Applications of the Discrete Choice Model in Industrial Organization

Instructor: Si Zuo*

1 Syllabus

1.1 Research Description

Industrial Organization is a field in applied microeconomics, studying the consumers' demand, firms' supply, the competition among firms, government's policies, people's education, and housing decisions. In this 5-week group research, we will learn the discrete choice demand model (logit model), the foundation of empirical industrial organization literature, from Berry (1994). Then we will see the applications of the discrete choice model in two recent AER (American Economics Review, Economics Top Journal) papers: one is about the consumers' choices in the local newspaper (Fan, 2013); the other is about the parents and children' choices on high schools (Kapor et al., 2020). Students could choose one AER paper for the simple replication exercise, including figures showing the data patterns and several regressions.

1.2 Research Objectives

Students are expected to understand and illustrate the discrete choice model in Berry (1994) and the paper students choose for the replication exercise. In addition, students should be able to finish the replication exercise using Stata, following the instructions from the instructor.

1.3 Research Period

Jan 17th - Feb 10th, five weeks

*Cornell University, New York, U.S.A. Email: sz549@cornell.edu. Website: https://www.si-zuo.com/.

1.4 Skills Required

Good English reading and writing. Data cleaning and data analysis with Stata.

1.5 Final Output Expected

Presentation Slides or PPT (no more than 30 pages) showing the data replication exercise results, including but not limited to paper summary and how demand estimation model, is used required and optional data descriptive tables and figures, required and optional regression tables and group reflection. After choosing the coding exercise, the instructor will give more details about required and optional tables and figures.

A report (latex or word, main content no more than 10 pages) including title, abstract, introduction, data description, empirical strategy, results, conclusion and reference. The report should include all the important graphs and tables and the explanation. Other graphs, tables and explanations should be in the appendix. The language needs to be polished and the writing needs to be clear.

1.6 Time Line

- Guided group research 1: Introduction of the demand estimation and the logit model. Explaining the coding assignment details and the data cleaning code. Students should begin to choose the data replication exercise. The instructor will distribute the sample codes of two coding exercises to all students.
- Guided group research 2: Illustration of Berry (1994). Students should decide on the data replication exercise and start the data cleaning and background information collecting in the group.
- Guided group research 3 & 4: Illustration of paper Fan (2013) and the sample code.
- Guided group research 5 & 6: Illustration of paper Kapor et al. (2020) and the sample code. Students need to finish the data cleaning and the background information collecting in class 5. Discussing about the problems in the data cleaning process.
- Guided group research 7 & 8: Replication exercise, report and slides preparation.

- Guided group research 9: Introduction of the nested logit model and the BLP model. Q & A and preparation for the group presentation.

1.7 Logistics

- Questions and communications: The beginning 10 minutes of each class (recommended), office hour (recommended), and email. (Writing emails is an important skill!)
- Data, sample code, and file posting: The instructor will post all materials (data, sample codes, two paper PDF, slides, and other things) by email.

2 Coding Exercise of Kapper 2020

2.1 Data Source

- National Center for Education Statistics
  `https://nces.ed.gov/ccd/elsi/expressTables.aspx`
- ACS data and Census
  `https://www.nber.org/research/data/census-sf1-zip-code-tabulation-area-zcta-data-2012`

2.2 Data Composition

- District Data: basic information, characteristics, enrollment basic, enrollment detail, teacher, expenditure, income. (half cleaned up)
- Demographics Data: 2015-2019 county-level demographic data. (cleaned up County Level dta data)

2.3 Topic and Aim

Topic: Preference Heterogeneity in the School District Choice
Aim: Understanding how parents and children value school district's characteristics (tuition, district agency type, teacher number, pupil/teacher ratio...). Furthermore, discovering the preference heterogeneity in different gender groups, race groups, and income groups.

2.4 Data Cleaning and Background Discovery

- Data Cleaning: following the instruction in the do-file
- Background Discovery: How are the school districts decided in the US? What are the school district agencies in the data? What factors affect parents' choices in schools from the newspaper reports? Try to understand the variables in the data. You could refer to the National Center for Education Statistics for details if needed.

2.5 Data Generation

- Generate the average tuition in every district (average tuition= tuition income/ enrollment student number)
- Generate the private school variables (private school student, private teacher) from the data. (district minus the public school variables)
- Sum all the variables to the county level.

2.6 Data Patterns

- Variable Summary Table: following the code in the do-file.
- Graphs (line) show the relation between total student and total teacher, total student and pupil-teacher ratio, total student and tuition. Show the above line graphs by different agency types.

2.7 Linear Regressions (OLS and Fixed Effect Regression)

- Using the county-year- level data, run the OLS regression for county i and year t:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \epsilon_{it}$$

where Y_{it} is the county education level variables (tuition, total student number, total teacher, public school student ratio, pupil-teacher ratio...). X_{it} is a vector including the county level demographics variables (population, average income, female ratio, adult group ratio, white ratio, Asian ratio, African American ratio...)

- Repeat the above OLS exercise using the Fixed effect regression. Add State FE and Year FE. Error cluster at the county level. (Check `http://scorreia.com/software/reghdfe/quickstart.html` for the code)

2.8 Demand Estimation (Simple Logit and Logit IV)

- Following the codes in `http://aguirregabiria.net/courses/eco310/tutorial_4.pdf`.
- Market is defined at county-year level.
- The utility of attending school j in county i year t is defined as :

$$U_{ijt} = \beta X_{jt} + \alpha p_{jt} + \epsilon_{ijt}$$

  where X_{jt} is a vector including logged total teacher number, public school teacher ratio, pupil teacher ratio and a constant. p_{jt} is the tuition. ϵ_{ijt} follows the type I extreme value distribution. The utility of attending an outside choice is normalized to 0. Estimate and α using the simple logit model.

- Repeat the estimation exercise above and use the average per student expenditure of other districts in the county, the number of schools in the district, the number of schools in the county as the IV for p_{jt}. (Check `http://scorreia.com/software/reghdfe/quickstart.html` for the code)
- Repeat the exercise above but including household heterogeneity in the utility:

$$U_{ijt} = \beta X_{jt} + \alpha p_{jt} + \delta X_{jt} * Z_{it} + \epsilon_{ijt}$$

  where Z_{it} includes dummy variables $(0, 1)$ whether $county_i$ is of high income, woman majority, Asian majority, or African American majority. You could define the cutoff as the national average. That is, for example, $county_i$ is regarded as a high-income county if the average income is higher than the national average income.

3 Coding Exercise of Fan 2013

3.1 Data Source

- MBL (Norway Newspaper Open Data Source, newspaper reader number, and circulation number by county)
  `https://www.mediebedriftene.no/tall-og-fakta/`
- Statistics Norway
  `https://www.ssb.no/en/statbank/table/08921/tableViewLayout1/`

3.2 Data Composition

- Newspaper Circulation Data: circulation, paper newspaper circulation, digital newspaper circulation by county, newspaper and year. (half cleaned up)
- Demographics Data: 2015-2019 county-level demographic data. (half cleaned up)

3.3 Topic and Aim

Topic: Does the Digital Newspaper Cannibalize the Printed Newspaper?
Aim: Exploring what kinds of newspapers are more likely to have high digital circulation. (national newspaper, new entrants...) Calculating and comparing the elasticity of paper newspaper supply on digital news supply by different markets. Exploring how people's demographics (income, age, and education level) affect the local newspaper supply type.

3.4 Data Cleaning and Background Discovery

- Data Cleaning: following the instruction in the do-file. (optional) Collecting the major national newspaper price (subscription fee) and type information (sports, health...).
- Background Discovery: What factors affect consumers' decision of paper newspaper or digital newspaper? Try to understand the variables in the data. You could refer to MBL for details if needed.

3.5 Data Generation

- Generate newspaper-year level total circulation, paper newspaper circulation, digital newspaper circulation and digital-paper ratio.
- Generate the circulation, paper newspaper circulation, digital newspaper circulation, digital-paper ratio and elasticity of paper newspaper supply on digital news in county i in year t following:

$$logs_{it}^{paper} = \gamma_0 + \gamma s_{it}^{digital} + \epsilon_{it}$$

  where s^{paper} is the paper newspaper circulation, s^{paper} is the digital newspaper circulation and γ is the elasticity.
- Generate the circulation, paper newspaper circulation, digital newspaper circulation, digital-paper ratio, and elasticity of paper newspaper supply on digital news of the whole country in year t.
- Generate a variable indicating how many counties the newspaper is serving and a dummy variable indicating whether the newspaper is a local newspaper. (only serving one county)
- Generate a variable indicating how many years the newspaper is serving from 2015.
- Generate a variable indicating whether the newspaper is a paper-dominating newspaper or digital-dominating newspaper. You could use the average ratio as the cutoff.

3.6 Data Patterns

- Variable Summary Table: following the code in the do-file.
- A Table presenting the country level elasticity of paper newspaper supply on digital news by year.
- A Table presenting the county level elasticity of paper newspaper supply on digital news by year in the top 5 counties.
- Graphs (line or bar) show the country level circulation, paper newspaper circulation, digital newspaper circulation, the digital-paper ratio by year.
- Graphs (line or bar) show the country level circulation, paper newspaper circulation, digital newspaper circulation, the digital-paper ratio by year in the top 5 counties. (by circulation)

3.7 Linear Regressions (OLS and Fixed Effect Regression)

- Using the county-year-level data, run the OLS regression for county i and year t:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \epsilon_{it}$$

where Y_{it} is the county education level variables (circulation, paper newspaper circulation, digital newspaper circulation, digital-paper ratio, elasticity of paper newspaper supply on digital news). X_{it} is a vector including the county level demographics variables (population, average income, female ratio, adult group ratio, children ratio, old ratio, high education adult ratio.)

- Repeat the above OLS exercise using the Fixed effect regression. Add County FE or Year FE. Error cluster at the county level. (Check `http://scorreia.com/software/reghdfe/quickstart.html` for the code)

3.8 Demand Estimation (Simple Logit)

- Following the codes in `http://aguirregabiria.net/courses/eco310/tutorial_4.pdf`.
- Market is defined at county-year level.
- The utility of choosing newspaper j in county i year t is defined as :

$$U_{ijt} = \beta X_{jt} + \delta Digital_{jt} + \epsilon_{ijt}$$

where X_{jt} is a vector including how many counties newspaper j is serving, how many years newspaper j is serving and a constant. $Digital_{jt}$ is the digital-paper ratio of the newspaper in year t. The utility of choosing an outside choice is normalized to 0. Estimate , δ using the simple logit model.

- Repeat the exercise above but including household heterogeneity in the utility:

$$U_{ijt} = \beta X_{jt} + \delta Digital_{jt} + \delta_2 X_{jt} * Z_{it} + \epsilon_{ijt}$$

where Z_{it} includes dummy variables $(0, 1)$ whether $county_i$ is of high income, woman majority, high education majority, young majority, and old majority. You could define the cutoff as the national average. That is, for example, $county_i$ is regarded as a high-income county if the average income is higher than the national average income.

References

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, 242–262.

Fan, Y. (2013). Ownership consolidation and product characteristics: A study of the us daily newspaper market. *American Economic Review 103*(5), 1598–1628.

Kapor, A. J., C. A. Neilson, and S. D. Zimmerman (2020). Heterogeneous beliefs and school choice mechanisms. *American Economic Review 110*(5), 1274–1315.