Regression Analysis to model the demand of Natural Gas

SUBMISSION

Group Name:

1. Shwetank Pandey- 2235191
2. Mridul Sharma- 2235157

Abstract

Natural gas is one of the primary fuels for various industries, along with diesel, furnace oil and coal. But the demand for natural gas by industrial consumers is fluctuating due to various external factors like the price of the natural gas/diesel/furnace oil, or due to seasonality variations, or due to reduced demand of downstream products.

The objective of this report is to develop a regression model to understand the importance of above mentioned factors in determining the demand for natural gas.

Problem Objective

To determine if there is any statistically significant linear relationship between the Overall industry sales of natural gas and the other variables namely:

- Price of Diesel oil
- Price of Furnace oil
- Price of Natural gas
- Exchange rate
- Problem month
- Competitor's price
- Construction Indices

Model Assumptions

The regression model that is being modelled to capture the sales of natural gas should satisfy the following assumptions:

1. The sales of natural gas is linearly related to the other variables mentioned in the problem statement
2. The mean of residuals is zero
3. Homoscedasticity of residuals or constant variance
4. No autocorrelation of residuals
5. predictor variables and residuals are uncorrelated
6. Residual distribution is normal

Problem Solving Methodology- CRISP DM Framework

Business Objective : Explained Above Slide

Data : Present in Excel Format, Information about all companies and details about Investment in them and mapping of primary sectors to their main sectors

Data Preparation : Cleaning of the data available by replacing the NA values with meaningful values wherever necessary.

Modelling: Applying a Model following the constraints above and sorting the results step by step coming to a conclusion for completing the objective.

Evaluation: Evaluating the results in different tools, reviewing the process and summarizing the results keeping the business success constraints in mind(Above Slide)

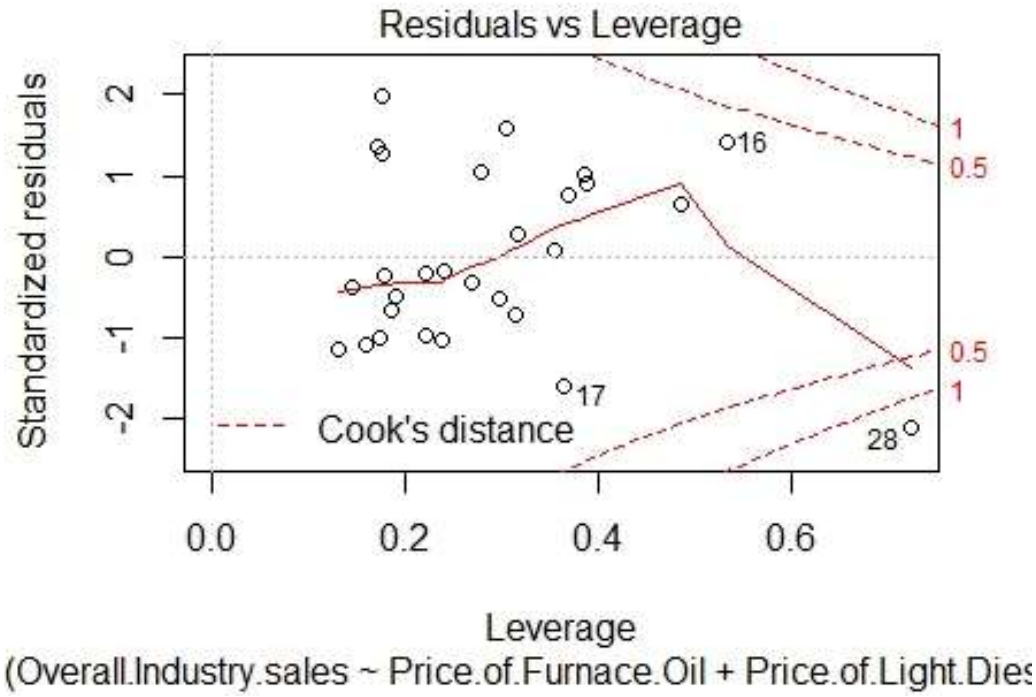
Tools Used : R, Excel

Outlier Removal and Converting Discrete Variables into Categorical

Strategy And Steps Followed

1. Making the box plot for every continuous variable and checking for the outliers.
2. Imputing the outliers using the Interquartile function.
3. Converting discrete variables into categorical variables using as.factor command in R.

Outlier Graph



Regression Modelling Approach

1. Run a linear regression with all the variables and perform ANOVA to test whether the F Statistic is significant , if yes then we proceed with the model
2. Check for collinearity among variables, we look for correlation among variables, assuming any pair having correlation over 0.9 is considered highly correlated, we either drop one of the variables from the model or try introducing an interaction term.
3. Perform stepwise regression and drop variables that are insignificant and with highest p value,
4. We finally arrive at the set of variables that give the max R square adjusted values
5. We test the model assumptions and if everything seems satisfactory then we accept the model as our competing model.
6. We attempt different models, excluding outliers, including interaction terms etc. and choose the model with the best performance parameters as our final recommendation.

Model validation will test the above listed assumptions of the competing models and choose the model with the best performance parameters as our final model.

The demand for natural gas is determined by a number of factors. Based on the data available from the past 28 months, a suitable regression model has been developed taking into consideration all the influential factors mentioned earlier. The collinear variables and outlier data have been identified, and adjusted to develop the regression model.

Price of Diesel: For every 1 Rupee increase in the price of diesel, the sale of natural gas increases by 3.98 million metric standard cubic meters (MMSCM).

Price of Furnace Oil: For every 1 Rupee increase in the price of furnace oil, the sale of natural gas decreases by 3.8 million metric standard cubic meters (MMSCM).

Price of Natural Gas: For every 1 Rupee increase in the price of natural gas, the sale of natural gas increases by 16.987 MMSCM.

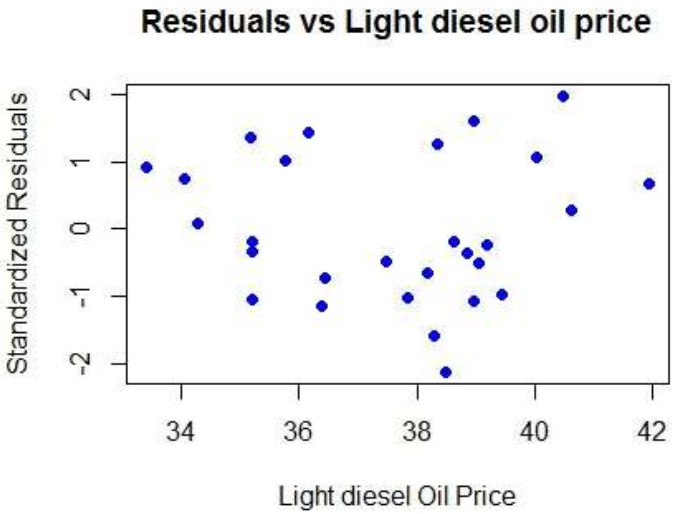
Exchange Rate: For every 1 Rupee increase in the exchange rate of Rupee against USD, the sale of natural gas increases by 11.8 MMSCM.

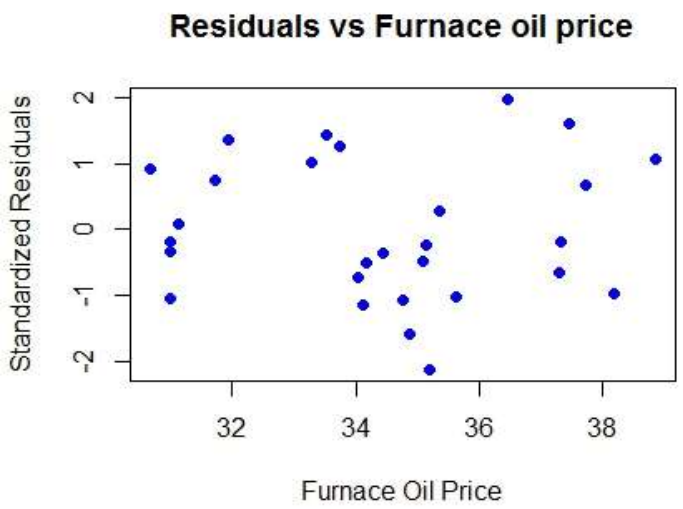
Exogenous factors: During the "problem month" where the consumption of natural gas is lowered due to exogenous factors, the sale of natural gas decreases by 8.09 MMSCM.

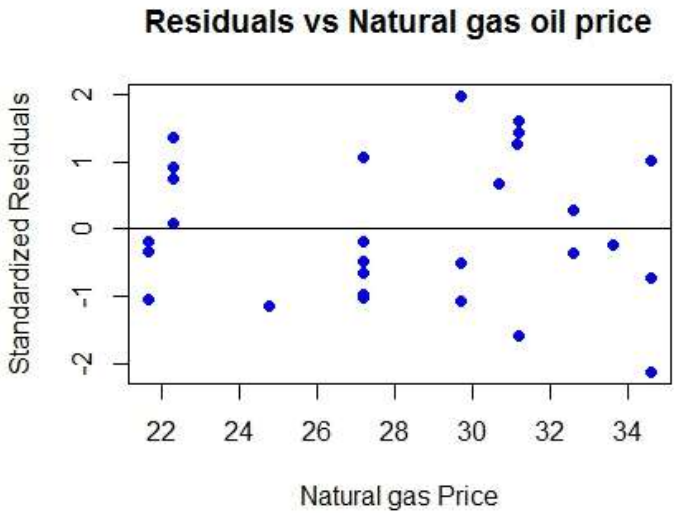
Construction Index: For every 1 unit increase in the construction index, there is a decrease in the sale of natural gas by 1.47 MMSCM.

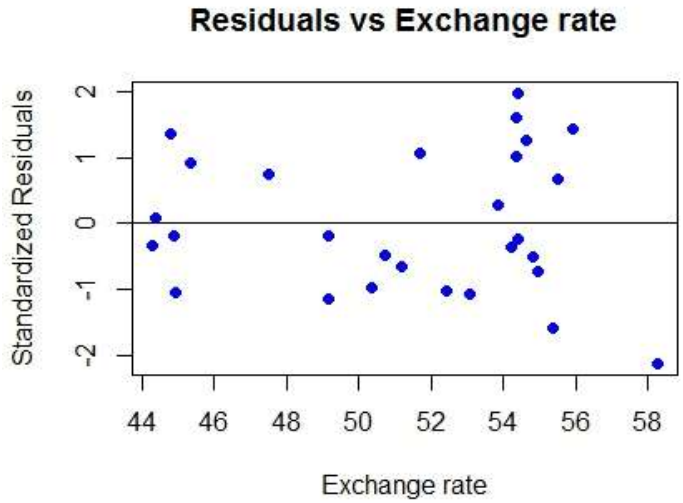
Model Evaluation

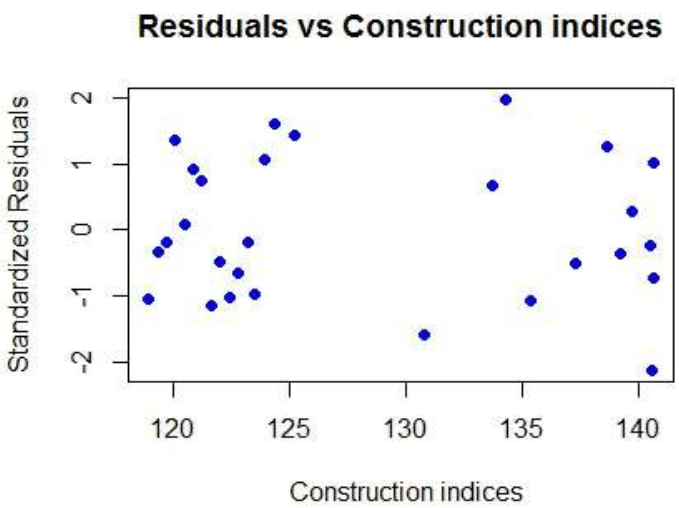
- The recommended regression **model is linear in parameters**, and that the individual predictor variables have a linear relationship with the response variable. This can be confirmed by looking at the graphs of "standardized residual errors" against the individual predictor variables.

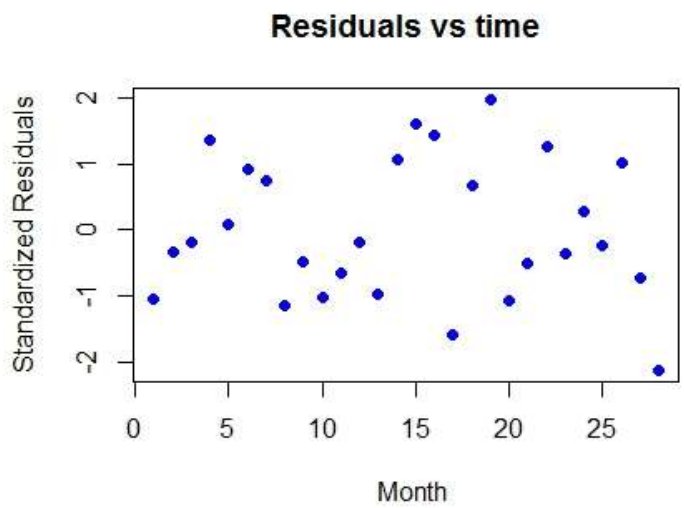




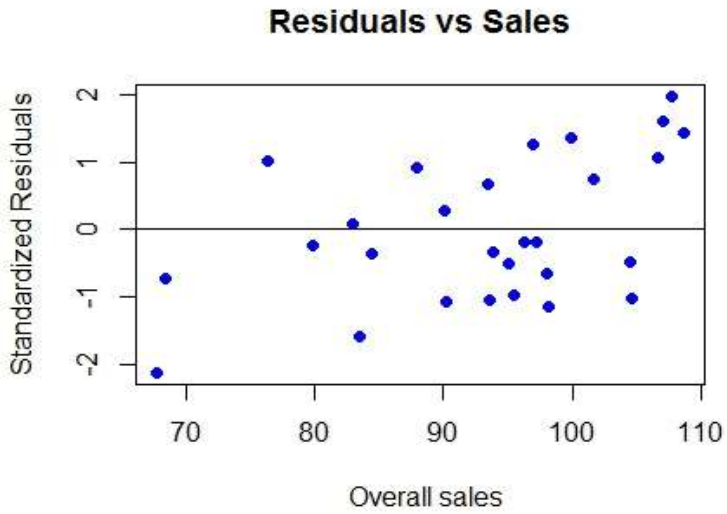

Residuals vs Natural gas oil price
Standardized Residuals
Natural gas Price
22
24
26
28
30
32
34
-2
-1
0
1
2

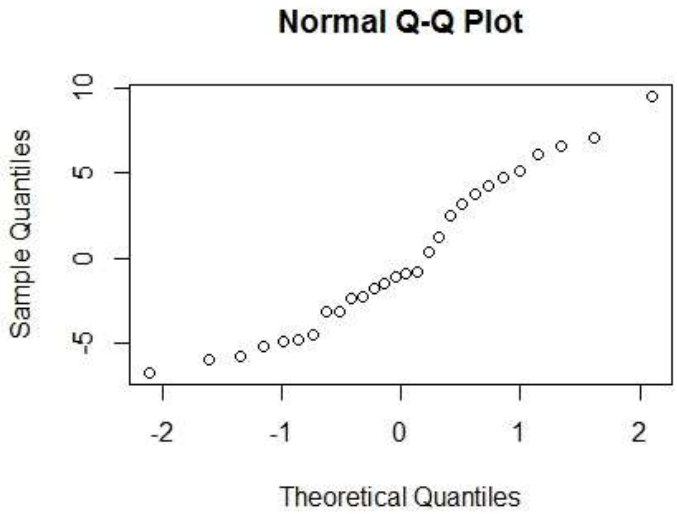

Residuals vs Exchange rate
Standardized Residuals
Exchange rate
44
46
48
50
52
54
56
58
-2
-1
0
1
2


Residuals vs Construction indices
Standardized Residuals
Construction indices
120
125
130
135
140
-2
-1
0
1
2


Residuals vs time
Standardized Residuals
Month
0
5
10
15
20
25
-2
-1
0
1
2

- The mean of the residuals is approximately zero. (Mean of residuals = 4.7e-17, i.e. ~=0).
- All **observations are random** and equally reliable. This can be confirmed by plotting the standardized residual errors against the fitted values.





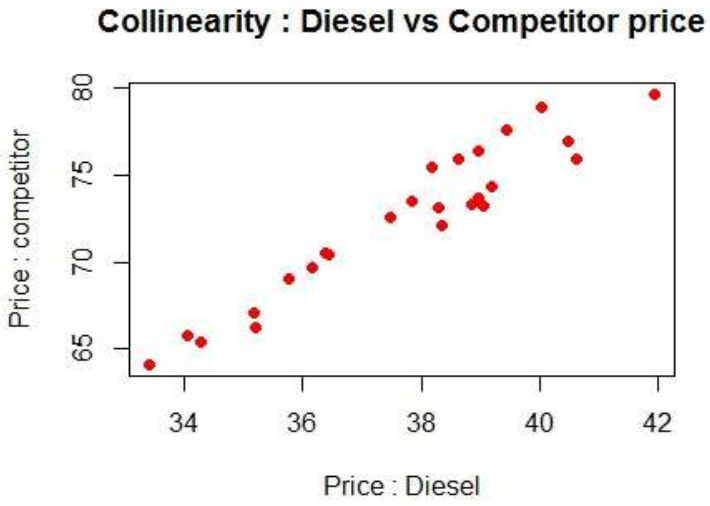
The plot shows that the residuals are randomly distributed with constant variance around mean, indicating homoscedasticity.
The normal QQ plot shows that all the points are falling close to the diagonal line indicating the **residuals follow a normal distribution**.

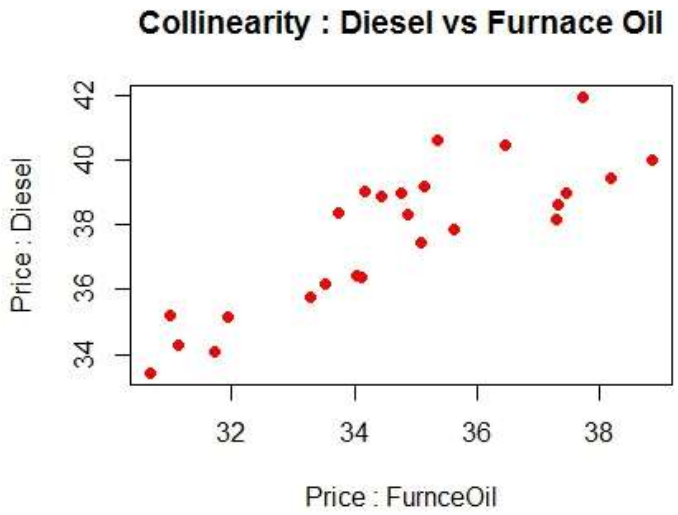
- The predictor **variables are linearly independent** of each other. This can be confirmed from the correlation matrix.

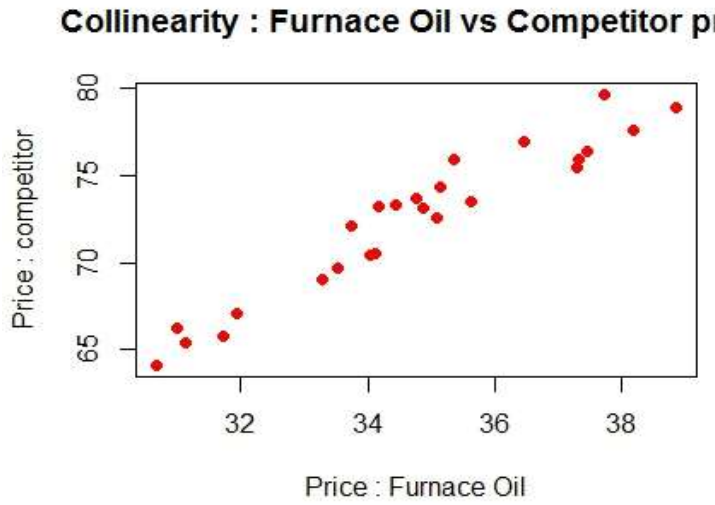
			Corelation coefficient Matrix				
	Price of Furnace Oil	Price of Diesel	Price of Natural gas	Exchange rate	Problem Month	Construction Indices	Competitor price
Price of Furnace Oil	X	0.86	0.51	0.6	-0.15	0.24	0.97
Price of Diesel	0.86	X	0.62	0.71	-0.1	0.52	0.96
Price of Natural gas	0.51	0.62	X	0.94	-0.13	0.87	0.59
Exchange rate	0.6	0.71	0.94	X	-0.1	0.77	0.68
Problem Month	-0.15	-0.1	-0.13	-0.1	X	-0.09	-0.13
Construction Indices	0.24	0.52	0.87	0.77	-0.09	X	0.34
Competitor price	0.97	0.96	0.59	0.68	-0.13	0.34	X

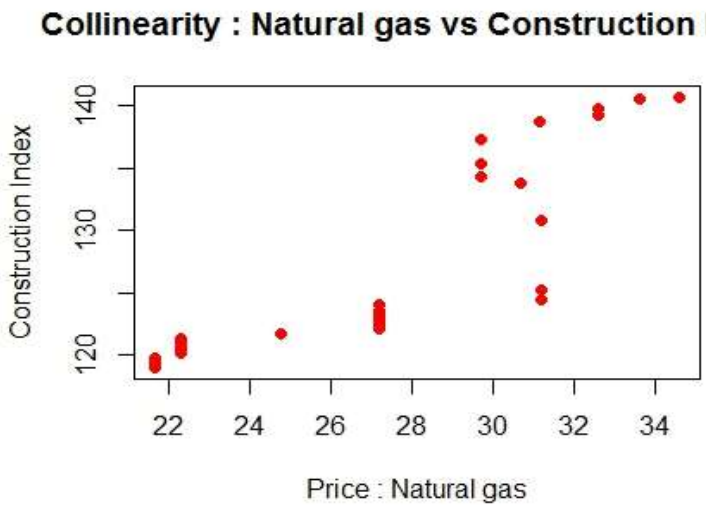
The matrix shows that the variables included in the regression model are not correlated to each other.

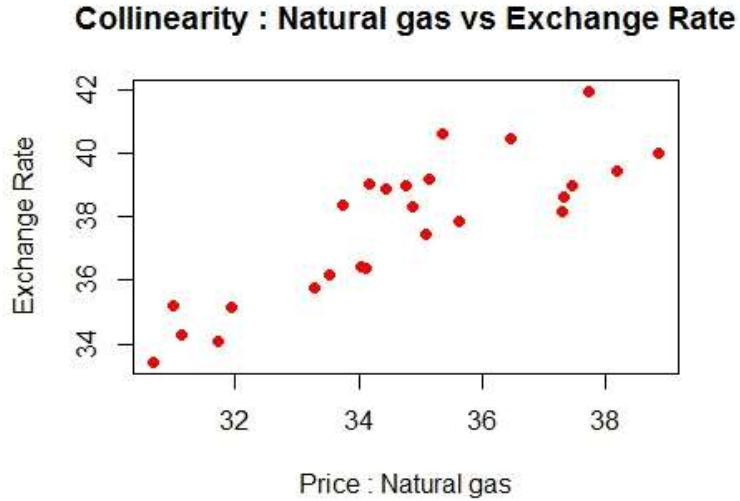
Collinearity Scatterplots



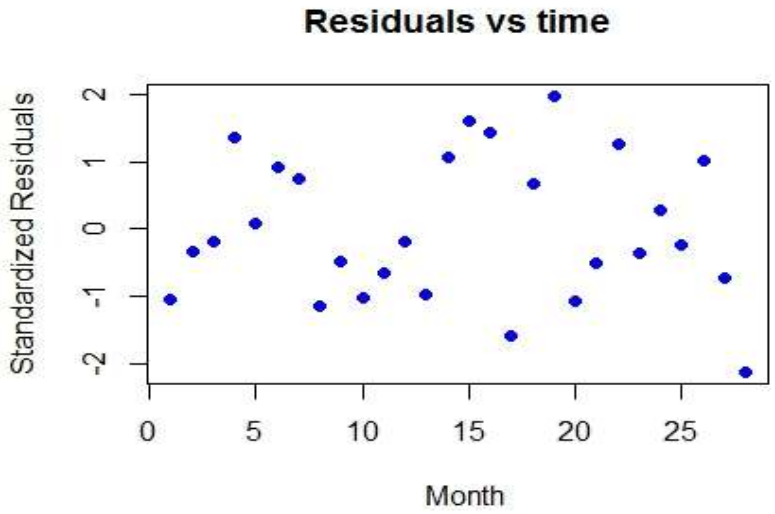



Collinearity : Furnace Oil vs Competitor price
Price : competitor
65
70
75
80
32
34
36
38
Price : Furnace Oil


Collinearity : Natural gas vs Construction Inde
Construction Index
120
130
140
22
24
26
28
30
32
34
Price : Natural gas

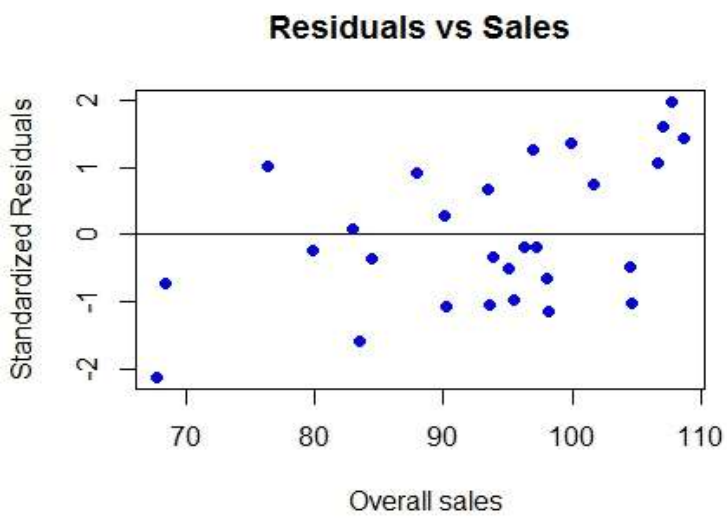

Collinearity : Natural gas vs Exchange Rate
Exchange Rate
34
36
38
40
42
32
34
36
38
Price : Natural gas

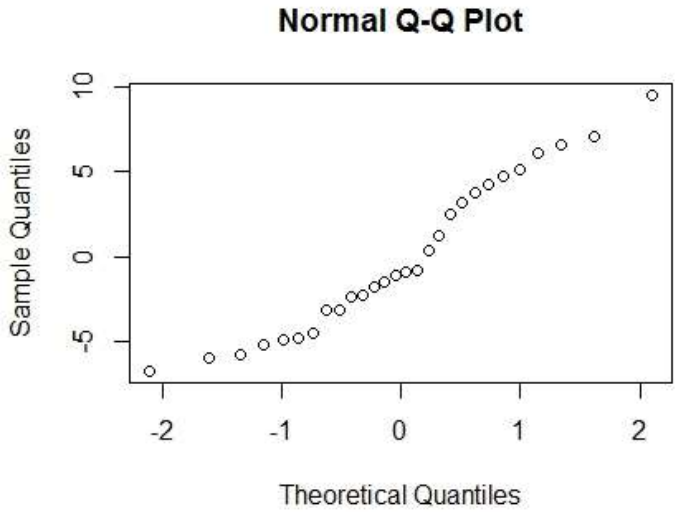
- The **observations are independent** of each other, and there is no autocorrelation issue. This can be confirmed by plotting the standardized residual errors against index



The observations are randomly scattered around mean and hence there is no autocorrelation.

- All the errors for the fitted model have the **same variance** and this can be confirmed by plotting the standardized residual errors against fitted values.
- The plot shows that the observations are randomly scattered around mean.





- No **autocorrelation** of residuals

The ACF plot of the residuals falls between the thresholds, indicating that the residuals are not correlated

- The plot of standardized residuals against **leverage points** show that there is an influential observation for the 28th month

Conclusion

We have done a comprehensive study on the demand/sales of natural gas with respect to various external factors and have presented a linear regression model to understand their relationship. We have provided the model assumptions and the steps that were followed in developing this model. We developed and examined various models and chose the model with the best performance parameters as our final model. The recommended model explains 83% of all variations in the observed data, and is a good fit for prediction of the sales of natural gas.

Accuracy Specificity and Sensitivity