Rummel Fuels Data Analysis

Junyuan Ke

Chen Fang

Senhao Hu

1.Pattern Analysis

 Based on the given historical data, the mean is 250 gallons, with a standard deviation of 15 gallons. So the lower 95% is (mean - 2 \* standard deviation) 220 gallons; the upper 95% is (mean + 2 \* standard deviation ) 280 gallons. Therefore, 95% of the number of customer, which is 63.65(round up to 64) should fall in 220~280 range. According to the frequency table below, only 9+4+3 = 16 customers fall in 220-280 gallons range, which conflicts with the empirical rule. Therefore, based on the current sample, the usage patterns changed.

If the current sample still follows the patterns, 95% of the data will fall within the first two standard deviations.

Also, based on the summary statistics of Rummel Fuel Data, the mean is 253.67 gallons, which is close to the historical average usage per customer. However, the standard deviation increases sharply to 139.53 gallons. According to the empirical rule, the upper 95% of the data is mean + 2 standard deviations, which is 533 gallons. The lower 95% of the data is mean - 2 standard deviations, is -25.39 gallons, which crosses zero. 

Since we have already known the standard deviations and means both historically and currently, we can calculate Coefficient of Variation(CV) by using standard deviation divided by the mean. Historically, the CV is 15/250, which equals 6%; currently, the CV is 139.53/250, which equals 55.8%. As we all know, CV measures the volatility of the data. The higher the CV is, the more volatile the data is. As a result, knowing the historical average and standard deviation, the usage patterns changes based on the different volatilities. However, if we do not know the historical standard deviation, we can only be able to analyze the pattern by comparing the average fuel usage. We’ve noticed that the mean of Oil Usage only differs slightly with the historical data(250 and 253.6). Therefore, our understanding about the usage pattern does not change if only knowing the historical average.

2.Using degree days to predict oil use:



The model explains the relationship between Oil Usage and Degree Days. According to the regression statistics, there are 67 observations. The correlation coefficient between the observed value of usage of oil and predicted value of usage is 0.655, suggesting a strong positive relationship between Oil Usage and Degree Days. The model suggests that more than 43% of the Oil Usage in the sample can be explained by Degree Days(due to an R Square value of 0.4289), showing a strong relationship between Oil Usage and Degree Days. Since there is only one independent variable and the adjusted R Square is 0.4201, it indicates an extremely low penalty of adding a new independent variable, which can also be interpreted as the unbiased estimate of R Square. 

The Standard Error is 106, which measures the dispersion from prediction.The sum of squared regression is the total variation explained by the regression line, which is 551149. The sum of squared due to residual is the residual unexplained, which is 733803. In addition, the degree of freedom of regression is 1, which means there is only 1 independent variable. The degree of freedoms of error is 65. The mean squared regression means on average, 551149 variance are explained by each degree of freedom. The mean squared error means on average, 11289 variances remain unexplained by each degree of freedom. Also, the F value indicates that the proportion of explained variance over unexplained variance is 48.82052. The extremely low Significance F value 1.83E-9 indicates that it is a good model: if we assert that there is a relationship between Oil Usage and Degree Days, the probability that we are wrong is almost zero (Significant F = 1.83E-9).

The coefficient of the predicted regression line is 0.268 with an interception of 57, therefore the function is The usage of oil (gallons)=0.268\*degree days+57.452. The t Stat shows coefficient over standard error, which is 6.987 and by looking at the lower 95% and upper 95% of the confidence interval for coefficient, which is 0.1916 and 0.345, does not cross zero, therefore, it is an appropriate prediction.

 The model reveals that there is a strong and positive relationship between Oil Usage and degree days. Therefore, we are confident to say that as degree days go up, the oil usage will go up as well. A house with higher degree days tend to have a relatively higher oil usage. However, although the R Square value is promising since it is greater than 0.3, there could be other independent variables to further explain the variance of the data.

3. The Model Combined



The combined model explains the relationship between Oil Usage and Degree Days, Number of People, and Home Index combined. The correlation coefficient between the observed value of usage of oil and predicted value of usage is 0.848, indicating a strong correlation. The model suggests that more than 71% of the Oil Usage in the sample can be explained by Degree Days(due to an R Square value of 0.7186), Number of People, and Home Index. However, because we added two more predictives to the model, we have to issue a penalty to each additional data, and therefore modifying the Adjusted R Square to 70%. This reveals that there is a strong correlation between Oil Usage and the three independent factors.The Standard Error is reasonable comparing to the Mean (with a Coefficient Variation of 29.9%) as we are making 67 observations.

Having multiple independent variables (with a degrees of freedom of 3) results in a relatively large amount of explained sum of squares compared to unexplained sum of squares. The Mean Squared Regression, therefore, is a lot larger than the Mean Squared Error, leading to a satisfying proportion of explained variance over unexplained variance(F value of 53.6).

Meanwhile, we know the relationship is strong because the probability that we are wrong if we assert that there is a relationship is almost zero (supported by the significance F value of 2.5E-17).

Knowing the Coefficient and Standard Error of the regression lines, we are able to understand the increase in Oil Usage with respect to the other three variables. First of all, 1 degree increase in Degree Days is likely to cause 0.28(with the confidence interval of +- 2\*0.0028) increase in Oil Usage. Secondly, one level increase in Home Index will lead to a 61.5(with a confidence interval of +-7.64) rise in Oil Usage. Lastly, one more person in the house will likely lead to 1.22 (with a confidence interval of +-7.8) increase in Oil Usage. Comparing to the t-Stat of Degree Days and Home Index, the ratio between Coefficient and Standard Error of Number of People is relative close to “0”. We can therefore observe that, unlike the other two factors, the confidence interval of Number of people crosses zero. P-value of the three independent factors proves our thinking from another angle: unlike the P-values of Degree Days and Home Index which are fairly close to zero, the P-value of Number of people indicate that the chances that there is no relationship between Oil Usage and Number of People is as high as 88%. Consequently, it is unconvincing to assert that there is a relationship between Oil Usage and the Number of People.

The model reveals that there is a strong and positive relationship between Oil Usage and the three independent factors. Therefore, as Degree Days, Home Index and Number of People go up, we can confidently predict that the Oil Usage goes up as well. A house with large Degree Days, Home Index and Number of People is more likely to have a relative high Oil Usage, and therefore should have a higher priority in fuel delivering. As a result, Rummel should first prioritize houses with higher Degree Days, Home Index and Number of People when they are scheduling oil delivery.

However, the company should also be aware that the relationship between Oil Usage and Degree Days is stronger than the other two factors(judging from their separated single regression analysis). Rummel should accordingly put houses with higher degree days to higher priority.

4. Better Model





 One weakness of the last model is that the number of people in a house cannot predict oil usage as good as home index and degree days. Comparing to the promising P-values of the two other independent variables(1.13E-14 and 2.96E-11), the probability that we are wrong by asserting the relationship between oil usage and number of people is extremely high(with a P-value of 87.6%), meaing that if we assert there is a relationship between oil usage and number of people, there is 87.6% that we are wrong. Meanwhile, for the number of people as a variable, the lower 95% is -14.37, while the upper 95% is 16.82, which suggests the set of data would cross zero and part of it would be negative. Since the number of people could not be negative, the number of people cannot be valid as an independent variable.

Judging from the correlation matrix, the correlation between Oil Usage and the Number of People is weaker comparing to the other two factors. Therefore, we need to remove the number of people as an independent variable in order to establish a better model.

 We know the model is better also because there is no noticeable decrease in the percentage of oil usage explained after removing the number of people (in fact, the R-square only decreased about 0.01% and the adjusted R-square increased about 0.5%). Instead, it is reasonable to believe that the number of people barely determines the demand for fuel of the house. Therefore, houses with more people barely use up their oil more quickly.

 The updated model explained most oil usage data by combining degree days and home index. This model still shows a strong and positive relationship between Oil Usage and Degree Days and Home Index due to a high R square value 0.7185, meaning more than 70% of Oil Usage can be explained by Degree Days and Home Index. The correlation coefficient between the observed value of usage of oil and predicted value of usage is 0.848, revealing a strong and positive correlation. Because we want as less independent variables as possible, Adjusted R square increased to 71% due to the removal of Number of People as an independent variable.

Having multiple independent variables (with a degrees of freedom of 2) results in a large proportion of explained variance over unexplained variance(F value of 81.67). Comparing to the last model, neither Sum Squared Regression nor Sum Squared Error changed noticeably. Meanwhile, Mean Squared Regression is significantly larger than Mean Squared Error. Another reason that it is a better model than the former one is that it contains a much smaller Significance F value, 2.42E-18 in comparison with that of the former model, 2.46E-17, showing that if we assert there is a relationship between oil usage and home index and degree days is almost zero.

When we look at Degree Days and Home Index individually, two linear regression lines are formed. On one hand, 1 degree increase in Degree Days is likely to cause 0.28(with the confidence interval of +- 2\*0.0028) increase in Oil Usage. On the other hand, one level increase in Home Index will lead to a 61.5(with a confidence interval of +-7.64) rise in Oil Usage. We are confident with both variables because both P-values are extremely close to zero(2.86E-15 and 2.03E-11). In addition, the T-stats of both variables are relatively big, showing that Standard Error is relatively small comparing to the Coefficients.

**Appendix**

Descriptive Statistics of the data:

Histogram of Oil Usage:

|  |
| --- |
|  |

Bin Chart of the Histogram:

Scattered Plot of Oil Usage and Degree Days



Regression Analysis of Degree Days:



Regression Analysis of Number of People:



Regression Analysis of home index:



Regression Analysis of three variables combined



Regression Analysis of Degree Days and Home Index:



Correlation Matrix:

