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 1**

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

**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.
  + (Price-MargCosts)/Price=-1/Elasticity
* Different types of elasticities (own price, cross-price, income)
  + What happens if you raise your price / others raise their price / income increases
* Definition of extensive and intensive margin.
  + Intensive you change the price so what happens if you still sell the same amount at different price. Extensive is how much more/less you sell at the new price (What you gain by changing price)
* How does optimal quantity vary with MC for monopolist?
  + - They set it such that profit is maximised
  + More or less quickly than with perfect competition.
    - It happens quicker than with perfect competition as they have all the power.
* Intuition on the markup formula and the role of price sensitivity and available substitutes in determining markups.
  + When people are less price sensitive firms can mark up more as they lose less sales with the increase of price (Increase in intensive margin is greater than loses on extensive)
* 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?
  + You get a discount on one peoduct if you buy the other one.
* When is bundling a good idea?
  + When a value of good 1 us much lower if you already have good 2

**Optimal pricing with many prices**

* What is the Myerson-Satterthwaite Theorem and how does it relate to pricing?
  + If two parties do not know each other’s values of a certain item it will always result in one taking a loss. This can be seen if there is only one price for good, then people in DWL triangle could still be served profitably but instead are not.
* Why does one price leave “trades on the table”? Understand this graphically and in words.
  + That someone has higher value for this item, so one who posses this item and his value is less then another person who wants it and has higher value, is able to trade it for more than it is worth to him
* Direct price discrimination, definition and examples.
  + Charging higher price to certain part of population due to some factors, (schools and firms pay for Microsoft office)
* Indirect price discrimination, definition and examples.
  + There are several options and customers choose what they pay by their actions/choices ( Game, and the Ultimate version)
* Definition of two-part tariff, examples of two-part tariff in the market.
  + Golf club membership pay to be member, then pay for tee time (good clubs) Fixed fee + usage fee
* How does bundling work? What “valuation profiles” make bundling attractive? What are some examples of bundles in the marketplace.
  + Pairing good together, the price of the bundle is less than the price of goods separately, Netflix is an example.
* Related to the OJ HW, if the supermarket chain learned that different shoppers at different stores have different price sensitivity, how should it respond?
  + They should increase the price where the price sensitivity is lower.
* Elasticity by different brands and elasticity matrices.
  + How each brand should price according to what the other brands do.

**Empirical methods**

* In the demand curve modeling, what functions in R were we using?
  + lm
* What are interaction terms? What can they tell us? What are some examples of how we have used these in assignments?
  + Multiplication of regressors with each other. They tell us the impact one thing has with the relation of another for example featured\*brand
* What is a predicted value?
  + The Value of explained variable linear regression would predict based on regressors
* How do you get a predicted value in R?
  + Predict(reg)
* What is K-fold cross validation algorithm?
  + Split data set into K parts. Use K-1 of then for training (different combinations) K times. And then the remaining one part to test
* What is cross validation good for?
  + Getting the best module for given research problem (MSE minimized)
* 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(-,-))?
  + **Predicted and real y**
* What is the overfit problem? How does out-of-sample testing help solve it?
  + Uses too much regressors to fit to random noise in training set, we compare MSE in test sets since regsressions are not fit to thses
* General knowledge of LASSO and Ridge regression
  + LASSO kicks out regressors, Ridge sets them smaller than OLS
* What is objective function for LASSO?
  + Minimize OLS onj + k\* (sum(abs(Beta)) where k is punishment, for RIDGE it is beta^2 not abs
* Why does LASSO “kick features out”?
  + Since their predictive power is less than needed.
* Why is LASSO not good for hypothesis testing?
  + It does not have a interpretation for coefficients, cant use to test hypotesis
* What is needed for causal inference?
  + Bunch of OLS assumptions:
* What is omitted variable bias?
  + We do not include a important variable so their importance is assigned to another variable.
* What is reshaping data and how do you reshape data in R (e.g., go wide?)
  + Shaping data frames such that we can use it for needed regression
* Understand the distinction between measure of model fit (e.g. R-squared and variants) and estimating causal parameters.
  + R^2 how much of the actual data is explained by our regression.
* What is positive selection versus negative selection (e.g., education example)?
  + Negative selection: giving grants for students above some GPA to get them to collage.
  + **Positive selection is the majority of bias in our data.**
  + **Majority in selection bias**
* What type of data is often useful for structural demand modelling modeling?
  + Cross-sectional data
* 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

Practice Questions

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?

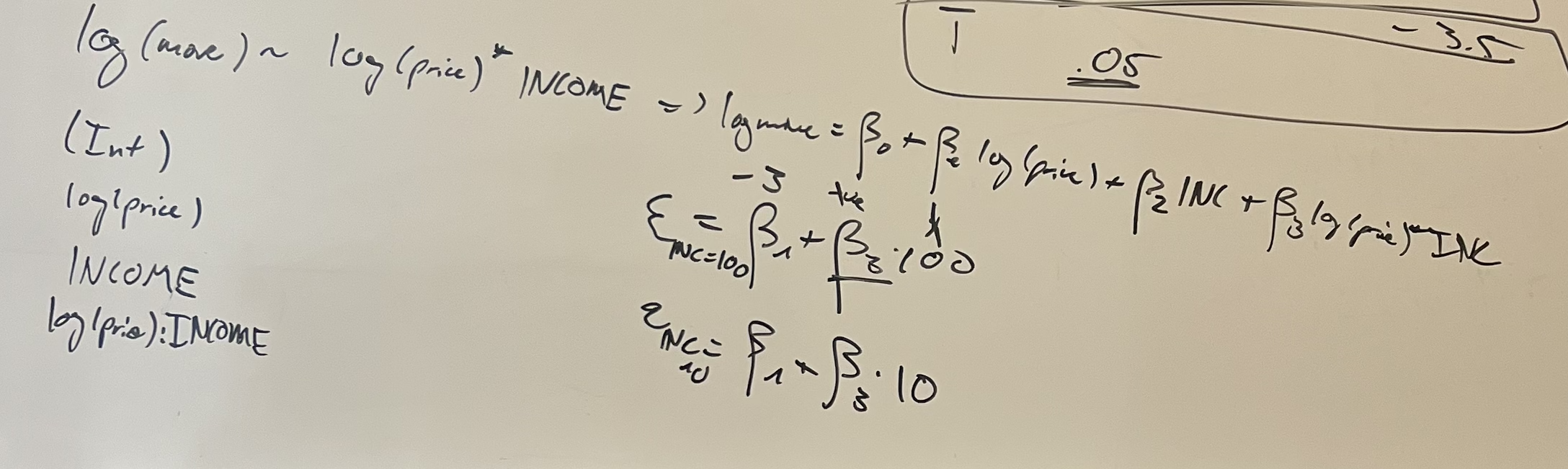
Decrease of 23%

* 1. What is the expected change in quantity (in percents) for a 10% increase in the price of MM?

Increase of 2.6%

* 1. Why should we be concerned about inferring causality from the coefficient on log(P\_Trop):feat?

Because Tropicana knows when it is feature it sets its prices accordingly so we don’t really know the effect of ‘featured’ itself. **FEAT IS NOT RANDOM, its endogenous**

* 1. If Trop is a normal good, what would you expect the sign of the interaction of INCOME and log(P\_Trop) to be?
  2. 

So Beta of interaction is positive since we would expect an increase of income to decrease elasticity.

1. Call:
2. glm(formula = logmove ~ log(P\_Dom) + log(P\_MM) + log(P\_Trop) \*
3. feat + AGE60 + EDUC + ETHNIC + INCOME + HHLARGE + WORKWOM +
4. HVAL150 + SSTRDIST + SSTRVOL + CPDIST5 + CPWVOL5, data = trop\_cross)
5. Deviance Residuals:
6. Min 1Q Median 3Q Max
7. -2.64300 -0.29531 -0.01999 0.26977 2.75790
8. Coefficients:
9. Estimate Std. Error t value Pr(>|t|)
10. (Intercept) 16.036404 0.405539 39.543 < 2e-16 \*\*\*
11. log(P\_Dom) 0.099452 0.021932 4.535 5.84e-06 \*\*\*
12. log(P\_MM) 0.265621 0.027778 9.562 < 2e-16 \*\*\*
13. log(P\_Trop) -2.318683 0.029309 -79.113 < 2e-16 \*\*\*
14. feat 1.736249 0.061270 28.337 < 2e-16 \*\*\*
15. AGE60 2.667011 0.156525 17.039 < 2e-16 \*\*\*
16. EDUC 1.597824 0.124527 12.831 < 2e-16 \*\*\*
17. ETHNIC 0.422341 0.045676 9.246 < 2e-16 \*\*\*
18. INCOME -0.509394 0.040865 -12.465 < 2e-16 \*\*\*
19. HHLARGE -1.310843 0.282861 -4.634 3.63e-06 \*\*\*
20. WORKWOM -0.778644 0.180130 -4.323 1.56e-05
21. HVAL150 1.090685 0.051409 21.216 < 2e-16 \*\*\*
22. SSTRDIST -0.023534 0.001803 -13.053 < 2e-16 \*\*\*
23. SSTRVOL -0.112973 0.011990 -9.422 < 2e-16 \*\*\*
24. CPDIST5 0.087307 0.007720 11.309 < 2e-16 \*\*\*
25. CPWVOL5 -0.534881 0.031534 -16.962 < 2e-16 \*\*\*
26. log(P\_Trop):feat -1.367762 0.066919 -20.439 < 2e-16 \*\*\*
27. ---
28. Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1
29. (Dispersion parameter for gaussian family taken to be 0.2164328)
30. Null deviance: 6927.8 on 9648 degrees of freedom
31. Residual deviance: 2084.7 on 9632 degrees of freedom
32. AIC: 12634
33. Number of Fisher Scoring iterations: 2
34. Subset the oj data to only look at Dominicks sales (to be turned in as Rmd output).
    1. Find the quartiles of INCOME.

Quan=quantile(oj$INCOME)

* 1. Average sales within each quartile.
  2. Construct the MSE within each quartile for the model described above?
  3. Which quartile has the lowest MSE?

Q3

* + 1. What does this mean about the distribution of sales within that quartile?

The Predicted distribution of logmove is close to the actual one.

* 1. Which quartile has the highest MSE?
     1. What does this mean about the distribution of sales with that quartile?