583 Project Report

Dataset Introduction and Hypotheses

The dataset we are using examines how the price of Mono-ethylene glycol (MEG) is related to upstream and downstream variables. Originally, there are 29 explanatory variables with a lot of missing data. With the help of the domain knowledge, We transformed them into 8 variables using the weighted method. We will be using the new transformed variables in our analysis, including the price of crude oil and coal, upstream MEG profit, MEG profits, MEG operating rate, downstream profits, downstream operating rate, downstream inventory, and inventory. The response variable is the price of MEG. The purpose of this report is to examine how the upstream and downstream variables affect the price of MEG, and to predict the price based on the variables.

The hypotheses are as follows: We expect that the price of MEG will be positively related to the price of crude oil and coal, the upstream MEG profit, MEG profit, downstream profits, and downstream operating rate. As the price of crude oil and coal is a measure of the cost of MEG, and the price will be higher when the cost is higher. Besides, the upstream MEG profit, MEG profit, downstream profits, and downstream operating rate are all measures of the demand of MEG, and the price will be higher when the demand is higher.

We also expect that the price of MEG will be negatively related to the inventory, as the inventory is a measure of the supply of MEG, and the price will be lower when the supply is higher.

We will also try to find an appropriate model to predict the price of MEG based on the variables.

Description of the Dataset

Variable Name	Unit of Measurement	Continuous vs Discrete
date	YYYY/M/D	Discrete
Price_MEG	CNY(¥)/ton	Continuous
Price_crude_oil_coal	USD($)/barrel	Continuous
Upstream_MEG_Profit	CNY(¥)/ton	Continuous
MEG_Profit	CNY(¥)/ton	Continuous
MEG_operating_rate	Percentage	Continuous
Downstream_Profits	CNY(¥)/ton	Continuous
Downstream_Operating_Rate	Percentage	Continuous
Downstream_Inventory	10,000 Tons	Continuous
Inventory	10,000 Tons	Continuous

Table 1: Description of the variables

As stated from the Table 1 above, the dataset contains 10 variables. The first variable is the date. Price of MEG is the response variable. All the other variables are explanatory variables. The date is a discrete variable, and the rest of the variables are continuous variables. The dataset collected weekly from 2018/1/5 to 2022/11/4, and there are 239 observations in this dataset. The dataset is collected from the Lili's previous employer, and Lili is allowed to use the dataset for this project.

Regression Analysis

Normality check for the response variable

We conduct a kernel density estimate for Price of MEG using an Epanechnikov kernel. The Epanechnikov kernel is a nonparametric kernel that is used to estimate the probability density function of a random variable.

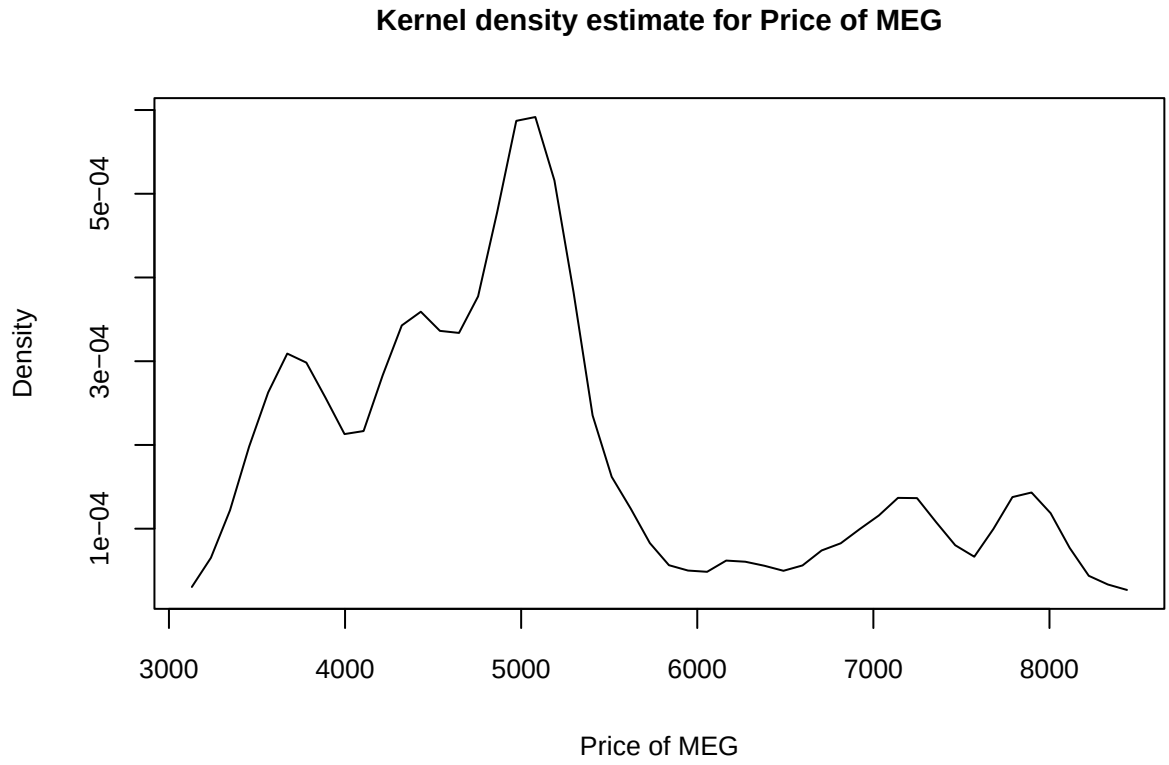


Figure 1: The kernel plot clearly shows the distribution of the

Figure 1 clearly shows the distribution of the response variable is not normal. We also conduct a Pearson test for normality. With an extremely small p-value, we reject the null hypothesis that the response variable is normally distributed. Therefore, we tried to transform the response variable to make it more normal first.

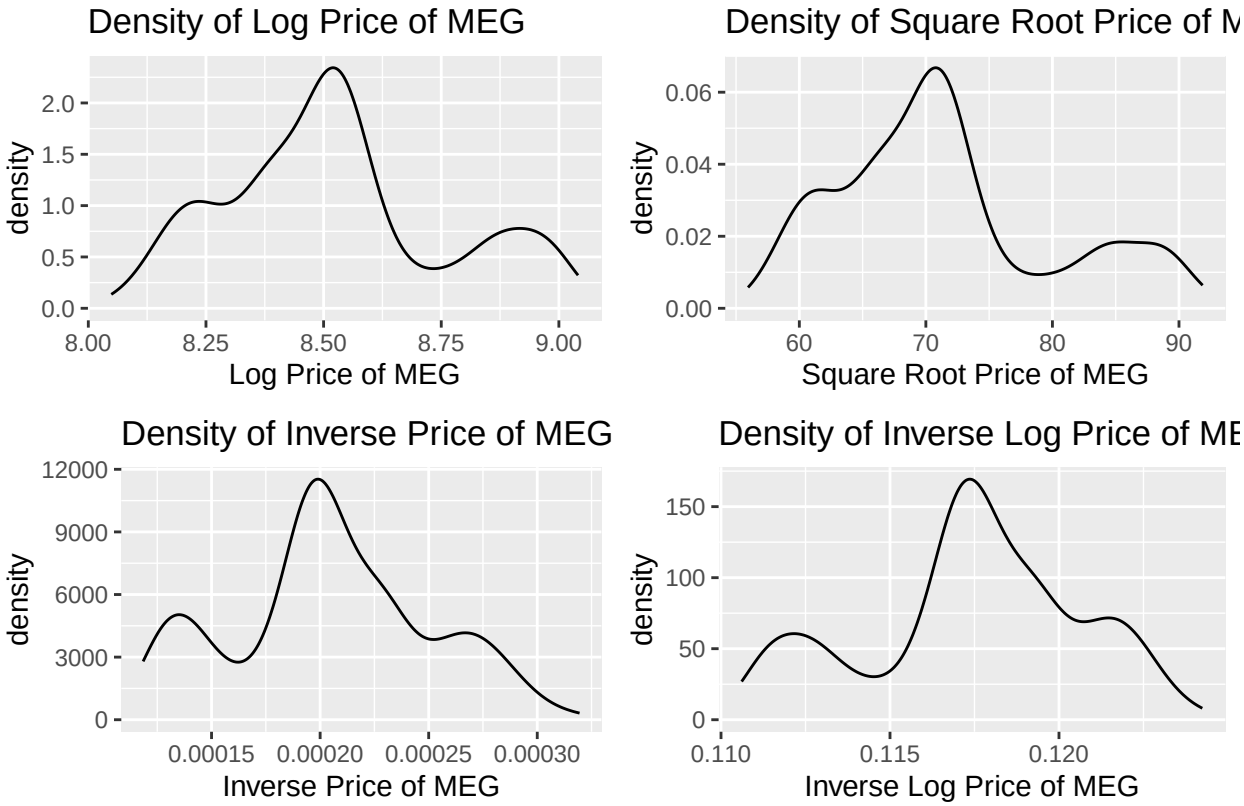


Figure 2: Density plots for different transformations of Price of MEG.

As shown in Figure 2, after using log, square root, inverse transformation, and inverse log transformation, we found that all of them can not handle the skewness and multi-peaks of the response variable. So we decided that nonparametric regression is the best choice for this dataset.

Check the correlation between variables

Before fitting a nonparametric regression model, we need to check the correlation between the response variable and the explanatory variables.

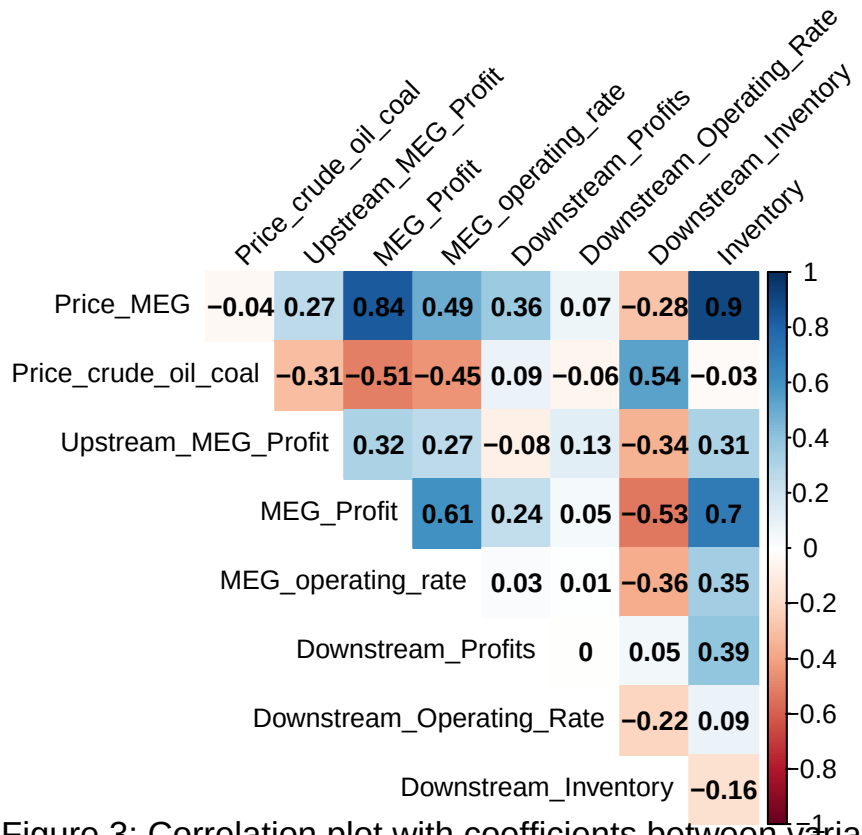


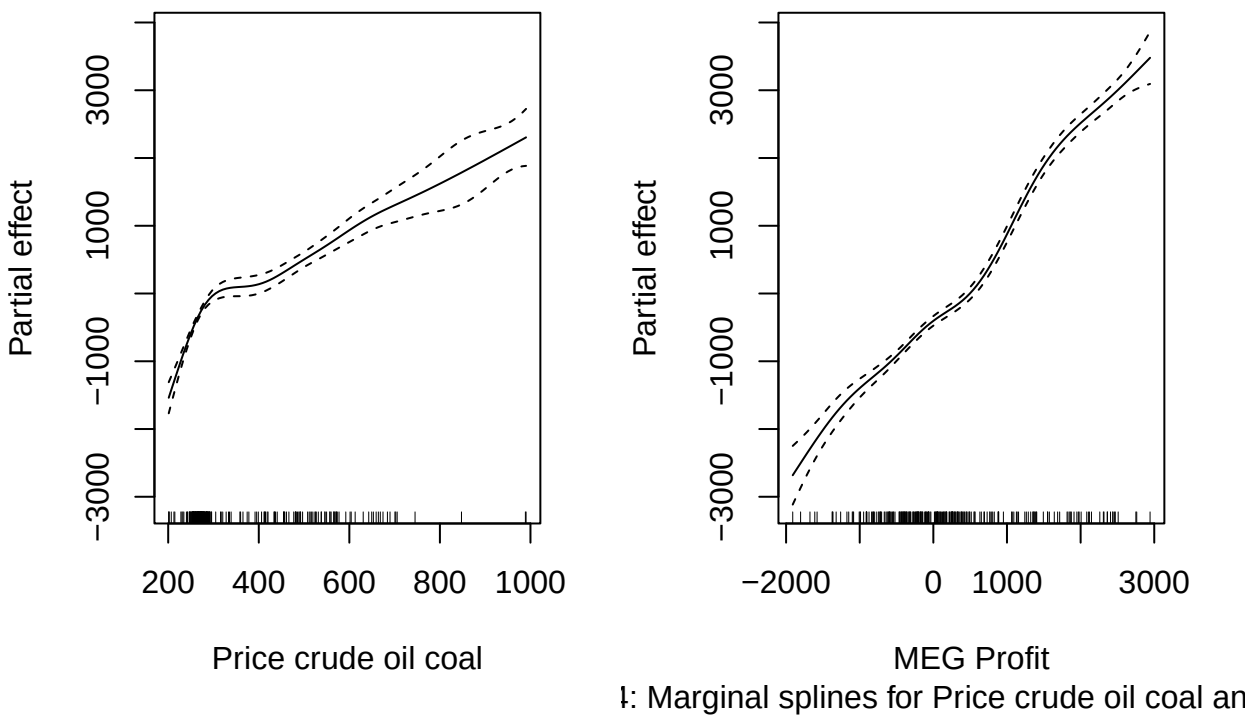
Figure 3: Correlation plot with coefficients between variables

Figure 3 shows that the price of MEG is positively highly correlated with MEG profit and inventory, and has almost no correlation with the price of crude oil and coal. Surprisingly, we expected that the price of MEG will be positively correlated with the price of crude oil and coal, and negatively correlated with the inventory.

Nonparametric local linear regression

After fitting the nonparametric local linear regression model, the summary of the model shows the R-squared is 0.997, which raises the concern of over-fitting. We then performed the consistent nonparametric test of significance, we found that the p-value of the price of crude oil and coal and MEG profit are extremely small, which means that they are significant. The p-value of the price of Upstream MEG profit, downstream operating rate and inventory are close to the 0.05 significance level, so we decided to drop them from the model. All the other variables are not significant according to the test.

Generalized additive model



ł: Marginal splines for Price crude oil coal an

We fitted a generalized additive model with the Price of MEG as the response variable and the Price of crude oil and coal and MEG profit as the explanatory variables. The summary of the model shows that the adjusted R-squared is 0.945, which is similar to the nonparametric local linear regression model.

This partial effect plots in Figure 4 can show the relationships between the response variable and the explanatory variable, while accounting for the effect of the other explanatory variable.

Both plots show that the Price of MEG increases as the Price of crude oil and coal, and MEG profit increase. The right plot shows the relationship between the Price of MEG and the MEG profit is more linear compared to the left plot. The left plot also has more uncertainty in the fitted smooth function, which means that the relationship between the Price of MEG and the Price of crude oil and coal is more complex.

Conclusion