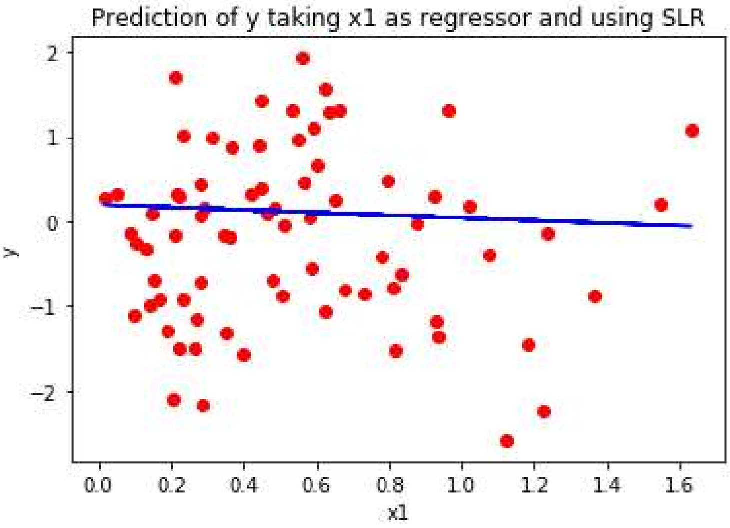
RTSM Assignment 2 (Data Analysis)
Name: - Sumit Kumar Yadav
Roll No: - 18CS30042

Question 1: Fit a best possible regression model to the data set “data1.txt”.

1. First perform the Simple linear Regression, we get the following prediction graphs:
   a. Prediction of y taking x1 as regressor and using SLR

   Equation of line:
   y = -0.1623232x1 + 0.1892645
   RMSE: 1.13265427

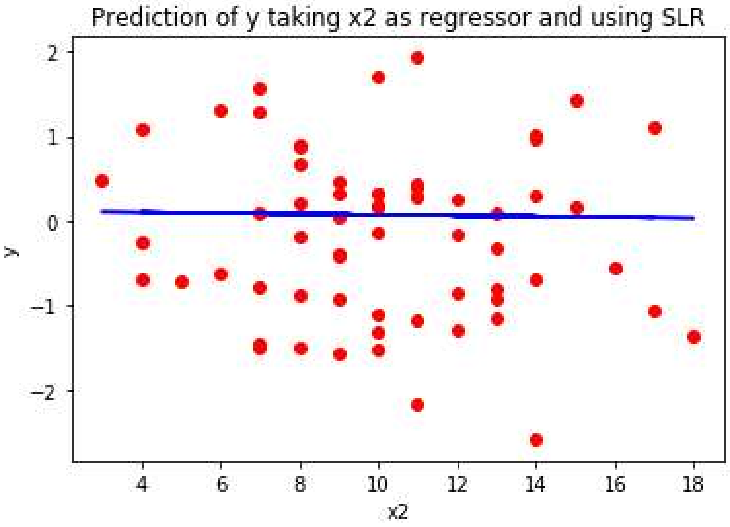


   b. Prediction of y taking x2 as regressor and using SLR

   Equation of line:

   y = -0.0046843x2 + 0.127845

   RMSE: 1.0094543954

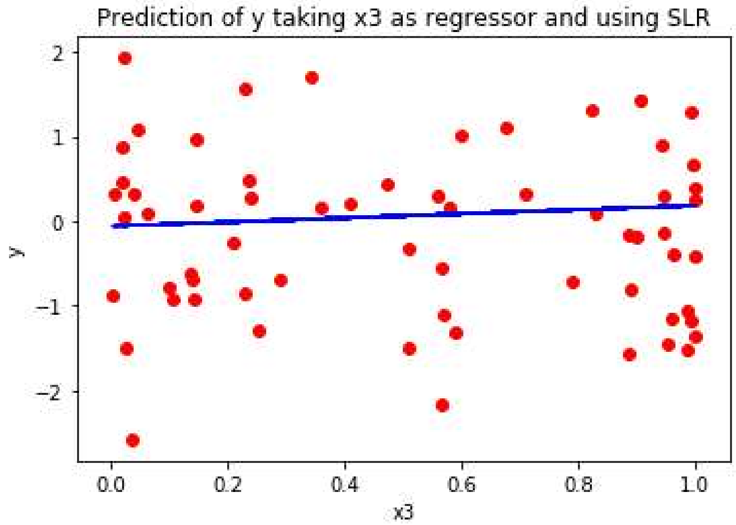


c. Prediction of y taking x3 as regressor and using SLR

Equation of line:

y = 0.2367432x3 - 0.0617643

RMSE: 1.01942333

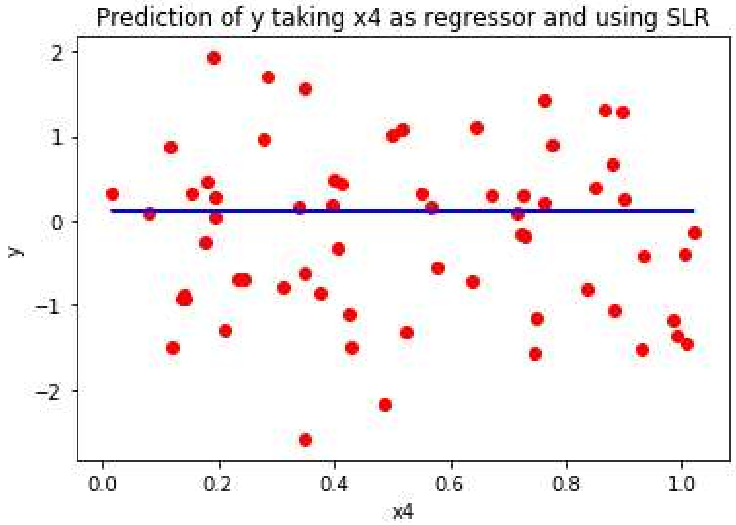


d. Prediction of y taking x4 as regressor and using SLR

Equation of line:

y = 0.14932x4 - 0.018463

RMSE: 1.02267335353



2. Performing a multiple linear regression using all variables, following results were obtained.

Multiple Linear Regression

```
==========================================================
                    R-squared:                   0.0069
                    Adj. R-squared:             -0.0089
                    F-statistic:                 0.4483
                    Prob (F-statistic):            0.761
==========================================================
                      coef          std err          t              P>|t|
------------------------------------------------------------------------------------------
   const            0.0988          0.244          0.405          0.686
   x1              -1.2312          1.852         -0.665          0.507
  x2              -0.0087           0.018         -0.477          0.633
   x3              -2.5098          4.348          0.577          0.564
   x4              3.7675          6.178           0.610          0.543
```

From the above data, we can see that the R^2 is very low and the p value of F-statistics is very high. So, the model is insignificant and fails to explain the observed data.

Now, To remove the multi-collinearity between x3 and x4, here we performing PCA and building a regression model on the first three components, then we obtain the following results.

```
                 ===================================
                     R-squared:                   0.006
                    Adj. R-squared:              -0.006
                    F-statistic:                 0.4818
                    Prob (F-statistic):          0.695
    ====================================================
                    coef          std err          t              P>|t|
   -------------------------------------------------------------------------------------
   const            0.0092          0.062          0.148          0.883
   x1               0.0188          0.044          0.432          0.666
   x2               0.0042          0.062          0.067          0.947
   x3              -0.0711          0.063         -1.120          0.264
```

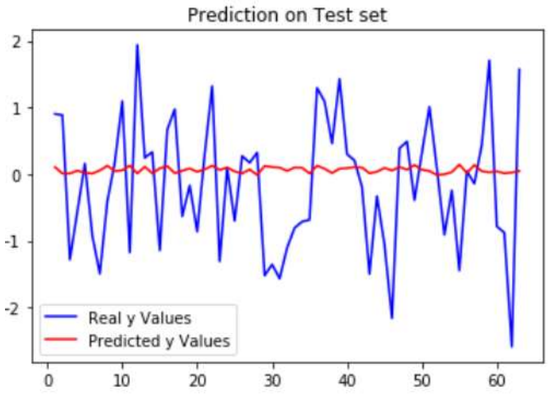
From the R^2 and p value of F-statistics, we can see that the model is still insignificant and fails to explain the observed data.
Therefore we can say that, polynomial regression with variables upto degree 2 also yielded similar results.

<u>Conculsion(Using Backward Elimination to find statistically significant regressors)</u>

1. Taking const,x1,x2,x3 and x4 as regressor:
    a. Since we found p value is very high therefore we remove it from our regression model.
2. Taking const,x1,x2 and x4 as regressor:
    a. Removing x3 affects as the r^2 value increased . Hence we remove constant and from our regression model.

3. Taking x2 and x4 as regressor:
    a. Since p value less than 0.5 we can accept this model.

Optimum MLR Model (using x2 and x4 as regressor):

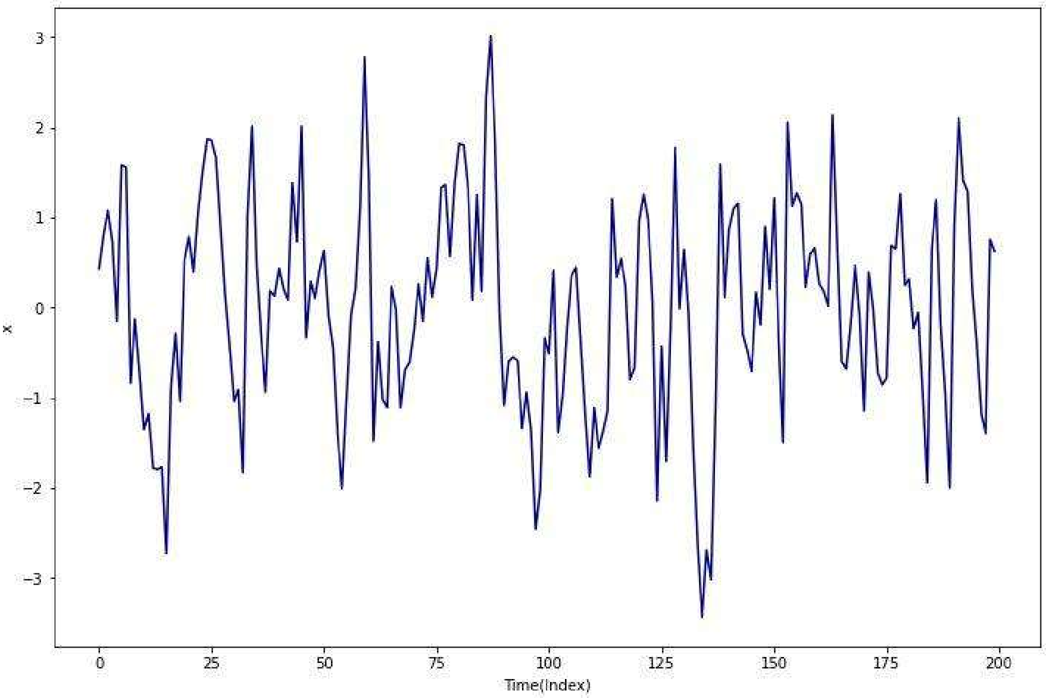


RMSE: 1.01465873

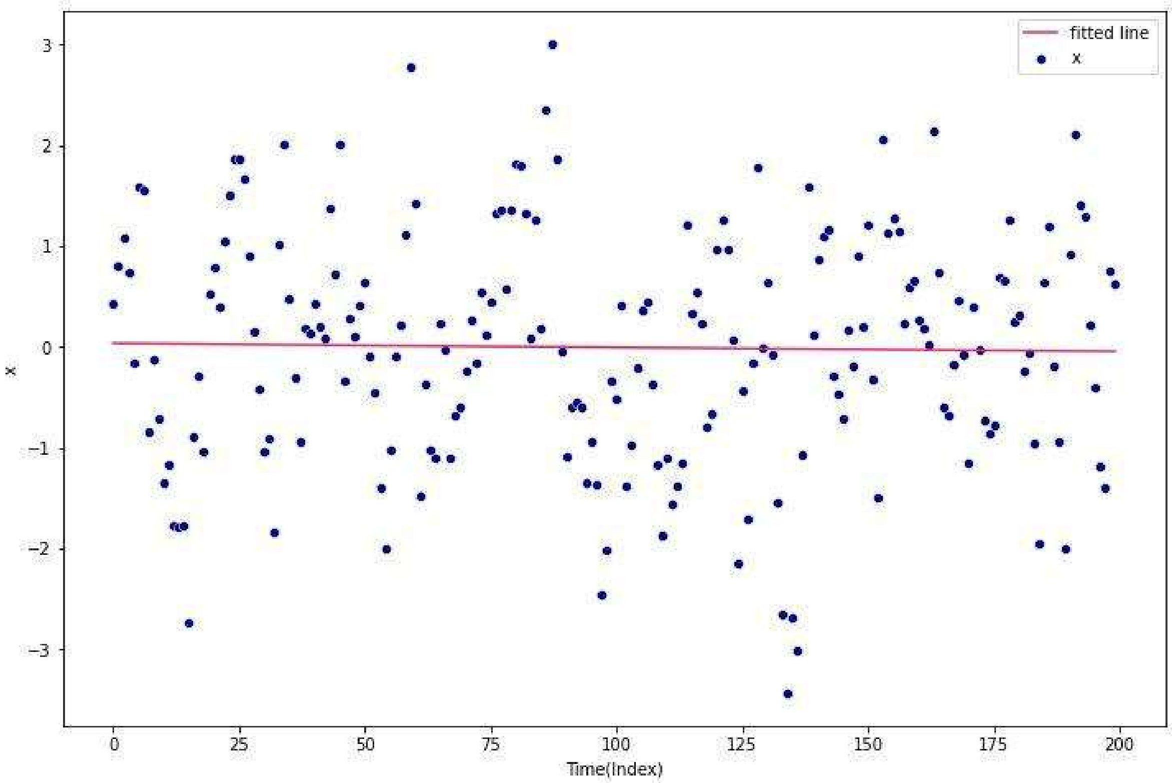
Question 2: Find the order and parameter of the time series for data set "data2.txt".

The Steps followed to find the order and parameters of the time series after importing the text file in python:-

Step 1- Plotting the time series.



Step 2: Now, check for the seasonality.

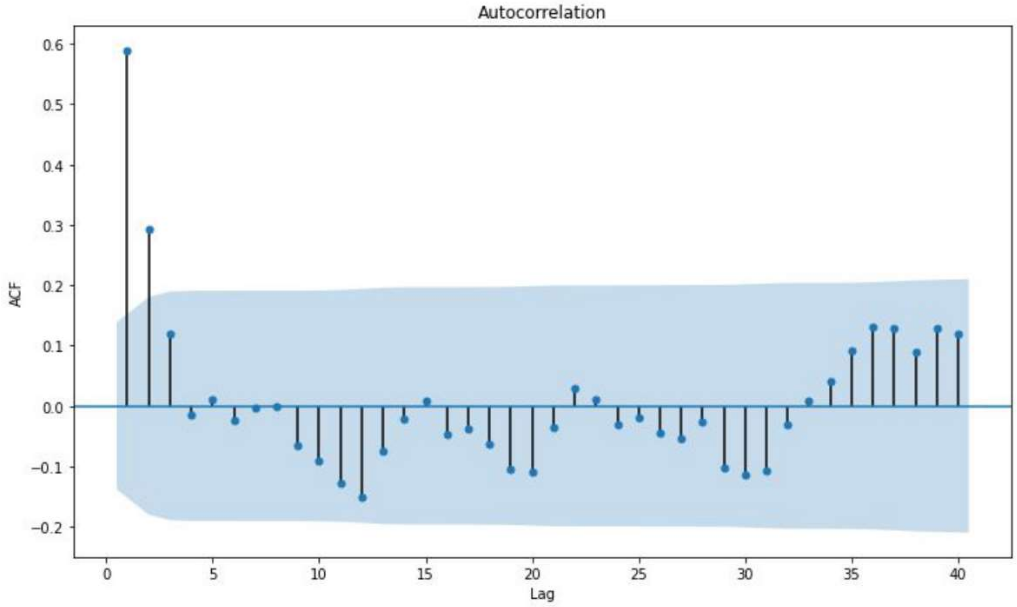


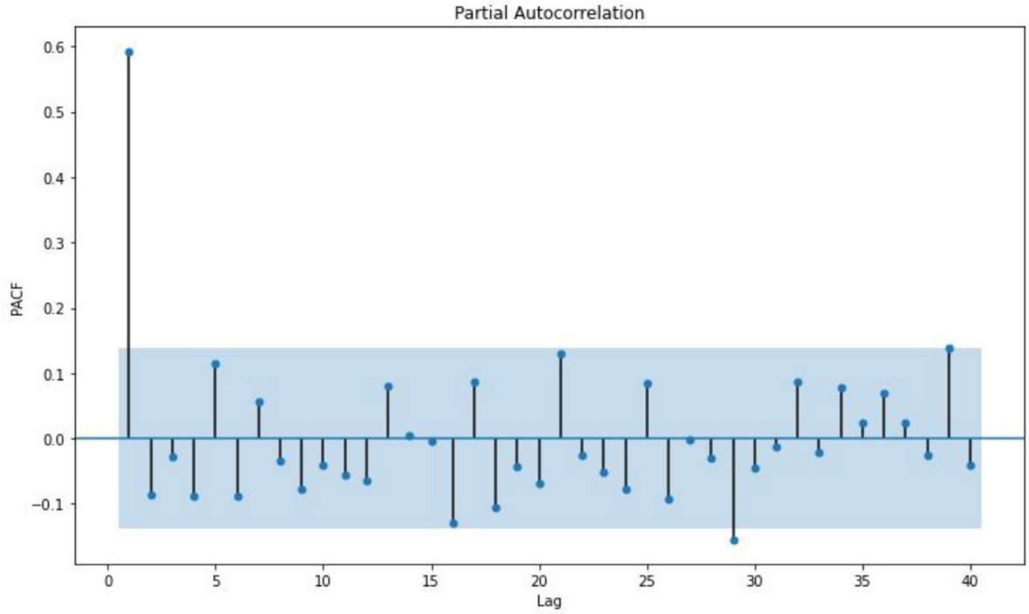
But here due to the lack of deterministic pattern, there exists no seasonality.

Step 3- Then Check whether the given series is stationary or not by using the dickey-fuller test, where the null hypothesis is that the series is not stationary. We find that p-value in the dickey-fuller test is 0 so we can reject the null hypothesis. Therefore, we can say that our given time series model is at least weakly stationary.

Step 4 -Now**,** Plot the Auto Correlation function (ACF) and Partial Auto Correlation Function (PACF) to determine whether the given time series is AR, MA or ARMA.

Clearly after the 1 lags the PACF values are insignificant as they lie in the blue region.
So it’s an AR(1) process.(ie, AR of order 1)





Step 5 – From the auto-covariances, we can obtain the model parameters phi and sigma^2. Since Φ= ρ(1) for AR(1), hence the value of the parameter is equal to the value of ACF at lag 1, which from the graph is approximately 0.58. To calculate the exact value, we can fit the AR(1) model and report the observed coefficient. On fitting the model the value of Φis found to be 0.5831.

We know that ,

gammax(h) = sigma^2*phi^h/(1-phi^2)

Hence our **parameter**is **0.5831**, which is **approximately 0.58**.

To verify this, fitting the given time-series to AR(1) , the following result were obtained:-

```
                         AutoReg Model Results
========================================================

No. Observations:                 200
Log Likelihood                    -271.619
AIC                               -0.097
BIC                               -0.067
========================================================
                  coef            std err           z            P>|z|
-----------------------------------------------------------------------------------------------
ar.L1             0.5831          0.053             11.021       0.000
sigma2            0.8731          0.089             9.890        0.000
```

Therefore the process is AR(1) with phi = 0.5831 and sigma^2 = 0.8731.

Result -Hence the order of the given time series is equal to 1 and the parameter is equal to 0.5831.