Data Science and Pricing

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

Emails: [jlarivi1@uw.edu](mailto:jlarivi1@uw.edu), [lghhager@uw.edu](mailto:lghhager@uw.edu) (TA)

# Course Assignments & Reading

Course assignments should be printed (code, output and descriptive answers) 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. 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. All work must be typed

Discussion questions do not need 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 5, due November 2**

This homework will also serve as a useful study guide for the midterm.

**Assignment to be turned in.** Please turn in your knitted Rmarkdown file (ideally HTML or PDF) and answers to the questions.

**Study guide:** You don’t need to turn in any answers to these questions. This is a study guide to help you, but it may not cover every concept on the midterm. Focus on the class notes, problem sets and our class discussions. Anything discussed in class is fair game.

**Optimal pricing basics**

* Setting up the monopolist’s profit maximization problem and deriving the optimal pricing formula for 1-good.
* Lerner index and optimal markup formula for one good.
* Different types of elasticities (own price, cross-price, income)
* Definition of extensive and intensive margin.
* How does optimal quantity vary with MC for monopolist?
  + More or less quickly than with perfect competition.
* Intuition on the markup formula and the role of price sensitivity and available substitutes in determining markups.
* You do not need to memorize the N-good optimal pricing formula, but you do need to know the intuition for how the optimal prices change when goods are complements vs. substitutes.
* In what way is a bundle a conditional discount?
* When is bundling a good idea?

**Optimal pricing with many prices**

* What is the Myerson-Satterthwaite Theorem and how does it relate to pricing?
* Why does one price leave “trades on the table”? Understand this graphically and in words.
* Direct price discrimination, definition and examples.
* Indirect price discrimination, definition and examples.
* Definition of two-part tariff, examples of two-part tariff in the market.
* How does bundling work? What “valuation profiles” make bundling attractive? What are some examples of bundles in the marketplace.
* Related to the OJ HW, if the supermarket chain learned that different shoppers at different stores have different price sensitivity, how should it respond?
* Elasticity by different brands and elasticity matrices.

**Empirical methods**

* When pricing new products, what is the most popular method firms rely on?
* In the demand curve modeling, what functions in R were we using?
* What are interaction terms? What can they tell us? What are some examples of how we have used these in assignments?
* What is a predicted value?
* How do you get a predicted value in R?
* What is K-fold cross validation algorithm?
* What is cross validation good for?
* Conceptually understand train/test split for samples to do out-of-sample evaluation of fit.
* What is our measure of “true R-squared” (e.g., corr(-,-))?
* What is the overfit problem? How does out-of-sample testing help solve it?
* General knowledge of LASSO and Ridge regression
* What is objective function for LASSO?
* Why does LASSO “kick features out”?
* Why is LASSO not good for hypothesis testing?
* What is needed for causal inference?
* What is omitted variable bias?
* What is reshaping data and how do you reshape data in R (e.g., go wide?)
* Understand the distinction between measure of model fit (e.g. R-squared and variants) and estimating causal parameters.
* What is positive selection versus negative selection (e.g., education example)?
* What type of data is often useful for structural demand modelling modeling?
* Value for regression (inference across different groups) & prediction (conditional averages)
* Basic R code for linear regression, summarizing data, LASSO, reshaping, merging.
* Interpreting R output.
* Writing down linear model which corresponds to a elasticity estimation and level versus slope shifters.
* Market versus product level demand modeling
  + Constant elasticity of supply.
* Changing impact of independent variables on outcomes.

**Intertemporal Elasticity**

* Same product in different time periods is substitute
  + True for non-durables.
  + For durables irrelevant since can consume in each period (e.g., pizza versus TV therefore only thing that matters is delaying consumption).
    - Key feature for durables is incentive to lower price in each time period to get larger market
  + Similar to cross price elasticity

**Midterm format**

* 90 minutes in length.
* Multiple Choice, T/F plus explanation, short answer, some calculation.
* Type of questions you should expect:
  + Conceptual: similar to discussion questions on HW and conversations we’ve had in class.
  + R Output interpretation (~20%): we provide you R output and you interpret this based on your understanding of the core concepts we have learned so far.
  + Pen/paper math: we give you key parameters and you use/derive formulas to produce the answers, similar to HW1.
  + Graphical: we show you figures and ask you to interpret what’s going on.
  + Methodological: we ask how you would test a certain hypothesis using the statistical methods we have used. Similar to the written questions in HW2-4.
  + R commands (~15%): straightforward questions on the main commands we have been using so far.
* Tips
  + Know and understand the key formulas we have highlighted above
  + Some questions will come almost directly from the HW assignments.
  + Word your answers concisely
  + Ask questions if anything is unclear
  + Justify your answers with core economic logic

Questions (to be turned in)

1. Consider the following regression and output in R. The data here is the reshaped data where we’re only looking at the Tropicana sales. Features include price of Dominicks, MM and Trop, though, and feat is only for Tropicana being featured.
   1. What is the expected change in quantity (in percents) for a 10% increase in the price of Trop?
   2. What is the expected change in quantity (in percents) for a 10% increase in the price of MM?
   3. Why should we be concerned about inferring causality from the coefficient on log(P\_Trop):feat?
   4. If Trop is a normal good, what would you expect the sign of the interaction of INCOME and log(P\_Trop) to be?
2. Call:
3. glm(formula = logmove ~ log(P\_Dom) + log(P\_MM) + log(P\_Trop) \*
4. feat + AGE60 + EDUC + ETHNIC + INCOME + HHLARGE + WORKWOM +
5. HVAL150 + SSTRDIST + SSTRVOL + CPDIST5 + CPWVOL5, data = trop\_cross)
6. Deviance Residuals:
7. Min 1Q Median 3Q Max
8. -2.64300 -0.29531 -0.01999 0.26977 2.75790
9. Coefficients:
10. Estimate Std. Error t value Pr(>|t|)
11. (Intercept) 16.036404 0.405539 39.543 < 2e-16 \*\*\*
12. log(P\_Dom) 0.099452 0.021932 4.535 5.84e-06 \*\*\*
13. log(P\_MM) 0.265621 0.027778 9.562 < 2e-16 \*\*\*
14. log(P\_Trop) -2.318683 0.029309 -79.113 < 2e-16 \*\*\*
15. feat 1.736249 0.061270 28.337 < 2e-16 \*\*\*
16. AGE60 2.667011 0.156525 17.039 < 2e-16 \*\*\*
17. EDUC 1.597824 0.124527 12.831 < 2e-16 \*\*\*
18. ETHNIC 0.422341 0.045676 9.246 < 2e-16 \*\*\*
19. INCOME -0.509394 0.040865 -12.465 < 2e-16 \*\*\*
20. HHLARGE -1.310843 0.282861 -4.634 3.63e-06 \*\*\*
21. WORKWOM -0.778644 0.180130 -4.323 1.56e-05 \*\*\*
22. HVAL150 1.090685 0.051409 21.216 < 2e-16 \*\*\*
23. SSTRDIST -0.023534 0.001803 -13.053 < 2e-16 \*\*\*
24. SSTRVOL -0.112973 0.011990 -9.422 < 2e-16 \*\*\*
25. CPDIST5 0.087307 0.007720 11.309 < 2e-16 \*\*\*
26. CPWVOL5 -0.534881 0.031534 -16.962 < 2e-16 \*\*\*
27. log(P\_Trop):feat -1.367762 0.066919 -20.439 < 2e-16 \*\*\*
28. ---
29. Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1
30. (Dispersion parameter for gaussian family taken to be 0.2164328)
31. Null deviance: 6927.8 on 9648 degrees of freedom
32. Residual deviance: 2084.7 on 9632 degrees of freedom
33. AIC: 12634
34. Number of Fisher Scoring iterations: 2
35. Subset the oj data to only look at Dominicks sales (to be turned in as Rmd output).
    1. Find the quartiles of INCOME.
    2. Average sales within each quartile.
    3. Construct the MSE within each quartile for the model described above?
    4. Which quartile has the lowest MSE?
       1. What does this mean about the distribution of sales within that quartile?
    5. Which quartile has the highest MSE?
       1. What does this mean about the distribution of sales with that quartile?