Data Analysis 2 : Term Project

Shah Ali Gardezi

Introduction

This report is a part of Term Project for MS in Business Analytics course, *Data Analysis 2* taught at Central European University (Budapest Campus). The project focuses on the evaluating the association of ***Item Sales*** with ***Price per unit*** of Low Fat Bread sold across the 10 stores of BigMart and aims to estimate how confounding variables such as *Item Visibility*, *Outlet Location Type* and *Outlet Store Type* influence this association.

Data and Data Mungaing

The data is extracted from **Kaggle** and it represents the sales of products in the year 2013, in 10 stores of BigMart. The data consists of 12 variables with total number of observations equaling 8523. Some of the variable of interest that are used for this analysis includes;

- Item MRP: This is the market retail price of the product
- Item Weight: This is the weight of the product sold
- Item Outlet Sales: This is sales of the product. This also be our dependent variable the y in our regression
- Item Visibility: This is the percentage of total display area of all products in a store allocated to the particular product
- Item Fat Content: This describes the product's fat content (Low Fat or Regular)
- Item Type: This is the category from which the item belongs to for example snacks, dairy, fruits etc
- Outlet Location Type: This is the type of city the store is located in
- Outlet Store Type: This describes the type of store, for example Supermarket Type1, Grocery store, Supermarket Type2 etc

There was a limitation in the data set with regards to the number of observations for each unique item. For instance the product with highest observations contained 10 observations only. This poses a problem for the quality of the association between the independent and dependent variable. This issue is addressed by creating a new variable named *Product* and making few assumptions about very similar products. The new created variable, *Products* is formed by combining *Item Fat Content* and *Item Type* variable after cleaning of these variables. While there is a lot differentiation in general between the products in the categories like snacks, fruits, seafood, the difference is small in the bread category. The choice of *Low Fat Bread* is made as the product of interest while assuming that all the breads in this category are of the same type with same ingredients etc. Moreover, these breads were all sold by different weight quantities so in order to normalize their price, the Price per Unit of weight is calculated. This is done by dividing the *Item MRP* by *Item Weight* and creating a new variable *Price per Unit*. This will be the independent variable, the x in our analysis.

The number of observations for the Low Fat Bread come out to be 140. However, at this point the missing values in the data of Low Fat Bread have not been account for. The missing values appear in the data for *Item Weight* as no observations are recorded for the weight of the breads sold and thus calculating the price per unit of bread is not possible. It is decided to drop them, and the number of observations come out to be 109.

In order fully uncover the association of *Item Sales* of Low Fat Bread with *Price per Unit* weight, confounding variables need to be selected. In order to have a better understanding of which variables to choose as confounders, the correlation of different variables with the y variable is to be calculated. However, it is noted that there are categorical variables in our data as well, such as *Outlet Location Type* and *Outlet Store Type* which could have a correlation with the *Item Sales*. To overcome this, these categorical variables are changed into Binary (numerical) variables. Three Binary variables for *Outlet Location Type*(*tier1*, *tier2*, *tier3*) and three for *Outlet Store Type* (*type1store*, *type2store*, *grocerystore*) are hence created, each representing a value of 1 for when their value is TRUE in respective former columns. The correlation is then calculated by creating a lower triangular matrix and heat map. The result of correlation is shown in Figure 1 of the appendix. The heat map shows that item visibility in stores, outlet location and store type indeed have a correlation in +/- 0.5 bounds with Item Sales. Hence our confounding variables will be *Item Visibility*, *Outlet Location Type*, and *Outlet Store Type*.

With the data now cleaned and all the variables in place, we observe the values in the data for each variable using *datasummary* (Shown in Table 1 in the appendix). Some degree of right skewness is observed in the x and y variables as the mean is greater than median. We plot the density plots for ***Item Sales*** with ***Price per Unit*** to visualize their distribution and take decision about whether computing the log for them will make the distribution normal distributed. At the same time we check for extreme values. There are extreme values in both sales (Item Sales = 6911) and price (Price per Unit = 32.6), but it decided to include them in this analysis. However, when computing the log, while the log of price helped solve the right skewness of Price per Unit, taking log of sales actually caused the distribution to be left skewed which is highly undesirable so we will proceed with Item Sales values without the log.

Regression Models

Before running on the actual regression models, we make a hypothesis about the expected result of association of Item Sales and Price per Unit weight of low fat bread. Our hypothesis estimates that in an unconditional model, as the Price per Unit weight tends to be higher by one unit, the sales of low fat bread tends to be lower on an average. The reasoning behind this can be thought of that, as the price goes higher, this may influence a lower demand by the consumers and hence lower sales. It is usually in the case of luxury products that the higher price tag item tends to perceived as more luxurious. But since this a necessity item we do not expect that behavior.

$$H_o : \beta_{true} < 0$$

We run a Non-Parametric *Lowess* regression to see how the association looks like. Figure 2 in the appendix shows the result which is contrary to our hypothesis. As the Price per Unit weight go higher the sales of the product tend to go higher as well but up until certain range which is 2.75 log of Price per Unit. This can be accredited to consumers being more health conscious and prefer the low fat bread even if the price goes higher and the higher price might suggest a better quality of the food item. Beyond this point as the price goes higher the sales of low fat bread tend to be lower on average. The *Lowess* also suggests that we can employ splines to better understand the association between the dependent and independent variables. When moving on with our regression we will use spline with knots at log price per unit of 2.75.

Model 1: Sales vs Log Price per Unit

$$Item\ Sales := \beta_0 + \beta_1 ln(Price\ per\ unit)$$

The first regression model is the level-log regression of Item Sales on Price per Unit (results summarized in Table Regression Model Summary in Appendix). The coefficient here shows that in the range of log Price per Unit of less than 2.75, the Item Sales of Low Fat Bread tends to be higher by 16.5 units on average, for one percent higher Price per Unit. This value is significant at 99.9%. While in the range beyond log Price per Unit of 2.75, the Item Sales tends to be lower by 7.6 units on average, for one percent higher Price per Unit. However this is not a significant value. The R square for this regression is 27.5% which refers to the percentage value of the variation in Item Sales that is explained by the Log of Price per Unit, the rest is left for residual variation. The intercept in this regression is meaningless. Its intercept coefficient estimates

that if the Log Price per Unit of Low Fat bread is 0, the Item Sales tends to be lower by 14.8% on average, which is meaningless to infer.

Model 2: Item Sales on Log Price per Unit and Item Visibility

$$Item\ Sales := \beta_0 + \beta_1 ln(Price\ per\ unit)\ + \beta_2(Item\ Visibility)$$

We introduce a confounding variable, the *Item Visibility* in this regression and conditioned Price per Unit on it. The coefficient of log Price per Unit estimates that for the range of log Price per Unit of less than 2.75, the Item Sales of low fat bread tends to be higher by 15.9 units on average for one percent higher Price per Unit keeping everything else constant. This value is significant at 99.9%.In the range of log Price per Unit beyond 2.75, the Item Sales tends to be lower by 5.4 units on average, for one percent higher Price per Unit, keeping everything else constant.However, since the standard error for this coefficient is around 7.8 units, the value of log Price per Unit includes 0 this not significant even at 80% confidence interval. The coefficient of *Item Visibility* estimates that Item Sales tends to be lower by 3782.6 units on average for one unit higher Item Visibility keeping everything else constant, but the value not significant and has a very hight standard error. The R square for the regression is 28.25% a slight increase from the previous regression. The intercept is again meaningless as it tells that at 0 log Price per Unit and with 0 Item Visibility, the Item Sales tends to be lower by 11.4 units on average.

Model 3: Item Sales on Log Price per Unit,Item Visibility and Outlet Location

$$Item\ Sales := \beta_0 + \beta_1 ln(Price\ per\ unit)\ + \beta_2(Item\ Visibility) + \beta_3(Tier\ 1\ Location) + \beta_4(Tier\ 3\ Location)$$

We include another confounding variable, the *Outlet Location* and estimate the results for log Price per Unit. Tier 2 Location is kept as a reference category for it has the highest number of observations. We see that the coefficient is very similar to Model 2 with similar significant levels for both splines. While the coefficient of Tier 1 Location is not important for our analysis, however it estimates that if the store is in Tier 1 Location, the Item Sales of Low Fat Bread tend to be higher 527 units on average than if the store was in Tier 2 Location, keeping everything else constant. This value is significant at 90%. The R square for the regression is 30.7% a little raised from Model 2.

Model 4: Item Sales on Log Price per Unit,Item Visibility and Store type

$$Item\ Sales := \beta_0 + \beta_1 ln(Price\ per\ unit) + \beta_2(Item\ Visibility) + \beta_3(Type\ 2\ Supermarket) + \beta_4(Grocery\ Store)$$

This 4th regression models accounts of another confounding variable which is store type in place of store location. Type 1 Supermarket is taken as the reference variable in this statistical model. The coefficient of log Price per Unit estimates Item Sales to be 18.7 units higher on average for one percent higher Price per Unit, keeping everything else constant. This value is significant at 99.9%. The value of coefficient for second spline is not significant however. The B4 coefficient estimates that if a store is a Grocery Store, the Item Sales of low fat bread tends to be lower by 2299.5 units on average,than Type 2 Supermarket, and this value is significant at 99.9%. The R square for this model is around 42% suggesting the percentage of the variation in Item Sales by independent variable and the rest which is left for residual variation.

Model 5: Item Sales on Log Price per Unit,Item Visibility, Store Location and Store type

$$Item\ Sales := \beta_0 + \beta_1 ln(Price\ per\ unit)\ + \beta_2(Item\ Visibility) + \beta_3(Tier\ 1\ Location)$$

$$+\beta_4(Tier\ 3\ Location) + \beta_5(Type\ 1\ Supermarket) + \beta_6(Grocery\ Store)$$

In this model, we include all the confounding variables and analyze the overall association and changes in the coefficients. The coefficient of log Price per Unit estimates that for the range of log Price per Unit of less than 2.75, the Item Sales of Low Fat Bread tends to be higher by 18.5 units on average for one percent higher Price per Unit keeping everything else constant. This value is significant at 99.9%. However, in the range of log Price per Unit beyond 2.75, the Item Sales tends to be lower by 11.8 units on average, for one percent higher Price per Unit, keeping everything else constant and this value is not significant. Moreover, with Tier 2 Location as our reference the coefficient of Tier 1 Location estimates that if the store is in Tier

1 Location, the Item Sales of Low Fat Bread tend to be higher 510 units on average than if the store was in Tier 2 Location, keeping everything else constant, and this value is significant at 90%. With the addition of store type in the confounding variable, keeping Type 1 Supermarket as the base variable, the coefficient of Grocery Store estimates that the Item Sales of Low Fat Bread tend to be lower by 2187.1 units on average than if a store is Type 1 Supermarket, on average keeping everything else constant. This value is significant at 99.9%. The R squared for this Model is 42.7%

Model 6: Item Sales on Log Price per Unit,Item Visibility, Store Location, Store type and interactions

$$Item\ Sales := \beta_0 + \beta_1 ln(Price\ per\ unit) + \beta_2(Item\ Visibility) + \beta_3(Tier\ 1\ Location) + \beta_4(Tier\ 3\ Location)$$
$$+ \beta_5(Type\ 1\ Supermarket) + \beta_6(Grocery\ Store) + \beta_7(Item\ Visibility * Type\ 2\ Supermarket)$$
$$+ \beta_8\ (Item\ Visibility * Grocery\ store)$$

In this last model of our analysis we add the interaction of store type with the item visibility. The coefficients of log Price per Unit for both ranges, store location do not show much change. The interaction coefficients of Item Visibility with Type 2 Supermarket estimates that if a store is Type 2 Supermarket, the Item Sales tends to be higher by 5836 units on average, if the Item Visibility goes higher by one unit. This value is not significant because of the high value of standard error. The interaction coefficients of Item Visibility with Grocery Store estimates that if a store is Grocery Store, the Item Sales tends to be lower by 5836 units on average, if the Item Visibility goes higher by one unit. This value is not significant as well because of the high value of standard error. The R square of this model is 42%.

Conclusion

Based on the results of running several regression models, we are better able to gauge the association of Item Sales with Price per Unit along with other confounding variables which could impact our independent variable. A step by step regression reveals that some of the confounding variables actually help us estimate a better association between the dependent and independent variable. The choice of model is based on the variables, which are better quality predictors and gives values of higher significance level for the coefficient. Our *Preferred Model* for this analysis is:

$$Item\ Sales := \beta_0 + \beta_1 ln(Price\ per\ unit) + \beta_2(Item\ Visibility) + \beta_3(Tier\ 1\ Location)$$
$$+/\beta_4(Tier\ 3\ Location) + \beta_5(Type\ 1\ Supermarket) + \beta_6(Grocery\ Store)$$

The choice of selection for this model is based on fact that as we subsequently incorporated confounders in our model, the Adjusted R squared began to increase from 27% to 42% up till Model 5. However, as we add the interaction in the model 6 we witness the adjusted R2 tends to decrease. To compare the goodness-of-fit for regression we compare Adjusted R2 square that contain differing numbers of variable, its value increase only when the new term improves the model fit more than expected.In our case when we add the interaction term of Item visibility with store type, the adjusted R square decreased suggesting this interaction was not contributing towards a better model fit. Moreover, when we see the p values for our regression while comparing model 5 & 6 we see that the level of significance is same for the coefficients but the p values in model 5 is comparatively lower. Therefore, considering these two factors we can conclude that model 5 provides a better simulation to compare with different estimators with respect to bias and mean squared error.

Lastly, our Null hypothesis is invalidated in the first spline. According to our model when Log Price per Unit gets higher, the Item Sales tends to go higher on average, up until our first spline knot (log Price per Unit of 2.75). This suggests that our product is highly elastic. Its not behaving as a necessity item should but rather mimics the behavior of a luxury product! While the idea of bread being a luxury seems peculiar, looking from another angle a Low Fat Bread is indeed a luxury. It has a niche market, not everyone wants it but those who want it are willing to pay a bit higher for it.The second spline however, validates the Null hypothesis the reason is that there is tipping point after which consumers tend not to buy the Low Fat Bread which has an higher price. Because at the end of the day it is a bread, people have the option of other healthier alternatives to switch to.

Appendix

Table 1: Descriptive statistics

	Mean	SD	Min	Max	Median	P95	P5	N
Item Sales	2169.61	1452.39	167.78	6911.00	1860.25	4818.53	4818.53	109
Price per unit	12.81	7.90	3.13	32.63	11.21	28.11	28.11	109
Item Visibility	0.06	0.05	0.00	0.15	0.05	0.14	0.14	109
Store Location:Tier 1	0.22	0.42	0.00	1.00	0.00	1.00	1.00	109
Store Location:Tier 2	0.43	0.50	0.00	1.00	0.00	1.00	1.00	109
Store Location:Tier 3	0.35	0.48	0.00	1.00	0.00	1.00	1.00	109
Supermarket:Type1	0.79	0.41	0.00	1.00	1.00	1.00	1.00	109
Supermarket:Type2	0.15	0.36	0.00	1.00	0.00	1.00	1.00	109
Grocery Store	0.06	0.25	0.00	1.00	0.00	1.00	1.00	109

Figure 1:

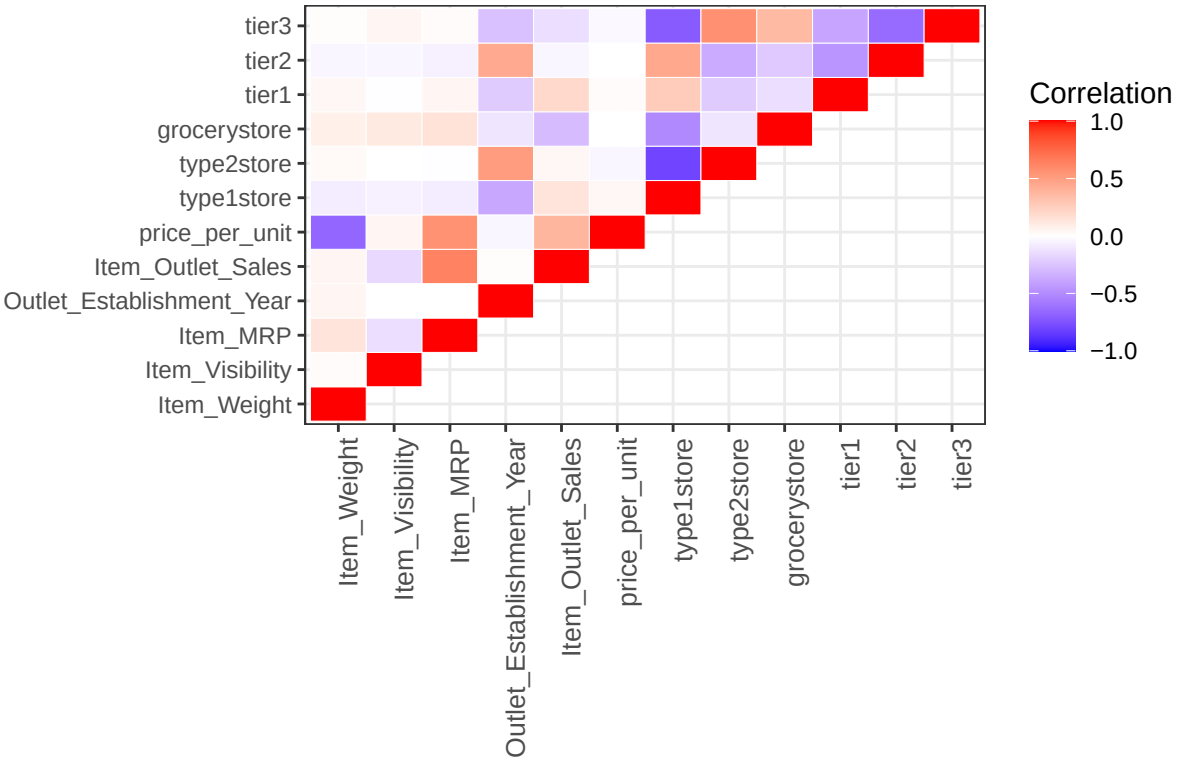

Figure 2: Non−Parametric Lowess − Item Sales & Log Price per Unit

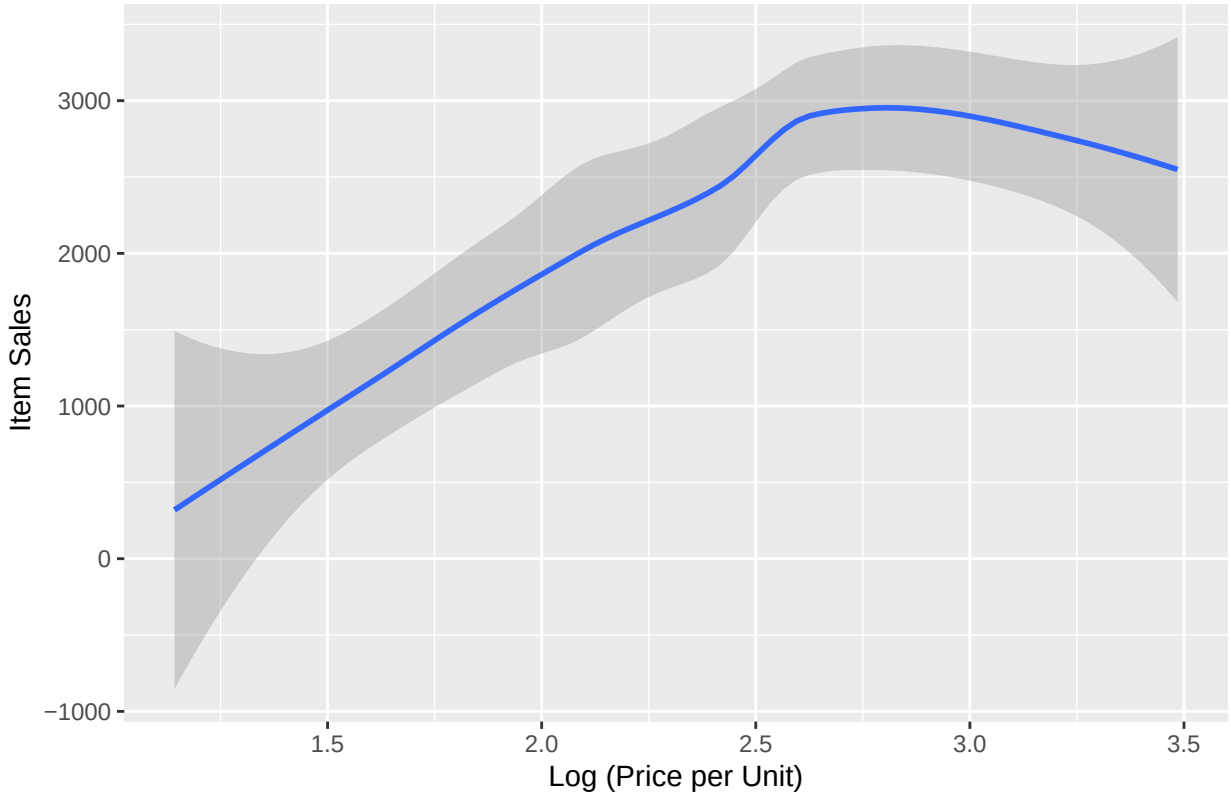

Table 2: Regression Model Summary

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	−1488**	−1140*	−1170*	−1718**	−1810**	−1773**
	(422)	(481)	(467)	(428)	(427)	(424)
ln(price per unit) < 2.75	1649**	1593**	1586**	1877**	1851**	1855**
	(236)	(238)	(230)	(226)	(223)	(224)
ln(price per unit) > 2.75	−763	−545	−582	−1219	−1178	−1130
	(731)	(753)	(781)	(771)	(776)	(785)
Item Visibility		−3783	−3591	−1780	−1875	−2679
		(2488)	(2416)	(2053)	(2060)	(2317)
Tier1 Location			527		510	511
			(281)		(276)	(279)
Tier3 Location			−227		51	53
			(275)		(271)	(277)
Type 2 supermarket				133	232	−138
				(348)	(395)	(512)
Grocery Store				−2300**	−2187**	−2158**
				(272)	(322)	(389)
(Item Visibility) x (Type 2 supermarket)						5836
						(5441)
(Item Visibility) x (Grocery store)						−119
						(3757)
Num.Obs.	109	109	109	109	109	109

* p < 0.05, ** p < 0.01