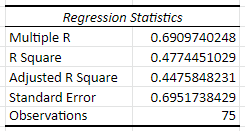
June:



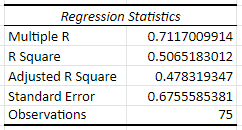
July:



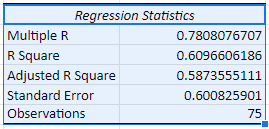
August:



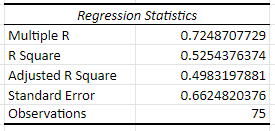
September:



October:



November:



From this analysis report we found that, after comparing to other months we found that, there is a significant difference in R2 value on the month October. Therefore, for feature selection, we considered the rainfall and temperature value of the month October.

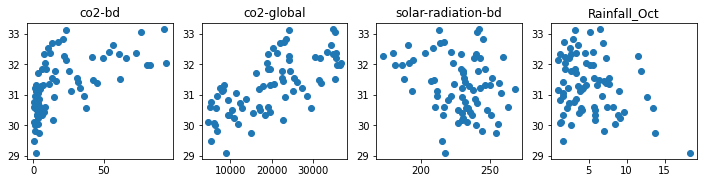
## Assumptions of Linear Regression

Here we are going to cover all six assumptions that we need to check when doing a linear regression.

### Check linear relationship among feature and target variables

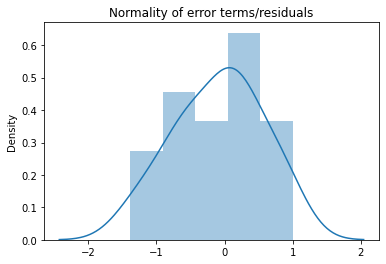
As it is logical and crucial to check after all, if there aren’t linear relationships why should we use a linear model as there are plenty of non-linear models. Basically, it shows the relationship between independent and dependent variables.

Let’s go for a scatter plot which is simple and good enough for checking linearity among variables.

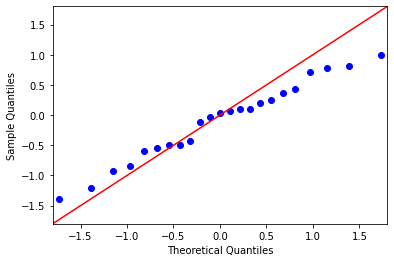


As we can see global CO2 emission has linear relations but among others there are no more linear relationships.

### Normality of errors



**Histogram**

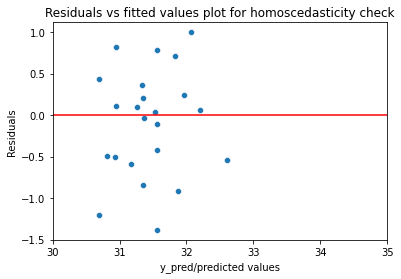


**QQ Plot**

Error means residual which is the difference between predicted and actual values. It should follow a normal distribution. Easy to test using histogram or QQ plot, although difficult to read but more precise to interpret.

### No heteroscedasticity (common variance)

Data should have homoscedastic means the residuals are equal across the regression line. While plotting residual vs fitted scaler plot, there shouldn’t be any pattern in error terms.



Now we know what this is, but to test whether heteroscedastic is present or not. We have to perform the **Goldfeld Quandt Test**.

**Null hypothesis:** Error terms are homoscedastic

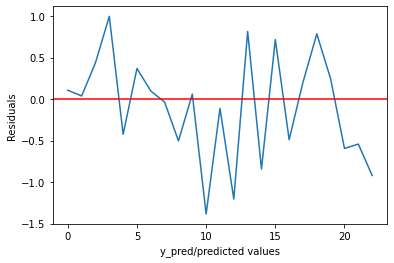
**Alternative hypothesis:** Error terms are heteroscedastic

As a result of p-value and F- statistics from the test, if p-value is more than 0.05 then we can’t reject the null hypothesis that error terms are homoscedastic which is good.

### 

### No autocorrelation

Error terms should not form any pattern and are independent. What autocorrelation means is that current value is dependent on historic values & definite unexplained patterns.

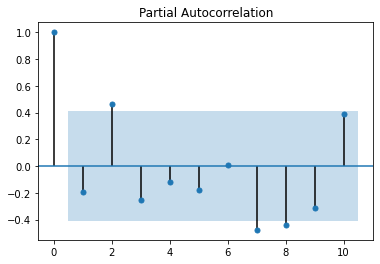


Now we are ensuring the absence of autocorrelation using the jungbox test.

**Null hypothesis:** Autocorrelation is absent

**Alternative hypothesis:** Autocorrelation is present

If p value is less than 0.05 we reject the null hypothesis that error terms are not autocorrelated. We got 0.1017, rejecting null hypothesis and accepting that autocorrelation is present.

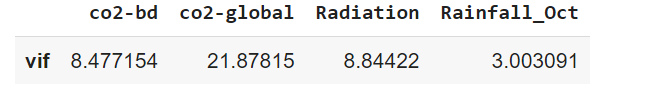


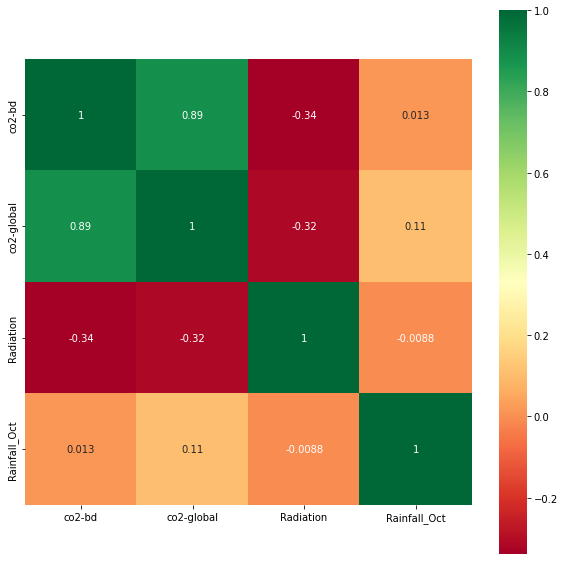
As spikes outside the confidence interval confirms the signs of autocorrelation and seasonality factor.

### 

### No multicollinearity

The extent to which multiple independent variables are correlated. It affects coefficient and p-values, but doesn’t influence the prediction. It clarifies the role of each independent variable and reduces severe multicollinearity.





With variance inflation factor and confusion, we can view the importance of each variable.

## 

## Selected Features:

Considering the collinearity relation, we considered not to use the CO2-bd feature. So, finally the featured we considered are:

1. CO2-world (CO2 emission of the world per year)
2. Solar Radiation (Solar Radiation of the world per year)
3. Rainfall (Rainfall data of October in Dhaka per year)
4. Target feature: Temperature (Temperature data of October in Dhaka per year)

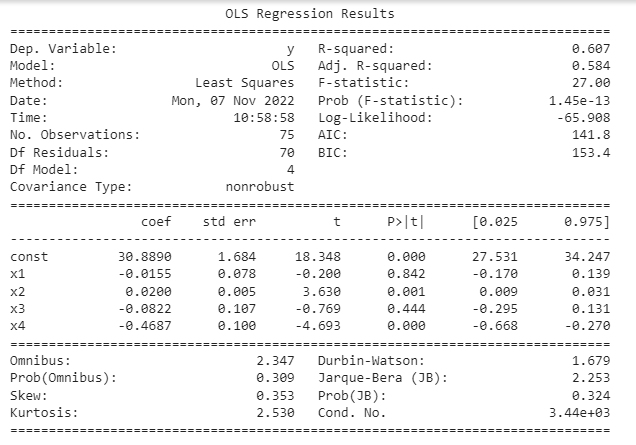
## Transformations

**Log transformation** of the CO2 emission data gave the best result of Global CO2 emission.

Transformation attempt:

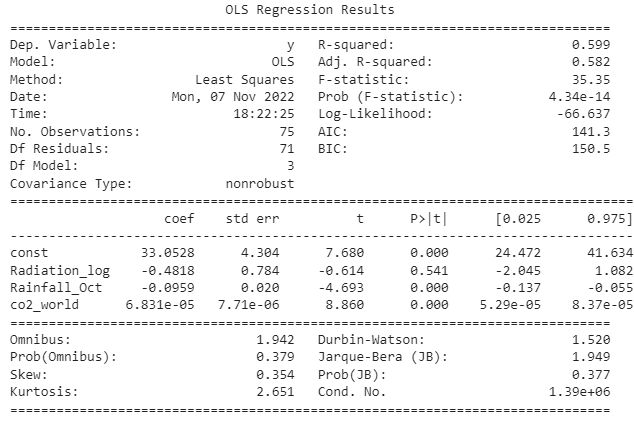
1. Log transformation
2. Square root transformation
3. Inverse transformation
4. Square-root inverse transformation

### **Result without any transformation:**



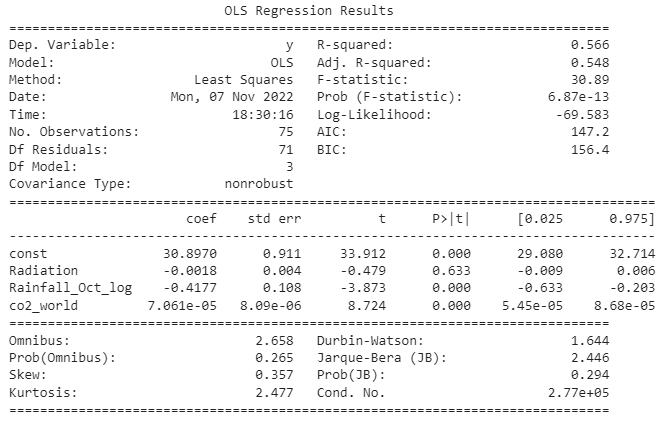
### **Log transformation on “CO2\_world” feature:**

### **Log transformation on “Solar\_Radiation” feature:**



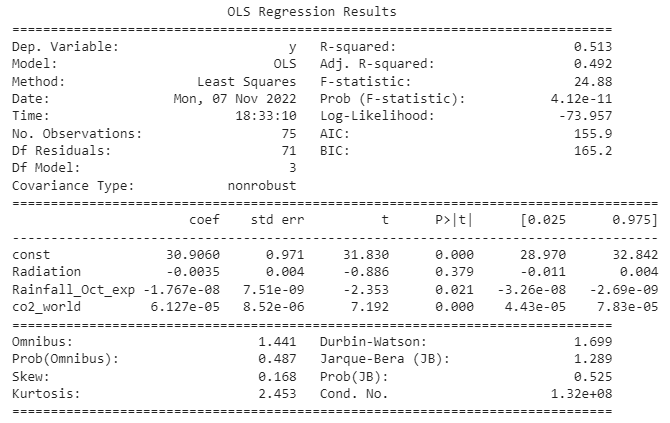
### 

### **Log transformation on “Rainfall” feature:**



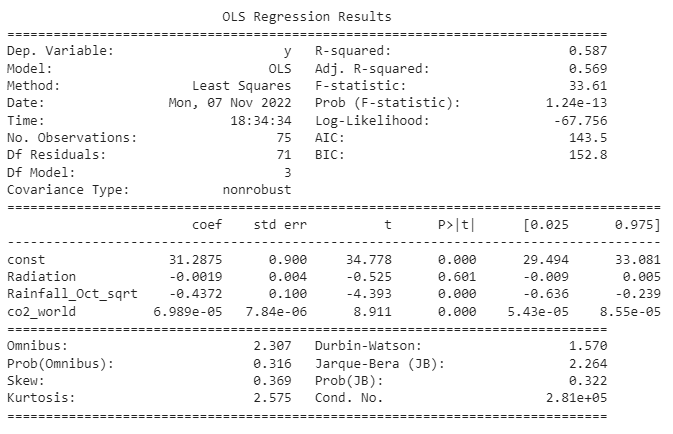
### 

### **Exponential transformation on “Rainfall” feature:**



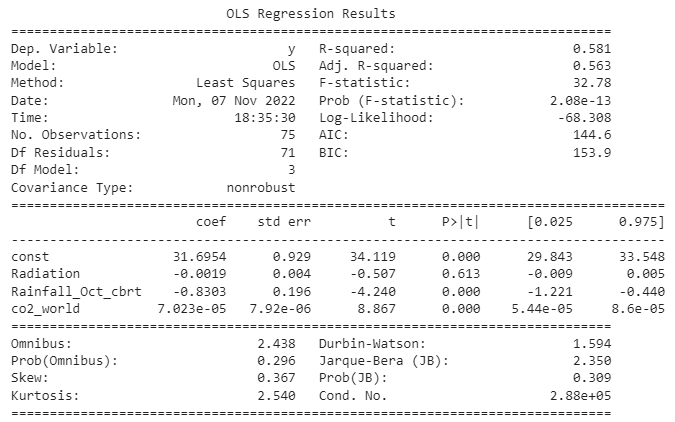
### 

### **Square-root transformation on “Rainfall” feature:**



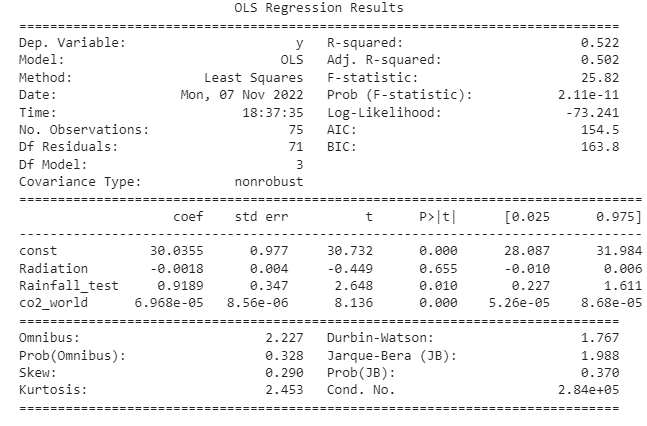
### 

### **Cube-root transformation on “Rainfall” feature:**



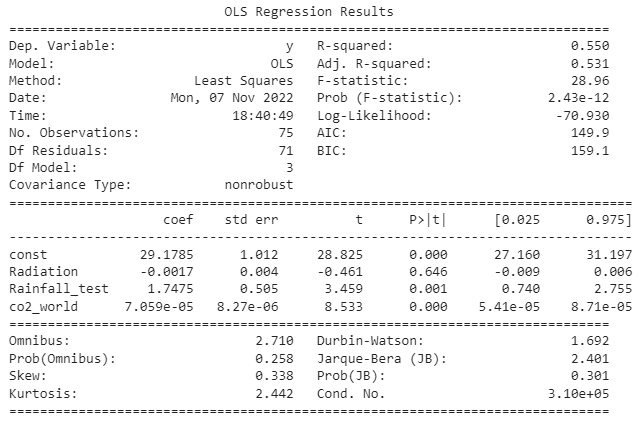
### 

### **Inverse(1/x) transformation on “Rainfall” feature:**



### 

### **Inverse() transformation on “Rainfall” feature:**



# 

# Results and Discussion:

We have finally worked on 3 yearly features : GlobalCOemission, Solar radiation of

Bangladesh, Rainfall in Bangladesh. Prior to any transformation, our the value was 0.607. We have computed various transformations on these features which we provided in detail in the previous section. Among them, log transformation on the feature “GlobalCOemission” has increased the value from 0.607 to 0.619. Log transformation, square-root transformation, cube-root transformation, exponential transformation and inverse transformation on the other features did not increase the value, rather decrease it.

|  | GlobalCOemission | Solar radiation BD | Rainfall BD |
| --- | --- | --- | --- |
| log | 0.619 | 0.599 | 0.566 |
|  | 0.583 | 0.541 | 0.513 |
|  | 0.537 | 0.508 | 0.587 |
|  | 0.551 | 0.569 | 0.522 |
|  | 0.597 | 0.485 | 0.550 |

# Final Equation

y(TEMPERATURE) = 𝛃0 + 𝛃1\*log(CO\_global) + 𝛃2(Solar\_Radiation\_BD) + 𝛃3(Rainfall\_BD)

# 

# Conclusion and Future Work

From this finding, we found that there is a relationship between CO2 emission of the world, rainfall and solar radiation with temperature. After performing linear regression, evaluating the p-value based on residuals less than 0.05 (significance level) rejects the null hypothesis which indicates impasse on the assumptions. We found the best possible R2 value to be 0.619, which is not very high but not too low. The value of R2 could be increased if we considered some other related features which are important factors of temperature change. Future work could also be - collecting more data of the considered features. For example, if we could collect monthly or even daily data of the following features, our results could have improved.