Customer Choice Model

Yan Li

Project Overview

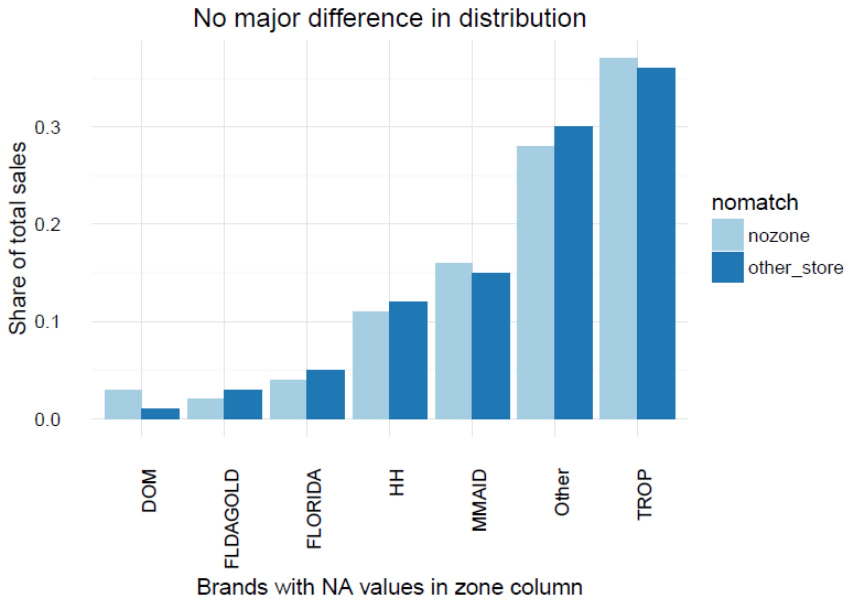
This dataset came from Dominick's Database, which included product sales, retail price, wholesale price and other promotion information per week per product by store in supermarket chain. In this project, we will use discrete choice model and instrumental variable approach to recognize the relationship between price and customer purchase choice and the price elasticity for juice market.

Data Overview

By initial exploration of the data, we can observe that the dataset is relative clean. We first join the category sales data with zone based on the store information.

Missing value

We find there're several NA's for column 'zone'. We have to check the importance of NA's then decide how to deal with them. The share of sales of these stores (zone = NA) in the overall chain is 7%. We also care if the top brands sold in these stores are very different from the other store.

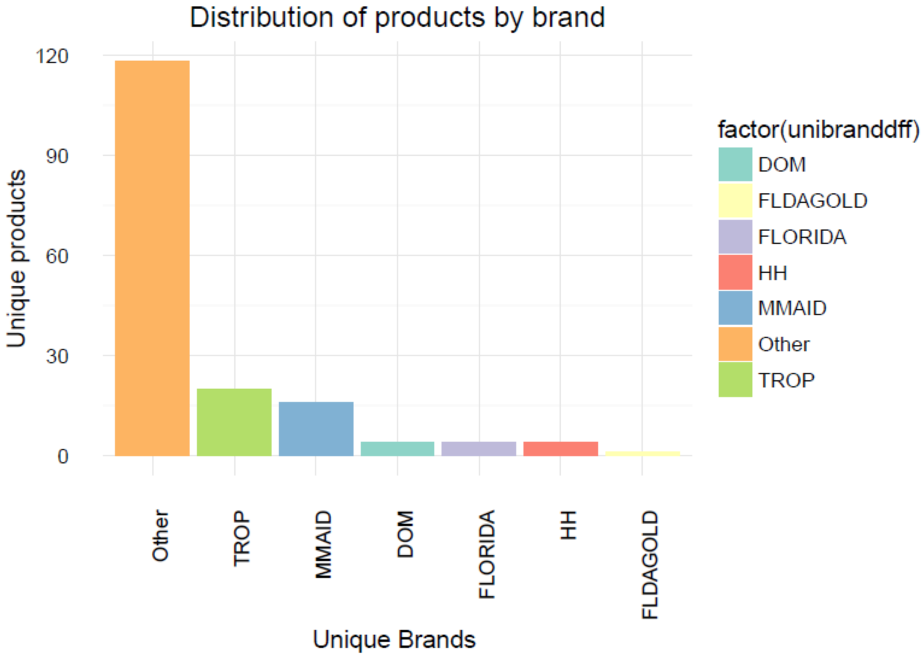


From the plot, we can see the shares for top brands sold in these stores(don't have zone) are not very different from shares in the other stores that belong to the chain.

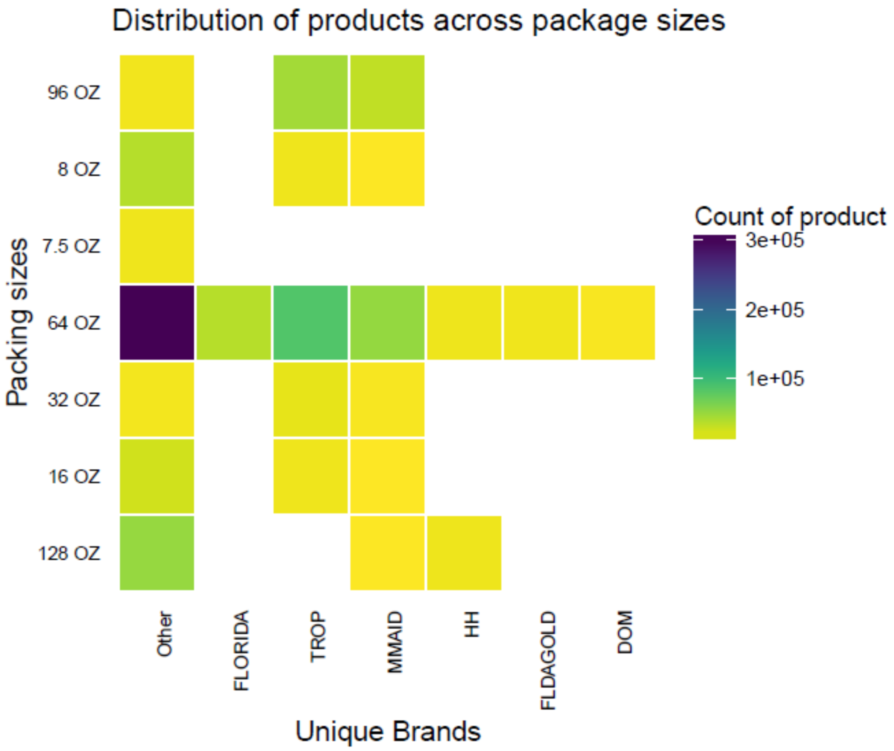
In conclusion, because the overall market share for stores without zone are not significant and the market shares for top brands sold in these stores are not very different from others, we believe remove these stores won't influence our analysis result. We decide to remove them.

Overall understanding of dataset

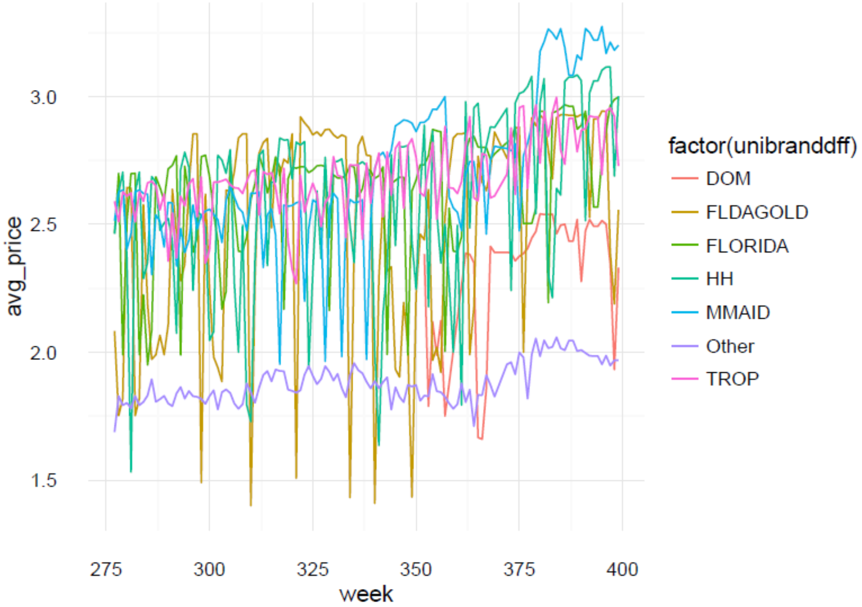
After the aggregation of data, we find there are 121 weeks, 81 stores and 7 unique brands. For each brand, the distribution of product is listed below. We can see other brands include lots of products whereas Florida and MMAID have fewer products.



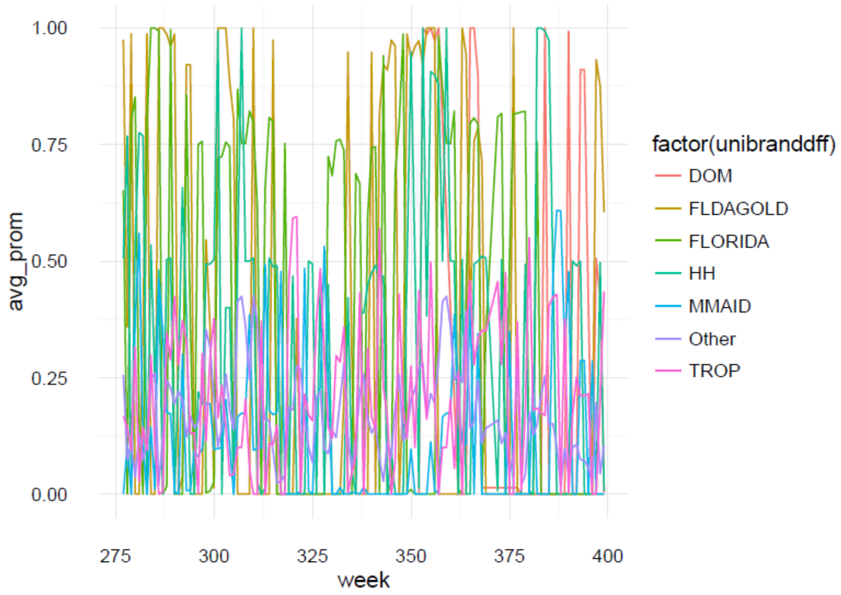
We're also interested with how products differ in terms of package size. From the plot below, we could see if the color is darker, this brand has more products in one size. For example, Florida only has one package size(64oz) whereas MMAID has six package sizes.



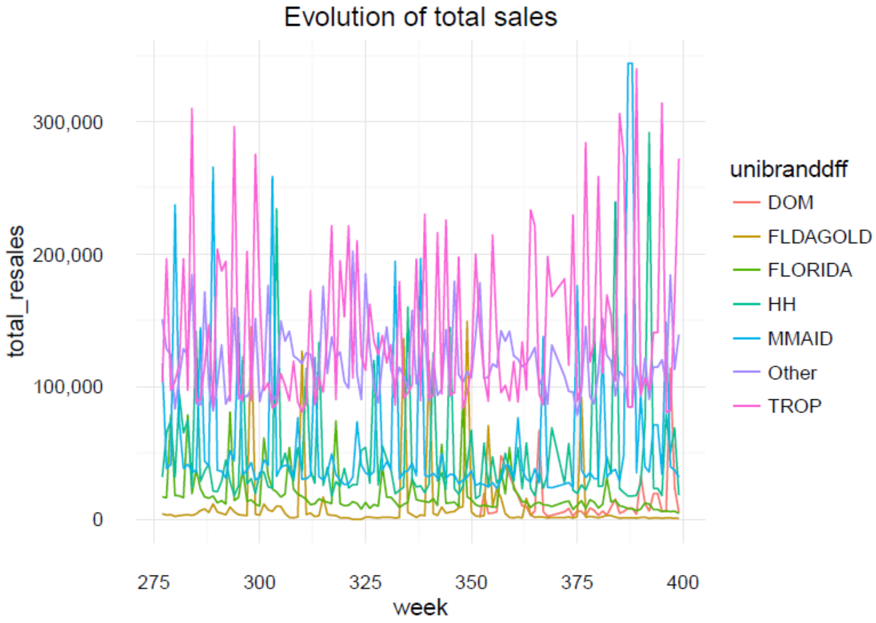
What's more, we also want to know if the average price, average promotion chance, total sales and market share will change with time. The plots below present the change pattern of them respectively.

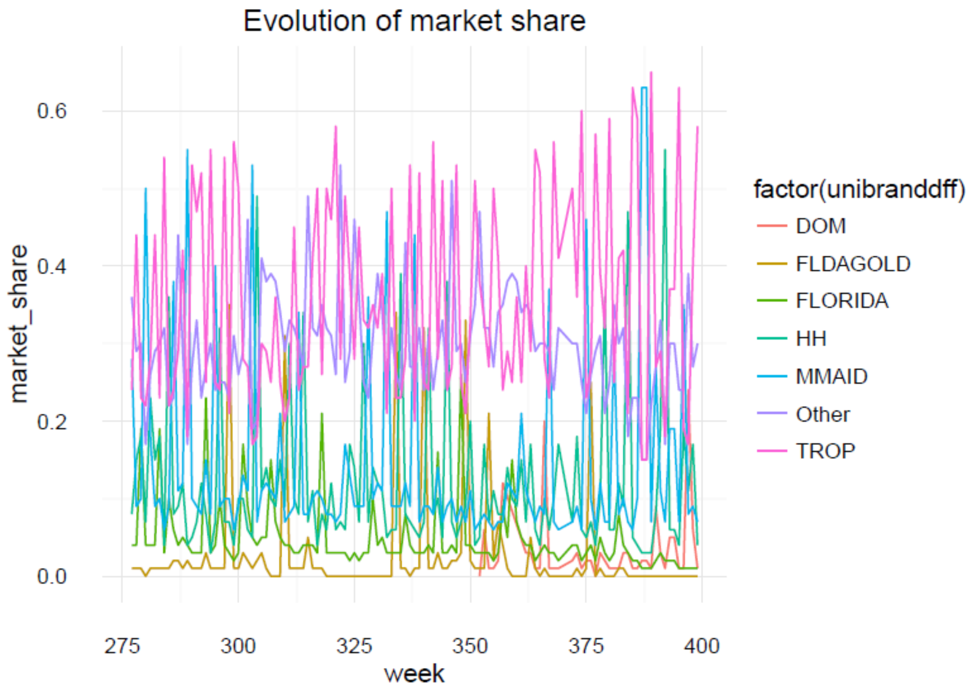


Note: Average price by brand



Note: Average promotion chance by brand





From the time series line, we could see there're seasonality and fluctuation for each brand in the whole period. However, the rank of market share is consistent- Trop and MMAID rank at first whereas HH and Florida rank at bottom. Same pattern for total sales.

Data Manipulation

After we understand the overall pattern for each brand, to use discrete choice model recognizing the relationship between choice and price, because we don't have customer panel data, we have to aggregate store level data.

Because in this assignment, we are only focusing on four major brands, Minute-Maid, Tropicana, HH and Florida, we added one more level to the brand column, 'No purc' to stands for all other purchases of all other brands. And we changed all the brands that are not these four major brands in to 'No purc'. This is also a preperation for the calculation of the s_{0wz}.

We calculated the price per oz and wholesale price of the sales data. Then we aggregate on the brand, week and zone level. For aggragation fuctions, we used mean on price per oz, resale price, promotion, wholesale price and wholesale price per oz. Here we are taking the average price of a brand in a week at s apecific zone as a regressor. The wholesale price and wholesale price per oz are used as Instrument variables in the Ivy-regression. We used sum for sales. This is for the calculation of the market share in the next step.

Then we calculated the s_{jwz}. We grouped on zone and week, calculating sum of sale of all the brands and calulated the market share of each individual brand. Here the brand No-purc's s_{jwz} is the s_{0wz} for the same week and zone.

After that, we generated the Huasman-type instruments, which is the average prices in other markets. Considering we have price and price per oz, we calculated the Huasman-type instrument for both price. The calulation is done by $\frac{sum(pricesinallzones)-price}{numofzones-1}$.

Finally, we assign all the s_{jwz} of the 'No purc' rows to the same brands s_{0wz} value that have same week and zone using a for loop.

Then we drop the intermediate columns generated during the process and all the 'No purc' row brands. And the dataframe is ready to use.

Model Building

We include different independent variable in our models to find the best model describing the relationship.

For OLS model (linear least model), we create six models with different variables.

1) OLS with price and promotion as independent variables.

2) OLS with price and promotion and brand dummies as independent variables.

3) OLS with price and promotion and brand and zone dummies as independent variables.

4) OLS with price per ounce and promotion as independent variables.

5) OLS with price per ounce and promotion and brand dummies as independent variables.

6) OLS with price per ounce and promotion and brand and zone dummies as independent variables.

Because in OLS model, it doesn't consider the effect that we might have omitted variables that is correlated with price and market share, we use IV model (Instrumental variable approach) to better describe the relationship between price and market share. For IV model, we create six models with different variables and wholesale price and Hausman instrument (average price in other markets) respectively.

Wholesale price

1) IV with price and promotion as independent variables and wholesale price as instrument for price.

2) OLS with price and promotion and brand dummies as independent variables and wholesale price as instrument for price.

3) OLS with price and promotion and brand and zone dummies as independent variables and wholesale price as instrument for price.

4) OLS with price per ounce and promotion as independent variables and wholesale price per ounce as instrument for price per ounce.

5) OLS with price per ounce and promotion and brand dummies as independent variables and wholesale price per ounce as instrument for price per ounce.

6) OLS with price per ounce and promotion and brand and zone dummies as independent variables and wholesale price per ounce as instrument for price per ounce.

Hausman instrument

1) IV with price and promotion as independent variables and Hausman as instrument for price.

2) OLS with price and promotion and brand dummies as independent variables and Hausman as instrument for price.

3) OLS with price and promotion and brand and zone dummies as independent variables and Hausman as instrument for price.

4) OLS with price per ounce and promotion as independent variables and Hausman per ounce as instrument for price per ounce.

5) OLS with price per ounce and promotion and brand dummies as independent variables and Hausman per ounce as instrument for price per ounce.

6) OLS with price per ounce and promotion and brand and zone dummies as independent variables and Hausman per ounce as instrument for price per ounce.

Model Result, Choice and Interpretation

Comparison of R^2

Bottle	**OLS**	**IV = wholesale price**	**IV2 = price of other stores**
Price + Promotion	0.045	-0.312	0.044
Price + Promotion + Brand	0.644	0.643	0.643
Price + Promotion + Brand + Zone	0.655	0.654	0.654

From the table above, based on price per bottle/box, we can clearly see that in all the 3 models (OLS and 2SLS models) when we only have price and promotion in the models, R-squared are lower than when we also have brand and zone dummy variables in the models. When we compare the two models, one adds only brand dummy variable, and another add two dummy variables, the R-squared of the models with 2 dummy variables (brand and zone) are relatively higher.

Therefore, in all 3 cases (OLS, 2SLS using wholesale price IV, 2SLS using price of other stores IV), Y ~ Price + Promotion + Brand + Zone model have the highest R-squared, which means this model fit more to the data.

Oz	**OLS**	**IV = wholesale cost**	**IV2 = price of other stores**
Price + Promotion	0.021	0.018	0.021
Price + Promotion + Brand	0.619	0.619	0.619
Price + Promotion + Brand + Zone	0.632	0.631	0.632

If we change the juice unit from per bottle/box to per oz, the results are the same. In all 3 cases (OLS, 2SLS using wholesale price IV, 2SLS using price of other stores IV), Y ~ Price + Promotion + Brand + Zone model have the highest R-squared, which means this model fit more to the data.

Next, we also compared the residual standard error for all the models.

Comparison Residual Standard Error

Bottle	**OLS**	**IV = wholesale cost**	**IV2 = price of other stores**
Price + Promotion	1.044	1.223	1.044
Price + Promotion + Brand	0.638	0.638	0.638
Price + Promotion + Brand + Zone	0.628	0.629	0.629

From the table above, based on price per bottle/box, we can clearly see that in all the 3 models (OLS and 2SLS models) when we only have price and promotion in the models, residual S.E. are higher than when we also have brand and zone dummy variables in the models. When we compare the two models with dummy variable, the residual S.E. of the models with 2 dummy variables (brand and zone) are relatively lower.

Therefore, in all 3 cases (OLS, 2SLS using wholesale price IV, 2SLS using price of other stores IV), Y ~ Price + Promotion + Brand + Zone model have the lowest residual S.E., which means this model fit the data better.

Oz	**OLS**	**IV = wholesale cost**	**IV2 = price of other stores**
Price + Promotion	1.057	1.059	1.057
Price + Promotion + Brand	0.659	0.659	0.659
Price + Promotion + Brand + Zone	0.649	0.649	0.649

If we change the juice unit from per bottle/box to per oz, the results are the same. In all 3 cases (OLS, 2SLS using wholesale price IV, 2SLS using price of other stores IV), Y ~ Price + Promotion + Brand + Zone model have the lowest residual S.E., which means this model fit better to the data.

To sum up, the conclusion we got from using R-squared as the measurement is consistent with the conclusion we got from using Residual S.E. as the measurement.

Y ~ Price + Promotion + Brand + Zone should be the best model in all 3 cases.

Here is the best model for OLS, IV1, IV2.

Table 7: The Best Price models

	Dependent variable:		
	log(sjwz/s0wz, base = exp(1))		
	OLS	*instrumental variable*	
	(1)	(2)	(3)
price	−0.665*** (0.030)	−0.535*** (0.052)	−0.703*** (0.033)
promotion	0.662*** (0.029)	0.722*** (0.035)	0.645*** (0.029)
unibranddffHH	1.033*** (0.021)	1.039*** (0.021)	1.031*** (0.021)
unibranddffMMAID	1.415*** (0.022)	1.422*** (0.022)	1.413*** (0.022)
unibranddffTROP	2.367*** (0.021)	2.372*** (0.021)	2.366*** (0.021)
zone2	0.002 (0.040)	0.011 (0.041)	−0.001 (0.040)
zone3	0.246*** (0.040)	0.248*** (0.040)	0.246*** (0.040)
zone4	−0.004 (0.040)	0.006 (0.041)	−0.007 (0.040)
zone5	−0.082** (0.040)	−0.072* (0.041)	−0.085** (0.040)
zone6	0.001 (0.041)	0.033 (0.043)	−0.009 (0.041)
zone7	0.050 (0.040)	0.047 (0.040)	0.051 (0.040)
zone8	0.003 (0.040)	0.014 (0.041)	−0.0004 (0.040)
zone10	−0.116*** (0.040)	−0.119*** (0.040)	−0.116*** (0.040)
zone11	−0.197*** (0.040)	−0.198*** (0.040)	−0.196*** (0.040)
zone12	−0.033 (0.040)	−0.021 (0.041)	−0.036 (0.041)
zone13	−0.066 (0.040)	−0.055 (0.041)	−0.070* (0.041)
zone14	0.140*** (0.040)	0.150*** (0.041)	0.137*** (0.040)
zone15	−0.197*** (0.041)	−0.180*** (0.041)	−0.202*** (0.041)
zone16	−0.095** (0.042)	−0.056 (0.044)	−0.107** (0.042)
Constant	−0.708*** (0.094)	−1.085*** (0.155)	−0.599*** (0.101)
PE Florida	-1.67	-1.34	-1.76
PE HH	-1.51	-1.21	-1.60
PE MMAID	-1.53	-1.23	-1.61
PE Trop	-1.16	-0.93	-1.22
Instrument variable	NA	wholesale price	hausman
Partial Fscore	NA	2765.410	27363.366
Observations	7,260	7,260	7,260
R^2	0.655	0.654	0.654
Adjusted R^2	0.654	0.653	0.654
Residual Std. Error (df = 7240)	0.628	0.629	0.629
F Statistic	721.953*** (df = 19; 7240)		

Note: *p<0.1; **p<0.05; ***p<0.01

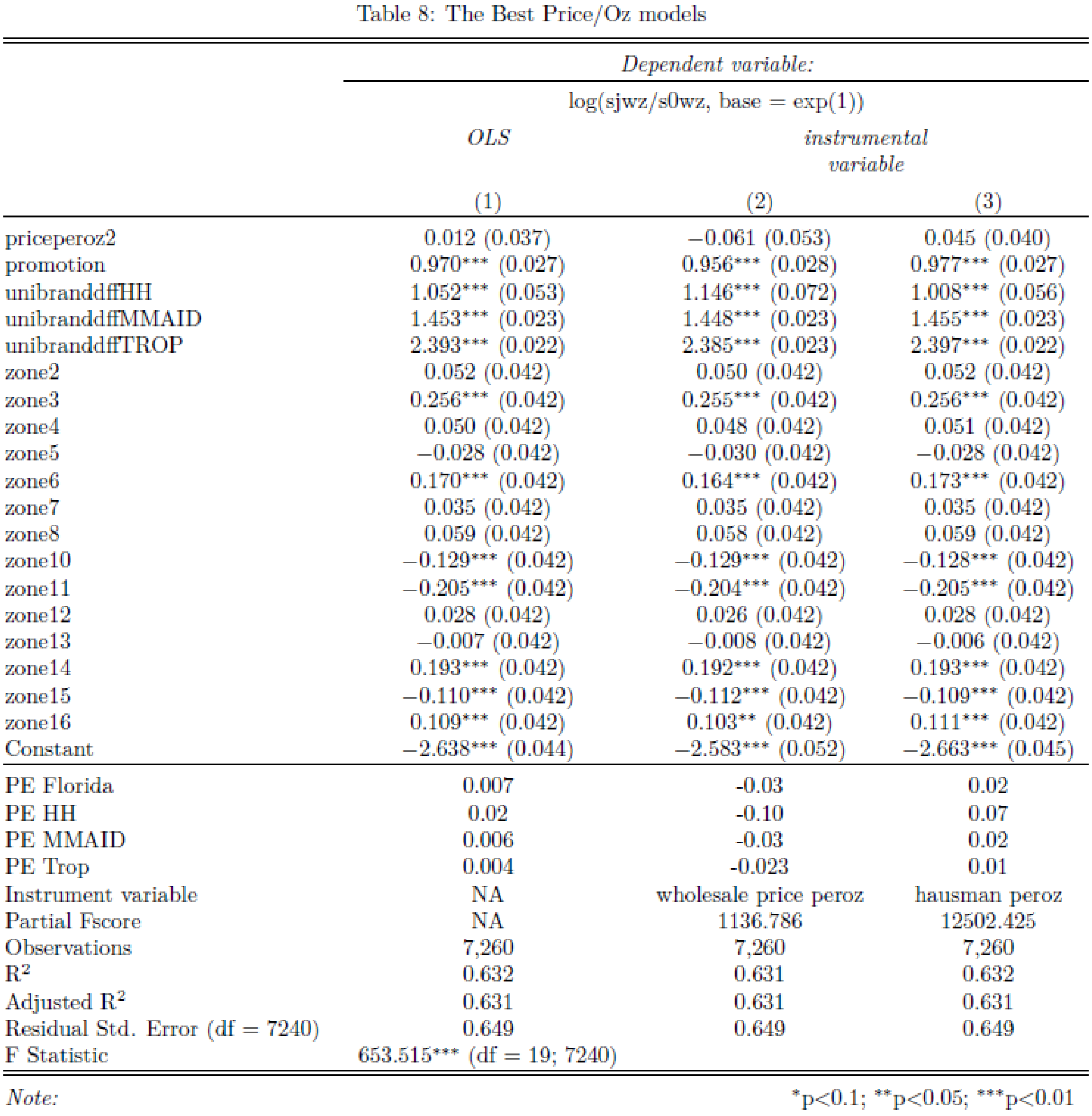


Table 8: The Best Price/Oz models

	Dependent variable:		
	log(sjwz/s0wz, base = exp(1))		
	OLS	*instrumental variable*	
	(1)	(2)	(3)
priceperoz2	0.012 (0.037)	−0.061 (0.053)	0.045 (0.040)
promotion	0.970*** (0.027)	0.956*** (0.028)	0.977*** (0.027)
unibranddffHH	1.052*** (0.053)	1.146*** (0.072)	1.008*** (0.056)
unibranddffMMAID	1.453*** (0.023)	1.448*** (0.023)	1.455*** (0.023)
unibranddffTROP	2.393*** (0.022)	2.385*** (0.023)	2.397*** (0.022)
zone2	0.052 (0.042)	0.050 (0.042)	0.052 (0.042)
zone3	0.256*** (0.042)	0.255*** (0.042)	0.256*** (0.042)
zone4	0.050 (0.042)	0.048 (0.042)	0.051 (0.042)
zone5	−0.028 (0.042)	−0.030 (0.042)	−0.028 (0.042)
zone6	0.170*** (0.042)	0.164*** (0.042)	0.173*** (0.042)
zone7	0.035 (0.042)	0.035 (0.042)	0.035 (0.042)
zone8	0.059 (0.042)	0.058 (0.042)	0.059 (0.042)
zone10	−0.129*** (0.042)	−0.129*** (0.042)	−0.128*** (0.042)
zone11	−0.205*** (0.042)	−0.204*** (0.042)	−0.205*** (0.042)
zone12	0.028 (0.042)	0.026 (0.042)	0.028 (0.042)
zone13	−0.007 (0.042)	−0.008 (0.042)	−0.006 (0.042)
zone14	0.193*** (0.042)	0.192*** (0.042)	0.193*** (0.042)
zone15	−0.110*** (0.042)	−0.112*** (0.042)	−0.109*** (0.042)
zone16	0.109*** (0.042)	0.103** (0.042)	0.111*** (0.042)
Constant	−2.638*** (0.044)	−2.583*** (0.052)	−2.663*** (0.045)
PE Florida	0.007	-0.03	0.02
PE HH	0.02	-0.10	0.07
PE MMAID	0.006	-0.03	0.02
PE Trop	0.004	-0.023	0.01
Instrument variable	NA	wholesale price peroz	hausman peroz
Partial Fscore	NA	1136.786	12502.425
Observations	7,260	7,260	7,260
R^2	0.632	0.631	0.632
Adjusted R^2	0.631	0.631	0.631
Residual Std. Error (df = 7240)	0.649	0.649	0.649
F Statistic	653.515*** (df = 19; 7240)		
Note:			*p<0.1; **p<0.05; ***p<0.01

Comparison cross OLS and 2SLS with different IVs

If we compare the best model cross OLS, 2SLS (with wholesale price as IV), 2SLS (with price of other stores as IV), even though their R-Squared and residual S.E. looks similar, but we can see the impact of introducing IV through Partial F-score. For example, if we look at the models using price per bottle/box, for the first 2SLS model, the best model has partial Fscore equals to 2765.41, and the second 2SLS model has partial Fscore of 27363.366. In both cases, partial Fscore is larger than 10, which means that IVs are having effect on directly on price, and only having indirect impacts on customers' odds to buy these brands' juice through its impact on price.

Therefore, after picking out the best model in 3 cases, which are all the model with both dummy variables for brand and zone. The two 2SLS models would explain data the best.

Interpretation of the best models

OLS model:

If we look at price per bottle/box, the impact of price on change of odds that customers choosing these 4 brands is significant. When we hold the brand, zone and week information fixed, with price increase by $1 for each bottle/box of juice, the odds that customers choose to buy these 4 juice brands would decrease by 0.514 (e^-0.665).

However, if we look at price per oz, the impact of price on change of odds that customers choosing these 4 brands is not significant.

As for own-price elasticity for the OLS model:

For a 1% change in price of brand Florida, there would be 1.672% change of the market share for brand Florida; For a 1% change in price of brand hh, there would be 1.517% change of the market share for brand hh; For a 1% change in price of brand mmaid, there would be 1.531% change of the market share for brand mmaid; For a 1% change in price of brand trop, there would be 1.161% change of the market share for brand trop.

Using wholesale cost as the IV in the 2SLS model:

If we look at price per bottle/box, the impact of price on change of odds that customers choosing these 4 brands is significant. When we hold the brand, zone and week information fixed, with price increase by $1 for each bottle/box of juice, the odds that customers choose to buy these 4 juice brands would decrease by 0.586 (e^-0.535).

However, if we look at price per oz, the impact of price on change of odds that customers choosing these 4 brands is also not significant under 2SLS model.

As for own-price elasticity for this model:

For a 1% change in price of brand Florida, there would be 1.344% change of the market share for brand Florida; For a 1% change in price of brand hh, there would be 1.219% change of the market share for brand hh; For a 1% change in price of brand mmaid, there would be 1.231% change of the market share for brand mmaid; For a 1% change in price of brand trop, there would be 0.934% change of the market share for brand trop.

Using prices of other stores as the IV in the 2SLS model:

If we look at price per bottle/box, the impact of price on change of odds that customers choosing these 4 brands is significant. When we hold the brand, zone and week information fixed, with price increase by $1 for each bottle/box of juice, the odds that customers choose to buy these 4 juice brands would decrease by 0.495 (e^-0.703).

Same as before, if we look at price per oz, the impact of price on change of odds that customers choosing these 4 brands is not significant under 2SLS model even if we change to another IV.

As for own-price elasticity for this model:

For a 1% change in price of brand Florida, there would be 1.767% change of the market share for brand Florida; For a 1% change in price of brand hh, there would be 1.603% change of the market share for brand hh; For a 1% change in price of brand mmaid, there would be 1.618% change of the market share for brand mmaid; For a 1% change in price of brand trop, there would be 1.227% change of the market share for brand trop.

Comparison with other research papers

Comparing the parameters and elasticities:

- The beer paper [1] starts with an extensive analysis of the earlier paper on the same topic and discusses how price elasticities have already been established for the alcoholic beverages market. They also acknowledge the claim that research has been conducted to show the pattern of beer buying among people under 21 years of age. The authors then proceed with identifying gaps such as there can still be unobservable product characteristics and the explanatory variables may not be all exogenous to the model, hence this is one problem which can be solved.

- We have also used a similar approach to analysis. We first assume certain variables which can explain the variability in product prices (and prices/oz) and later on we move to the IV analysis to find out the effect of unobserved variables. We can refine this further by studying if we have included any kind of bias by using information from the OLS models into the IV models

- When we compare the parameters, we can see that the explaining power of the model (R sq) increases when they use a MNL and NL model replace of the base OLS model. We have replace used different variations of the OLS model, each of which differs by the number of independent variables. Some of the parameters to compare are the improvement in R sq values - The paper improves the R sq value from 0.2 to 0.73. However, we had a drastic improvement from 0.045 to 0.655 by including the brand variable. Brand had a lot of impact in increasing the R sq but including the zone did not. In the author' paper also zone has a very low coefficient indicating very similar results to our analysis

- The elasticity values in the paper are only for the `Mass`, `Craft` and `Import` variables where as we have elasticities for different Brands. It is not advisable to compare them and we can look at a different paper

- FMCG paper [2] also talks about the problem with too many explanatory variables and how to deal with them -``there can be a large number of competing products at the UPC level: a typical product category such as Soft Drinks may contain hundreds of items of different flavours, package sizes, and brands which are all competitors with each other because they satisfy similar customer needs and wants''.

- This results in too many competitive explanatory variables. Hence most models go towards overfitting making any kind of predictive analysis futile. We should always aim towards the optimal set of variables, we can use many feature selection algorithms to get the features with the most amount of varience or the most explanatory power.

Comparing the data cleaning process:

- The FMCG paper starts the analysis by first developing statistics from cross tables of categorical data and then a summary table for the continuous variables of the data. The authors also defined a few variables -``For example, we choose 34 products from the Bottled Juice category. On average, these 34 products are being sold on promotion for 42 weeks during the 200 weeks considered (i.e. an intensity of 0.21, with a standard deviation of 0.09). ''.
- This becomes quite useful as the analysis can now be more focused once we get an idea of how the promotion of certain products is in the data. ``Take the Bottle Juice category as an example, the promotions in this category increase the sales of focal product by 169% on average compared to the baseline predicted sales assuming there was no promotion. ''
- This analysis can lead us to a discovery of a different range of promotions across products which me might have missed otherwise. The authors have finally used a range of statistical tests to ensure the validity of their assumptions - for example the Dickey Fuller test for stationary time series. We can also use t-tests or prop test to find statistical difference (if any exists) across promotions (discreet variable) or sales (continuous variable). This further helps us to improve the OLS models.

Comparing the difference in data:

- The data used in the FMCG paper is very similar to the data in this analysis. However, the objective is quite different. The former is mostly concerned about the accuracy of time series model predictions

- Behavioral Decision Making paper[3] focuses on how the change in packaging could have influenced the purchase behavior among customers. We could also have done a similar analysis which looks for change in pack sizes (or availability of pack sizes) for the same product across time. This can help us in creating a flag variable which may indicate the change in product size in the life time of the product

- The Rigidities paper [4] analyses a different dataset (2006-2008) but uses the Dominick's data as a validation dataset and have then compared the share of price and time spent across the primary and validation dataset.

References

[1] Daniel Toro-González, Jill J. McCluskey and Ron C. Mittelhammer [Beer Snobs do Exist: Estimation of Beer Demand by Type.] Journal of Agricultural and Resource Economics (JARE) (2014).

[2] Huang, Tao, Robert Fildes, and Didier Soopramanien. [The Value of Competitive Information in Forecasting FMCG Retail Product Sales and the Variable Selection Problem.] European Journal of Operational Research (2014).

[3] Jami, Ata and Himanshu Mishra. [Downsizing and Supersizing: How Changes in Product Attributes Influence Consumer Preferences] Journal of Behavioral Decision Making (2013).

[4] Eichenbaum, Martin, Nir Jaimovich and Sergio Rebelo. [Reference Prices, Costs, and Nominal Rigidities.] American Economic Review 101.1 (2011): 234-62.