Industrial Organization and Data Science

Instructors: Jacob LaRiviere, Lukas Hager

Emails: lghhager@uw.edu

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. **Work must be submitted on Canvas as a .html file!**

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 9

Optional: Varian's notes on price discrimination

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

Optional: Hoch, Stephen J., et al. "Determinants of store-level price elasticity." *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

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.

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.
    a. Make a box plot of price.

i. 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 = 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`.

b. Make a box plot of log price.
c. Make a box plot of price, but separate out each brand.
d. Do the same for log price.
e. **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?**

5) Visualizing the quantity/price relationship
   a. 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)))`
      i. **What do insights can you derive that were not apparent before?**
6) Estimating the relationship.
   a. Do a regression of log quantity on *log price* (you can use the lm or glm function to do this). **How well does the model fit? What is the elasticity, does it make sense?**
   b. 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?**
   c. 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?**
7) 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)
   a. What is the average price and featured rate of each brand? Hint: use group_by and summarise within dplyr.
   b. How should incorporate the feature variable into our regression? Start with an additive formulation (e.g. feature impacts sales, but not through price).
   c. Now run a model where features can impact sales and price sensitivity (e.g., the model we discussed in class).
   d. 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.

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

8) Overall analysis
   a. **Based on your work, which brand has the most elastic demand, which as the least elastic?**
   b. **Do the average prices of each good match up with these insights?**
   c. **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?**
9) Let's return to the orange juice assignment and investigate how store demographics are related to demand.
   a. Take one of the final models from (7) and add in the store demographics as linear features (e.g. + demo1 + demo2). Report your output.
   b. What demographics significantly (t-value>2) influence demand?
   c. Use the `predict` command to determine how well the model predicts `logmove` and create a new variable called `logmove_hat`. To do so construct the "fair r^2" covered in class. What is the improvement relative to the model without the demographic features?
   d. Rather than using fair r^2 lets now use a test set to determine which model gives the best out of sample prediction.
      i. Create a new dataframe which is a random subset of 80% of the data (look at `sample_n` from the `dplyr` package).
      ii. Estimate the model with and without demographic characteristics. Construct MSE for the training and test set for the models.
      iii. Compare the out of sample MSE for the models. Which is lower implying the model does a better job of fitting the data?