Final Project: Executive Summary

Submitted by – Debapratim Ghosh (dg19@illinois.edu) and Shantanu Solanki (solanki7@illinois.edu)

Problem Description

During a production process, a lot of machineries are involved in the production of items. Some machines have various parameters of operation and under different operating conditions, have different production rate.

It is of paramount importance that the production rate is maximum among the operating conditions chosen so that the items are produced in a timely and efficient manner.

In this problem, a Company XX is a manufacturer of several types of protective packaging, including bubble wrap sold in both retail and bulk. The objective of this project is to determine the best operating conditions for the bubble wrap lines to increase production capacity.

Variables such as extrusion rate, temperature, line speed, and percent loading of additives were key factors that were considered in the study. After preliminary experiments, the engineers decided that **line speed**, and **percent loading of additives** were the most significant factors and thus a complete randomized (also called factorial) design was implemented with these two factors:

Factors	Levels		
Line Speed (m/mm)	36	37	38
Loading of Additives	0	2	4

The response variable was the production rate measured in **lbs/hr**. The experiment was replicated 3 times and the randomization order for each replication was also recorded. Our goal is to find the optimal combination of line speed and percent load of additives that results in the highest production rate in **two cases**:

- **(i) When all observations are considered valid (Balanced Anova)**
- **(ii) When some observations are removed (Unbalanced Anova)**

Data Dictionary

replication: The indicator for the serial number of trial, values = 1, 2, 3.
run_order: randomization order for each replication, values = 1, 2, …, 9.
line_speed: as explained above, with three categorical values = 36, 37, 38.

loading: percentage of loading of additives as explained above, values = 0, 2, 4.
rate: our response variable, i.e. rate of production, continuous variable in lbs/hr.

Objectives and Methods

Our objective is to ascertain whether we can estimate the optimal combination of line speed and percent load of additives that results in the highest production rate for the two scenarios mentioned above solely on the basis of the data we have and if yes, find out the said optimal combination.

We would be using ANOVA models to reach our conclusions. The first part involves a balanced ANOVA model while we explore the unbalanced ANOVA in the second part.

As usual, we start with the exploratory data analysis to ascertain what types of data we are dealing with. Since, we have a factor analysis, we convert the said variables into factors and observe the interaction between the factors through the Interactions Plot. If we find visible interaction, we fit a two-way ANOVA model with the interaction term and check for the significance of different terms.

As with any linear model, we check for unusual observations, equal spread of the error terms and normality in the distribution of errors. Although we are not bothered about the linearity of the relationship.

For unusual observations, we would be using the Cook's distance and if required, fit alternative models after setting aside those values.

For checking normality, we would use Shapiro-Wilk test (since we have a small sample). If we find the assumption not holding true, we would be employing Box-Cox transformation to get the error terms approximately normal. However, with such a small sample, we expect normality assumption to not hold true.

We can use both Breusch-Pagan test and the Levene's test for ascertaining if the errors have a constant variance. In the latter, we would regress the residuals obtained from our model on the predictors to check for the significance of the variables.

For the second part, we would be removing the invalid observations and refitting the model in the same way and performing model diagnostics. However, this time we can't check for the significance of the predictor variables using a normal ANOVA test as the residuals will not be additive (orthogonal). We would be using a Type III ANOVA in this case.

Summary of Findings

Part-I:

1. The interactions plot (Fig. 1) shows that the interaction is present. Therefore, we choose the model with the interaction term. However, upon checking the ANOVA summary of the model, the interaction term is found to be insignificant.

2. We move forward with the additive model and check for unusual observations (Fig. 2), constant variance assumption (Fig. 3) and normality of errors assumption (Fig. 4). We don't find any unusual observations and the constant variance assumption also holds good. However, the normality assumption fails on the Shapiro-Wilk Test.
3. To remedy the failure of the normality assumption, we perform a transformation of the response and re-perform the tests. We perform the diagnostics for the error variance (Fig. 5) and normality (Fig. 6) and observe significant improvement in the plots. The model also 'passes' the tests for constant variance and normality of error terms.
4. Having finalized the model, we finally plot the Tukey Confidence intervals for line speed (Fig. 7) and loading (Fig. 8) to find out that there is significant difference in the response based on different levels of both the factors.

Part-II:

1. The interactions plot (Fig. 9) shows that the interaction is present. Therefore, we choose the model with the interaction term. However, upon checking the type III ANOVA summary of the model, the interaction term is found to be insignificant. We remove it and refit the additive model.
2. We check the significance of each factor against the additive model with both the factors and find that both line_speed and loading are statistically significant in explaining the variation of production rate. So we decide to keep the additive model .
3. We then perform model diagnostics on the additive model and find that there are no unusual observations (Fig. 10) and the constant variance assumption is met (Fig 11), but the normality assumption is violated (Fig 12). Since the sample size is small, this was expected.
4. In order to check the statistical significance of our factors (without any assumption of an underlying distribution), we perform permutation tests on each of our factors. We find that only **loading percentage** factor turns out to be statistically significant. However, since the sample size is small, we the non-normality of residuals was expected. Therefore, to avoid loss of information, we keep both the factors and use the additive model to perform pairwise comparison. If line speed is truly insignificant, then all the pairwise comparisons of the factor levels wrt to production rate would contain 0 in their confidence intervals.
5. Finally, we perform Tukey Pairwise comparisons (Fig.13 and Fig.14) on each of our factors to find the difference in response due to each of the factors. We find that loading percentage of 4% is statistically significant in having the highest production rate while none of the line speeds have any significant effect on production rate.

Conclusion:

Part-I:

We conclude that the combination of line speed of 37 m/mm and loading 4% has significantly better response than any other combination as per the available data. Therefore, we recommend the above combination.

Part-II:

In case of unbalanced ANOVA, we removed a few datapoints before performing our analysis. The optimal conditions for maximising production rate comes out to be a loading percentage of 4% . The different line speeds to operate the production line turns out to be statistically similar to each other and therefore any line speed can be chosen under 4% loading percentage to achieve the highest production rate.

Appendix :

Part I

Figure 1 : Interaction Plot between Line Speed and Loading

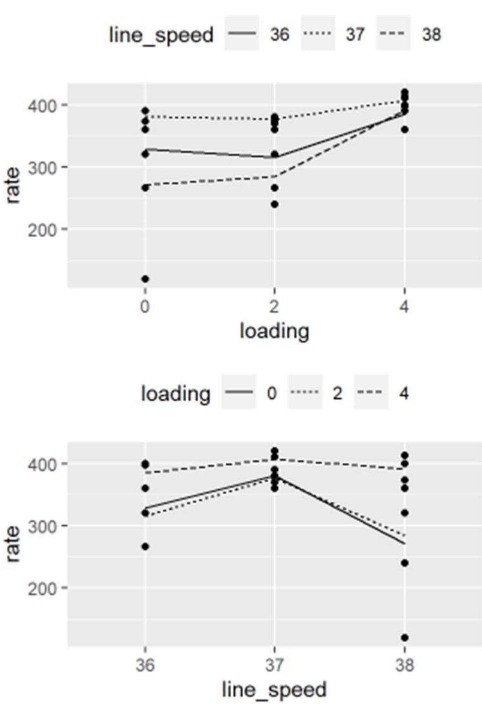


Figure 2 : Presence of Unusual Observations through Cook's distance

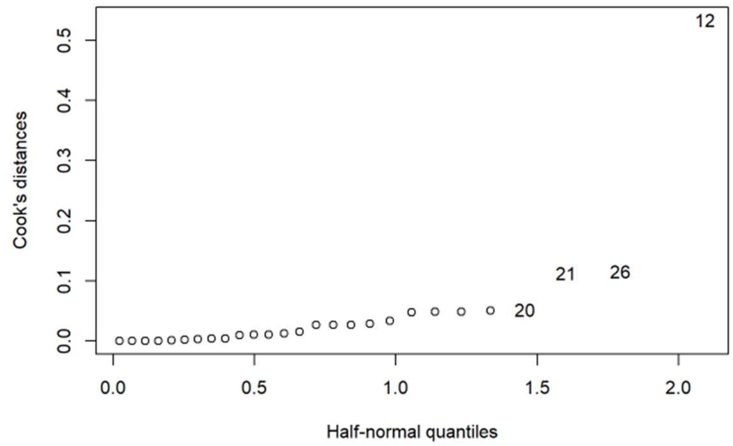


Figure 3 : Constant Variance Assumption Check for Untransformed Model

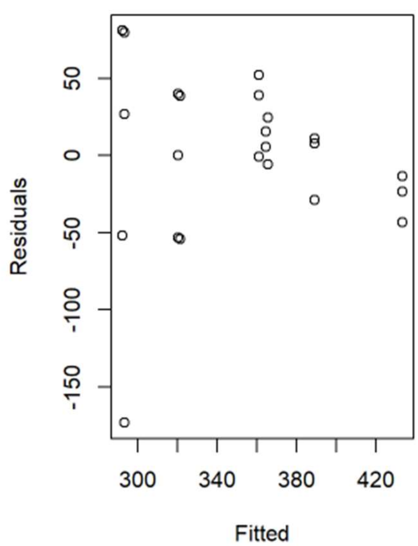


Figure 4 : Normality Assumption Check for Untransformed Model

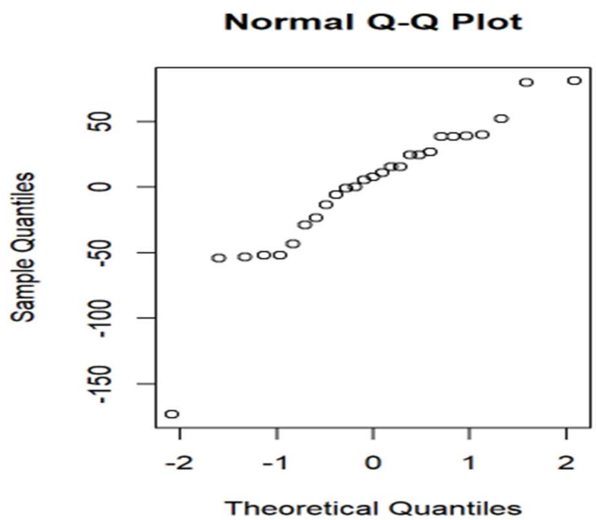


Figure 5 : Constant Variance Assumption Check for Transformed Model

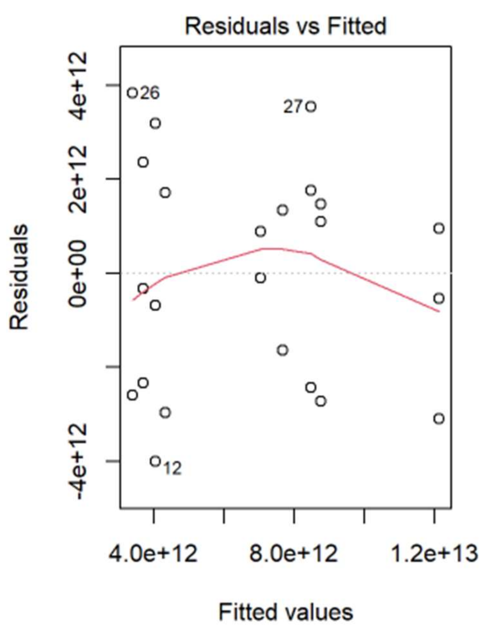


Figure 6 : Normality Assumption Check for Transformed Model

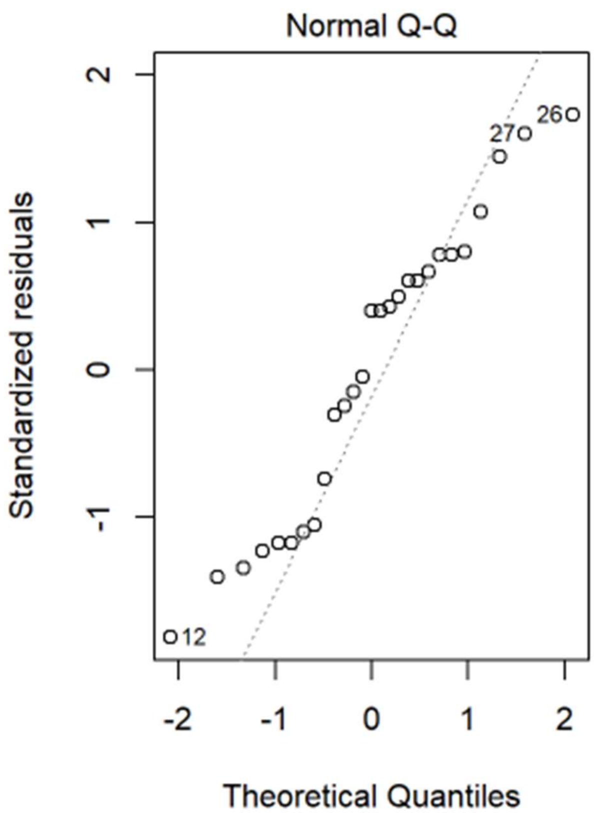


Figure 7 : Tukey Confidence Intervals for Line speed

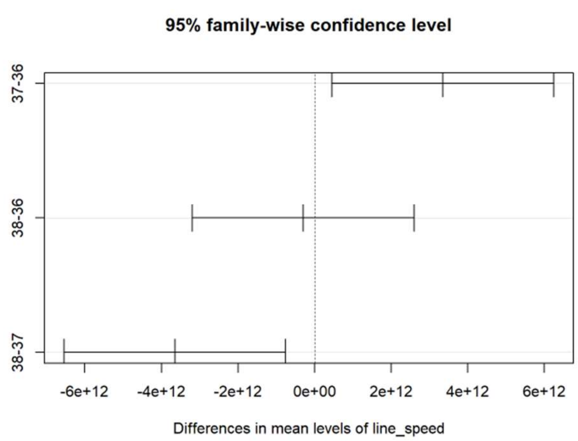
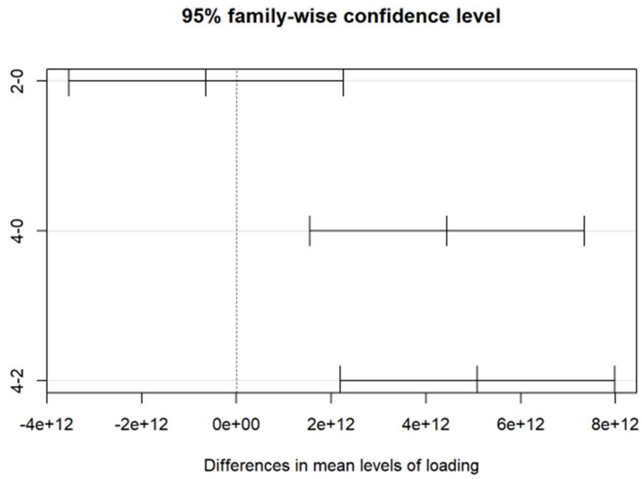


Figure 8 : Tukey Confidence Intervals for Loading Percentage



Part II

Figure 9 : Unbalanced ANOVA- Interaction Plot between Line Speed and Loading

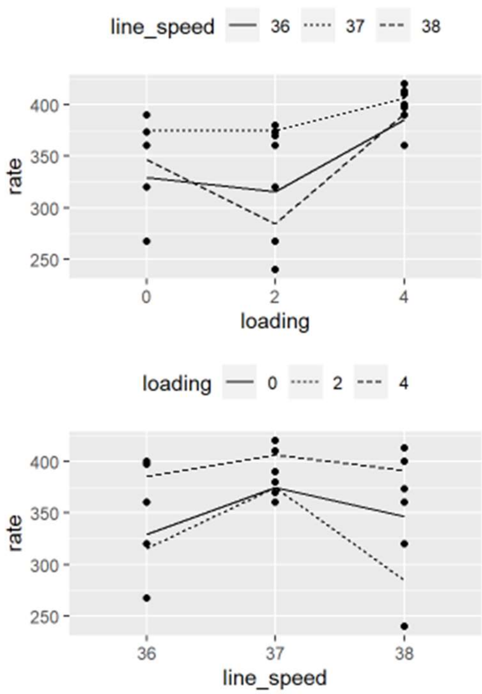


Figure 10 : Unusual Observations Check for Unbalanced ANOVA Model

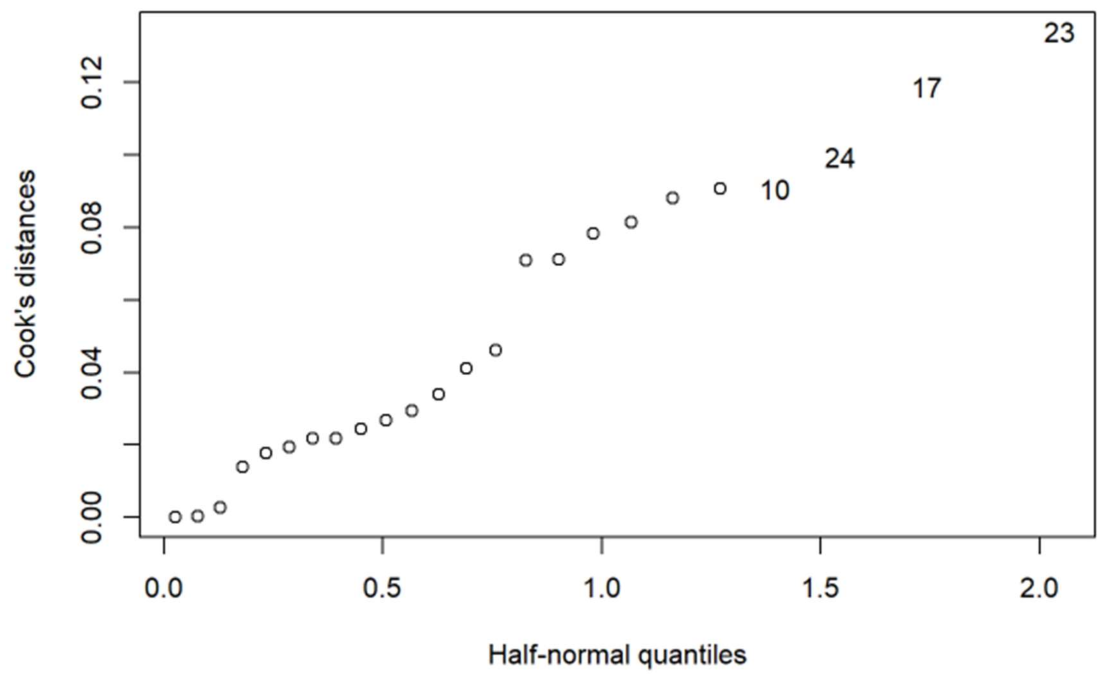


Figure 11: Constant Variance Assumption for Unbalanced ANOVA Model

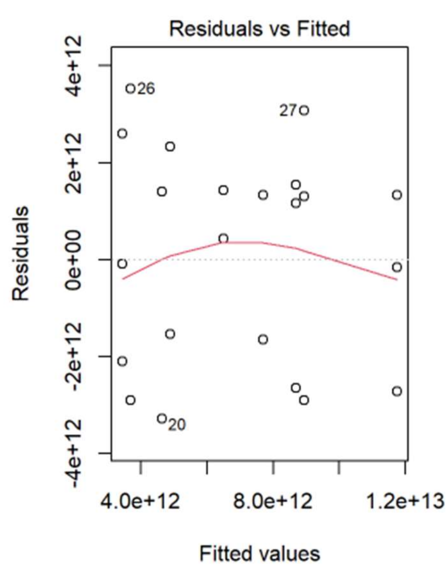


Figure 12: Normality Test for Unbalanced ANOVA Model

```
shapiro.test(model2.additive$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model2.additive$residuals
## W = 0.9016, p-value = 0.02328
```

Figure 13: Tukey Pairwise Comparison for Loading Percentage (Unbalanced ANOVA Model)

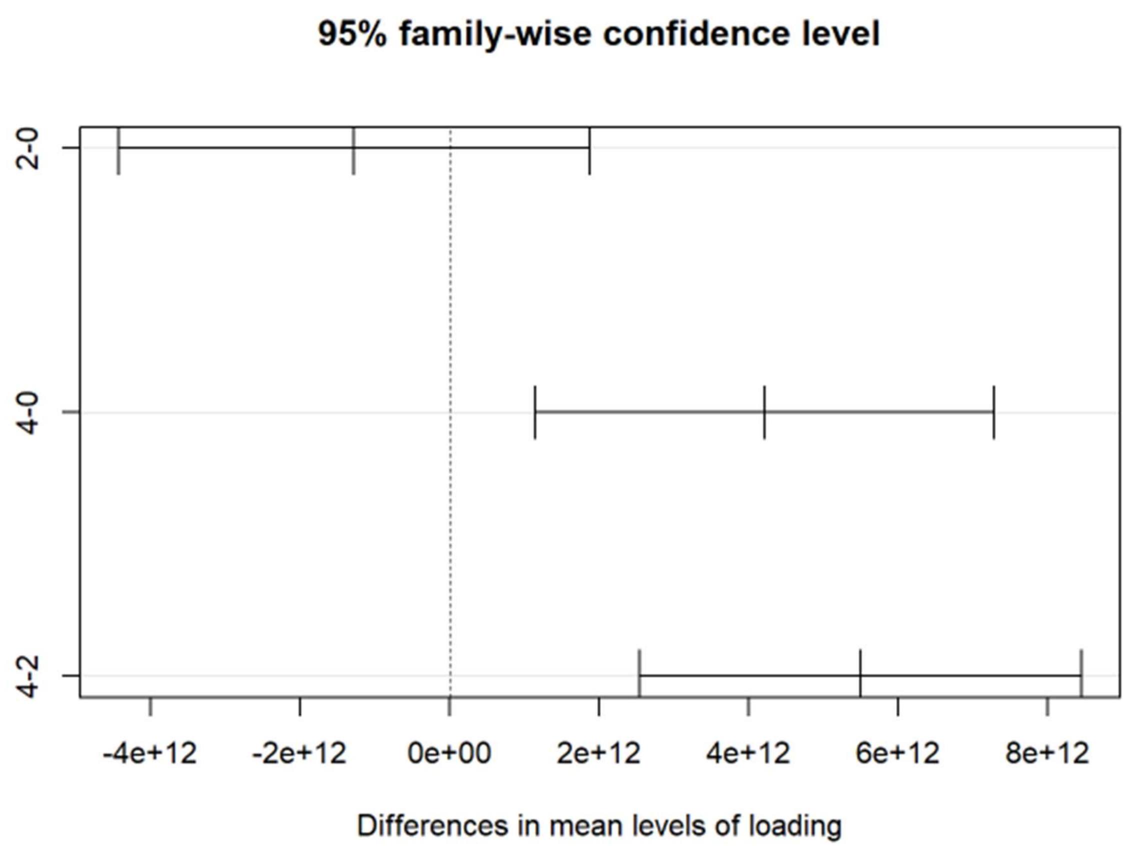


Figure 14: Tukey Pairwise Comparison for Line Speed (Unbalanced ANOVA Model)