Industrial Organization and Data Science

Instructors: Jacob LaRiviere, Affiliate Professor & Senior Principal Researcher, Microsoft

Emails: [jlariv@microsoft.com](mailto:jlariv@microsoft.com)

# Course Assignments & Reading

Course assignments should be printed (code, output and descriptive answers) and turned in by the start of class on canvas unless otherwise noted. Feel free to work in groups but everyone is required to turn in their own work with answers written in your own words. In both calculations and complex ideas, write down each step of logic used in reaching your conclusion. Keep in mind that in most cases a good answer is one precise sentence; quality is heavily favored over quantity. This will be graded on a full credit, half credit and no credit basis.

Discussion questions do not need to be written out ahead of time. At the beginning of each class the professors will lead a discussion around these questions. Students will be called on, potentially at random, to add their insight. This part of class will contribute heavily to your course participation grade.

**Week 2, due Oct 13**

**Optional:** [Varian’s notes](http://firstmonday.org/ojs/index.php/fm/article/view/473/394) on price discrimination

**Optional:** McAfee Ch. 3 (understanding each equation not needed).

**Optional:** Hoch, Stephen J., et al. "[Determinants of store-level price elasticity](https://research.chicagobooth.edu/marketing/databases/dominicks/docs/1995_Determinants_of_Store-Level_Price_Elasticity.pdf)." *Journal of Marketing Research* (1995): 17-29. *NOTE: This is the same dataset we’ll be using.*

**Optional:** Ch 2-5.3 in [Hermalin’s lecture notes](http://faculty.haas.berkeley.edu/hermalin/LectureNotes201b_v5.pdf)

**Assignment to be turned in.** Please turn in your typed-up theory answers and use Rmarkdown for the R script/output.

Note: you will probably not know all the relevant commands of the top of your head. Simply search “command in R” or “command in R examples” etc. in a search engine, and this will almost always give the answer.

**Theory Section**

1. Assume that you are the new manager of a firm selling a good. Historically, your firm has charged a single price for every unit sold as opposed to price discriminate.
   1. Consider a strategy of using two-part tariffs (2nd degree price discrimination) to sell. Say that every consumer had quantity demanded of 10 units at a price of $2.
      1. If you could charge consumers an access fee for the right to purchase your good, do you have enough information to say what the lowest access fee consumers would be willing to pay?

*The lowest access fee would be ~~$2.~~ $0*

* + 1. What would the demand curve look like to dictate such an access fee?

*The demand curve would ~~shift down by $2~~. be flat.*

* + 1. In what situation could you charge your customers a higher access fee: when they have elastic demand curves or inelastic demand curves? In one sentence and/or a picture, why?

*If they have a more inelastic demand curve, we could charge more become the demand responds less to a change in price if it is inelastic, meaning that customers are less price sensitive.*

* 1. Now assume the firm sells a homogenous good with lots of competition. (A homogenous good is one where there is lots of perfect- or almost perfect- substitutes. Examples include gasoline, conventional potatoes, and filtered bottled water).
     1. What would happen if you charged a two-part tariff now?

*We would lose all customers because they would go to another firm to purchase the same good without the tariff.*

* + 1. Assume that due to your customers’ lack of information, only half of them knew there were other firms selling your good.
       1. Under what circumstances would it still be optimal to use the two-part tariff pricing structure?

*It would be optimal to still use the two-part tariff structure only if the additional customer surplus that we could captured would be greater than the revenue loss from having only half of the customers.*

* + - * 1. If all your consumers had inelastic demand curve, would you be more or likely to go forward with the two-part tariff? In one sentence, why?

*I would be more likely to go forward with the two-part tariff because inelastic demand means that customers would be less price sensitive thus would mind less if we charge with a tariff.*

* + - * 1. What if the informed consumers had more inelastic demand than the uninformed consumers? In one sentence, why?

*~~Then it would be optimal to use a two-part tariff because the informed consumer would be less likely to leave for another seller due to their inelastic demand and we can retain more customer.~~*

*Less optimal because uniformed consumer has less consumer surplus*

**Empirical Section**

1. Download the orange juice data from the course website and create an R script for this assignment.
2. Change the working directory so that R knows where to look for the data (tip: create a Econ487 folder and save datasets there). See setwd(). [You can type ?setwd to see the help file.]
3. Read in the data, see read.csv. oj is a data frame with many variables. You can click on the dataframe in the top right corner of Rstudio to explore. You can refer to any variable with oj$var\_name where “var\_name” is the variable of interest. We will also refer to df as a generic term for a “dataframe”
4. Visualizing price.
   1. Make a box plot of price.
      1. Use the ggplot2 package to do this. ggplot2 is kind of quirky but powerful package. You’ll need to start by calling the package once you’ve installed it:

library(ggplot2)

ggplot(df, aes(factor(var\_name1), var\_name2)) + geom\_boxplot(aes(fill = factor(brand)))

The first line above calls the ggplot and tells it to use the dataframe df.

aes is short for “aesthetics”

the term factor(var\_name1) tells it to create a unique plot by each unique value in var\_name1.

the second variable listed var\_name2 tells it to use that variable in creating the boxplot.

The second part of the line + geom\_boxplot(aes(fill = factor(var\_name1))) tells it to make a boxplot and color each one by var\_name1.

* 1. Make a box plot of log price.
  2. Make a box plot of price, but separate out each brand.
  3. Do the same for log price.
  4. **What do these graphs tell you about the variation in price? Why do the log plots look different? Do you find them more/less informative?**

*The mean of all the orange juice prices is around $2.2, while Tropicana has a higher average price of $3 and a higher variance of prices, Minute Maid in the middle having $2.2, and Dominick’s having a lower price of around $1.6. The log plots make skewed data more even, therefore, they show more outliers in both graphs, which is more informative if we want to target on studying the price changes like discounts.*

1. Visualizing the quantity/price relationship
   1. Plot logmove (log quantity) vs. log(price) for each brand. For this one the appropriate second part of the ggplot command will be: + geom\_point(aes(color = factor(var\_name)))
      1. **What do insights can you derive that were not apparent before?**

*This graph shows information about the change in quantity sold in accordance with the change in price, and we can tell that amongst all three brands, shoppers who buy Dominick’s is the most price-sensitive – when they are having a discount, the quantity sold increases the most.*

1. Estimating the relationship.
   1. Do a regression of log quantity on *log price*. **How well does the model fit? What is the elasticity, does it make sense?**

*The model fits well, with a p-value < 0.001. The elasticity is -1.6 which make sense because the lower the price, the higher the quantity sold.*

* 1. Now add in an intercept term for each brand (add brand to the regression), **how do the results change? How should we interpret these coefficients?**

*Two more statically significant coefficient was added, the brand of minute maid and tropicana~~, each with a positive elasticity of 0.87 and 1.53 respectively.~~ This means that the brand being Tropicana leads to significantly higher quantity of sales, and Minute Maid also correlates to a slightly more sales. ~~Dominick’s, on the other hand, has no correlation with the change of quantity of sales.~~ The elasticity of price also dropped to -3.14, showing a more accurate correlation between price discount and the increase of quantity sold.*

*Dominick’s is the base line. For the same price, Minute maid would have a greater quantity of sales, and the same for tropicana. Here, the elasticity is the same across brands (because we didn’t allow it to change) is -2.53.*

* 1. Now figure out a way to allow the elasticities to differ by brand. Search “interaction terms” and “dummy variables” if you don’t remember this from econometrics. Note the estimate coefficients will “offset” the base estimates. **What is the insights we get from this regression? What is the elasticity for each firm? Do the elasticities make sense?**
     1. *Now, we can see the price elasticities of each brand by looking at the interaction terms. The price elasticity of Tropicana is -2.13, Minute Maid is -2.65 and Dominick’s is -3.92. This make sense because for each unit change of price, the sales of Dominick’s is being impact the most because its customers are the most price-sensitive.*

1. Impact of “featuring in store”. The “feat” variable is an indicator variable which takes the value of one when a product is featured (e.g., like on [an endcap display](https://easyshiftapp.zendesk.com/hc/en-us/articles/206348625-Endcap-Display))
   1. Which brand is featured the most? **Make a ggplot to show this**. Hint: using position = "jitter", within the aes(color = factor(var\_name)) of ggplot is one way to do this.

*Minute Maid and Dominick’s are featured the most.*

* + 1. What is the average price and featured rate of each brand? Hint:

aggregate(df[, x:y], list(df$var\_name), mean) where x and y are the column numbers of the two variables you care about.

See if you can do this with the dplyr or ddplyr package.

|  |
| --- |
|  |
| **Brand** | **Average feat rate** | **Average price** |  |  |
| dominicks | 0.2570215 | 1.735809 |  |  |
| minute.maid | 0.2885273 | 2.241162 |  |  |
| tropicana | 0.1662348 | 2.870493 |  |  |

* 1. How should incorporate the feature variable into our regression? Start with an additive formulation (e.g. feature impacts sales, but not through price).
  2. Now run a model where features can impact sales and price sensitivity (e.g., the model we discussed in class).
  3. Now run a model where each brand can have a different impact of being featured and a different impact on price sensitivity.

**Produce the regression results for this regression brand with brand level elasticities.**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 10.53479 0.01868 563.919 < 2e-16 \*\*\*

feat 1.33948 0.03188 42.020 < 2e-16 \*\*\*

log(price):branddominicks -2.97375 0.03220 -92.350 < 2e-16 \*\*\*

log(price):brandminute.maid -2.08461 0.02375 -87.772 < 2e-16 \*\*\*

log(price):brandtropicana -1.51291 0.01841 -82.191 < 2e-16 \*\*\*

log(price):branddominicks:feat -0.96414 0.06492 -14.851 < 2e-16 \*\*\*

log(price):brandminute.maid:feat -0.32409 0.04661 -6.953 3.65e-12 \*\*\*

log(price):brandtropicana:feat -0.73270 0.03966 -18.474 < 2e-16 \*\*\*

* 1. Now add what you think are the most relevant sociodemographic controls and **produce the regression results from that regression as well.**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 12.44679 0.16429 75.761 < 2e-16 \*\*\*

feat 1.30513 0.03131 41.680 < 2e-16 \*\*\*

INCOME -0.18064 0.01471 -12.277 < 2e-16 \*\*\*

HHLARGE -1.80132 0.14381 -12.526 < 2e-16 \*\*\*

AGE60 1.47739 0.07068 20.902 < 2e-16 \*\*\*

log(price):branddominicks -3.04421 0.03169 -96.063 < 2e-16 \*\*\*

log(price):brandminute.maid -2.13589 0.02337 -91.388 < 2e-16 \*\*\*

log(price):brandtropicana -1.55275 0.01811 -85.721 < 2e-16 \*\*\*

log(price):branddominicks:feat -0.91045 0.06376 -14.280 < 2e-16 \*\*\*

log(price):brandminute.maid:feat -0.28421 0.04578 -6.208 5.43e-10 \*\*\*

log(price):brandtropicana:feat -0.70173 0.03895 -18.016 < 2e-16 \*\*\*

1. Overall analysis
   1. **Based on your work, which brand has the most elastic demand, which as the least elastic?** 
      1. *Dominick’s has the most elastic demand, and Tropicana has the least elastic.*
   2. **Do the average prices of each good match up with these insights?**
      1. *Yes, the most price-elastic brand has the lowest price among the three, and vice versa.*
   3. **Take average prices for each brand. Use the elasticity pricing formula (you can use average values from your analysis above) to “back out” unit costs for each brand. Do the unit costs appear to be the same or different? What are your insights/reactions?**

*The average price for Tropicana is around $3, for Minute Maid, around $2.2, and for Dominick’s around $1.6. The unit costs appear to be quite similar, with Dominick’s on a lower end, which make sense to me because orange juice is a commodity that wouldn’t have too much of a difference in its unit cost, especially when it is being mass-produced.*

*Elasticity pricing formula: (P – MC) / P = 1 / |elasticity|*