Corey Huang
STAT 425 Final Project - Executive Summary
Due: December 13, 2021

Overview:

In this study, the ‘bubblewrap.csv’ dataset has been given, which consists of data from a study comparing different operating conditions in relation to the process of bubble wrap production. I was tasked with fitting a model to find the optimal combination of ‘line speed’ and ‘percent load of additives’ factor levels to find the highest ‘production rate’. In this report, you will uncover my findings pertaining to data exploration and manipulation, creating and testing the validity of different models through diagnostic tests, and obtaining factor level means through confidence intervals in order to discover the optimal combination of factor levels in producing the highest production rate.

Report:

I begin by understanding the dataset and its variables before fitting any model. From the prompt, the factor ‘line_speed’ and ‘loading’ were the most significant factors in the study. Therefore, we must separate the data into factor levels, with the ‘loading’ factor containing the levels ‘0’, ‘2’, and ‘4’ and the ‘line_speed’ factor containing ‘36’, ‘37’, and ‘38’ (Figure 2). After checking each variable type, we see that ‘line_speed’ and ‘loading’ are now seen as ‘factor’ rather than ‘integer’ (Figure 4). Next, I output the first five rows of the dataset (Figure 1) and summary statistics for each variable (Figure 3) to get a broad understanding of the distribution of the response variable. Another data exploration tool I utilized was boxplots, where I plotted each factor level against rate to understand how the distribution of rate differs from each factor. The ‘line_speed’ boxplot shows that the factor level with 37 m/mm has the least variability and highest average in rate (Figure 5), and the ‘loading’ boxplot shows that the factor level with 4% loading of additives had the highest mean and lowest variability in rate (Figure 6). From the results of the boxplot, I make the initial prediction that the optimal combination that produces the highest production rate is with a line speed of 37 m/mm and a 4% loading of additives. After, I use interaction plots to see if interaction effects are present between factor levels. From the ‘loading’ and ‘line_speed’ interaction plots, we can see that there are interactions and thus interaction effects in both the ‘loading’ plot (Figure 7) and the ‘line_speed’ plot (Figure 8). Having interaction effects in the data may signify that there is an interaction term that is significant to the model, therefore we begin model selection by fitting the full model with an interaction term present.

To create the best fitting ANOVA model for this relationship, I begin by creating a full model with all the factors and levels included along with the interaction term, titled Model 1. The summary output of Model 1 states that the p-value of the ‘loading:line_speed’ interaction term is 0.6829 (Figure 9), meaning that the interaction term is not significant to the model and it can be dropped from the model. I then try the additive model, which drops the interaction term from the model, and output its summary (Figure 10). The significance of ‘loading’ and ‘line_speed’ factors increased in the new additive model, meaning that the additive model is a much better fit than the full model. Next, I run diagnostics on the model to see if the constant variance and normality assumptions are still met.

After finding the best fitted model, I check for constant variance by plotting the residuals against the fitted line (Figure 11). From the plot, it seems that constant variance is satisfied because there are no clear patterns that say otherwise. Additionally, I conduct the Breusch-Pagan and Levene’s Test for constant variance and obtain 0.117 (Figure 13) and 0.6196 (Figure 14) for p-values of both, respectively. Since the p-values are not statistically significant, this confirms that constant variance is satisfied in the model. After, I test for normality by plotting the Normal Q-Q plot (Figure 12) and histogram of the residuals (Figure 15). It is unclear from the Q-Q plot whether the normality assumption is satisfied or not, but the histogram shows that the distribution of residuals is left skewed and therefore is likely to not follow a normal distribution. This is confirmed in the Shapiro-Wilk test for normality (Figure 16), where the p-value is 0.01321. Because the normality assumption is not satisfied, it may make sense to perform a Box-Cox transformation on the model. However, by the Central Limit Theorem, the normal distribution is seen in generally large sample sizes, specifically of sample size of 30 or greater. This data set contains only 27 observations, which is not enough for the Central Limit Theorem to occur, hence the normality assumption is dismissed in this design.

Next, I look for any unusual observations by finding outliers and highly influential points. I find that the critical value to identify outliers is a value more extreme than 3.497923, and the one observation that exceeded this was the 12th observation (Figure 17). Additionally, I calculated Cook’s Distances for all the observations and ordered them by the highest absolute value (Figure 18). From this, I found no highly influential points because none of the observations were greater in absolute value than 1. Since there was one outlier in the dataset, I fit a new model, Model 3, with the outlier dropped from the data set and test if it made improvements to the model.

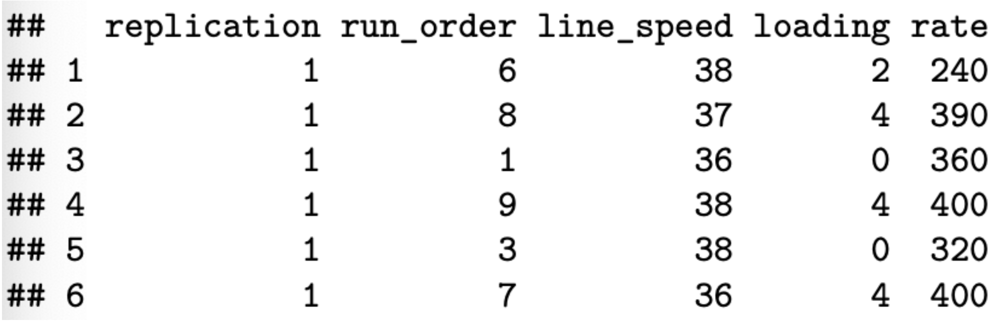
From the summary output of Model 3, I notice that both the ‘loading’ and ‘line_speed’ p-values are slightly less in Model 3 (Figure 19) than the additive model, meaning that both factors are even more significant in Model 3 than in the previous additive model. Afterwards, I look at the diagnostic plots for Model 3 (Figure 21) and notice that the diagnostics plots are very similar to the additive model. By comparing Levene’s Test statistic for the additive model (Figure 14) with Model 3 (Figure 22), we see that the p-value increases from 0.6196 to 0.9. This means that the constant variance assumption has been improved by dropping the outlier. Additionally, the p-value of the Shapiro-Wilk test for normality increased to 0.1824 (Figure 24), signifying that the normality assumption is now met. However, it must be noted that the normality assumption must not be given the importance it has in large scale samples due to the small sample size of 27 in this particular design. Now that we have the optimal model for representing the effect the two factors have on production rate, I can now use Tukey's test to find the specific factor levels that together produce the highest production rate.

After fitting the Tukey interval with Model 3, I output two plots that represent the differences within factor level means of both factors. Additionally, I output the numerical and p-values of each Tukey interval in order to get the most accurate measures in the test. The 'loading' Tukey interval displays from the plot (Figure 25) and p-values (Figure 27) that the '4' factor level results in the highest production rate on average for the 'loading'. Additionally, the 'line_speed' Tukey interval displays from the plot (Figure 26) and p-values (Figure 28) that the factor level '37' results in the highest production rate on average for the 'line_speed' factor.

Conclusion:

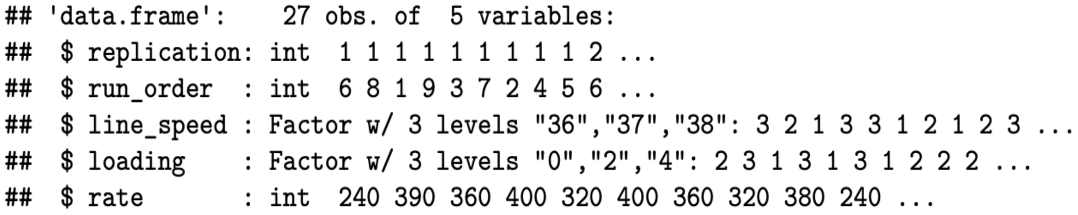
To summarize, I was given the task of finding the optimal combination of factor levels that results in the highest production rate. During model selection, where I built the best model for representing the association between the two factors and production rate, I found Model 3, the additive model with the 12th observation dropped from the data set, to be the best model for this design. This was the best model because it was the only model to follow both the constant variance and normality assumptions with higher accuracy than both the original additive model and full model. However, it is important to note that the normality assumption may not be significant for this design because the sample size is less than the desired 30 for normal distributions to occur. After conducting Tukey intervals on each factor, I find the factor levels that contain the highest production rate within each factor. Thus, the optimal combination of 'loading' factor levels and 'line_speed' factor levels that will produce the highest production rate is with a 4% loading of additives and a line speed of 37 m/mm.

Appendix:



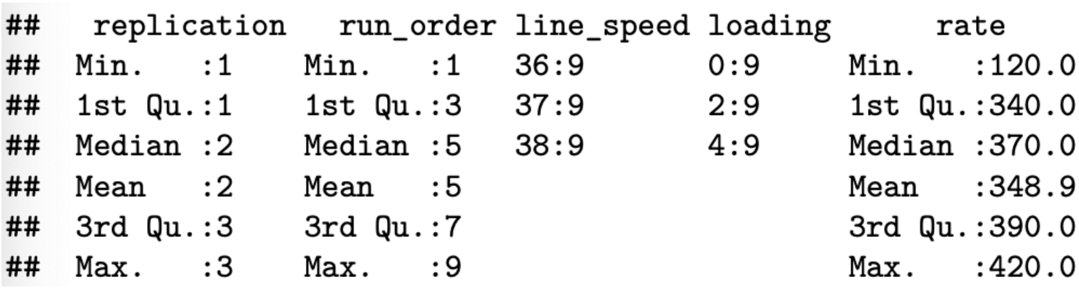
```
##   replication run_order line_speed loading rate
## 1           1         6         38       2  240
## 2           1         8         37       4  390
## 3           1         1         36       0  360
## 4           1         9         38       4  400
## 5           1         3         38       0  320
## 6           1         7         36       4  400
```

Figure 1: First 5 Rows of 'bubblewrap.csv'



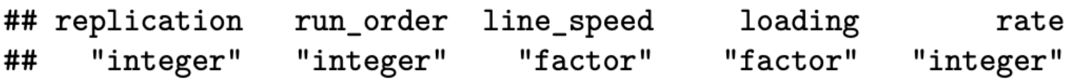
```
## 'data.frame':    27 obs. of  5 variables:
##  $ replication: int  1 1 1 1 1 1 1 1 1 2 ...
##  $ run_order  : int  6 8 1 9 3 7 2 4 5 6 ...
##  $ line_speed : Factor w/ 3 levels "36","37","38": 3 2 1 3 3 1 2 1 2 3 ...
##  $ loading    : Factor w/ 3 levels "0","2","4": 2 3 1 3 1 3 1 2 2 2 ...
##  $ rate       : int  240 390 360 400 320 400 360 320 380 240 ...
```

Figure 2: Understanding Data Types in Variables



```
##   replication   run_order  line_speed loading      rate
##  Min.   :1    Min.   :1   36:9       0:9     Min.   :120.0
##  1st Qu.:1    1st Qu.:3   37:9       2:9     1st Qu.:340.0
##  Median :2    Median :5   38:9       4:9     Median :370.0
##  Mean   :2    Mean   :5                      Mean   :348.9
##  3rd Qu.:3    3rd Qu.:7                      3rd Qu.:390.0
##  Max.   :3    Max.   :9                      Max.   :420.0
```

Figure 3: Summary Statistics of Each Variable



```
## replication   run_order  line_speed     loading        rate
##   "integer"   "integer"    "factor"    "factor"   "integer"
```

Figure 4: Another Example of Understanding Data Types in Variables

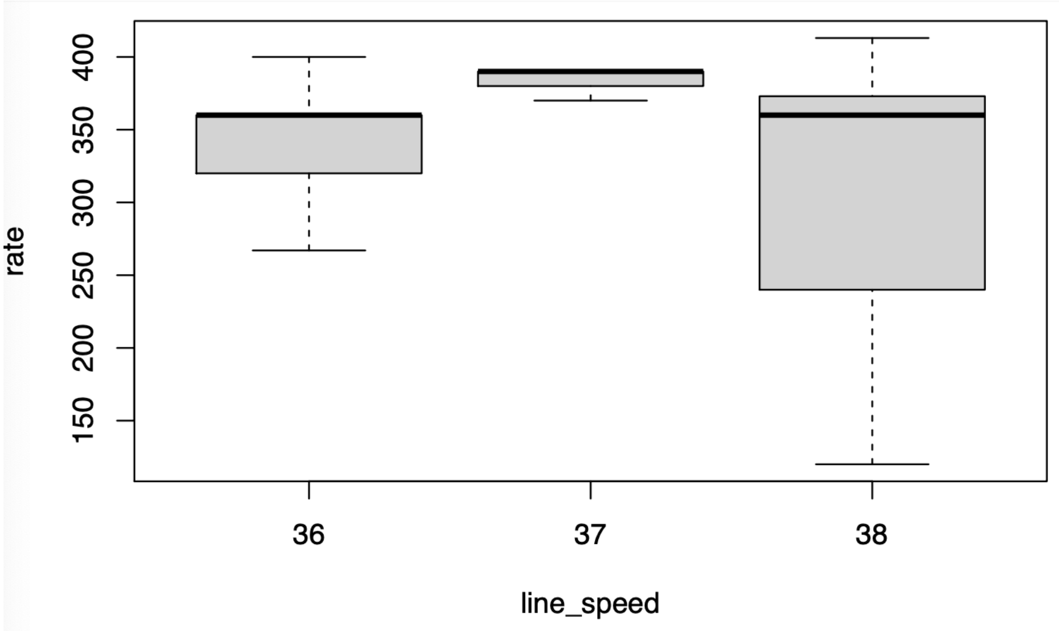


Figure 5: Boxplot for the 'line_speed' factor

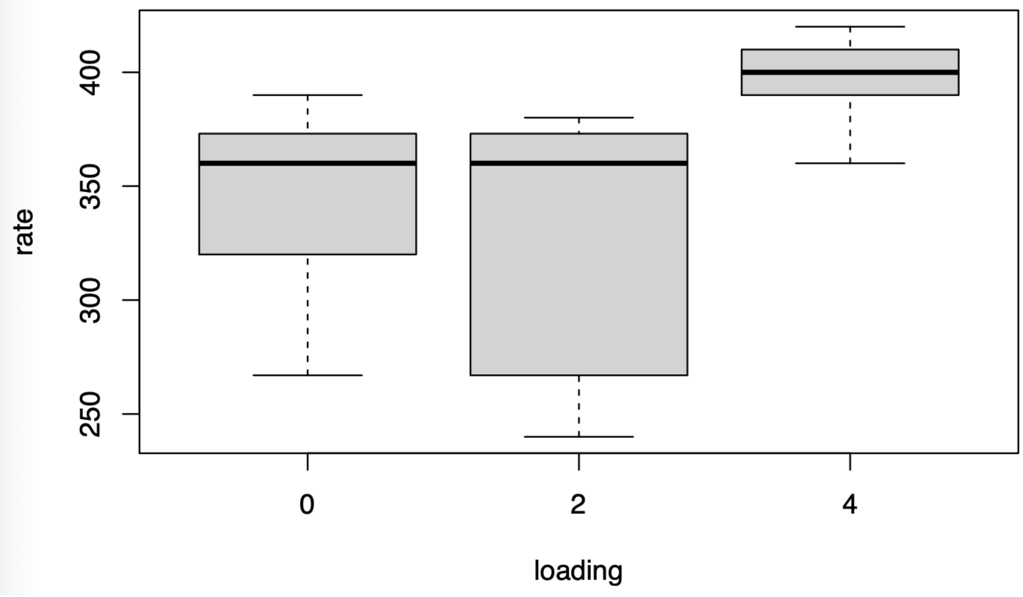


Figure 6: Boxplot for the 'loading' factor

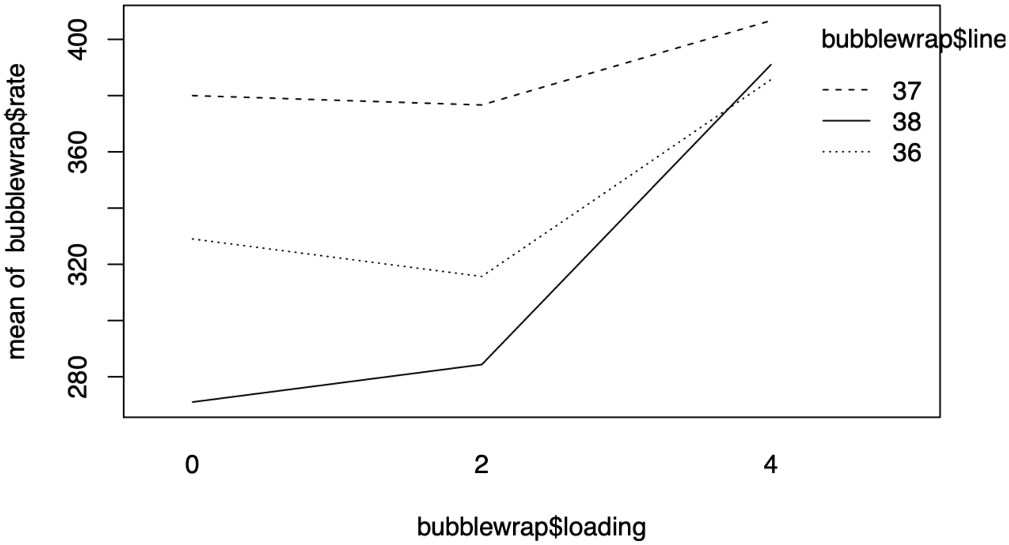


Figure 7: Interaction plot for the 'loading' factor

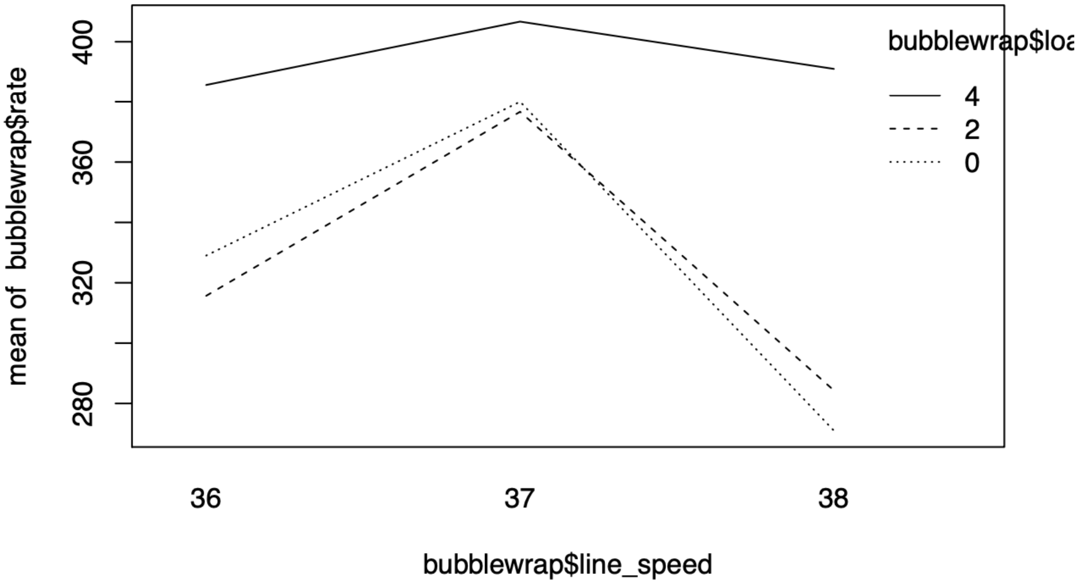
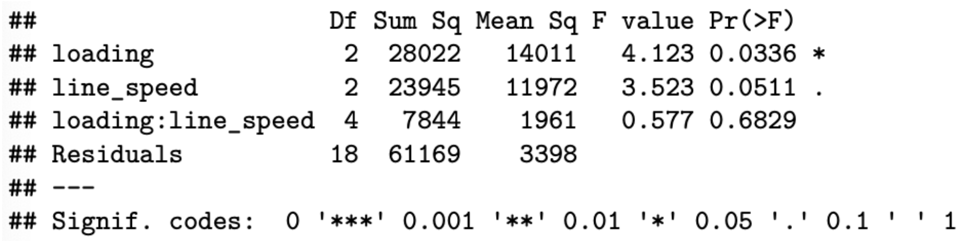
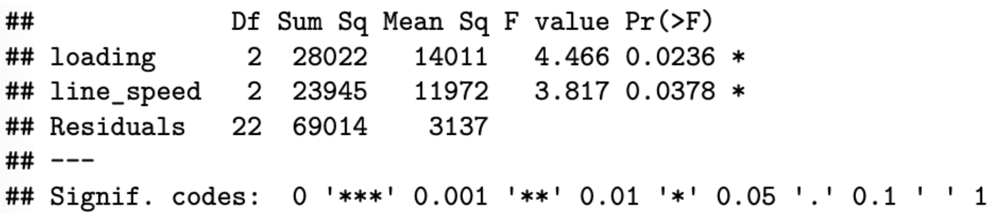


Figure 8: Interaction plot for the 'line_speed' factor



```
##                     Df Sum Sq Mean Sq F value Pr(>F)
## loading              2  28022   14011   4.123 0.0336 *
## line_speed           2  23945   11972   3.523 0.0511 .
## loading:line_speed   4   7844    1961   0.577 0.6829
## Residuals           18  61169    3398
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 9: Summary Output of Full Model with Interaction Term Included (Model 1)



```
##             Df Sum Sq Mean Sq F value Pr(>F)
## loading      2  28022   14011   4.466 0.0236 *
## line_speed   2  23945   11972   3.817 0.0378 *
## Residuals   22  69014    3137
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 10: Summary Output of Additive Model (Interaction Term Not Included)

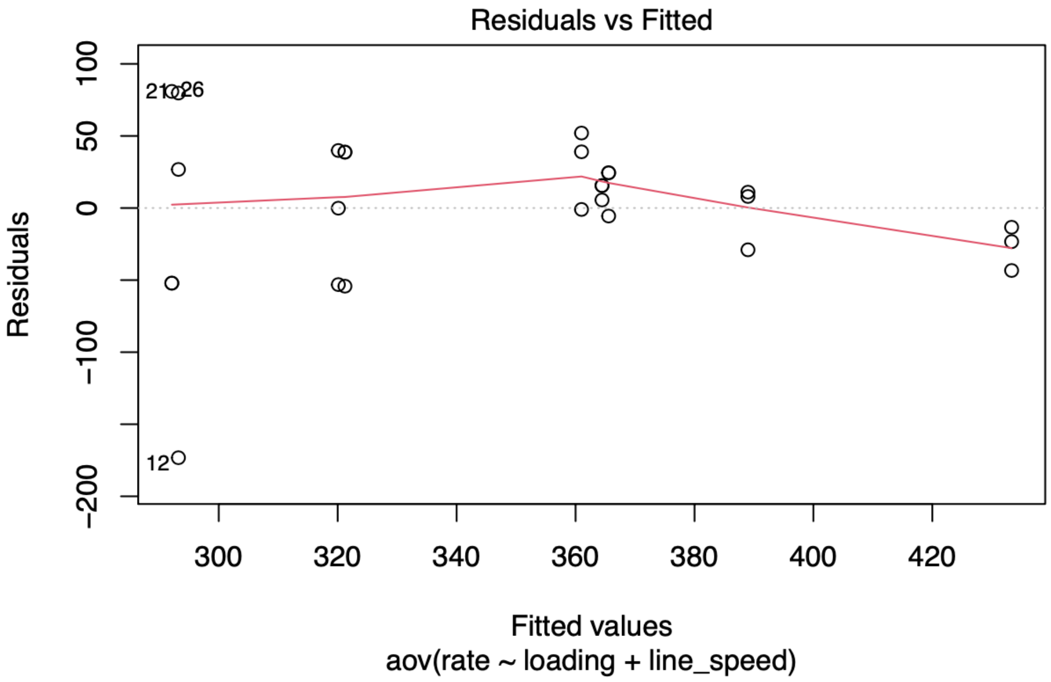


Figure 11: Residual Plot of Additive Model

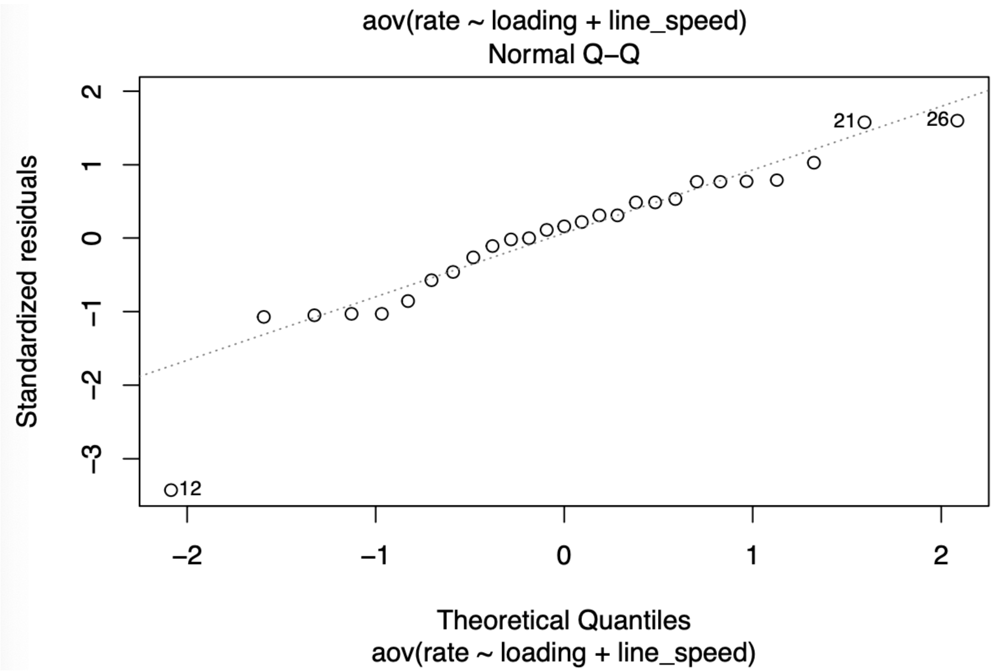
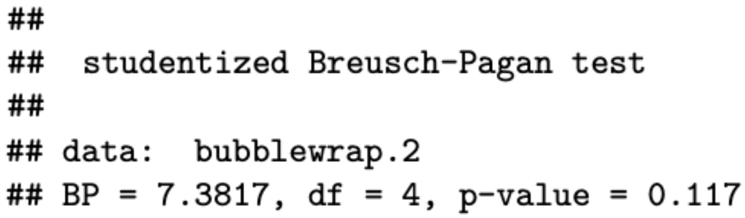
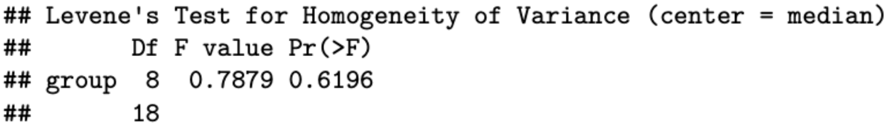


Figure 12: Normal Q-Q plot of Additive Model



```
##
##  studentized Breusch-Pagan test
##
## data:  bubblewrap.2
## BP = 7.3817, df = 4, p-value = 0.117
```

Figure 13: Model 2 Breusch-Pagan Test



```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  8  0.7879 0.6196
##       18
```

Figure 14: Additive Model Levene's Test

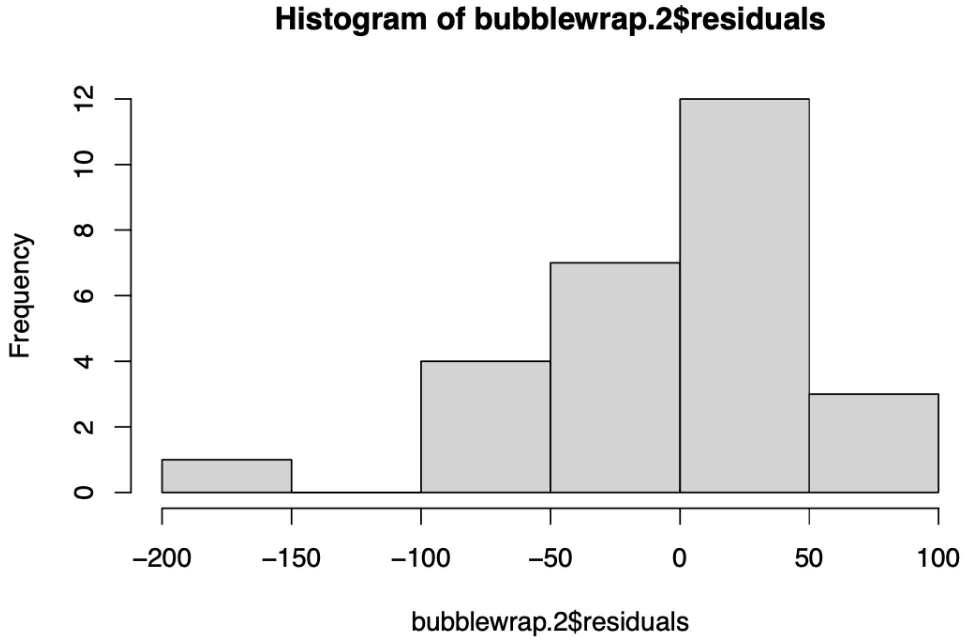
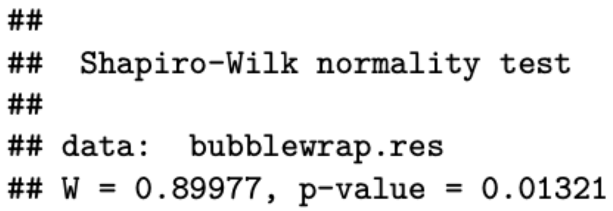
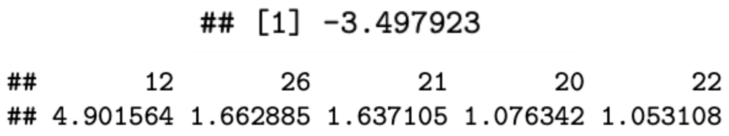


Figure 15: Histogram of Additive Model Residuals



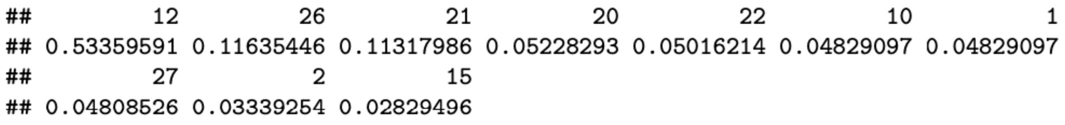
```
##
##  Shapiro-Wilk normality test
##
## data:  bubblewrap.res
## W = 0.89977, p-value = 0.01321
```

Figure 16: Additive Model Shapiro-Wilk Test



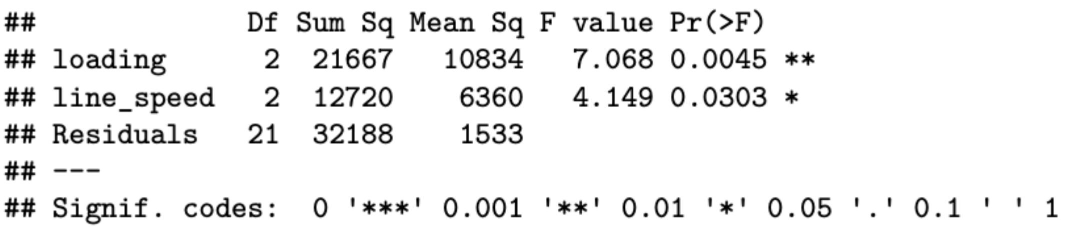
```
           ## [1] -3.497923

##       12       26       21       20       22
## 4.901564 1.662885 1.637105 1.076342 1.053108
```

Figure 17: Bonferroni Outlier Test Procedure



```
##          12         26         21         20         22         10          1
## 0.53359591 0.11635446 0.11317986 0.05228293 0.05016214 0.04829097 0.04829097
##          27          2         15
## 0.04808526 0.03339254 0.02829496
```

Figure 18: Cook's Distance



```
##             Df Sum Sq Mean Sq F value Pr(>F)
## loading      2  21667   10834   7.068 0.0045 **
## line_speed   2  12720    6360   4.149 0.0303 *
## Residuals   21  32188    1533
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 19: Summary Output of Model 3

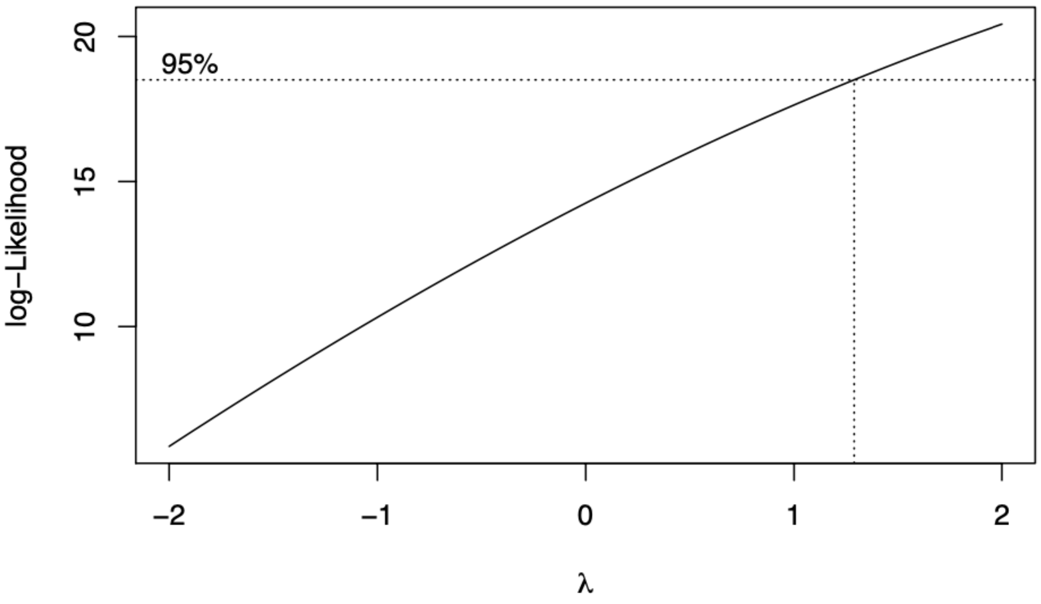


Figure 20: Box-Cox of Model 3

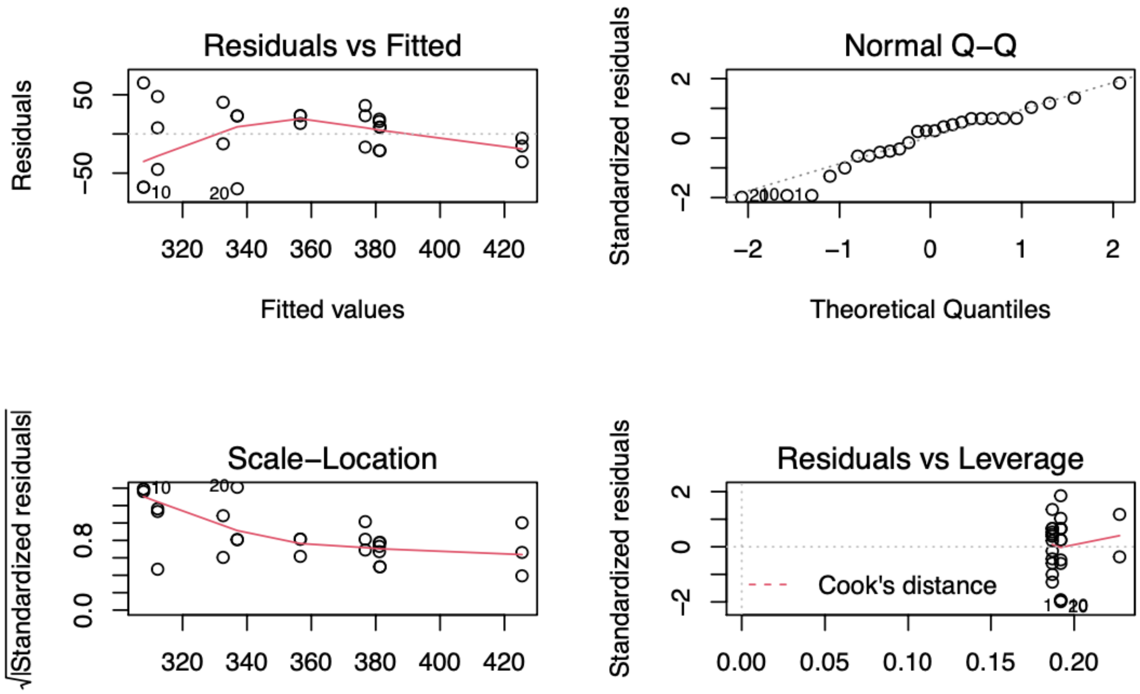
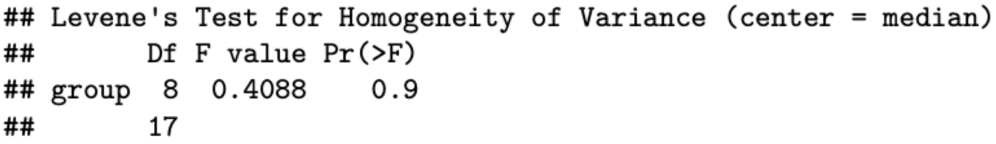


Figure 21: Diagnostic Plots for the Model 3



```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  8  0.4088    0.9
##       17
```

Figure 22: Levene’s Test for Model 3

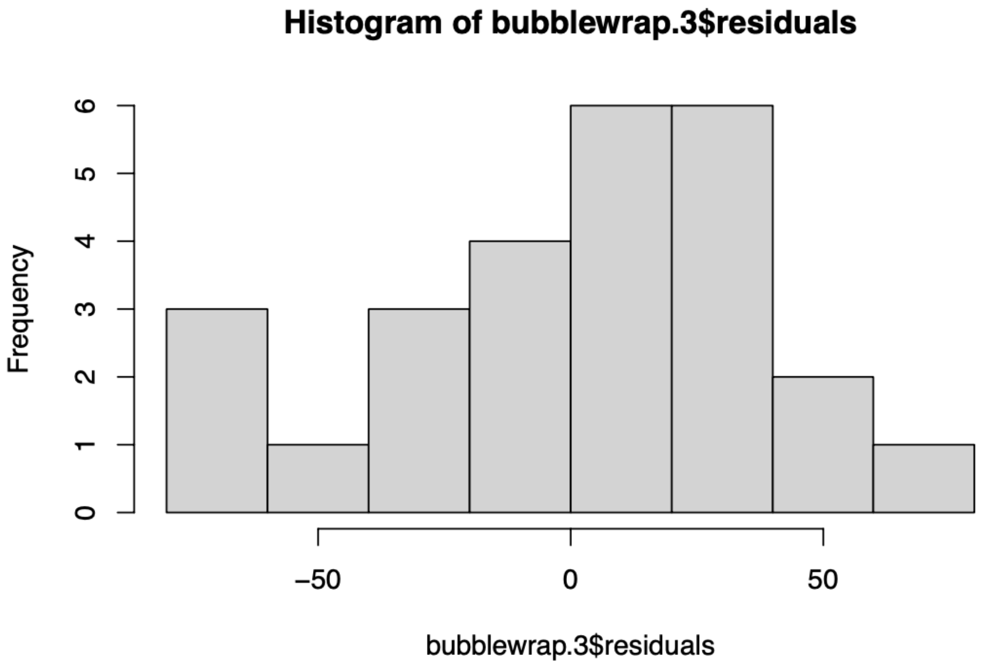
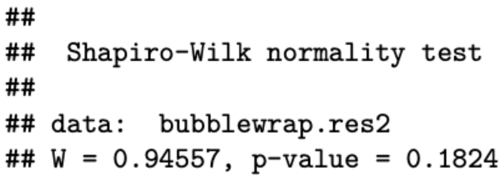


Figure 23: Histogram of Model 3



```
##
##  Shapiro-Wilk normality test
##
## data:  bubblewrap.res2
## W = 0.94557, p-value = 0.1824
```

Figure 24: Model 3 Shapiro-Wilk Test

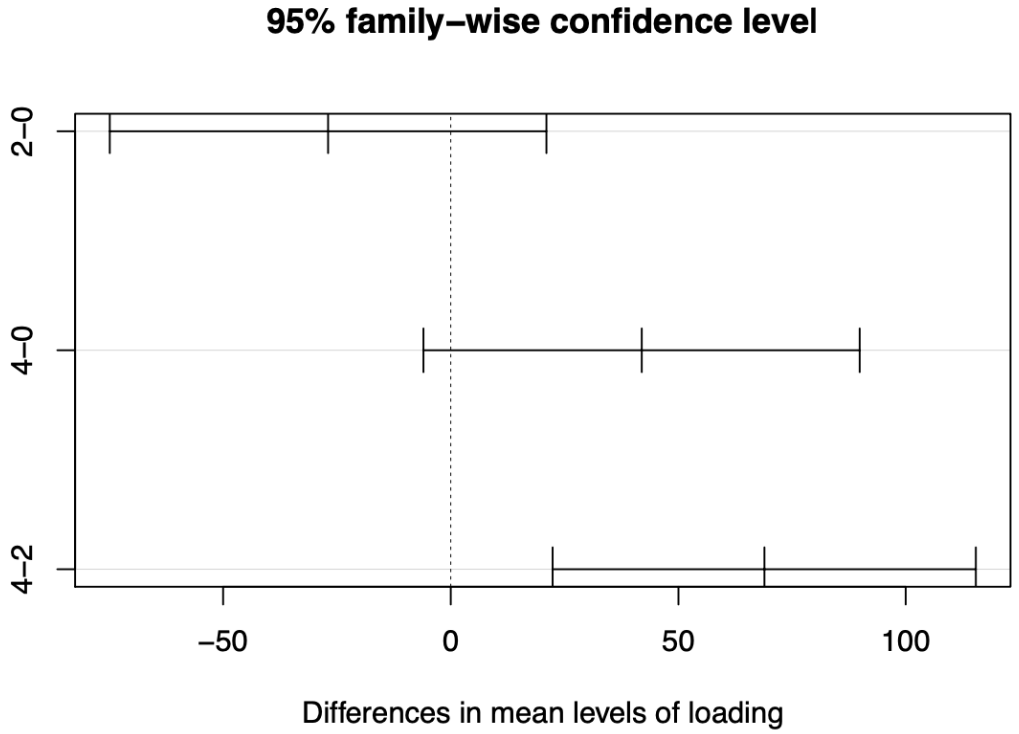


Figure 25: Tukey Interval of the 'loading' factor

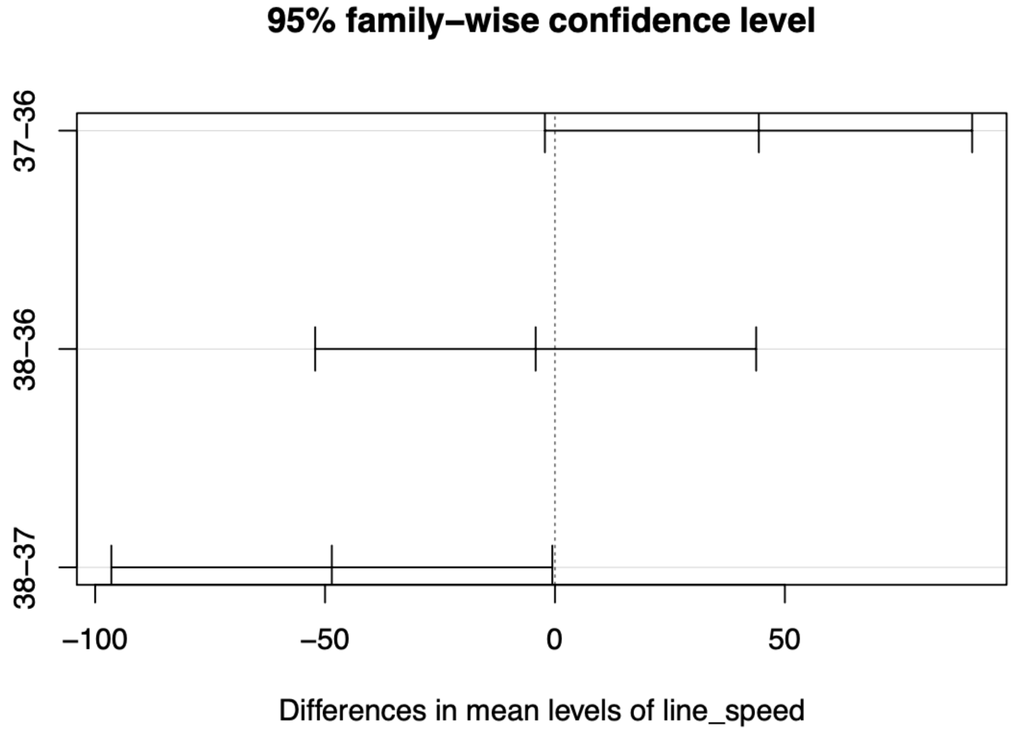
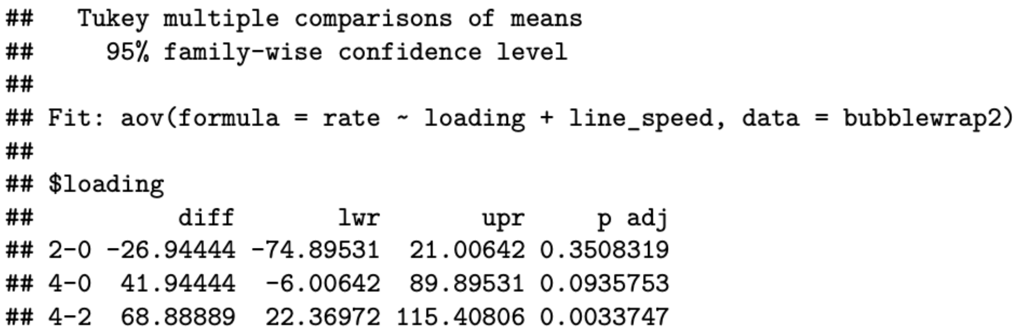
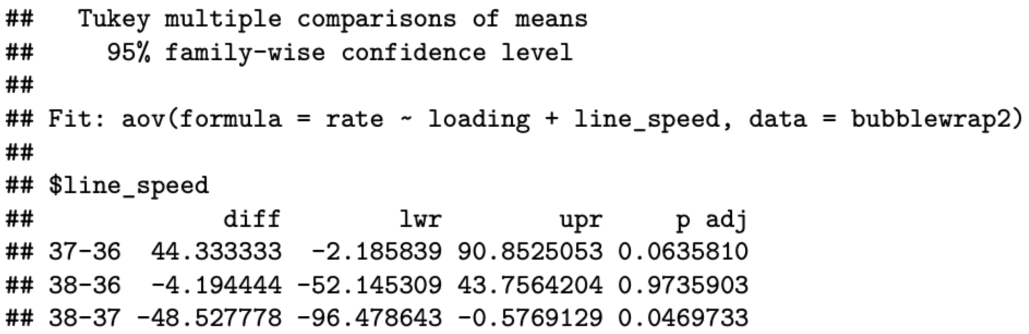


Figure 26: Tukey Interval of the 'line_speed' factor



```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = rate ~ loading + line_speed, data = bubblewrap2)
##
## $loading
##          diff       lwr       upr     p adj
## 2-0 -26.94444 -74.89531  21.00642 0.3508319
## 4-0  41.94444  -6.00642  89.89531 0.0935753
## 4-2  68.88889  22.36972 115.40806 0.0033747
```

Figure 27: Tukey Interval Numerical and P-value Outputs for 'loading' factor



```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = rate ~ loading + line_speed, data = bubblewrap2)
##
## $line_speed
##            diff        lwr        upr     p adj
## 37-36  44.333333  -2.185839 90.8525053 0.0635810
## 38-36  -4.194444 -52.145309 43.7564204 0.9735903
## 38-37 -48.527778 -96.478643 -0.5769129 0.0469733
```

Figure 28: Tukey Interval Numerical and P-value Outputs for 'line_speed' factor