Data Science for Game Theory and Pricing

Instructor: Jacob LaRiviere, Director of Economics and Data Science at Amazon

Emails: [jlarvi1@uw.edu](mailto:jlarvi1@uw.edu), [lghhager@uw.edu](mailto:lghhager@uw.edu) (TA)

# Course Assignments & Reading

Course assignments R Markdown files (.html or .rmd) and turned in at 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. 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. All work must be typed.

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 3, due Oct 19**

**Assignment to be turned in.** Please turn in your Rmarkdown output and answers to the questions typed up and turned in on canvas. Make sure you do the theory portion of the assignment as well.

1. Let’s return to the orange juice assignment and investigate how store demographics are related to demand.
   1. Take one of the final models from HW2 (logmove ~ log(price)\*brand\*feat) and add in the store demographics as linear features (e.g. + demo1 + demo2). Report your output (past into your answer document).
   2. What demographics significantly (t-value>2) influence demand?
   3. 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 r2” covered in class. What is the improvement relative to the model without the demographic features?
   4. Rather than using fair r2 lets now use a test set to determine which model gives the best out of sample prediction.
      1. Create a new dataframe which is a random subset of 80% of the data (look at sample\_n from the dplyr package).
      2. Estimate the model with and without demographic characteristics. Construct MSE for the training and test set for the models.
      3. Compare the out of sample MSE for the models. Which is lower implying the model does a better job of fitting the data?
2. Let’s focus on two variables HHLARGE (“fraction of households that are large”) and EDUC (“fraction of shoppers with advanced education”).
   1. What are the means and percentiles of each of these variables?

**HINT:** summary(oj$EDUC)

* 1. Using your coefficient estimates from the regression in 1b:
     1. If we move from the median value of HHLARGE to the 75th percentile (3rd quartile), how much does log(quantity) change each week on average?

**HINT:** using coef(reg\_output)["var\_name"] exports the coefficient on “var\_name” from the regression model “reg\_output”.

Similarly, summary(df$var\_name) will output a bunch of summary statistics for the variable var\_name in data frame df. Using summary(df$var\_name)["3rd Qu."] will take the level of the 3rd quantile from the summary of var\_name.

Note: if we wanted to assess the changes in levels, you’d want to take the exponent of everything.

* + 1. If we move from the median value of EDUC to the 75th percentile (3rd quartile), how much does log(quantity) change each week on average?
    2. Base on this analysis, which is the more important predictor of demand?
  1. Now let’s see if these variables impact price sensitivity. Add two interaction terms (with logprice) to the model to test this.
     1. What are the coefficients on the interaction terms?
     2. Does the sign of your estimates make sense based on your intuition?
     3. What are the coefficient estimates on the variables EDUC and HHLARGE that aren’t part of the interaction term? How do they compare to your regression from 1b?
     4. Similar to 2b, if we move from the median value of each variable to the 3rd quartile, how much does elasticity change? Based on this, which is more important to price sensitivity?
  2. You should notice that the coefficients on EDUC and HHLARGE have flipped sign once we include interaction terms with price. HHLARGE now appears to be a positive demand shifter and increases price sensitivity. Explain in words or pictures what is going on.

1. Create make a new dataframe which takes the previous week’s prices as a variable on the same line as the current week. This would enable you to see if there is *intertemporal* substitution.
   1. There are going to be a couple of steps. First is creating a new dataframe which is like the old one except that the week variable will change by a single week
      1. Df1 <- oj
      2. Df1$week <- Df1$week+1
         1. This will replace week with week+1
      3. The next step will use the merge function.
         1. Df2 <- merge(oj, df1, by=c("brand","store","week"))
         2. Investigate the Df2 and rename the lagged store values needed for a lagged price within the same store
   2. Now run a regression with this week’s log(quantity) on current and last week’s price.
   3. What do you notice about the previous week’s elasticity? Does this make sales more or less attractive from a profit maximization perspective? Why?

BONUS: Do (3) but using the same week’s prices of *other* brand’s OJ in the same regression. Here’s a hint from base R: dcast(oj\_prices, store + week ~ brand). See if you can do the same but with dplyr or another R package. Try doing this for only a single brand of orange juice as a first step.

Theoretical Questions:

1. Go back to lecture 2’s slides on Value Based Pricing. List each type of value-based pricing (e.g., 2-part tariffs, bundling, etc.).
   1. Give a one sentence definition of each.
   2. Try to think of one practical problem with implementing each type of value based pricing. This could be either competitive (e.g., competing firms) or information deficiencies.
2. Assume that in addition to orange juice, you also observe demand for bananas.
   1. What regression would you run to determine if bananas and orange juice are compliments or substitutes? What is the coefficient of interest (i.e. on what variable) that would inform you?
   2. Assume you find they are substitutes. What would the sign of the coefficient be? Would you be more or less likely to bundle these products if they are substitutes?
      1. Explain why with an equation, figure or a sentence or two.
      2. Would the price of the bundle be less than or more than the sum of the two independent prices? (Not a trick question; verifying you understand bundles.)
   3. During a sale for orange juice, should you continue to offer the bundle? Why or why not? HINT: who is price sensitive for orange juice? Who comes into market? Would you want to offer the bundle at a lower price than before?