ANALYSIS, MODELLING AND REPORT FOR TASSPAPERMILL

by ANALYTICS7

The TassPaperMill TPM anticipates a slowdown in the next five years due to changes in customer preference. TPM hired ANALYTICS7 to help them build a solid customer base by conducting a customer survey, building models and forecasting.

The analysis is based on 200 records of customer data. Forecasting is based on TPM’s last 10 years’ turnover data. We provide a model to estimate order quantity, a model to predict the likelihood a customer to sign a contract, a model to forecast the next 4 quarters

We divide our cases into sections, each explaining the technique used and displaying the process and interpretation of the results.

In the conclusion section, we summarise our findings and give recommendations.

# 1. Overview of order quantity and contract

Order Quantity

Dependent variable Order Quantity is numerical continuous data in tonnes. The Descriptive analysis details can be found in Appendix 1, Table 1.1

On average customers order 7.67 tonnes of paper from TPM, with a minimum order of 4.30 tonnes and a maximum order of 9.9 tonnes.

The distribution plot in Appendix 1, figure 1.1 shows the order quantity is normally distributed. skewing to the left (skewness -0.21).

There are no outliers in Order quantity (Appendix 1, figure 1.2)

Contract

Contract is a categorical binary variable. The value of Contract is 1 if a customer has a contract with TMP, otherwise the value is 0.

From 200 records, there are almost as many customers with contracts (101) and those without one (99).

|  |  |
| --- | --- |
|  |  |

# 2. Modelling Order Quantity

### 2.1. Influencing factors on Order quantity

We apply the multiple regression backward elimination method. We put all relevant factors then eliminate them one at a time until all factors are significant (p-value < 0.05). We then obtain the final model.

The dependent variable is **Order Quantity**.

To decide which variables are influencing factors, we visualise each independent variable against Order Quantity. The scatter plots in Appendix 2 shows that all independent variables have linear relationships with Order Quantity.

In addition to having linear relationships, we measure the strength and direction of the relationship using a correlation table (Appendix 1, Table 2.1)

According to the general rule, the Correlation coefficient values < 0.3 is considered week; 0.3-0.7 moderate and value >0.7 strong

The correlation table demonstrates:

1. loyalty, dist\_channel, quality, prdct\_line, brand\_image, order\_fullfilment, shipping speed, shipping\_cost and contract have **positive** **medium** strength linear relationship with quantity
2. sm\_presence, advertising have **positive** **weak** linear relationship with quantity
3. cust\_type, region, comp\_price, flex\_price have **negative** **weak** linear relationship with quantity

Now we look at the relationships between independents. we note that there are possible multicollinearity between the following variables:

* brand\_image and social media presence (0.79)
* order fulfilment and shipping speed (0.77)
* shipping cost and shipping speed (0.84)

We remove one of the offending variables, based on its correlation coefficient to Order Quantity. Brand\_image has 0.34 and Social Media Presence has 0.24, therefore we exclude Social Media Presence

However, we also exclude shipping speed because it is strongly correlated with both shipping cost and order fulfilment.

After eliminations, the possible factors that influence order quantity are **Loyalty, Cust\_Type, Region, Dist\_Channel, Quality, Advert, Prdct\_Line, Brand\_Image, Comp\_Pricing, Order\_Fulfillment, Flex\_Price, Shipping\_Cost, Contract**

### 2.2. a model to estimate Order\_Qty

We run regression with all the possible factors. The result is the first model, which contains R2 = 0.56. However, some of the p-values of the independent variables are insignificant (> 5%), we exclude the most insignificant independent variables and re-run the regression. We repeat the regressions until all the independent variables are significant (p-value < 0.05) to get the final model.

|  |  |
| --- | --- |
| **Iteration** | **Highest p-value** |
| 1 | Cust\_Type |
| 2 | Prdct\_Line |
| 3 | Region |
| 4 | Dist\_Channel |
| 5 | Advert |
| 6 | Flex\_Price |
| 7 | Comp\_Pricing |
| 8 | Order\_Fulfillment |

**The final model**

Hypothesis

H0: βj=0 no linear relationship

H1: βj≠0 linear relationship exists between x and y

From the anova table (Appendix 1, Table 2.2), the significance F is 0 (< 0.05) therefore Reject H0. Overall the model has predictive power.

Based on the regression coefficient (Appendix 1, Table 2.3*)*, the interpretation of the estimated multiple regression model is as follows:

**Y = 3.53 + 0.05 x1 + 0.21 x2 + 0.13 x3 + 0.22 x4 + 0.34 x5**

The interpretation based on the assumption that other variables held constant (=0)

|  |  |
| --- | --- |
| Y | Predicted order quantity |
| b0 | other variables = 0, the average order quantity is expected to be 3.53 tonnes |
| b1 (Loyalty) | other variables = 0, for each year of increase in loyalty, on average the order quantity increases by 0.05 tonnes. |
| b2 (Quality) | other variables = 0, for each increase in scale of Quality, on average the order quantity increases by 0.21 tonnes. |
| b3 (brand\_image) | other variables = 0, for each increase in scale of Brand image, on average the order quantity increases by 0.13 tonnes. |
| b4 (Shipping\_Cost) | other variables = 0, for each increase in scale of Shipping Cost (cost more affordable), on average the order quantity increases by 0.22 tonnes. |
| b5 (Contract) | other variables = 0, when a customer changes from no contract to with contract, on average the order quantity increases by 0.34 tonnes. |

R2 in regression statistic is 0.54 (Appendix 1, Table 2.4), which means 54% variation in order quantity can be explained by loyalty, contract, quality, brand\_image and shipping\_cost

The remaining 46% of variation in order quantity would be explained by other factors, not included in the model.

The R2 (0.54) for the final model is slightly less than the first model (R2 = 0.56). Adjusted R factors in the number of independents. In the first model, we reduce 13 independent variables down to 5 in the final model. However, the adjusted R stays at 0.53, which is quite good.

Assumptions:

**Linearity**: Scatter plots (Appendix 2) show all independent variables have linear relationships with dependent variable

**Normality**: if the plot is almost a straight line, we assume the normality assumption is validated; However, we cannot conclude that our model follows this assumption. (Appendix 3)

**Homoscedasticity**: In all our residual plots, points are randomly dispersed, there is no evidence pattern, so the error is independent (Appendix 3)

**Independence**: We assume that observations are independent from each other.

**Outliers**: Using a conservative cutoff of 2, we have 8 potential outliers. (Appendix 1, Table 2.5)

# 3. Interaction Effects brand image on quality and quantity

We establish there is relationship between quality and quantity. Adding brand image to the relationship, we examine whether the brand image has a special effect on the relationship quality and order quantity.

The result after the regression is as follow:

In the Anova table (Appendix 1, Table 3.1), the model is statistically significant with Significance F = 0 (< 0.05). Overall the model has some predictive power.

Observation from regression coefficient table (Appendix 1, Table 3.2)

* All individual independent variables (including the interaction term) are individually significant at p < .05
* Significance of the interaction term (p-value = 0.04) indicates that brand image interacts with quality in predicting quantity. At 95% confidence, there is sufficient evidence to conclude that the interaction term is statistically significant in the model.

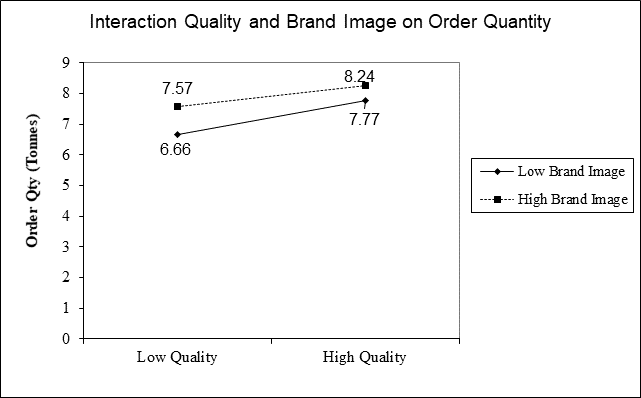


Figure 1.1: Interaction plot

Customers who have low perception of Quality but high perception of Brand Image on average order more papers than customers who have low perception of Quality and negative perception of Brand Image.

Customers who have high perception of Quality and have positive perception of Brand Image on average order more papers than customers who have high perception of quality but negative perception of Brand Image.

Overall, perception of brand image interacts with quality and order quantity, whose relationship is significantly stronger for high brand image than for low brand image

With p-value < 5%, there is sufficient evidence to conclude that the interaction term is statistically significant in the model

# 4. The likelihood of customers signing a contract with TPM.

### 4.1. Regression and Modelling

The outcome is contract. The predictors are quality, Prdct\_Line, brand\_image, flex\_price, comp\_pricing

In the first model regression coefficient (Appendix 1, Table 4.1), the competitive price is insignificant (p-value = 0.77 > 0.05), hence we exclude that variable and run the second regression.

In the second regression, all independent variables are now significant (p-value < 0.05), and we have the final model

**The final model:**

In regression coefficient, all independent variables have p-value 0 (<0.5) (Appendix 1, Table 4.2), so we reject H0 and are confident that there are relationships between each of the independent variables with order quantity.

From the classification table (Appendix 1, Table 4.3) and classification accuracy rate, we determine the practical significance of this model.

* The overall classification accuracy (hit ratio) is **77 %**, meaning 77% of predictors are accurately classified by the model. The remaining 23% could be captured if more relevant predictors had been included in the model.
* Of the 101 customers who signed contracts (observed), 79 were correctly predicted to sign contracts, 22 were mis-classified (predicted not to sign contracts)
* Of the 99 customers who did not sign contracts, 25 were correctly predicted not to sign contracts, 74 were mis-classified (predicted to sign contracts)
* We conclude the practical significance of the logistic model providing an accuracy rate of **77%,** which is greater than PCC hit ratio **50%**
* Our accuracy rate of 77% is greater than standard ratio 63%, It provides evidence for practical significance of the logistic model.

Overall fit & LL Value (Appendix 1, Table 4.4)

* LL is related to deviance. The lower the deviance, the less unexplained variation and the better our model is. Compared to the baseline model, the final model significantly reduced the LL value. (from -138.62 to -97.87).
* The p-value associated with the chi-square is 0, meaning the model has predictive power.
* According to Cox and Snell, R2 = 33% of variation in signing a contract can be explained by our model. The 67% of variation in signing a contract can be explained if we include other factors.
* According to Nagelkerke, R2 = 45% of the variation in the dependent variable can be explained by the model, the other 55% can be explained if we include other factors

ROC Curve (Appendix 1, Figure 4.1)

The model fits the data well, it is explained by the area under the curve (AUC), which is very close to 1.0.

The curve is good distant from diagonal, indicating the model's ability to differentiate between success (sign a contract) and failure (not sign a contract) is not due to chance.

The logistic regression model:

**Y = -16.83 + 0.81 q + 0.61 pl + 0.76 bi + 0.64 fp**

Assuming we hold the other variables constant

|  |  |
| --- | --- |
| q | 1 unit increase in the perception of quality increases the likelihood of signing a contract by 81% |
| pl | 1 unit increase in the perception of product line increases the likelihood of signing a contract by 61% |
| bi | 1 unit increase in the perception of brand image increases the likelihood of signing a contract by 76% |
| fp | 1 unit increase in the perception of flex price increases the likelihood of signing a contract by 64% |

**Predicted probabilities by varying factors**

We predict the probabilities of a customer signing a contract with TPM when

- brand image = 5

- product line = 5

- quality = 1 to 10

- flex price = 5 to 10

From the probability calculation (Appendix 1, Table 4.5), we draw the chart below

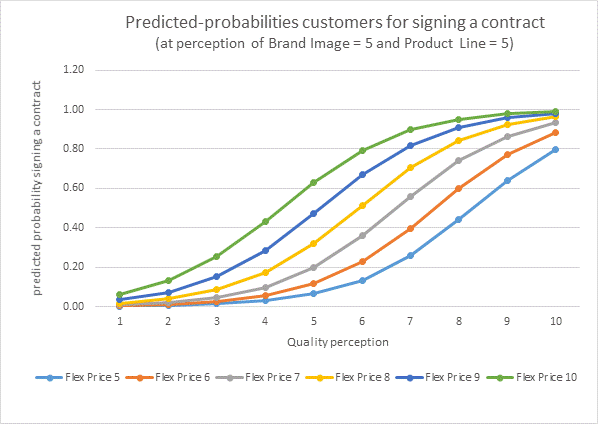


Figure 1.2: Predicted-probabilities

The graph shows that with neutral perception of brand image and product line; the higher the quality perception, the higher probability a customer will sign a contract.

This is also true for brand image perception; the higher the brand image perception, the higher probability a customer will sign a contract.

TPM must improve their customer’s perception on flexible price and quality if they want their customers to sign contracts with them.

# 5. TPM's turnover forecast

TPM wants to forecast the demand for 4 quarters in 2020. Time series analysis helps to forecast the order quantity.

**View the data, trend and pattern**

* Dependent variable is order quantity. Independent variable is time series of 10 years data represented quarterly
* The **trend** seems to be **upward linear**; the demand is increasing every year. The data contains **trend** component
* There is an unusual jump in 2013 then back to the normal pattern in 2014. There was an unusual drop in 2018, which recovered in 2019. It shows the **random** components
* From the data set, we can see the pattern that orders usually at the highest in quarter 2 and at the lowest in quarter 3 within a year. This shows the **seasonal** component (Appendix 1, Table 5.1)

**Visualise the data** (charts can be seen in Appendix 1, Figure 5.1*)*

* Order quantities is the dependent variable and is represented by an orange line in our chart below.
* A blue line represents a trend after the smoothing
* A gray line represents de-seasonalized data
* A yellow line represents forecasting

**Apply the smoothing technique**

Before we can see the real trend, we must smooth the line. We use centered-4 period moving average (MA) because our data is divided per 4 quarters. We calculate 4 MA then apply centered-4 period MA on the data.

The trend still contains seasonal data, we must de-seasonalize it. We divide the observed data with their seasonal index.

**Equation for the forecasting model**

On de-seasonalized trend, we determine the trend as upward linear non-stationary trend.

Our trend model can be represented as:

**y = 17.092x + 1097.4**

which y is Order quantity

x is time period

**Forecast the next 4 quarters**

Based on the trend model above. we enter our next 4 quarters period value to x then we get our non-seasonal forecast. We re-seasonalize the forecast by multiple it by its seasonal index.

After we establish the trend, we predict the future quantity based on our model. Here is our forecast.

For year 2020, TMP will have orders of 1726.97 tonnes in quarter 1, 2126.73 tonnes in quarter 2, 1603.75 tonnes in quarter 3 and 1766.73 for the final quarter (appendix 1, table 5.2)

**Accuracy of the model**

We used past data to calculate Mean Absolute Percent Error

MAPE result is 0%; it shows that the model is accurate. See the visualization in Appendix 1, Figure 5.2

# Conclusion

Contract has a positive linear relationship with Order Qty. On average a customer who signed a contract with TMP ordered 0.97 tonnes more than who did not. Therefore, TPM must improve the likelihood for their customers to sign a contract by increasing their customer perceptions on brand image, flexible price, quality and product line.

According to our multiple regression models, customers order more products if they have higher perception of quality, brand image and shipping cost, have contracts with TPM, have high loyalty (years with TPM).

We suggest that TPM investigate further to build a better model by considering potential outliers and including other factors to improve the R2

We establish that brand image interacts with quality and order quantity, which the relationship is significantly stronger for high brand image than for low brand image

Based on our forecasting model, the demand for TPM product will be higher in the next 4 quarters in 2020. As the seasonal trend from previous data, the first quarter will slightly increase, jump up in the second quarter then drop in the third quarter and back to normal in the fourth quarter.

# Appendices

### **Appendix 1**

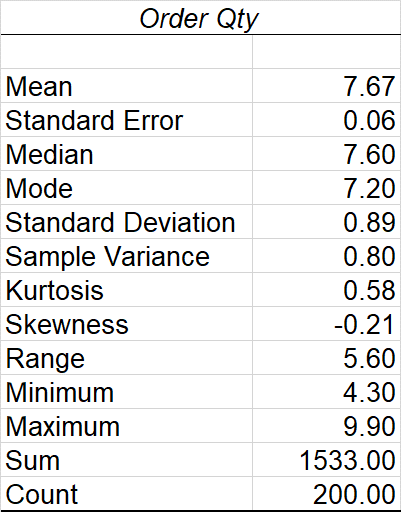


Table 1.1 Order Quantity Descriptive Statistic

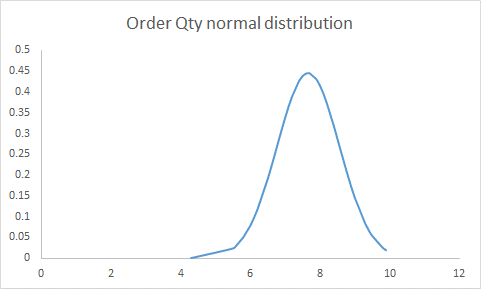


Figure 1.1: Order Quantity Normal Distribution Plot

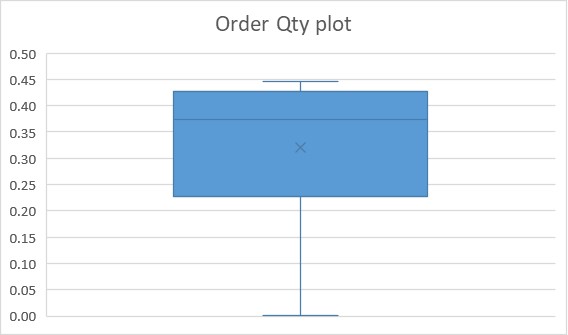


Figure 1.2: Order Qty boxplot



Table 2.1: Correlation table

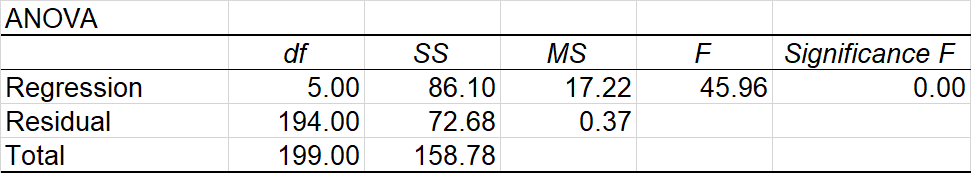
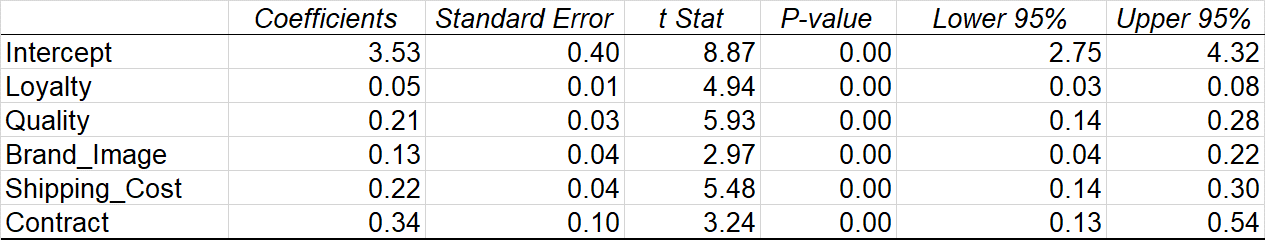


Table 2.2: Anova table – final model

Table 2.3: Regression Coefficients table – final model

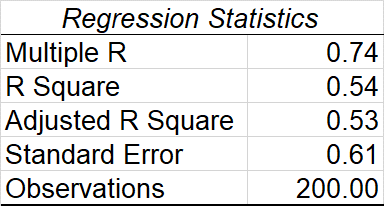


Table 2.4: Regression Statistic – final model

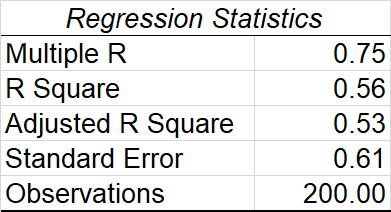


Table 2.5: Regression Statistic – first model

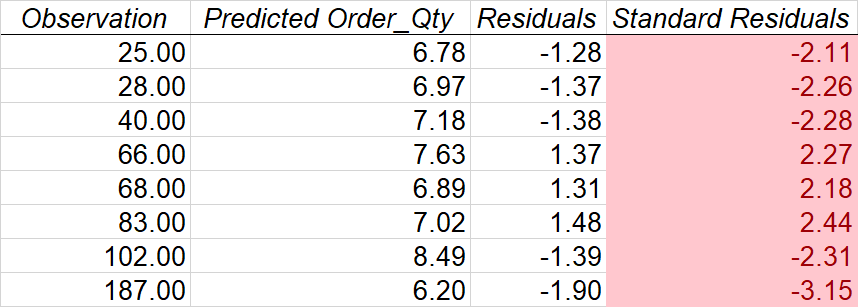


Table 2.6: Potential outliers

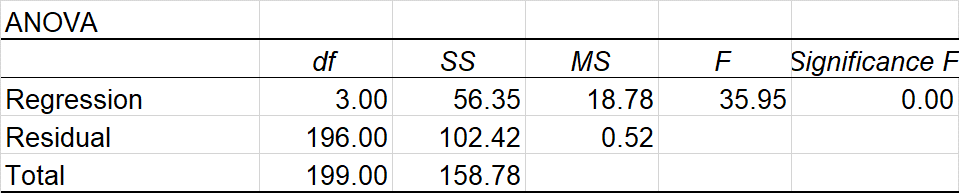


Table 3.1: Anova table

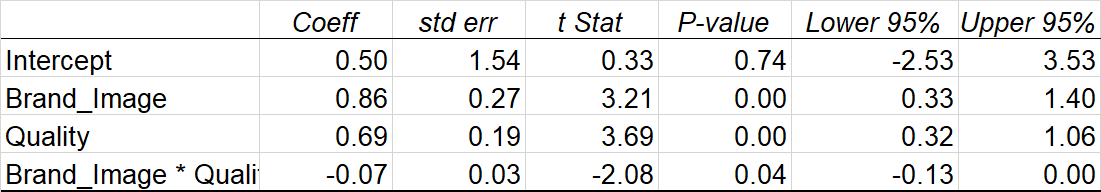


Table 3.2: Regression Coefficient

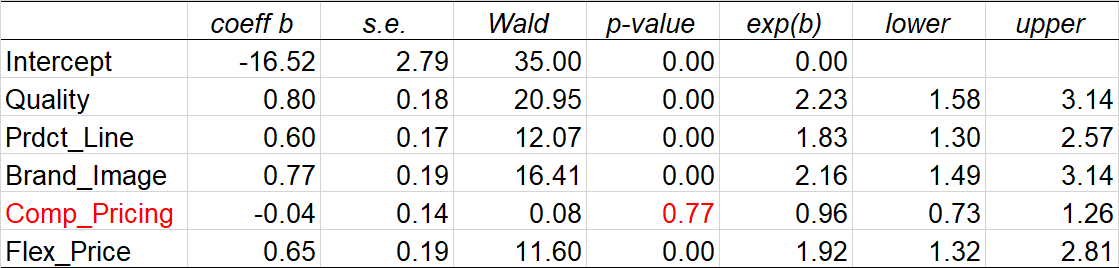


Table 4.1: Regression Coefficient of first model

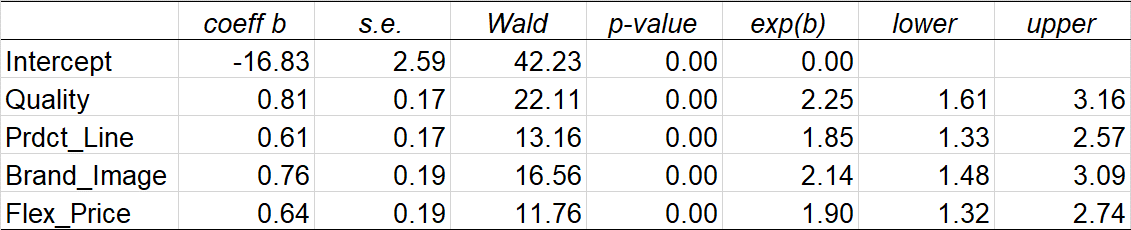


Table 4.2: Regression Coefficient of final model

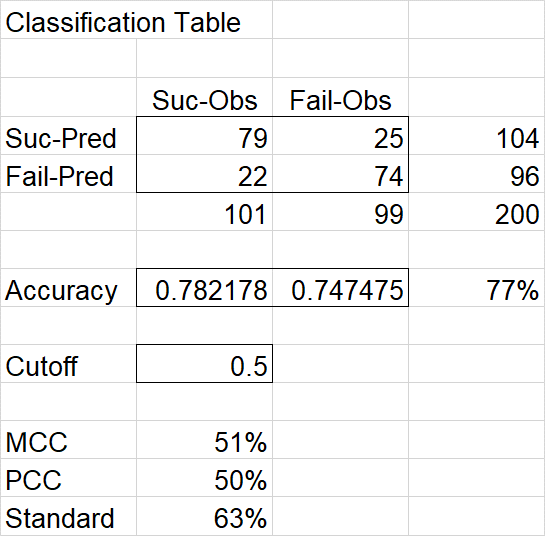
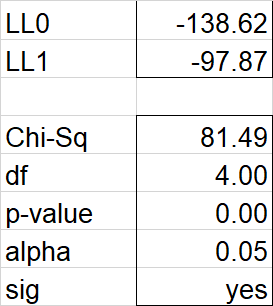


Table 4.3: Classification table



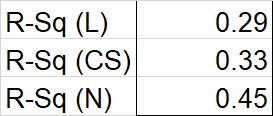


Table 4.4: LL value, Overall fit, Pseudo R2

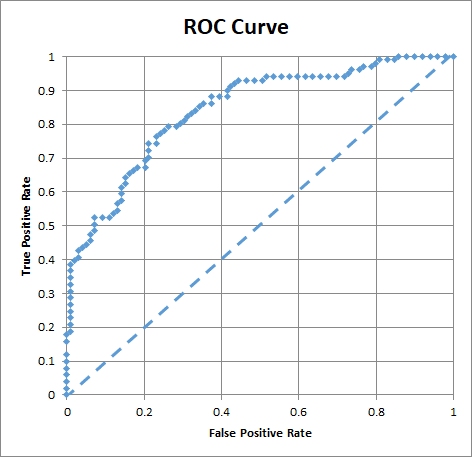


Figure 4.1: ROC Curve

Y = -16.83 + 0.81 q + 0.61 pl + 0.76 bi + 0.64 fp

bi = Brand image = 5 pl = Product line = 5

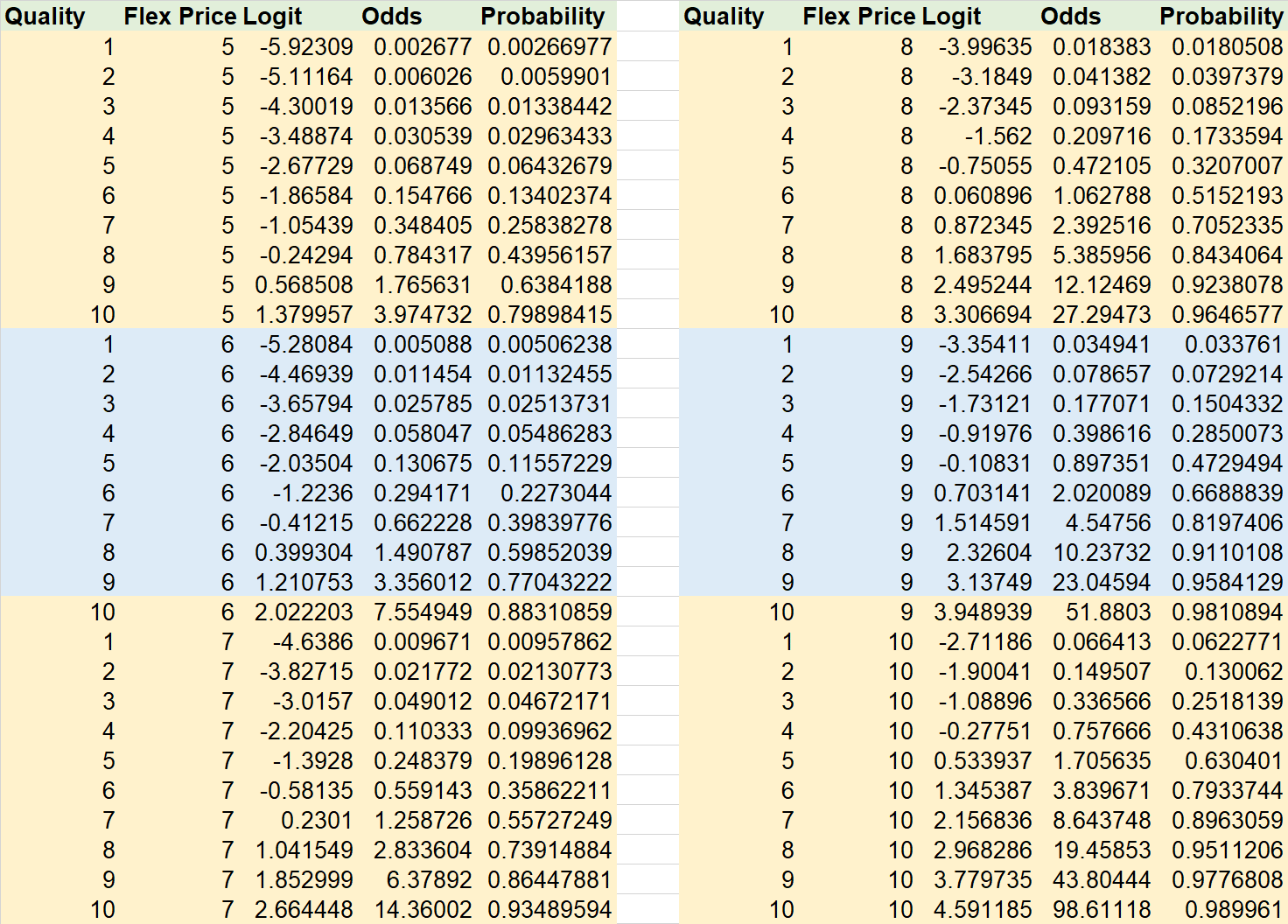


Table 4.5: Probabilities table

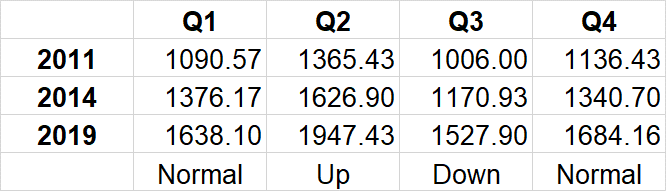


Table 5.1: Sample Order Qty comparation

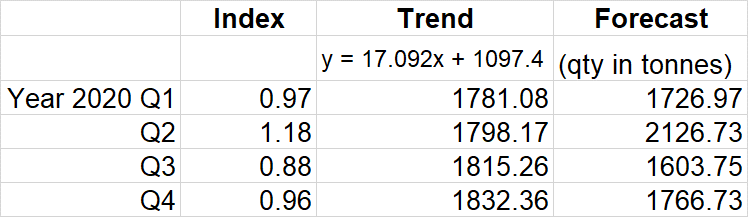


Table 5.2: Forecast table

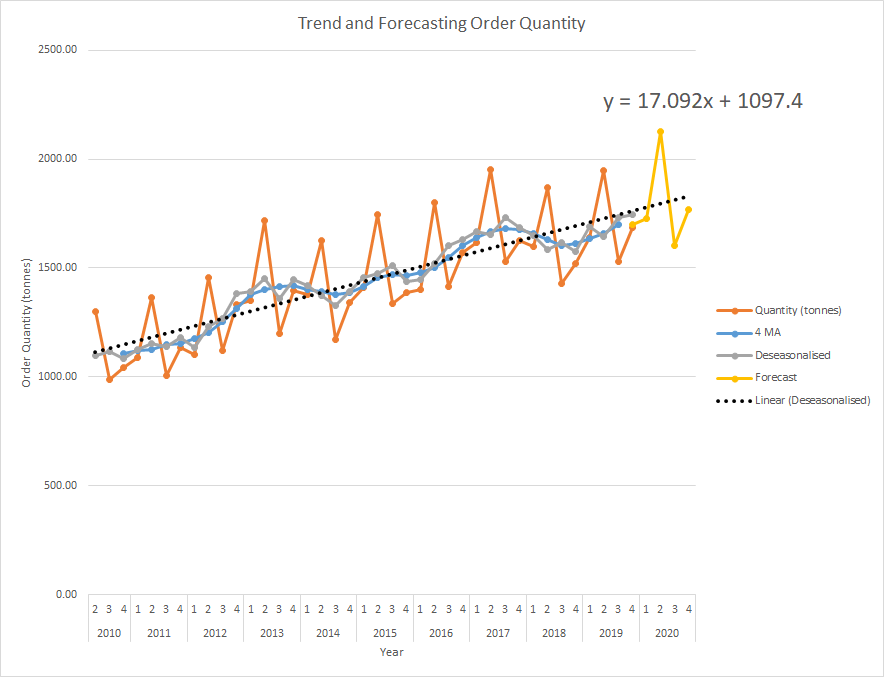


Figure 5.1: Trend and forecasting order quantity

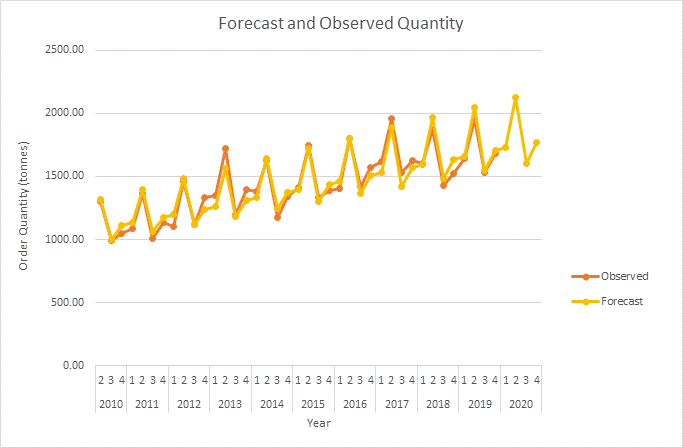
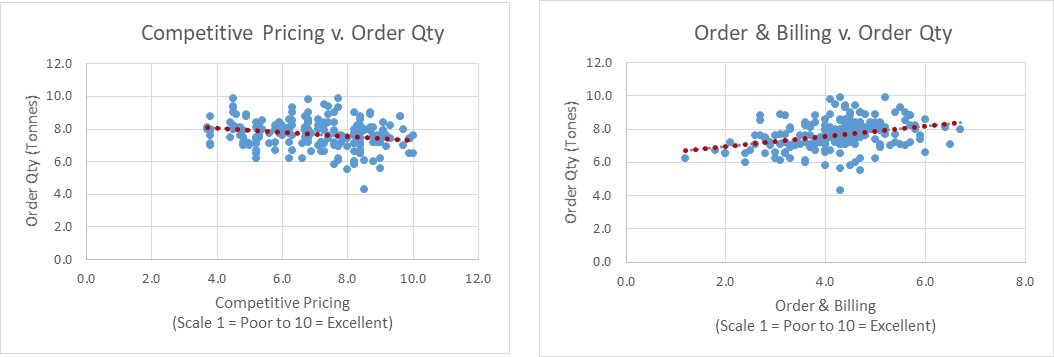
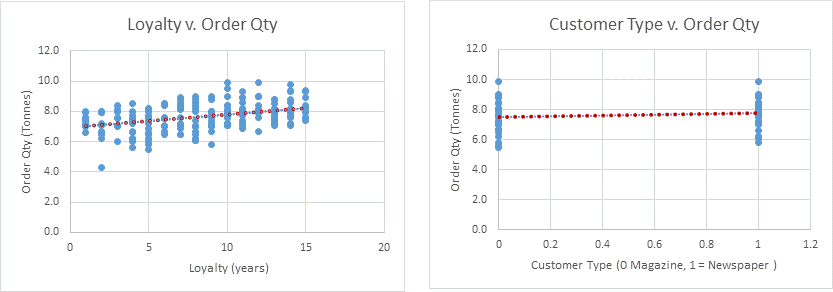
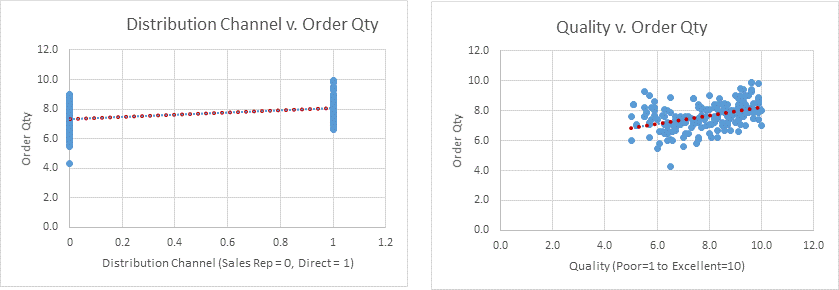


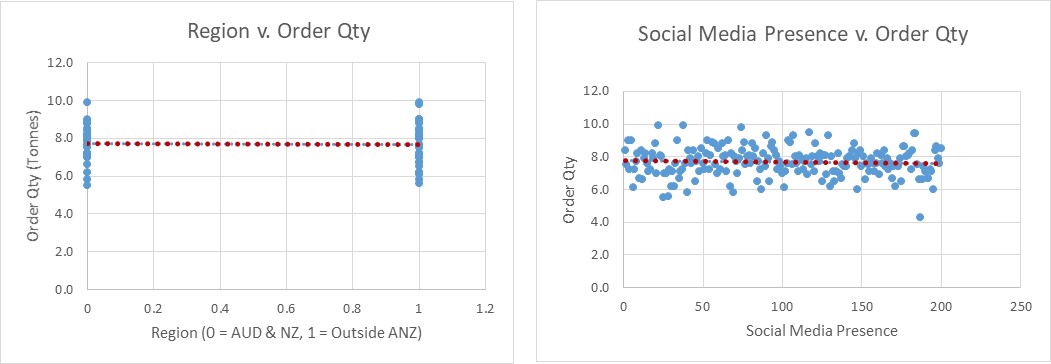
Figure 5.2: Forecast and Observed Order quantity

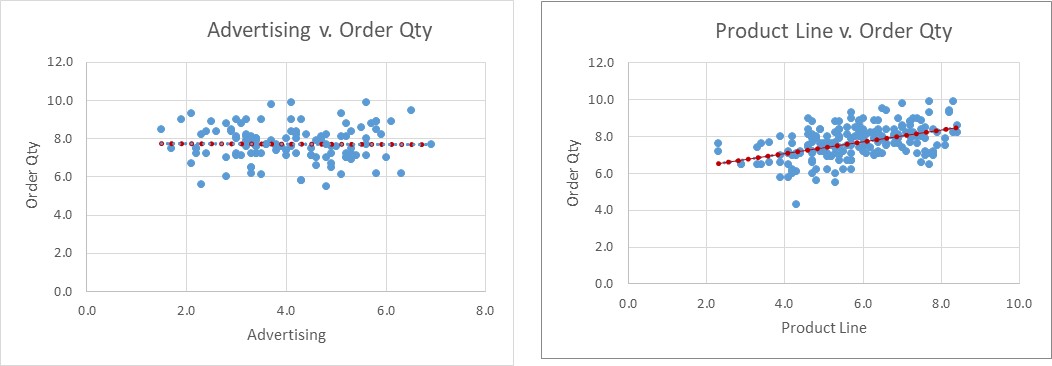
### **Appendix 2**

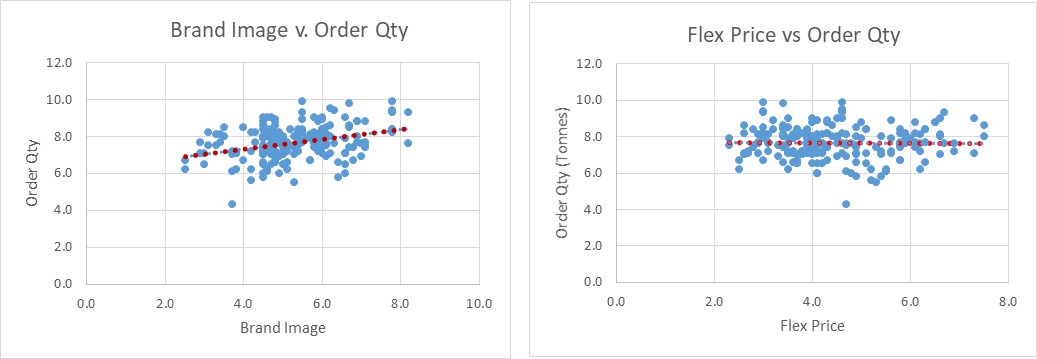
****

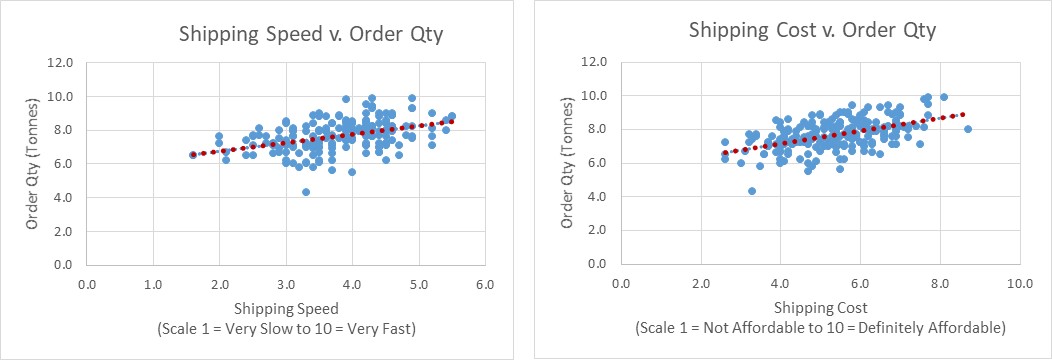
****

****

****

****

****

****

****

### **Appendix 3**

