Mariama SOUMAHORO

PROJECT

ABSTRACT

The data presented to us contain 13 observations with three variables: Foam height(centimeters), Time (seconds) and Liquid height. It describes how the height of the foam (dependent variable) is being affected by time and Liquid's height(centimeters). A multiple linear regression model was fit to the data in order to predict the foam height by establishing a statistically significant relationship with the different factors. After fitting both linear and non-linear regression models using the information given, a polynomial regression had the best fit. Both independent variables turned out to be important determinants. The techniques used give helpful insights into the set-up of the experiment.

2 DATA VISUALIZATION AND EXPLORATION

Diving into the data and taking a closer look at the information given helps us get more familiar. All the variables are continuous. We first look at the summary statistic of all the variables. The highest the foam can reach is 17.400 cm. The lowest height is 2.900 cm. The average height that the foam can reach is 8.481cm. We then plot the data to see if there exists a relation between the dependent variable foam's height and independent variables.

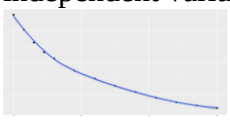
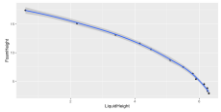


Figure 1 represents the scatterplot of the independent variable: Time against Foam's height. It is clear that the relationship between the two variables is not linear. Checking for the correlation, we get a value of -0.9515949. This shows that, although not linear, there is strong negative relationship between the Height of the foam and Time. As time increases, the height decreases.

Next, we look at the relation between the height of the foam and the liquid height by plotting the data and checking for the correlation.



The correlation is equal to -0.968875 which shows a strong negative nonlinear relationship between the two variables. Looking at the plot, we can see that as the height of the liquid increases, the height of the foam continues to decrease.
After viewing the data, the next step is to fit a multiple linear regression.

3 MODEL SELECTION AND INTERPRETATION

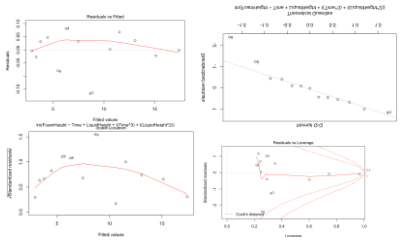
With section 2, we were able to establish some patterns between foam's height and the predictors. In this section, we will summarize the patterns using multiple linear and nonlinear regressions modeling.
We first fit a multiple linear regression with all the predictors. We get the F- statistic: F= 1904; it can be concluded that there is a relationship between predictors and response variable. Based on the 'p-value', all predictors are significant for the model. R^2 (multiple-R-squared) value indicates how much variation is captured by the model. R2 closer to 1 indicates that the model explains the large value of the variance of the model and hence a good fit. In this case, the value is 0.9974 (closer to 1) meaning 99.74% of the data is explained by the model. The AIC is equal to -27.76605 and percentage error is: 3.134328. Although the value of the R^2

is high, it does not capture entirely the relation between the dependent variable and independent variables. We plot the residuals against each predictor to see if the predictors require any transformations in the model. The results show us that we need to include polynomial terms in the model. We therefore refit the model using Stepwise selection. A better R^2(0.998) and Adjusted R^2(0.997) are found meaning that 99.8% of the data is being explained. The AIC is -23.16035 which is very low. All the p-values of the predictor variables are significant. All those observations are indication that the model fitted is the best model. We also fit multiple nonlinear regressions models to see if we can get better results. Out of those models, the model with equation: y = (a*x1 + b*x2)/(b+x1+d+x2) turns out to be the model that fits the data closely. Indeed, it had the smallest residual standard error (0.53) and the smallest AIC.

The next step is to decide between the polynomial model and the nonlinear model. By doing so, we compare the AIC, and observe that the polynomial model had the smallest AIC. Consequently, our final model is: $\widehat{Y}$ ***= 1.802e+01 -2.438e-02(Time) - 9.820e-01 (LiquidHeight) + 5.272e-08(Time^3) - 7.653e-02(LiquidHeight^2)***.

Regarding the coefficient of Time, for a unit increase in Time, there is a decrease in the height of the foam by -2.438e-02. The same can be said for Height of Liquid. For the polynomial factors, we note that the height increases and decreases as the predictor increases.

We continue our analysis by checking for the residual plots which help us detect if there is an anomality with the model chosen.



Starting with the Residual vs fitted graph for the Nonlinearity of the data, we observe that the line is approximately a curve. It suggests that there is a relationship between the predictors and the outcome variable.

Non constant variance: using the scale location plot, we have that most of the residuals are scattered around the curve line, this suggests that there is a constant variance in the residual's errors. The Q-Q plot shows most of the points are on the line. Also using the Shapiro test, the p-value is not significant, therefore we conclude that the residuals are normally distributed. The residuals vs Leverage help us find the influential points. Indeed, we have 4 influential points with no outliers.

LIMITATIONS

At the end of our study, after performing both multiple linear and nonlinear regression, we were able to determine a good model to fit the data and understand the relationship between foam's height and the covariates. All the predictors have a significant impact on the dependent variable. Around 99.8% of variation in the data was explained. Although the R^2 is a great, the size of the data (rather small) could be misleading. Perhaps more predictors should be added or more transformations of the data are required. There is still room for improvement as we can apply advanced techniques to further improved the accuracy of the model.