Industrial Organization and Data Science

Instructors: Justin Rao, Affiliate Professor & Senior Researcher, Microsoft

Jacob LaRiviere, Affiliate Professor & Senior Researcher, Microsoft

Emails: justin.rao@microsoft.com

jlariv@microsoft.com

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 3, due April 20

Hastie Ch. 1 and 3.1-3.3

Thirty Years of Conjoint Analysis

Measuring Customer's Reaction to Price

Assignment to be turned in. Please turn in your R script and anything in bold below.

Note: you will probably not know all the relevant commands of the top of your head. Nobody does. Simply search "command in R" or "command in R examples" etc. in a search engine, and this will almost always give the answer.

Data from a conjoint experiment in which two partial profiles of credit cards were presented to 946 respondents. The variable bank$choiceAtt$choice indicates which profile was chosen. The profiles are coded as the difference in attribute levels.

- If choice is 1, then 1's indicate that the chosen option had the attribute
- If choice is 0 (they chose the other option), then -1's give the attributes for the "passed over" option (e.g. the one represented by the row).
- 0's indicate the two choices were the same on this dimension.
- Choice: =1 if that left profile was chosen, =0 if the other option was chosen
- Med_Fint: medium fixed interest rate.
- Low_Fint: low fixed interest rate.
- Med_Vint: medium variable interest rate.
- Reward_X: reward plan X was offered.
- Med_fee: medium fee

- Low_fee: = low fee
- Bank_A/Bank_B/Outofstate: signifying which bank
- Med/High_rebate: =signifying rebate level
- High_CredLine: high credit line
- Long_grace: long grace period for payment penalties

Install the bayesm package and load the library (`library(bayesm)`). The manual is available here https://cran.r-project.org/web/packages/bayesm/bayesm.pdf

1) We'll be using the bank card conjoint data described on p.3. To load the data just type `data(bank)`.
2) Assign the choiceatt and demo to dataframes. E.g. `choiceatt <- bank$choiceatt`.
3) The demographic data and choice data are in different data frames. It will be useful to join them together. Use the `merge` command to do so. You should have a combined data frame that has 14,799 observations and 19 variables.
4) Run a baseline logit regression that predicts choice as a function of the attributes. Note: use a constant (do not put ~0, although the results will be very similar either way. If you suppressed the constant initially, don't worry about it).
    a. **Make a table of coefficients and put them in your writeup.**
    b. **What factors have the biggest impact on choices?**
    c. **Do people seem to care more about fees or interest rates? Is this rational?**
5) My hypothesis is that people who care about low fees care more about low interest rates. Create an interaction term to test my hypothesis. **Was I correct? What is the p-value on the interaction term?**
6) Now let's try to estimate demand parameters for each individual. Since we have approximately 16 choices per person, we cannot estimate many parameters. Let's run a simple model that includes: fees and interest rates.
    a. **Write down your estimating equation.** There are 5 variables, but how many can we include in the regression?
        i. **Answer:** Normally we can't include all the variables in a given category because then the constant is not defined (e.g. the male/female example from the lecture, assuming everyone identifies as either a male or a female). However these data are a bit different. The reason being is the feature values are the *differences of the two choices presented.* For example, if Med_Fee=1 this does not mean Low_Fee = 0 (it cannot be 1 since we know the chosen option has medium fee, but it could be 0 or -1).
        ii. If we wanted to include fees and interest rates, we should include a constant and all the features.
        iii. For part(c), you can choose to run the regression including all the fees variables, all the interest variables, or both. Explore what works and what does not.
    b. Run your estimating equation on the all the data (as in 4). **Report the results summary. What is different? How has the R-squared changed?**

c. Now let's run your estimating equation for each individual. There are many ways to this in R. Let's explore one such way.
   i. We need to run a regression on a subset of the data, namely we want to restrict to a given subject, store the data, then run it again. Step 1, create a data frame that is just subject 1's data. Hint: one way to do this is the `subset` command.
   ii. Let's create a data frame to store our results. It needs to have as many rows as we have subjects and 3 columns (one for each variable in our regression). Figure out how to create this data frame, all the cells can be empty to start with. Give each column a name for the variables, and be sure to use the same order in (iii) below.
   iii. Now let's run a regression for subject 1. Use the `coef` function and place the coefficients in the first row of our results matrix (make sure your columns match up).
   iv. Now we want to run this regression for each subject. Use a *for loop* to do so. The "inside" of your loop has already been created (subset the data, run the regression, store the results). At each stage, save the results to the right part of the results data frame, at the end, you should have results for each subject.

d. Plot the distribution of each demand parameter (e.g. histogram). What does these tell us? Are there identifiable types?
   i. If you used just the fees variables, we want to know if some people look like they really care about fees and some people don't care at all.
   ii. Same applies if you used the interest rate variables.
   iii. Bonus: if you have estimates for both fees and interest rates, does it look like these demand parameters are correlated within an individual?

e. Now that we have results for each individual, let's see if demand parameters depend on demographics. Merge in the demographics by subject id.
   i. Regress each demand parameter on the demographics.
      1. Example: Regress the parameter for "low fee" on demographics. This would say "what demographics predict really liking a low fee."
   ii. What results are statistically significant?
   iii. What have you learned?

7) **Bonus:** perform the individual level regressions using the `sapply` or `lapply` function.
8) **Bonus:** For the individual level regressions, was it better to do a separate fee and interest rate regression, or do it in a combined way? What is your argument? Hint, check the model standard errors. An easy way to do this is the coefficients object in summary.

```
summary(df1)$coefficients
```

   a. Column 2 gives the standard errors, so you can record them as var1_se=

```
summary(df1)$coefficients[1,2]
```