IOT Project 2: Regression Analysis

Ojas Barve - ovbarve@ncsu.edu

Task 1:

This task involved the basic preliminary analysis before carrying out any regression.

Task 1.1:

This involved the plotting of histogram of each independent variable and calculate their means and standard deviations(variance). Table 1 below gives the value of mean and variance of each variable. The Figure 1 below show the histogram of each variable plotted with bin size 30.

Table 1: Mean and variance v/s each variable

	Mean	Variance
X1	3.193837113333333	752.4794063480422
X2	3.546478960000001	705.7700312946481
X3	8.253569083333334	834.4016768926834
X4	9.810877666666666	649.2749440814695
X5	9.208723308	792.2815368819033
Y	1307.0356689999999	2535443.2559497557

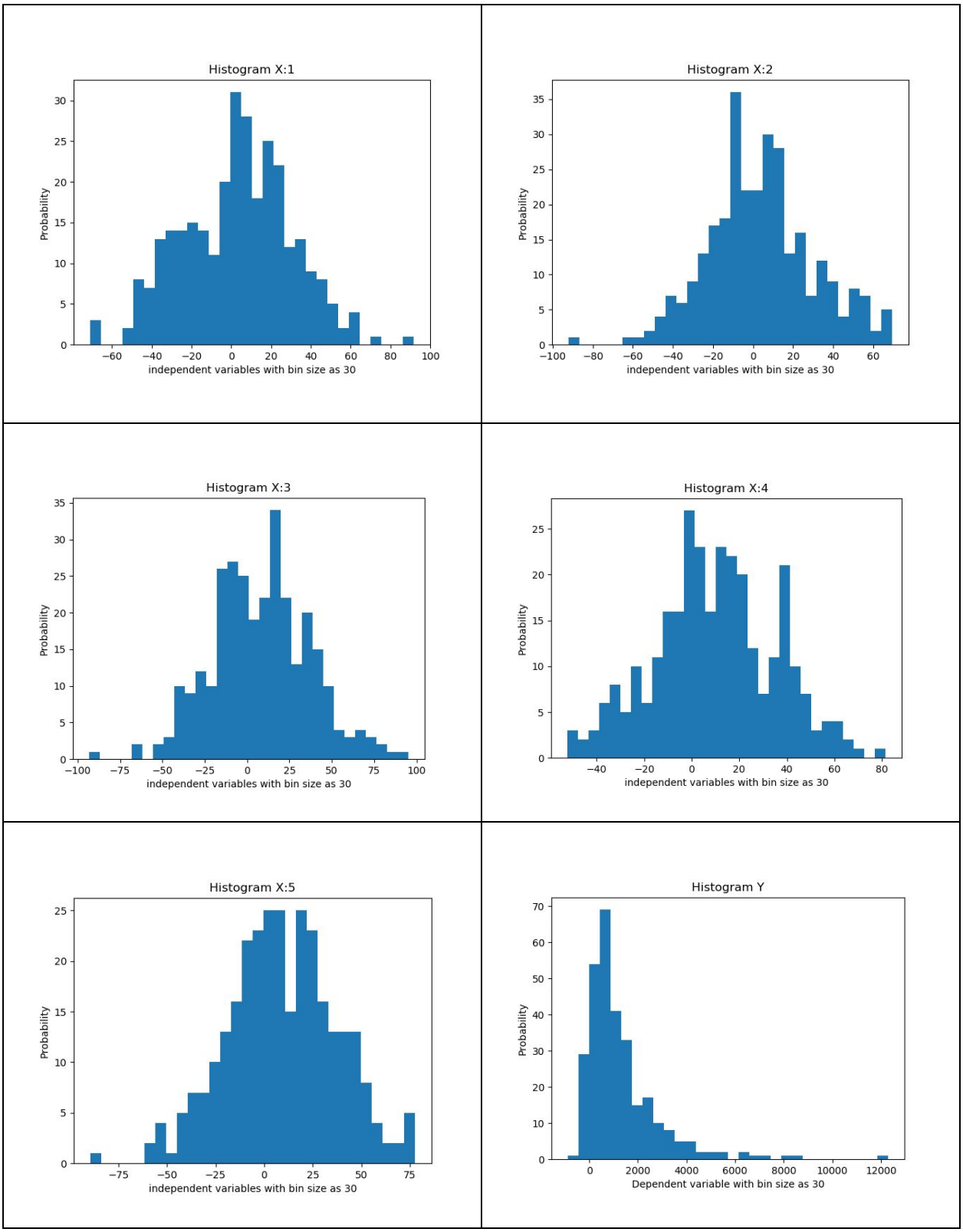


Figure 1. Plots of histogram of each independent variable followed the dependent variable on bottom right corner. Bin size is 30.

Task 1.2:

The value of Z scores was used to remove the outliers from the data set and all the different regression tasks done below were carried out with and without removing the outliers and not much difference or improvement was observed in the best fit obtained so the data given was quite consistent. To remove outliers the values beyond 3 standard deviations were dropped and the resulting shapes of the input array was :

Before (300, 6)

After (289, 6)

Task 1.3:

Table 2: Correlation Matrix

	X1	X2	X3	X4	X5	Y
X1	1.00	-0.001908	0.032828	0.079785	-0.033442	0.140241
X2		1.00	-0.039332	-0.091800	0.058964	0.033356
X3			1.00	-0.037926	0.023654	0.222028
X4				1.00	0.094189	0.117790
X5					1.00	0.263953
Y						1.00

From the correlation matrix we can conclude that the dependencies between the independent variable and the Y is not very high as overall for each combination the correlation is very less. The variables X3 and X5 ahave roughly similar correlation coefficients with Y followed by X1 and X4. From the table we can do a preliminary analysis and conclude that the independent variable X2 is most likely not going to affect the Y values and can be dropped during the regression fitting. All the other variables have more or less equal effect on Y.

Also below we (Figure 2) plot the correlation matrix scatter plot for each variable.

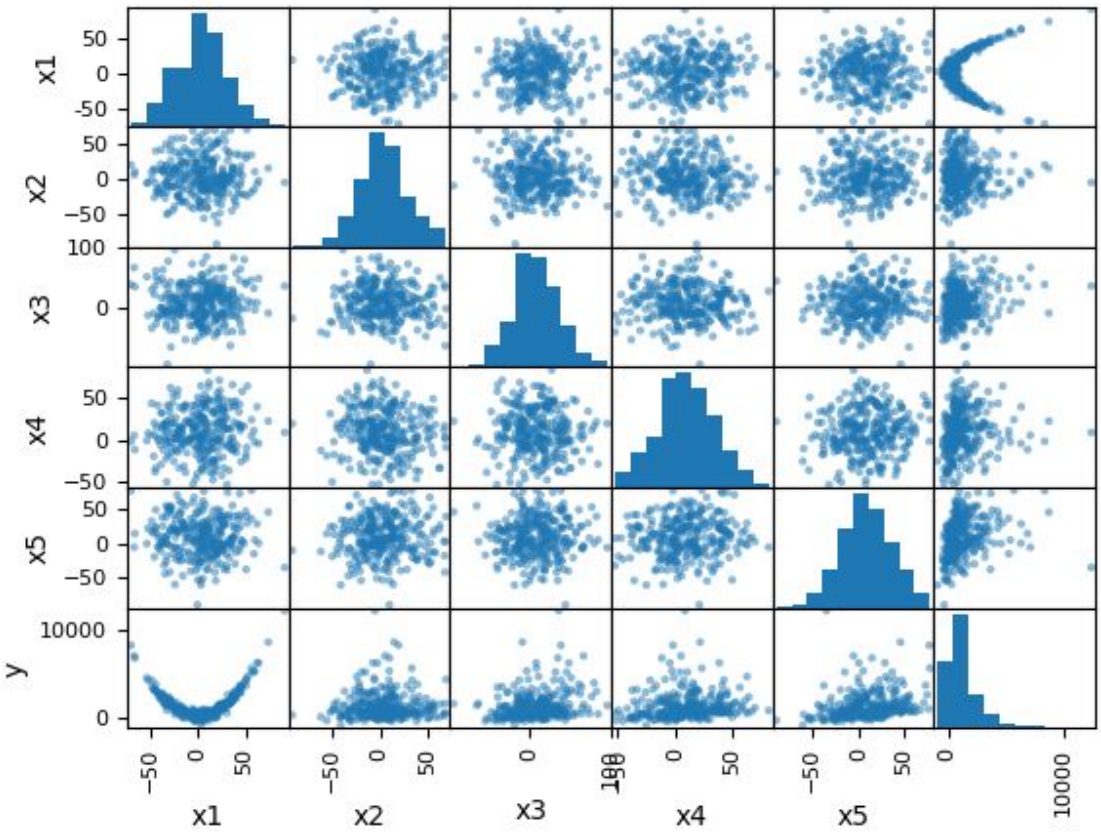


Figure 2: Scatter plot of all variables

As we can see from the scatter plot of the independent variables with respect to the dependent variable we can conclude that the relationship between X1 and Y seems to be non linear.

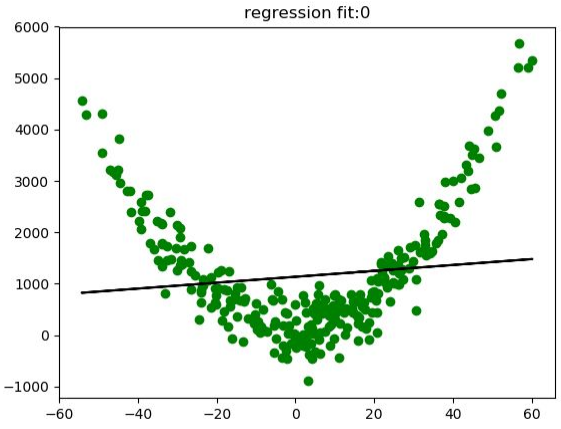
From here we can conclude that we can use polynomial regression to fit the data for X1 and Y as the relationship seems quadratic from looking at the trends in the scatter plot.

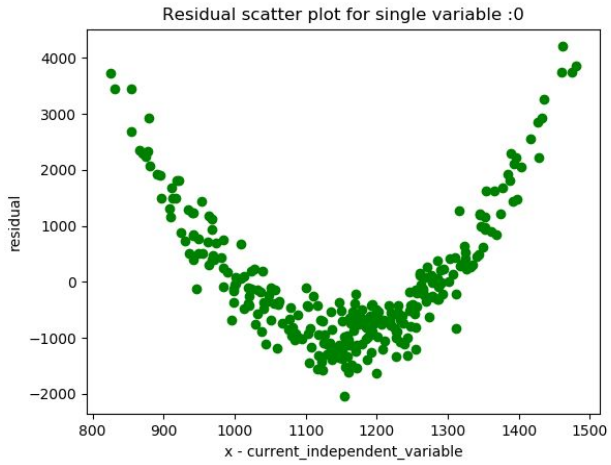
Task 1.4:

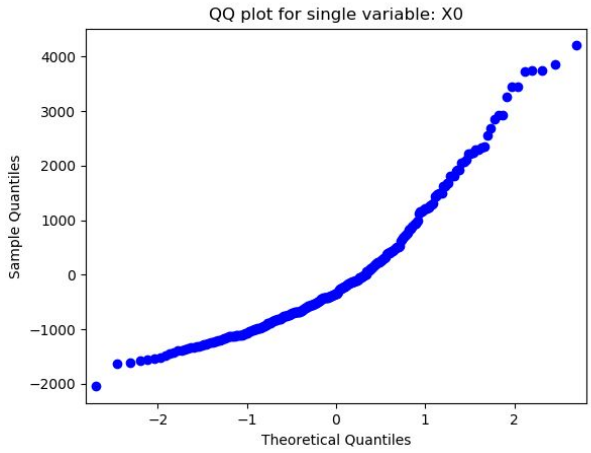
The preliminary analysis of the data shows that the the independent variable X2 is the least related to the dependent variable Y. Also relation between X1 and Y seems to be quadratic or a higher power and a polynomial regression might be more better to fit.

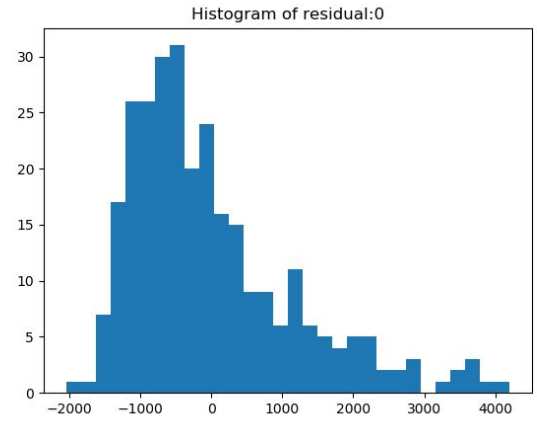
Task 2 Linear regression:

With Variable x1: y = a0 + a1*X1









	Coefficient values	P values
const	1135.9939	0.00
X1	5.7332	0.038

R-Squared	0.015

F-statistic	4.364
Prob (F-statistic):	0.0376
Chi square test: p-value	1.0075962558898167e-13
Variance of residual	1398362.3040994583

Comment:

We see that the regression fit line obtained is very poor. Also we see the residual scatter plot has a non linear trends. Also when we run chi square test on the residuals and do a null hypothesis on the value obtained we see that the p value is almost zero. This concludes that the null hypothesis fails and hence we can say that the residuals do not follow a normal distribution. Also we see that the qq plot which has been obtained is non linear. Hence all the tests on residuals have failed.

Also we can see from the P- values for the coefficients that the null hypothesis is not valid. That is the alternative hypothesis is accepted. This proves that the coefficients are not zero and hence we can say that the independent variable does have correlation with the dependent variable. But the fit obtained in this case is bad which leads us to try polynomial regression.

We can also see that the R Squared value is very less. This shows that the regression fit is not good. Usually when the R value is close to 100% it is a good fit while if it is near 0% then it is a bad fit. In this case it is close to 0.

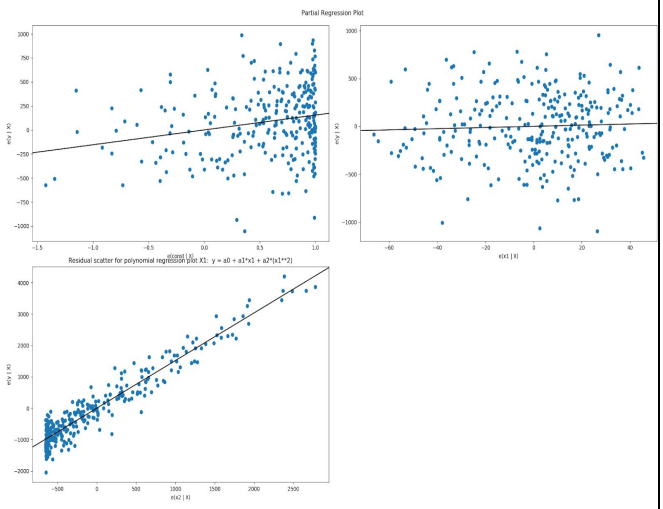
The value of Prob(F) is the probability that the null hypothesis for the full model is true (i.e., that all of the regression coefficients are zero). For example, if Prob(F) has a value of 0.01000 then there is 1 chance in 100 that all of the regression parameters are zero. This low a value would imply that at least some of the regression parameters are nonzero and that the regression equation does have some validity in fitting the data (i.e., the independent variables are not purely random with respect to the dependent variable). Hence here the probability is low hence we can conclude that the regression coefficients do have a significance but the regression fit is not good as both the chi square test and the qq plot test have failed.

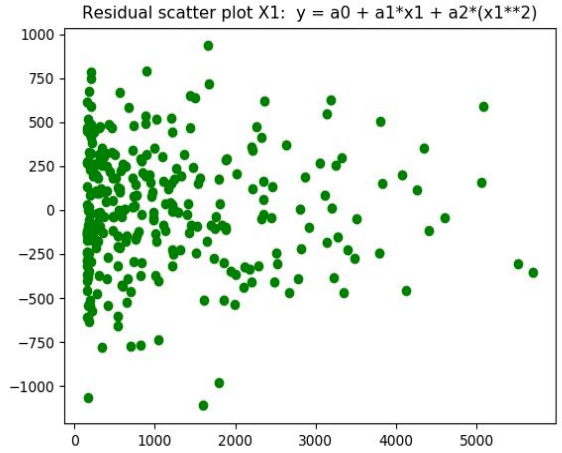
Task 2 Polynomial regression:

Polynomial Linear regression outputs :

With Variable x1: y = a0 + a1*X1 + a2 * (X12)**

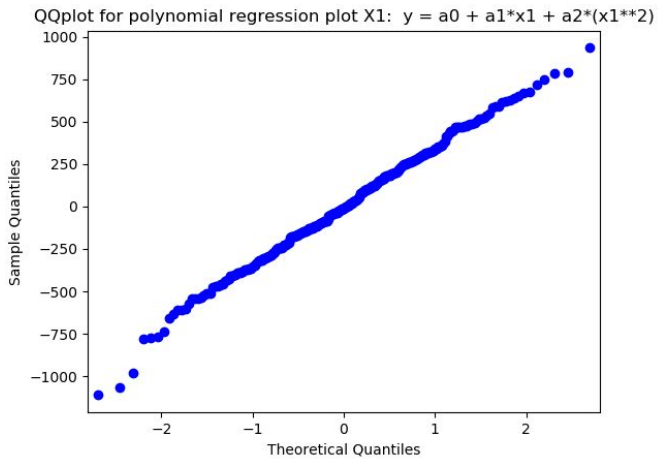
Regression fit line on all independent variables

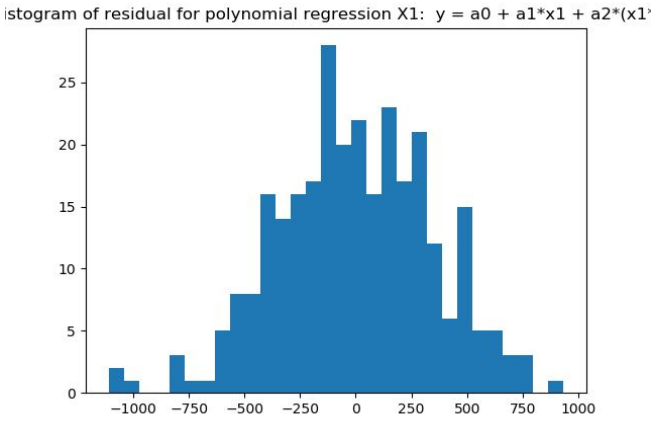




Residual scatter plot

Qqplot





Residual histogram

	Coefficient values	P values
const	154.2509	0.000
X1	0.6146	0.450
X2	1.5192	0.000

R-Squared	0.915
F-statistic	1542.
Prob (F-statistic):	6.25e-154
Chi square test: p-value	0.5761257471962815
variance of residuals is :	120452.93919540454

Comment:

We see that the regression fit line obtained is better in this case. Also we see the residual scatter plot has a no trends in this case so we can conclude that the scatter plot of residuals is random. Also when we run chi square test on the residuals and do a null hypothesis on the value obtained we see that the p value is 0.576. This concludes that the null hypothesis has passed and hence we can say that the residuals are following a normal distribution. Also we see that the qq plot which has been obtained is linear. Hence all the tests on residuals have passed.

Also we can see from the P- values for the coefficients that the null hypothesis is not valid except for x1. That is the alternative hypothesis is accepted for x1^2 and the constant term. This proves that the coefficients are non zero and hence we can say that the independent variable when squared does have correlation with the dependent variable. But the x1 term can be dropped.

We can also see that the R Squared value is 0.915. This shows that the regression fit is good. Usually when the R value is close to 100% it is a good fit while if it is near 0% then it is a bad fit. In this case it is close to 100 hence a good fit.

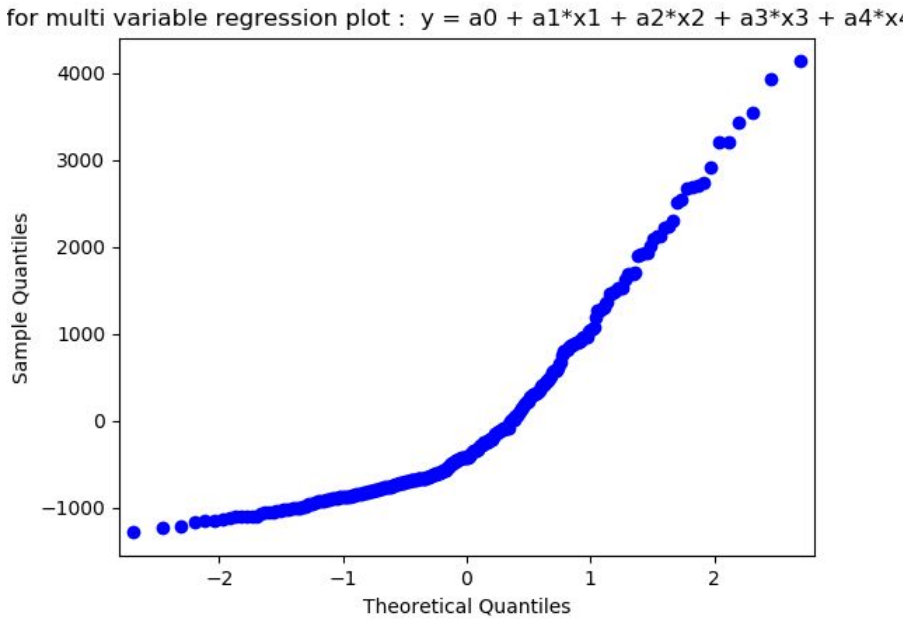
The value of Prob(F) is the probability that the null hypothesis for the full model is true

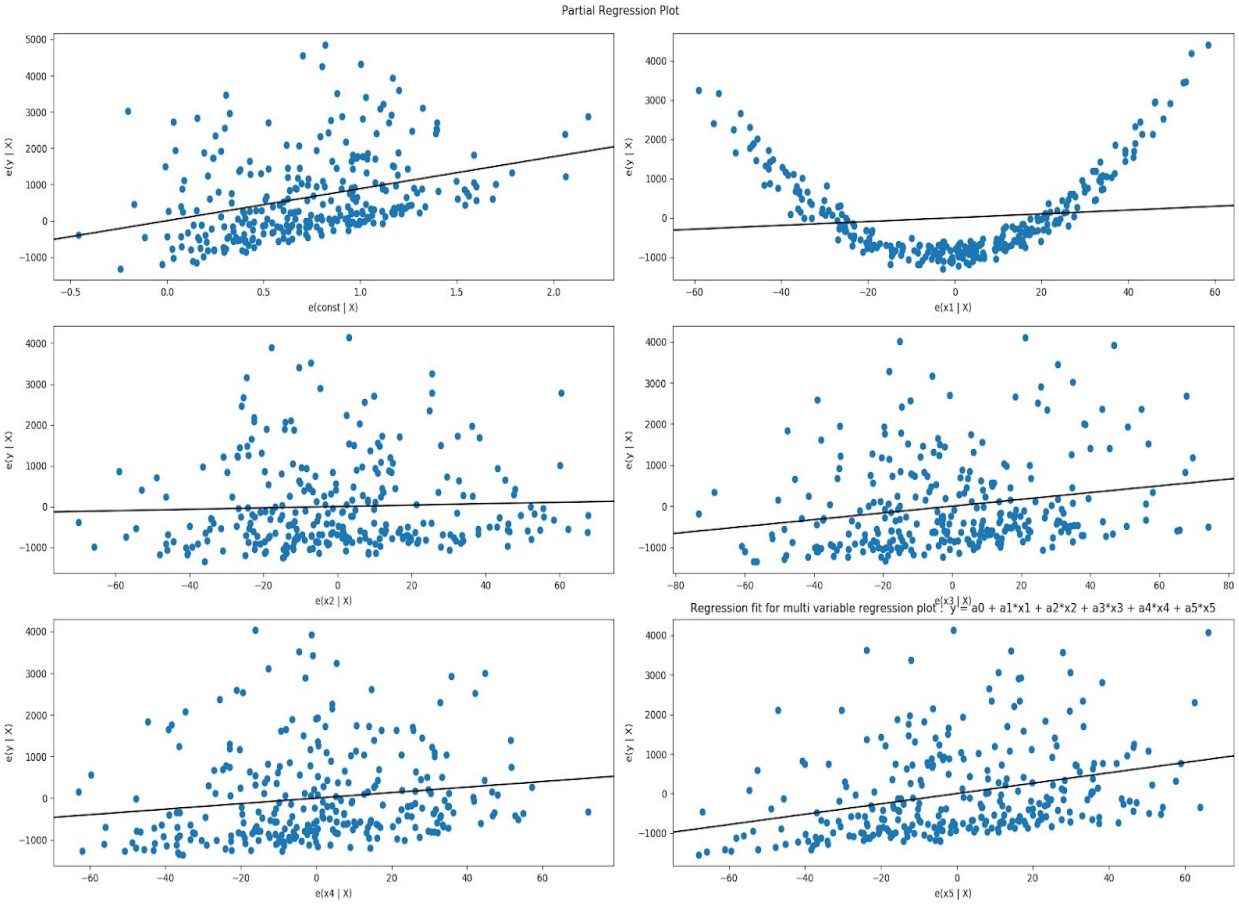
(i.e., that all of the regression coefficients are zero). For example, if Prob(F) has a value of 0.01000 then there is 1 chance in 100 that all of the regression parameters are zero. Thus low a value would imply that at least some of the regression parameters are nonzero and that the regression equation does have some validity in fitting the data (i.e., the independent variables are not purely random with respect to the dependent variable). Hence here the probability is low hence we can conclude that the regression coefficients do have a significance and the regression fit is good as both the chi square test and the qq plot test have passed.

Task 2 Multi Variable regression:

Multivariable Linear regression outputs :

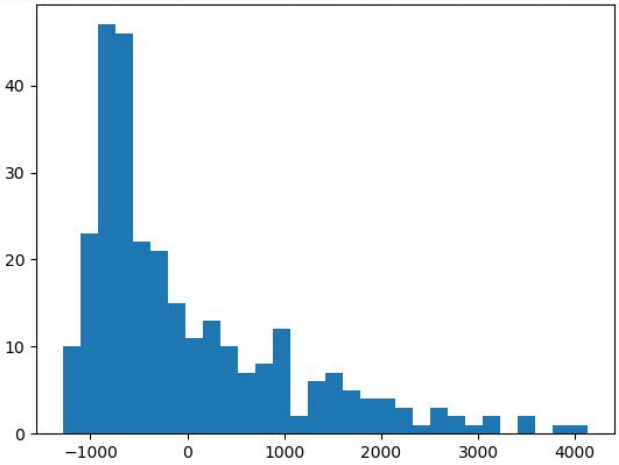
With Variable: y = a0 + a1*x1 + a2*x2 + a3*x3 + a4*x4 + a5*x5



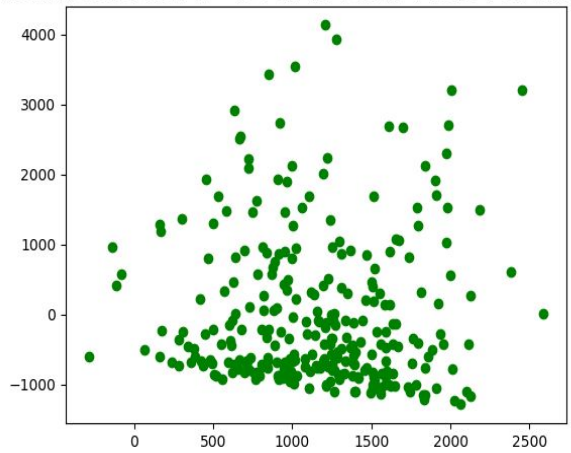


Regression fit for multi variable regression plot : y = a0 + a1*x1 + a2*x2 + a3*x3 + a4*x4 + a5*x5

of residual for polynomial regression : y = a0 + a1*x1 + a2*x2 + a3*x3 + a



Residual scatter plot : y = a0 + a1*x1 + a2*x2 + a3*x3 + a4*x4 + a5*x5



	Coefficient value	P value
constant	882.1428	0.000
x1	4.8624	0.056
x2	1.7279	0.484
x3	8.2400	0.000
x4	6.6252	0.009
x5	13.0604	0.000

R-squared	0.178
F-statistic	12.24
Prob (F-statistic):	9.42e-11
Chi square test: p-value	1.725738838908708e-16
variance of residuals is :	1167226.571237647

Comment:

We see that the regression fit line obtained is not good in this case. We see the residual scatter plot has a no trends in this case so we can conclude that the scatter plot of residuals is random. When we run chi square test on the residuals and do a null hypothesis on the value obtained we see that the p value is 0.00. This concludes that the null hypothesis has failed and hence we can say that the residuals are not following a normal distribution. Also we see that the qq plot which has been obtained is non linear. Hence all the tests on residuals have have failed and we conclude that the regression fit is poor.

Here we observe that the null hypothesis for the regression coefficients has passed only in the case of x2 (p- value is 0.484). Hence we can conclude that our preliminary hypothesis for correlation coefficients was correct as from here we can say that the variable x2 is not related to y. For all other variables we can say that the alternative

hypothesis has passed and all other variables do have some correlation with the dependent variable y.

We can also see that the R Squared value is 0.178. This shows that the regression fit is bad. Usually when the R value is close to 100% it is a good fit while if it is near 0% then it is a bad fit. In this case it is close to 0% hence a bad fit.

The value of Prob(F) is the probability that the null hypothesis for the full model is true (i.e., that all of the regression coefficients are zero). For example, if Prob(F) has a value of 0.01000 then there is 1 chance in 100 that all of the regression parameters are zero. Thus low a value would imply that at least some of the regression parameters are nonzero and that the regression equation does have some validity in fitting the data (i.e., the independent variables are not purely random with respect to the dependent variable). Hence here the probability is low hence we can conclude that the regression coefficients do have a significance. But from other statics we can say that the overall regression is a bad fit. This is because the qq plot and chi square tests have failed. And also R squared value is also low.

Also after removing the variable x2:

OLS Regression Results

==========

Dep. Variable:	y	R-squared:	0.176
Model:	OLS	Adj. R-squared:	0.165
Method:	Least Squares	F-statistic:	15.20
Date:	Sat, 27 Oct 2018	Prob (F-statistic):	2.82e-11
Time:	23:08:38	Log-Likelihood:	-2429.0
No. Observations:	289	AIC:	4868.
Df Residuals:	284	BIC:	4886.
Df Model:	4		
Covariance Type:	nonrobust		

==========

==

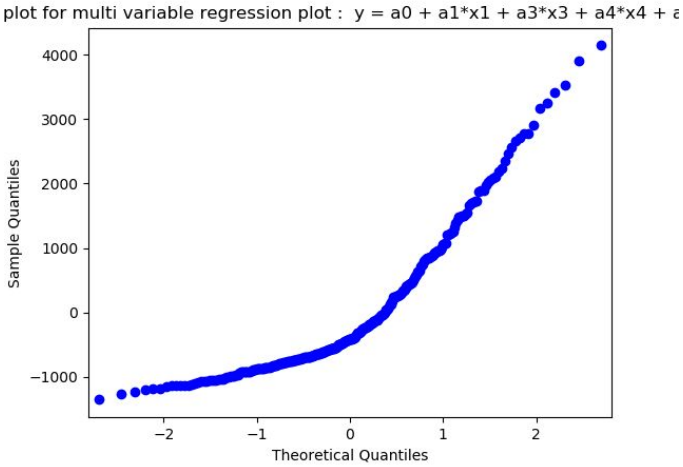
	coef	std err	t	P>|t|	[0.025	0.975]
const	890.2962	73.550	12.105	0.000	745.524	1035.068
x1	4.9019	2.534	1.935	0.054	-0.085	9.889
x2	8.1533	2.326	3.505	0.001	3.574	12.733
x3	6.4572	2.511	2.571	0.011	1.514	11.401
x4	13.1630	2.363	5.571	0.000	8.512	17.814

==

Omnibus:	71.806	Durbin-Watson:	1.927
Prob(Omnibus):	0.000	Jarque-Bera (JB):	123.583
Skew:	1.399	Prob(JB):	1.46e-27
Kurtosis:	4.561	Cond. No.	37.4

==

Here also we see that the regression fit is poor and the corresponding qq plots and chi square tests produce bad results on residuals.



Chis square test p - value = 2.556298649531486e-16

Here we see that chi square test fails and regression fit is bad.

Final Comment:

From the regression fits we can conclude that the variable x4 is not related to the dependent variable y. Also there are non linearities in the data and we can say that a quadratic or a higher order polynomial might be used as a better fitting criteria. Hence when we fit a quadratic model we see a very good fit compared to linear or multivariable regression.

For other regression models the code can be run and modified to obtain different regression fits.