University of Melbourne

# Project Report

MAST90072 – Data and Decision Making

Mengfei Hu – 719434

2016-5

mergfeih@student.unimelb.edu.au

## 1. Context of the problem, including hypotheses.

According to recent business research, the online retailers tend to use competition-based pricing strategy which means to set their prices based on competitors' prices (Marchall, et al., 2014). An Ebay (ebay.com.au) seller "Big Bang Sale" is trying to know if the competition-based strategy could work for them and they want to start with an observation study on their best sales product category "photo frame". So the question this observational study would solve is that *is there a relationship between item sold and price difference or shipping fee difference from competitor of photo frames in bigbangsale*, and if so, *to what extent they are related and what is the statistic model for that*.

There are two stages in this analysis, first stage is to determine the strength of relationship between price difference/shipping fee difference and the number of item sold; second stage is to fit a statistic model for the number of item sold.

Hypotheses including:

- The null hypotheses: the price difference/shipping fee difference have no effect on sales data
- The alternative hypotheses: smaller price difference/smaller shipping fee difference from average price/shipping fee of competitors could lead to better sales data

## 2. The statistical model (or models)

We try to fit a liner regression model for sales data (item sold) so we have tried following models in this project:

$$y = ax1+bx2 + e$$

where y = Item sold, x1 = price difference, x2 = shipping fee difference and e = error.

## 3. Experimental design

(a) Type of experiment.

This is an observational study that comes from real data of industry case. Terapeak platform (www.tarapeak.com) have provide data source on ebay platform for sellers to do product research and competitor research. All data provided are calculated as average with key words searching. The key word product name "photo frame" and seller ID "bigbangsale" are being used for data collection.

(b) Sampling procedure.

The photo frames being sold on ebay.com.au should be the samples of this study. With intention of focusing on one particular seller, data collection tool have gathered average data including price and shipping fee of all photo frames being sold in "bigbangsale" and among all competitors on ebay.com.au.

So random sampling is replaced by full observation thanks to the ecommerce data collection tool. Although the sample size, which is the number of photo frames being sold, is different in different sellers, the calculated average price could be compared.

As time range could be set in terapeak, data from April,2015 to April,2016 monthly could been collected. So there are 12 rows in total. Pay attention to that all prices and shipping fee shown are already calculated by terapeak on average level using time frame, product key words and seller ID as input.

(c) Data collected.

| Competitor Price | Competitor Shipping Fee | Bigbangsale Price | Bigbangsale Shipping Fee | Bigbangsale Item Sold | Month |
|---|---|---|---|---|---|
| 16.66 | 11.08 | 25.39 | 15.61 | 133 | 2015/5 |
| 17.03 | 10.65 | 23.16 | 14.42 | 197 | 2015/6 |
| 17.79 | 10.64 | 22.27 | 13.86 | 226 | 2015/7 |
| 17.2 | 10.41 | 23.4 | 13.24 | 206 | 2015/8 |
| 18.46 | 11.06 | 22.11 | 14.32 | 213 | 2015/9 |
| 17.26 | 11.21 | 21.27 | 14.57 | 212 | 2015/10 |
| 18.41 | 10.75 | 21.31 | 12.07 | 290 | 2015/11 |
| 21.12 | 11.61 | 21.57 | 12.47 | 288 | 2015/12 |
| 17.36 | 11.61 | 21.93 | 12.19 | 243 | 2016/1 |
| 17.66 | 11.02 | 22.4 | 12.42 | 134 | 2016/2 |
| 16.31 | 10.7 | 21.76 | 9.8 | 92 | 2016/3 |
| 17.33 | 10.77 | 21.4 | 12.19 | 127 | 2016/4 |

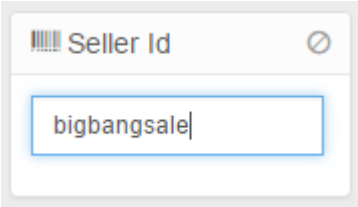
## 4. Data collection and management

Data collection and management tool of ecommerce analytics "Terapeak" is being used. There are main three steps to collect each data in the cell of the table:

Step 1: Constraint product as "photo frame"


Step 2: Constraint time within one-month period


Step 3: Choose to apply filter on seller name of "Bigbangsale" or not: if collect competitor data, do not apply any filter because then it will be average data from all sellers selling "photo frame"; if collect big bang sale data, then apply the filter to get average data only belongs to this seller.




## 5. Exploratory analysis

(a) Compute Data

Based on raw data given as competitor price and bigbangsale price, price difference and shipping fee difference between bigbangsale and competitor should be calculated as

Price difference = bigbangsale price – competitor price

Shipping fee difference = bigbangsale shipping fee – competitor shipping fee

(b) Graph Analysis

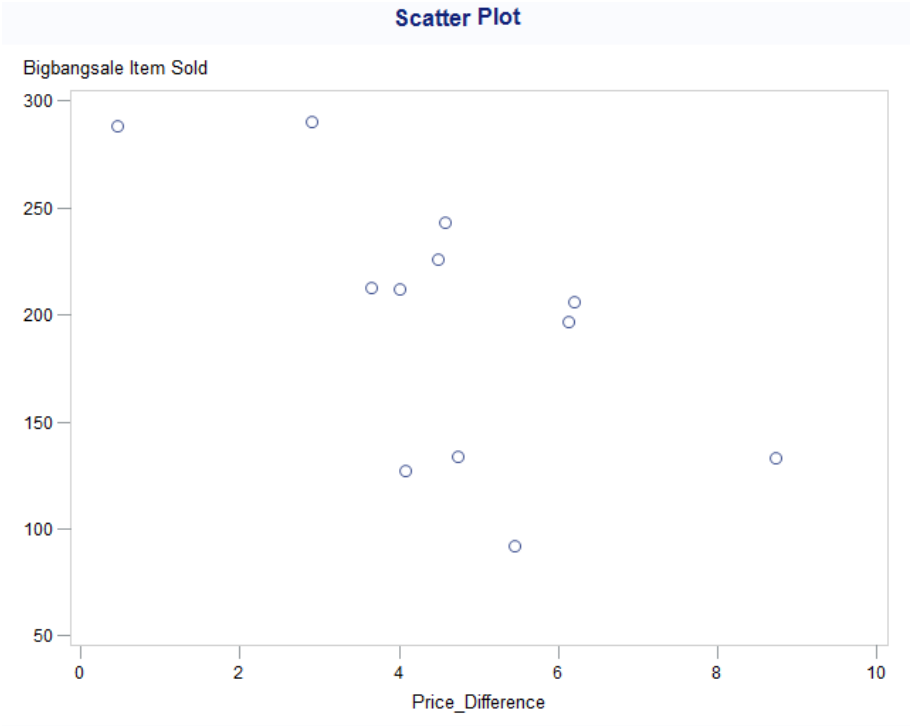



*Figure 1 - Scatter Plot of Price Diff(h) - Item Sold(v)*

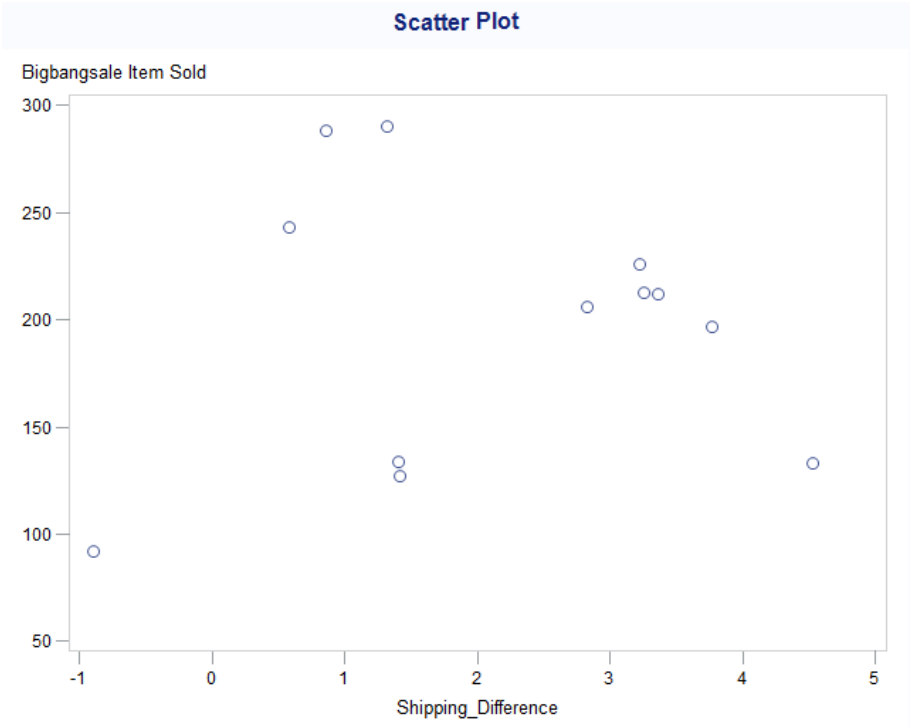



*Figure 2 - Scatter Plot of Shipping Diff(h) - Item Sold(v)*

It appears to be a liner relationship between price difference and item sold while no obvious liner relationship between shipping fee difference and item sold. So we need to check for the strength of liner relationship in the next step.

(c) The strength of liner relationship

**Correlation Analysis**

The CORR Procedure

| 3 Variables: | Shipping Difference | Price Difference | Bigbangsale Item Sold |
|---|---|---|---|

**Simple Statistics**

| Variable | N | Mean | Std Dev | Sum | Minimum | Maximum | Label |
|---|---|---|---|---|---|---|---|
| Shipping Difference | 12 | 2.13750 | 1.58992 | 25.65000 | -0.90000 | 4.53000 | Shipping Difference |
| Price Difference | 12 | 4.61500 | 2.00809 | 55.38000 | 0.45000 | 8.73000 | Price Difference |
| Bigbangsale Item Sold | 12 | 196.75000 | 63.55974 | 2361 | 92.00000 | 290.00000 | |

**Pearson Correlation Coefficients, N = 12**
**Prob > |r| under H0: Rho=0**

| | Shipping Difference | Price Difference | Bigbangsale Item Sold |
|---|---|---|---|
| Shipping Difference | 1.00000 | 0.45566 | 0.06522 |
| | | 0.1366 | 0.8404 |
| Price Difference | 0.45566 | 1.00000 | -0.62359 |
| | 0.1366 | | 0.0303 |
| Bigbangsale Item Sold | 0.06522 | -0.62359 | 1.00000 |
| | 0.8404 | 0.0303 | |

*Figure 3 – Correlation Analysis*

As shown in the result of multivariate analysis, the correlation coefficient of price difference and item sold is -0.62359 and 0.06522 for shipping difference and item sold. So it means there is a strong negative liner relationship between price difference and item sold while the slight liner relationship between shipping fee difference and item sold.

There is an interesting finding on the high possibility of positive liner relationship between shipping difference and price difference which may lead to error of this study. It will be further discussed in final report section later.

## 6. Confirmatory analysis

(a) Liner Regression of price difference/shipping fee difference – item sold

**Linear Regression Results**

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Bigbangsale Item Sold

| | |
|---|---|
| Number of Observations Read | 12 |
| Number of Observations Used | 12 |

**Analysis of Variance**

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 2 | 24126 | 12063 | 5.34 | 0.0295 |
| Error | 9 | 20312 | 2256.92841 | | |
| Corrected Total | 11 | 44438 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 47.50714 | R-Square | 0.5429 |
| Dependent Mean | 196.75000 | Adj R-Sq | 0.4413 |
| Coeff Var | 24.14594 | | |

**Parameter Estimates**

| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| Intercept | Intercept | 1 | 279.51124 | 35.98108 | 7.77 | <.0001 |
| Shipping_Difference | Shipping_Difference | 1 | 17.62615 | 10.12096 | 1.74 | 0.1156 |
| Price_Difference | Price_Difference | 1 | -26.09689 | 8.01335 | -3.26 | 0.0099 |

*Figure 4 –Liner Regression 1*

As shown above, Pr>|t| of shipping difference is 0.1156 > 0.05 while Pr>|t| for price difference is 0.0099 < 0.05. So price difference is significant for item sold while shipping fee is not significant. Then we remove shipping fee difference to redo liner regression.

(b) Liner Regression of price difference – item sold

**Linear Regression Results**

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Bigbangsale Item Sold

| | |
|---|---|
| Number of Observations Read | 12 |
| Number of Observations Used | 12 |

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 1 | 17281 | 17281 | 6.36 | 0.0303 |
| Error | 10 | 27158 | 2715.76033 | | |
| Corrected Total | 11 | 44438 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 52.11296 | R-Square | 0.3889 |
| Dependent Mean | 196.75000 | Adj R-Sq | 0.3278 |
| Coeff Var | 26.48689 | | |

Parameter Estimates

| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | 95% Confidence Limits | |
|---|---|---|---|---|---|---|---|---|
| Intercept | Intercept | 1 | 287.84050 | 39.11922 | 7.36 | <.0001 | 200.67744 | 375.00355 |
| Price Difference | Price Difference | 1 | -19.73792 | 7.82469 | -2.52 | 0.0303 | -37.17241 | -2.30343 |

*Figure 5- Liner Regression 2*

**y = -19.73792x+287.84050**

According to Pr>|t|=0.0303 < 0.05 and R-Square to be 0.3889 which suggest a strong liner relationship. This should be the statistic model we are looking for in the end. And it indicates that the closer to competitor average price could lead to higher number on item sold in this case.

The distribution of residuals and fit plot of this model could also support the conclusion.

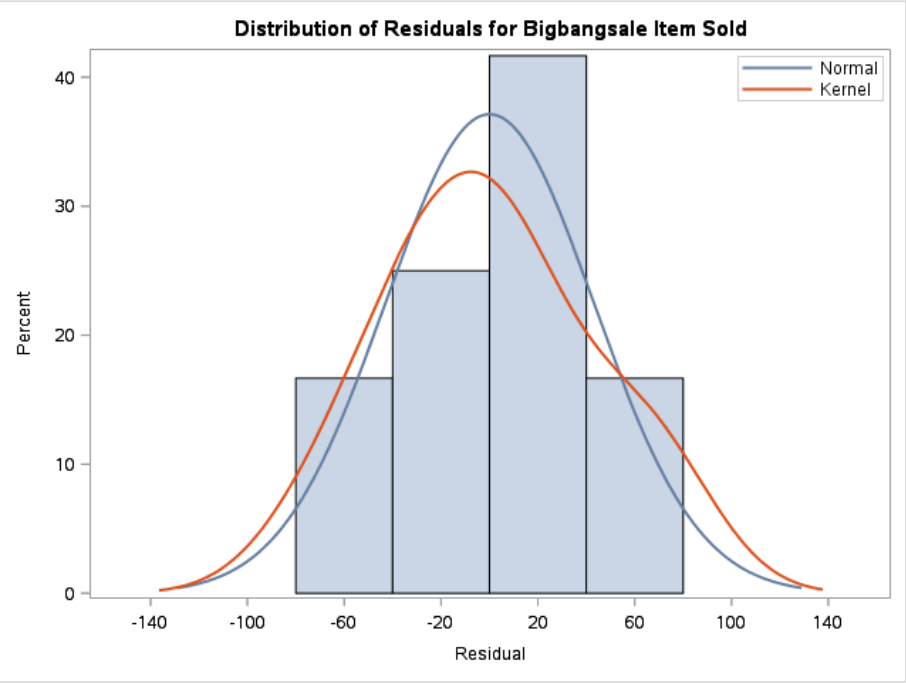



*Figure 6- Distribution of Residuals*

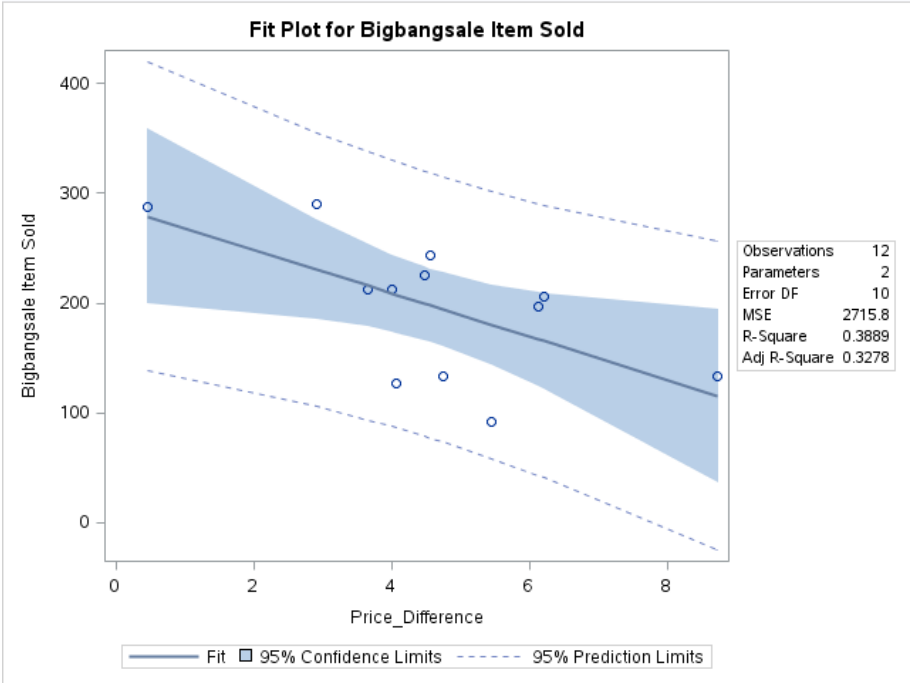



*Figure 7- Fit Plot*

The histogram shows the distribution of residuals are normally distributed and the fit plot shows 95% prediction limits has included all results without outliers.

## 7. Final report

(a) Execute summary

According to this observational study on prices, shipping fee and the number of item sold, we have identified that the sales number of an online retailer would be affected by the price difference between this particular seller and competitor average price of one specific product. It shows the item sold would increase when price difference become narrow with 95% confidence interval. The execution of this observational study project has followed the lab 6 document of subject MAST90072 – Data and Decision Making, 2016 semester 1. It finds a suitable liner regression model in three main steps in SAS after data collection and necessary calculation:

Step 1. Draw scatter plots of possible liner relationship data

Step 2. Correlation analysis of variables

Step 3. Liner regression execution and validate the liner regression model

(b) Potential issues

As mentioned before, in the correlation analysis stage of the project, the possible relationship between price difference and shipping fee difference has been identified. We conclude in later liner regression stage that shipping fee difference should be removed and regarded as an error term in this project, so here occurs a problem on endogeneity as mentioned in other research (J. Miguel Villas-Boas & RUssel S. Winer, 1999) because we are regressing historical sales data on price in a simple way. So a more accurate way to explore competition-based pricing strategy of online retailer could have control on price and other possible factors to conduct an experiment on more than one seller of random products choices etc.

(c) Further improve experiment design

If consider an experiment of pricing strategy on bigbangsale, here is a paired design experiment purposed as a recommendation for further improvement.

During the following 12 months, keep control of average price on all photo frames based on last month average competitor price.

Two level of average price could be defined, for example:

level 0 = (last month competitor price) * 1.1

level 1 = (last month competitor price) * 1.5

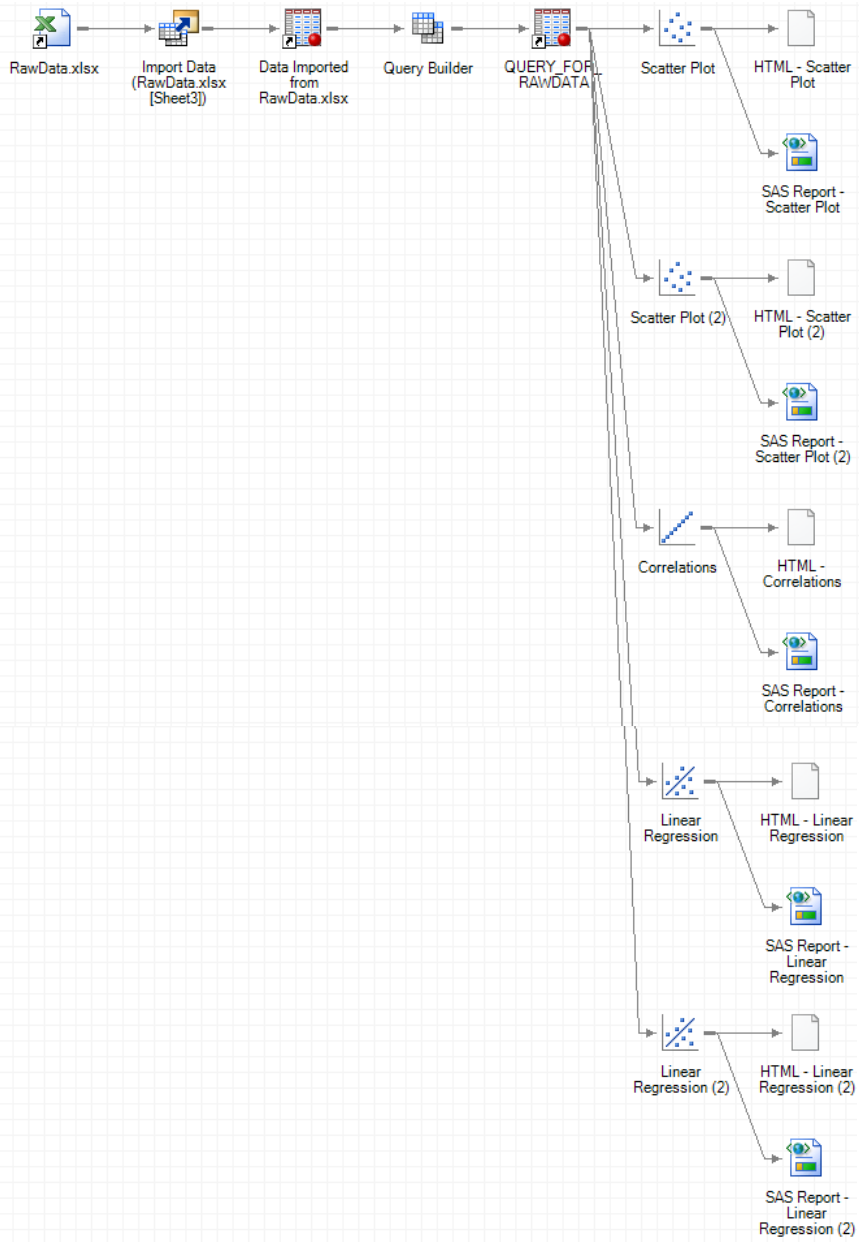
Adjust price of each photo frames randomly in order to keep pricing on the same level during one month. Then change into another price level in the next level and change the price level monthly.

The data gathered in the experiment would be:

| Price level | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item sold | | | | | | | | | | | | |

Therefore, paired t-test could be used to analyze which price level could achieve more sales.

## 8. SAS process flow



## 9. Reference

J. Miguel Villas-Boas & RUssel S. Winer, 1999. Endogeneity in Brand Choice Models. Management Science, 45(10), pp. 1324-1338.

Marchall, F., Santiago, G. & Li, J., 2014. Competition-Based Dynamic Pricing in Online Retailing: A Methodology Validated with Field Experiments,.