Regression Analysis to model the demand of Natural Gas

**1. Introduction**

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 statement**

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

1. Price of Diesel oil
2. Price of Furnace oil
3. Price of Natural gas
4. Exchange rate
5. Problem month
6. Competitor’s price
7. 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

**Regression Modelling Steps:**

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.

**2. Regression Modelling**

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.

The best model that could be developed to understand the demand of natural gas, based on the historical data is:

Sales of Natural Gas = -307.441

+ 3.982 \* (Price of Diesel)

-3.80\*(Price of Furnace Oil)

+17.3431 \* (Price of natural gas)

+ 11.8 \*(Exchange rate) –

-8.09\*(problem Month)

- 1.47\* (Construction Index)

-0.356\*(Price of Natural Gas\*Exchange rate)

**The model explains 83% of all variations in the observed data.**

Interpretation

The demand/sales of natural gas is determined by the following 5 factors:

|  |  |  |
| --- | --- | --- |
| 1 | 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). |
| 2 | 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). |
| 3 | 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. |
| 4 | 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. |
| 5 | 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. |
| 6 | 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. |

**3. Regression Approach**

3.1. As a first step to develop the **full regression model**, the entire set of predictor variables are analyzed and a linear model is generated. Based on the full model, the demand for natural gas could be expressed in the following equation:

Sales = 131.5219 - 531.6\*(price of Furnace Oil) - 526.1\*(Price of Diesel)

- 1.5\*(price of natural gas) + 529.5\*(Price of Competitor)

+ 2.6\*(Exchange Rate) - 9.5\*(Problem Month) – 1.5(Construction Index)

* In the above model, all available predictor variables are taken into consideration.

However, some of the variables are insignificant in this model.

* The price of furnace oil, the price of diesel, and the price of natural gas have no significant impact on the model.
* The above model explains 76% of all variations in the observed data

3.2 The next step is to identify the **collinear relationship** between variables in the model. We have calculated the pair-wise co-relation coefficient of variables and also plotted the scatter-plot to determine the relationship. It is assumed that correlation above 0.9 to be significantly high. It has been observed that:

* The price of furnace oil and price of diesel seem to be highly related to each other (correlation coefficient: 0.96), and also highly related to the competitor price.
* Price of Natural gas and exchange rate show a correlation of .96
* Please refer to Annexure: Correlation coefficient Matrix.

Based on the collinearity relationship, some of the variables might be redundant in the final regression model.

3.3. As we have observed that some of the variables are insignificant in the full model and that some of the variables have collinear relationship, we attempted to develop other models with reduced variables using the **“Backward elimination procedure”.**

3.4. **Reduced Model**

The competitor price is the sum of price of furnace oil and the price of diesel and hence it can be considered as a derived variable. And moreover, the variable has insignificant impact in the full model discussed above. Hence competitor price can be removed from our analysis to find a better model for the demand of natural gas

When the variable “Competitor price” is not included in the model, the regression model is in the form of:

Sales = 120.5 - 1.9\*(Price of Furnace Oil) + 3.6\*(Price of Diesel)

- 1.7\*(Price of Natural Gas) + 2.5\*(Exchange Rate)

- 9.7\*(Problem Month) – 1.4\*(Construction Index)

* The model explains 74% of all variations.
* The variables “price of furnace oil” is insignificant in this model.

3.5 Since some of the variables are not significant in the previously developed model, we attempt to still reduce the regression model by exploring other correlated factors. Considering that the “price of natural gas” and “exchange rate” have a high correlation and both of them show high significance to the model, an **interaction term** for these variables is added to the model. The regression model is of the form:

Sales of Natural Gas = -307.441 + 3.982 \* (Price of Diesel) -3.80\*(Price of Furnace Oil)

+17.3431 \* (Price of natural gas) + 11.8 \*(Exchange rate) –

-8.09\*(problem Month) - 1.47\* (Construction Index)

-0.356\*(Price of Natural Gas\*Exchange rate)

* ThThe model explains 83.05% of all variations.
* All the variables are significant in this model( P-value <.05, assuming alpha=.05)

3.6. Based on the results above, it can be inferred that the model developed in [3.5] is better compared to other models [3.1 & 3.4], and it explains 83.05% of all variations in the data

1. **Goodness of fit**

For the selected model (described in 3.5), we perform the regression diagnostics to determine the goodness of fit. The coefficient of determination of the model(R square) is 0.8305, which implies that the model accounts for **83.05%** **of all variations** in the data. The adjusted R-square value is also high at 0.7712.

Model Validations

1. 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.

*Appendix: Linearity sheet: contains the plot of std. Residual errors against predictor variables*

All the plots show a linear relationship.

1. The mean of the residuals is approximately zero.

(Mean of residuals = 4.7e-17, i.e. ~=0).

1. All **observations are random** and equally reliable. This can be confirmed by plotting the standardized residual errors against the fitted values.

*Appendix: Random sample sheet:*

*Contains the plot of std. Residual errors against 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**.

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

*Appendix: collinearity matrix, collinearity scatter plot*

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

1. 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

*Appendix: Independent observations sheet:*

*Contains the plot of std. Residual errors against index.*

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

1. 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.

*Appendix: Random sample sheet:*

*Contains the plot of std. Residual errors against fitted values.*

The plot shows that the observations are randomly scattered around mean.

1. **No autocorrelation** of residuals

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

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

**5. Comparison with other models**

The best model that we have developed with the observed data is given below:

**M1: Best model**

Sales of Natural Gas = -307.441 + 3.982 \* (Price of Diesel) -3.80\*(Price of Furnace Oil)

+17.3431 \* (Price of natural gas) + 11.8 \*(Exchange rate) –

-8.09\*(problem Month) - 1.47\* (Construction Index)

-0.356\*(Price of Natural Gas\*Exchange rate)

* R square : 0.8305
* Adjusted R square : 0.7712

Apart from the chosen model M1, we have analyzed various other models and present here two of the best competing models. In all of the models we have analyzed, it is observed that the following variables are always significant:

* price of Diesel
* exchange rate
* problem month
* construction index

**M2: Competing Model 1**

Sales = 57.08 + 1.20\*(Price of Diesel) – 3.16\*(Price of Natural Gas)

+ 3.59\*(Exchange Rate) – 10.10\*(Problem Month) – 0.79\*(Construction Index

* R square : 0.8221
* Adjusted R square : 0.7797
* The “price of natural gas” is included, while the “price of furnace oil” is not included in the model. The “price of furnace oil” and “price of diesel” are highly collinear and it is redundant to include “price of furnace oil” in the model.
* We believe that the “price of natural gas” is a significant factor in determining the sales of natural gas, and hence should be part of the model
* This model treats the last observation (of month 28) as an outlier and the regression model is developed based on the remaining 27 observations
* The R-square and adjusted R-square are lesser than models M1.

**M3: Competing Model 2**

Sales = 173.8 – 2.98\*(Price of Furnace Oil) +4.3\*(Price of Diesel)

+ 1.99\*(Exchange Rate) – 10.05\*(Problem Month) – 1.86\*(Construction Index)

* + - R square : 0.781
    - Adjusted R square : 0.7288

In this competing model,

* + - The “price of natural gas” is not included and does not have an impact. However, the “price of furnace oil” is included in the model.
    - The “price of furnace oil” and “price of diesel” are highly collinear and it is redundant to include “price of furnace oil” in the model.
    - We believe that the “price of natural gas” is a significant factor in determining the sales of natural gas, and hence should be part of the model
    - This model treats the last observation (of month 28) as an outlier and the regression model is developed based on the remaining 27 observations
    - The R-square and adjusted R-square are lesser than models M1 and M2.
    - Based on the above considerations, we feel that the model M3 is not better than M1 or M2.

**6. 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.

**6. Appendix**

* 1. Appendix\_Natural\_Gas\_Sales\_Regression.xlsx

This supplementary document contains the following information in each of the sheet:

* Collinearity matrix

This table shows the collinearity among predictor variables

* Collinearity scatter plot

This plot shows the collinearity trend, if present.

* Regression Models

The various models that were studied for this report

* Outlier Data

This plot shows a possible influential data in the dataset.

* Linearity

This plot shows that the linearity assumptions hold good in this model

* Independent observations

This plot shows that the observations are independent

* Randomness

This plot shows that the same is random and normally distributed.

* Auto-correlation

This plot shows that there is no auto-correlation in the residuals

* Homoscedasticity
  1. R- Code used to develop the regression models.
* Natural\_Gas\_Sales\_Regression\_Code.R
* Natural\_Gas\_Sales\_Regression\_output.txt