

Time Series Analysis Of Air Pollution



Submitted By:

Bratati Ghosh

Government College Of Engineering &

Ceramic Technology

3rd Year(Information Technology)

Reg no:171130110052 OF 2017-18

Roll No:GCECTB-R17-2014

Acknowledgement

I take this opportunity to express my profound gratitude and deep regards to my faculty (Sarbani Roy) for her exemplary guidance, monitoring and constant encouragement throughout the course of this project. The help and guidance given by her time to time shall carry me a long way in the journey of life on which I am about to embark.

I am grateful for her cooperation during the period of my assignment.

BRATATI GHOSH

(Teacher’s signature)

**Time Series Forecasting of Air pollutants using machine learning**

1.Introduction:

1.1 *Overview*

A **time series** is a sequence of observations taken sequentially in **time**. **Time series** forecasting involves taking models then fit them on historical data then using them to predict future observations.

The Air Pollution Prediction dataset describes different air pollutants at multiple times of a day for two consecutive months(January ,February) and requires a prediction of air pollutant measurements .

Specifically, different air pollutant observations such as PM2.5, PM10, CO , NO , NO2 etc are provided hourly for each days of two months January and February 2020. The objective is to predict air pollutants’ measurements for any hour using previous hours .

This dataset is described the measurements of air pollutants of Jadavpur area hourly for each day of the months January and February 2020 .

*1.2 Motivation*

Time series forecasting is an important area of machine learning. It is important because there are so many prediction problems that involve a time component. However, while the time component adds additional information, it also makes time series problems more difficult to handle compared to many other prediction tasks. Time series data, as the name indicates, differ from other types of data in the sense that the temporal aspect is important. On a positive note, this gives us additional information that can be used when building our machine learning model — that not only the input features contain useful information, but also the changes in input/output over time.

This factors of a time series forecasting problem motivates me and makes me curious to solve such kind of problem .

2. Proposed Approach/methodology :

2.1 *problem Statement:*

The dataset “Air Pollution(Jadavpur)describes the values of different air pollutants like PM2.5 , PM10 , CO , NO ,NO2 etc hourly for each day of two consecutive months January and February. This dataset has 1440 rows and 11 columns. The problem is to forecast the value of the air pollutants by viewing the trend of their values.

*2.2 Work-Flow:*

*2.3 Air Pollution Data Collection:*

* Station data:
* Station name: Jadavpur
* Time Period: Two consecutive months January and February 2020
* Lag intervals: 1 hour
* No of Pollutants: 9 (PM2.5 , PM10 , CO , NO , NO2 , WS , RH , SO2, Ozone)
* Number of total Samples: 1440

*2.4 Machine Learning based Forecasting models:*

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the [linear relationship](https://www.investopedia.com/terms/l/linearrelationship.asp) between the explanatory (independent) variables and response (dependent) variable.

In essence, multiple regression is the extension of ordinary least-squares (OLS) [regression](https://www.investopedia.com/terms/r/regression.asp) that involves more than one explanatory variable.

The Formula for Multiple Linear Regression Is

​*yi*​=*β*0​+*β*1​*xi*1​+*β*2​*xi*2​+...+*βp*​*xip*​+*ϵ*

**where, for***i*=*n***observations**

**:***yi*​=dependent variable

*xi*​=expanatory variables

*β*0​=y-intercept (constant term)

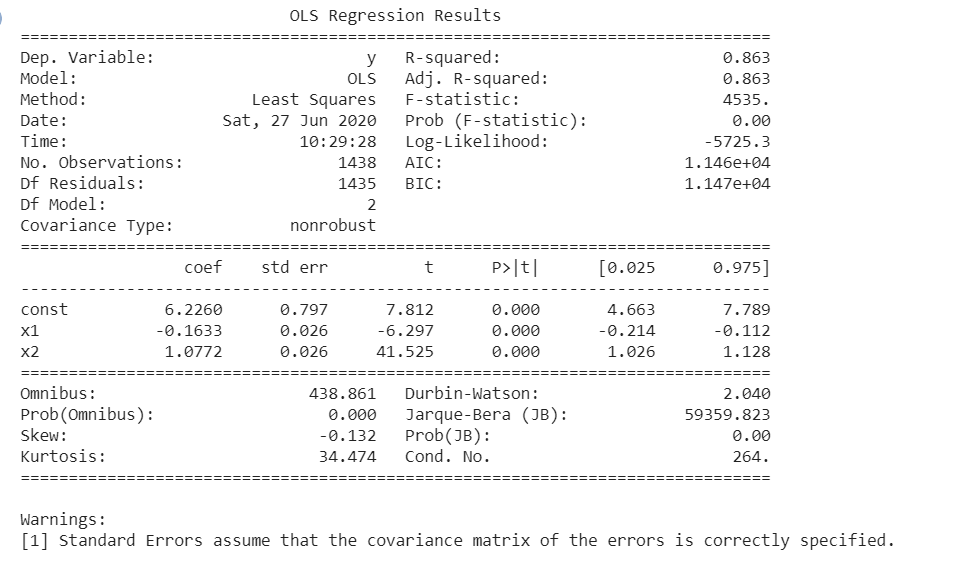
*βp*=slope coefficients for each explanatory variable

*ϵ*=the model’s error term (also known as the residuals)​

* For PM2.5:

By analysing the values of PM2.5 pollutant a trend is observed . It is observed that from 1 a.m to 5 p.m the value is decreasing but from 5 p.m to 12 a.m the value keeps increasing. That means the graph of the trend is parabolic in nature.

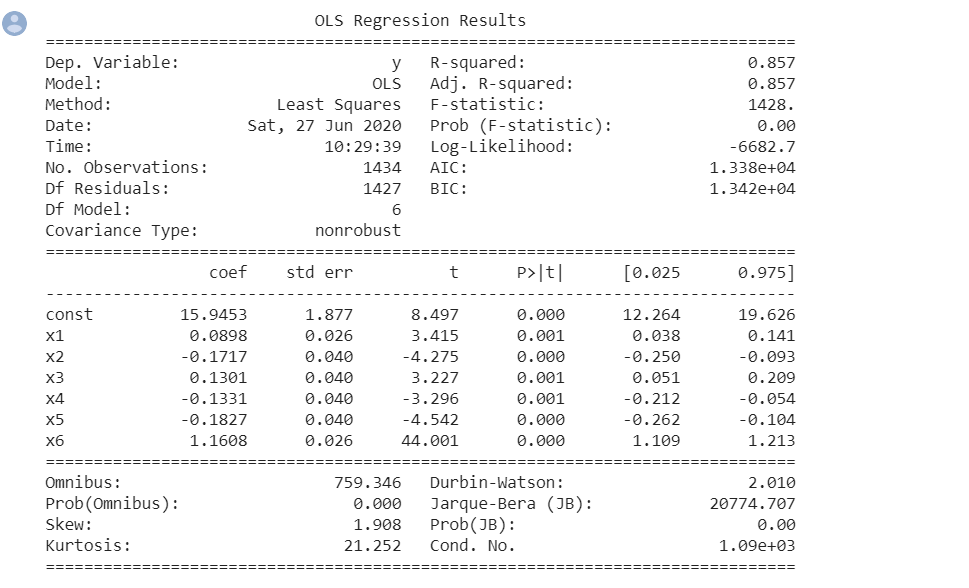
So, here multiple linear regression is used to forecast value of PM2.5.



* For PM10:

By analysing the values of PM10 pollutant a trend is observed . It is observed that in a certain time period of a day the value is decreasing but other time the value keeps increasing. That means the graph of the trend is parabolic in nature.

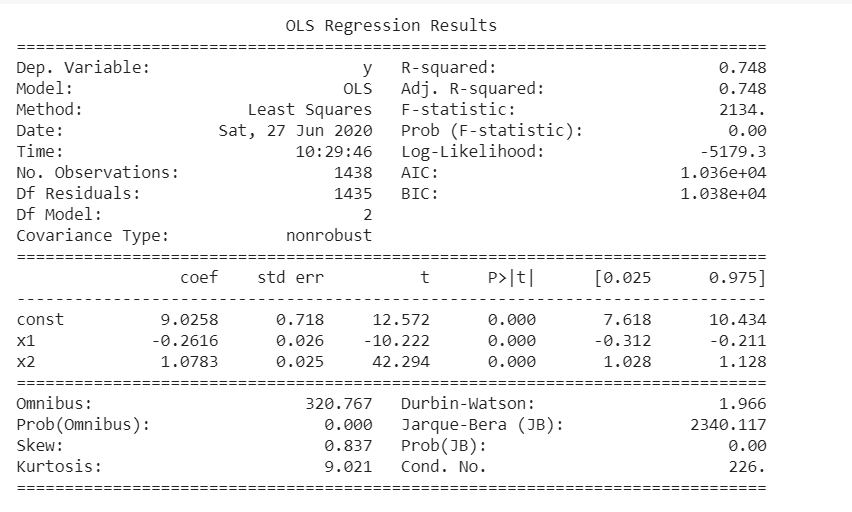
So, here multiple linear regression is used to forecast value of PM10.



* For NO2:

By analysing the values of NO2 pollutant a trend is observed . It is observed that in a certain time period of a day the value is decreasing but other time the value keeps increasing. That means the graph of the trend is parabolic in nature.

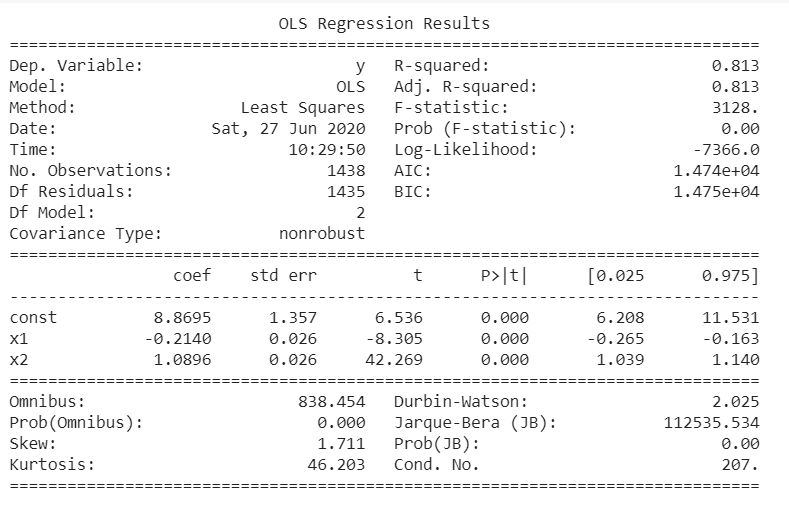
So, here multiple linear regression is used to forecast value of NO2.



* For NO:

By analysing the values of NO pollutant a trend is observed . It is observed that in a certain time period(1 a.m to 1 p.m) of a day the value is decreasing but other time the value keeps increasing. That means the graph of the trend is parabolic in nature.

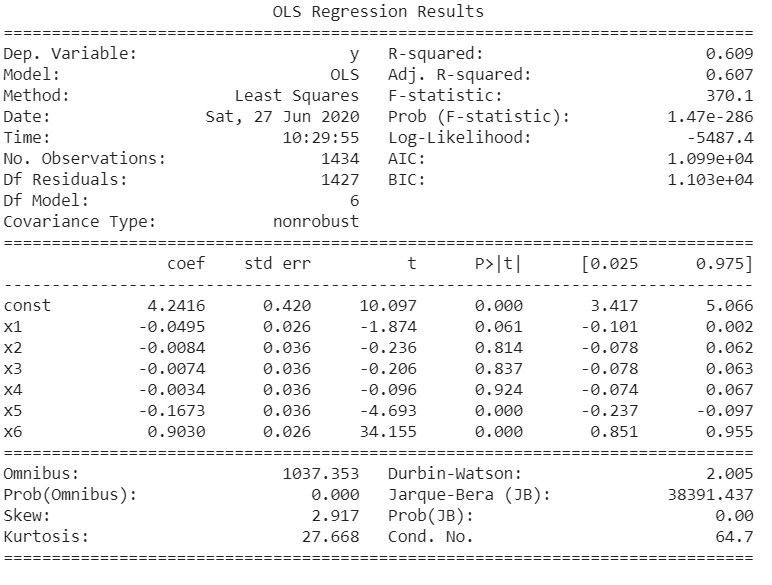
So, here multiple linear regression is used to forecast value of NO.



* For Ozone:

By analysing the values of Ozone pollutant a trend is observed . It is observed that in a certain time period of a day the value is decreasing but other time the value keeps increasing , for two times. That means the graph of the trend is like sine curve.

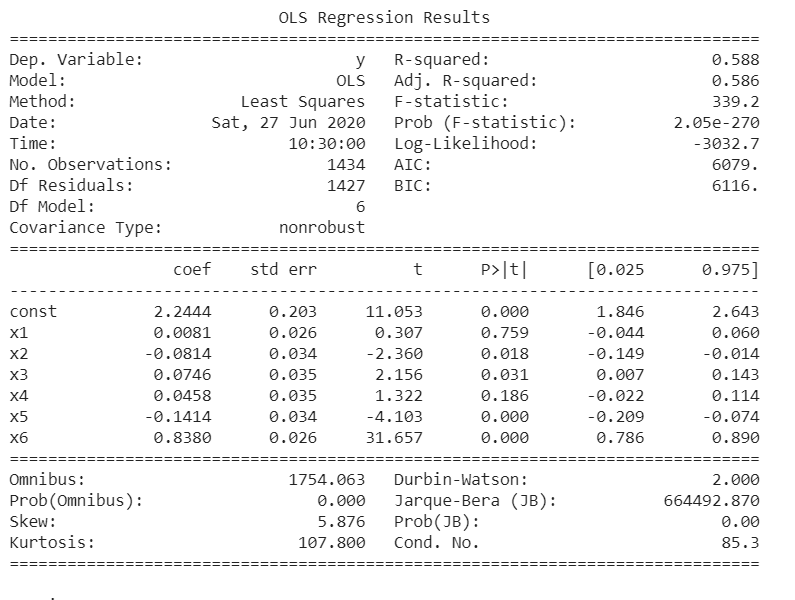
So, here multiple linear regression is used to forecast value of Ozone.



* For SO2:

By analysing the values of SO2 pollutant a trend is observed . It is observed that in a certain time period of a day the value is decreasing but other time the value keeps increasing. That means the graph of the trend is parabolic in nature.

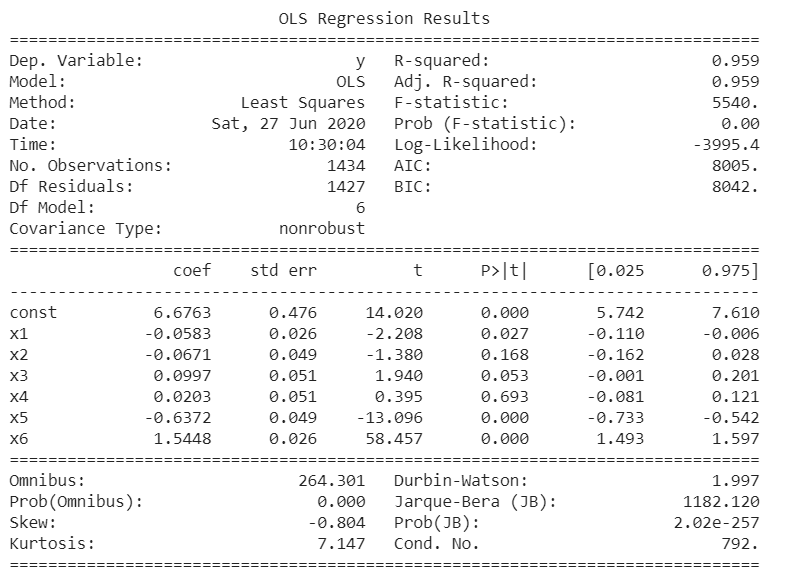
So, here multiple linear regression is used to forecast value of SO2.



* For RH:

By analysing the values of RH pollutant a trend is observed . It is observed that in a certain time period of a day the value is decreasing but other time the value keeps increasing , for many times. That means the graph of the trend is like sine curve.

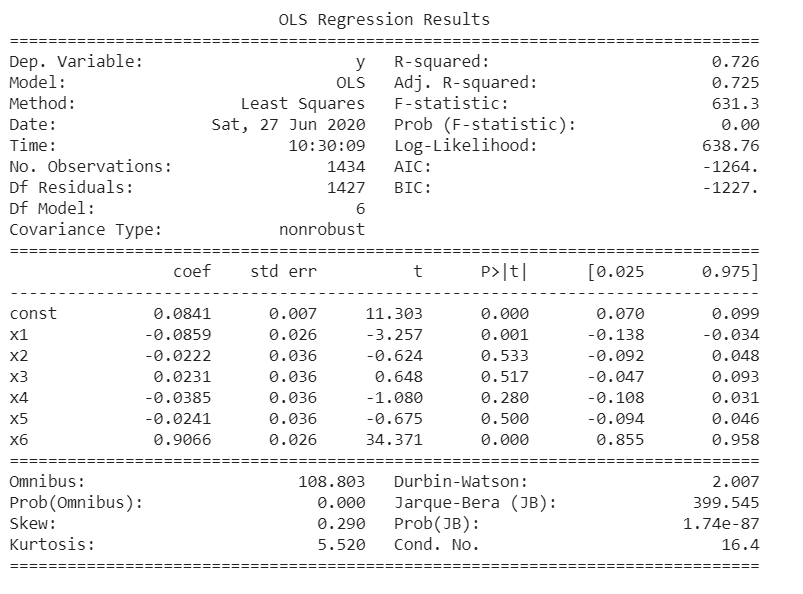
So, here multiple linear regression is used to forecast value of RH.



* For WS:

By analysing the values of WS pollutant a trend is observed . It is observed that in a certain time period of a day the value is decreasing but other time the value keeps increasing , for many times. That means the graph of the trend is like sine curve.

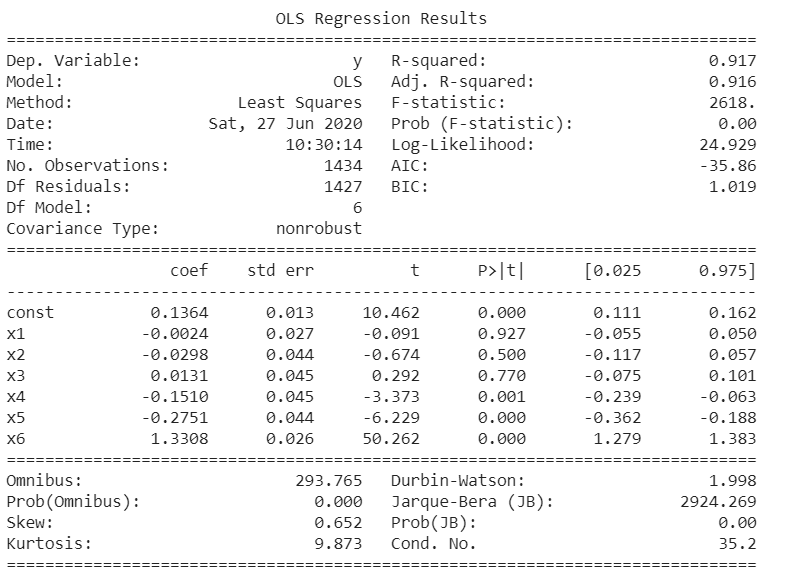
So, here multiple linear regression is used to forecast value of WS.



* For CO:

By analysing the values of CO pollutant a trend is observed . It is observed that in a certain time period of a day the value is decreasing but other time the value keeps increasing. That means the graph of the trend is parabolic in nature.

So, here multiple linear regression is used to forecast value of CO.



3.Results and Analysis:

*3.1 Experimental Setup:*

To solve this forecasting problem I used Python 3.7 as programming language. For this I required a system with –

Processor : 1 gigahertz (GHz) or faster processor or **System** on a Chip (SoC)

RAM : 2 GB for 64-bit

Hard drive space: 32 GB for 64-bit OS

I used some libraries of python to solve this problem. They are –

1. Numpy

2. Pandas

3.Sklearn

4 .Matplotlib

5. Seaborn

*3.2 Results:*

* PM2.5:

R-squared: 0.863

* PM10:

R-squared: 0.857

* NO2:

R-squared: 0.748

* NO:

R-squared: 0.813

* SO2:

R-squared: 0.588

* Ozone:

R-squared: 0.609

* RH:

R-squared: 0.959

* WS:

R-squared: 0.726

* CO:

R-squared: 0.91

4. Conclusion And Future Work:

The regulation of air pollutant levels is rapidly becoming one of the most important tasks. It is important that people know what the level of pollution in their surroundings is and takes a step towards fighting against it.

The results show that machine learning models (multiple linear regression ) can be efficiently used to forecast the value of air and predict the level of PM2.5 in the future.

The proposed system will help common people as well as those in the meteorological department to detect and predict pollution levels and take the necessary action in accordance with that. Also, this will help people establish a data source for small localities which are usually left out in comparison to the large cities.

**Certificate**

This is to certify that Ms. BRATATI GHOSH of GOVT. COLLEGE OF ENGINEERING AND CERAMIC TECHNOLOGY,

registration number: 171130110052 OF 2017-18, has successfully completed a project(Time Series Forecasting of Air pollutants using machine learning) on using MACHINE LEARNING under the guidance of Mrs. Sarbani Roy.

---------------------------------------

[Name of your faculty]

Jadavpur University