MULTIVARIATE REGRESSION PROJECT REPORT Sl-422

Submitted by:
Sonu Gupta 22N0062
Deepak Kumar 22N0058
Atul Verma 22N0052

Supervised by:
Prof. Siuli Mukhopadhyay
Associate Professor
Department of Mathematics
IIT Bombay

Team 13

- Sonu Gupta (22N0062)
- Deepak Kumar (22N0058)
- Atul Verma (22N0052)

Regression Analysis Final Project:

The goal of the final project is to apply what we have learned in this course to conduct a statistical analysis.

This project is a part of our coursework - Applied Regression Analysis. In this project, our aim was to find the relationship between Independent and dependent variable.

Problem statement:

In this project of simple linear regression analysis, we are trying to determine the relationship between one response variable (**Y**) and predictor variables (X_1, X_2, X_3, …..X_{13}, X_{14}, X_{15}). We want to determine how the different values of all the predictor variables affect the value of the response variable. And among all predictor variable, which variable shows the most linear relationship with the response.

Data Description:

This data has 17 columns and 440 rows in such way that

- 9 Integer data type (one is id column, Y column and rest are X_i {where 1<= i <= 15}) columns
- 6 Float type data columns
- 2 Object type data columns

Here Actual value of dependent variable **(Y)** is in Integer.

Data Cleaning:

First we have 'id' column in our data which have unique value in each row which is just counting the row and not necessary to have it so we drop that columns.

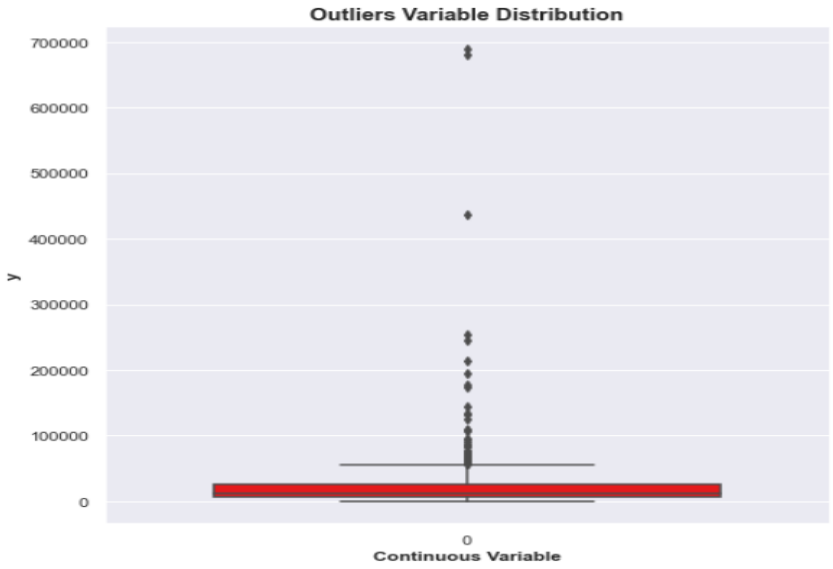
Also we have two object data type columns one is X_1 which has 373 unique name in 440 rows and second is X_2 which has 48 unique name in 440 rows which is not served it's purpose in model fitting even if we add in our analysis it will increase variable which might be led to overfitting our data and also it arise computational problem so it's better to drop this columns .

Data have no null value in each columns also there is no duplicates rows.

So after dropping the columns we have X_3, X_4, ……X_{15} , y columns in our data.

Exploratory Data Analysis:

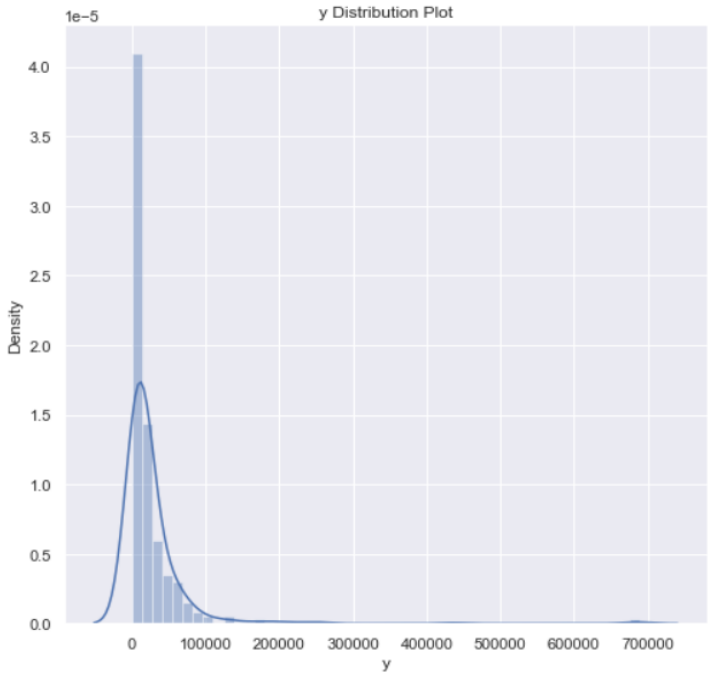
1. **Boxplot of Y variable**



Insights:

- There are some y values above 300000 which can be termed as outliers. But we have small data so removing outliers will be significant to statistically but not practically because we want that our data would work nicely on training data or new data too. So we see first rows should be remove or not.

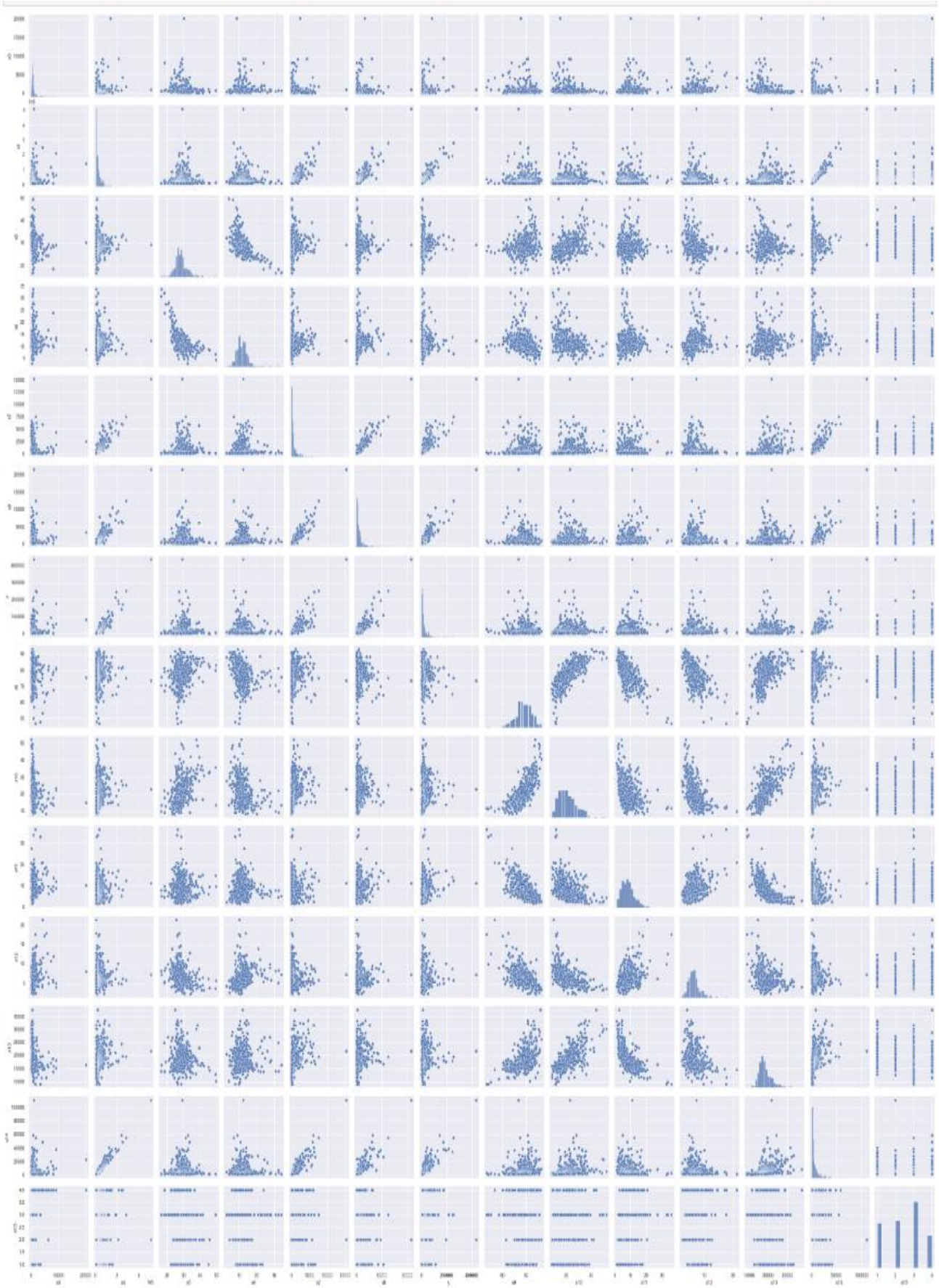
2. **Visualizing the distribution of y**



Insights:

- Here we can say our response variable is normally distributed and positively skewed. As we see that our maximum response variable is less than 100000 after that their long right tail exist which is due to outliers. But we try to keep 99.7 percentage data because we have small data set and it will not good idea for practically interpretation.
- Above 99.7 percentile which is 603608.49 we drop the outliers from data.

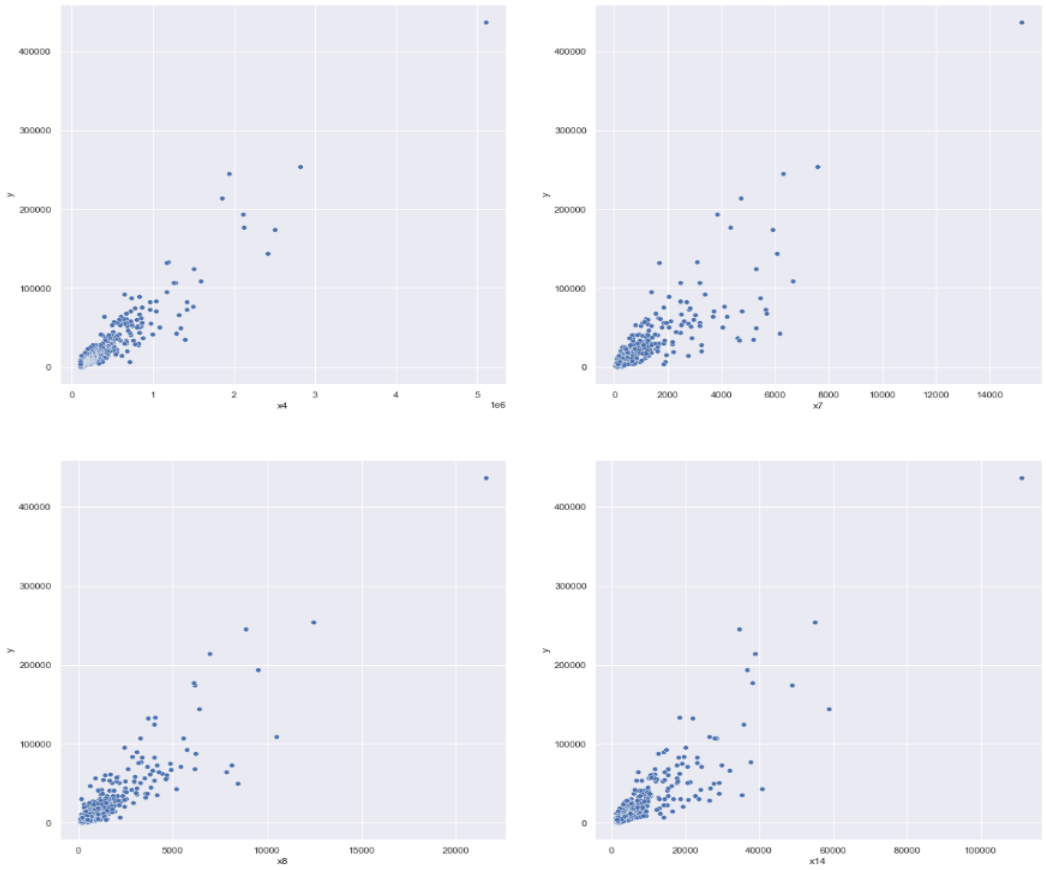
3. Pairplot of all the numeric variables



Insights:

- By Pair plot we visualise the correlation with independent variable and depended variable also simultaneously between each independent variable.
- y seems to have a positive correlation with X_4, X_7, X_8 and X_{14}.
- Pair of any two of X_4 , X_7, X_8 and X_{14} are highly positive correlated.
- X_{13} is positive correlation with X_{10} and X_9.

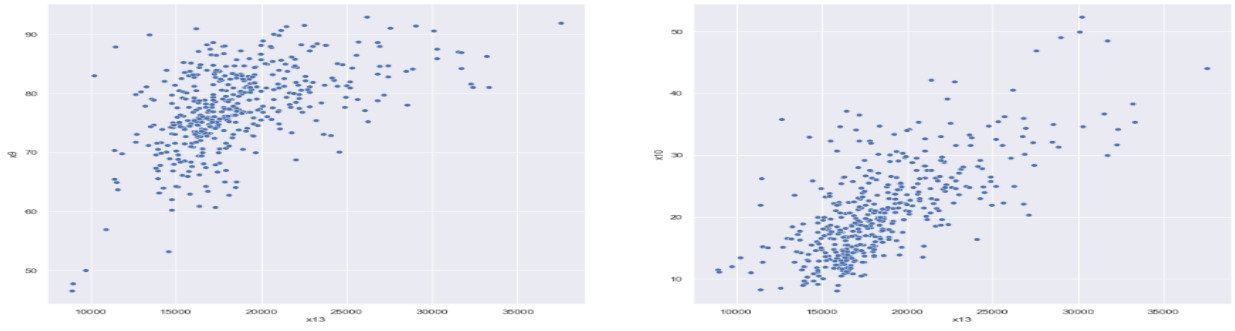
4. Scatterplot with y variable and X_4 , X_7, X_8 and X_{14} respectively



Insights:

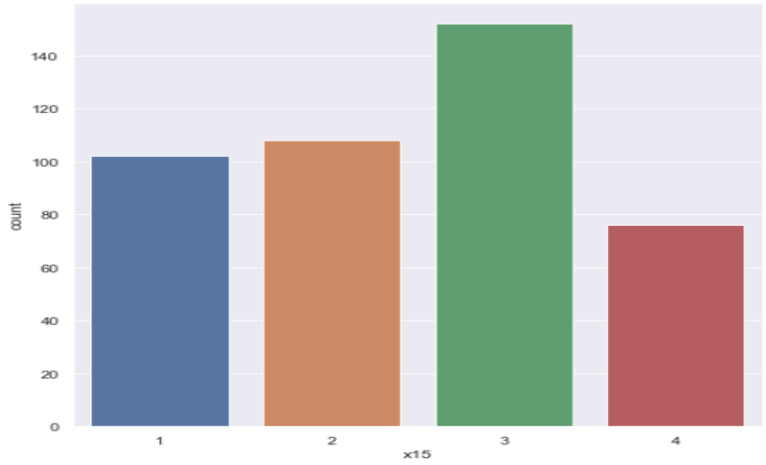
- Here, we can clearly see that Sub Scatterplot with y variable and X_4 , X_7, X_8 and X_{14} respectively have high correlation with y variable.

5. Multicollinearity between X_{13} and X_9, X_{10} respectively



- Here, We can remove the highly correlated independent variables x4, x7 and, x8.
- We can also remove x10 and x9. They are highly correlated with x13.
- But we try to not remove any independent variable until we have strongly significant reason to remove it because our data set is small.

6. **Bar-plot of variable X_{15}**



- Here, we create bar-plot of X_{15} because X_{15} is a discreate column.
- Here, Count of 3 maximum and count of 4 is minimum.

7. **Significant variables after Visualization**
   - Here Significant variable will be X_3, X_4, X_5 , X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}, X_{14}, X_{15} and y

Data Preprocessing:

So first some row of independent and dependent variable is look like below table

	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	y
1	946	5105067	29.2000	12.4000	15153	21550	73.4000	22.8000	11.1000	7.2000	21729	110928	2	436936
2	1729	2818199	31.3000	7.1000	7553	12449	74.9000	25.4000	12.5000	5.7000	19517	55003	3	253526
3	4205	2498016	33.5000	10.9000	5905	6179	81.9000	25.3000	8.1000	6.1000	19588	48931	4	173821
4	790	2410556	32.6000	9.2000	6062	6369	81.2000	27.8000	5.2000	4.8000	24400	58818	4	144524
6	9204	2122101	29.2000	12.5000	4320	6104	81.5000	22.1000	8.8000	4.9000	18042	38287	4	177593

Splitting the Data into Training and Testing Sets

- Here, we split data in 70/30 ratio mean we divide our data randomly in such way that 70 percentage data use for training data and 30 percentage data use for testing after model fitting. Few columns of training data is look like below

	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	y
153	521	317471	27.8000	9.3000	473	1263	77.7000	21.4000	5.0000	7.5000	21684	6884	2	16721
262	868	178386	25.8000	13.8000	600	1330	81.8000	22.4000	7.2000	4.8000	19601	3497	2	11929
142	771	349660	28.5000	11.5000	1510	2785	79.0000	23.5000	10.5000	5.8000	18225	6373	3	42404
219	399	217399	26.9000	10.5000	950	1592	75.8000	19.3000	12.5000	7.0000	17744	3858	3	18586
119	1000	403662	28.5000	11.4000	930	1840	82.4000	22.2000	8.3000	4.5000	19276	7781	2	34071

Rescaling the Features

For Simple Linear Regression, scaling doesn't impact model. But here some variable have large range compare to small range variable So it is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. Having features on a similar scale will help the gradient descent converge more quickly towards the minima. There are two common ways of rescaling:

1. Min-Max scaling
2. Standardisation (mean-0, sigma-1)

Here, we will use Standardisation Scaling.

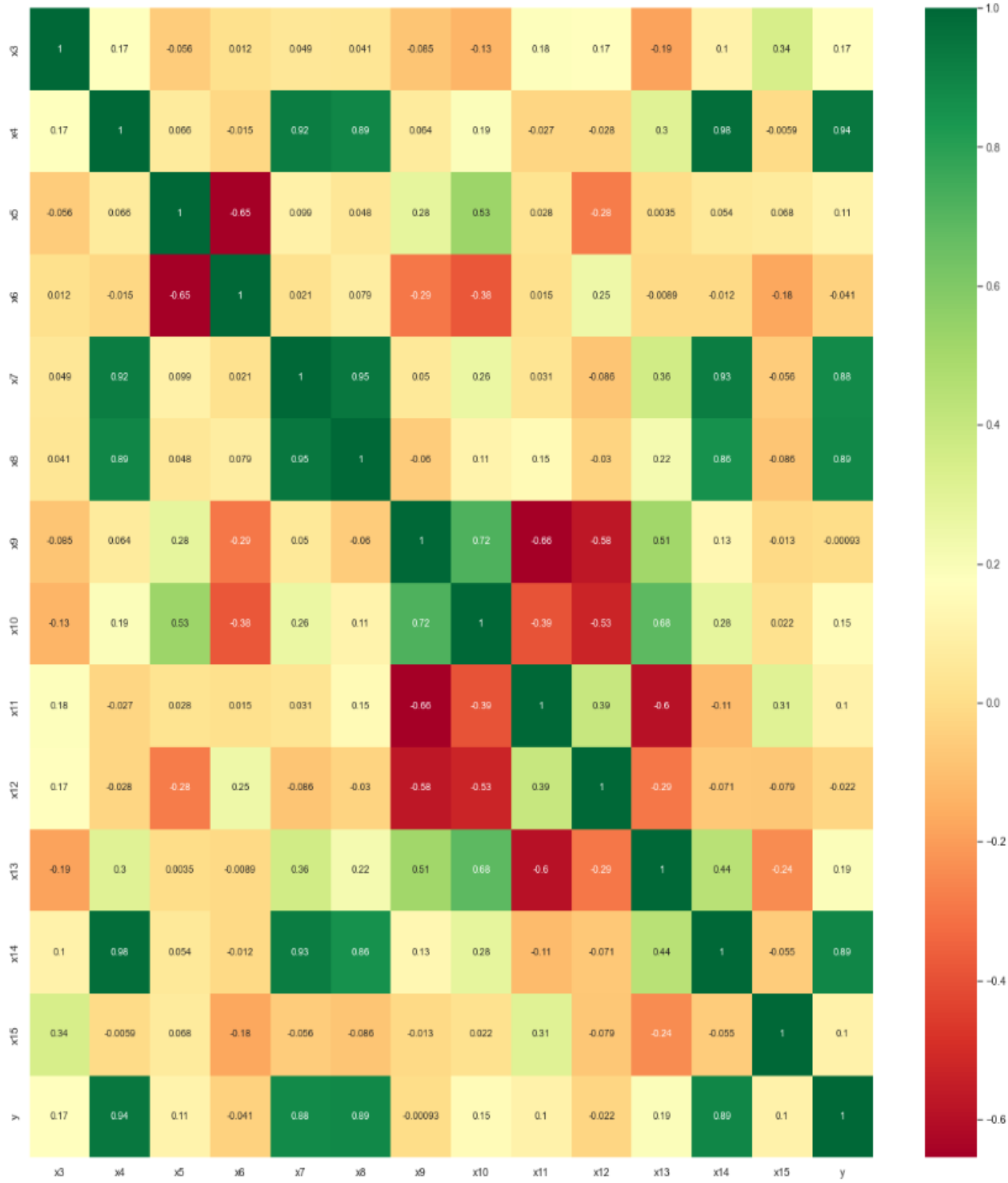
After Standard scaler few rows of data look like as below table:

	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	y
153	-0.3257	-0.0934	-0.1884	-0.6830	-0.3035	-0.0417	0.0022	-0.0054	-0.7965	0.4880	0.7618	-0.0294	-0.4623	-0.1623
262	-0.1200	-0.4193	-0.6763	0.3842	-0.2134	-0.0054	0.5782	0.1230	-0.3133	-0.7345	0.2526	-0.3834	-0.4623	-0.3030
142	-0.1775	-0.0180	-0.0177	-0.1613	0.4318	0.7840	0.1849	0.2643	0.4115	-0.2817	-0.0838	-0.0828	0.5001	0.5916
219	-0.3980	-0.3279	-0.4080	-0.3984	0.0348	0.1368	-0.2647	-0.2750	0.8508	0.2616	-0.2014	-0.3456	0.5001	-0.1076
119	-0.0418	0.1085	-0.0177	-0.1850	0.0206	0.2713	0.6625	0.0974	-0.0717	-0.8704	0.1731	0.0643	-0.4623	0.3470

Also we can see data description as below

	count	mean	std	min	25%	50%	75%	max
x3	306.0000	0.0000	1.0016	-0.6256	-0.3742	-0.2487	-0.0725	11.2569
x4	306.0000	0.0000	1.0016	-0.6028	-0.5083	-0.3375	0.1614	11.1216
x5	306.0000	-0.0000	1.0016	-2.9690	-0.5787	-0.0909	0.4457	5.0800
x6	306.0000	-0.0000	1.0016	-2.1771	-0.5407	-0.1376	0.3545	5.1272
x7	306.0000	0.0000	1.0016	-0.6112	-0.4988	-0.3666	0.0679	10.1053
x8	306.0000	-0.0000	1.0016	-0.6770	-0.5037	-0.3114	0.0832	10.9646
x9	306.0000	-0.0000	1.0016	-4.1982	-0.5140	0.0093	0.6871	2.1376
x10	306.0000	0.0000	1.0016	-1.7132	-0.7630	-0.1851	0.5500	3.9625
x11	306.0000	0.0000	1.0016	-1.4994	-0.7087	-0.1596	0.4719	5.5072
x12	306.0000	0.0000	1.0016	-1.9118	-0.6440	-0.1459	0.3974	6.7365
x13	306.0000	0.0000	1.0016	-2.3454	-0.6087	-0.1977	0.4659	4.6379
x14	306.0000	0.0000	1.0016	-0.6296	-0.5023	-0.3493	0.1534	10.8429
x15	306.0000	-0.0000	1.0016	-1.4247	-0.4623	0.5001	0.5001	1.4625
y	306.0000	0.0000	1.0016	-0.6367	-0.4661	-0.3193	0.1102	12.1741

Heatmap to see correlation between variable and response variable



Insights:

- Here, we can see response variable and X_4, X_7, X_8 and X_{14} are highly corelated.
- As we see previous that there are highly corelated with independent variable, we can also see their.

Building a Univariate model to multivariate model

Stepwise selection

In statistics, **stepwise selection** is a procedure we can use to build a regression model from a set of predictor variables by entering and removing predictors in a stepwise manner into the model until there is no statistically valid reason to enter or remove any more.

Forward model selection(Manual Method)

Forward selection is a popular method in regression analysis for selecting a subset of predictor variables to include in a multiple regression model. The basic idea behind forward selection is to start with a model containing only the intercept, and then sequentially add one predictor variable at a time, selecting the variable that provides the most improvement in the model's fit.

Here are the steps for the forward selection method:

1. Start with a model containing only the intercept.

2. Fit all possible simple linear regression models between the response variable and each of the predictor variables.

3. Choose the predictor variable with the highest correlation (or lowest p-value) with the response variable, and add it to the model.

4. Fit all possible multiple linear regression models between the response variable and the chosen predictor variable, along with each of the remaining predictor variables.

5. Choose the predictor variable that, when added to the chosen predictor variable(s), provides the most improvement in the model's fit (e.g., the variable with the highest increase in R-squared).

6. Repeat steps 4 and 5 until no further improvement in the model's fit is achieved (e.g., no predictor variable provides a significant improvement in the model's fit).

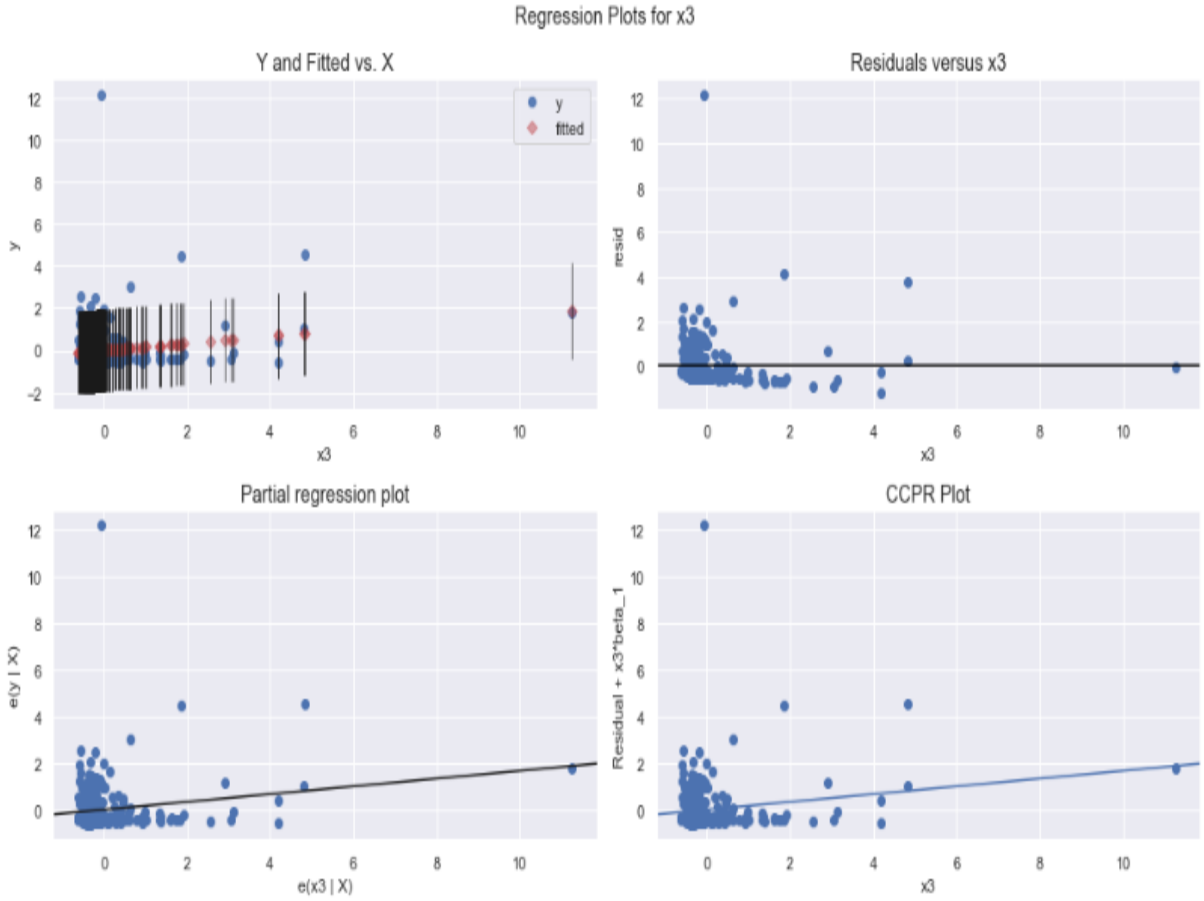
7. Check the assumptions of the model and interpret the results.

It's important to note that while forward selection can be a useful method for identifying a subset of predictor variables to include in a model, it can also lead to overfitting if not used carefully. Therefore, it's important to balance model fit with model simplicity and interpretability**.**

Implementation of Forward Feature Selection

1. **Fitting independent variable 'X_3' to dependent variable 'y'**

   Visualise the data with a scatter plot and the fitted regression line



Insights:

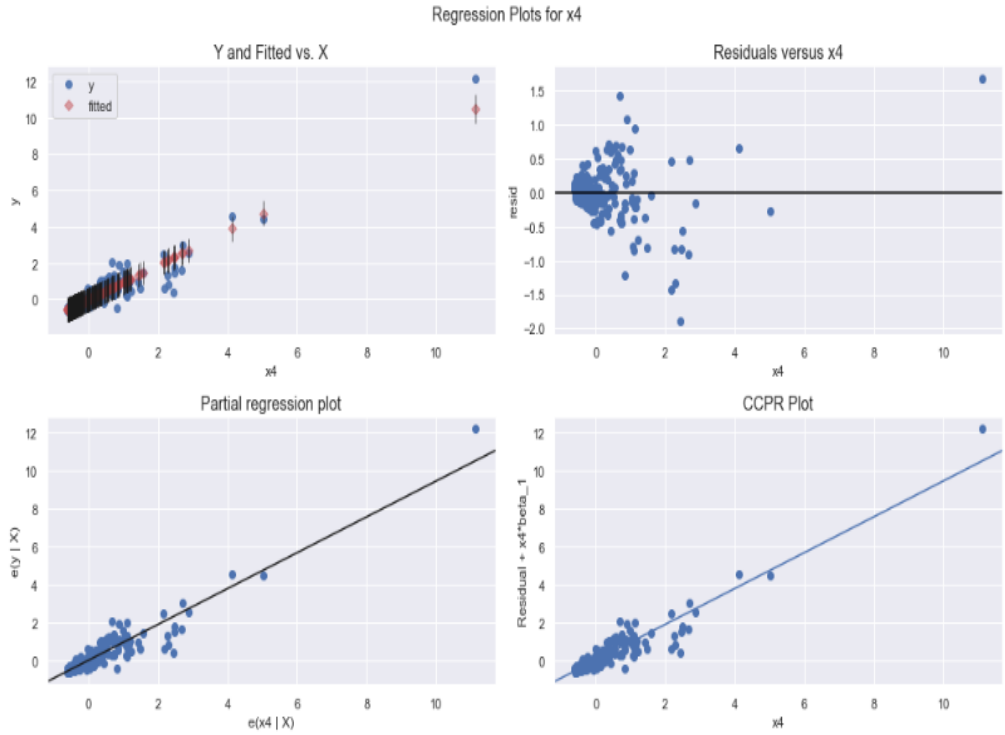
- As we can see by scatterplot there are high error between actual value and fitted value. Let's visualise it by statistically

R- squared	**0.028**
Adj. R- squared	**0.024**

- So we have less R- squared value obtained is **0.028**. Since we have so many variables, we will not add X_3 in our model.

2. Fitting independent variable 'X_4' to dependent variable 'y'

a. Visualise it by graph

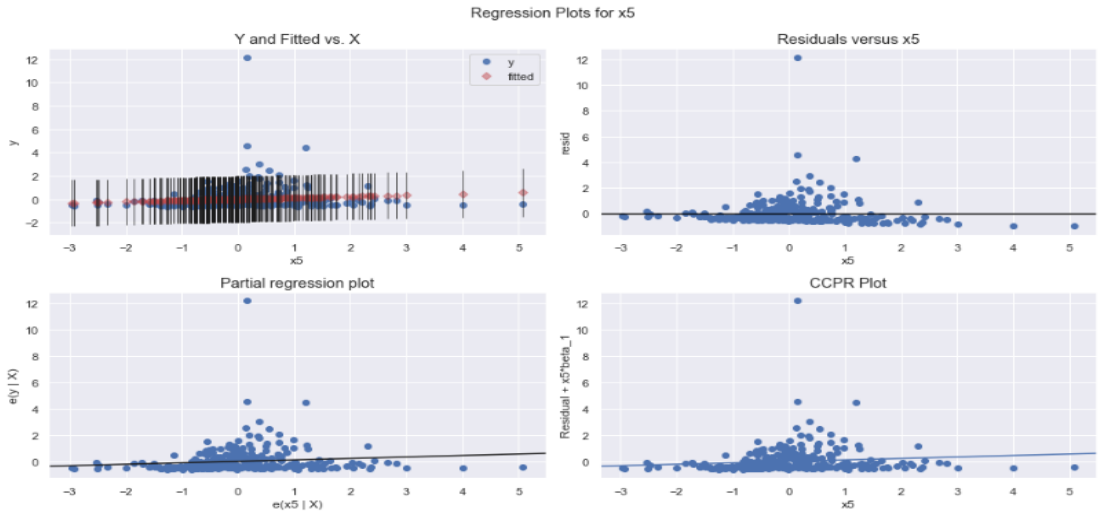


- Visualise it by statistically

R- Squared	**0.890**
Adj. R-squared	**0.890**

- The R-squared value obtained is 0.890. So we will add 'X_4' in our model.

3. Fitting independent variable 'X_5' to dependent variable 'y'



- Visualise it by graph

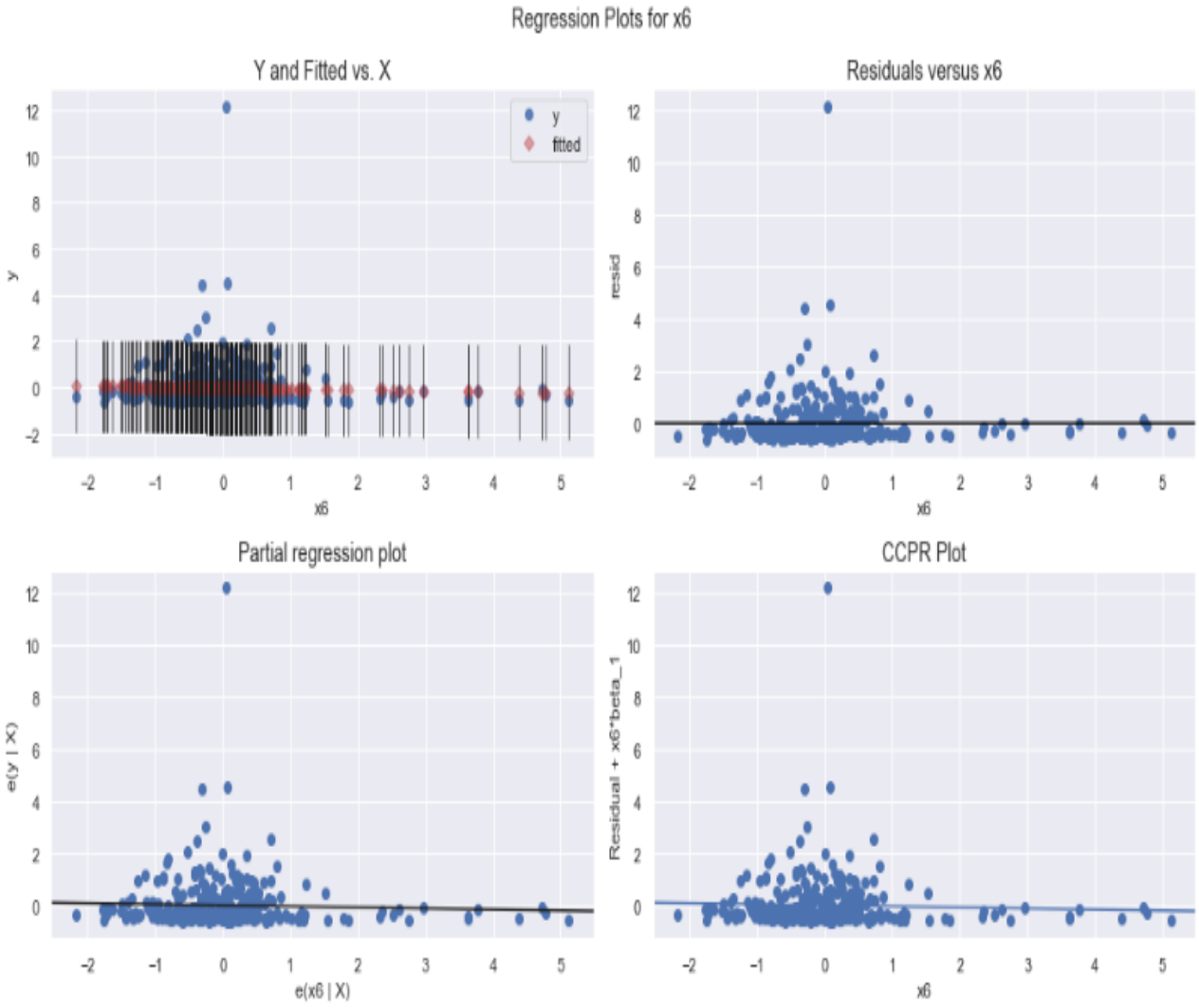
    - Visualise it by statistically

R- Squared	**0.012**
Adj. R- Squared	**0.009**

    - Here, R-squared value obtained is 0.012. Since we have so many variables, we will not add 'X_5' in our model.

4. **Fitting independent variable 'X_6' to dependent variable 'y'**

    - Visualise it by graph



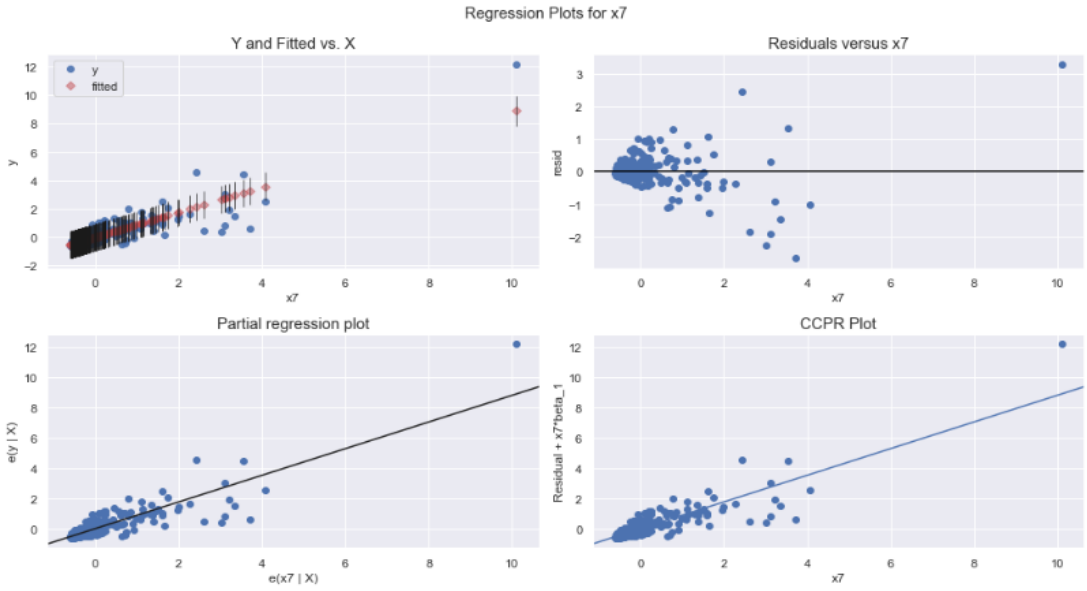
- Visualise it by statistically

R- Squared	**0.002**
Adj. R – Squared	**-0.002**

- The R-squared value obtained is 0.002. Since we have so many variables, we will not add 'X_6' in our model.
- A negative adj. R-squared can occur when the model is overfitting the data, meaning it is trying to fit noise rather than the underlying pattern in the data. It can also occur when the model is misspecified or if the sample size is too small.
- In summary, a negative adj. R-squared value indicates that the model is not a good fit for the data and should not be used for prediction or interpretation.

5. Fitting independent variable 'X_7' to dependent variable 'y'

- Visualise it by graph



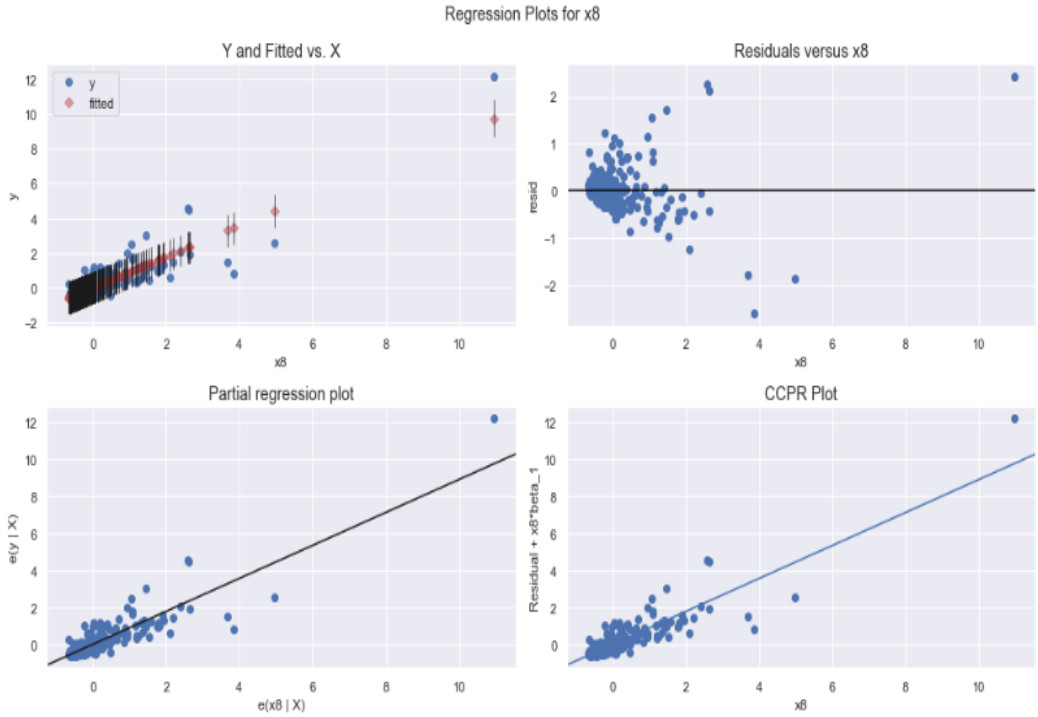
- Visualise it by statistically

R-squared	**0.773**
Adj. R-squared	**0.773**

- The R- squared value obtained is 0.773. Since we have so many variables, we will add 'X_7' in our model.

6. Fitting independent variable 'X_8' to dependent variable 'y'

- Visualise it by graph



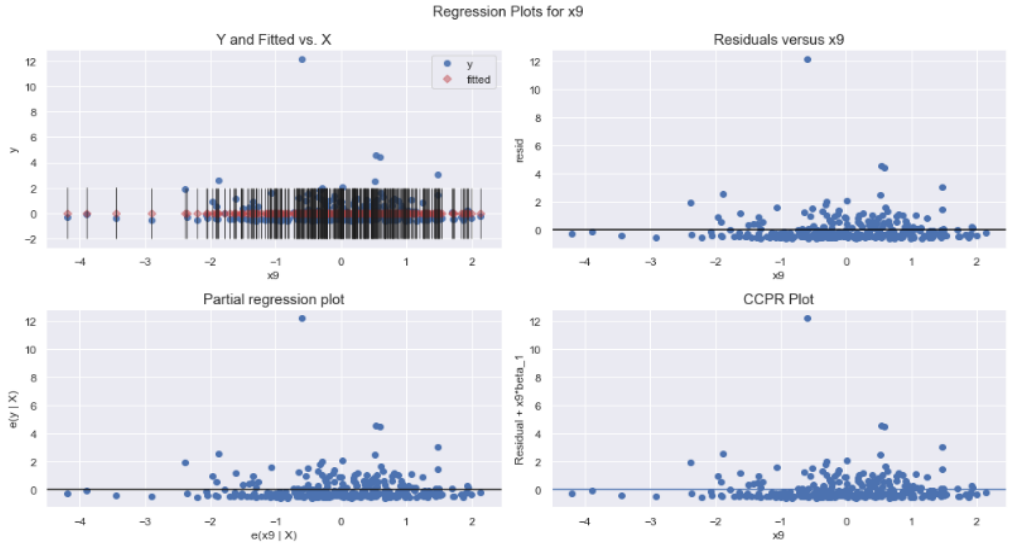
- Visualise it by statistically

R- squared	**0.790**
Adj. R- squared	**0.789**

- The R- squared value obtained is 0.790. So we will add 'X_8' in our model.

7. Fitting independent variable 'X_9' to dependent variable 'y'

- Visualise it by graph



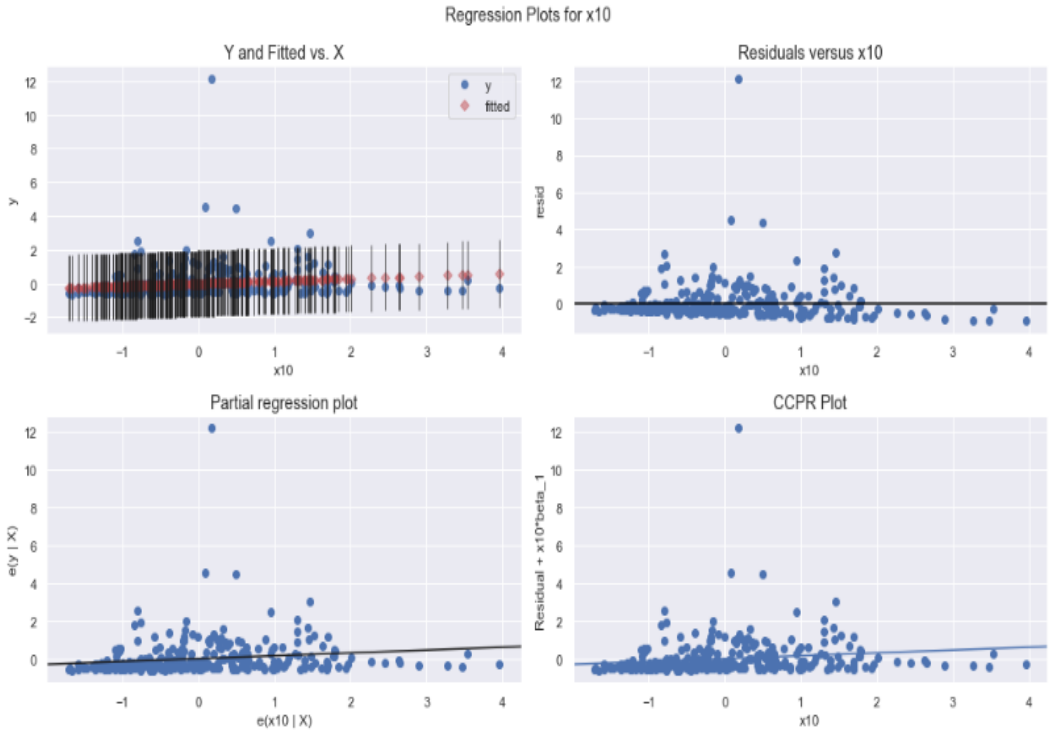
- Visualise it by statistically

R- squared	**0.000**
Adj. R- squared	**-0.003**

- The R- squared value obtained is 0.000. Since we have so many variables, we will not add 'X_9' in our model.
- A negative adj. R-squared can occur when the model is overfitting the data, meaning it is trying to fit noise rather than the underlying pattern in the data. It can also occur when the model is misspecified or if the sample size is too small.
- In summary, a negative adj. R-squared value indicates that the model is not a good fit for the data and should not be used for prediction or interpretation.

8. Fitting independent variable 'X_{10}' to dependent variable 'y'

- Visualise it by graph



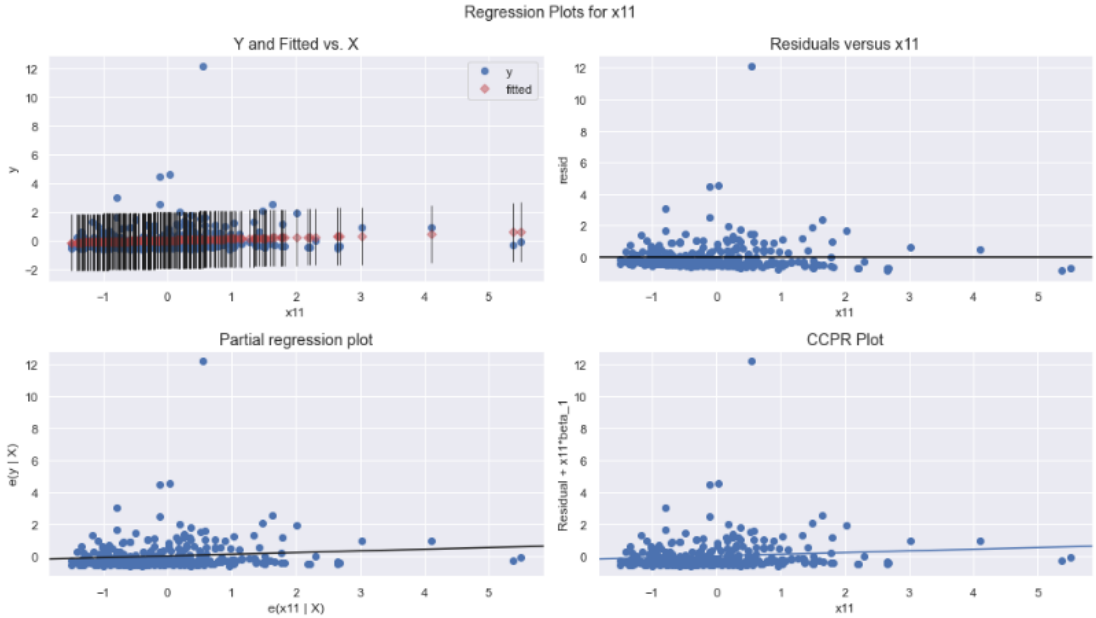
- Visualise it by statistically

R- squared	**0.023**
Adj. R- squared	**0.019**

- The R- squared value obtained is 0.023. So we will add 'X_{10}' in our model.

9. Fitting independent variable 'X_{11}' to dependent variable 'y'

- Visualise it by graph



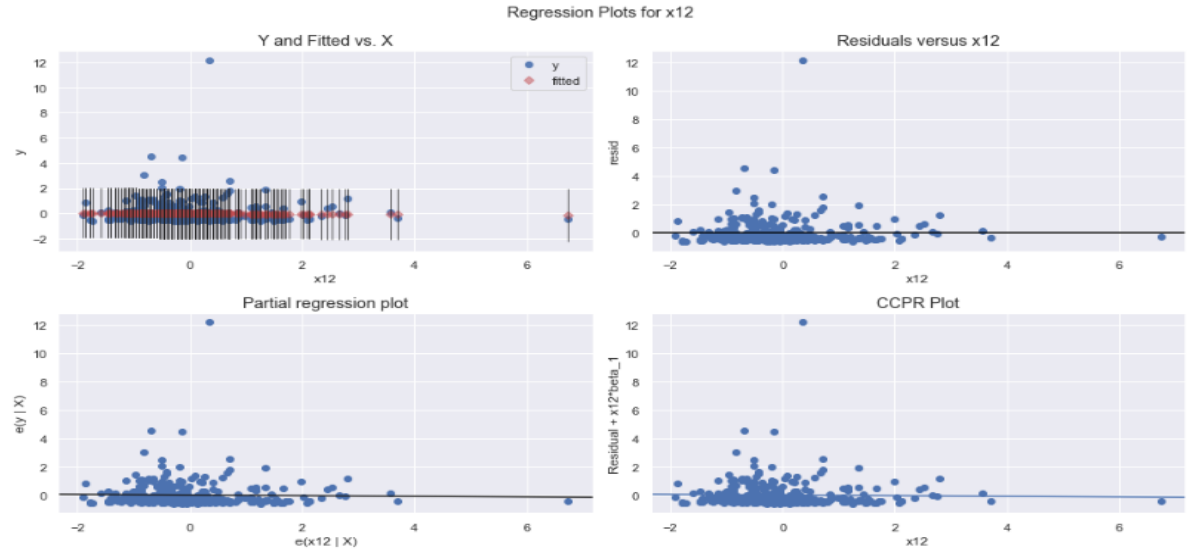
- Visualise it by statistically

R- squared	**0.011**
Adj. R- squared	**0.008**

- The R- squared value obtained is 0.011. Since we have so many variables, we will not add 'X_{11}' in our model.

10. Fitting independent variable 'X_{12}' to dependent variable 'y'

- Visualise it by graph



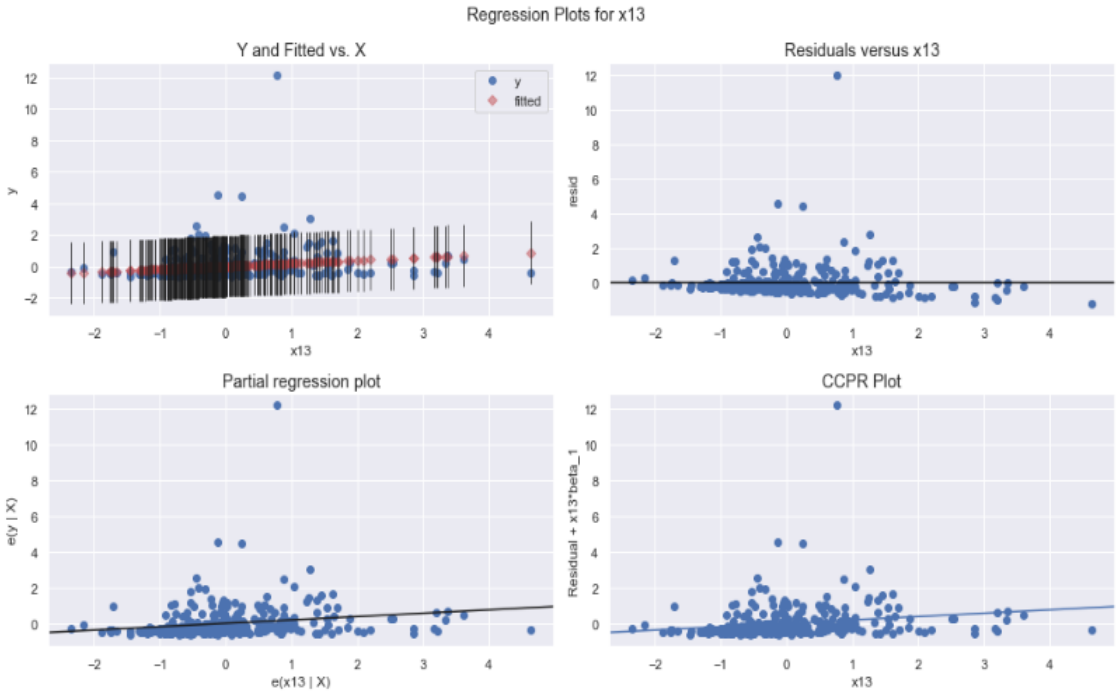
- Visualise it by statistically

R- squared	**0.000**
Adj. R- squared	**-0.003**

- The R- squared value obtained is 0.000. Since we have so many variables, we will add 'X_{12}' in our model.
- A negative adj. R-squared can occur when the model is overfitting the data, meaning it is trying to fit noise rather than the underlying pattern in the data. It can also occur when the model is misspecified or if the sample size is too small.
- In summary, a negative adj. R-squared value indicates that the model is not a good fit for the data and should not be used for prediction or interpretation.

11. Fitting independent variable 'X_{13}' to dependent variable 'y'

- Visualise it by graph



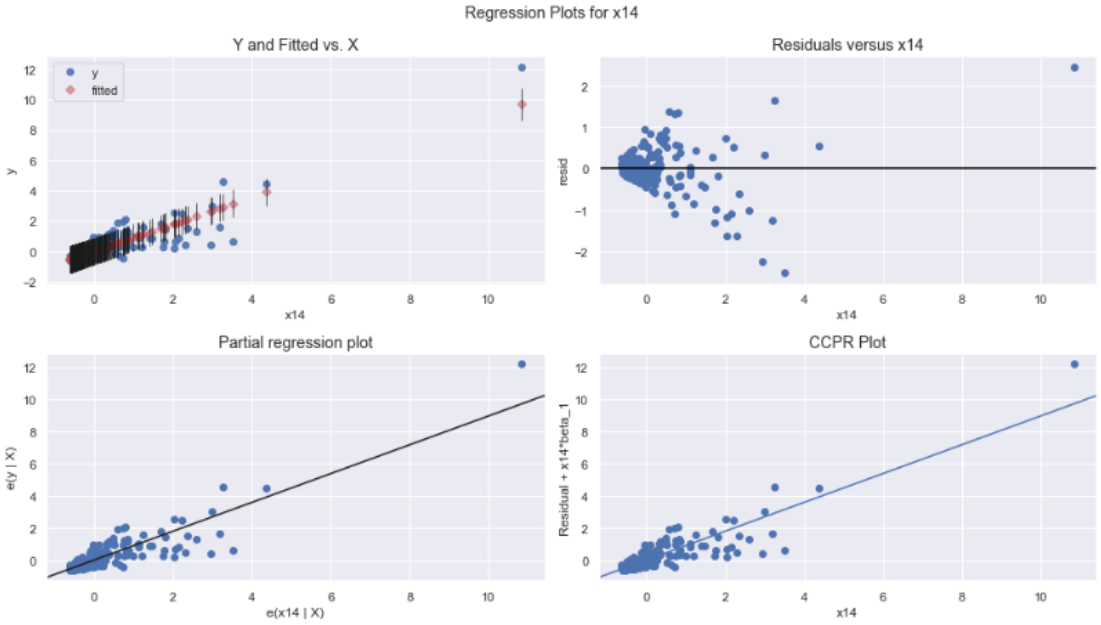
- Visualise it by statistically

R- squared	**0.036**
Adj. R- squared	**0.032**

- The R- squared value obtained is 0.036. Since we have so many variables, we will not add 'X_{13}' in our model.

12. Fitting independent variable 'X_{14}' to dependent variable 'y'

- Visualise it by graph



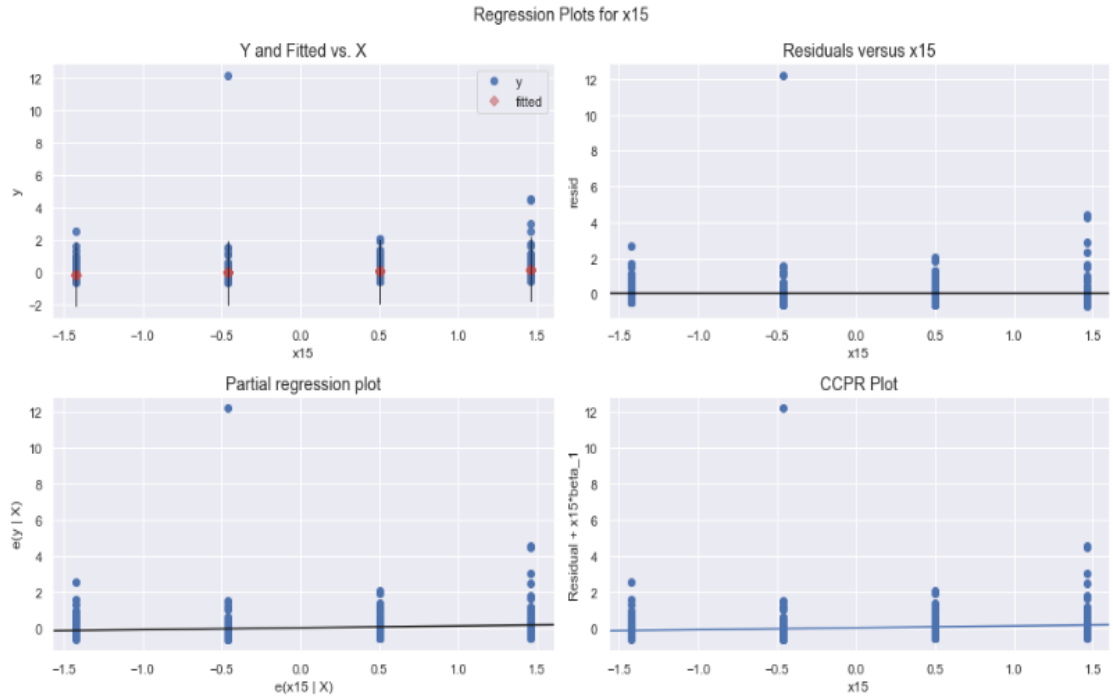
- Visualise it by statistically

R- squared	**0.801**
Adj. R- squared	**0.800**

- The R- squared value obtained is 0.801. So we will add 'X_{14}' in our model.

13. Fitting independent variable 'X_{15}' to dependent variable 'y'

- Visualise it by graph



- Visualise it by statistically

R- squared	**0.011**
Adj. R- squared	**0.008**

- The R- squared value obtained is 0.011. Since we have so many variables, we will not add 'X_{15}' in our model.

Remark On Negative Adj. R-squared:

- The adjusted R-squared (adj. R-squared) is a modification of the R-squared value that adjusts for the number of predictor variables in a multiple regression model. The adj. R-squared can range from negative infinity to 1, with higher values indicating a better fit of the model to the data.
- However, it's not common to have a negative adj. R-squared value, as it would imply that the model is worse than a model that simply uses the intercept to predict the response variable. A negative adj. R-squared suggests that the model is not useful for making predictions and that the model does not explain the variation in the response variable better than the mean of the response variable.
- A negative adj. R-squared can occur when the model is overfitting the data, meaning it is trying to fit noise rather than the underlying pattern in the data. It can also occur when the model is misspecified or if the sample size is too small.
- In summary, a negative adj. R-squared value indicates that the model is not a good fit for the data and should not be used for prediction or interpretation.

CONCLUSION:

- The R-squared value of 'X_4', 'X_7', 'X_8' and 'X_{14}' obtained are 0.890, 0.773, 0.790 and 0.801 respectively.
- The Variables 'X_7' and 'X_8' are highly correlated So, we will add 'X_7' from our model.
- The model will contain the variables 'X_4', 'X_8' and 'X_{14}'.

Adding the variable in forward selection

R- squared of 'X_4'	**0.890**
R- squared of 'X_7'	**0.773**
R- squared of 'X_{14}'	**0.801**

- The R-squared value obtained is 0.890. Since we have so many variables, we can clearly do better than this. So let's go ahead and add the other highly correlated variable, i.e. 'X_{14}'.

R- squared of 'X_4'	**0.890**
R- squared of 'X_{14}' and 'X_{14}'	**0.909**

- The R-squared of 0.909 good, Since we have so many variables, we can clearly do better than this. So lets add another correlated variable, i.e. 'X_8'.

R-squared of 'X_4'	**0.890**
R-squared of 'X_4' and 'X_{14}'	**0.909**
R-squared of 'X_4', 'X_8' and 'X_{14}'	**0.917**

- We have achieved a R- squared of 0.917 by manually picking the highly correlated variables.

RMSE Score of test data

The R2 score of Training set is 0.917 and Test set is 0.945 which is quite close. Hence, We can say that our model is good enough to predict the Car prices using below predictor variables

- x4
- x8
- x14

Equation of Line to predict for y value

$$y=1.3679*x4+0.1922*x8-0.6087*x14$$

Model I Conclusions:

- R-squared and Adjusted R-squared - 0.917 and 0.916 - 91% variance explained.
- F-stats and Prob(F-stats) (overall model fit) - 1107.0 and 1.50e-162(approx. 0.0) - Model fit is significant and explained 91%
  variance is just not by chance.
- p-values - p-values for all the coefficients seem to be less than the significance level of 0.05. - meaning that all the
  predictors are statistically significant.

Forword model selection(Computerised method)

To finding the best subsets for each number of features by computer method

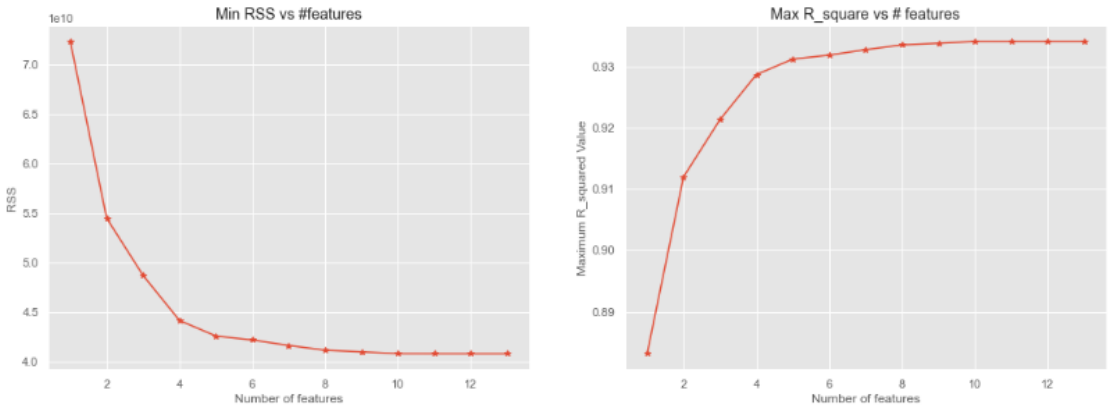
- Using the smallest RSS Value, or the largest R-squared value
- Finding the best subsets for each number by minimum RSS value

	numb_features	RSS	R_squared	features
1	1	72391134361.1159	0.8831	(x4,)
34	2	54525980580.6152	0.9120	(x4, x14)
189	3	48743839976.8954	0.9213	(x4, x8, x14)
726	4	44189372377.3233	0.9287	(x4, x8, x14, x15)
1221	5	42621121063.2930	0.9312	(x3, x4, x8, x14, x15)
2463	6	42218715797.0634	0.9318	(x3, x4, x5, x8, x14, x15)
4534	7	41671117884.2695	0.9327	(x3, x4, x8, x11, x13, x14, x15)
6026	8	41188039110.5191	0.9335	(x3, x4, x5, x8, x9, x13, x14, x15)
7243	9	41021944716.4308	0.9338	(x3, x4, x5, x7, x8, x9, x13, x14, x15)
7910	10	40841275622.9899	0.9341	(x3, x4, x5, x7, x8, x9, x11, x13, x14, x15)
8140	11	40836856746.9007	0.9341	(x3, x4, x5, x7, x8, x9, x11, x12, x13, x14, x15)
8182	12	40833594073.2697	0.9341	(x3, x4, x5, x6, x7, x8, x9, x11, x12, x13, x14, x15)
8190	13	40832909427.1169	0.9341	(x3, x4, x5, x6, x7, x8, x9, x10, x11, x12, x13, x14, x15)

- Finding the best subsets for each number by maximum R-squared value

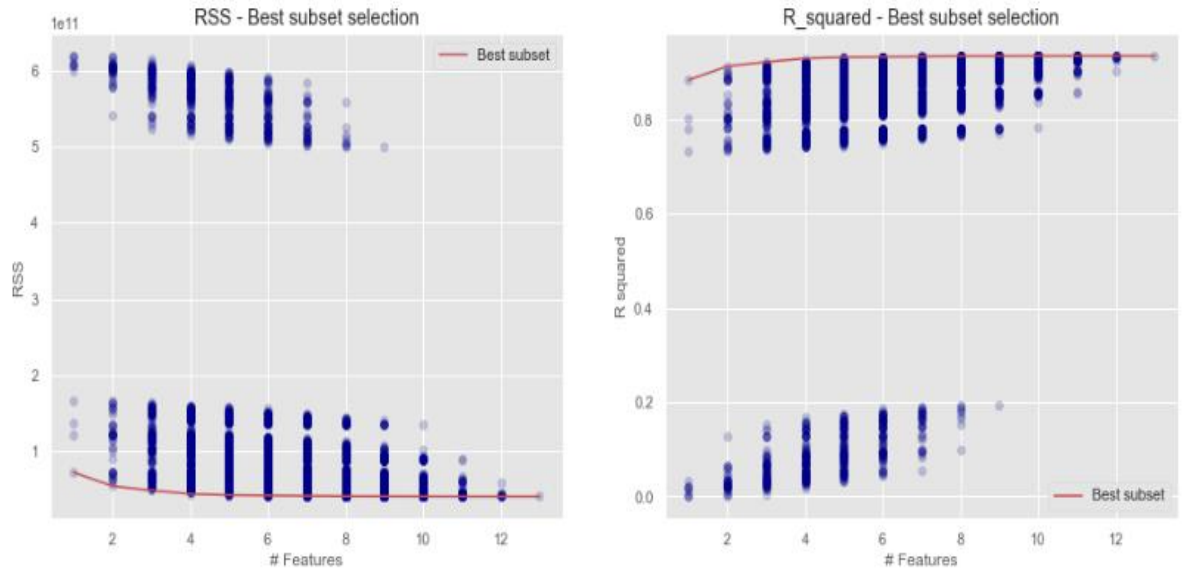
	numb_features	RSS	R_squared	features
1	1	72391134361.1159	0.8831	(x4,)
34	2	54525980580.6152	0.9120	(x4, x14)
189	3	48743839976.8954	0.9213	(x4, x8, x14)
726	4	44189372377.3233	0.9287	(x4, x8, x14, x15)
1221	5	42621121063.2930	0.9312	(x3, x4, x8, x14, x15)
2463	6	42218715797.0634	0.9318	(x3, x4, x5, x8, x14, x15)
4534	7	41671117884.2695	0.9327	(x3, x4, x8, x11, x13, x14, x15)
6026	8	41188039110.5191	0.9335	(x3, x4, x5, x8, x9, x13, x14, x15)
7243	9	41021944716.4308	0.9338	(x3, x4, x5, x7, x8, x9, x13, x14, x15)
7910	10	40841275622.9899	0.9341	(x3, x4, x5, x7, x8, x9, x11, x13, x14, x15)
8140	11	40836856746.9007	0.9341	(x3, x4, x5, x7, x8, x9, x11, x12, x13, x14, x15)
8182	12	40833594073.2697	0.9341	(x3, x4, x5, x6, x7, x8, x9, x11, x12, x13, x14, x15)
8190	13	40832909427.1169	0.9341	(x3, x4, x5, x6, x7, x8, x9, x10, x11, x12, x13, x14, x15)

Plotting the minimum RSS and Maximum R-square Vs Number of features



COMMENT: Both subsets of selection process are given similar

Plotting the best subset selection process



Forward stepwise selection

For computational reasons, the best subset cannot be applied for any large n due to the 2^n complexity. Forward Stepwise begins with a model containing no predictors, and then adds predictors to the model, one at the time. At each step, the variable that gives the greatest additional improvement to the fit is added to the model.

Algorithm

Let M_0 denote the null model which contains no predictors

- For k=1,2,...,n−1
  - Consider all n−k models that augment the predictors in M_k with one additional predictor
  - Choose the among these n−k models, and call it M_k+1
  - Select the single best model among M0,M1,...,Mn using cross validated predict error, Cp,BIC,adjustedR2 or any other method.

Comparing models: AIC, BIC, Mallow's CP

The training set Mean Squared Error (MSE) is generally an underestimate of the test MSE. This is because when we fit a model to the training data using least squares, we specifically estimate the regression coefficients such that the training RSS is minimized. In particular, the training RSS decreases as we add more features to the model, but the test error may not. Therefore the training RSS and R2 may not be used for selecting the best model unless we adjust for this underestimation.

Mallow's Cp

Mallow's Cp is named after Colin Lingwood Mallows and is defined as:

$Cp=(1/m)*(RSS+2d\sigma'^2)$

where σ'^2 is an estimate of the variance of the error ϵ associated with each response measurement. Typically σ'^2 is estimated using the full model containing all predictors.

Akaike's Information Criteria (AIC)

The AIC criterion is defiend for a large class of models fit by maximum likelihood. In the case of a linear model with Gaussian errors, MLE and least squares are the same thing and the AIC is given by

$AIC=(1/(m\sigma'^2))*(RSS+2d\sigma'^2)$

Bayesian Information Criteria (BIC)

BIC is derived from a Bayesian point of view, and looks similar to the Cp and AIC- it is defined (up to irrelevant constants) as:
$BIC=(1/m\sigma'2)*(RSS+log(m)d\sigma'2)$
Like Cp and AIC, the BIC will tend to take small values for a model with low test error.

Adjusted R2

Since the R2 always increases as more variables are added, the adjusted R2 accounts for that fact and introduces a penalty. The intuition is that once all the correct variables have been included in the model, additional noise variables will lead to a very small decrease in RSS, but an increase in k and hence will decrease the adjusted R2. In effect, we pay a price for the inclusion of unnecessary variables in the model.
$R_a^2=1-(RSS/(m-k-1))/(TSS/(m-1))=1-((1-R2)(m-1)/(m-k-1))$

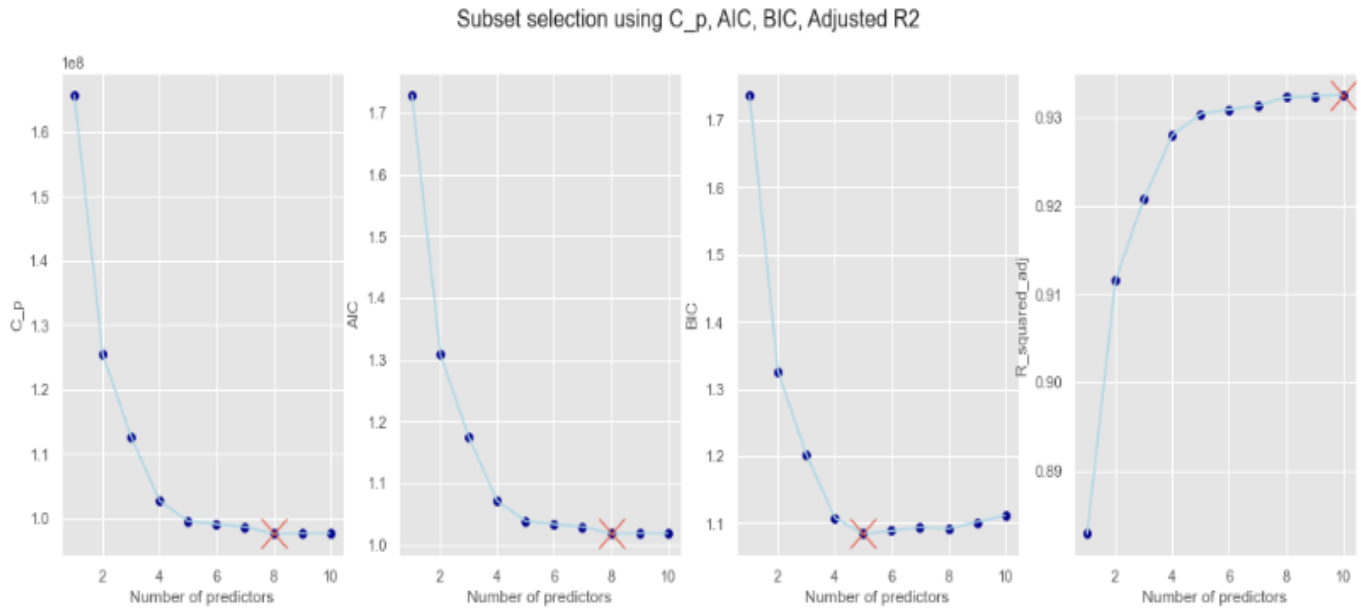
Theoretical justification

Cp, AIC, BIC all have rigorous theoretical justification that rely on asymptotic arguments, i.e. when the sample size m grows very large, whereas the adjusted R2, although quite intuitive, is not as well motivated in statistical theory.

Computing the C_p, AIC, BIC and R- square adjusted

	features	RSS	R_squared	numb_features	C_p	AIC	BIC	R_squared_adj
1	[x4]	72391134361.1159	0.8831	1	165714332.0529	1.7285	1.7378	0.8829
2	[x4, x14]	54525980580.6152	0.9120	2	125364079.3075	1.3076	1.3263	0.9116
3	[x4, x14, x8]	48743839976.8954	0.9213	3	112600614.6370	1.1745	1.2025	0.9208
4	[x4, x14, x8, x15]	44189372377.3234	0.9287	4	102640056.3687	1.0706	1.1079	0.9280
5	[x4, x14, x8, x15, x3]	42621121063.2930	0.9312	5	99497343.5012	1.0378	1.0844	0.9304
6	[x4, x14, x8, x15, x3, x5]	42218715797.0634	0.9318	6	99016379.6012	1.0328	1.0887	0.9309
7	[x4, x14, x8, x15, x3, x5, x13]	41863611004.6015	0.9324	7	98643407.6504	1.0289	1.0942	0.9313
8	[x4, x14, x8, x15, x3, x5, x13, x9]	41188039110.5191	0.9335	8	97538775.6504	1.0174	1.0920	0.9323
9	[x4, x14, x8, x15, x3, x5, x13, x9, x7]	41021944716.4308	0.9338	9	97597334.2895	1.0180	1.1019	0.9324
10	[x4, x14, x8, x15, x3, x5, x13, x9, x7, x11]	40841275622.9899	0.9341	10	97622617.3594	1.0183	1.1115	0.9325

Plotting the computed values as a function of number of features



COMMENT:

We found the optimum subsets for each number of features up to 10 variables using the computerised forward selection approach, with associated R2 scores. Other statical parameters such as Mallow's Cp, Akaike's Information Criteria (AIC), Bayesian Information Criteria (BIC), and Adjusted R2 will be found for each number of features. And will show the relationship between the number of predictors and Mallow's Cp, Akaike's Information Criteria (AIC), Bayesian Information Criteria (BIC), and Adjusted R2. And we may infer that the model with 5, 8, and 10 cannot be used. Mallow's Cp and Akaike's Information Criteria (AIC) both reject the model with 8 predictors. Furthermore, the Bayesian Information Criteria (BIC) and Adjusted R2 will reject the model with 5 and 10 predictors, respectively.

Backword model selection

Backward selection is another popular method in regression analysis for selecting a subset of predictor variables to include in a multiple regression model. The basic idea behind backward selection is to start with a full model containing all predictor variables, and then sequentially remove one predictor variable at a time, selecting the variable that provides the least reduction in the model's fit.

Here are the steps for the backward selection method:

1. Start with a full model containing all predictor variables.

2. Fit the multiple linear regression model and calculate the p-value for each predictor variable.

3. Choose the predictor variable with the highest p-value and remove it from the model.

4. Fit the multiple linear regression model again, but this time with the predictor variable removed.

5. Repeat steps 2-4 until no further reduction in the model's fit is achieved (e.g., no predictor variable has a p-value above the significance level).

6. Check the assumptions of the model and interpret the results.

Similar to forward selection, it's important to balance model fit with model simplicity and interpretability when using backward selection. Backward selection can also lead to overfitting if not used carefully, so it's important to validate the final model on independent data.

RFE (Recursive feature elimination) :

RFE stands for Recursive Feature Elimination. It is a feature selection technique used in machine learning to select the most important features in a dataset. RFE works by recursively removing attributes from a dataset and building a model using the remaining attributes. The importance of each attribute is then determined based on how much the performance of the model decreases when that attribute is removed. This process is repeated until the desired number of features is reached or until no further improvement in the model's performance is observed. RFE can be used with a variety of machine learning algorithms, including regression, classification, and clustering algorithms.

We select the variables which are in support and by RFE we get 'x3', 'x4', 'x5', 'x7', 'x8', 'x10', 'x11', 'x13', 'x14', 'x15'

After passing the arbitrary selected columns by RFE we will manually evaluate each models p-value and VIF value. Unless we find the acceptable range for p-values and VIF we keep dropping the variables one at a time based on below criteria.

- High p-value High VIF : Drop the variable
- High p-value Low VIF or Low p-value High VIF : Drop the variable with high p-value first.
- Low p-value Low VIF : accept the variable

OLS Regression Results

Dep. Variable:	y	R-squared:	0.932
Model:	OLS	Adj. R-squared:	0.930
Method:	Least Squares	F-statistic:	405.7
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	6.22e-166
Time:	16:05:57	Log-Likelihood:	-22.406
No. Observations:	306	AIC:	66.81
Df Residuals:	295	BIC:	107.8
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.776e-17	0.015	1.83e-15	1.000	-0.030	0.030
x3	-0.0526	0.018	-2.881	0.004	-0.089	-0.017
x4	1.3099	0.129	10.154	0.000	1.056	1.564
x5	0.0388	0.023	1.722	0.086	-0.006	0.083
x7	-0.1485	0.074	-2.018	0.044	-0.293	-0.004
x8	0.2757	0.065	4.228	0.000	0.147	0.404
x10	0.0030	0.032	0.093	0.926	-0.060	0.066
x11	0.0498	0.022	2.218	0.027	0.006	0.094
x13	0.0433	0.036	1.199	0.232	-0.028	0.114
x14	-0.4904	0.131	-3.753	0.000	-0.748	-0.233
x15	0.1112	0.018	6.046	0.000	0.075	0.147

Omnibus:	92.246	Durbin-Watson:	1.974
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1371.931
Skew:	-0.769	Prob(JB):	1.23e-298
Kurtosis:	13.258	Cond. No.	25.1

Insights:

- Looking at the p-values, it looks like some of the variables aren't really significant (in the presence of other variables) and we need to drop it.

P-value and VIF:

Managing p-values and VIFs are essential parts of statistical analysis in regression modeling. Here are some general tips for managing p-values and VIFs:

For p-values: The p-value is a statistical measure that determines the significance of a regression coefficient. Generally, a p-value of less than 0.05 is considered significant, which means that the corresponding independent variable is likely to have a significant impact on the dependent variable. If the p-value is greater than 0.05, it suggests that the independent variable is not statistically significant and may need to be removed from the model.

For VIFs: The VIF measures the degree of multicollinearity among the independent variables in a regression model. Generally, a VIF of less than 5 is considered acceptable, while a VIF greater than 10 indicates that multicollinearity is present and the corresponding independent variable may need to be removed from the model.

To manage p-values and VIFs in regression analysis, follow these steps:

a. Check the p-values for each independent variable in the regression model. If a p-value is greater than 0.05, consider removing the corresponding independent variable from the model.

b. Check the VIF values for each independent variable in the regression model. If a VIF is greater than 10, consider removing the corresponding independent variable from the model.

c. If you remove an independent variable from the model, re-run the regression analysis and check the new p-values and VIFs. Repeat this process until you have a final model with all significant independent variables and acceptable VIFs.

d. Be careful not to overfit the model by including too many independent variables, as this can result in high VIFs and low p-values, but the model may not generalize well to new data.

Overall, managing p-values and VIFs is a critical step in ensuring the accuracy and reliability of regression models.

Checking VIF:

VIF stands for Variance Inflation Factor. It is a statistical measure used to detect multicollinearity in regression analysis. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other, making it difficult to assess the true effect of each variable on the dependent variable. The VIF measures the degree to which the variance of the estimated regression coefficient for a particular independent variable is increased due to multicollinearity.

A high VIF value (greater than 10) suggests that multicollinearity is present and the corresponding independent variable may need to be removed from the model. In contrast, a low VIF value (less than 5) indicates that there is little or no multicollinearity and the regression coefficients are reliable.

VIF is a commonly used tool in regression analysis to assess the multicollinearity of independent variables, and it is used to ensure that the regression model is stable and reliable.

Variance Inflation Factor or VIF, gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model. The formula for calculating `VIF` is:

$$VIF_i = \frac{1}{1-{R_i}^2}$$

Now get a dataframe that will contain the names of all the feature variables

	Features	VIF
8	x14	74.3100
1	x4	72.4300
3	x7	23.5700
4	x8	18.5000
7	x13	5.6800
5	x10	4.4700
2	x5	2.2100
6	x11	2.2000
9	x15	1.4700
0	x3	1.4500

We generally want a VIF that is less than 5. So there are clearly some variables we need to drop. Dropping the variable and updating the model**.** We start dropping with high pvalue first.

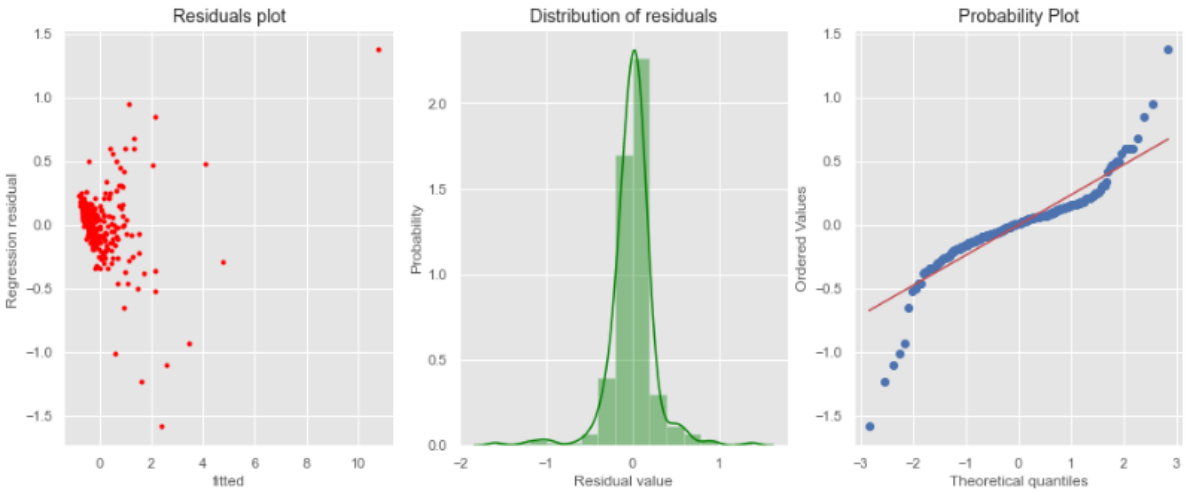
OLS Regression Results

Dep. Variable:	y	R-squared:	0.932
Model:	OLS	Adj. R-squared:	0.930
Method:	Least Squares	F-statistic:	452.3
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	2.85e-167
Time:	16:06:00	Log-Likelihood:	-22.410
No. Observations:	306	AIC:	64.82
Df Residuals:	296	BIC:	102.1
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.776e-17	0.015	1.83e-15	1.000	-0.030	0.030
x3	-0.0527	0.018	-2.891	0.004	-0.089	-0.017
x4	1.3104	0.129	10.184	0.000	1.057	1.564
x5	0.0403	0.016	2.539	0.012	0.009	0.071
x7	-0.1476	0.073	-2.026	0.044	-0.291	-0.004
x8	0.2751	0.065	4.245	0.000	0.148	0.403
x11	0.0498	0.022	2.219	0.027	0.006	0.094
x13	0.0455	0.027	1.670	0.096	-0.008	0.099
x14	-0.4914	0.130	-3.780	0.000	-0.747	-0.236
x15	0.1117	0.018	6.356	0.000	0.077	0.146

Omnibus:	92.056	Durbin-Watson:	1.974
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1370.066
Skew:	-0.766	Prob(JB):	3.12e-298
Kurtosis:	13.252	Cond. No.	24.6

	Features	VIF
7	x14	73.7900
1	x4	72.2900
3	x7	23.1900
4	x8	18.3300
6	x13	3.2500
5	x11	2.2000
0	x3	1.4500
8	x15	1.3500
2	x5	1.1000



COMMENT:- Dropping 'X13' because it's p- value is 0.096 and we want p-value less than 0.05 and hence rebuilding the model.

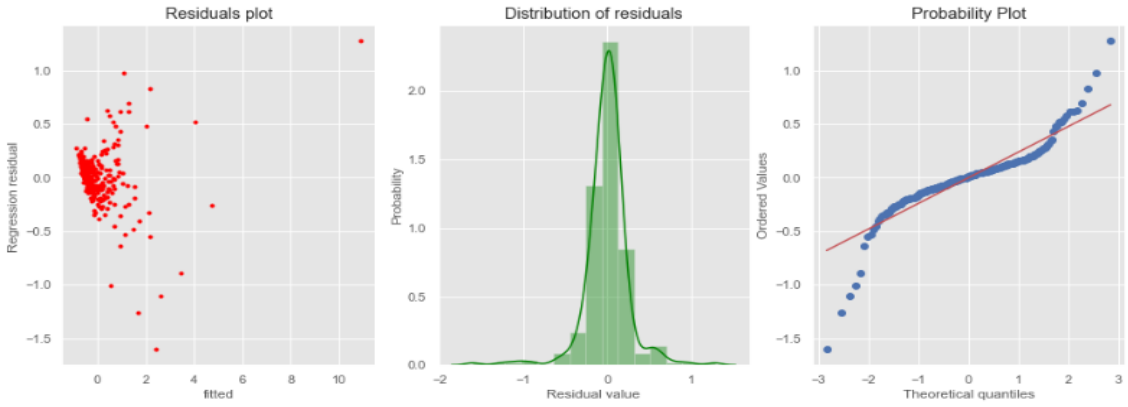
OLS Regression Results

Dep. Variable:	y	R-squared:	0.932
Model:	OLS	Adj. R-squared:	0.930
Method:	Least Squares	F-statistic:	505.4
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	4.91e-168
Time:	16:06:00	Log-Likelihood:	-23.845
No. Observations:	306	AIC:	65.69
Df Residuals:	297	BIC:	99.20
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.776e-17	0.015	1.83e-15	1.000	-0.030	0.030
x3	-0.0500	0.018	-2.748	0.006	-0.086	-0.014
x4	1.1895	0.107	11.151	0.000	0.980	1.399
x5	0.0415	0.016	2.613	0.009	0.010	0.073
x7	-0.1450	0.073	-1.985	0.048	-0.289	-0.001
x8	0.2892	0.064	4.488	0.000	0.162	0.416
x11	0.0301	0.019	1.573	0.117	-0.008	0.068
x14	-0.3701	0.108	-3.423	0.001	-0.583	-0.157
x15	0.1129	0.018	6.409	0.000	0.078	0.148

Omnibus:	95.365	Durbin-Watson:	1.972
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1258.739
Skew:	-0.857	Prob(JB):	4.66e-274
Kurtosis:	12.787	Cond. No.	20.5

	Features	VIF
6	x14	50.7500
1	x4	49.3900
3	x7	23.1800
4	x8	18.0200
5	x11	1.5900
0	x3	1.4400
7	x15	1.3500
2	x5	1.0900



COMMENT: Dropping ‘X11’ because it’s p-value is 0.117 and we want p-value less than 0.05 and hence rebuilding the model

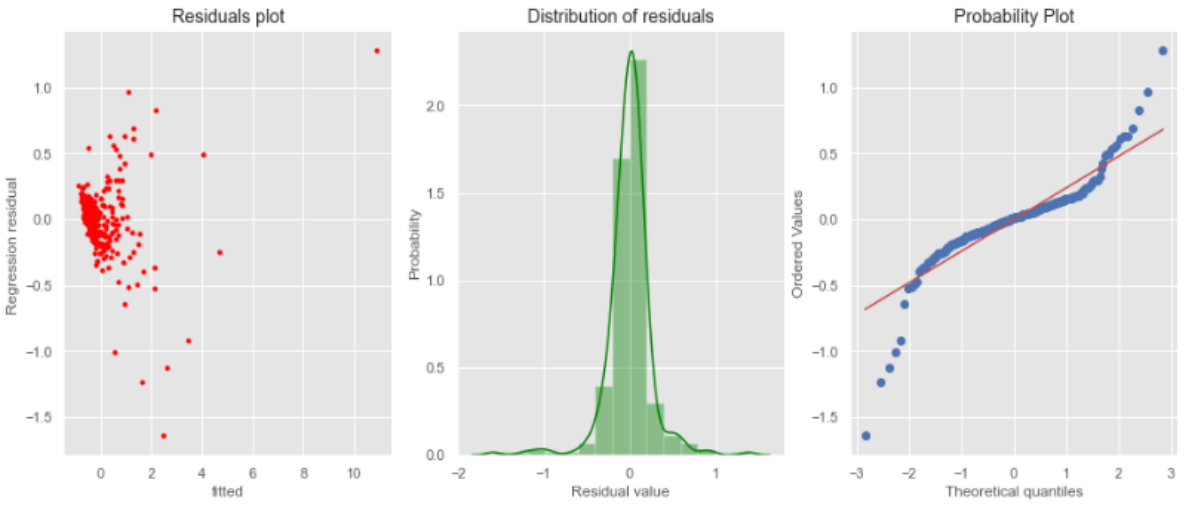
OLS Regression Results

Dep. Variable:	y	R-squared:	0.931
Model:	OLS	Adj. R-squared:	0.929
Method:	Least Squares	F-statistic:	574.5
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	6.76e-169
Time:	16:06:02	Log-Likelihood:	-25.114
No. Observations:	306	AIC:	66.23
Df Residuals:	298	BIC:	96.02
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.776e-17	0.015	1.82e-15	1.000	-0.030	0.030
x3	-0.0465	0.018	-2.567	0.011	-0.082	-0.011
x4	1.2005	0.107	11.251	0.000	0.991	1.411
x5	0.0420	0.016	2.638	0.009	0.011	0.073
x7	-0.1450	0.073	-1.980	0.049	-0.289	-0.001
x8	0.3166	0.062	5.089	0.000	0.194	0.439
x14	-0.4078	0.106	-3.857	0.000	-0.616	-0.200
x15	0.1212	0.017	7.191	0.000	0.088	0.154

Omnibus:	98.341	Durbin-Watson:	1.943
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1325.148
Skew:	-0.893	Prob(JB):	1.77e-288
Kurtosis:	13.037	Cond. No.	20.4

	Features	VIF
1	x4	49.1800
5	x14	48.2700
3	x7	23.1800
4	x8	16.7100
0	x3	1.4200
6	x15	1.2300
2	x5	1.0900



COMMENT: Dropping ‘X4’ because its VIF is 49.1800 and we want VIF less than 5 and hence rebuilding the model

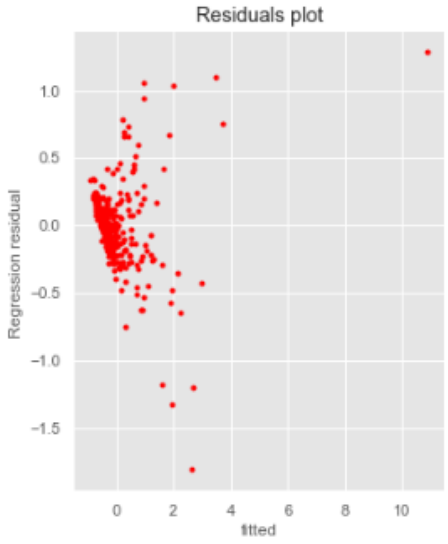
OLS Regression Results

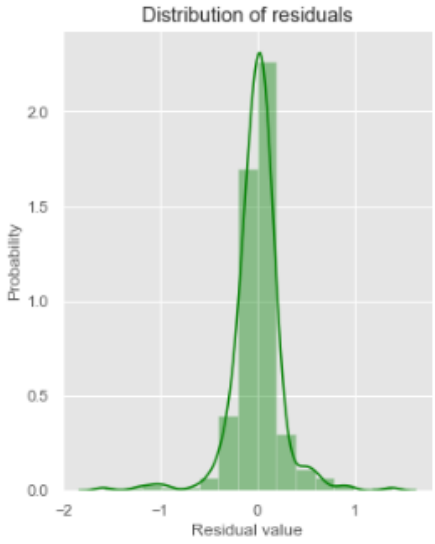
Dep. Variable:	y	R-squared:	0.902
Model:	OLS	Adj. R-squared:	0.900
Method:	Least Squares	F-statistic:	457.1
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	2.27e-147
Time:	16:06:02	Log-Likelihood:	-79.280
No. Observations:	306	AIC:	172.6
Df Residuals:	299	BIC:	198.6
Df Model:	6		
Covariance Type:	nonrobust		

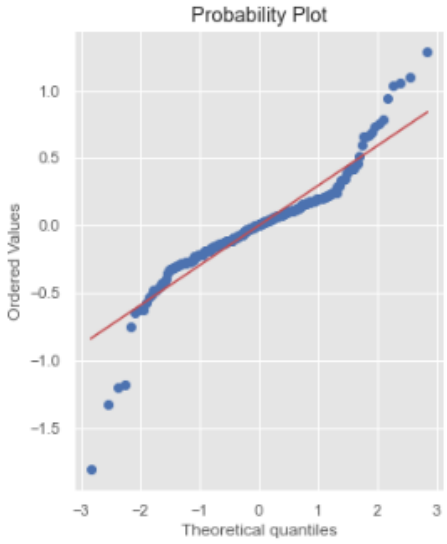
	coef	std err	t	P>|t|	[0.025	0.975]
const	2.776e-17	0.018	1.53e-15	1.000	-0.036	0.036
x3	0.0371	0.020	1.886	0.060	-0.002	0.076
x5	0.0750	0.019	4.024	0.000	0.038	0.112
x7	-0.4687	0.080	-5.838	0.000	-0.627	-0.311
x8	0.7551	0.058	13.073	0.000	0.641	0.869
x14	0.6833	0.050	13.623	0.000	0.585	0.782
x15	0.1627	0.020	8.303	0.000	0.124	0.201

Omnibus:	61.219	Durbin-Watson:	1.972
Prob(Omnibus):	0.000	Jarque-Bera (JB):	612.412
Skew:	-0.428	Prob(JB):	1.04e-133
Kurtosis:	9.878	Cond. No.	9.24

	Features	VIF
2	x7	19.6000
3	x8	10.1500
4	x14	7.6500
0	x3	1.1800
5	x15	1.1700
1	x5	1.0600







COMMENT: Dropping 'X3' because its p-value is 0.060 and we want p-value less than 0.05 and hence rebuilding the model.

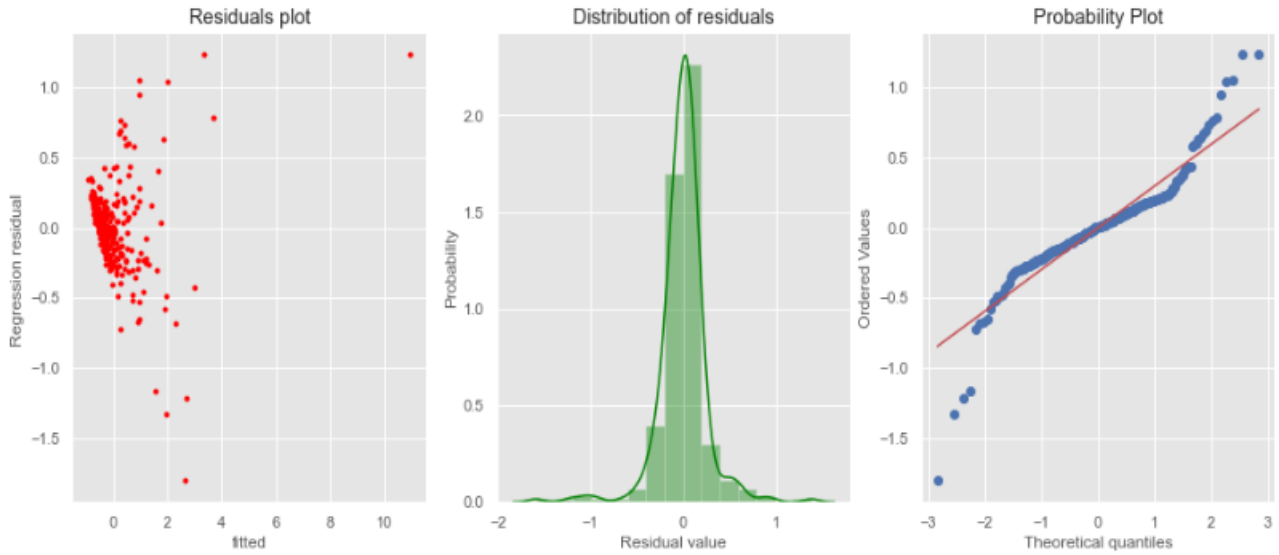
OLS Regression Results

Dep. Variable:	y	R-squared:	0.901
Model:	OLS	Adj. R-squared:	0.899
Method:	Least Squares	F-statistic:	543.2
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	5.41e-148
Time:	16:06:03	Log-Likelihood:	-81.089
No. Observations:	306	AIC:	174.2
Df Residuals:	300	BIC:	196.5
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.776e-17	0.018	1.52e-15	1.000	-0.036	0.036
x5	0.0725	0.019	3.882	0.000	0.036	0.109
x7	-0.4835	0.080	-6.026	0.000	-0.641	-0.326
x8	0.7590	0.058	13.092	0.000	0.645	0.873
x14	0.6984	0.050	14.046	0.000	0.601	0.796
x15	0.1759	0.018	9.577	0.000	0.140	0.212

Omnibus:	59.088	Durbin-Watson:	1.941
Prob(Omnibus):	0.000	Jarque-Bera (JB):	591.936
Skew:	-0.386	Prob(JB):	2.90e-129
Kurtosis:	9.770	Cond. No.	9.18

	Features	VIF
1	x7	19.4200
2	x8	10.1300
3	x14	7.4600
0	x5	1.0500
4	x15	1.0200



COMMENT: Dropping ‘X7’ because its VIF is 19.4200 and we want VIF less than 5 and hence rebuilding the model.

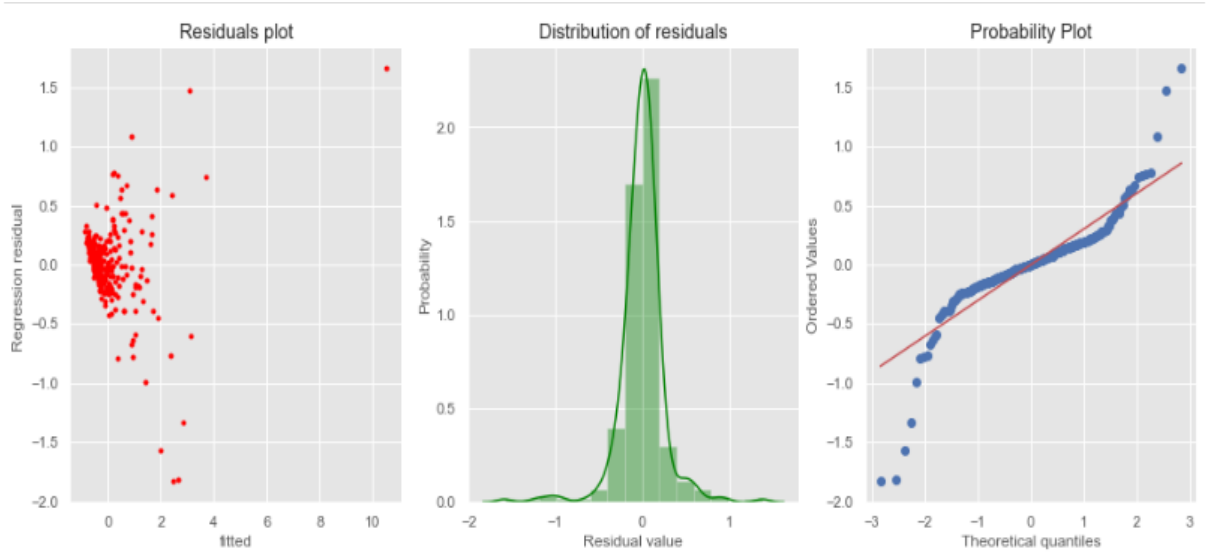
OLS Regression Results

Dep. Variable:	y	R-squared:	0.888
Model:	OLS	Adj. R-squared:	0.887
Method:	Least Squares	F-statistic:	599.6
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	5.64e-142
Time:	16:06:03	Log-Likelihood:	-98.571
No. Observations:	306	AIC:	207.1
Df Residuals:	301	BIC:	225.8
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.776e-17	0.019	1.44e-15	1.000	-0.038	0.038
x5	0.0496	0.019	2.565	0.011	0.012	0.088
x8	0.4825	0.037	12.879	0.000	0.409	0.556
x14	0.4879	0.037	13.045	0.000	0.414	0.561
x15	0.1693	0.019	8.734	0.000	0.131	0.207

Omnibus:	95.493	Durbin-Watson:	1.905
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1391.925
Skew:	-0.824	Prob(JB):	5.59e-303
Kurtosis:	13.318	Cond. No.	3.63

	Features	VIF
1	x8	3.7900
2	x14	3.7800
0	x5	1.0100
3	x15	1.0100



COMMENT: Let's drop 'X5' and see if there is any drastic fall in R squared. If not we can drop 'X5'. Our aim is to explain the maximum variance with minimum variable.

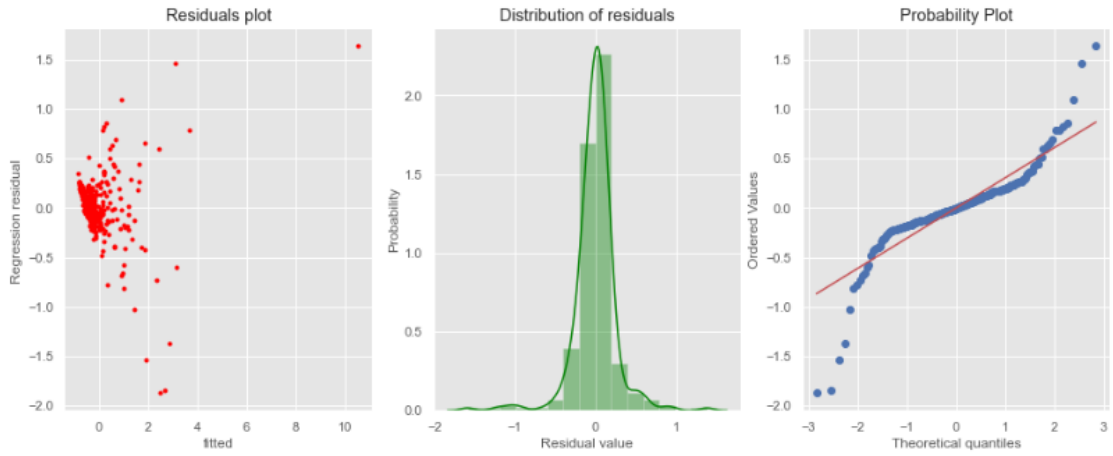
OLS Regression Results

Dep. Variable:	y	R-squared:	0.886
Model:	OLS	Adj. R-squared:	0.885
Method:	Least Squares	F-statistic:	782.8
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	4.80e-142
Time:	16:06:05	Log-Likelihood:	-101.88
No. Observations:	306	AIC:	211.8
Df Residuals:	302	BIC:	226.7
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.776e-17	0.019	1.43e-15	1.000	-0.038	0.038
x8	0.4833	0.038	12.781	0.000	0.409	0.558
x14	0.4901	0.038	12.989	0.000	0.416	0.564
x15	0.1729	0.020	8.860	0.000	0.134	0.211

Omnibus:	95.454	Durbin-Watson:	1.883
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1359.577
Skew:	-0.832	Prob(JB):	5.91e-296
Kurtosis:	13.191	Cond. No.	3.62

	Features	VIF
1	x14	3.7900
2	x15	3.7800
0	x8	1.0100



COMMENT: Now the VIFs and p-values both are within an acceptable range. So we can go ahead and make our predictions using model “lm_rfe6”(‘X8’, ‘X14’, ‘X5’ and ‘X15’) and “lm_rfe7” (‘X8’, ‘X14’ and ‘X15’).

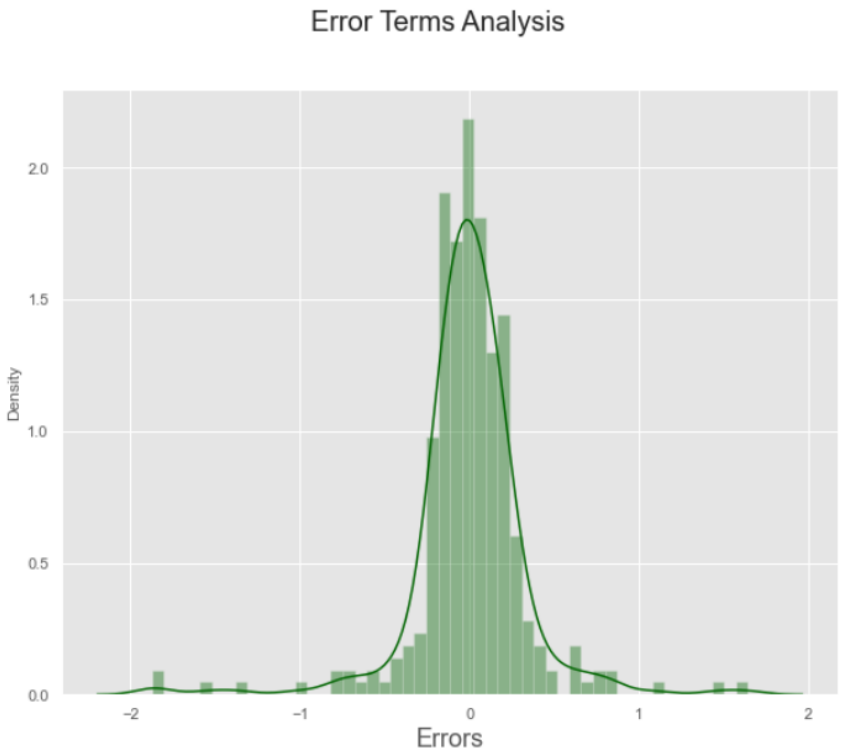
Here, we are proposing Business 2 Models which can be used to predict y

MODEL 1

- With lm_rfe7 which has basically 3 predictor variables(‘X8’, ‘X14’ and ‘X15’)

Residual Analysis of the train data:

One of our assumptions in linear regression that error terms are normally distributed, let’s visualise it by histogram of it.



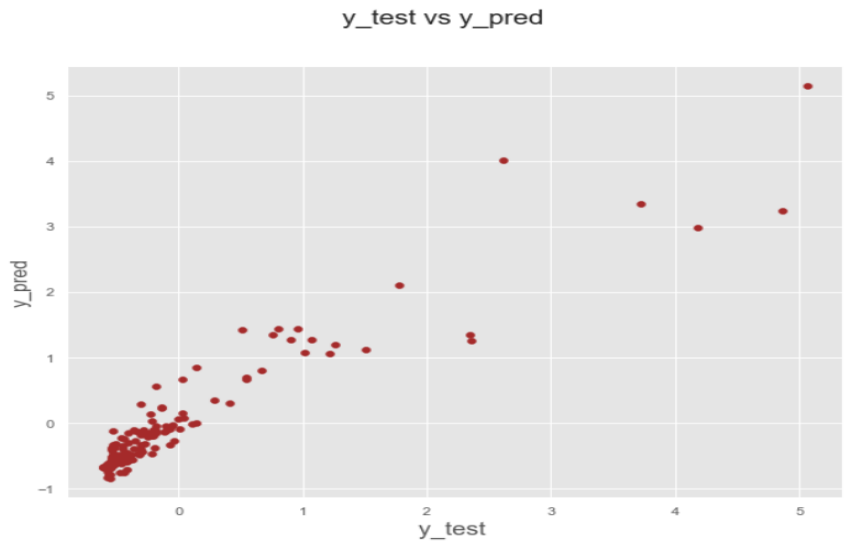
COMMENT: Now we can see errors follow normality, So it's satisfy one of linear regression assumptions.

Making Predictions Using the Final Model

- Now that we have fitted the model and checked the normality of error terms, it's time to go ahead and make predictions using the final model.
- Dividing test set into X test and Y test

Model Evaluation

Let's now plot the graph for actual versus predicted values.



COMMENT: So we can see our prediction lying approximate to y=x line. Most of data give accurate predicted value.

RMSE Score:

The R2 score of Training set is 0.886 and Test set is 0.852 which is quite close. Hence, We can say that our model is good enough to predict the Car prices using below predictor variables

- x8
- x14
- x15

Equation of Line to predict for y value

$$y=0.4833*x8+0.4901*x14+0.1729*x15$$

Model I Conclusions:

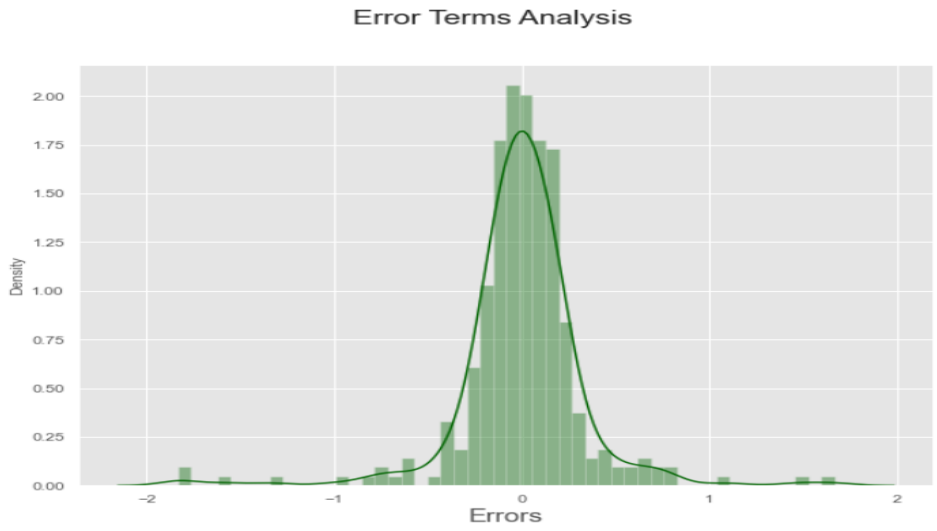
- R-squared and Adjusted R-squared - 0.886 and 0.885 - 88% variance explained.
- F-stats and Prob(F-stats) (overall model fit) - 782.8 and 4.80e-142(approx. 0.0) - Model fit is significant and explained 90%
  variance is just not by chance.
- p-values - p-values for all the coefficients seem to be less than the significance level of 0.05. - meaning that all the
  predictors are statistically significant.

MODEL II

- With lm_rfe6 which has basically 4 predictor variables('X8', 'X14', 'X5' and 'X15').

Residual Analysis of the train data

- So, now to check if the error terms are also normally distributed (which is in fact, one of the major assumptions of linear regression), let us plot the histogram of it.

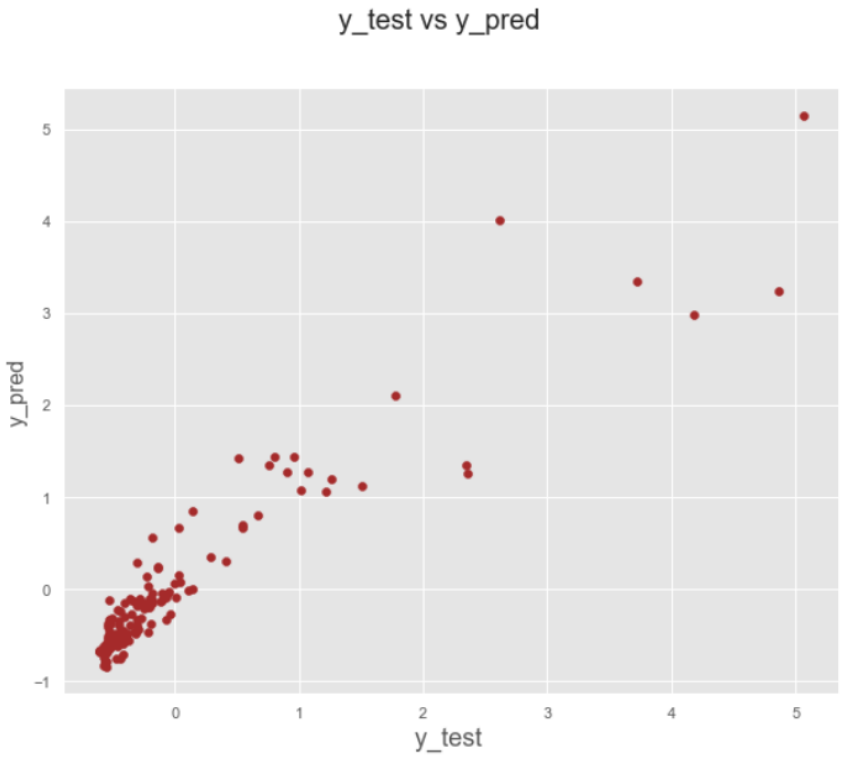


Making Predictions Using the Final Model

- Now that we have fitted the model and checked the normality of error terms, it's time to go ahead and make predictions using the model.
- Dividing test set into X test and Y test

Model Evaluation

- **Let's now plot the graph for actual versus predicted values.**



RMSE Score:

The R2 score of Training set is 0.888 and Test set is 0.882 which is quite close. Hence, We can say that our model is good enough to predict the Car prices using below predictor variables

- x5
- x8
- x14
- x15

Equation of Line to predict for y value

y=0.0496∗x5+0.4825∗x8+0.4879∗x14+0.1693∗x15

Model I Conclusions:

- R-sqaured and Adjusted R-squared - 0.888 and 0.852 - 88% variance explained.

- F-stats and Prob(F-stats) (overall model fit) - 599.6 and 5.64e-142(approx. 0.0) - Model fit is significant and explained 90%
  variance is just not by chance.
- p-values - p-values for all the coefficients seem to be less than the significance level of 0.05. - meaning that all the
  predictors are statistically significant.

Conclusion:

- Both the models are good enough to predict y-value which explains the variance of data upto 90% and the model is significant.

Closing Statement:

- Both the Forward (Manual and Computerised) and Backward approaches are good enough to forecast y values that explain up to 90% of the variation in data, and the model is significant. We received the same single model with three variables from both manual and computerised forward. As a result, we obtained one model in the forward technique and two models in the backward way.