IOT Analytics Project 2

Task 1. Basic statistics analysis

1.1. For each variable Xi, i.e. column in the data set corresponding to Xi, calculate the following: Histogram, mean, variance.

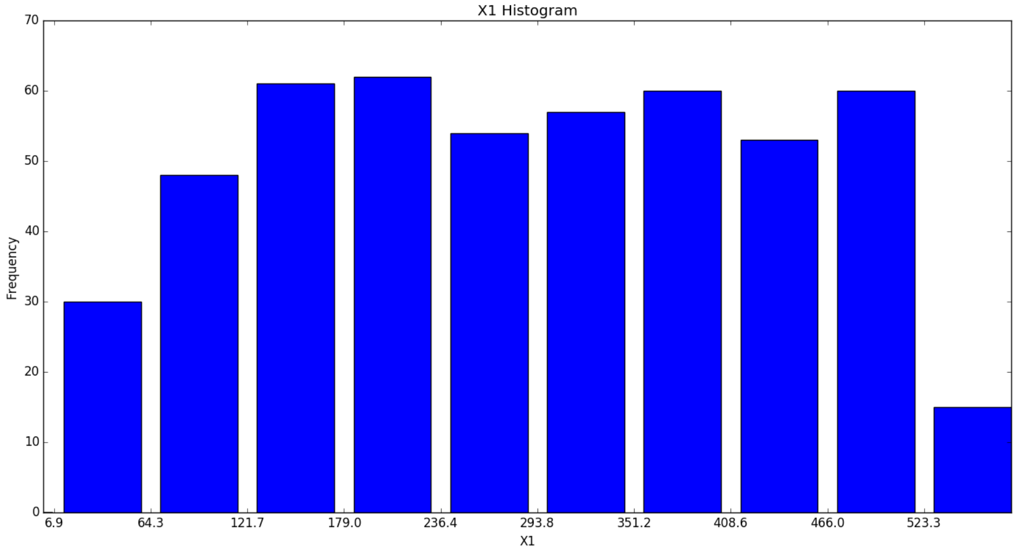
1.2 Calculate the correlation matrix Σ among all variables, i.e., Y, X1, X2, X3, X4 and X5. Draw conclusions related to possible dependencies among these variables.

1.3 Comment on the results

Solutions

1.1 Following is histogram (with 10 bins), mean and variance output generated from the python code

```
X1 Statistics
Mean: 290.124089121
Variance: 20950.37758
Histogram:
```



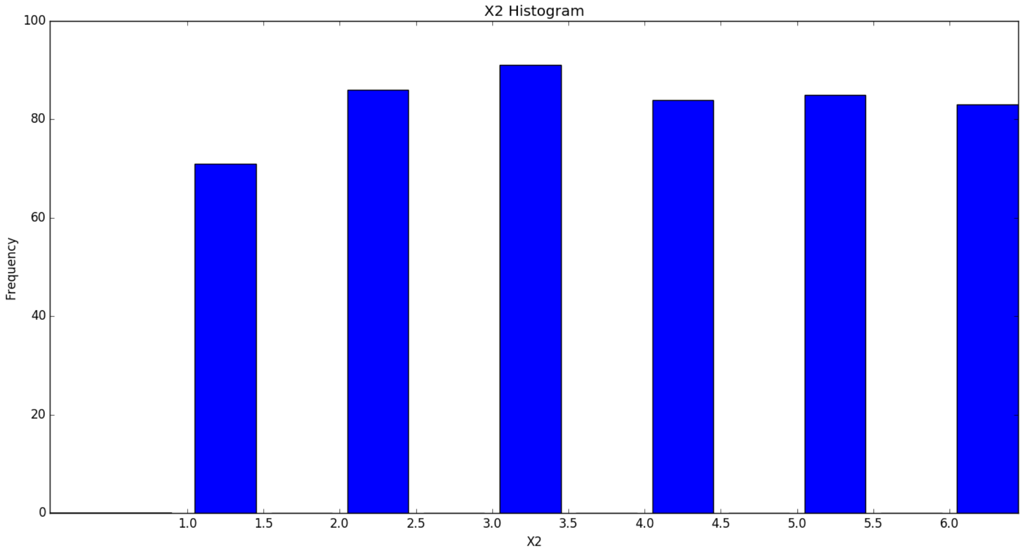
Comments:
The X1 predictor values have almost a uniform distribution across the range 6.88-580.72. Hence variance is also high

X2 Statistics

Mean: 3.55

Variance: 2.7795

Histogram:



Comments:

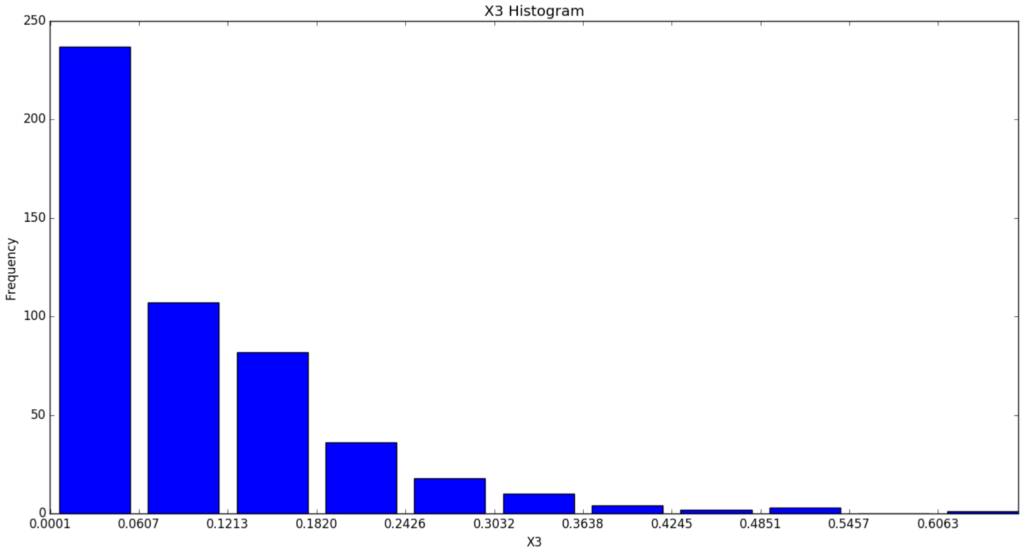
X2 values show a uniform distribution across the range 1-6 (It takes only integer values in the range 1-6)

X3 Statistics

Mean: 0.0961538665037

Variance: 0.00850874896488

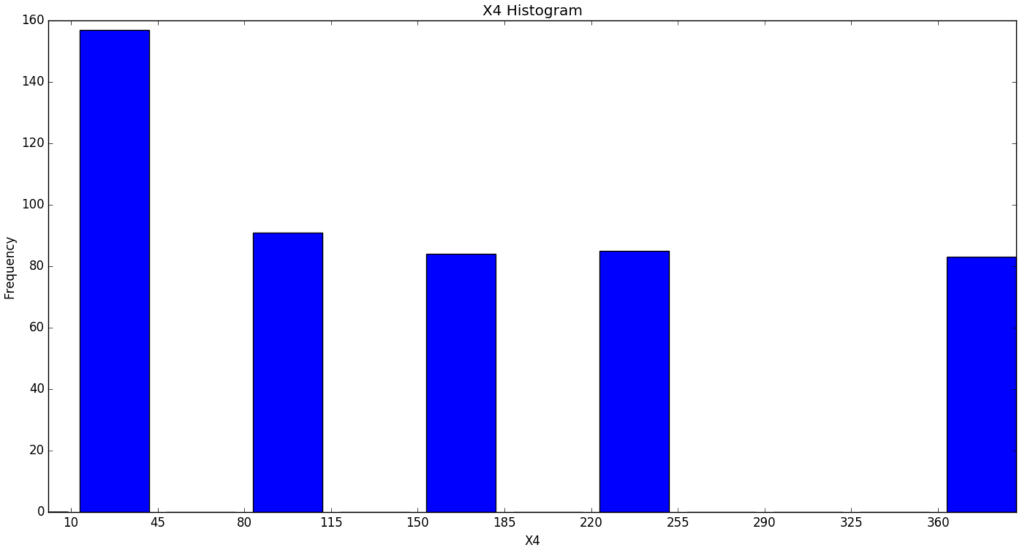
Histogram:



Comments:

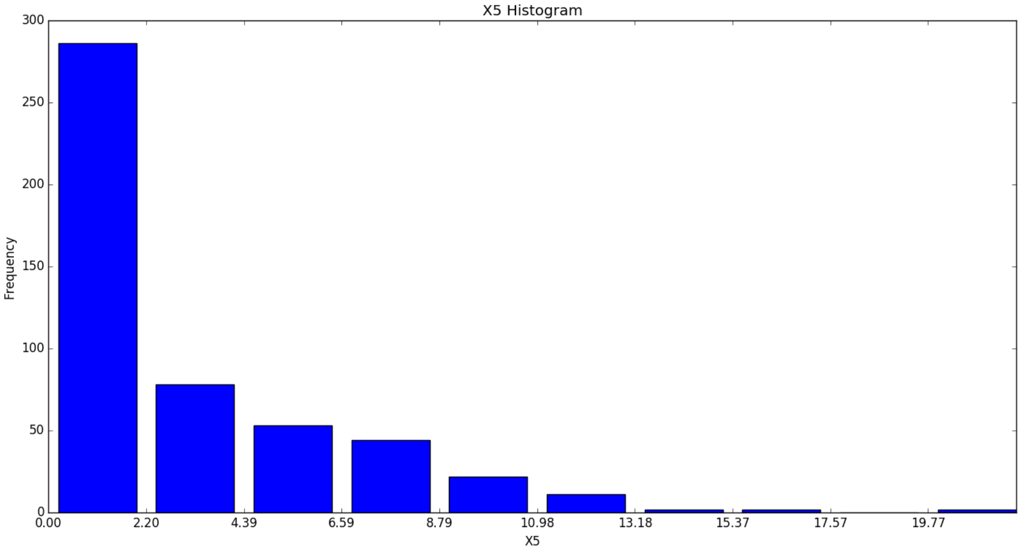
X3 shows a decreasing frequency towards the higher values in the range 0.000092-0.606

X4 Statistics
Mean: 153.82
Variance: 14542.4076
Histogram:



Comments:
X4’s distribution looks to be concentrated at uniform intervals. The values lie in the range 10-360. Hence variance is also high

X5 Statistics
Mean: 2.94755715125
Variance: 12.4413830852
Histogram:



Comments:

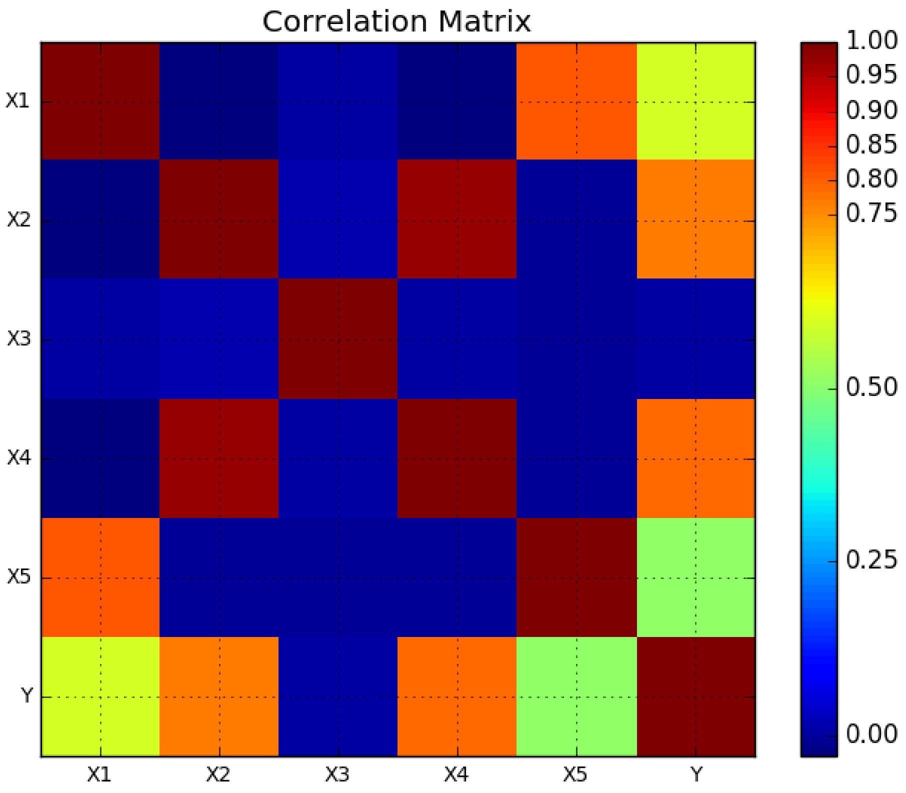
The frequency of X5 is decreasing towards the higher values in the range 0-21.96

1.2 Correlation matrix Σ among all variables, i.e., Y, X1, X2, X3, X4 and X5. Draw conclusions related to possible dependencies among these variables.

Solution

```
Correlation Matrix
           X1        X2        X3        X4        X5         Y
X1  1.000000 -0.028226  0.008334 -0.030691  0.808730  0.594091
X2 -0.028226  1.000000  0.014436  0.979061 -0.003626  0.769769
X3  0.008334  0.014436  1.000000  0.004659 -0.002230  0.009335
X4 -0.030691  0.979061  0.004659  1.000000 -0.005422  0.784811
X5  0.808730 -0.003626 -0.002230 -0.005422  1.000000  0.506689
Y   0.594091  0.769769  0.009335  0.784811  0.506689  1.000000
```

Pictorially it can be represented as



Conclusions:
As can be clearly seen from the matrix,
There is a high correlation between X1 and X5 (0.808).
Similarly, X2 and X4 have a very high correlation of 0.97
All variables are weakly correlated with X3. As a matter of fact, X3 is also not correlated with Y.

Regarding correlation with Y, all independent variables except X3 have a fairly high (> 0.5) correlation with Y. X3 and Y are highly uncorrelated. So, Y does not show dependence on X3.
It is also obvious that all variables are 100% correlated with themselves.

1.3 Overall Comments

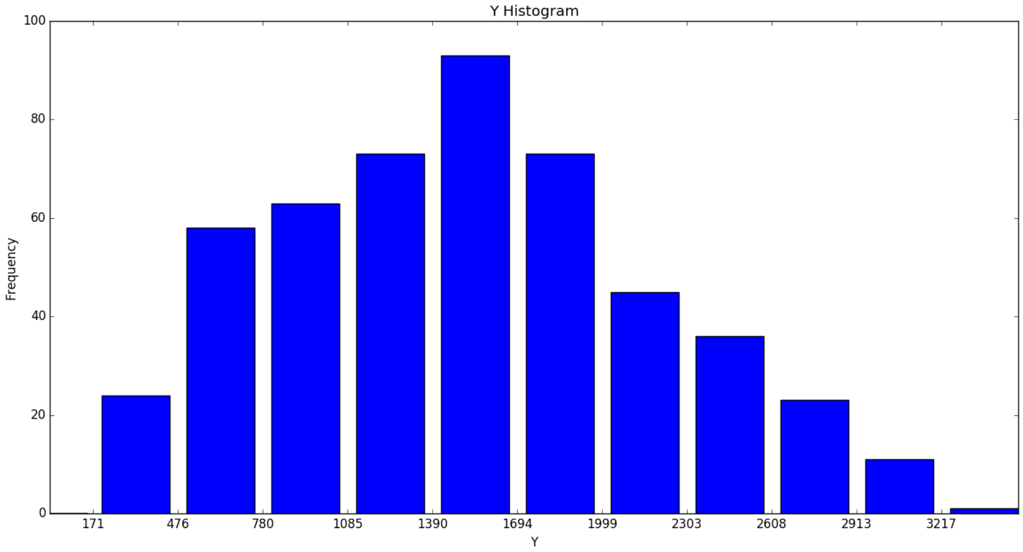
Here is a quick statistical summary of all the variables.

```
               X1          X2          X3          X4            X5  \
count  500.000000  500.000000  500.000000  500.000000  5.000000e+02
mean   290.124089    3.550000    0.096154  153.820000  2.947557e+00
std    144.887413    1.668853    0.092335  120.712678  3.530767e+00
min      6.888046    1.000000    0.000092   10.000000  2.023632e-18
25%    167.161179    2.000000    0.028700   40.000000  2.410634e-01
50%    291.101319    4.000000    0.067882  160.000000  1.366149e+00
75%    410.419896    5.000000    0.136746  250.000000  4.909353e+00
max    580.721180    6.000000    0.606321  360.000000  2.196329e+01

                Y
count   500.000000
mean   1514.667647
std     675.695478
min     171.257672
25%    1003.640591
50%    1500.081343
```

Comments on correlation matrix:
We observe that X3 is not correlated with Y. Considering very less dependency of Y on X3, we should exclude it when predicting the value of Y (when doing Multiple linear regression). When included, it may result in the case of overfitting.
Also, X3 shows minimal correlation with other independent variables X1, X2, X4 and X5 which makes it a good candidate for zero contribution to multi-collinearity when performing multivariate linear regression. We will analyze it more in the third task discussed below.

We also, see that the independent variables, X1, X2, X4 and X5 are highly correlated with Y. Hence, they are good candidates for predictor variables. But there is a high correlation between X1 and X5 and also X2 and X4 which can result in multicollinearity problem
We will analyze the effect in task 3 below

Comments on Y distribution: Y has a high mean. It is much likely that it has only positive correlation with predictor variables. Also, on plotting Y’s histogram, it looks close to a normal distribution



Task 2: Linear regression

Before proceeding with the multiple regressions, you will carry out a simple linear regression to estimate the parameters of the model: $Y = a0 + a1X + \varepsilon$, where $X = X1$.

2.1 Determine the values for a0, a1, and s^2.

2.2 Check the p-values, R^2, F value to determine if the regression coefficients are meaningful.

2.3 Plot the regression line against the data.

2.4 Do residuals analysis:

a. Do a Q-Q plot of the pdf of the residuals against N (0, s2) Alternatively, draw the residuals histogram and carry out a χ 2 test that it follows the N $(0, s^2)$.

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

2.7 Use a higher-order polynomial regression, i.e., $Y = a0 + a1X + a2X2 + \varepsilon$, to see if it gives better results.

2.8 Comment on your results in a couple of paragraphs.

Solutions

2.1 This was achieved using the python functions written in project_P2.py as submitted and it was verified using the scipy and pandas.stats library in python

Slope (a1) = 2.77059784726, intercept (a0) = 710.850469793

s^2 = 294831.701777

2.2 R^2 = 0.352944381494

F-Value = 271.640175832

p-value = 5.0366022946e-49 = 0(approximately)

```
   Quick summary of Regression Analysis for X1 using pandas.stats library

-------------------------Summary of Regression Analysis-------------------------

Formula: Y ~ <X1> + <intercept>

Number of Observations:         500
Number of Degrees of Freedom:   2

R-squared:         0.3529
Adj R-squared:     0.3516

Rmse:            544.0733

F-stat (1, 498):   271.6402, p-value:     0.0000

Degrees of Freedom: model 1, resid 498

-----------------------Summary of Estimated Coefficients------------------------
      Variable       Coef    Std Err     t-stat    p-value    CI 2.5%   CI 97.5%
--------------------------------------------------------------------------------
            X1     2.7706     0.1681      16.48     0.0000     2.4411     3.1001
     intercept   710.8505    54.5035      13.04     0.0000   604.0237   817.6772
   ---------------------------------End of Summary---------------------------------
```

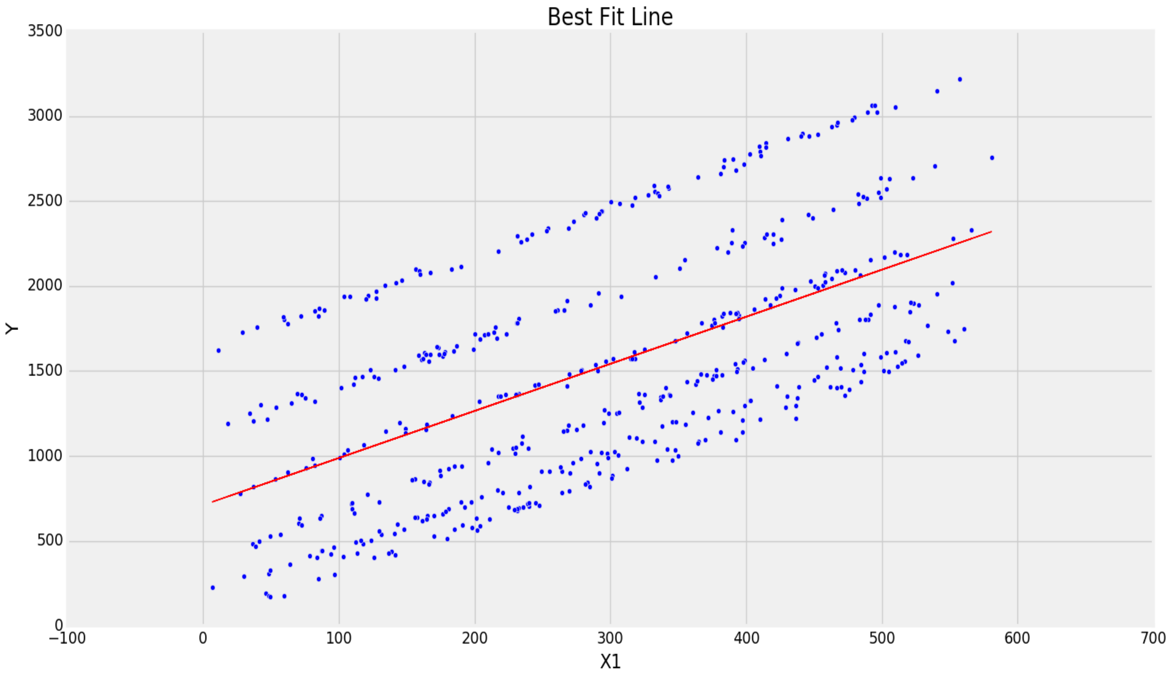
Comments:

R^2 value is around 0.3529 which is low in order to ensure goodness of the model to fit the data. Low R shows there is more scatter around the regression line. So it is possible that if we add more predictor variables to predict Y value, the model may fit better.

P value corresponding to the F-test is 0.000 which is significant (considering 0.025 as the cutoff for significance) enough to state that our model provides a better fit than intercept only model. So X1 is a meaningful addition. Higher F value and low p would mean a meaningful predictor

When considering p value of X1 we find that it is low (around 0), which implies good variability of the response variable (Y) with changing X1. So there is a good relationship between X1 and Y. So, predictor X1 is a meaningful addition to our model.

2.3 Regression Line is shown in the following figure



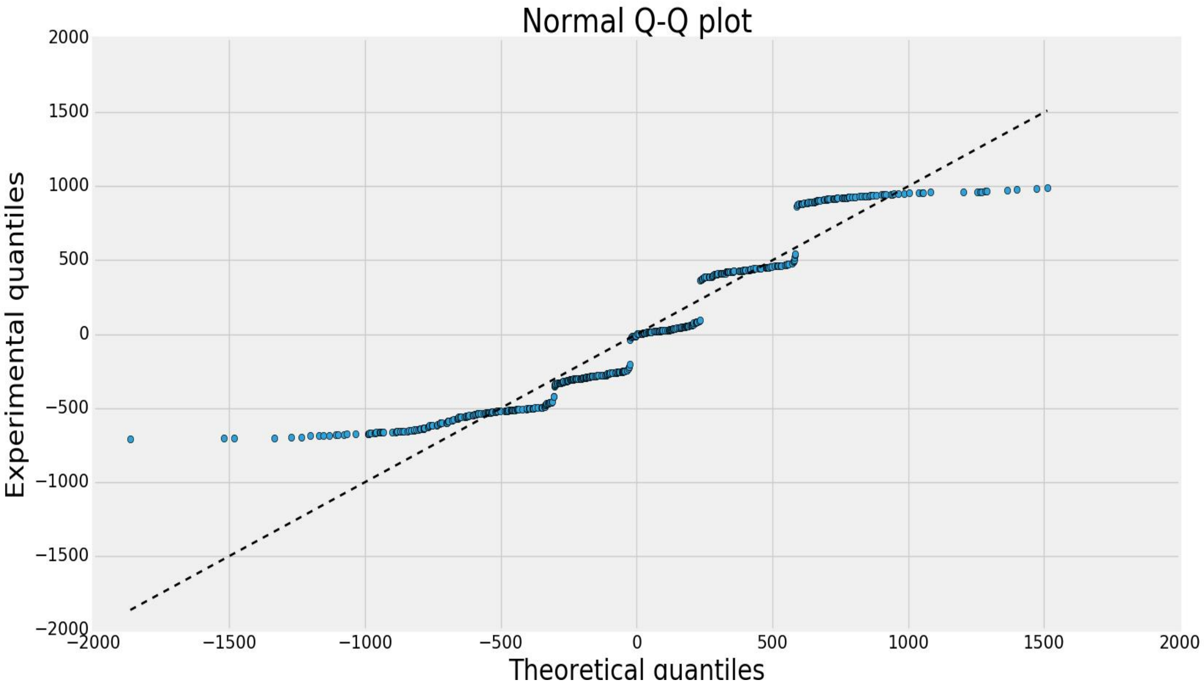
Comments: As can be seen and also as analyzed above, the line has a lot of scatter around it which accounts for a small R_squared value

2.4 Residual Analysis

In residual analysis, we observe the residual values distribution and will assess that the residuals are consistent with stochastic errors. It implies that the residuals should not be either symmetrically high or low. So their average should be zero throughout the range of fitted values. Also, the residuals are assumed to be normally distributed in ordinary least squares context.
Any non-randomness in residuals implies that the deterministic portion (predictor variables) of the model is not capturing some information that may be leaking into the residuals.
In addition, the residuals should not be correlated with any predictor variables and also not auto-correlated.

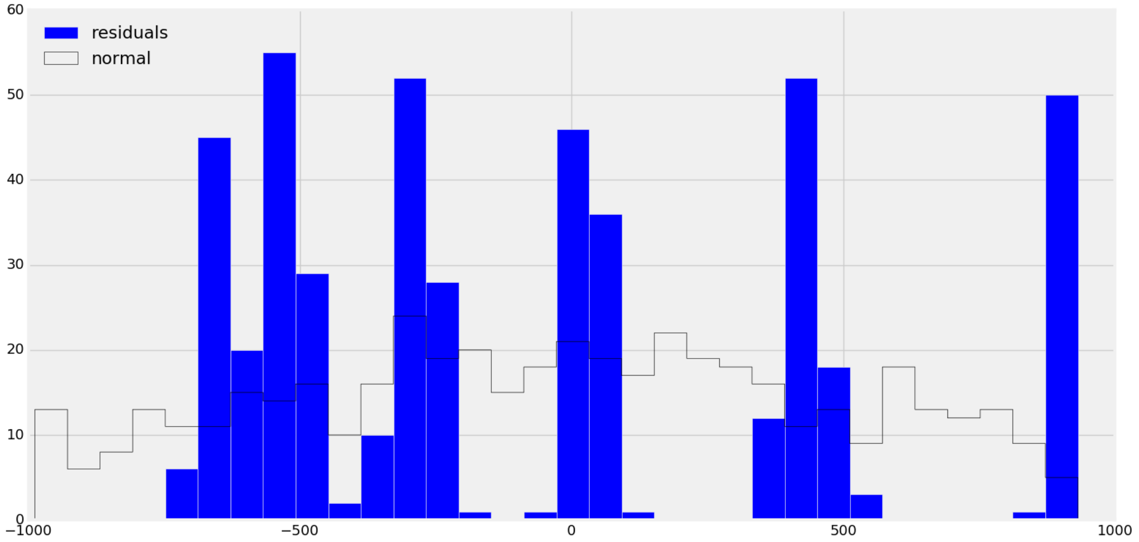
2.4 (a)

Q-Q plots of the residuals



As can be seen, the distribution of residuals is close to the normal distribution line but it has tails and high modality. There are multiple breaks (camel humps like structure) in residual when compared with a normal distribution (shows a multi modal behavior). So, residuals distribution here doesn't resemble a normal distribution perfectly.

Comparing the normal distributions histograms



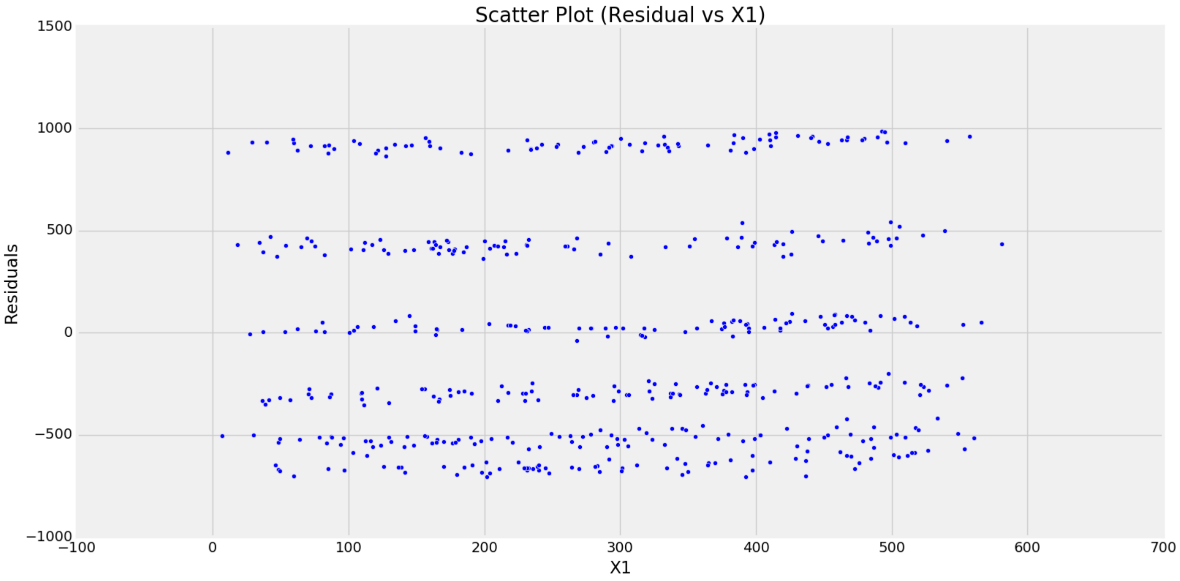
Normal distribution for the residuals is very discontinuous (multi-modal) and is not a perfect representation of the normal distribution $N(0,s^2)$

We perform the chi-square test on these two histograms data
H0: the data are normally distributed
Ha: the data are not normally distributed

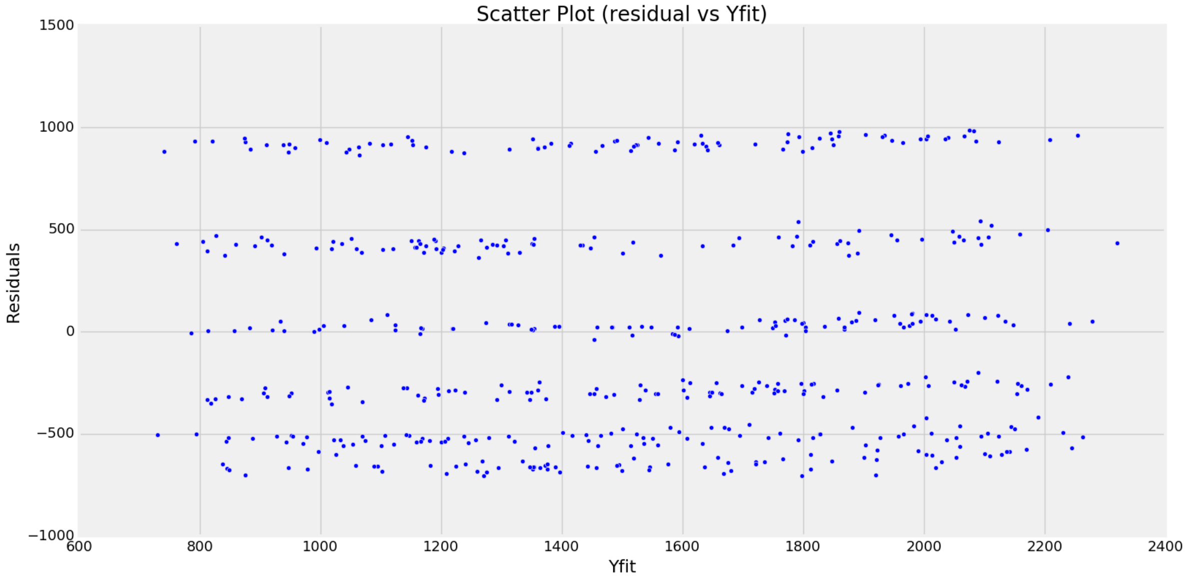
From the program and using scipy library, the calculated chi_square with 32 bins value is
X^2 = 1121.23323618
Calculating critical value for chi_square statistics with
X^2_crit = 42.557
Since X^2 > X^2_crit, we can reject the hypothesis, and know that the residual data is not normally distributed

2.4(b)

Scatter plots of residuals with X1



Scatter plot of residuals with Yfit (best fit values)



Comments:

From the scatter plot, it can be seen that the residuals average out to zero value and they have no correlation with the X1 and predicted values. X1 is a good candidate for the model

2.7 Analysis using higher order polynomial

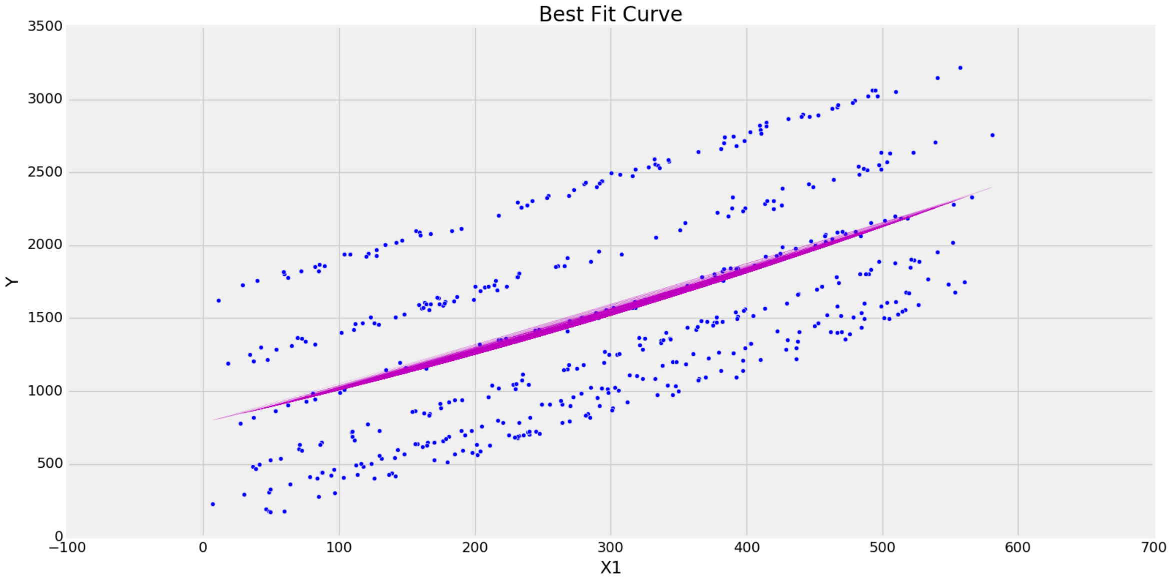
Summary of univariate linear regression with order 2

```
-------------------------Summary of Regression Analysis-------------------------

Formula: Y ~ <X1_sq> + <X1> + <intercept>

Number of Observations:         500
Number of Degrees of Freedom:   3

R-squared:         0.3542
Adj R-squared:     0.3516

Rmse:            544.1052

F-stat (2, 497):   136.2750, p-value:     0.0000

Degrees of Freedom: model 2, resid 497

-----------------------Summary of Estimated Coefficients------------------------
      Variable       Coef    Std Err     t-stat    p-value    CI 2.5%   CI 97.5%
--------------------------------------------------------------------------------
         X1_sq     0.0012     0.0012       0.97     0.3323    -0.0012     0.0036
            X1     2.0799     0.7314       2.84     0.0046     0.6463     3.5134
```

```
     intercept    785.4800    94.2628        8.33     0.0000    600.7248    970.2351
--------------------------------End of Summary---------------------------------
```

We can observe that the R squared value has not changed much from the 1st order simple linear regression. Also, the standard error for X1 has increased significantly. P-value for X1_sq coefficient is high and implies that including X1_sq coefficient is not better than the model with just X1 and intercept

Graphically also, the model doesn't improve

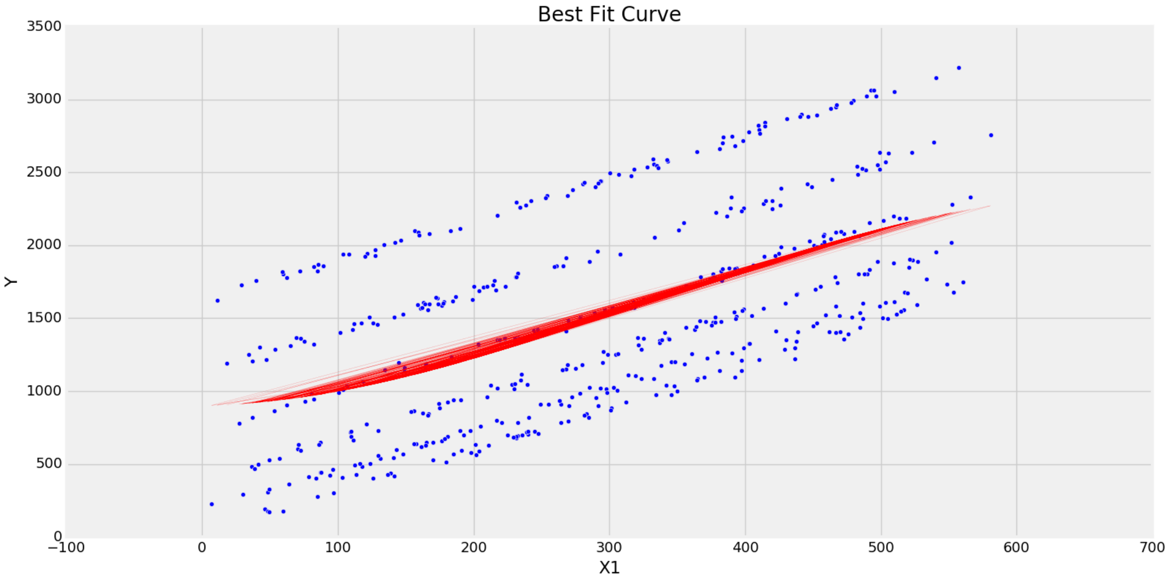


We shall try with a 3rd order polynomial and analyze the statistics

Summary is as follows

```
-------------------------Summary of Regression Analysis-------------------------

Formula: Y ~ <X1_cube> + <X1_sq> + <X1> + <intercept>

Number of Observations:         500
Number of Degrees of Freedom:   4

R-squared:          0.3556
Adj R-squared:      0.3517

Rmse:             544.0310

F-stat (3, 496):    91.2533, p-value:     0.0000

Degrees of Freedom: model 3, resid 496

-----------------------Summary of Estimated Coefficients------------------------
      Variable       Coef    Std Err     t-stat    p-value    CI 2.5%   CI 97.5%
```

```
---------------------------------------------------------------------------------
       X1_cube    -0.0000     0.0000      -1.07     0.2871    -0.0000     0.0000
         X1_sq     0.0092     0.0076       1.21     0.2272    -0.0057     0.0241
            X1     0.1481     1.9548       0.08     0.9396    -3.6833     3.9795
     intercept   900.5107   143.3041       6.28     0.0000   619.6347  1181.3867
   ---------------------------------End of Summary---------------------------------
```



The model is still not improving, with standard error for X1 going very high. Also the p-values for these intercept are representing that the model is getting worse.

2.8 Overall Comments:

For Linear Fit
X1 alone provides a bad predictor model. The R_squared value is quite low and standard error in X1 is also high to be considered as the best model.
The residuals do not have a normal distribution and show multi-modal behavior.
This can be seen using Q-Q plots and histograms.

For Polynomial Fit
X1^2 and X1^3 both fail to improve the model and cannot provide any better fit or prediction model. In fact, the standard errors for the predictor variables is increased with polynomial fit clearly indicating that the model gets worse with increasing order of X1. Also, the p values start getting out of significance zone with polynomial fits.

Hence, we need to consider other predictor variables to decide the best regression model for the given data. This analysis is done in multiple linear regression in task 3.

Task3. Multivariate regression

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

Using pandas.stats library in python following summary is presented for the model with multivariate linear regression

```
-------------------------Summary of Regression Analysis-------------------------

Formula: Y ~ <X1> + <X2> + <X3> + <X4> + <X5> + <intercept>

Number of Observations:         500
Number of Degrees of Freedom:   6

R-squared:         0.9988
Adj R-squared:     0.9987

Rmse:             23.9039

F-stat (5, 494): 79644.5847, p-value:     0.0000

Degrees of Freedom: model 5, resid 494

-----------------------Summary of Estimated Coefficients------------------------
      Variable       Coef    Std Err     t-stat    p-value    CI 2.5%   CI 97.5%
--------------------------------------------------------------------------------
            X1     2.7693     0.0126     220.29     0.0000     2.7447     2.7940
            X2     2.4599     3.1537       0.78     0.4358    -3.7214     8.6411
            X3     4.7774    11.6047       0.41     0.6808   -17.9678    27.5226
            X4     4.4627     0.0436     102.36     0.0000     4.3772     4.5481
            X5     5.8932     0.5156      11.43     0.0000     4.8826     6.9038
--------------------------------------------------------------------------------
     intercept    -1.7921     5.6475      -0.32     0.7511   -12.8612     9.2771
---------------------------------End of Summary---------------------------------
```

Values of all coefficients are as shown in the table above.

Coefficient of X1 = 2.7693

Coefficient of X2 = 2.4599

Coefficient of X3 = 4.7774

Coefficient of X4 = 4.4627

Coefficient of X5 = 5.8932

Intercept = -1.7921

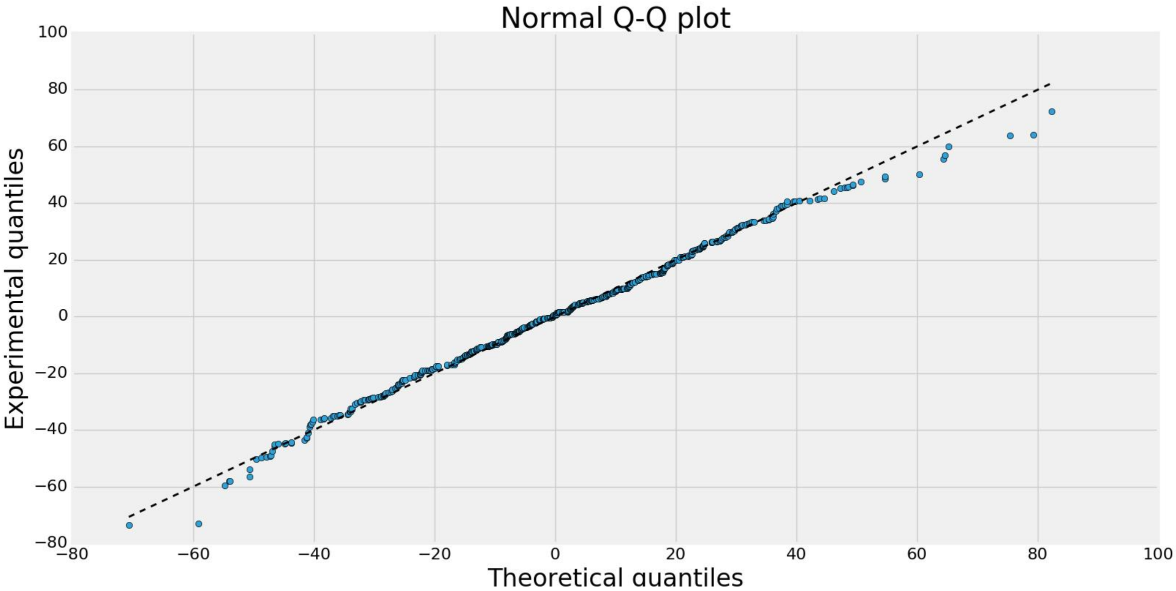
RMSE presented by the summary is 23.9039

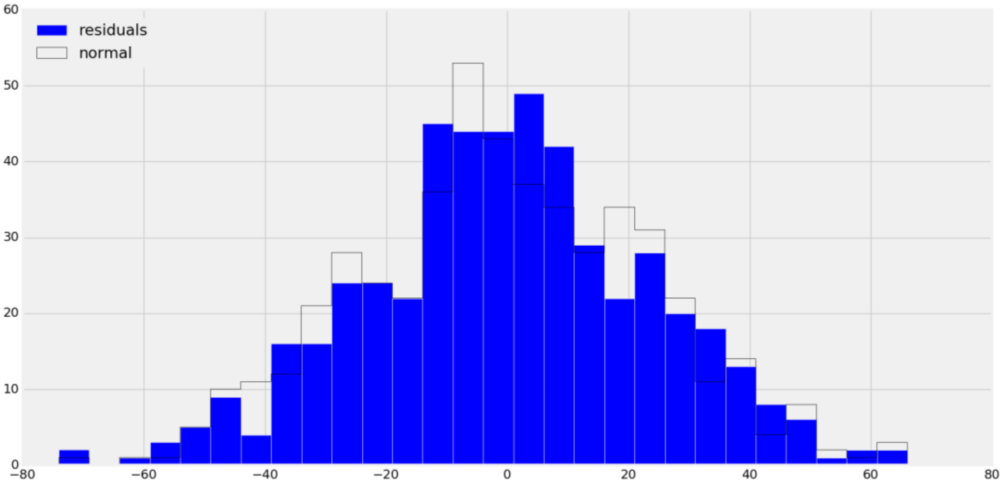
From the summary we have the following observations:

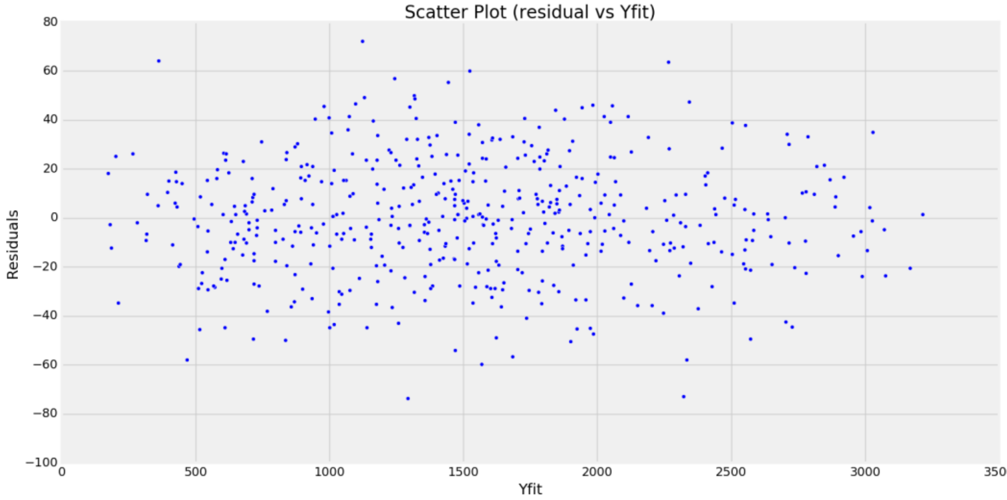
R-squared value is very high which is indicative of good fit of the model. But it can also be a case of overfitting. As we can see, the p-values for X2 and X3 are very high which clearly indicates that the model is bad with all the variables included. The standard error in these independent variables (X2 and X3) is also very high. Overall, the model looks very promising with following values on residual analysis

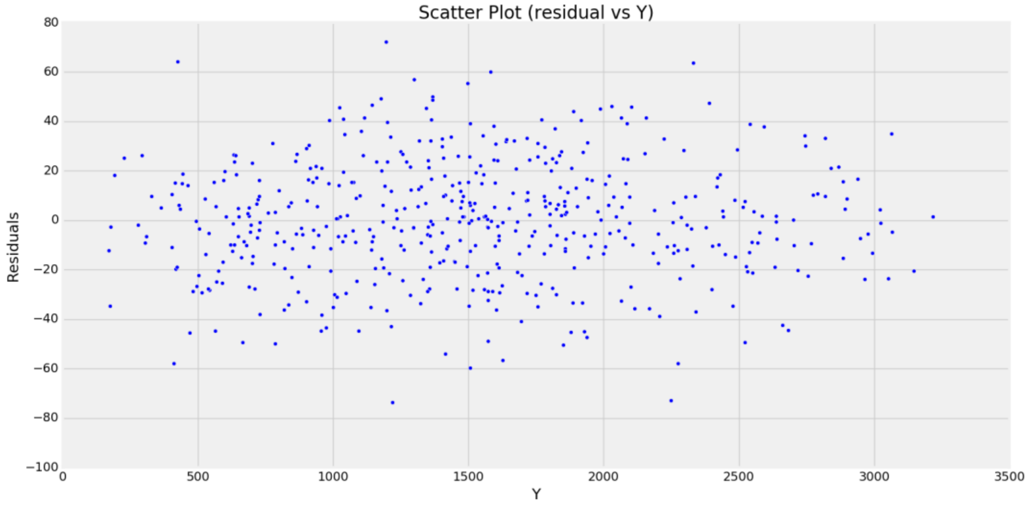
```
MultiVariate Linear regresssion model X1,X2,X3,X4,X5, Y
s_square= 564.540245174
RMSE= 23.8077188129
R_squared: 0.998761025577
Critical value for Chi-squared test: 37.8280426087
Critical value: 42.5569678043
```

Graphically too, it appears to be an excellent regression model. But due to the reasons indicated above and observing correlation matrix, we can say that it is case of overfitting.




residuals
normal
60
50
40
30
20
10
0
−80
−60
−40
−20
0
20
40
60
80


Scatter Plot (residual vs Yfit)
Residuals
80
60
40
20
0
−20
−40
−60
−80
−100
0
500
1000
1500
2000
2500
3000
3500
Yfit


Scatter Plot (residual vs Y)
Residuals
80
60
40
20
0
−20
−40
−60
−80
−100
0
500
1000
1500
2000
2500
3000
3500
Y

3.2 Based on the p-values, R2 , F value, and correlation matrix Σ, identify which independent variables need to be left out (if any) and go back to step 3.1

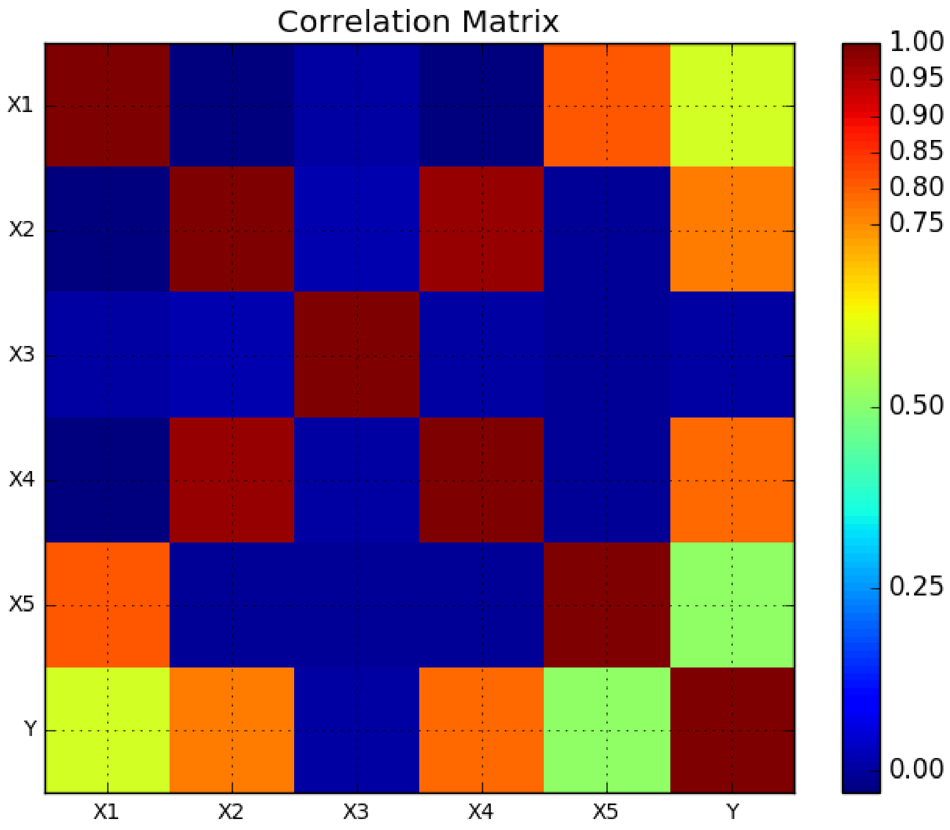
3.3 Do a residual analysis:

a. Do a Q-Q plot of the pdf of the residuals against N(0, s 2). Alternatively, draw the residuals histogram and carry out a χ 2 test that it follows the N(0, s 2).

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

Based on the above observation of R_sq, the model looks very promising and good fit but high values of p values (beyond significance zone) for X2 and X3 indicate that model might be better off even without their contribution.

Now let us see the correlation matrix from task 1 once again



Here, X3 is not correlated to Y. So it may be an additional entry without contributing to the regression model. It may be causing the problem of overfitting. So we should drop X3 predictor variable

Other predictor variables i.e. X1, X2, X4 and X5 have a fairly high correlation with the response variable Y. So their contribution can be significant in predicting Y.

Also it can be seen that X1 and X5 are highly correlated (0.808) and similarly X2-X4 have a high correlation (0.979). So their mutual correlation may cause the issue of multi collinearity. As a precaution for our best prediction model, we will need to drop one variable from each of the two pairs.

Let’s do it step by step.

First, we will analyze the model without X3. Then we can test the model by dropping some variables that are correlated with other variables. Let’s follow the following order and predict the best model.

1. Drop X3. Use X1, X2, X4 and X5
2. Use X1,X2
3. Use X1,X4
4. Use X2,X5
5. Use X4,X5

1. **Drop X3, Use X1,X2,X4,X5**
   Perform multivariate regression

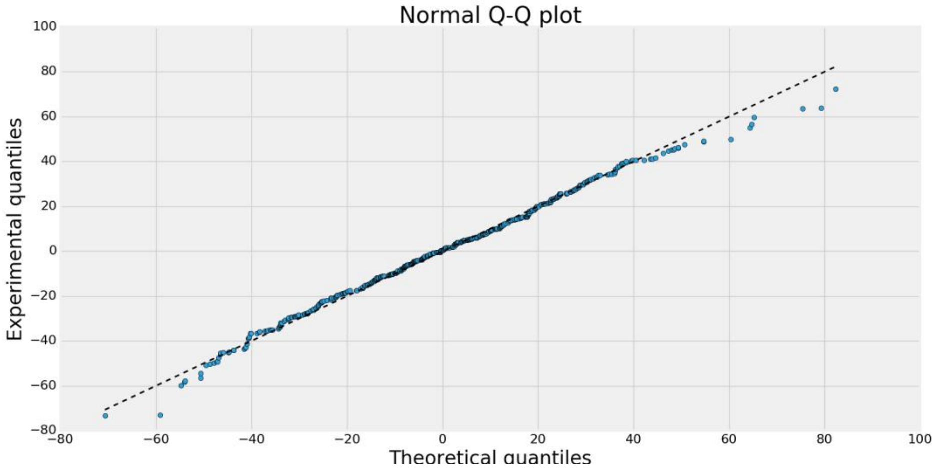
```
-------------------------Summary of Regression Analysis-------------------------

Formula: Y ~ <X1> + <X2> + <X4> + <X5> + <intercept>

Number of Observations:         500
Number of Degrees of Freedom:   5

R-squared:         0.9988
Adj R-squared:     0.9988

Rmse:             23.8838

F-stat (4, 495): 99723.0057, p-value:     0.0000

Degrees of Freedom: model 4, resid 495

-----------------------Summary of Estimated Coefficients------------------------
      Variable       Coef    Std Err     t-stat    p-value    CI 2.5%   CI 97.5%
--------------------------------------------------------------------------------
            X1     2.7694     0.0126     220.52     0.0000     2.7448     2.7940
            X2     2.5228     3.1473       0.80     0.4232    -3.6460     8.6916
            X4     4.4618     0.0435     102.53     0.0000     4.3765     4.5471
            X5     5.8899     0.5151      11.43     0.0000     4.8803     6.8996
     intercept    -1.4444     5.5793      -0.26     0.7958   -12.3798     9.4911
      ---------------------------------End of Summary---------------------------------
```

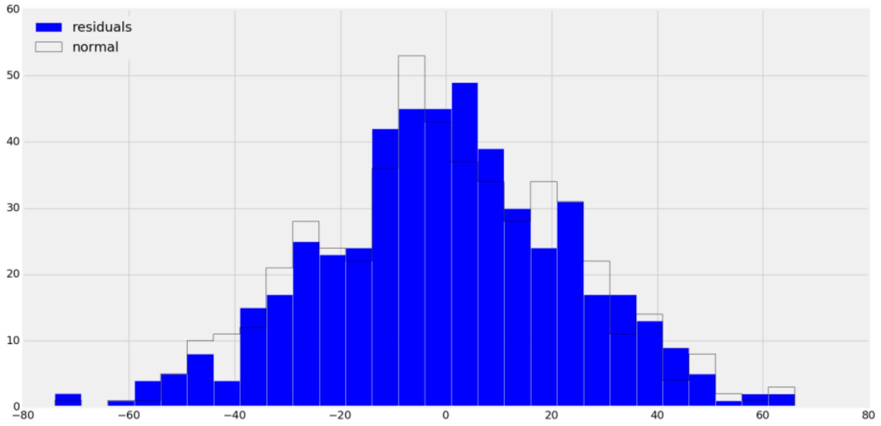
Comments: The model is still good without X3. So it can be dropped
But p value for X2 is outside the significance zone. Also, its standard error is high. X5 also has high standard error. It shows that we can have a better model without X2 and X5.

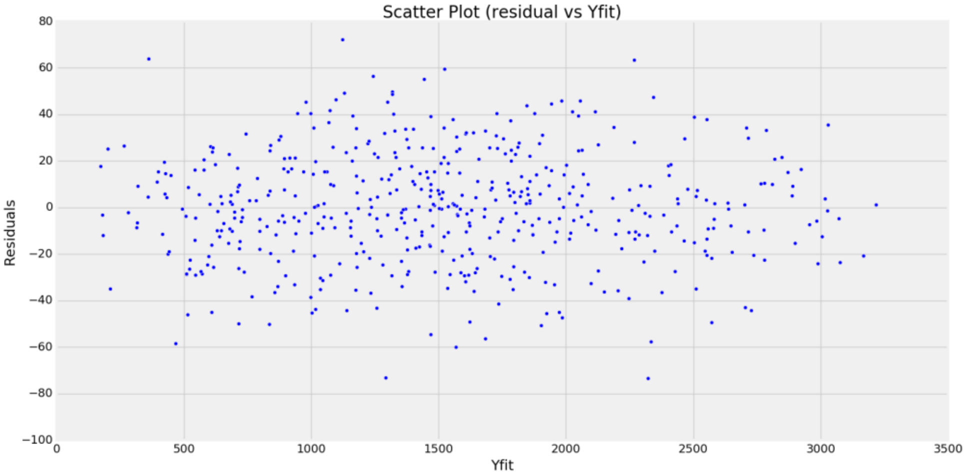
Let’s perform a residual analysis in this case

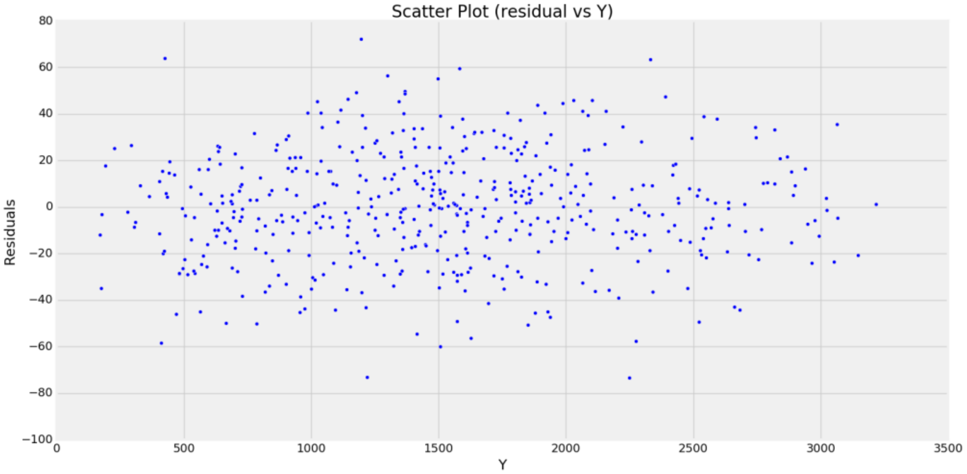
```
Linear regresssion model X1,X2,X4,X5, Y
s_square= 564.734001975
RMSE= 23.8118040059
R_squared: 0.998760600347
Critical value for Chi-squared test: 40.6120064988
Critical value: 42.5569678043
```

The residual fits pretty well with normal distribution. It can be seen through plots too









```
Scatter plots show no correlation trends
```

2. Use X1,X2 (correlation -0.028)

```
-------------------------Summary of Regression Analysis-------------------------

Formula: Y ~ <X1> + <X2> + <intercept>

Number of Observations:         500
Number of Degrees of Freedom:   3

R-squared:         0.9721
Adj R-squared:     0.9720

Rmse:            113.1310

F-stat (2, 497):  8651.9033, p-value:     0.0000

Degrees of Freedom: model 2, resid 497
```

```
-----------------------Summary of Estimated Coefficients-----------------------
      Variable       Coef    Std Err     t-stat    p-value    CI 2.5%   CI 97.5%
--------------------------------------------------------------------------------
            X1     2.8742     0.0350      82.20     0.0000     2.8057     2.9428
            X2   318.7123     3.0359     104.98     0.0000   312.7620   324.6627
     intercept  -450.6407    15.8381     -28.45     0.0000  -481.6834  -419.5980
---------------------------------End of Summary---------------------------------
```

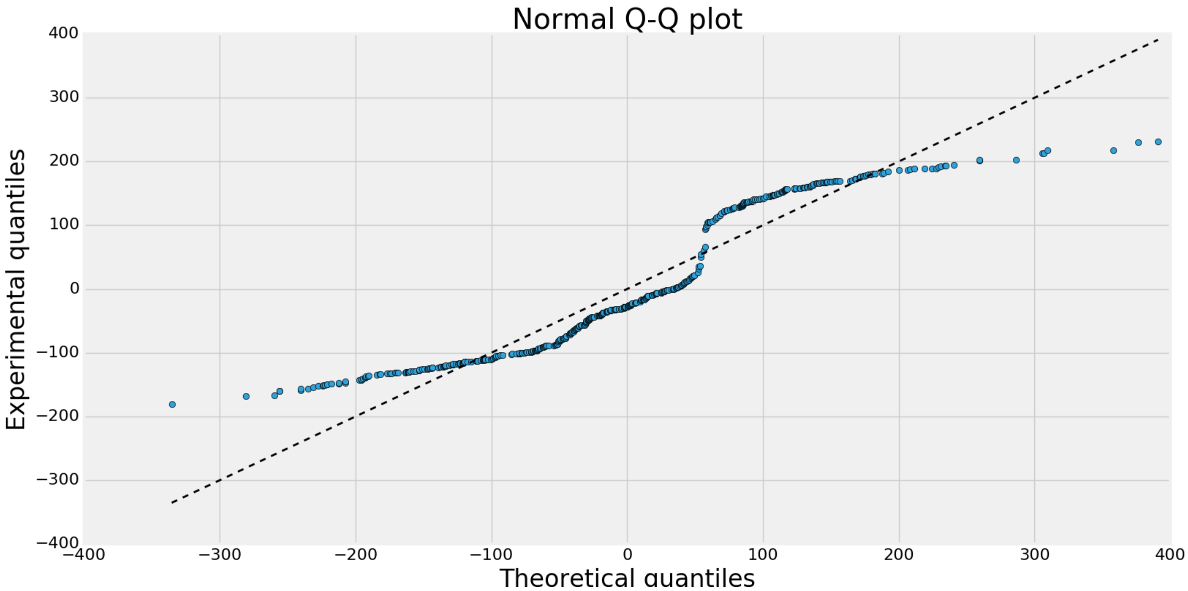
Comments: R_sq value has slightly reduced but still a good fit. p-value for X1 and X2 are inside significance zone. But since we see the Std error for X2 is still high we can use some other predictor variable.

Let’s perform the residual analysis.

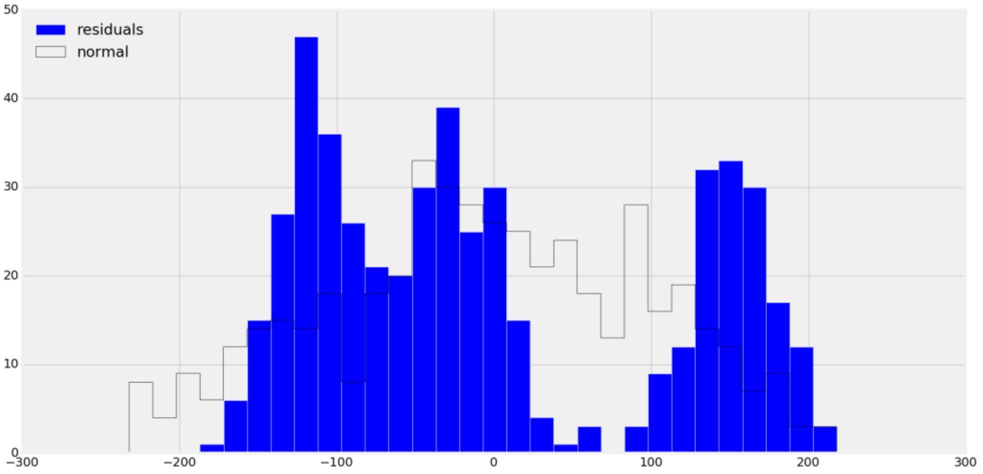
```
Linear regresssion model X1, Y
s_square= 12721.8208115
RMSE= 113.01731045
R_squared: 0.972079916834
Critical value for Chi-squared test: 441.375194177
Critical value: 42.5569678043
```

Using R_square, the fit appears to be fine but, the normal distribution of residuals is highly distorted (We can say that using chi_sq values). So this model should be rejected

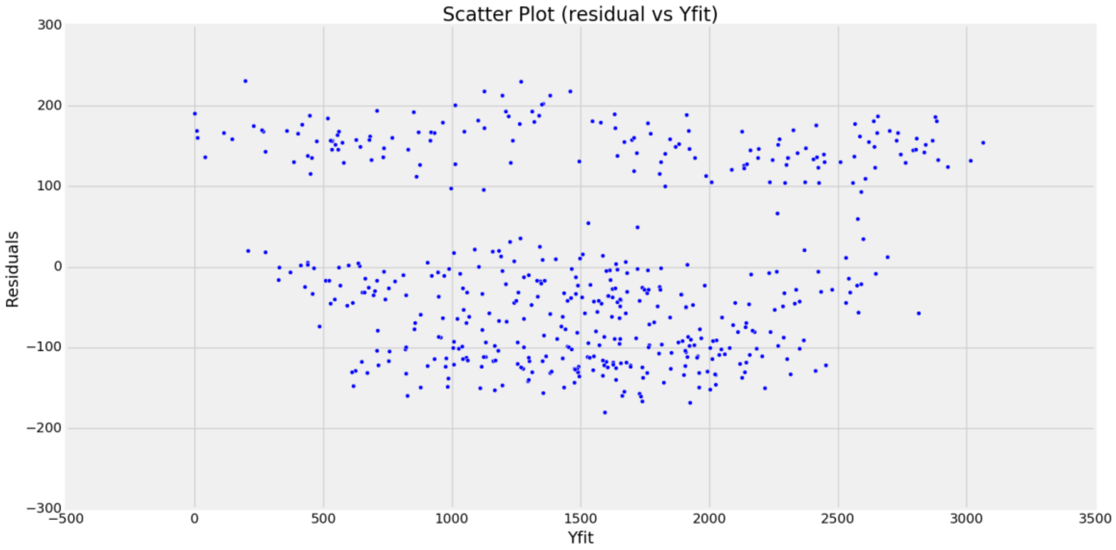
Let’s see the plots

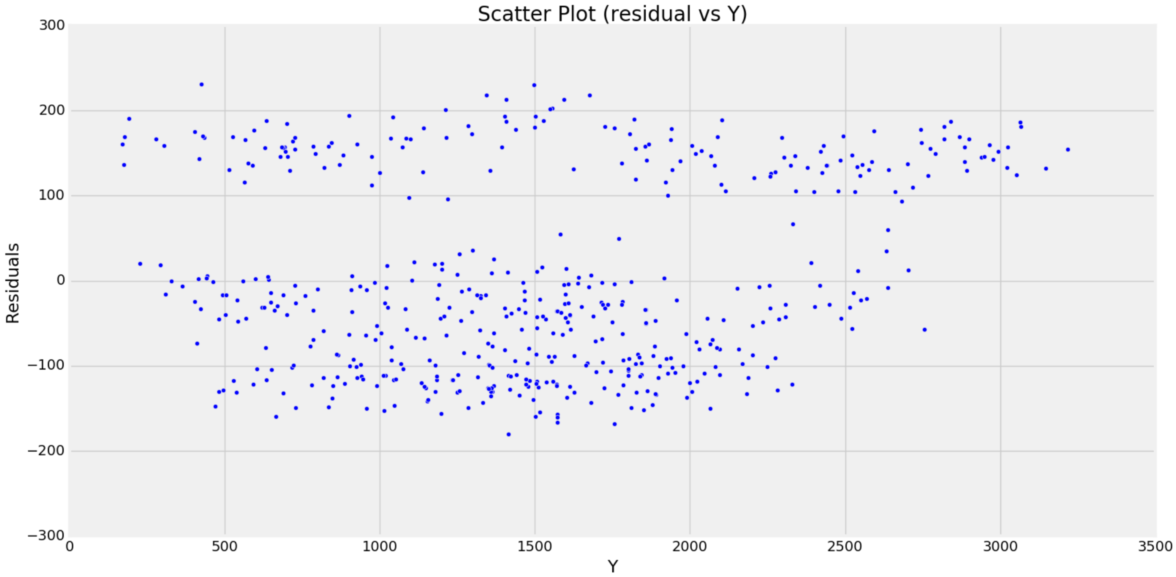


Clearly, Q-Q plot shows the bi-modal behavior and hence the residuals in our models do not represent the normal distribution. It can be further seen using histogram



Following are the scatter plots for the residuals





Scatter plots however, do not show any correlation trends

3. **Using X2,X5 (Correlation -0.003)**
   Summary

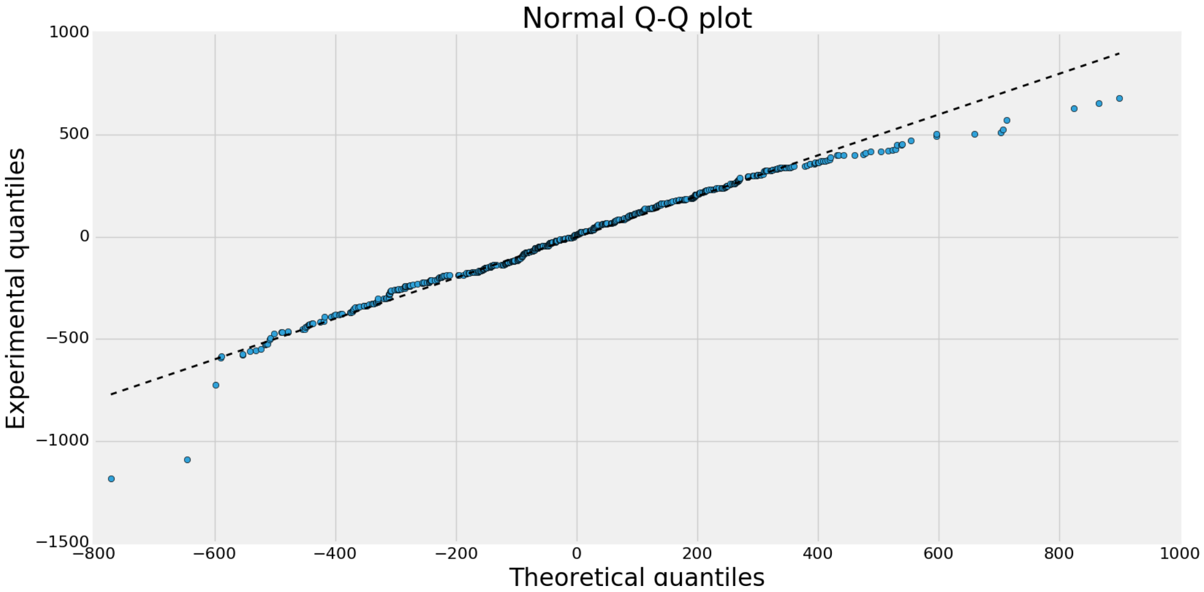
```
-------------------------Summary of Regression Analysis-------------------------

Formula: Y ~ <X2> + <X5> + <intercept>

Number of Observations:         500
Number of Degrees of Freedom:   3

R-squared:         0.8521
Adj R-squared:     0.8515

Rmse:            260.3641

F-stat (2, 497):  1431.8914, p-value:     0.0000

Degrees of Freedom: model 2, resid 497

-----------------------Summary of Estimated Coefficients------------------------
      Variable       Coef    Std Err     t-stat    p-value    CI 2.5%   CI 97.5%
--------------------------------------------------------------------------------
            X2   312.4168     6.9842      44.73     0.0000   298.7278   326.1058
            X5    97.5023     3.3011      29.54     0.0000    91.0320   103.9725
     intercept   118.1945    29.0988       4.06     0.0001    61.1607   175.2282
      ---------------------------------End of Summary-------------------------------
      --
```

Comments: R_squared values have reduced and hence it can be concluded that, this model does not give a good fit for the predictor when compared with previous models. Also, the standard error for coefficients has increased. So we will reject this model

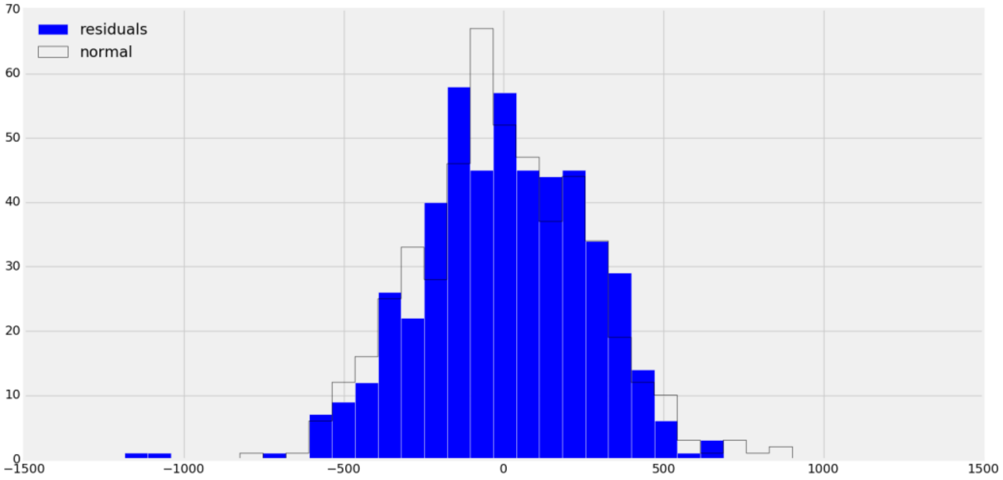
Let’s perform the residual analysis for this model

```
Linear regresssion model X1, Y
s_square= 67382.7162153
RMSE= 260.102536576
R_squared: 0.852117784979
Critical value for Chi-squared test: 43.563257628
Critical value: 42.5569678043
```

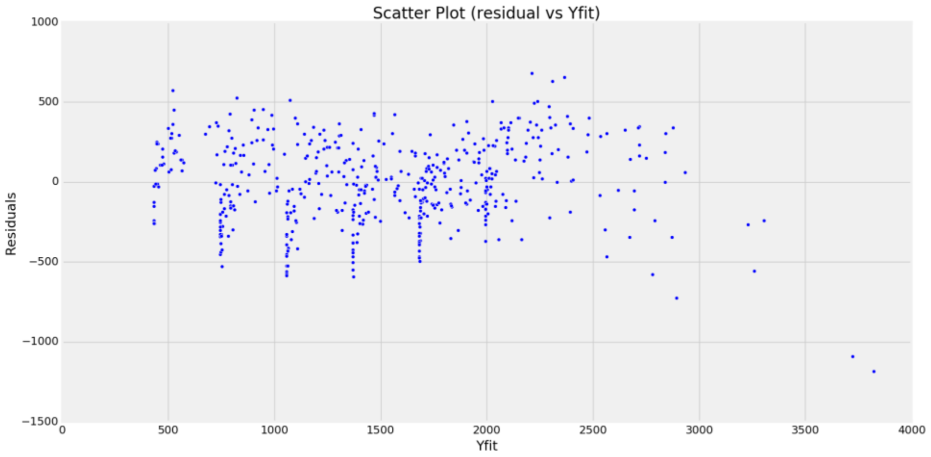
Comments: Using the chi_squared test, here we should reject the hypothesis for normal distribution of the residuals. But it is quite close to normal distribution hypothesis and can be seen in the plots too. But due to the above mentioned reasons, we will reject this hypothesis



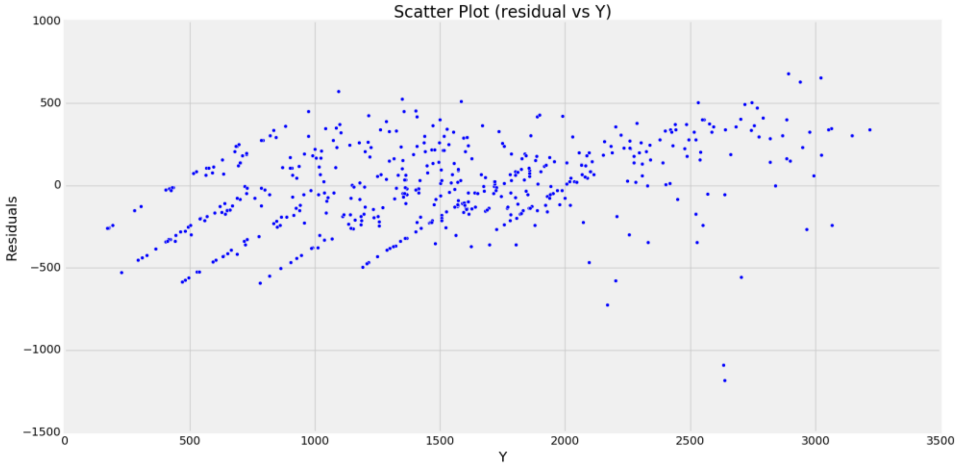
Q-Q plot has a long tail. Possibly, that is causing the deviation from normal behavior



The histogram shows skewness in the residual model



Scatter plot shows no correlation but more values are concentrated in the positive side.



4. X4,X5 (Correlation -0.005422)

Summary

```
-------------------------Summary of Regression Analysis-------------------------

Formula: Y ~ <X4> + <X5> + <intercept>

Number of Observations:         500
Number of Degrees of Freedom:   3

R-squared:         0.8770
Adj R-squared:     0.8765

Rmse:            237.4522

F-stat (2, 497):  1771.8196, p-value:     0.0000

Degrees of Freedom: model 2, resid 497
```

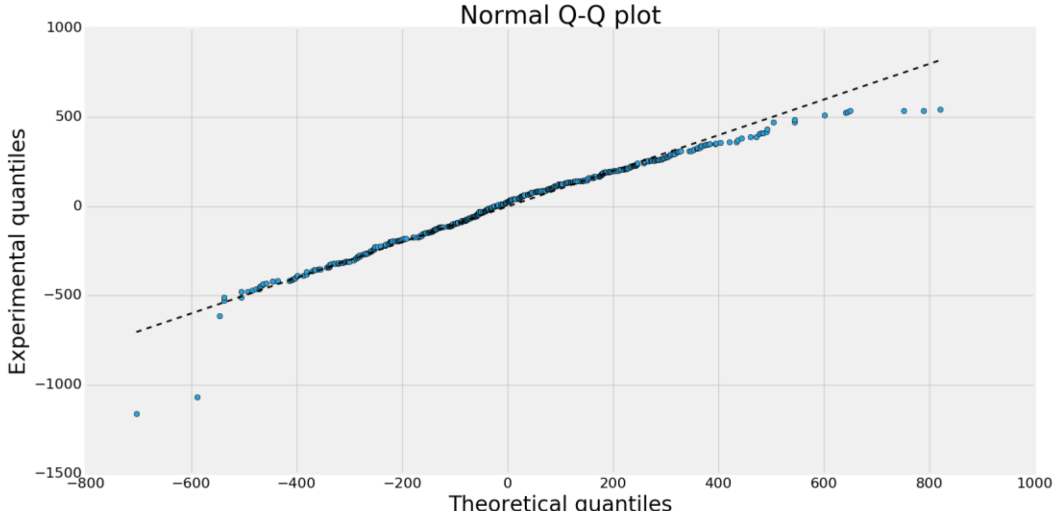
```
-----------------------Summary of Estimated Coefficients------------------------
      Variable       Coef    Std Err     t-stat    p-value    CI 2.5%   CI 97.5%
--------------------------------------------------------------------------------
            X4     4.4085     0.0881      50.06     0.0000     4.2359     4.5811
            X5    97.7841     3.0107      32.48     0.0000    91.8832   103.6850
     intercept   548.3237    19.3984      28.27     0.0000   510.3028   586.3447
      ---------------------------------End of Summary---------------------------------
      --
```

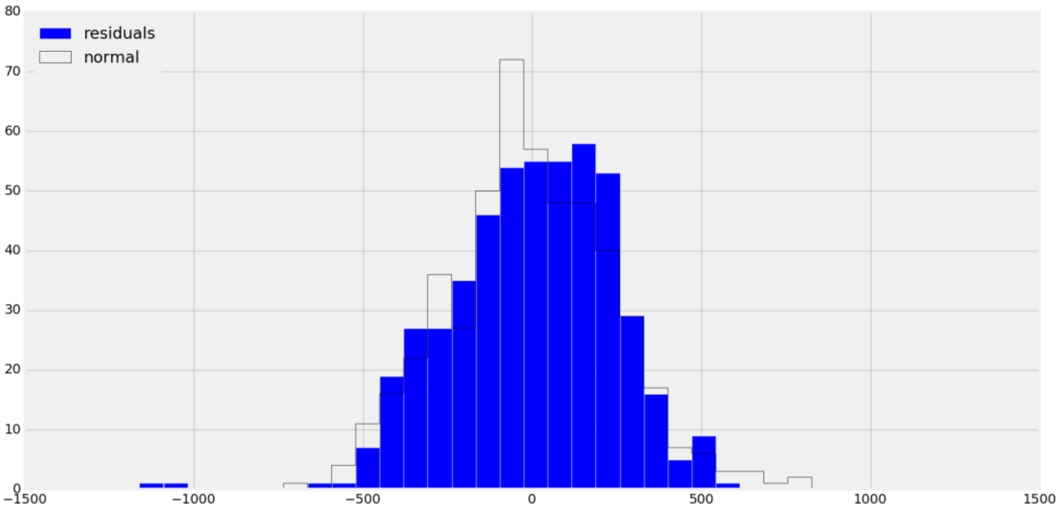
Comments: R_squared value is decreased. Also the standard error for X5 is still high. So we reject this model

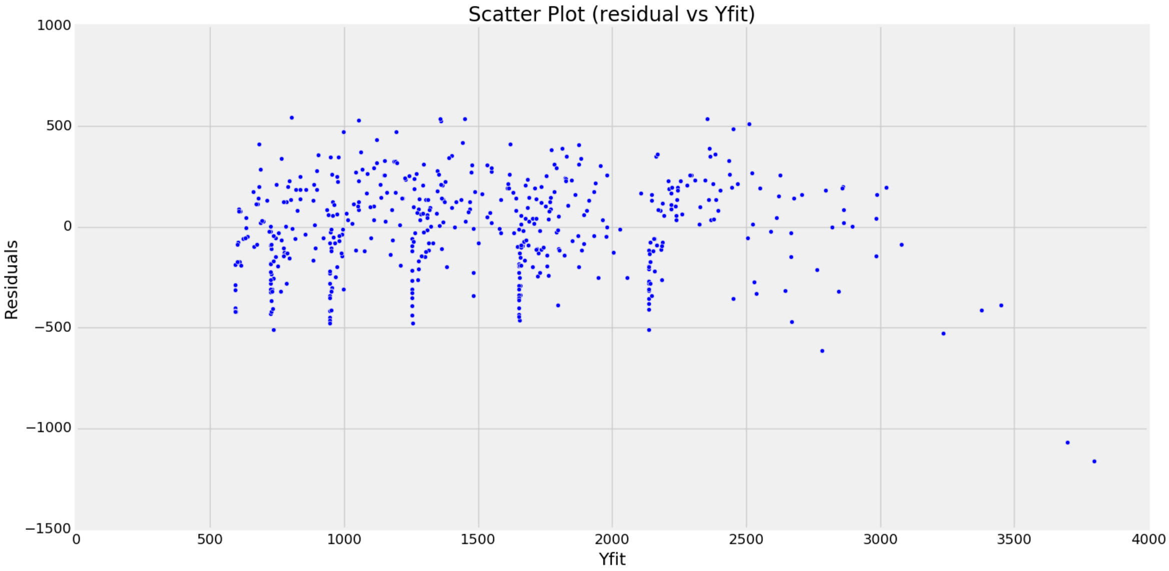
Let’s do the residual analysis

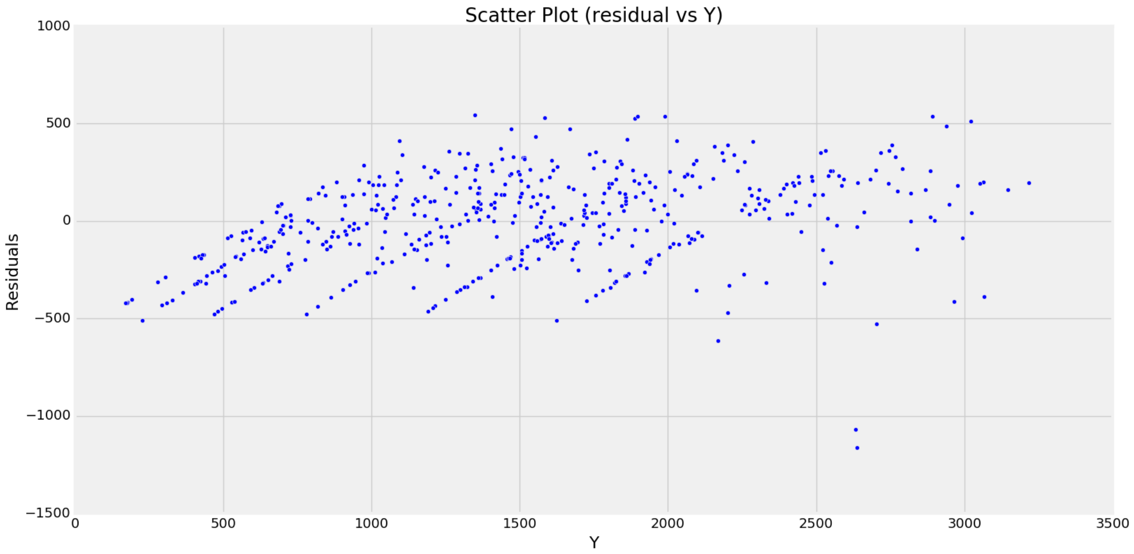
```
Linear regresssion model X4,X5 Y
s_square= 56045.2580913
RMSE= 237.213699961
R_squared: 0.876999661434
Critical value for Chi-squared test: 32.7067070007
Critical value: 42.5569678043
```

The residual analysis gives a match to a fairly normal distribution for the residuals. It can be seen from the plots as well




residuals
normal
80
70
60
50
40
30
20
10
0
−1500
−1000
−500
0
500
1000
1500


Scatter Plot (residual vs Yfit)
Residuals
1000
500
0
−500
−1000
−1500
0
500
1000
1500
2000
2500
3000
3500
4000
Yfit



5. X1,X4 (Correlation -0.030691)

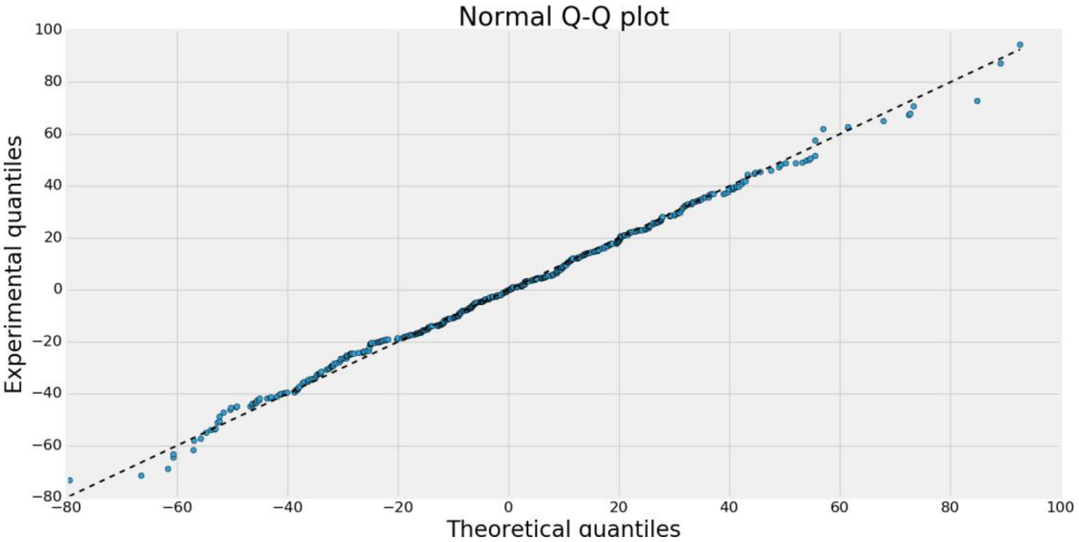
```
-------------------------Summary of Regression Analysis-------------------------

Formula: Y ~ <X1> + <X4> + <intercept>

Number of Observations:         500
Number of Degrees of Freedom:   3

R-squared:         0.9984
Adj R-squared:     0.9984

Rmse:             26.8136

F-stat (2, 497): 158190.0479, p-value:     0.0000

Degrees of Freedom: model 2, resid 497

-----------------------Summary of Estimated Coefficients------------------------
      Variable       Coef    Std Err     t-stat     p-value    CI 2.5%   CI 97.5%
--------------------------------------------------------------------------------
            X1     2.8856     0.0083     348.15      0.0000     2.8694     2.9019
            X4     4.4993     0.0099     452.26      0.0000     4.4798     4.5188
     intercept   -14.6123     3.1286      -4.67      0.0000   -20.7444    -8.4802
       ---------------------------------End of Summary-------------------------------
       --
```

Comments: This y model has a high R_squared value. Also, it has p-values for X1 and X4 within zone of significance and also standard erros for independent variables is highly reduced. F value is also high and over all p value is also less. Hence this is the best model for our data and **this model should be accepted as the predictor model**

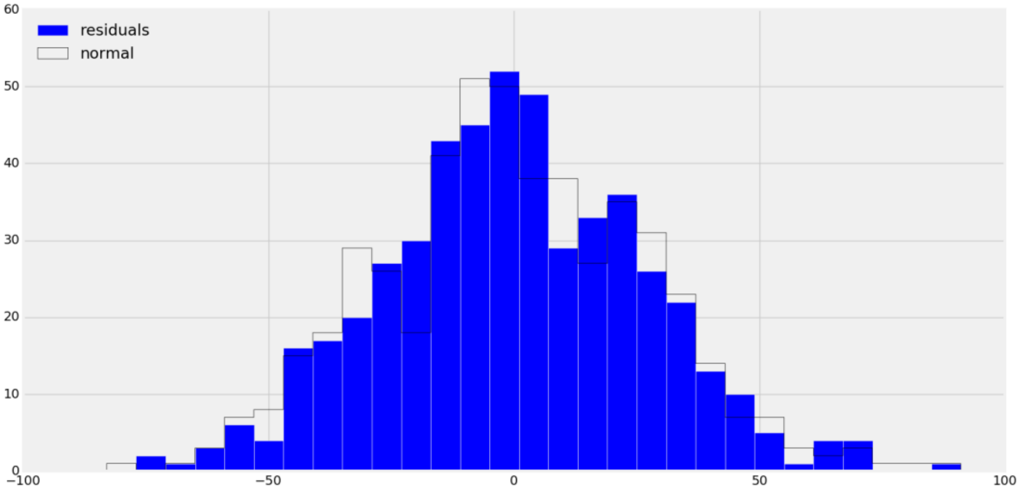
Let’s do the residual analysis for this model

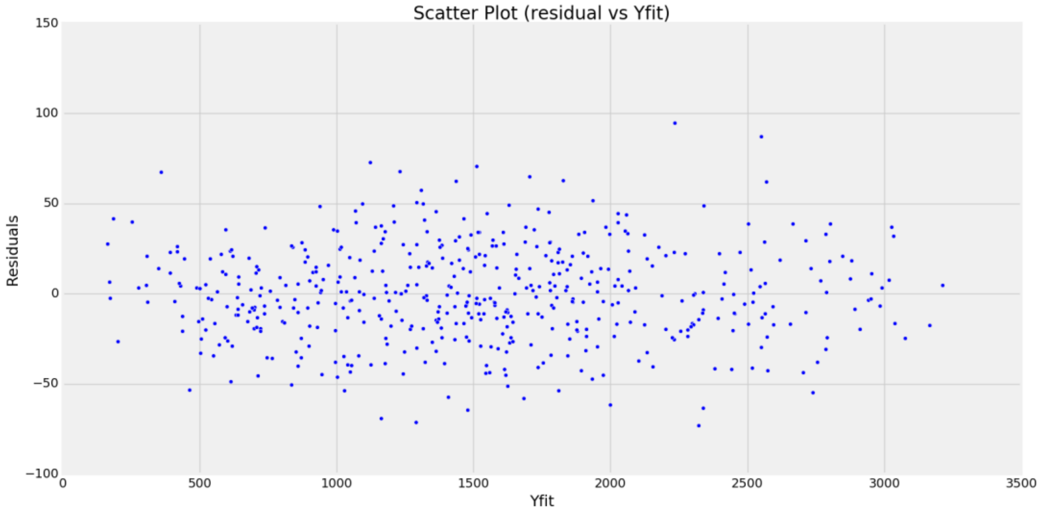
```
Linear regresssion model X1,X4, Y
s_square= 714.658034604
RMSE= 26.8136507361
R_squared: 0.998431567929
Critical value for Chi-squared test: 30.102949885
Critical value: 42.5569678043
```

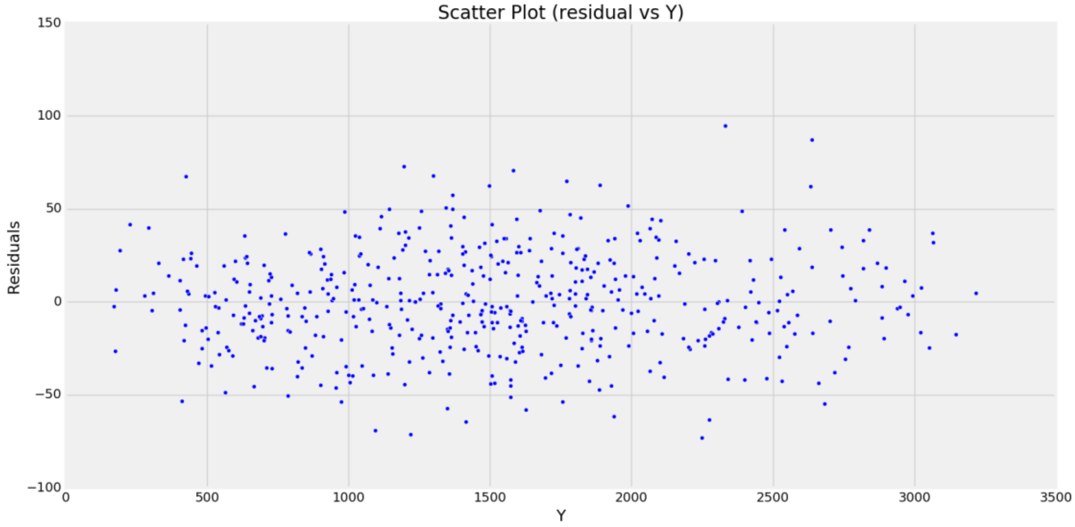
Clearly, the chi_squared value lies well below critical value and hence the hypothesis that the residuals have a normal distribution holds true. Also, this can be seen using the plots



From Q-Q plot and histograms it follows that residuals have a normal distribution







Also, the scatter plots for the residuals average out to zero with no correlation trends.

Hence, this is the best regression model for the given data (consists of predictor variables X1 and X4)