Data Science for Game Theory and Pricing

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

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

# Course Assignments & Reading

Course assignments should be printed and stapled Rmarkdown files and turned in at the start of class 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 20**

**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 the final model 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).

Call:

glm(formula = logmove ~ log(price) \* brand \* feat + AGE60 + EDUC +

ETHNIC + INCOME + HHLARGE + WORKWOM + HVAL150, data = oj)

Deviance Residuals:

Min 1Q Median 3Q Max

-5.0461 -0.3990 -0.0028 0.3892 3.0766

Coefficients:

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

(Intercept) 11.96596 0.31077 38.504 < 2e-16 \*\*\*

log(price) -2.88199 0.03740 -77.057 < 2e-16 \*\*\*

brandminute.maid 0.13369 0.04487 2.980 0.00289 \*\*

brandtropicana 0.81040 0.04888 16.580 < 2e-16 \*\*\*

feat 1.04612 0.03664 28.553 < 2e-16 \*\*\*

AGE60 1.52744 0.12570 12.151 < 2e-16 \*\*\*

EDUC 0.47262 0.09922 4.763 1.91e-06 \*\*\*

ETHNIC 0.58163 0.03579 16.251 < 2e-16 \*\*\*

INCOME -0.13953 0.03110 -4.487 7.27e-06 \*\*\*

HHLARGE -1.18884 0.23164 -5.132 2.88e-07 \*\*\*

WORKWOM -1.33364 0.14165 -9.415 < 2e-16 \*\*\*

HVAL150 0.40076 0.03996 10.029 < 2e-16 \*\*\*

log(price):brandminute.maid 0.71365 0.05904 12.087 < 2e-16 \*\*\*

log(price):brandtropicana 0.69152 0.05463 12.657 < 2e-16 \*\*\*

log(price):feat -0.39170 0.07123 -5.499 3.85e-08 \*\*\*

brandminute.maid:feat 1.09687 0.07879 13.922 < 2e-16 \*\*\*

brandtropicana:feat 0.72157 0.09492 7.602 3.00e-14 \*\*\*

log(price):brandminute.maid:feat -1.04438 0.11750 -8.889 < 2e-16 \*\*\*

log(price):brandtropicana:feat -0.97440 0.11928 -8.169 3.23e-16 \*\*\*

* 1. What demographics significantly (t-value>2) influence demand?

***All demographics of the stores here are significant, including Age, Education, Ethnic, Income, Household size, Working Women, and Household income.***

* 1. 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?

***The fair R-squared increased from 0.52 to 0.55.***

* 1. 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?

***The model with demographic characteristics did a better job of fitting the data because its MSE is 0.45 in comparison to 0.49.***

1. 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)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.01351 0.09794 0.11122 0.11560 0.13517 0.21635

* 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.

***The log(quantity) would increase by 0.97 each week on average.***

* + 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?

***The log(quantity) would increase by 1.01 each week on average.***

* + 1. Base on this analysis, which is the more important predictor of demand?

***Based on this analysis, education is a more important predictor of demand because it would have the greater influence on quantity sold.***

* 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?

***The coefficient for HHLARGE is -5.15, and that for EDUC is 3.55.***

* + 1. Does the sign of your estimates make sense based on your intuition?

***The signs make sense to me because the larger the household, the more price-sensitive people would be since they would be buying in large quantity. And for education, those who have a college degree are more likely to have greater income and thus would be less price sensitive.***

* + 1. 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?

***The coefficient estimates of HHLARGE is now 2.63 and of EDUC is -2.60. Compared to 1b, they used to be -1.19 for HHLARGE and 0.47 for EDUC.***

* + 1. 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?

***Elasticity of HHLARGE changed by 0.88, while that of EDUC changed by 1.22. This means that education is more important to price sensitivity.***

* 1. 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.

***This means that without the factors of price-sensitivity, a large household would buy more quantity of orange juices and an educated person would buy less, which would make sense.***

***If we imagine this as HHLARGE as a demand curve with logmove on the y-axis and log(price) on the x-axis, then the coefficient on HHLARGE is now the intercept and the coefficient of log(price): HHLARGE is the slope. In this case, it’s a downward sloping curve with a positive intercept, whereas before when we didn’t include the interaction terms, it’s just a vertical shifter and we can think of it as the mean of this downward sloping curve. That’s why the coefficient for HHLARGE beforehand is negative.***

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?

***From the coefficients, we can see that decreases in lagged prices leads to decrease in this week's quantity sold, which make sales less attractive. Moreover, this phenomenon is shown mostly for Dominick's (0.65), less for Minute Maid (0.52) and the least with Tropicana (0.38), which is consistent with our findings about the elasticity for these brands so far.***

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.

***From the coefficients, we can see that Minute Maid has twice the cross-price elasticity as Dominick's, which makes sense because orange juice in one brand is substitute of other brands.***

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.
      1. Perfect price discrimination: ***when a business charges the maximum possible price for each unit consumed and captures the entire consumer surplus.***
      2. Product-based: ***when a business offers multiple versions of goods and allow consumers to self-select.***
      3. 2-part tariff: ***when there is a flat fee and a per unit price.***
      4. Direct price discrimination: ***when a business sets different prices based on observable and legal demographics of its consumer.***
      5. Bundling: ***It’s a conditional discount where if you buy one product, then you get another product at a discounted rate.***
   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.
      1. Perfect price discrimination: ***it would be very hard to get information on each customer’s maximum willingness to pay.***
      2. Product-based: ***to determine how to price different versions of the good in a profit maximizing way.***
      3. 2-part tariff: ***this would face big competition because the per unit price of all items would have been increased.***
      4. Direct price discrimination: ***one problem that businesses face would be that they must verify the identity of each consumer in terms of its association of one demographic group.***
      5. Bundling: ***It would take a lot of trial and error to figure out the profit-maximizing way of creating a bundle.***
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?
      1. ***log(quantity demand for orange juice) ~*** ***log(price of orange juice) + log(price of bananas)***
      2. ***If the coefficient of log(price of bananas) is positive, this means that they are substitutes. Otherwise, if it is negative, then they are complements.***
   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.
         1. ***The sign would be positive, and you would be more likely to bundle the products because the value of bananas decreases when people by orange juices if they are substitutes of each other. The bundle would force people to purchase more for both items.***
      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.)
         1. ***The price of the bundle needs to be less than the sum because otherwise people would not buy the bundle.***
   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?
      1. ***You should continue to offer the bundle because a sale in orange juice would be attractive to the price sensitive customers. Moreover, you would want to also offer the bundle at a lower price because the price of orange juice has gone down.***