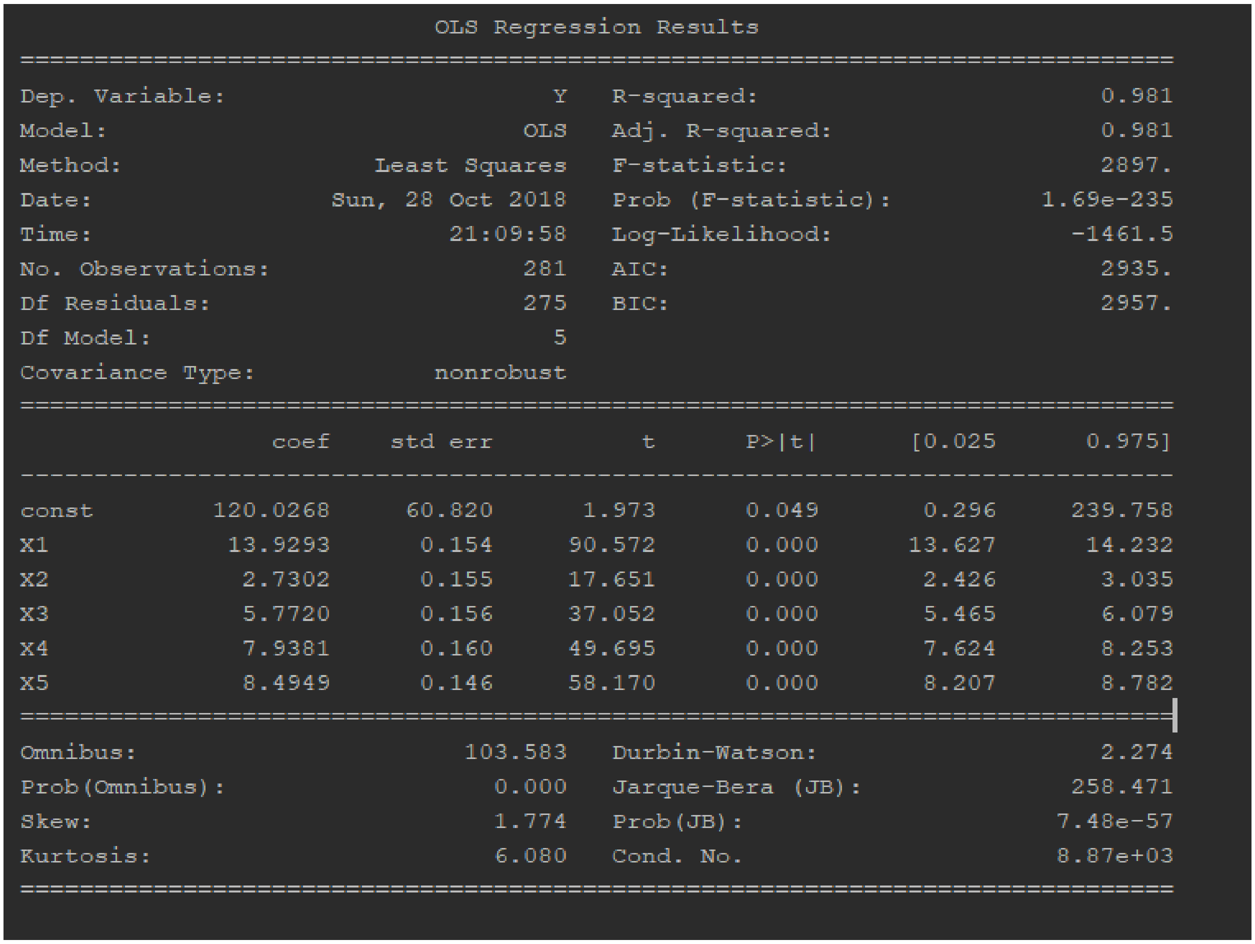
ECE 592-005 IOT Analytics

Project 2 : Regression Task 3

Sreeraj Rajendran   Email: srajend2@ncsu.edu   ID: 200210462

Task 3. Linear Multivariable Regression

3.1 Carry out a multivariable regression on all the independent variables, and determine the values for all the coefficients, and σ^2.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.981
Model:                            OLS   Adj. R-squared:                  0.981
Method:                 Least Squares   F-statistic:                     2897.
Date:                Sun, 28 Oct 2018   Prob (F-statistic):          1.69e-235
Time:                        21:09:58   Log-Likelihood:                -1461.5
No. Observations:                 281   AIC:                             2935.
Df Residuals:                     275   BIC:                             2957.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        120.0268     60.820      1.973      0.049       0.296     239.758
X1            13.9293      0.154     90.572      0.000      13.627      14.232
X2             2.7302      0.155     17.651      0.000       2.426       3.035
X3             5.7720      0.156     37.052      0.000       5.465       6.079
X4             7.9381      0.160     49.695      0.000       7.624       8.253
X5             8.4949      0.146     58.170      0.000       8.207       8.782
==============================================================================
Omnibus:                      103.583   Durbin-Watson:                   2.274
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              258.471
Skew:                           1.774   Prob(JB):                     7.48e-57
Kurtosis:                       6.080   Cond. No.                     8.87e+03
==============================================================================
```

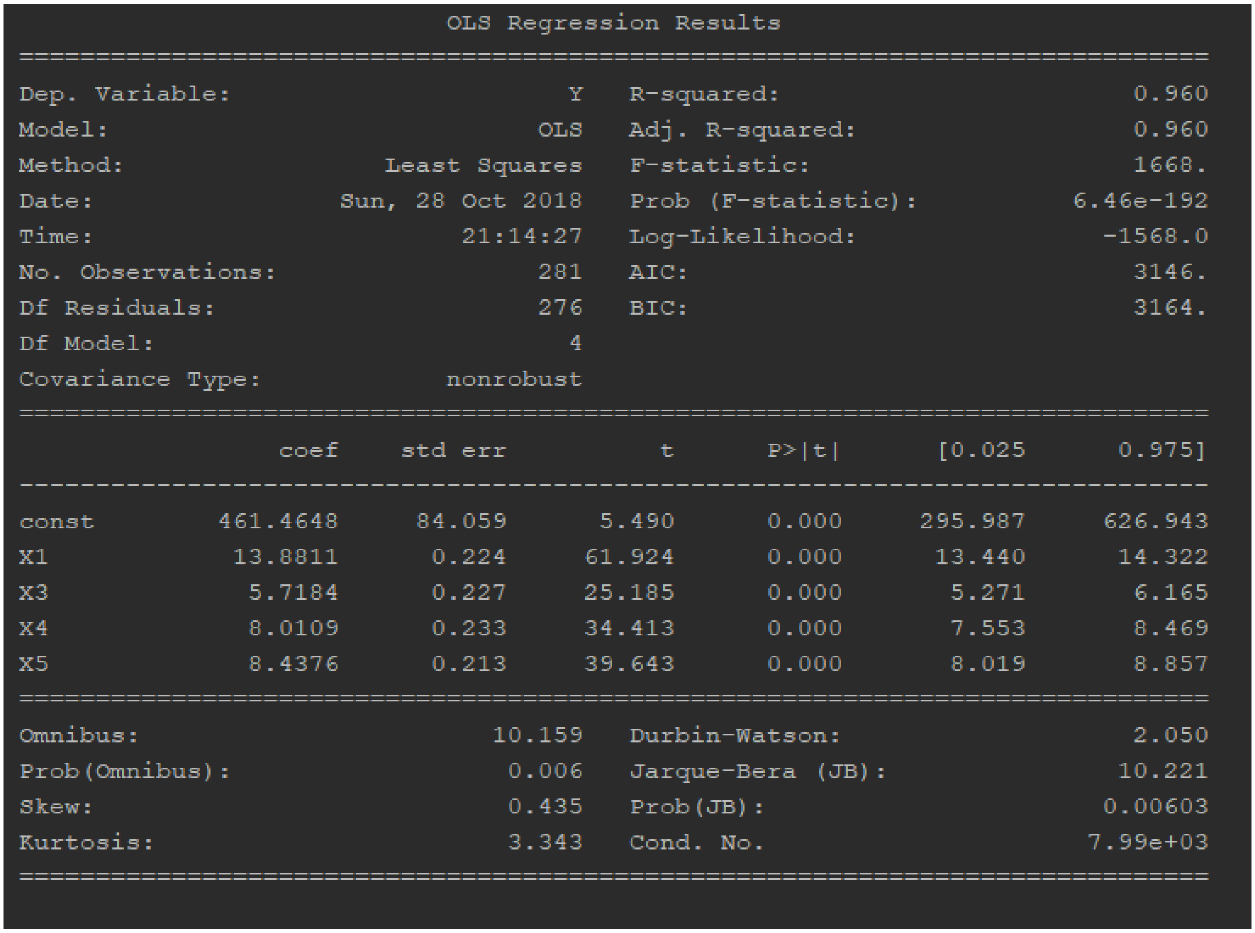
R squared = 0.981, F value= 2897, p values=0 for X1 through X5, coefficients are listed in table
variance =1928.5843098079024
Comments: R squared value is closer to one and F value is high indicating that the model is good. The p values are 0 for all X variables, but high for constant (>0.01). Hence the model seems to be good.

Task 3.2: Based on the *p*-values, *R_2*, *F* value, and correlation matrix, identify which independent variables need to be removed (if any) and go back to step 3.1.

Covariance Matrix:

```
      X1        X2       X3              X4         X5          Y
X1  1.000000 -0.021343 -0.000240 -0.120413  0.015228  0.705355
X2 -0.021343  1.000000 -0.015942  0.028434 -0.023076  0.125173
X3 -0.000240 -0.015942  1.000000  0.076431 -0.073155  0.300592
X4 -0.120413  0.028434  0.076431  1.000000 -0.073015  0.316279
X5  0.015228 -0.023076 -0.073155 -0.073015  1.000000  0.436696
Y   0.705355  0.125173  0.300592  0.316279  0.436696  1.000000
```

Considering covariance matrix, X2 has least influence on Y. Removing X2 gives following result:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.960
Model:                            OLS   Adj. R-squared:                  0.960
Method:                 Least Squares   F-statistic:                     1668.
Date:                Sun, 28 Oct 2018   Prob (F-statistic):          6.46e-192
Time:                        21:14:27   Log-Likelihood:                -1568.0
No. Observations:                 281   AIC:                             3146.
Df Residuals:                     276   BIC:                             3164.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        461.4648     84.059      5.490      0.000     295.987     626.943
X1            13.8811      0.224     61.924      0.000      13.440      14.322
X3             5.7184      0.227     25.185      0.000       5.271       6.165
X4             8.0109      0.233     34.413      0.000       7.553       8.469
X5             8.4376      0.213     39.643      0.000       8.019       8.857
==============================================================================
Omnibus:                       10.159   Durbin-Watson:                   2.050
Prob(Omnibus):                  0.006   Jarque-Bera (JB):               10.221
Skew:                           0.435   Prob(JB):                      0.00603
Kurtosis:                       3.343   Cond. No.                     7.99e+03
==============================================================================
```

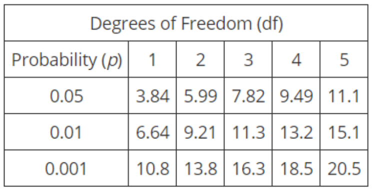
p-value is zero for all and R-squared and F-values are good too. Variance= 4113.514794117363
Hence, this model obtained by removing independent variable X2, is finalized.

3.3 Do a residuals analysis:

a. Do a Q-Q plot of the pdf of the residuals against N(0, *s*2). In addition, draw the residuals histogram and carry out a χ2 test that it follows the normal distribution N(0, *s*2).

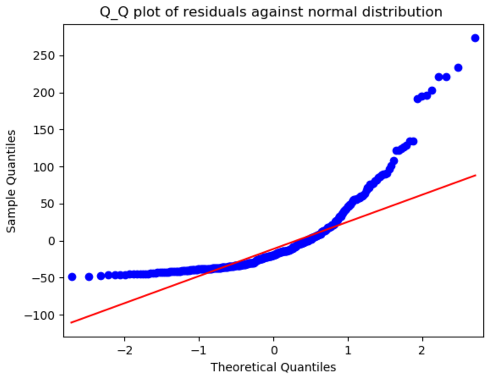
For 99percentile, z value should be greater than 2.58.

Critical Chi squared values for different degrees of freedom are shown below. Obtained Chi squared value needs to be lower than critical value.

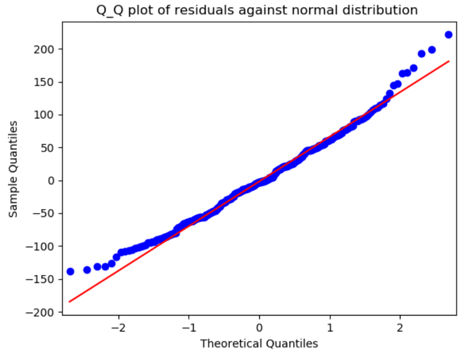


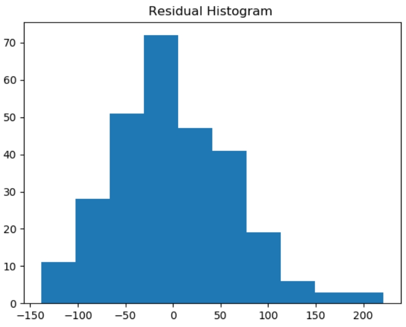
	Degrees of Freedom (df)				
Probability (*p*)	1	2	3	4	5
0.05	3.84	5.99	7.82	9.49	11.1
0.01	6.64	9.21	11.3	13.2	15.1
0.001	10.8	13.8	16.3	18.5	20.5

Q-Q plot should not be diverging from reference line for good fit.
It was decided from previous test that X2 will be omitted. But the Q_Q plot without any omition is shown here for reference. Huge deviations can be seen at lower and higher ranges indicating really bad fit.



X2 omitted

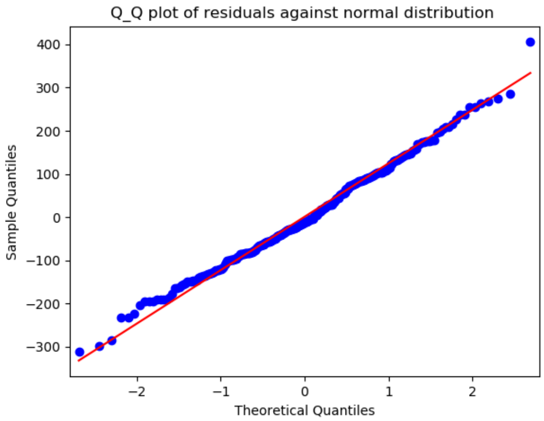


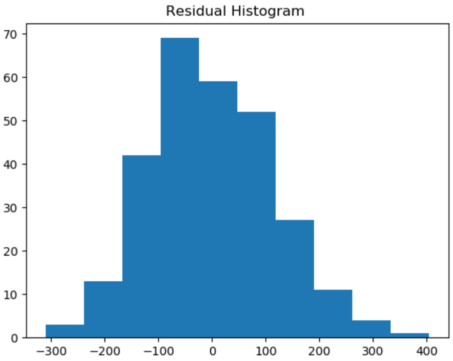


Q_Q plot and histogram indicates a good fit.

Z value 10.159203976565179 > 2.58 hence feasible
chi squared probability for the hypothesis test 0.006222385105335757 < 9.21 hence feasible

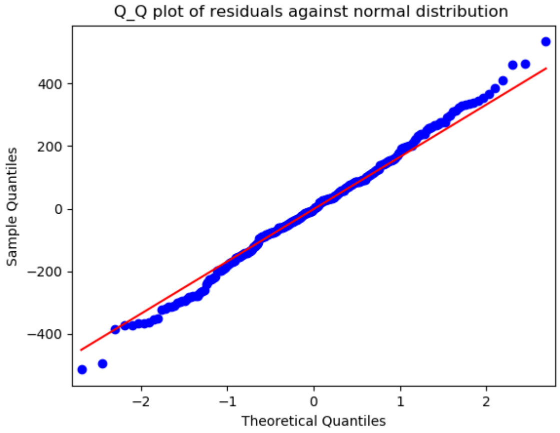
If X2 andX3 omitted

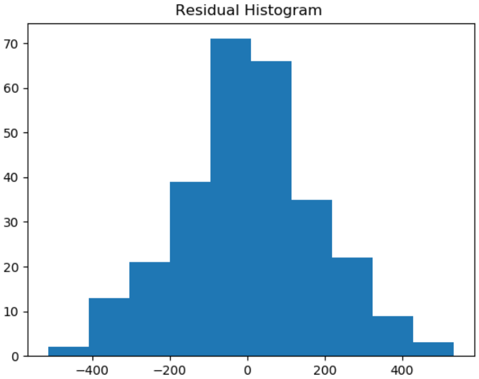




Z value 2.733425987611487 > 2.58 hence feasible.
chi squared probability for the hypothesis test 0.25494358494600877 < 11.3 Hence feasible

If, X2,X3,X4 are omitted





Q_Q plot is better compared to X2 omission and relatable to X2+X3 omission.

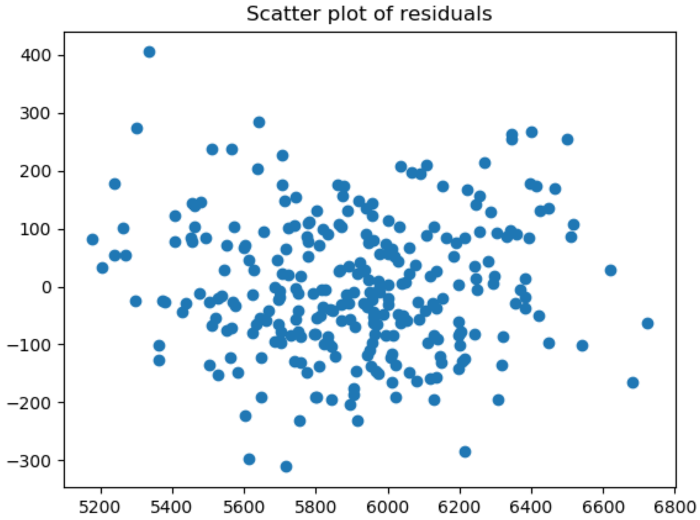
Z value 0.017769466080486202 <2.58 . **Hence not feasible**
chi squared probability for the hypothesis test 0.9911546195683092 < chi critical

It can be concluded that best results are obtained on omitting X2 and X3 independent variables.

b. Do a scatter plot of the residuals to see if there are any trends.

For X2 and X3 independent variables omitted:

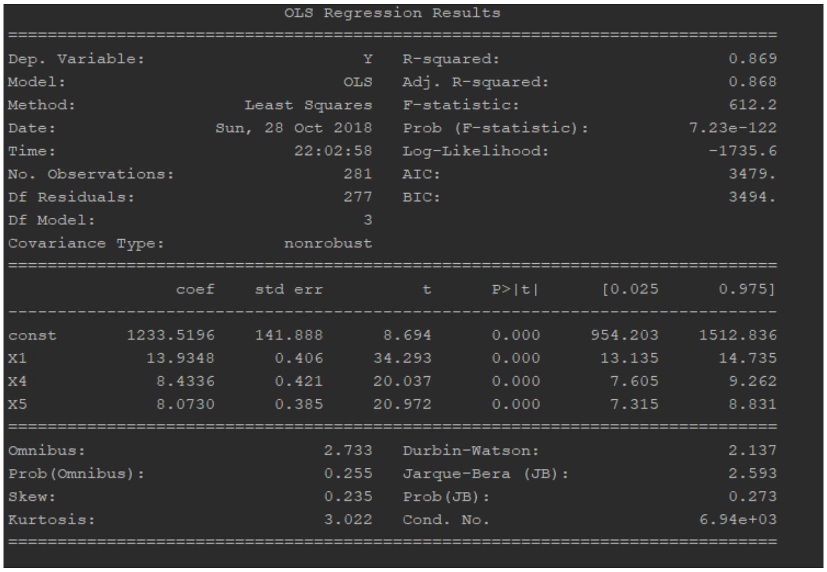
Y axis: Residuals
X Axis: predicted Y value



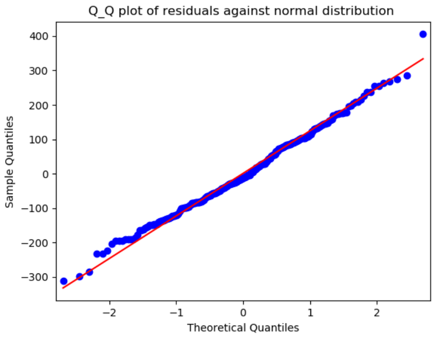
No trends of positive or negative correlation can be seen. Hence The model is good.

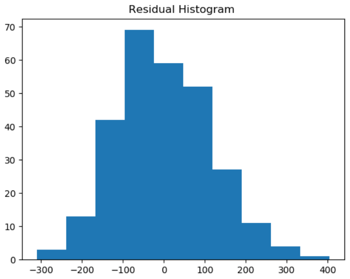
Final Result:

Y=const + a1*X1+a4*X4+a5*X5 + residual
Variance of residual = 13567.097185288696

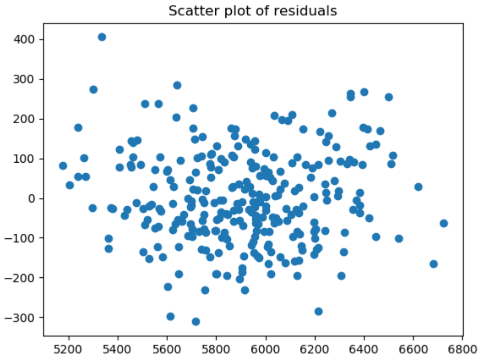


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.869
Model:                            OLS   Adj. R-squared:                  0.868
Method:                 Least Squares   F-statistic:                     612.2
Date:                Sun, 28 Oct 2018   Prob (F-statistic):          7.23e-122
Time:                        22:02:58   Log-Likelihood:                -1735.6
No. Observations:                 281   AIC:                             3479.
Df Residuals:                     277   BIC:                             3494.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1233.5196    141.888      8.694      0.000     954.203    1512.836
X1            13.9348      0.406     34.293      0.000      13.135      14.735
X4             8.4336      0.421     20.037      0.000       7.605       9.262
X5             8.0730      0.385     20.972      0.000       7.315       8.831
==============================================================================
Omnibus:                        2.733   Durbin-Watson:                   2.137
Prob(Omnibus):                  0.255   Jarque-Bera (JB):                2.593
Skew:                           0.235   Prob(JB):                        0.273
Kurtosis:                       3.022   Cond. No.                     6.94e+03
==============================================================================
```





Z value 2.733425987611487 > 2.58 hence feasible.
chi squared probability for the hypothesis test 0.25494358494600877 < 11.3 Hence feasible


Scatter plot of residuals