ECON 209
Introduction to Econometrics: Honors
Immigration and Inequality

Casper Neo

15 March 2017

1 Introduction

This article is my final project in ECON 209: Honors Econometrics. It is a partial replication and extension of David Card's *Immigration and Inequality*. Like in Card's paper, I estimate the elasticity of substitution between education and income groups using a constant elasticity of substitution (CES) model and linear regression. Special attention on selective inference and multiple testing corrections will be applied to bound the false discovery rate and false coverage rates.

Contents

2 Original Paper

Immigration and Inequality by David Card was published in 2009 in the American Economic Review. It explores the relationship between immigration status, education, and wages. Card uses the 1980-2000 Census data combined with the 2005-2006 American Community Survey data. In contrast I use the 2007-2015 American Community Survey. Card explores his hypothesis with time series and panel data as well. I select from his paper the following hypothesis to test and replicate.

1. Workers with below high school education are perfect substitutes for those with a high school education.
2. High school-equivalent and college-equivalent workers are imperfect substitutes with an elasticity of substitution between 1.5-2.5;
3. Within education groups, immigrants and natives are imperfect substitutes with an elasticity of substitution on the order of 20

As a main criticism which I will attempt to address is the fact that Card explores multiple hypothesis and does not address the fact that potentially many more hypothesis were considered (implicitly or otherwise) but not included in his paper. For example he may have checked the elasticity of substitution between immigrants and natives, grouped by race, geography or decade of entry.

3 Data Overview

Data from 'https://www.census.gov/programs-surveys/acs/data/pums.html'.

I download the 1-year ACS surveys from 2007-2015 however only few columns are used. I identified the following variables as relevant from the *PUMS Data Dictionary 2011-2015*. I map these variables from the values in the raw data into units usable by my models. Between the 8 years there are 27'725'196 people surveyed. Though after I filter for working age (18-60) and those in the labor force there are 11'994'312 people. The procedure and models are developed a random 5% sample then run on the remaining 95%.

Table 1: Data Used

Variable Key	Description
WAGP	Wages or salary income past 12 months
CIT	Citizenship status
NATIVITY	Native or Foreign Born
AGEP	Age
SCHL	Years of Schooling
ESR	Employment Status Recode
DECADE	Decade of entry into the United States
ST	State

4 Multiple Testing Concerns and Selective Inference

When the same dataset is used to generate hypothesis as the dataset used to test the hypothesis, researchers run into a multiple testing problem where, while generating their model, they mentally filter out models that seem not to fit the data and in the end use models that fit well. It is then little surprise that the models chosen have a strong fit to the data. Since both the question and the answer to the question are functions of (and so are dependent on) the same data, knowing the question reveals information about the answer; i.e. the hypothesis and the resulting p-values or confidence intervals are not

independent. The simplest resolution to this is to use different data to generate the hypothesis and to test the hypothesis. Hence, while I am using more recent data than David Card, whose hypothesis I test, I write my procedure using a random partition of my data and run tests on the rest.

I apply the Benjamini Hochberg False Discovery Rate (BH-FDR) control algorithm to bound the false discovery rate when studying the elasticity of substitution between groups. This has been shown [CITATION] to work when the test statistics are independent or exhibit positive regression dependancy. Since I am comparing the elasticity of substitution between groups. This assumption is probably valid. If education changes the elasticity of substitution for natives, intuitively it should do so for immigrants as well due to similar effects on marginal productivity.

5 Estimating Elasticity of Substitution Between Groups

Assuming a 1 sector framework with a Constant Elasticity of Substitution (CES) model

$$y = [\alpha_J L_{J\rho} + \alpha_K L_{K\rho}]^{\frac{1}{\rho}} \tag{1}$$

Where J and K are separate groups (for example high school educational equivalents and college equivalents), $\rho = 1 - \frac{1}{\sigma}$ and σ is the between group elasticity of substitution. L is the labor input, estimated by $1-$ the unemployment rate within the group. This is estimated by the linear model

$$\log \frac{w_j}{w_k} = \log \frac{\alpha_j}{\alpha_k} - \frac{1}{\sigma} \log \frac{L_j}{L_k} \tag{2}$$

We observe labor inputs, and wages then use a linear model to estimate $\sigma = \frac{1}{\beta_1}$. Our data has sufficient detail to know if the observed is a non-citizen, naturalized citizen, or born citizen (the former two are considered immigrants). We can also distinguish years of education, which state and even public use micro-data area code (By the census definition). The latter designate areas of 100'000 or more population. Each data point in the CES model is over one micro-data area in one year. Given the granularity of my data there are many potential hypothesis attempting to find the elasticity of substitution split along the following lines;

1. Immigrant vs native
2. Non-citizen vs citizen
3. Non-citizen vs naturalized citizen
4. Non-citizen vs born citizen
5. Naturalized citizens vs born citizens
6. less than high school educated vs only high school educated
7. less than high school educated vs college educated
8. less than high school educated vs at least high school educated
9. high school educated vs college educated

10. less than college educated vs college educated.

fifteen more hypothesis splitting across education while fixing citizenship and twelve more hypothesis splitting across citizenship while fixing education. I choose these 37 hypothesis since they seem the most natural to me.

With a set of 3 education levels there are 7 ways to include education levels in groups. Analogously for citizenship levels. Hence there are 49 potential groups. That means when comparing the elasticity of substitution between groups there are

$$\binom{(2^3-1)*(2^3-1)}{2} = \binom{49}{2} = 1176$$

ways to split the data of varying degrees of reasonability. The combinatoric explosion gets exponentially worse if I also chose to include race, decade of entry, geography, or any other potential categories.

6 Conclusion

7 Appendix: My code

Github link: https://github.com/CasperN/Immigration-and-Inequality.git

To download and clean the data, run in a unix terminal *bash getdata.sh*. Note *cleanData.py* should be in the same directory when this happens. This will automatically download and clean the data as I use it, and randomly split the data into a set for hypothesis generation and another for testing.

List of Figures

List of Tables