Humboldt University Berlin

Institute of Marketing

Prof. Dr. Daniel Klapper

Advanced Marketing Modeling

SS 2021

Special Work Performance 2: Estimating Price Elasticities and Promotion Uplift

This is non-graded individual work.

Your answers including all tables and graphs must not exceed 10 pages. Please start a new page when providing your report to a new subtask. Please use typeface Times Roman in 12pt with 1.15 line spacing (in tables and graphs you may use 10pt and 1.0 line spacing) and 1 inch space on all sides. Do not forget to report your name and student number and a page number on each page starting with number one on the first answering page.
Do not include a title page or content page.
Send your report as pdf to my email address daniel.klapper@hu-berlin.de not later than June, 11, 2021, 4:00pm. Please report in the subject line "AMM SWP2 and your name".

Download the dataset "dat.coke.pepsi.csv" from the Moodle course page. The data are store-level scanner data from two local markets, reporting sales, revenue measures and marketing activities of the following brands: "Coke Classic", "Diet Coke", "Coke Zero", "Pepsi", and "Diet Pepsi". The brands are offered in the following package formats: "12 cans of 355ml each", "24 cans of 355ml each", "2 liter plastic bottle", "592ml plastic bottle". The data come from 13 stores in two different local markets ("EAU CLAIRE", "PITTSFIELD") and cover the period from January 2004 until December 2006. Your dataset has 31681 (first row reports the variable names) data rows and 20 columns. The columns provide information about the following criteria:

1. IRI_KEY
2. YEAR
3. WEEK
4. PRODUCT.TYPE
5. FLAVOR.SCENT
6. L4
7. L5
8. VOL_EQ
9. PACKAGE
10. UNITS
11. DOLLARS
12. price
13. display_minor
14. display_major
15. feature_small
16. feature_medium

17. feature_large
18. display_all
19. feature_all
20. MARKET

Additional information about the data is discussed in class.

This special work performance is designed that learn how to estimate the effect on price on sales via price elasticities and the uplift of sales due to display activities and retailer feature advertising.

Therefore, regress log of units on log price and promotional instruments and potential seasonal and other sales shifters. Do the analysis separately for each brand and package format combination and document the estimation results in a few tables that contain all important estimation results and that allow an intuitive understanding of the key similarities and discrepancies of the estimation results across brands and package formats. Interpret the results carefully and document your estimation strategy in some detail. Do not report R-codes and edit the estimation results you obtained with R. Also, make use of your econometric knowledge when estimating parameters of a linear additive model. Use tables and graphs to support your description and explain in words the key facts of the data set.

Advanced Marketing Modelling SWP 02

This assignment is meant to be the 2nd Special Work Performance for the course Advanced Marketting Modelling Summer semester 2021, in which we are essentially expected to calculate different price elasticities of various product combinations of brands, stores, volumes etc. while at the same time estimating the effect of display activities and retailer feature advertising on the uplift of sales. To provide a more specific outline of the assignment, first, the data set used for analysis will be introduced while also applying some explanatory analysis in order to get a better grasp of it. This will be useful for the following sections in the assignment. Then, in the next part, various aspects of the assignment will be elaborated on in detail such as the building of the marketing response model used for parameter estimation approach. This will focus on which, why and how a combination of subset of the data were analysed and the key insights we learn from those. Lastly, in the last part, a summary of what has been done, possible limitations of the data set and what other approaches that could be and can be taken for further analysis will be provided.

Introduction and Explanatory Analysis

The given data set for the assignment are based on the products of the two biggest rival brands in the carbonated beverage market from the year 2004 to 2006 on a weekly basis in different stores in EAU Claire and Pittsfield, US. In the data set, we have several additional variables, one indicating the volume equivalent of the products, one indicating whether the product is sold in cans or plastic bottles, two indicating whether the product had a premium display or relatively lower display in the store (namely and respectively, major_display and minor_display) and lastly three others conveying the information whether the specific store has allocated a small, medium or a large budget for advertisement of the specific product, if any. Also, a check has been done to understand whether the very low sales of some stores stems from them being opened recently, but it was observed that that is not the case. Since these stores having exceptionally low sales might harm the quality and the consistency of the results being produced and the data analysis being carried on, store 259111 and 266596 are excluded from further consideration. Additionally, although display and feature variables are supposed to be binary ones, they included some values in between 1 and 0. So, those values have been rounded to the closest integer to avoid any potential problems in the future. Lastly, one other practical aspect to know about the data set is that certain volume equivelants correspond to only certain packaging namely 0.1042 and 0.3521 to plastic bottles small and big respectively and 0.75 and 1.5 to 12 pack and 24 pack cans.

Some facts about the data:

	Units	Revenue
Coca Cola Co	2013546	5206665
Pepsi Inc	1547157	3740460

Figure 1. Table showing the market shares of the two competitors

Some general argument why there is such a difference between two brands' overall sales could be that Coca Cola has a price mean of 2.606 where as Pepsi's price mean is 3.017. Also, Coca cola's overall display and feature mean is 0.4237 and 0.2144 respectively whereas Pepsi's are 0.4182 and 0.2047, which are slightly lower than of Coca Cola and might constitute a potential argumentation of difference in sales between two brands.

Vol Eq./Products	Coke Classic	Diet Coke	Diet Pepsi	Pepsi
0.1042	123,675 (Same)	128,578 (Same)	104,661 (Same)	115,554 (Same)
0.3521	333,341 (2nd)	212,215 (2nd)	194,631 (2nd)	359,404 (1st)
0.75	453,068 (1st)	428,072 (1st)	259,189 (1st)	319,995 (2nd)
1.5	122,957 (Same)	154,290 (Same)	86,387 (Same)	107,336 (Same)

Figure 2. Table showing brand's product/volume combination focus in dollars of revenue

We observe that Pepsi classic product focuses its sales predominantly on 2 liter plastic bottles contrary to other products and brands. However, this could be a mistake because of the high profitability of the cans as can be seen in the table below:

	Units	Revenue
Can	1,961,502	7,122,801
Plastic Bottle	1,599,201	1,824,324

Figure 3.Units sold and revenue in dollars distribution of packaging formats

So, in order to decide whether a specific product volume has the highest sales in a product mainly because the consumer prefers that over the alternative volumes of the same product or because company also wants to deliberately target that product and volume combination. We, for instance, multiplied with the number of Pepsi 0.3521 units sold with Pepsi 0.3521 display all mean, and did the same for every product and volume combination (here in total 16 combinations). In the end, what was found is that the highest sale product volume combinations have the highest display all mean. So, it is not only consumers preference but companies also target those specific combination of products.

Year/Vol Eq	0.1042	0.3521	0.75	1.5
2004	166675	378607	347264	244499
2005	169199	378530	556624	124169
2006	150632	355558	585258	103688

Figure 4. Table combining volume and display_all means over time

From the table we can also see that there is a general trend towards 12 packed cans in general whereas the other packaging/volume combinations decrease in units over time.

One other unexpected fact is that only in Store 652159, Pepsi sold more units and made more revenues than Coca Cola. We now investigate why that is the case.

	Display minor	Display major	Feature small	Feature medium	Feature large	Display all	Feature all	Price
Coca Cola	0.1120	0.1955	0.0112	0.0601	0.1233	0.3075	0.1947	2.677
Pepsi	0.1516	0.2002	0.0032	0.1061	0.1393	0.3518	0.2486	3.345

Figure 5. Brand specific marketing activity and price mean comparisions

From previous tables we know that Coca Cola has higher unit sales and revenue than Pepsi; however, only in store 652159, Pepsi has higher units and revenue than Coca Cola does. Investigating the reason for this, figure 5 shows us the mean display, feature and price in store 652159 for both of the brands. What we observe based on the table is that extra display place in the store and additional advertisement, as one would expect, helped Pepsi to surpass Coca Cola in sales and overall revenue in that store.

Building of the regression model

When building the marketing response model, it would make sense to include all variables except the ones such as market or packaging as these would already be included with the inclusion of other variables. Also, I interacted log of price with products, volume equivelant, store ids and years as a factor. The reason for this is that when including these variables to the regression, most of the variables come out as significant. Second reason is that it makes sense to argue that price elasticity would depend on different stores, brands, volumes as well as years as it might change over time. However, when feeding the regression model with such numbers of variables, it gets very hard to interpret. In some cases, the coefficents cannot even be seen. So, because of that reason I omitted the year factor from the interaction chain in the model.

```
Call:
lm(formula = log(UNITS) ~ log(price) + display_all + feature_all +
    spring + summer + as.factor(L4) + as.factor(L5) + as.factor(VOL_EQ) +
    as.factor(MARKET) + as.factor(YEAR) + week + as.factor(IRI_KEY) +
    log(price):as.factor(L5):as.factor(VOL_EQ):as.factor(IRI_KEY),
    data = dat.cola)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3877 -0.2955 -0.0061  0.2849  3.4660

                Estimate Std. Error  t value Pr(>|t|)
(Intercept)  4.3416008  0.0295965  146.693  < 2e-16 ***
log(price)  -4.2693488  0.0792303  -53.885  < 2e-16 ***
display_all  0.6649756  0.0103463   64.272  < 2e-16 ***
feature_all  0.1217706  0.0130279    9.347  < 2e-16 ***
spring      -0.0365309  0.0258363   -1.414 0.157392
summer       0.0557191  0.0251840    2.212 0.026941 *

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5853 on 28331 degrees of freedom
Multiple R-squared:  0.7838,    Adjusted R-squared:  0.782
F-statistic: 431.6 on 238 and 28331 DF,  p-value: < 2.2e-16
```

Figure 6. Linear regression summary estimating the price elasticity and uplift effect of the sales

The regression output seen above is the one that is used to estimate the uplift parameters and the price elasticity as well as some time interaction terms. Even though, the interpretation gets almost impossible given the amount variables in the model, since I believe price elasticity differs from store to store thus the interaction term for that had to be included because stores apply different display and feature activities as well as different pricing strategies, which would cause every stores' price elasticity to differ. The same way of thinking applies to inclusion of product and volume level interaction too. One other reason to use the above specified variables in the model is that the fact that cola as a product in general is a quite homogenous product where people cannot accurately differentiate between products in blind tests and this leaves the existing differences between the products to be explained by either outside factors such as the ones above or some other marketing activities. Issues like whether consumers are more price sensitive toward Pepsi or Coca Cola or towards certain volume levels are inferred from plots, mean comparisions and simpler regression models, which will be elaborated on later.

Coming to the interpretation of the specific regression above, in the model whole data was used as it is better to interpret the price elasticity and effects of marketing efforts with more data as much data as possible. The model is able to explain the 78.38% of the variation in the data set and is highly significant looking at the F-statistic and p-value. Around 95% of the variables that are not shown here including the interaction terms are found to be at the thousand level significance.

So, the reason why this specific model was chosen is that because it is consistent and in line with what one would expect certain factors to affect the sales in real life. For example, the price elasticity is not smaller than -1, which would have caused the brands to increase price as the total revenue from increasing the prices would outcompansate the loss of sales that would result from a price increase. In fact, when we examine the price elasticity here, we end up with the interpretation that a 1 percent increase in price would result in 4.27 decrease in sales. In the additional regressions, we found that Pepsi's price elasticity is more negative than Coca Cola's meaning that their sales decrease more than Coca Cola's in case of a price increase in their products as their consumers are more price sensitive. This could result from potentially brand equity that the company has built over the course of years. We also find that as volume increases price sensitivity increase. This is likely to result from the fact that as volume increases the price increases as well. And one last cross-price elasticity is that the fact that Pittsfield is being found to be more price sensitive than EAU Claire.

Coming to the display and feature activity, in the regressions the confidence intervals of small display and large display was barely different and confidence intervals of feature variations were different to a extent that in the model just feature_all variable could be used because the models aim are rather than representing the reality to a full extent, which is impossible, to provide us with meaningful insights. Throughout the analysis, we have observed the impact of additional display area to be higher on sales than retailer making advertisement, which is confirmed by the model. During the model building process, feature variable were the variable that is most prone to giving false coefficients as it would indicate most of the time that additional advertisement of the retailer would decrease the sales, which is counter intuitive. However, here we observe that it contributes to the sales in a positive way and in a way that is less than what additional display would contribute,

which makes the model consistent with out findings. To be more specific, when we take the exponent of the coefficients of display and feature variables, we learn that at least having an additional display area in the store for a certain product would cause the sales to almost double with a value of 1.94 and the retailer somehow advertising would increase the sales 12.95% in total.

Also, at least in ten weeks of the data set we would observe a very sharp increase sometimes backed up with discounts and additional marketing effort and sometimes without anything, which would point out to us that there could be potentially some seasonal shifters affecting the data. There is a few that might be affecting certain weeks, which we will talk about in the coming part but in the regression model I have checked a very straight forward assumption about the sales. The assumption is that since cola is a beverage consumed cold, people would tend to buy it more during the months where the temperature is higher than the others. Thus, I included dummy variables for each season and found only summer to have a significant effect on the sales at the 5 percent significance level, which is a indication that it may match with the assumption that higher degrees correlates positively with the sales of cola products. So, when we take the exponent of the summer time shifter's coefficient we observe that when it is summer the sales increase 74.58%. Additionally the reason why spring may not have been found to be significant is that US being in the north, temperature at spring might still be not enough to envoke the need of refreshment with a cold beverage. Figure 7 provides additional support of the role of seasons in sales of cola products.

	Winter	Spring	Summer	Fall
Unit Sales	65979	69961	72812	66861

Figure 7. Seasonal effects on sales

Vol Eq	0.1042 (Bottle)					0.3521 (Bottle)				
	Coca Cola	Pepsi	Diet Coke	Diet Pepsi	Coke Zero	Coca Cola	Pepsi	Diet Coke	Diet Pepsi	Coke Zero
Display All	-0.1695 **	-0.0521	-0.1407*	-0.0357	0.8575** *	0.3082* **	0.4308 ***	0.3419** *	0.3858 ***	0.5801 ***
Feature All	0.0525	0.0074	-0.2749*	0.1172	0.8583*	0.0287	0.1589 ***	0.1054** *	0.2361 ***	0.1323 .
Log(Price)	-1.9436 ***	-0.1932	-1.6871 ***	0.2472	-0.8288 .	-2.5982* **	-2.928***	-2.5812** *	-2.9506 ***	-2.5255 ***

Vol Eq	0.75 (Can)					1.5 (Can)				
	Coca Cola	Pepsi	Diet Coke	Diet Pepsi	Coke Zero	Coca Cola	Pepsi	Diet Coke	Diet Pepsi	Coke Zero
Display All	0.7550** *	0.9667 ***	0.7402 ***	0.9063 ***	0.7026 ***	0.5050 ***	0.7666 ***	0.6096 ***	0.7969 ***	0.7969 ***
Feature All	0.0141	0.0569	0.0151	0.0935 .	-0.0157	0.4223 ***	0.1396*	0.3995** *	0.1253 .	0.1253 .
Log(Price)	-3.055 ***	-3.828 ***	-3.1620 ***	-3.781 ***	-3.538 ***	-7.237 ***	-4.6406 ***	-7.3619 ***	-4.2226 ***	-4.2226 ** *

Figure 8. Regression model iterations on combinations of products and volume equivelants

- (lm(log(UNITS)~log(price) + display_all + feature_all +spring+summer +

  + as.factor(MARKET) + as.factor(YEAR)+week+ as.factor(IRI_KEY) +

  log(price):as.factor(IRI_KEY), data= "Changes in each product volume comb."))

Insights that one could derive from figure 8:

1- Display of 600 ml bottles doesn't seem to provide additional sales except for Coke Zero.
2- With all product/volume combinations we observe the increasing price elasticity as volume increases or price.
3- With the product that have larger volume, we observe higher percentage that the feature activity gets significant.
4- In all comparisions Pepsi's price elasticity is higher. So, the consumer acts in a more price sensitive way when Pepsi increases prices. Likely because Coca Cola has a stronger brand equity.
5- However, probably because of the same reason Pepsi takes advantage of its budget spent on display and feature activity more than Coca Cola does. So, its Investment to Return ratio seems to be higher than of Coca Cola.
6- Regarding Coke Zero, people seem to respond less, for example, to a store's brochure having Coke Zero in it, rather than seeing the actual product in the store. So, for Coke Zero focusing on keeping the display area as a given at first stage is probably logical.

```
> aggregate(cbind(UNITS,DOLLARS)~display_all+feature_all,data = cola, mean, na.rm = TRUE) #
  display_all feature_all     UNITS  DOLLARS
1           0           0  60.73789 125.9607
2           1           0 171.21805 572.0237
3           0           1 122.50963 297.8625
4           1           1 299.25906 683.3065
> aggregate(cbind(UNITS,DOLLARS)~display_all+feature_all,data = pepsi, mean, na.rm = TRUE) #
  display_all feature_all     UNITS  DOLLARS
1           0           0  50.49676 104.2171
2           1           0 126.49671 429.9637
3           0           1 133.03441 276.2325
4           1           1 310.04568 631.8705
```

Figure 9. Display and feature interaction effects on sales and revenue

This finding also supports the findings above such as that of display affects Cola's sales more than feature. However, the same relationship is less strong with Pepsi. In Pepsi we observe; however a bigger increase in sales when display and feature is combined.

Brand, volume & store specific analysis

To be able to understand the effect of price, display and feature activity on the uplift of sales, in this part of the special work performance we take a more detailed look in a combination of brand, volume and store and its interactions with sales and other promotional activities. This combination is Coca Cola of 12 cans in store 653776 because this combination of brand, volume and store had the highest revenues among all of the others.

As one would expect first of all 12 packs of Coca Cola products have a negative correlation of between its price and the revenue it brings of -0.26.

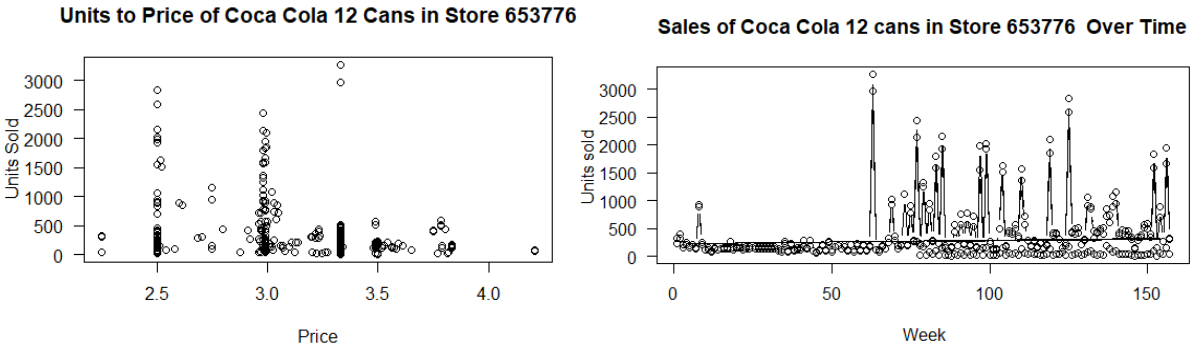

Figure 10. Analysis of 12 pack of Coca Cola in store 653776

When observing the two plots, on the plot on the right, we see from week 63 on a sudden increase in variation of unit sales and in the amount as well as frequency. On the plot on the right, we see that in prices such as 2.5, 3 or 3.33 sales sometimes explode, which drives us to investigate the causing force.

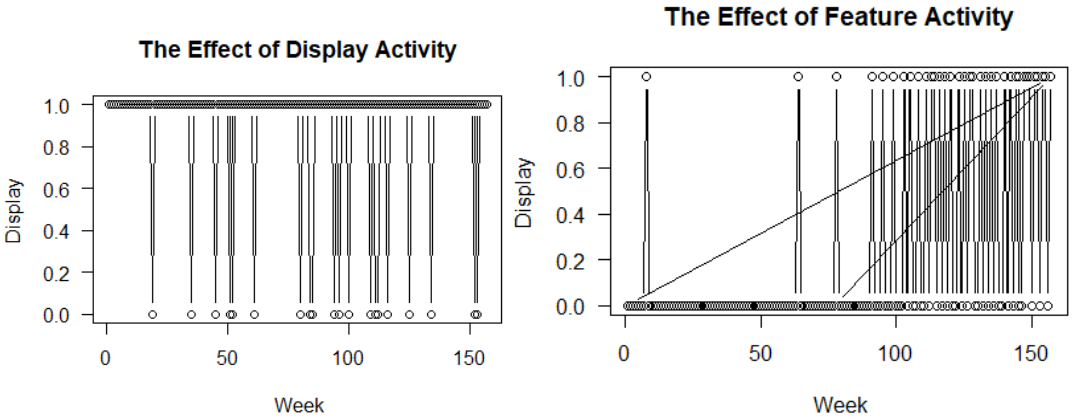

Figure 11. Display and feature distribution of the same product

On the plot on left we see almost all of the sales are with the help of display activity, to be more specific 94.43%. However, on the plot on the right side we see a potential positive correlation between the increasing frequency of retailer advertisement and unit sales.

Means of Units Sold	Feature 1	Feature 0
Display 1	445	345
Display 0	91	112

Figure 12. Interaction effects of feature and display on the specific product

In figure 12, we see how the means of the units sold changes with different combinations of display and feature activity for 12 cans of Coca Cola products in store 653776. Other than illustrating positive effects of feature and display activities this table provides us one more insight about the

interaction between display and feature values. Above, we observe that when there is only advertising of the retail store the sales unrealisticly decrease; however, the advertisement of the store actually increases the sales further when there the product is displayed also in a premium way. The same relationship is observed when we compare these using the whole data, using only Coca Cola products. Only in aggregated Pepsi products we observe that the mean of the units sold are higher when there is only feature activity than in the case when there is only display activity.

One insight that could be derived from examining this relationship is that in week 63 the sales are the highest and the 2nd highest, however there is only large display as 90% of the regular sales and the price is as usual in 3.33. However, one clue is that the two purchases in the same week consists of around 3000 of Coca Cola Classic and again around 3000 of Diet Coke. Although I was unable to find any events during the week of 63 (March 7 2005 – March 13 2005), there could be some local event like the university of EAU Claire having a sports tournament or symposium or some local fair in the neighborhoud etc., in which the organizer prepares for by offering the visitors both the option to drink a Classic Coca Cola and a diet one.

On the other hand, the 3rd and the 4th highest sales that is both during week 125 is likely to stem from a price reduction to 2.5 and major display plus a medium feature activity.

Coca Cola vs. Pepsi

Now in this section I compare Coca Cola's and Pepsi's sales in general, analysing the drivers of the sudden changes in the sales and the possible time-specific effects that might help explaining them.

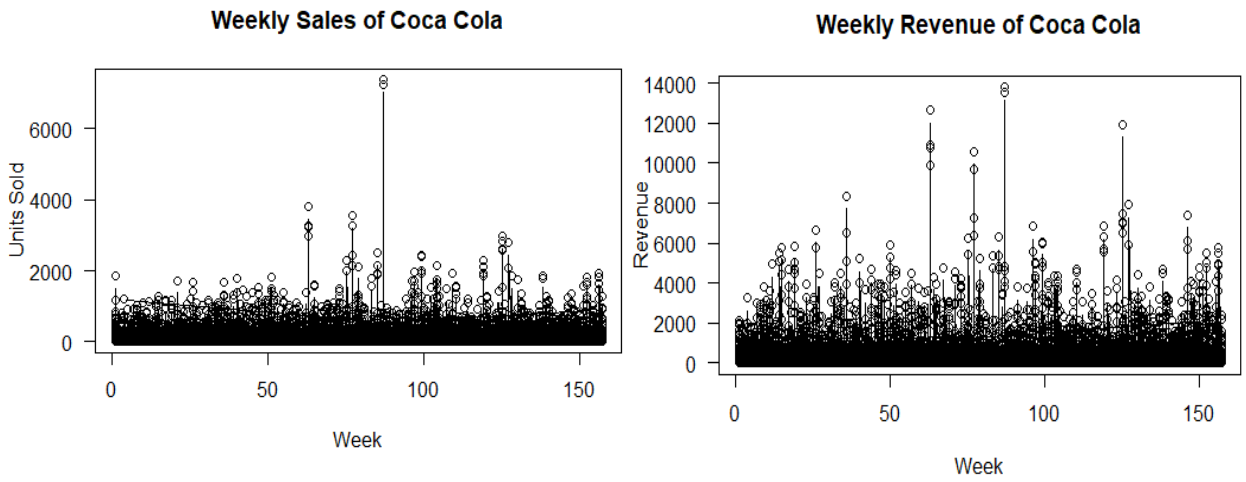

Figure 12. Analysis of Coca Cola in general over time

So, again as we have seen before, in figure 12, we observe that there are some unusual peaks in sales of Coca Cola brand. We will only try to explain some of the extreme sales that took place in one week all together. So, taking a look at the sales, the 1st and the 2nd highest sales take place in the week 87, which is in between 22 and 28th of August, 2005. I was unable to find out what event was there in this particular in EAU Claire or in US in general, but again two purchases of 7500 of

Coke Classic and 7500 of Diet Coke was made, which points out to an event organization, likely a local one. In 125th week the 4th highest sale took place, which is just a high volume transaction of 0.75 cans for a price of 4.60 in Pittsfield although its average price is 3.48 $. In week 77, there has been the 7th highest sales. This week corresponds to June 13 – June 19 2005, which includes Father's Day. So, families might gather for dinner, do some outdoor events such as barbeque etc. as the weather is probably also warm. So, in this week the prices of the classic and diet 0.75 Coca Cola pack has been reduced to 2.98 from a mean of 3.48, that could be a Father's Day campaign. But the very week is the week where Coke Zero has been introduced, so additional promotion that was meant for Coke Zero might have affected other products of Coca Cola too.

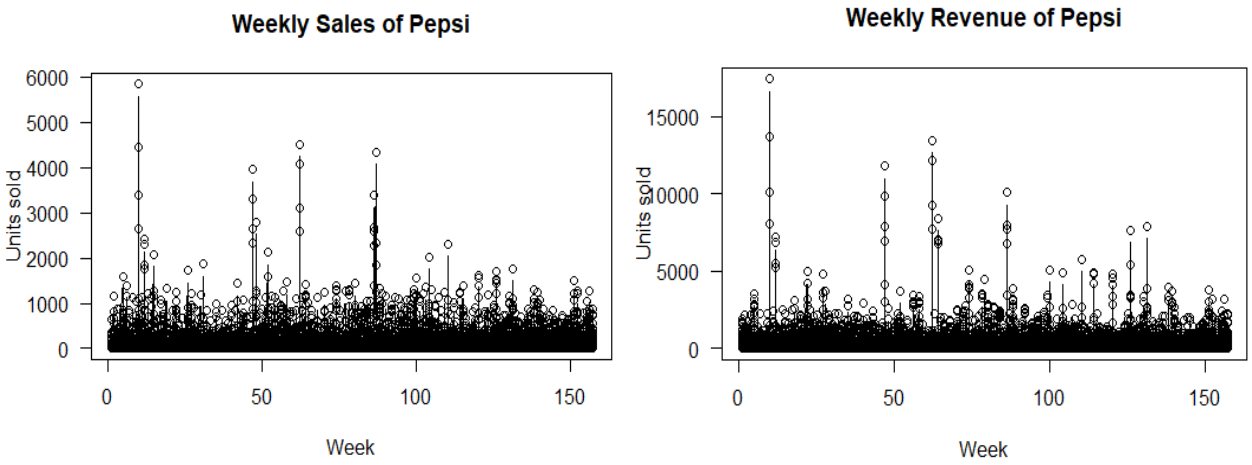

Figure 13. Analysis of Pepsi in general over time

In week 62, that is one week before Coca Cola's peek with the same campaign as Pepsi applied and where Pepsi had its 3rd and 4th highest sales, Pepsi applies the same campaign of decreasing its 12 cans to a price of 2.98, which had a mean price of 3.50 $ before. In week 86, Pepsi had its 7th highest sales by again the same strategy and one week before Coca Cola applies the same exact strategy. So, it seems like by acting earlier Pepsi tries to collect some of the profit in the market, that become available for some reason before Coca Cola does. Pepsi also does the same campaign in week 47, which corresponds to the week before Thanksgiving, that may potentially explain the high sales in that particular week because of preparations for the national holiday. And lastly, unlike Coca Cola, Pepsi seems to have targeted week 10 additionally. There is again a huge purchase of normal 12 cans of Pepsi and Diet Pepsi, hinting a preparation for an event.

Conclusion

Overall, IRI Academic data set is a very extensive data set that could be divided into subsets in so many different ways, which would enable us to control for not only the effect of price changes of a competitor's product but also some complementary products such as chips, that is probably one of the supplement products of coke that is bought together frequently. Also, feature value behaving in a unreliable way made the regression estimation harder. Another issue that didn't allow us to investigate a lot is that the introduction of Coke Zero being from only the middle of the 2nd year.

Additionally, price elasticities tend to be non-linear and dynamic as price and other factors change, which we haven't took into account here but in reality is likely to play a role in estimation results. Lastly, autocorrelation in the models was checked but since it wasn't much of a problem it was not dealt with. However, if one wants to possibly improve the estimation results he or she should check the presence of the regression assumptions holding true and other potential problems such as outliers creating some bias in the model or inconstant and varying error term over time, which I checked for but didn't apply any specific transformations or procedures for.