SI 422 Regression Analysis

Topic:- Multiple Linear Regression

Supervised by:-
Prof. Siuli Mukhopadhyay
Department of Mathematics

Submitted by:-
Gagan Kumar Mishra (23N0064)
Amit Kumar Gupta (23N0070)
Raj Sharma (23N0068)
Bithal Baibhav Nayak (23N0066)

Table of Contents

Acknowledgment

I extend my sincere appreciation to our project supervisor, **Prof. Siuli Mukhopadhyay**, for their consistent support and expert guidance throughout this endeavor. Their valuable advice, feedback, and constructive criticism have significantly influenced the trajectory of this project, enhancing its quality. I am deeply thankful for their generous investment of time and expertise.

Additionally, I wish to express profound gratitude to all those who have stood by me during this project. Your encouragement, direction, and aid have been indispensable, in facilitating the successful completion of this endeavor.

Lastly, I am grateful to all the resources and references that have contributed to shaping the ideas and concepts presented in this project. The knowledge and insights gleaned from these sources have played a pivotal role in its success.

Once again, I extend my heartfelt thanks to everyone who has supported me throughout this journey. Your unwavering encouragement and assistance have made this project achievable.

Gagan Kumar Mishra (23N0064)
Amit Kumar Gupta (23N0070)
Raj Sharma (23N0068)
Bithal Baibhav Nayak (23N0066)

Date:- 02/04/2024

Chapter 1

Objective

The project aims to develop regression models that best explain the variation in the response variable, denoted as Y, using 10 potential predictor variables. Initially, the individual predictor variables will undergo analysis to determine if any unusual pattern exists that could impact model accuracy, potentially necessitating transformations.

The next step is to fit a comprehensive regression model to the data, incorporating any necessary transformations and all 10 predictor variables. The effectiveness of each predictor variable in explaining variation will be assessed. Variance inflation factors will be computed to identify and eliminate weak predictors as necessary, with the ultimate goal of identifying the most influential factors.

In this process, the full data set is utilized to select the best models corresponding to each criterion: Adjusted R-square, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Mallows' statistics, and Predicted Residual Sum of Squares (PRESS residuals). Model selection techniques such as backward elimination, forward selection, and stepwise selection are employed to optimize the model. This approach ensures the most accurate and reliable model is selected for interpreting the factors influencing the response variable.

To ensure the model meets the assumptions of linear regression, a thorough residual analysis will be conducted. This analysis will involve examining quantile plots and scatterplots to assess normality and constant variance of errors, as well as identifying and addressing outliers using Cook's Distance. Of particular interest will be identifying outliers with significant impacts on the regression.

Chapter 2

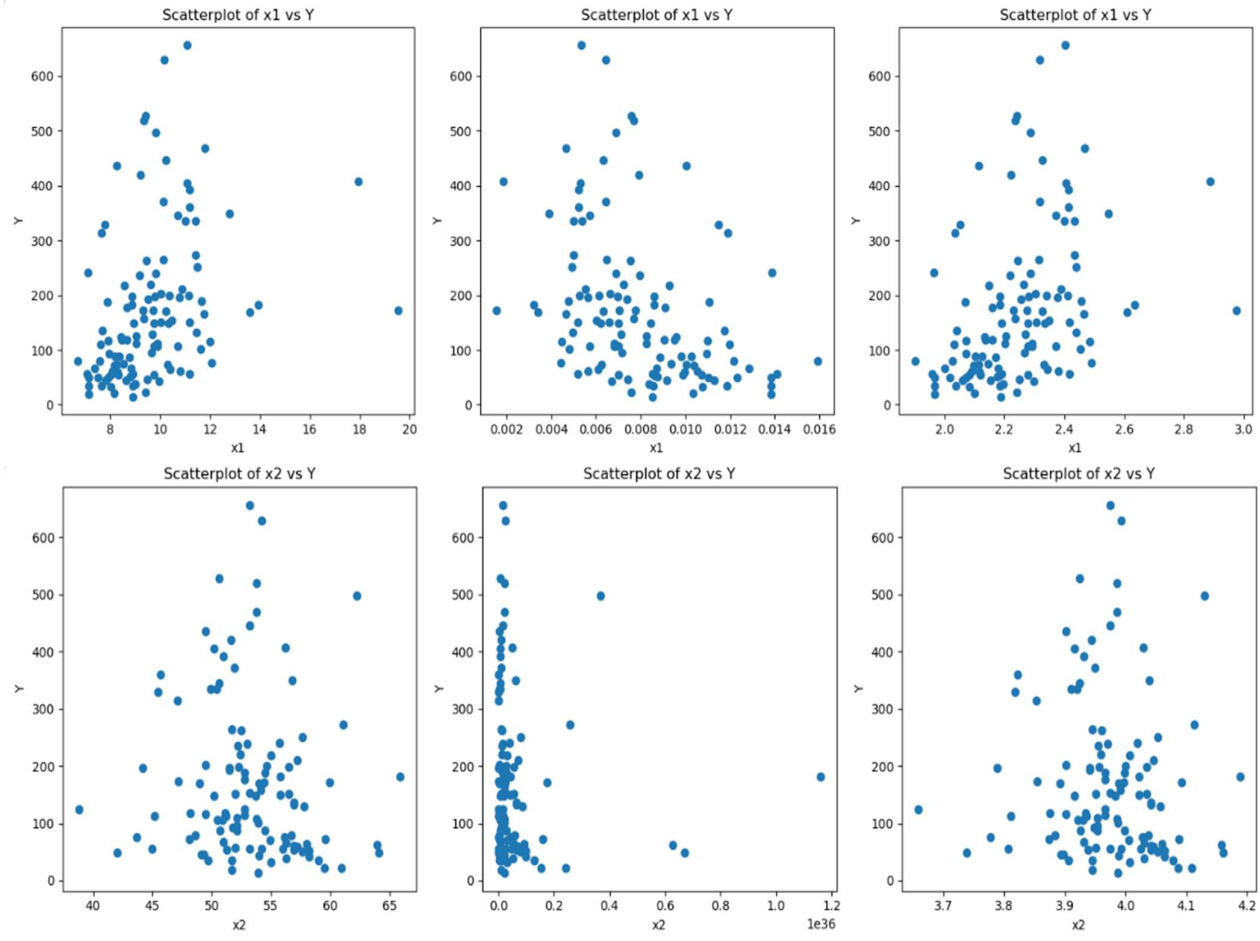
Exploratory Data Analysis

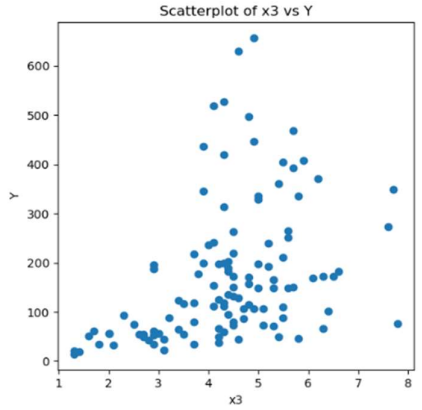
In this part firstly we check the shape of the data and it was came out to be 113 rows and 11 columns. Also we check if there are **Missing values and Duplicate values** in the data set, improper handling of missing observations can lead to inaccurate inferences about the data. Since there were no duplicate and missing values we can safely proceed to further analysis.

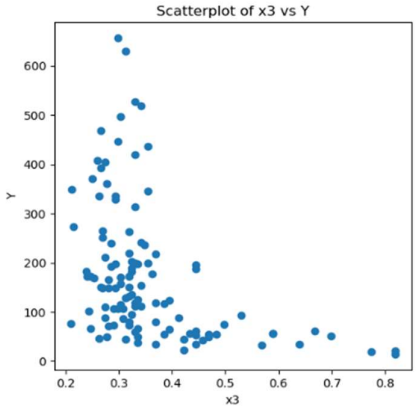
After that we have plotted the scatter plot between independent variable and all other predictor variables. Where we have side by side plotted individual variable using raw data, after applying Box Tidwell and log transformations.

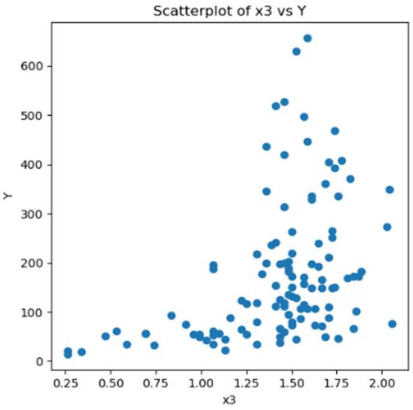
Univariate Analysis for Numerical Columns

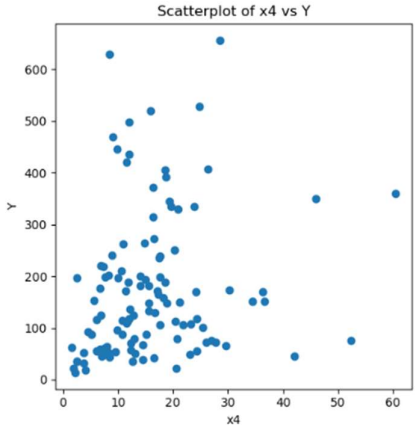
Raw data | **Box Tidwell** | **Log X**

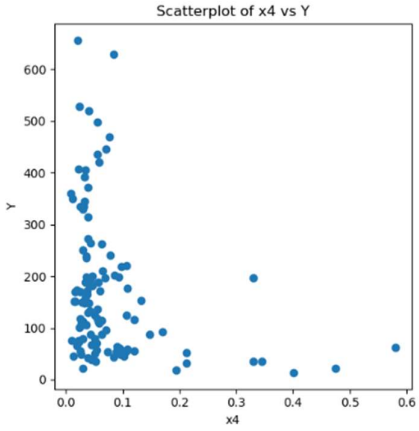


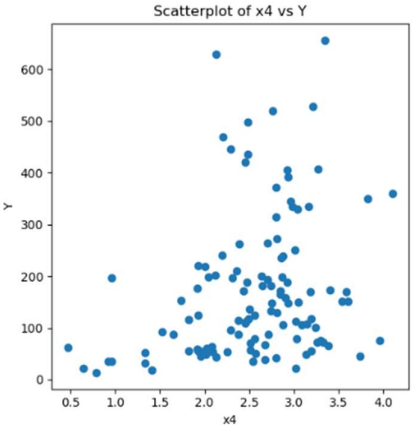

Scatterplot of x3 vs Y
Y
x3

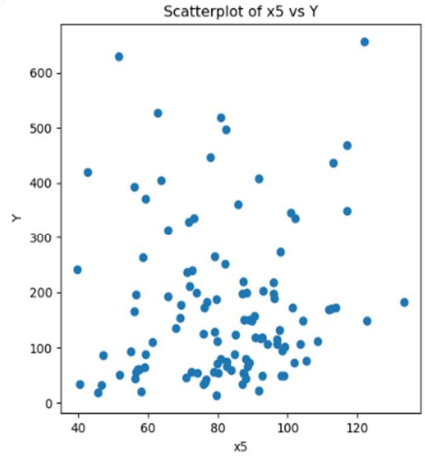

Scatterplot of x3 vs Y
Y
x3

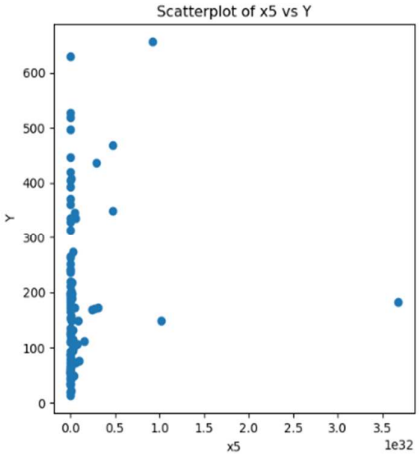

Scatterplot of x3 vs Y
Y
x3

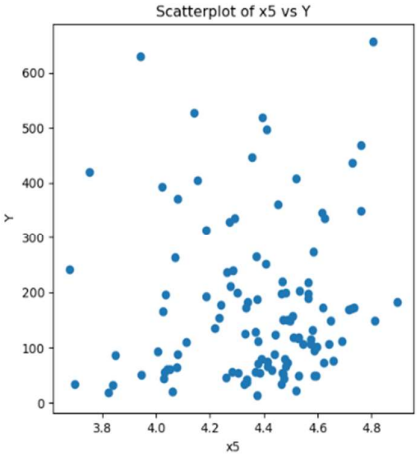

Scatterplot of x4 vs Y
Y
x4

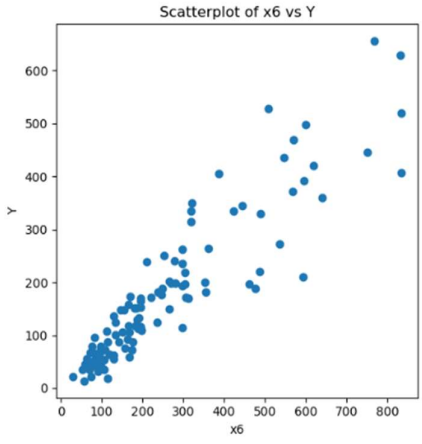

Scatterplot of x4 vs Y
Y
x4

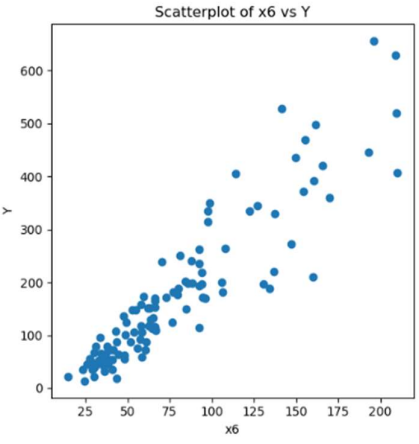

Scatterplot of x4 vs Y
Y
x4

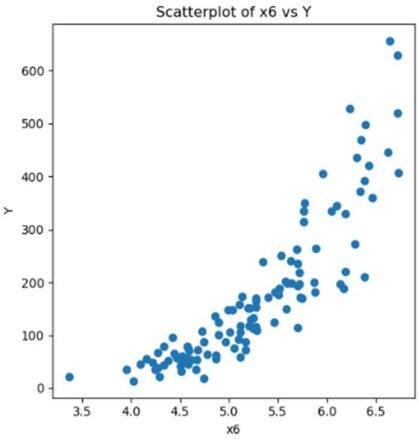

Scatterplot of x5 vs Y
Y
x5

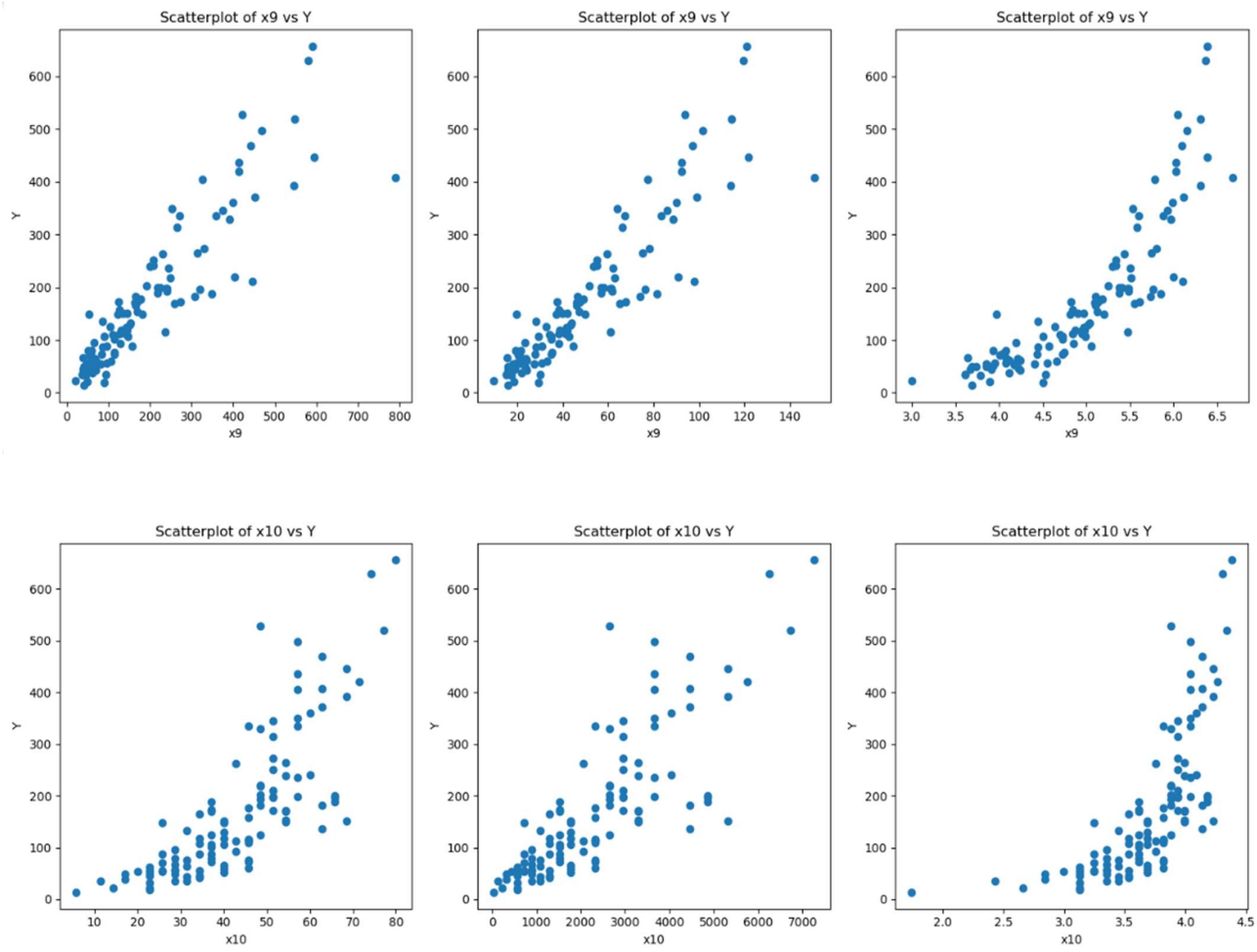

Scatterplot of x5 vs Y
Y
x5
1e32


Scatterplot of x5 vs Y
Y
x5


Scatterplot of x6 vs Y
Y
x6


Scatterplot of x6 vs Y
Y
x6


Scatterplot of x6 vs Y
Y
x6



Conclusions:-

As we can see from the graphs that box Tidwell was effective over some variables and over some variables log transformations was effective and also there are some variables where none of the transformations were effective.

So we have decided that over X1 and X4 we will use log transformation and over X6, X9 and X10 we will use Box Tidwell transformations.

Categorical Analysis

There were two categorical columns namely X7 and X8 where X7 was containing categories 1 and 2. Whereas X8 was containing 1,2,3 and 4.

So corresponding to each column we have made C-1 dummy variables where C is number of categories in that column.

- **For X7**

$$X7\_1 = \begin{cases} 1 & ,when\ X7 = 1 \\ 0 & ,when\ X7 = 2 \end{cases}$$

- **For X8**

$$X8\_1 = \begin{cases} 1 & ,when\ X8 = 1 \\ 0 & ,otherwise \end{cases}$$

$$X8\_2 = \begin{cases} 1 & ,when\ X8 = 2 \\ 0 & ,otherwise \end{cases}$$

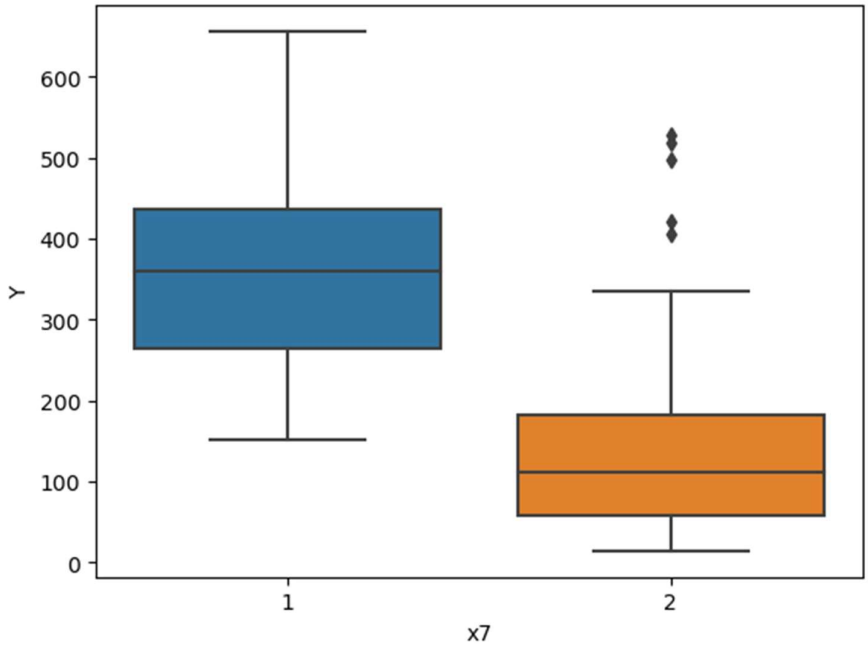
$$X8\_3 = \begin{cases} 1 & ,when\ X8 = 3 \\ 0 & ,otherwise \end{cases}$$

And the table for encoded variables is attached below.

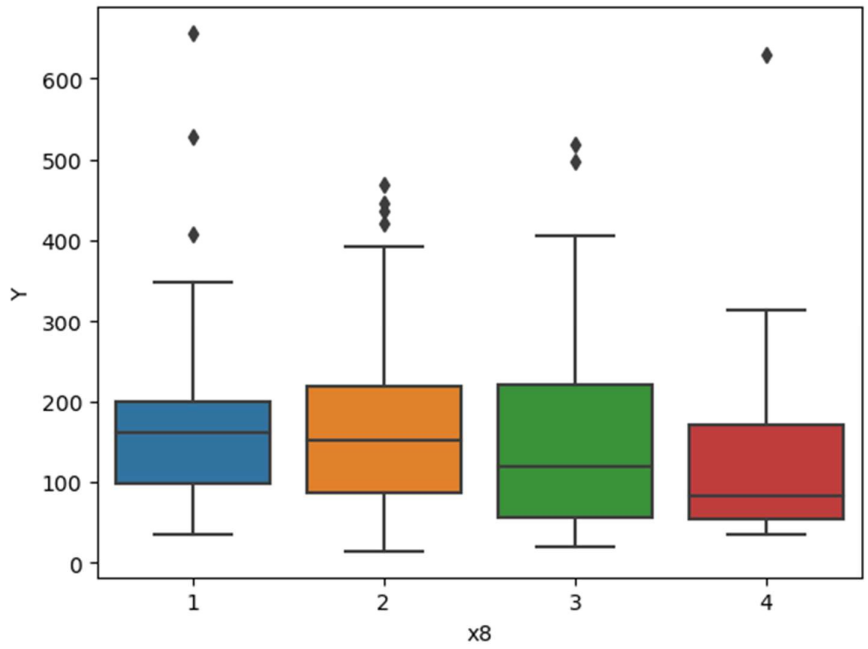
x7_1	x8_1	x8_2	x8_3
0	0	0	0
0	0	1	0
0	0	0	1
0	0	0	0
0	1	0	0

Univariate Analysis for categorical variables

Boxplot of Y with respect to X7



Boxplot of Y with respect to X8



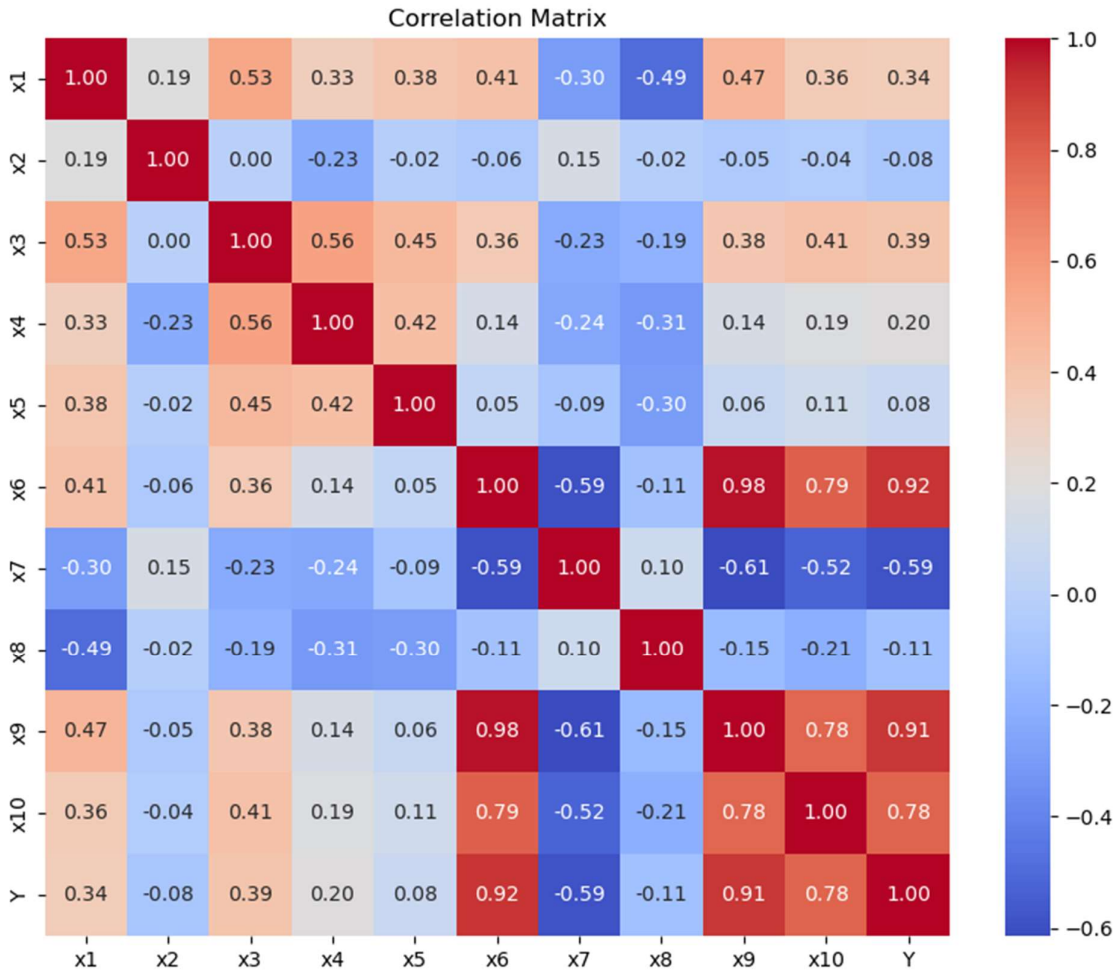
Conclusions:-

It can be seen from the above boxplots that for column X7 there were two categories

And the average for y variable in case of category 1 is larger than category 2. And for X8 column we can see that there is downward trend in the average of y for categories 1,2,3 and 4

Heat map

Here we have also plotted heatmap to see which variables are highly correlated with y and also among themselves.



Conclusions:-

X9 is related to both X6 and X10 higly so there is possibility of multicollinearity

Also X6 and X10 are also correlated highly.

And y is correlated highly with X6,X9 and X10 so they will be important in analysis.

Chapter 3

Data Preprocessing

Outlier and Influential Observation Analysis using Cook's Distance:-

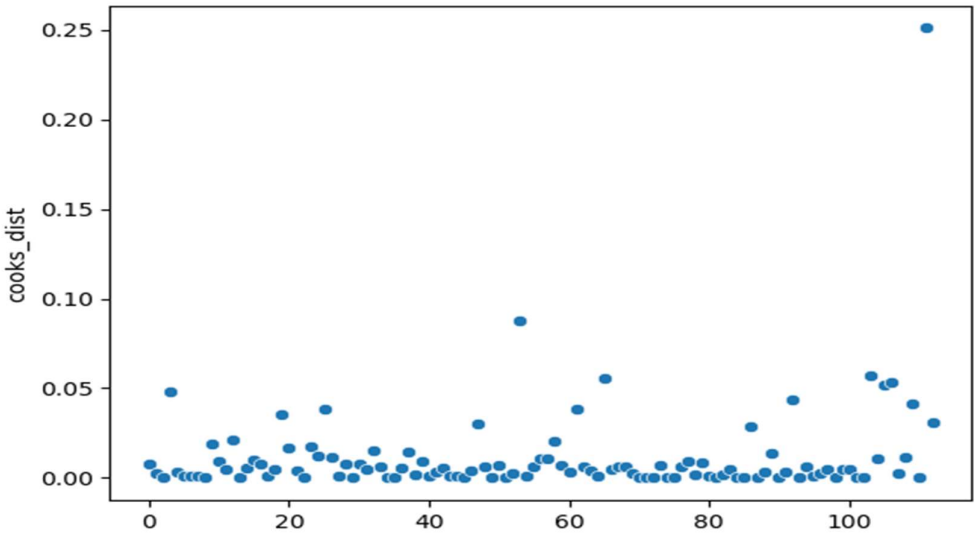
Cook's distance for the *i*th observation can be calculated using the following formula:

$$Di = \frac{\sum_{j=1}^{n}(\widehat{Yj} - \widehat{Yj(i)})^2}{p * MSE}$$

- $\widehat{Yj}$ is the predicted value for the *j*th observation based on the full model.
- $\hat{Y}\ j(i)$ is the predicted value for the *j*th observation based on the model with the *i*th observation removed.
- p is number of parameters and MSE is the mean squared error of full model.
- Identify observations with Cook's distance greater than a chosen threshold (e.g., 4/n, where n is the sample size) as potentially influential.

Analysis:

When we used usual threshold that is 4/n (=0.03) then we got 11 influential observation and removing them will lead to significant loss in data. So we decided to increase the threshold to 0.06 and we got only 2 influential observation. And after deleting that we moved to further analysis.



Outlier and Influential Observation Analysis using DFFITS method :-

The DFFITS (difference of fits) is another approach to deal with influential observations.

For the point 'i' we calculate DFFITS values as follows:

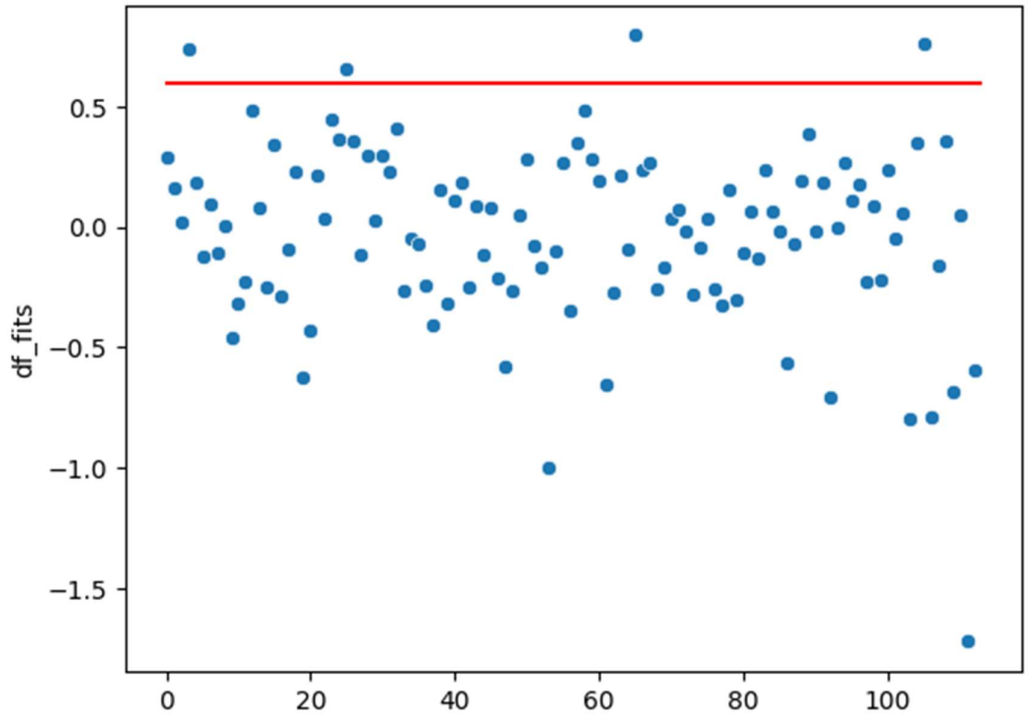
$$DFFITS_i = \frac{y_i^{(i)} - y_i}{s_{(i)}\sqrt{h_{ii}}}$$

where:

- yi(i) is the predicted value for point i when point i is left out of the regression.
- yi is the predicted value for point i with point i included in the regression.
- s(i) is the standard error estimated without the point in question.
- hii is the leverage for the point

Typically, observations that have DFFITS values greater than a threshold of $2\sqrt{\frac{p}{n}}$ are investigated, where p is the number of predictor variables used in the model and n is the number of observations used in the model.

The Graph for DFFITS with the threshold is shown below.



Chapter 4

Model Fitting

Univariate Model Fitting

1. Y vs X1

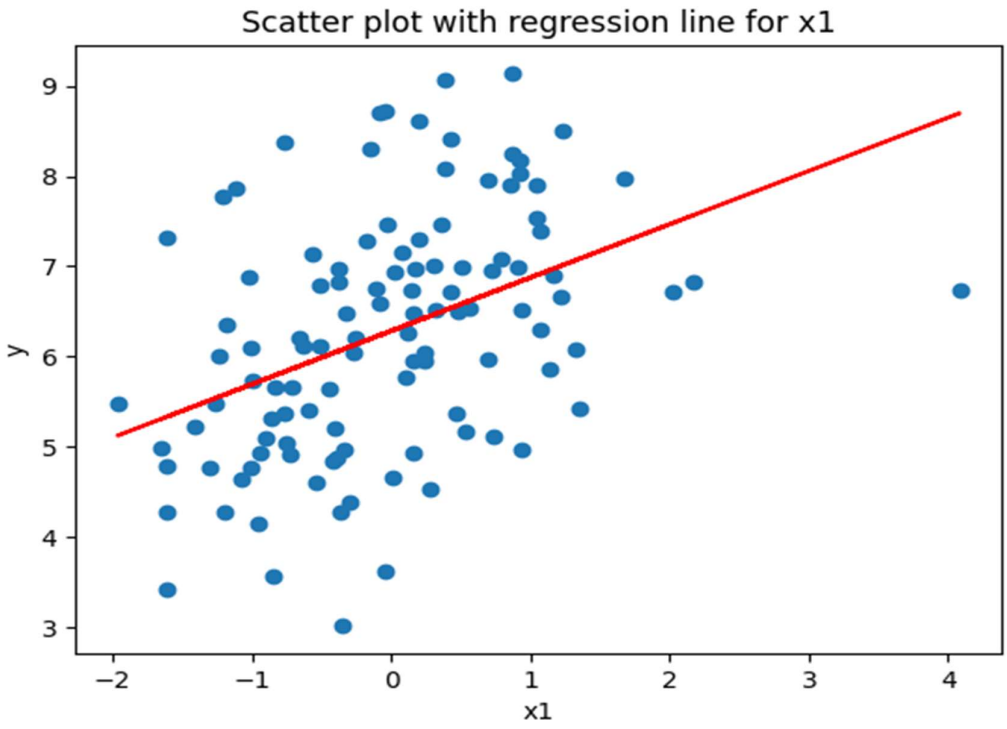
Equation of fitted model is :- $Y = 6.2838 + 0.5908 * \text{X1}$

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.175
Model:	OLS	Adj. R-squared:	0.167
Method:	Least Squares	F-statistic:	23.27
Date:	Sat, 27 Apr 2024	Prob (F-statistic):	4.56e-06
Time:	05:38:00	Log-Likelihood:	-180.60
No. Observations:	112	AIC:	365.2
Df Residuals:	110	BIC:	370.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.2838	0.116	54.278	0.000	6.054	6.513
x1	0.5908	0.122	4.824	0.000	0.348	0.833

Omnibus:	0.765	Durbin-Watson:	2.003
Prob(Omnibus):	0.682	Jarque-Bera (JB):	0.868
Skew:	0.111	Prob(JB):	0.648
Kurtosis:	2.631	Cond. No.	1.07



2. Y vs X2 :-

Equation for the fitted model is :- $Y = 6.2637 + -0.1881 * \text{X2}$

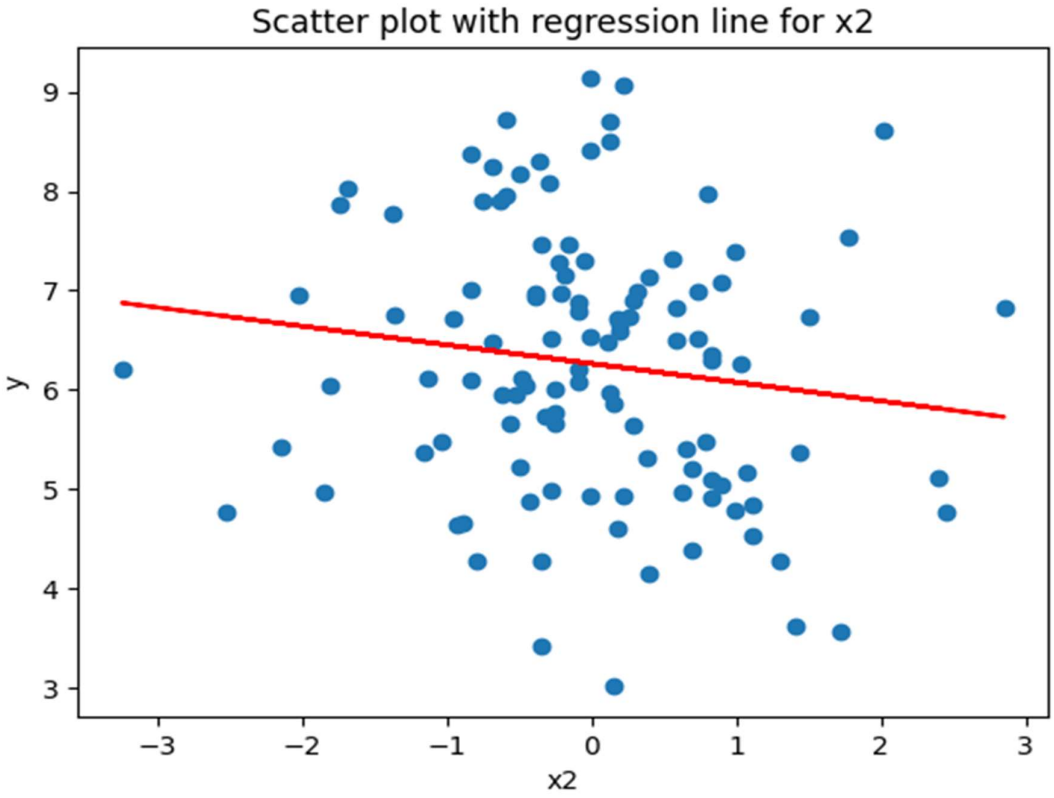
Regression results for x2:

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.020
Model:	OLS	Adj. R-squared:	0.011
Method:	Least Squares	F-statistic:	2.237
Date:	Sat, 27 Apr 2024	Prob (F-statistic):	0.138
Time:	05:38:00	Log-Likelihood:	-190.21
No. Observations:	112	AIC:	384.4
Df Residuals:	110	BIC:	389.9
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.2637	0.126	49.680	0.000	6.014	6.514
x2	-0.1881	0.126	-1.496	0.138	-0.437	0.061

Omnibus:	2.065	Durbin-Watson:	2.009
Prob(Omnibus):	0.356	Jarque-Bera (JB):	1.500
Skew:	0.033	Prob(JB):	0.472
Kurtosis:	2.437	Cond. No.	1.01



3. Y vs X3 :-

Equation for the fitted model is :- $Y = 6.2721 + 0.7040 * \text{X3}$

Regression results for x3:

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.277
Model:	OLS	Adj. R-squared:	0.270
Method:	Least Squares	F-statistic:	42.13
Date:	Sat, 27 Apr 2024	Prob (F-statistic):	2.54e-09
Time:	05:38:01	Log-Likelihood:	-173.18
No. Observations:	112	AIC:	350.4
Df Residuals:	110	BIC:	355.8
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.2721	0.108	57.914	0.000	6.058	6.487
x3	0.7040	0.108	6.491	0.000	0.489	0.919

Omnibus:	1.188	Durbin-Watson:	2.096
Prob(Omnibus):	0.552	Jarque-Bera (JB):	1.253
Skew:	0.234	Prob(JB):	0.534
Kurtosis:	2.776	Cond. No.	1.01

4. Y vs X4

Equation for the fitted model is :- $Y = 6.2694 + 0.4901 * \text{X4}$

Regression results for x4:

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.135
Model:	OLS	Adj. R-squared:	0.127
Method:	Least Squares	F-statistic:	17.09
Date:	Sat, 27 Apr 2024	Prob (F-statistic):	6.97e-05
Time:	05:38:01	Log-Likelihood:	-183.25
No. Observations:	112	AIC:	370.5
Df Residuals:	110	BIC:	375.9
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.2694	0.118	52.912	0.000	6.035	6.504
x4	0.4901	0.119	4.135	0.000	0.255	0.725

Omnibus:	2.318	Durbin-Watson:	2.097
Prob(Omnibus):	0.314	Jarque-Bera (JB):	1.789
Skew:	0.141	Prob(JB):	0.409
Kurtosis:	2.449	Cond. No.	1.01

5. Y vs X5

Equation for the fitted model is :- $Y = 6.2658 + 0.2006 * \text{X5}$

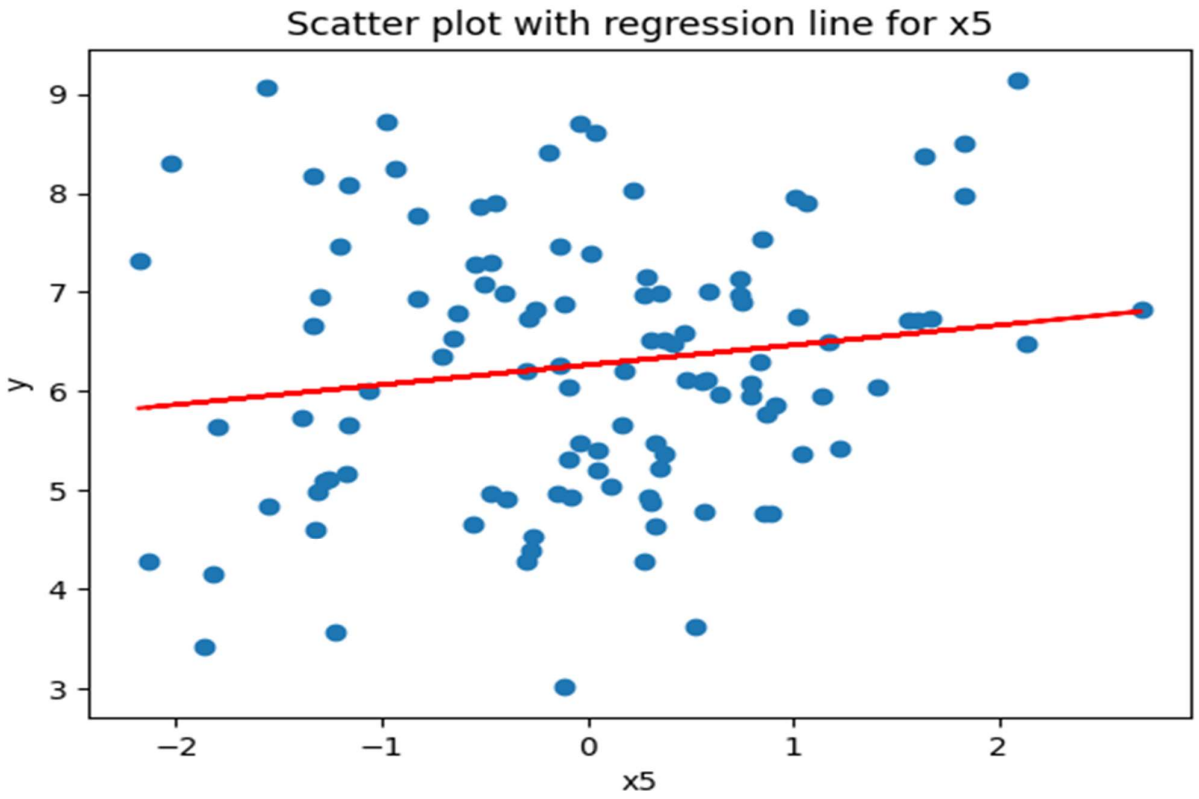
Regression results for x5:

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.023
Model:	OLS	Adj. R-squared:	0.014
Method:	Least Squares	F-statistic:	2.554
Date:	Sat, 27 Apr 2024	Prob (F-statistic):	0.113
Time:	05:38:01	Log-Likelihood:	-190.06
No. Observations:	112	AIC:	384.1
Df Residuals:	110	BIC:	389.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.2658	0.126	49.766	0.000	6.016	6.515
x5	0.2006	0.126	1.598	0.113	-0.048	0.449

Omnibus:	2.173	Durbin-Watson:	2.033
Prob(Omnibus):	0.337	Jarque-Bera (JB):	1.667
Skew:	0.117	Prob(JB):	0.435
Kurtosis:	2.450	Cond. No.	1.01



6. Y vs X6

Equation for the fitted model is :- $Y = 6.2956 + 1.2151 * X6$

Regression results for x6:

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.775
Model:	OLS	Adj. R-squared:	0.773
Method:	Least Squares	F-statistic:	379.3
Date:	Sat, 27 Apr 2024	Prob (F-statistic):	1.92e-37
Time:	05:38:01	Log-Likelihood:	-107.76
No. Observations:	112	AIC:	219.5
Df Residuals:	110	BIC:	225.0
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.2956	0.060	104.222	0.000	6.176	6.415
x6	1.2151	0.062	19.476	0.000	1.091	1.339

Omnibus:	6.233	Durbin-Watson:	1.871
Prob(Omnibus):	0.044	Jarque-Bera (JB):	5.769
Skew:	-0.538	Prob(JB):	0.0559
Kurtosis:	3.283	Cond. No.	1.04

7. Y vs X9

Equation for the fitted model is :- $Y = 6.3035 + 1.2602 * \text{X9}$

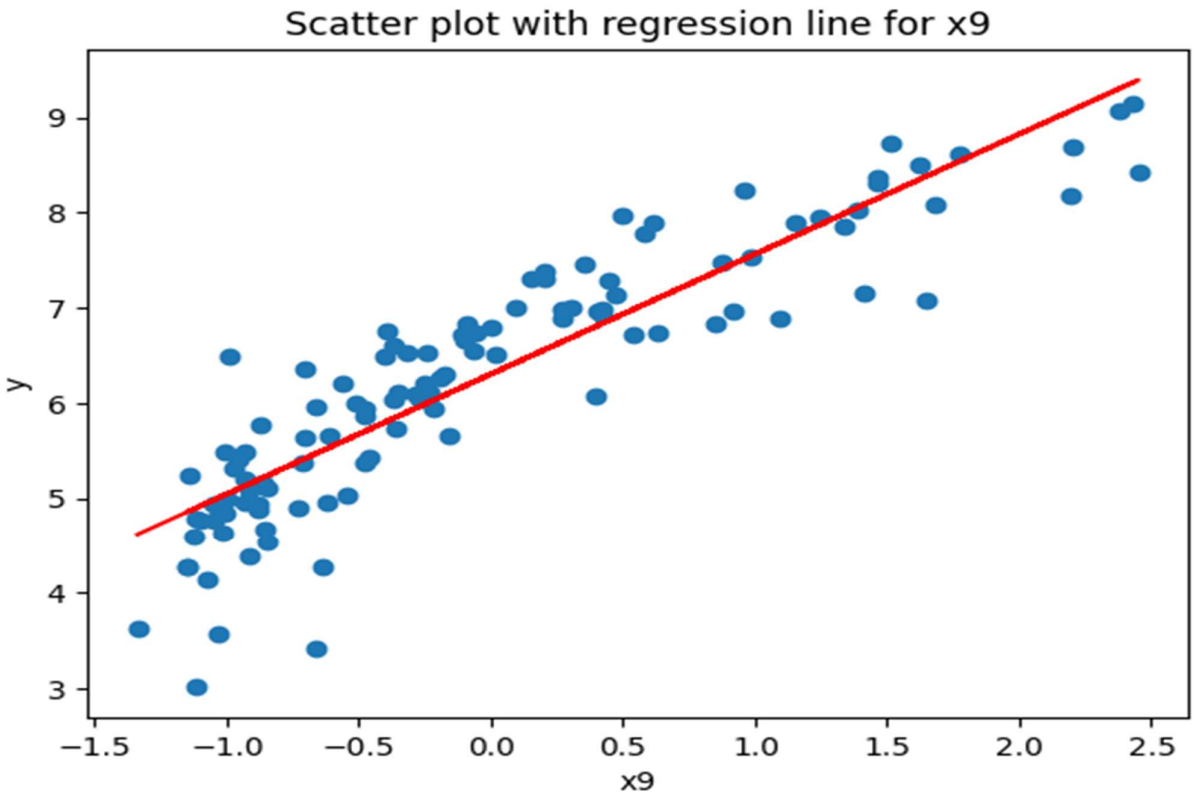
Regression results for x9:

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.804
Model:	OLS	Adj. R-squared:	0.802
Method:	Least Squares	F-statistic:	450.4
Date:	Sat, 27 Apr 2024	Prob (F-statistic):	1.09e-40
Time:	05:38:02	Log-Likelihood:	-100.17
No. Observations:	112	AIC:	204.3
Df Residuals:	110	BIC:	209.8
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.3035	0.056	111.656	0.000	6.192	6.415
x9	1.2602	0.059	21.222	0.000	1.143	1.378

Omnibus:	13.755	Durbin-Watson:	1.918
Prob(Omnibus):	0.001	Jarque-Bera (JB):	15.718
Skew:	-0.736	Prob(JB):	0.000386
Kurtosis:	4.097	Cond. No.	1.06



8. Y vs X10 :-

Equation for the fitted model is :- $Y = 6.2783 + 1.0928 * \text{X10}$

Regression results for x10:

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.664
Model:	OLS	Adj. R-squared:	0.661
Method:	Least Squares	F-statistic:	217.3
Date:	Sat, 27 Apr 2024	Prob (F-statistic):	8.34e-28
Time:	05:38:02	Log-Likelihood:	-130.28
No. Observations:	112	AIC:	264.6
Df Residuals:	110	BIC:	270.0
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.2783	0.074	85.030	0.000	6.132	6.425
x10	1.0928	0.074	14.741	0.000	0.946	1.240

Omnibus:	0.074	Durbin-Watson:	1.684
Prob(Omnibus):	0.964	Jarque-Bera (JB):	0.161
Skew:	0.058	Prob(JB):	0.923
Kurtosis:	2.856	Cond. No.	1.01

Full Model Fitting

Here we have fitted the model considering all the variables with necessary transformations like Box Tidwell and log on continuous type X variables, Boxcox on the output Y variable. We have also encoded the categorical variables into dummy variables.

The model summary is as shown below.

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.877
Model:	OLS	Adj. R-squared:	0.862
Method:	Least Squares	F-statistic:	58.30
Date:	Wed, 01 May 2024	Prob (F-statistic):	4.03e-39
Time:	20:46:26	Log-Likelihood:	-73.573
No. Observations:	111	AIC:	173.1
Df Residuals:	98	BIC:	208.4
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.5386	0.155	42.156	0.000	6.231	6.846
x1	-0.0192	0.079	-0.241	0.810	-0.177	0.138
x2	-0.0897	0.055	-1.633	0.106	-0.199	0.019
x3	0.2168	0.079	2.754	0.007	0.061	0.373
x4	0.1347	0.071	1.898	0.061	-0.006	0.276
x5	-0.0139	0.058	-0.237	0.813	-0.130	0.102
x6	-0.2147	0.297	-0.724	0.471	-0.804	0.374
x7_1	-0.2608	0.174	-1.500	0.137	-0.606	0.084
x8_1	-0.2063	0.203	-1.015	0.313	-0.610	0.197
x8_2	-0.1779	0.175	-1.018	0.311	-0.525	0.169
x8_3	-0.2679	0.173	-1.548	0.125	-0.611	0.076
x9	1.1745	0.305	3.851	0.000	0.569	1.780
x10	0.2835	0.095	2.978	0.004	0.095	0.472

Omnibus:	4.205	Durbin-Watson:	2.084
Prob(Omnibus):	0.122	Jarque-Bera (JB):	3.978
Skew:	-0.402	Prob(JB):	0.137
Kurtosis:	2.539	Cond. No.	16.9

As it can be observed many of the variables seem insignificant according to their p-values. So we proceeded towards selecting the subset of variables that best explain the variance of the response variable. This was done on the basis of the following criteria

- Adjusted R-squared
- Mallow Statistics
- AIC and BIC criterion
- PRESS Residuals

Chapter 5

Model Selection

Part A :- Based on criteria

Adjusted R Squared

Adjusted R-squared is a statistical measure used in the context of multiple regression analysis. It represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

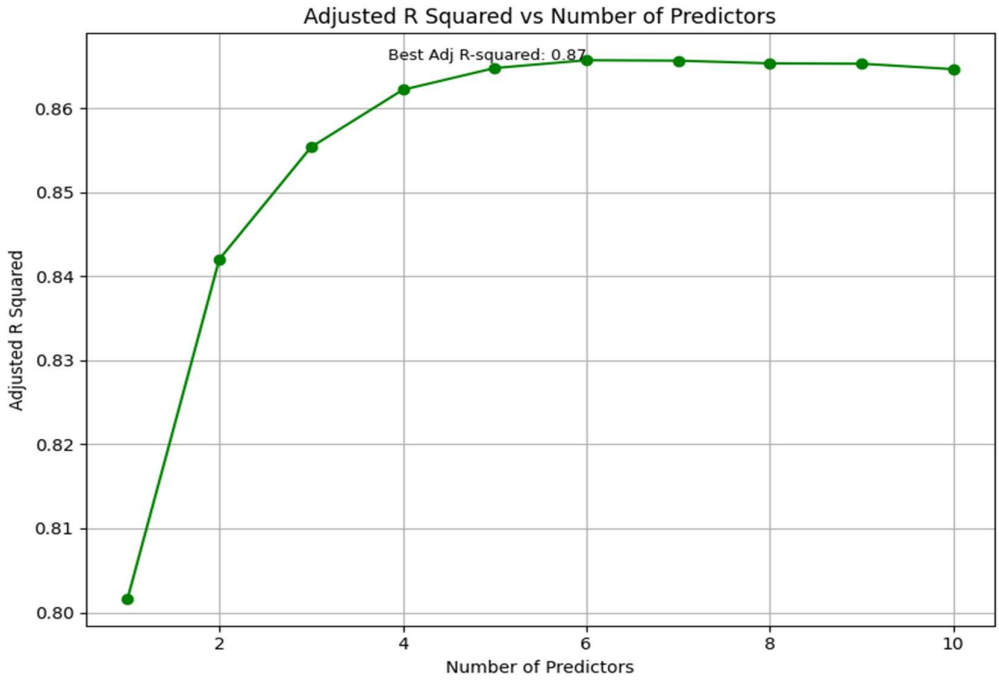
The mathematical formula for Adjusted R-squared is:

$$\text{Adjusted } R^2 = 1 - \left[\frac{(1 - R^2)(n - 1)}{n - k - 1}\right]$$

where:

- R2 is the R-squared of the model.
- n is the number of observations.
- k is the number of predictor variables.

Below is the graph that plots the maximum adjusted R squared value obtained among all the models for a particular number of regressors.



Mallow's Statistic

Mallows' Cp, named after Colin Lingwood Mallows, is a statistical measure used in the context of multiple regression analysis to assess the fit of a regression model that has been estimated using ordinary least squares. It is particularly useful in model selection, where a number of predictor variables are available for predicting an outcome, and the goal is to find the best model involving a subset of these predictors.

The mathematical formula for Mallows' Cp is:

$$Cp = \frac{SSE_P}{MSE_F} - (N - 2p)$$

SSE_p: The sum of squared errors for the reduced or potential model.

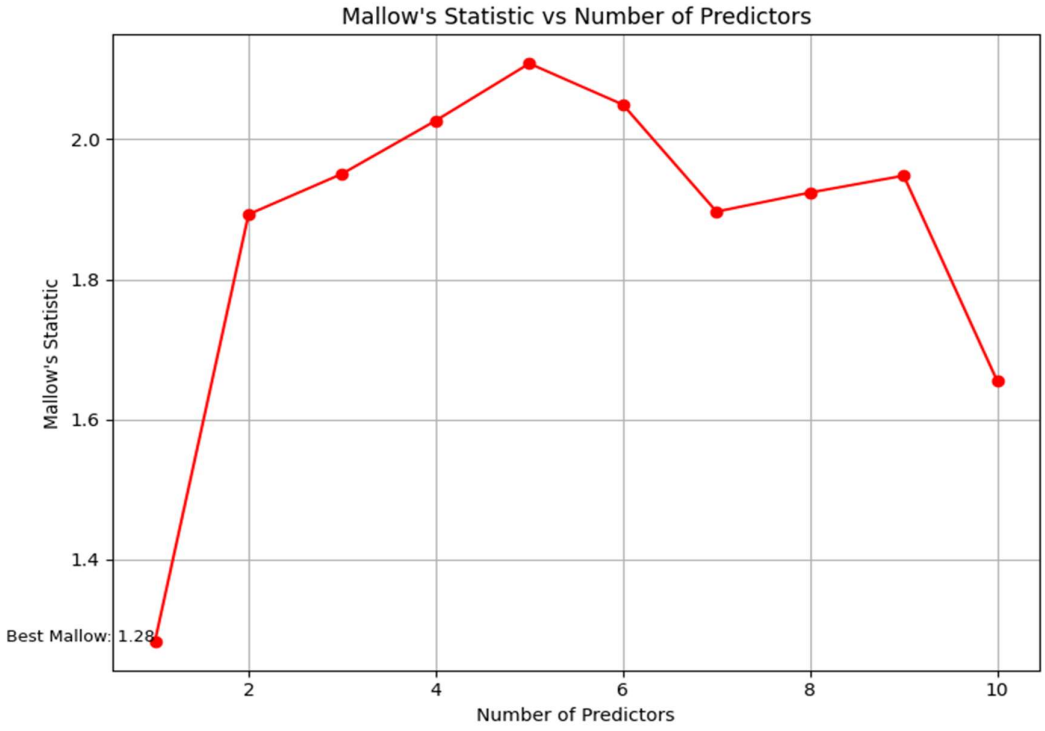
MSE_f: The mean squared error for the full model.

p: The count of predictors. The penalty term "n-2p" imposes a cost on models that incorporate a higher number of predictors.

N: Number of observations.

The "best" regression model is identified as the model with the lowest Cp value that is closest to P+1, where P is the number of predictor variables in the model.

Below is the graph that plots the lowest |Cp-p| value obtained among all the models for a particular number of regressors.



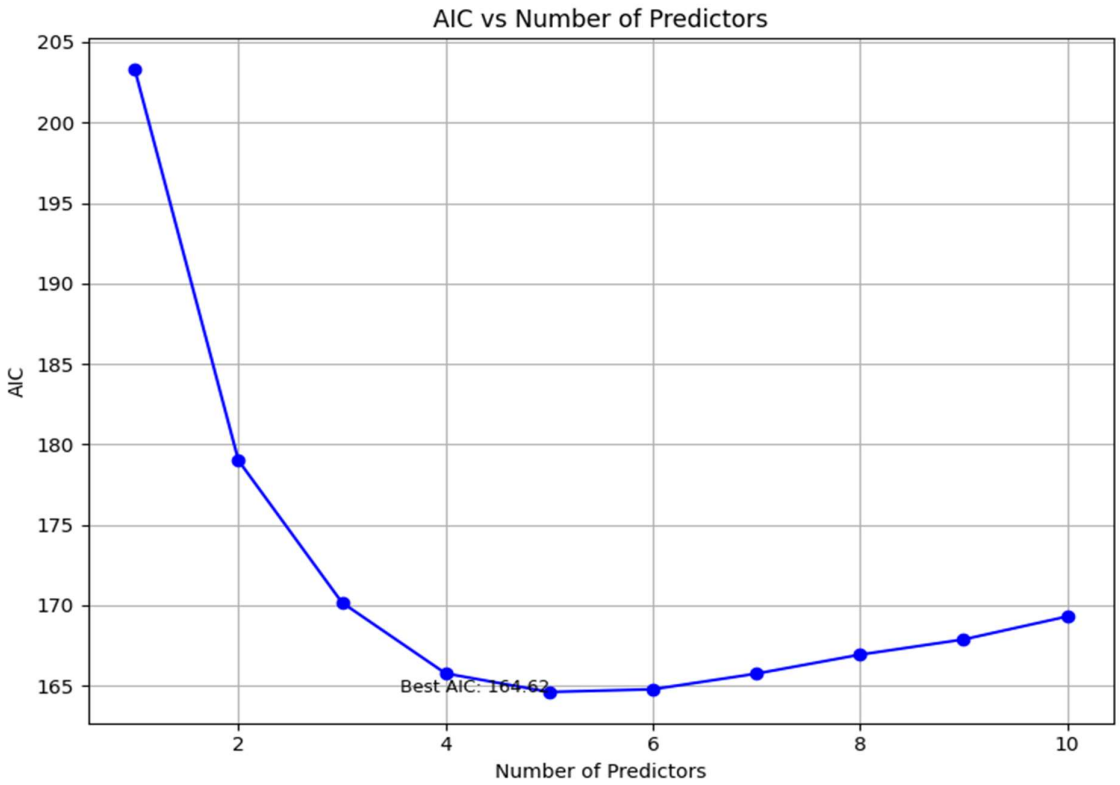
Akaike Information Criterion (AIC)

- It provides a means for model selection by estimating the quality of each model.
- AIC deals with the trade-off between the goodness of fit of the model and the simplicity of the model.
- Mathematical form of AIC:

$$AIC_p = n \ln SSE_p - n \ln n + 2p$$

- The decision rule for AIC is to choose the model with the minimum AIC_p value for each p and and for overall selection choose the one with smallest AIC value.

Below is the graph that plots the lowest AIC value obtained among all the models for a particular number of regressors.



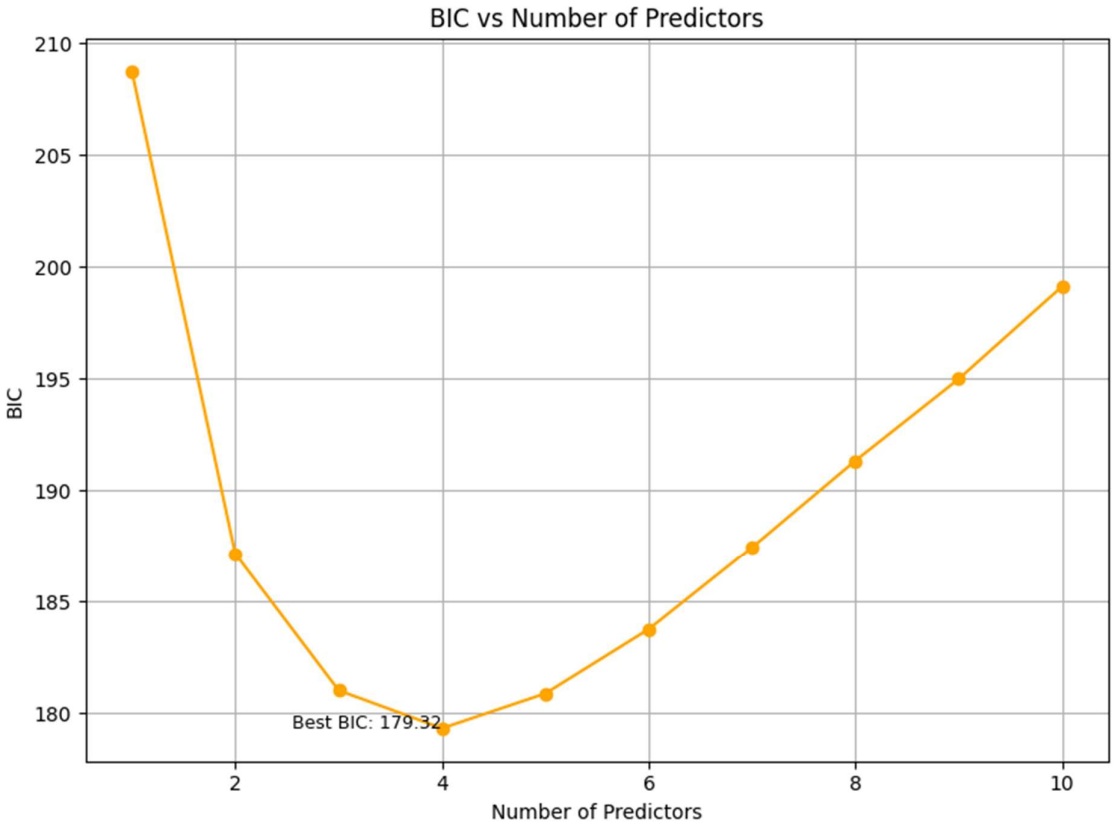
Bayesian Information Criterion (BIC)

- BIC is also a criterion for model selection among a finite set of models.
- Its mathematical form is :-

$$SBC_p = n \ln SSE_p - n \ln n + [\ln n]p$$

- In the above expression there is an expression for SSE_p which varies with p and it decreases with increase in p and so is $\ln(SSE_p)$.
- Models with lower BIC are generally preferred.
- The decision rule for BIC is to choose the model with the minimum BIC_p value for each p.

Below is the graph that plots the lowest BIC value obtained among all the models for a particular number of regressors.



PRESS residuals

The Predicted Residual Error Sum of Squares (PRESS) is a form of cross-validation used in regression analysis to provide a summary measure of the fit of a model to a sample of observations that were not themselves used to estimate the model. It is a model validation method used to assess a model's predictive ability.

The mathematical formula for PRESS is

$$e_{(i)} = y_i - \hat{y}_{(i)}$$

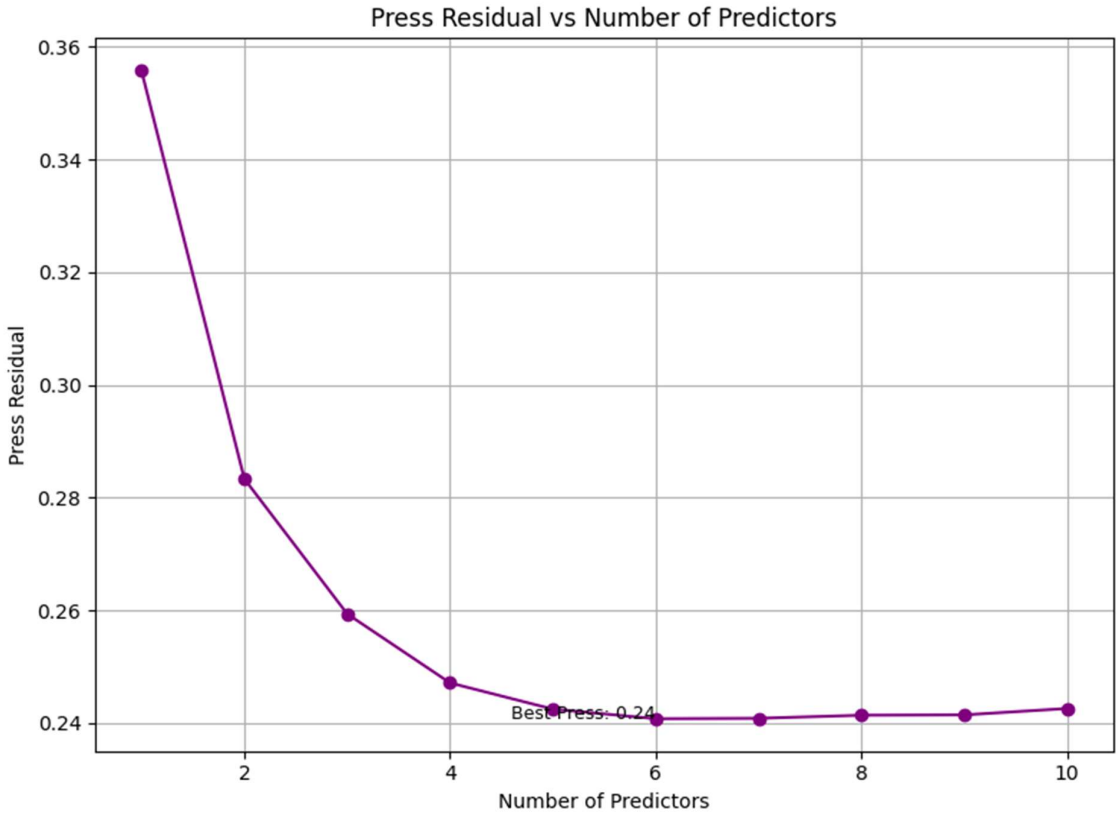
$$PRESS = \sum \left(y_i - \hat{y}_{(i)} \right)^2$$

where:

- yi is the observed value of the response variable for the ith observation.
- y^i(i) is the predicted value of the response variable for the ith observation, obtained from a model fitted to all observations except the ith observation.

In general, we prefer the model with a smaller the PRESS value.

Below is the graph that plots the lowest PRESS value obtained among all the models for a particular number of regressors.



We have tabulated the most preferred criterion value obtained for a particular number of regressor variables in the model (varies from 1 to 10).

	Number of Predictors	AIC	BIC	Adjusted R Squared	Mallow's Statistic	Press Residual
0	1	203.289176	208.708237	0.801642	1.283084	0.355785
1	2	179.002943	187.131534	0.842018	1.892746	0.283339
2	3	170.173976	181.012097	0.855363	1.950699	0.259383
3	4	165.770808	179.318459	0.862183	2.026560	0.247131
4	5	164.617441	180.874623	0.864767	2.108134	0.242476
5	6	164.785296	183.752007	0.865702	2.049427	0.240778
6	7	165.756747	187.432989	0.865649	1.896632	0.240851
7	8	166.936611	191.322383	0.865330	1.923897	0.241399
8	9	167.881882	194.977184	0.865283	1.947962	0.241460
9	10	169.308375	199.113207	0.864637	1.654627	0.242594

We summarize the findings of this section in the below table.

Criterion	Most appropriate number of features (p)	Features selected
Adjusted R squared	6	x2, x3, x4, x9, x10
Mallow's statistic	1	x3
AIC	5	x2, x3, x4, x9, x10
BIC	4	x2, x3, x9, x10
PRESS residuals	6	x2, x3, x4, x7_1, x9, x10

Part B :- Feature Selection Techniques

Forward Selection

1. Start with model having no regressors.
2. All possible models with one regressors are considered and F statistic for each regressor is computed. The regressor having highest F statistic value is added to the model if $F > F_{\alpha,1,\text{res. df.}}$.
3. Partial F statistic are computed for all of the remaining regressors in the presence of previously selected regressors and the one yielding the highest F is added to the model if $F > F_{\alpha,1,\text{res. df.}}$.
4. The process terminates when either the highest F value $< F_{\alpha,1,\text{res. df.}}$ or when the last candidate regressor is added to the model.

Backward Elimination

1. Start with full model.
2. Compute the partial F statistic for each regressor in the presence of other regressors in the model.
3. The regressor with smallest partial F value is removed from the model if $F < F_{\alpha,1,\text{res. df.}}$
4. Partial F statistic are computed for this new model and process repeats until the smallest partial $F > F_{\alpha,1,\text{res. df}}$ or all the variables have been removed.

Stepwise Selection

1. Start with model having no regressors.
2. All possible models with one regressors are considered and F statistic for each regressor is computed. The regressor having highest F statistic value is added to the model if $F > F_{\alpha,1,\text{res. df.}}$.
3. Partial F statistic are computed for all of the remaining regressors in the presence of previously selected regressors and the one yielding the highest F is added to the model if $F > F_{\alpha,1,\text{res. df.}}$.
4. Simultaneously, assess if existing variables become insignificant with a new addition. If so, eliminate that variable.
5. The process terminates when either the highest F value $< F_{\alpha,1,\text{res. df.}}$ or when the last candidate regressor is added to the model.

We now provide the findings of the feature selection techniques in the table below.

Technique	Most appropriate number of features (p)	Features selected
Forward Selection	4	x2, x3, x9, x10
Backward Elimination	4	x2, x3, x9, x10
Stepwise Selection	4	x2, x3, x9, x10

Conclusion: All three of the feature selection techniques resulted in the same subset of regressors.

Chapter 6

Validating assumptions of linear regression

The key assumptions of linear regression are:

1. *Linearity*: There exists a linear relationship between the independent variables (predictors) and the dependent variable (response).
2. *Homoscedasticity*: The variance of the errors (residuals) is constant across all levels of the predictors.
3. *Normality of residuals*: The residuals (errors) should be normally distributed. This means that the distribution of the residuals should follow a normal distribution with mean zero.
4. *Independence*: This means that there is no correlation between the errors (residuals) of different observations.
5. *No Perfect Multicollinearity:* There should be no perfect multicollinearity among the predictor variables. Perfect multicollinearity occurs when one predictor variable is a perfect linear function of other predictor variables, leading to unstable estimates of the coefficients.

We have utilized few methods to validate the above mentioned assumptions.

1. Brown Forsythe test to check for homogeneity of variance among residual populations.
2. Y_pred vs residual plot to illustrate homogeneity of variance. We demand residuals to be scattered and lie in a horizontal band about 0.
3. Quantile-quantile (QQ) plot to illustrate normality of residuals.
4. Variance Inflation Factor (VIF) as a measure of multicollinearity.

   $$\mathbf{VIF}_i = \frac{1}{1-R_i^2}$$

   where, R_i^2 is the coefficient of multiple determination when X_i is regressed over other p-2 input variables.

We now proceed to implement these validation procedures on each of the "best" models obtained by different criteria and feature selection techniques.

Model 1: (Adjusted R^2 criterion)

Y = 6.34 -0.10*x2 + 0.24*x3 + 0.11*x4 - 0.22*x7_1 + 0.90*x9 + 0.30*x10

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.871
Model:	OLS	Adj. R-squared:	0.865
Method:	Least Squares	F-statistic:	141.7
Date:	Thu, 02 May 2024	Prob (F-statistic):	5.07e-45
Time:	00:57:07	Log-Likelihood:	-76.309
No. Observations:	111	AIC:	164.6
Df Residuals:	105	BIC:	180.9
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.3098	0.047	134.178	0.000	6.217	6.403
x2	-0.0958	0.051	-1.893	0.061	-0.196	0.005
x3	0.2401	0.068	3.525	0.001	0.105	0.375
x4	0.1108	0.064	1.739	0.085	-0.016	0.237
x9	0.8788	0.092	9.560	0.000	0.697	1.061
x10	0.2806	0.085	3.307	0.001	0.112	0.449

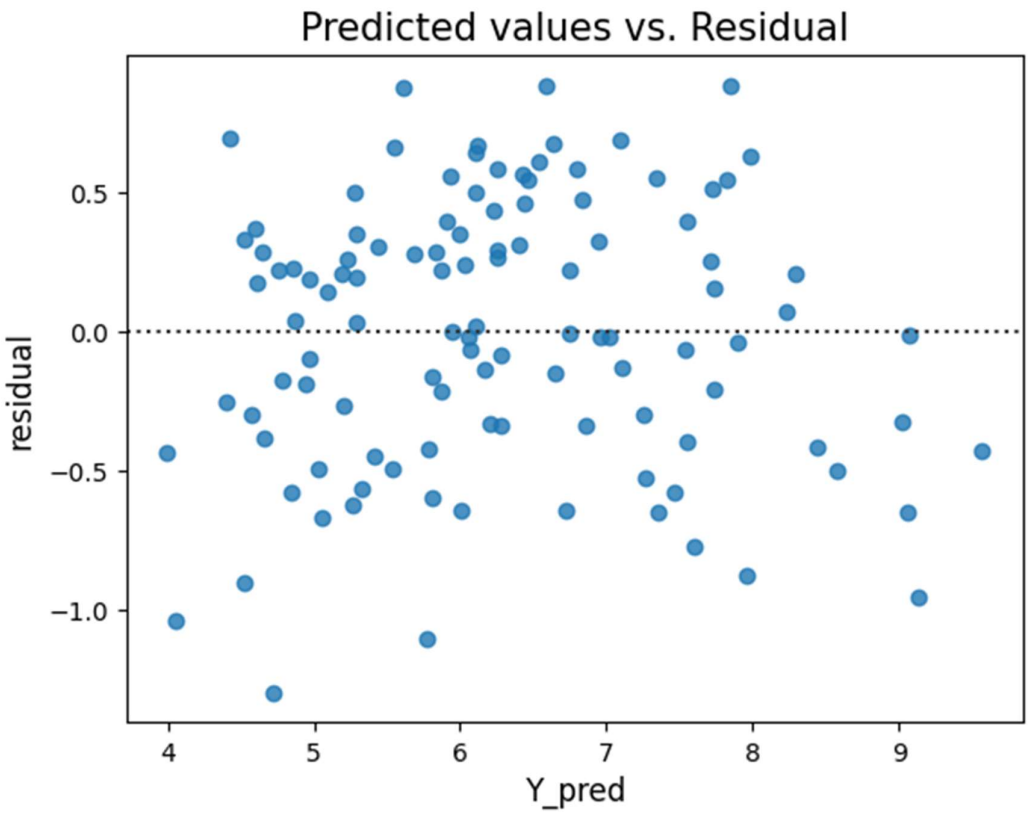
Omnibus:	4.312	Durbin-Watson:	2.070
Prob(Omnibus):	0.116	Jarque-Bera (JB):	3.603
Skew:	-0.340	Prob(JB):	0.165
Kurtosis:	2.436	Cond. No.	3.89

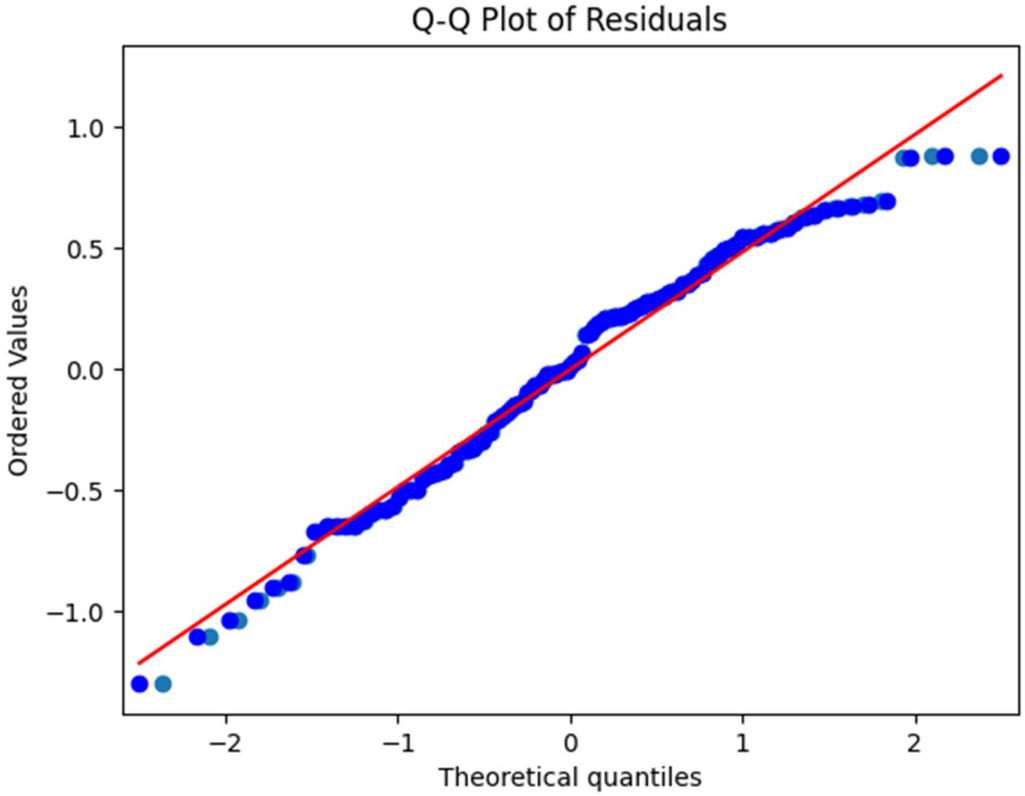
P value of Brown Forsythe test: 0.6874926342333663

Conclusion - Since P value is less than 0.05 we say there is **homoscedasticity.**

	Features	VIF
0	x2	1.129482
1	x3	1.989442
2	x4	1.786237
3	x9	3.482366
4	x10	3.244594

Clearly, we can see that all the VIF values are less than 5 so there is little or no multicollinearity present in the selected subset of features.


Predicted values vs. Residual
residual
0.5
0.0
−0.5
−1.0
4
5
6
7
8
9
Y_pred


Q-Q Plot of Residuals
Ordered Values
1.0
0.5
0.0
−0.5
−1.0
−2
−1
0
1
2
Theoretical quantiles

Model 2: (Mallow's Statistic)

Y = 6.29 + 0.77*x3

OLS Regression Results

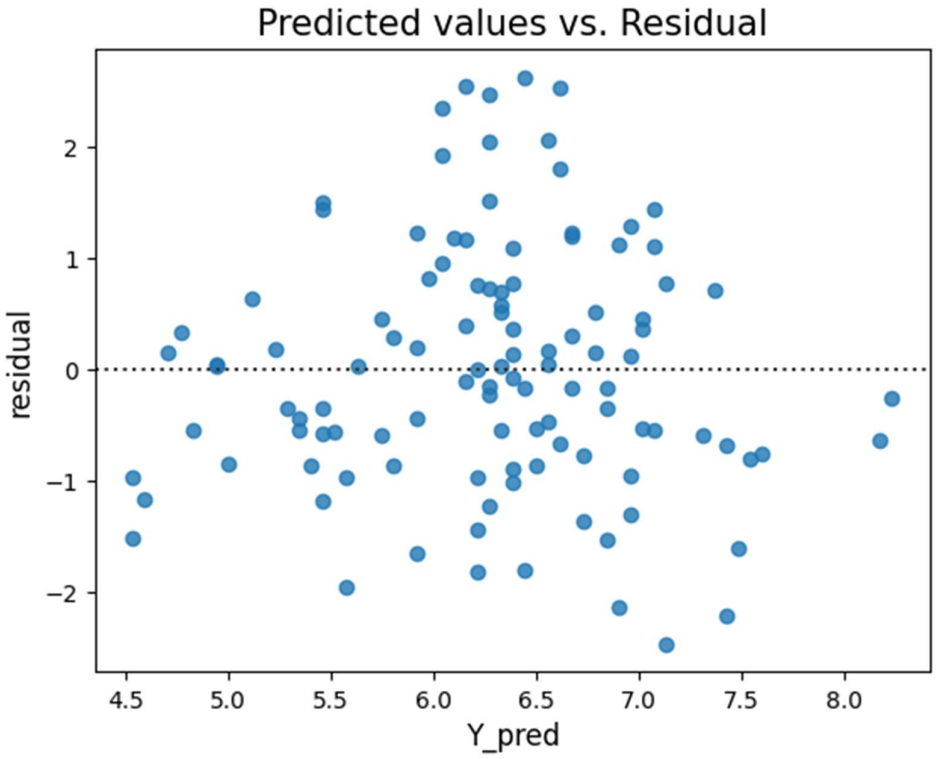
Dep. Variable:	Y	R-squared:	0.313
Model:	OLS	Adj. R-squared:	0.306
Method:	Least Squares	F-statistic:	49.58
Date:	Thu, 02 May 2024	Prob (F-statistic):	1.79e-10
Time:	00:57:10	Log-Likelihood:	-169.12
No. Observations:	111	AIC:	342.2
Df Residuals:	109	BIC:	347.7
Df Model:	1		
Covariance Type:	nonrobust		

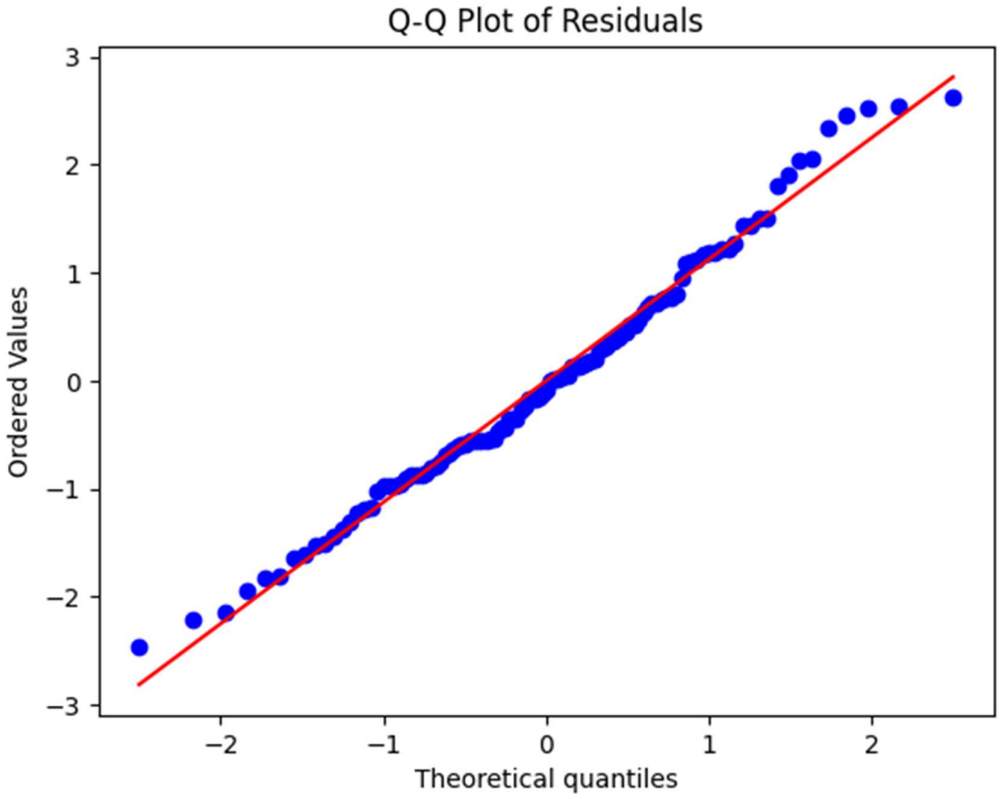
	coef	std err	t	P>|t|	[0.025	0.975]
const	6.2984	0.106	59.185	0.000	6.087	6.509
x3	0.7705	0.109	7.042	0.000	0.554	0.987

Omnibus:	1.828	Durbin-Watson:	2.157
Prob(Omnibus):	0.401	Jarque-Bera (JB):	1.868
Skew:	0.293	Prob(JB):	0.393
Kurtosis:	2.756	Cond. No.	1.05

P value of Brown Forsythe test: 0.41351332477232605

Conclusion – Since P value is less than 0.05 we say ,There is **homoscedasticity.**




Q-Q Plot of Residuals
Ordered Values
Theoretical quantiles
3
2
1
0
−1
−2
−3
−2
−1
0
1
2

Model 3: (AIC criterion)

Y = 6.30 - 0.09*x2 + 0.24*x3 + 0.11*x4 + 0.88*x9 + 0.28*x10

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.871
Model:	OLS	Adj. R-squared:	0.865
Method:	Least Squares	F-statistic:	141.7
Date:	Thu, 02 May 2024	Prob (F-statistic):	5.07e-45
Time:	00:57:07	Log-Likelihood:	-76.309
No. Observations:	111	AIC:	164.6
Df Residuals:	105	BIC:	180.9
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.3098	0.047	134.178	0.000	6.217	6.403
x2	-0.0958	0.051	-1.893	0.061	-0.196	0.005
x3	0.2401	0.068	3.525	0.001	0.105	0.375
x4	0.1108	0.064	1.739	0.085	-0.016	0.237
x9	0.8788	0.092	9.560	0.000	0.697	1.061
x10	0.2806	0.085	3.307	0.001	0.112	0.449

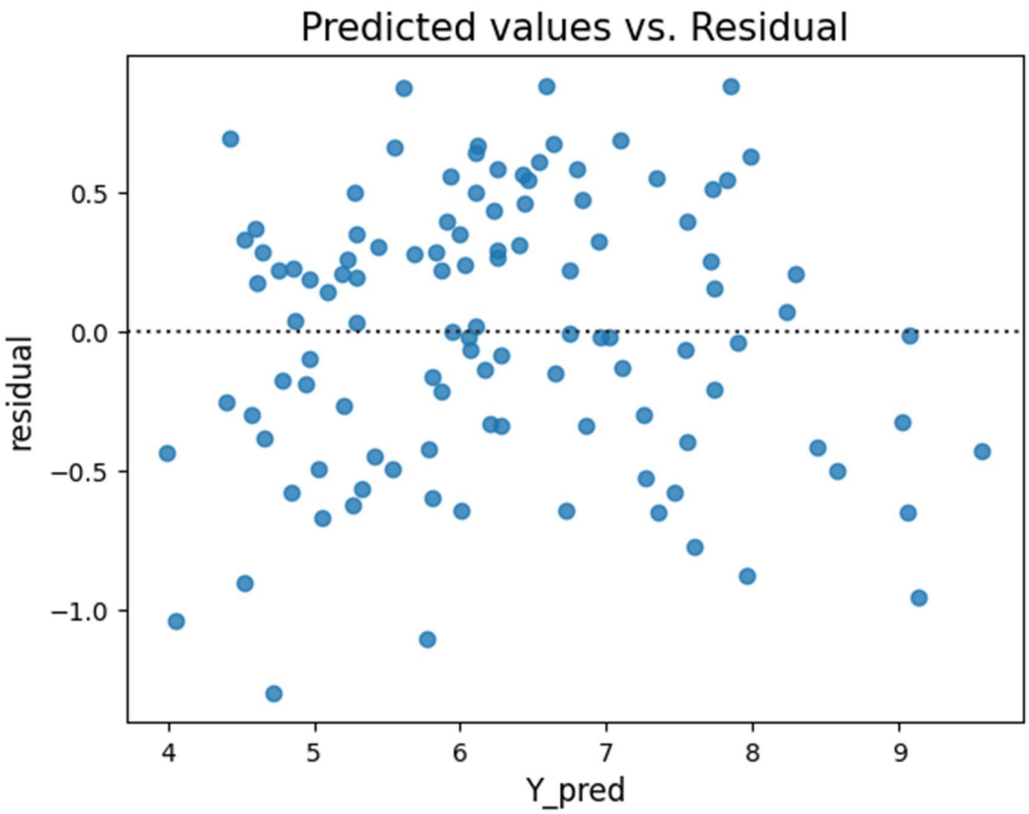
Omnibus:	4.312	Durbin-Watson:	2.070
Prob(Omnibus):	0.116	Jarque-Bera (JB):	3.603
Skew:	-0.340	Prob(JB):	0.165
Kurtosis:	2.436	Cond. No.	3.89

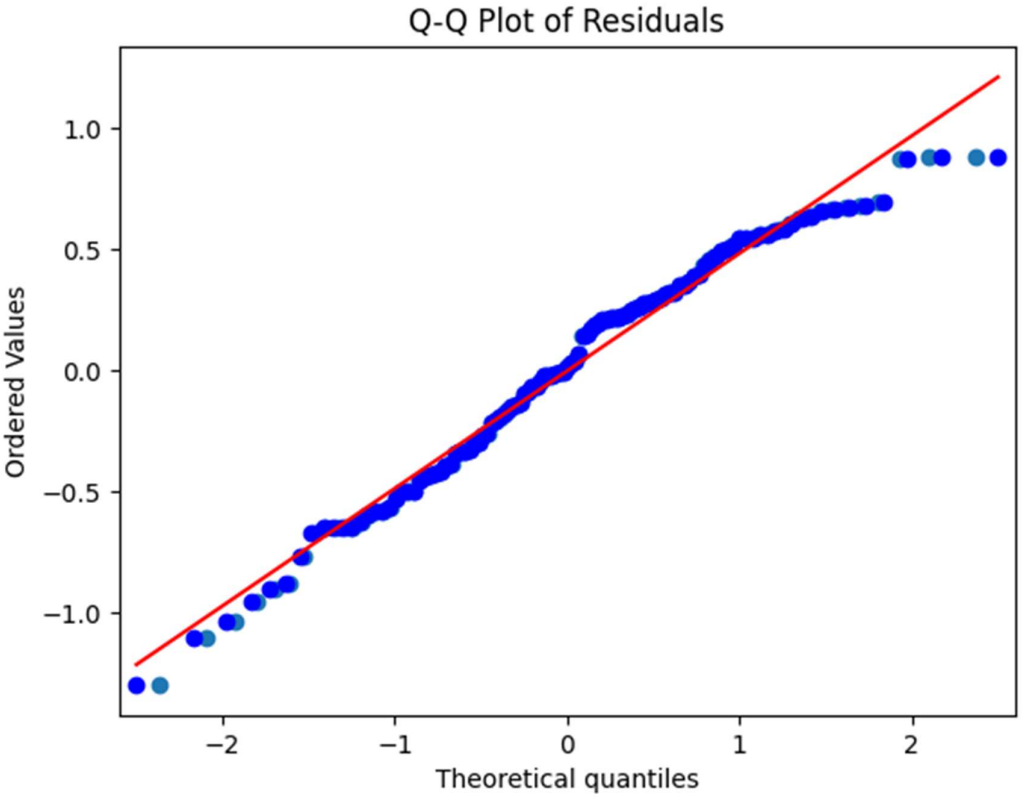
P value of Brown Forsythe test: 0.7220658649641678

Conclusion – Since P value is less than 0.05 we say ,There is **homoscedasticity.**

	Features	VIF
0	x2	1.129482
1	x3	1.989442
2	x4	1.786237
3	x9	3.482366
4	x10	3.244594

Clearly, we can see that all the VIF values are less than 5 so there is little or no multicollinearity present in the selected subset of features**.**


Predicted values vs. Residual
residual
0.5
0.0
−0.5
−1.0
4
5
6
7
8
9
Y_pred


Q-Q Plot of Residuals
Ordered Values
1.0
0.5
0.0
−0.5
−1.0
−2
−1
0
1
2
Theoretical quantiles

Model 4: (BIC criterion)

Y = 6.30 - 0.12*x2 + 0.31*x3+ 0.85*x9 + 0.30*x10

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.867
Model:	OLS	Adj. R-squared:	0.862
Method:	Least Squares	F-statistic:	173.0
Date:	Thu, 02 May 2024	Prob (F-statistic):	1.59e-45
Time:	00:57:09	Log-Likelihood:	-77.885
No. Observations:	111	AIC:	165.8
Df Residuals:	106	BIC:	179.3
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.3089	0.047	132.903	0.000	6.215	6.403
x2	-0.1223	0.049	-2.509	0.014	-0.219	-0.026
x3	0.3134	0.054	5.805	0.000	0.206	0.420
x9	0.8512	0.091	9.313	0.000	0.670	1.032
x10	0.3013	0.085	3.553	0.001	0.133	0.469

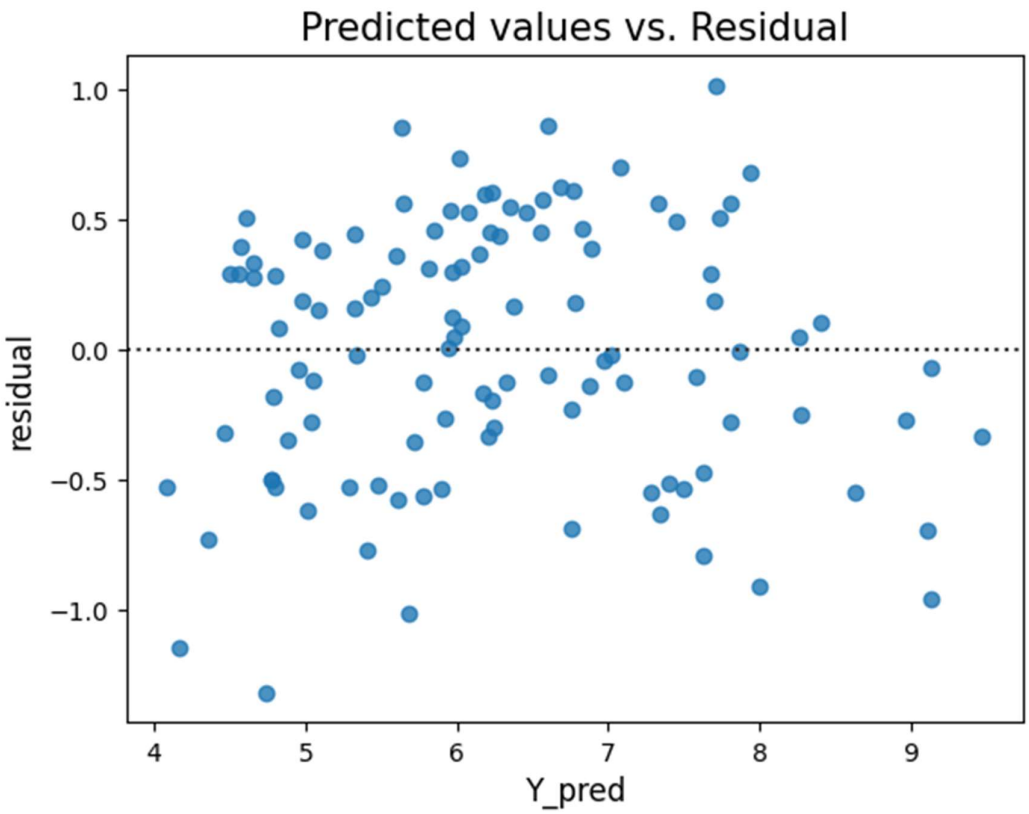
Omnibus:	4.478	Durbin-Watson:	2.013
Prob(Omnibus):	0.107	Jarque-Bera (JB):	3.594
Skew:	-0.328	Prob(JB):	0.166
Kurtosis:	2.411	Cond. No.	3.55

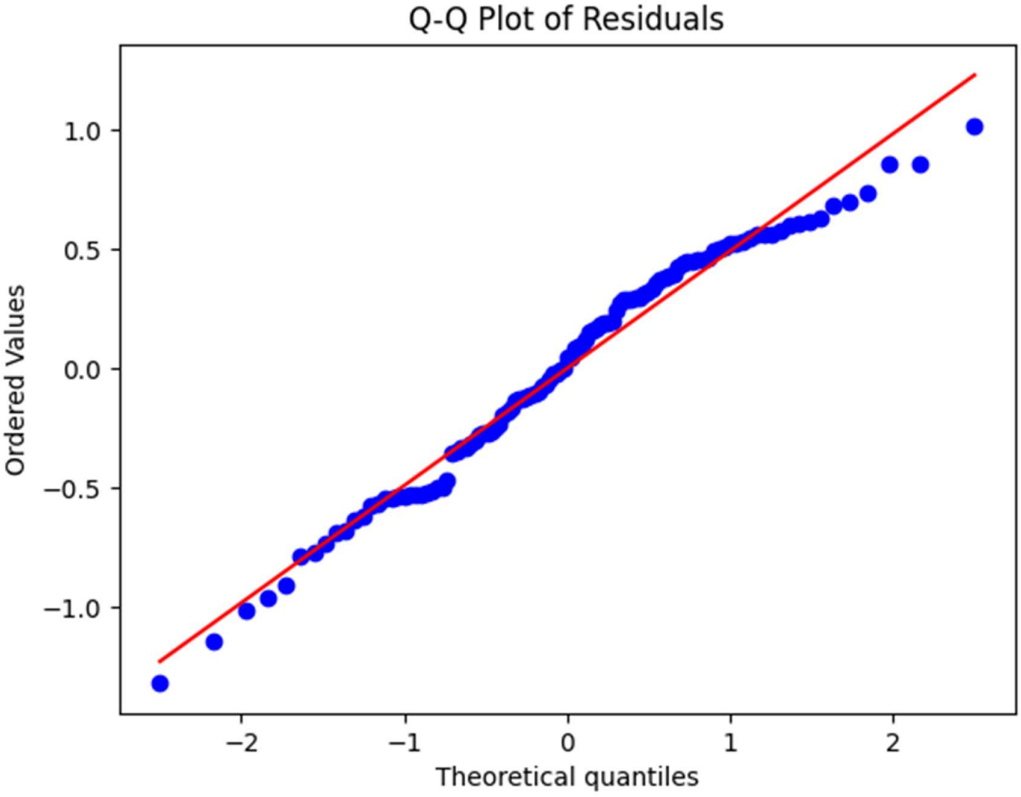
P value of Brown Forsythe test: 0.7033500909962516

Conclusion – Since P value is less than 0.05 we say ,There is **homoscedasticity.**

	Features	VIF
0	x2	1.027332
1	x3	1.226233
2	x9	3.378690
3	x10	3.180744

Clearly, we can see that all the VIF values are less than 5 so there is little or no multicollinearity present in the selected subset of features**.**


Predicted values vs. Residual
residual
Y_pred
1.0
0.5
0.0
−0.5
−1.0
4
5
6
7
8
9


Q-Q Plot of Residuals
Ordered Values
Theoretical quantiles
1.0
0.5
0.0
−0.5
−1.0
−2
−1
0
1
2

Model 5: (PRESS residual)

Y = 6.34 - 0.10*x2 + 0.23*x3 + 0.11*x4 - 0.22*x7_1 + 0.90*x9 + 0.30*x10

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.873
Model:	OLS	Adj. R-squared:	0.866
Method:	Least Squares	F-statistic:	119.2
Date:	Thu, 02 May 2024	Prob (F-statistic):	2.71e-44
Time:	00:58:49	Log-Likelihood:	-75.393
No. Observations:	111	AIC:	164.8
Df Residuals:	104	BIC:	183.8
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.3425	0.053	119.548	0.000	6.237	6.448
x2	-0.1050	0.051	-2.061	0.042	-0.206	-0.004
x3	0.2381	0.068	3.507	0.001	0.103	0.373
x4	0.1135	0.064	1.786	0.077	-0.012	0.239
x7_1	-0.2202	0.167	-1.316	0.191	-0.552	0.112
x9	0.9066	0.094	9.643	0.000	0.720	1.093
x10	0.3013	0.086	3.503	0.001	0.131	0.472

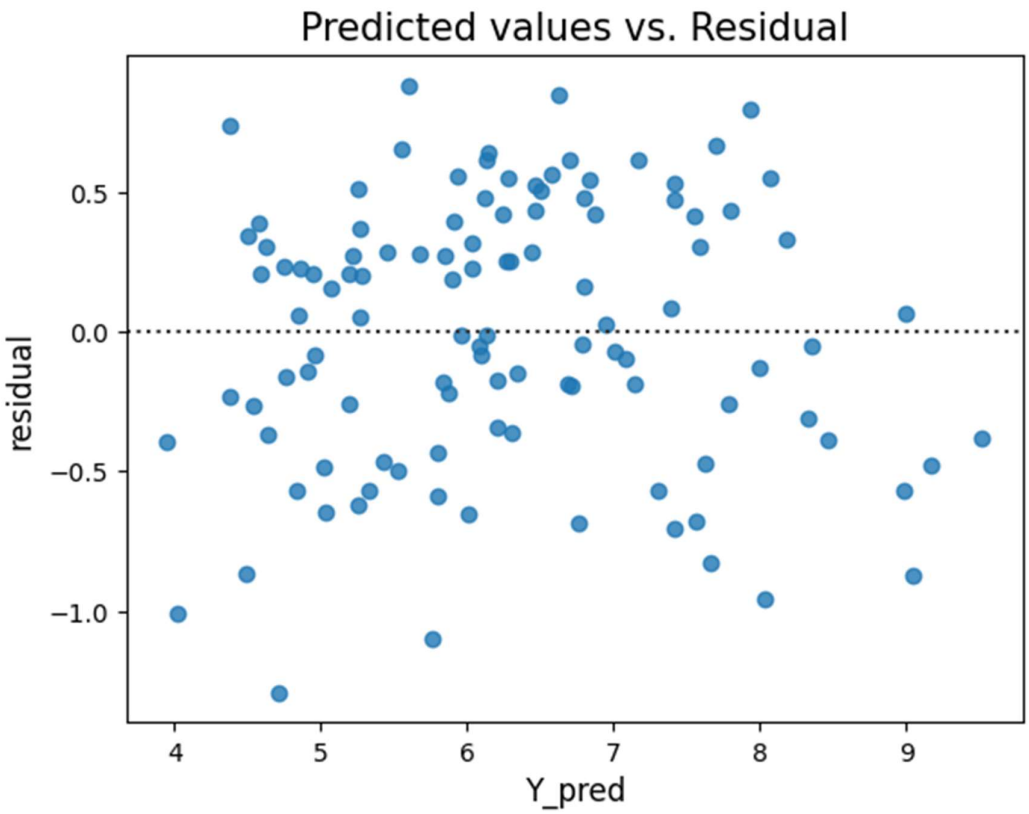
Omnibus:	5.276	Durbin-Watson:	2.067
Prob(Omnibus):	0.071	Jarque-Bera (JB):	4.361
Skew:	-0.385	Prob(JB):	0.113
Kurtosis:	2.408	Cond. No.	5.57

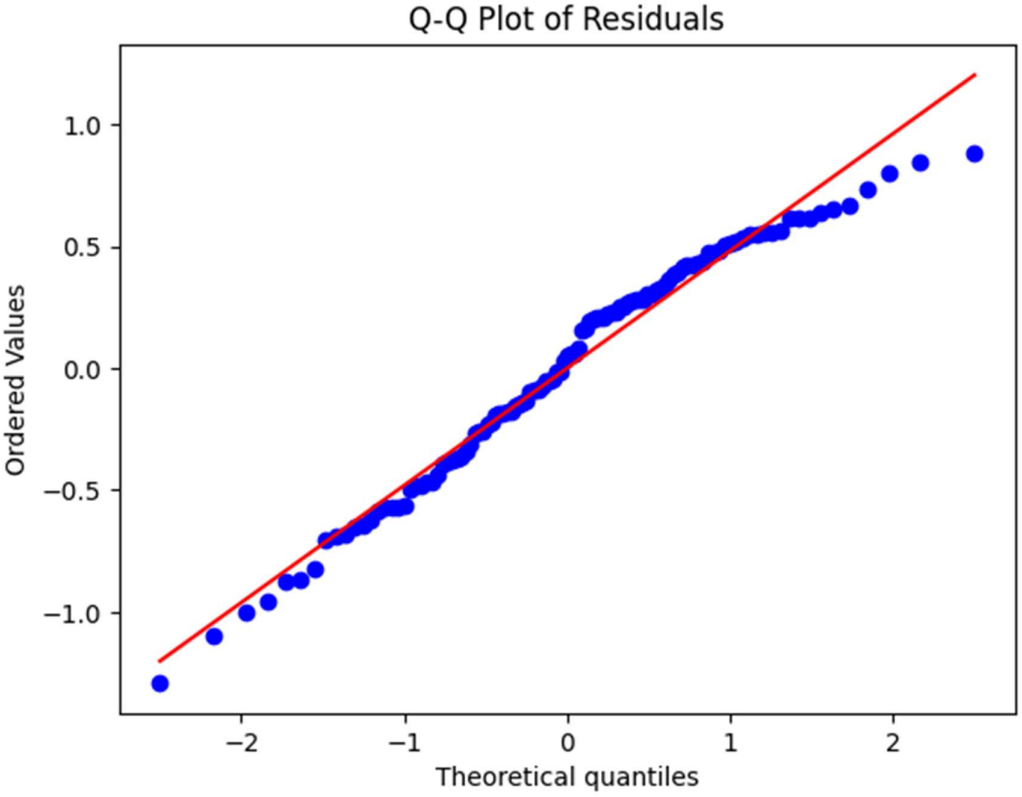
P value of Brown Forsythe test: 0.6874926342333663

Conclusion – Since P value is less than 0.05 we say, there is **homoscedasticity.**

	Features	VIF
0	x2	1.145411
1	x3	1.990881
2	x4	1.787174
3	x7_1	1.439004
4	x9	3.605062
5	x10	3.350544

Clearly, we can see that all the VIF values are less than 5 so there is little or no multicollinearity present in the selected subset of features**.**


Predicted values vs. Residual
residual
0.5
0.0
−0.5
−1.0
4
5
6
7
8
9
Y_pred


Q-Q Plot of Residuals
Ordered Values
1.0
0.5
0.0
−0.5
−1.0
−2
−1
0
1
2
Theoretical quantiles

Model 6: (Forward selection, Backward elimination, Stepwise selection)

Y = 6.30 - 0.12*x2 + 0.31*x3+ 0.85*x9 + 0.30*x10

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.867
Model:	OLS	Adj. R-squared:	0.862
Method:	Least Squares	F-statistic:	173.0
Date:	Thu, 02 May 2024	Prob (F-statistic):	1.59e-45
Time:	00:58:14	Log-Likelihood:	-77.885
No. Observations:	111	AIC:	165.8
Df Residuals:	106	BIC:	179.3
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.3089	0.047	132.903	0.000	6.215	6.403
x2	-0.1223	0.049	-2.509	0.014	-0.219	-0.026
x3	0.3134	0.054	5.805	0.000	0.206	0.420
x9	0.8512	0.091	9.313	0.000	0.670	1.032
x10	0.3013	0.085	3.553	0.001	0.133	0.469

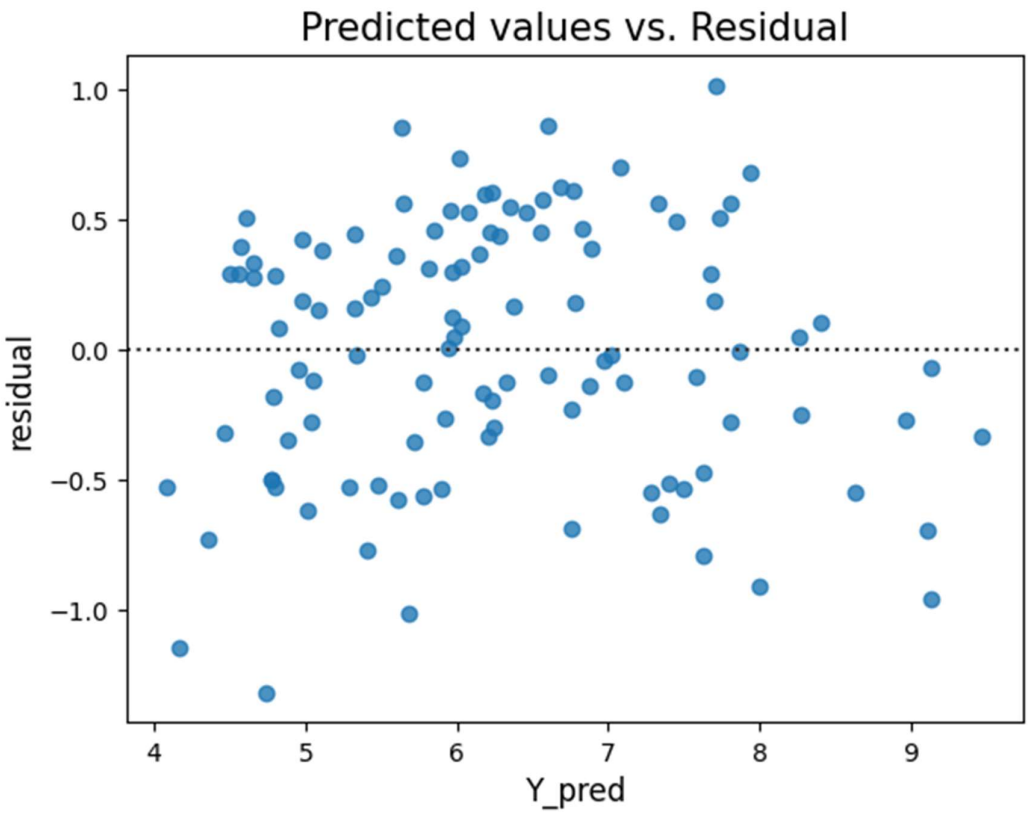
Omnibus:	4.478	Durbin-Watson:	2.013
Prob(Omnibus):	0.107	Jarque-Bera (JB):	3.594
Skew:	-0.328	Prob(JB):	0.166
Kurtosis:	2.411	Cond. No.	3.55

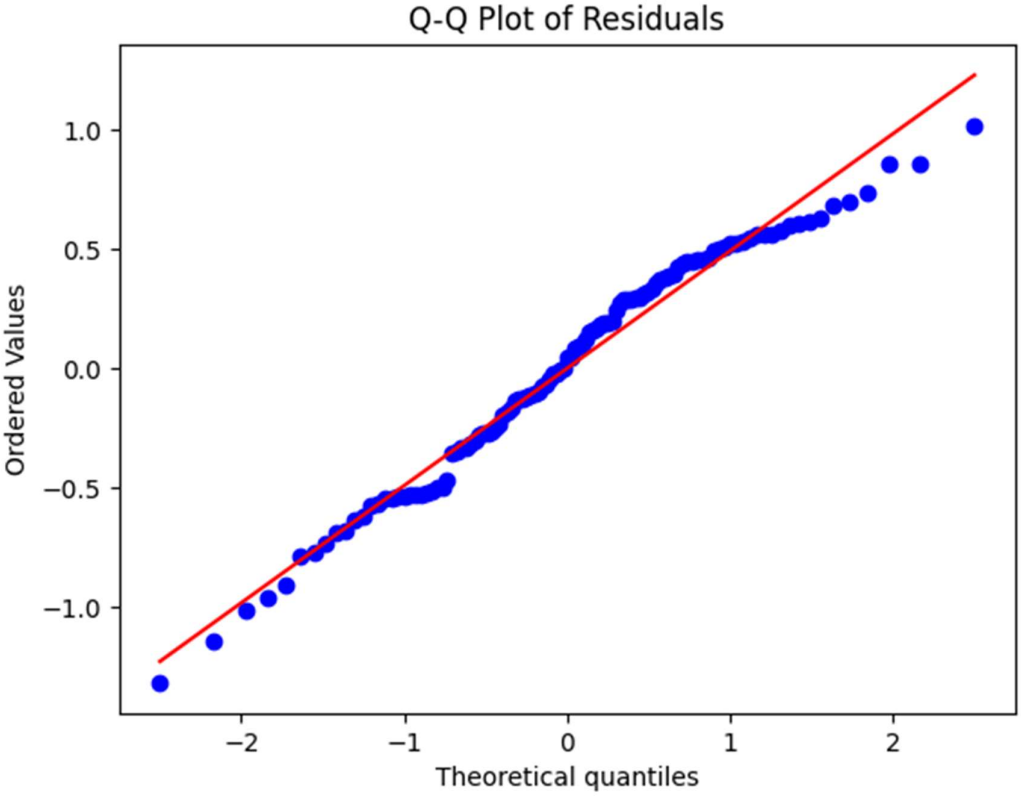
P value of Brown Forsythe test: 0.7033500909962536

Conclusion – Since P value is less than 0.05 we say ,there is **homoscedasticity.**

	Features	VIF
0	x2	1.027332
1	x3	1.226233
2	x9	3.378690
3	x10	3.180744

Clearly, we can see that all the VIF values are less than 5 so there is little or no multicollinearity present in the selected subset of features**.**


Predicted values vs. Residual
residual
Y_pred
1.0
0.5
0.0
−0.5
−1.0
4
5
6
7
8
9


Q-Q Plot of Residuals
Ordered Values
Theoretical quantiles
1.0
0.5
0.0
−0.5
−1.0
−2
−1
0
1
2

Chapter 7

Cross-validation

K-Fold Cross-Validation is a statistical method used to evaluate the performance of machine learning models. Here's a concise description:

1) The entire dataset is divided into 'K' equally sized folds or segments.

2) The model is trained and evaluated 'K' times, each time using a different fold as the validation set.

3) The remaining 'K-1' folds are used for learning.

4) Performance metrics from each fold are averaged to estimate the model's generalization performance.

This method helps to ensure that the model selected for deployment is robust and generalizes well to new data. It is particularly useful when the dataset is limited.

In our project we have taken K to be 5.

A) Model (Adjusted R^2):

Adjusted r square for the 5 models are: [0.82831641, 0.79667261, 0.73431099, 0.77781539, 0.84141914]

Mean Cross-Validation Adjusted R-squared: 0.7957069100761631

Standard Deviation of Cross-Validation Adjusted R-squared: 0.03806175311809686

B) Model (Mallow's Cp):

Mallow's Cp for the 5 models are: [1.66514169, 1.09062788, 1.36285585, 1.42608101, 1.51946205]

Mean Cross-Validation Mallows's Cp: 1.4128336968350261

Standard Deviation of Cross-Validation Mallows's Cp: 0.1905354503644557

C) Model AIC:

AIC for the 5 models are: [133.50264007, 135.28693128, 140.52770216, 120.2914209, 135.09458949]

Mean Cross-Validation AIC: 132.94065678080005

Standard Deviation of Cross-Validation AIC: 6.753172311204462

D) Model BIC:

BIC for the 5 models are: [146.71390792, 147.9154427, 154.07973981, 132.22671652, 148.69822987]

Mean Cross-Validation BIC: 145.9268073658734

Standard Deviation of Cross-Validation BIC: 7.299558576826607

E) Model PRESS:

PRESS for the 5 models are: [5.14013916, 5.07201868, 5.5520314, 8.53423047, 4.96200642]

Mean Cross-Validation PRESS: 5.852085226111191

Standard Deviation of Cross-Validation PRESS: 1.3558426694423873

Conclusion:

In conclusion, our project embarked on a comprehensive journey through data analysis and model building. We began with an exploratory data analysis (EDA), which allowed us to understand the structure, patterns, and relationships within our data. We also checked for influential observations based on Cook's distance and DFFITS criteria.

We then fitted simple linear regression models as a baseline for our analysis. To improve upon this, we employed various model selection techniques to find the best subset models. These techniques included the use of criteria such as Adjusted R-squared, Mallows' Cp, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Predicted Residual Sum of Squares (PRESS).

To further refine our models, we identified the most significant features using backward elimination, forward selection, and stepwise selection methods.

We also ensured that our models met all the necessary assumptions of linear regression. This step was crucial in guaranteeing the validity and reliability of our models.

Finally, we validated our models using the K-fold cross-validation technique. This method provided us with an understanding of how our models would perform on unseen data.