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

1 Introduction **1**

2 Original Paper **2**

3 Data Overview **2**

4 Multiple Testing Concerns and Selective Inference **3**

5 Estimating Elasticity of Substitution Between Groups **3**

6 Conclusion and Results **5**

7 Appendix: My code **6**

8 Appendix: 60 Plausible sounding null hypothesis **7**

List of Figures

1 Slope Parameter 5
2 Elasticity of Substitution 6

List of Tables

1 Data Used 2
2 R-Squared Values 6

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. He also restricts his data to the 124 largest metropolitan areas while I use all all areas in the country. 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.

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. I only look at similar hypothesis within the space of education and immigration, however the multiple testing procedure will be valid if more variables are added too.

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 between generating hypothesis and testing. 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 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.

Let n be the number of p-values, and α the false discovery rate you want to guarantee. The algorithm is to first find k^* such that

$$k^* = \max\{k : \text{at least } k \text{ p-values } < \alpha\frac{k}{n}\}$$

Then reject the k^* smallest p-values. By adjusting the threshold with k, the procedure accounts for both many weak signals and few strong signals. The procedure also comes with a simple process for constructing confidence intervals such that false coverage rate is controlled. Simply construct $(1 - \alpha\frac{k^*}{n})$ confidence intervals.

The advantage of this procedure over the well known Bonferroni correction is power at the cost of flexibility. The Bonferroni procedure does not assume anything about the distribution or correlation of p-values however it is too conservative and only the stongest of signals will reject a null hypothesis. The BH-DFR procedure takes advantage of positive dependance and also rejects null hypothesis when there are many weak signals. However a drawback is that while expected proportion of discoveries that are false is bounded, the variance of the proportion grows with positive correlation.

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. Observe as $\rho \to 1$, $\sigma \to \infty$, a percent change in group J's labor input corresponds to infinite change in group K's labor input. That is, they are perfect substitutes. When $\rho \to -\infty$, $\sigma \to 0$ which means every percent change in group J's labor input should correspond to

a percent change in group K's labor input. That is, they are perfect complements. When $\rho = 0$, $\sigma = 1$ we reach the threshold at which if $\rho \downarrow$ then $\sigma \downarrow$ and we have complements, and if $\rho \uparrow$ then $\sigma \uparrow$ and we have substitutes. As a null hypothesis I assume $\sigma = 1$ so we may test if the groups are substitutes or complements.

$\rho = 1$, $\sigma = \infty$ and we have perfect substitutes. As $\rho \to \infty$ we have $\sigma \to 1$ which means every percent change in group J's labor input should correspond to a percent change in group K's labor input. That is, they are perfect complements. As a null hypothesis for arbitrary groups I will assume they are perfect substitutes and $\sigma = \infty$. This production function may be estimated by the 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. If one of the groups is empty (which is possible given a large number of public micro-data areas, years, and categories) a divide by zero error will be reached. I throw out such data points and any x, y pairs with infinite values. Given the granularity of my data there are many potential hypothesis attempting to find the elasticity of substitution split along the following lines;

- Immigrant vs native
- Non-citizen vs citizen
- Non-citizen vs naturalized citizen
- Non-citizen vs born citizen
- Naturalized citizens vs born citizens
- less than high school educated vs only high school educated
- less than high school educated vs college educated
- less than high school educated vs at least high school educated
- high school educated vs college educated
- less than college educated vs college educated.

25 more hypothesis splitting across education while fixing citizenship and 25 more hypothesis splitting across citizenship while fixing education. I choose these 60 hypothesis since they seem the most natural. Their corresponding data are generated algorithmically and are listed in Section 8.

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

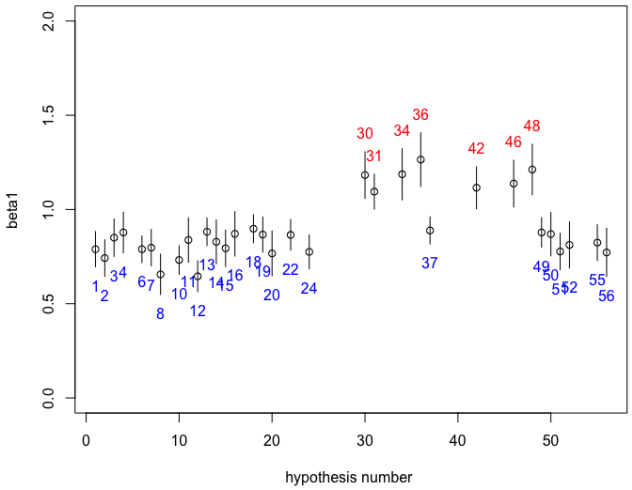
$$\binom{(2^3 - 1) * (2^3 - 1)}{2} = \binom{49}{2} = 1176$$

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 and Results

After running, for every hypothesis;$H^1, ..., H^{60}$; a two sided $T-$test testing the null $H_0^i : \beta_1^{H^i} = 1$ we get 60 p values. We run the $BH - FDR$ procedure and find 33 nulls were rejected while 27 were not. A surprising number of nulls where rejected however this is more to do with the dependance between the overlapping hypothesis.

Figure 1: Slope Parameter



Of the 33 rejections, we find 26 of them are considered gross substitutes while 7 of them are considered compliments. Interestingly, since I generated the hypothesis automatically we a kind of structure in the hypothesis space. Odd numbered hypothesis are comparing citizenship level groups (immigrant vs born citizen or non-citizen vs citizen) while holding level of education constant.

Unlike Card's first conclusion that those with below high school education are perfect substitutes, I do not find that result. All of my observations of β_1 lay close to 1.

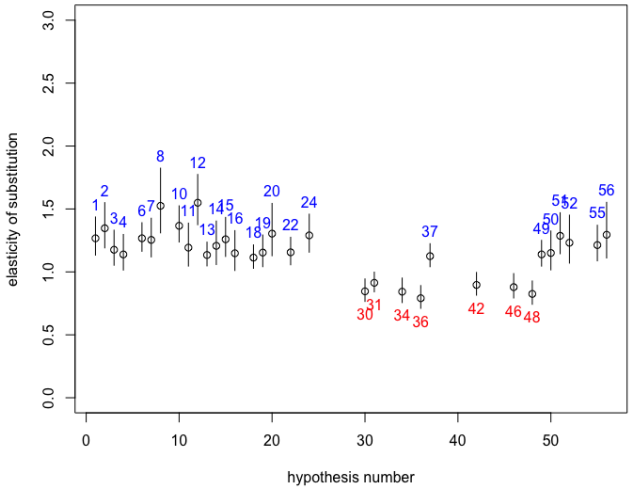
For hypothesis $H^{37}, H^{39}, H^{41}, H^{43}$ which correspond to Card's conclusion; that within education groups, immigrants and natives are imperfect substitutes with a large but finite elasticity of substitution; we fail to reject the null hypothesis that that $\sigma = 1$.

The differences between Card's and my results probably result from a combination of different data 1980-2005 versus 2007-2015, and the fact that I do not limit my analysis to the 124 largest metropolitan area. Also I find the R-squared values of the linear models to be incredibly weak.

Table 2: R-Squared Values

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.02518	0.04105	0.04935	0.05473	0.06089	0.11060

Figure 2: Elasticity of Substitution



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.

generateHypothesis.py will read either the training or testing data then process the data into simple xs and ys to be compared in a simple linear model. These data are saved in a *hypothesis/* directory. *analysis.R* reads the hypothesis and performs the multiple testing corrections.

References

[1] Card, David. 2009. "Immigration and Inequality." American Economic Review, 99(2): 1-21. http://www.aeaweb.org/articles?id=10.1257/aer.99.2.1

[2] Benjamini, Yoav; Yekutieli, Daniel. The control of the false discovery rate in multiple testing under dependency. Ann. Statist. 29 (2001), no. 4, 1165–1188. doi:10.1214/aos/1013699998. http://projecteuclid.org/euclid.aos/1013699998.

8 Appendix: 60 Plausible sounding null hypothesis

H1: For below high school workers, the elasticity of substitution between non-citzens and naturalized citizens is ∞
H2: For non-citizen workers, the elasticity of substitution between below high school and high school equivalent education is ∞
H3: For high school equivalent workers, the elasticity of substitution between non-citzens and naturalized citizens is ∞
H4: For naturalized citizen workers, the elasticity of substitution between below high school and high school equivalent education is ∞
H5: For college equivalent workers, the elasticity of substitution between non-citzens and naturalized citizens is ∞
H6: For born citizen workers, the elasticity of substitution between below high school and high school equivalent education is ∞
H7: For below college workers, the elasticity of substitution between non-citzens and naturalized citizens is ∞
H8: For immigrant workers, the elasticity of substitution between below high school and high school equivalent education is ∞
H9: For high school or more workers, the elasticity of substitution between non-citzens and naturalized citizens is ∞
H10: For all citizen workers, the elasticity of substitution between below high school and high school equivalent education is ∞
H11: For all workers, the elasticity of substitution between non-citzens and naturalized citizens is ∞
H12: For all workers, the elasticity of substitution between below high school and high school equivalent education is ∞
H13: For below high school workers, the elasticity of substitution between non-citizens and born citizens is ∞
H14: For non-citizen workers, the elasticity of substitution between below high school and college equivalent is ∞
H15: For high school equivalent workers, the elasticity of substitution between non-citizens and born citizens is ∞
H16: For naturalized citizen workers, the elasticity of substitution between below high school and college equivalent is ∞
H17: For college equivalent workers, the elasticity of substitution between non-citizens and born citizens is ∞
H18: For born citizen workers, the elasticity of substitution between below high school and college equivalent is ∞
H19: For below college workers, the elasticity of substitution between non-citizens and born citizens is ∞
H20: For immigrant workers, the elasticity of substitution between below high school and college equivalent is ∞
H21: For high school or more workers, the elasticity of substitution between non-citizens and born citizens is ∞
H22: For all citizen workers, the elasticity of substitution between below high school and college equivalent is ∞
H23: For all workers, the elasticity of substitution between non-citizens and born citizens is ∞
H24: For all workers, the elasticity of substitution between below high school and college equivalent is ∞
H25: For below high school workers, the elasticity of substitution between naturalized and born citizens is ∞
H26: For non-citizen workers, the elasticity of substitution between high school equivalent and college equivalent education is ∞
H27: For high school equivalent workers, the elasticity of substitution between naturalized and born citizens is ∞
H28: For naturalized citizen workers, the elasticity of substitution between high school equivalent and college equivalent education is ∞
H29: For college equivalent workers, the elasticity of substitution between naturalized and born citizens is ∞
H30: For born citizen workers, the elasticity of substitution between high school equivalent and college equivalent education is ∞
H31: For below college workers, the elasticity of substitution between naturalized and born citizens is ∞
H32: For immigrant workers, the elasticity of substitution between high school equivalent and college equivalent education is ∞
H33: For high school or more workers, the elasticity of substitution between naturalized and born citizens is ∞
H34: For all citizen workers, the elasticity of substitution between high school equivalent and college equivalent education is ∞
H35: For all workers, the elasticity of substitution between naturalized and born citizens is ∞
H36: For all workers, the elasticity of substitution between high school equivalent and college equivalent education is ∞
H37: For below high school workers, the elasticity of substitution between immigrants and born citizens is ∞
H38: For non-citizen workers, the elasticity of substitution between below college and college equivalent education is ∞
H39: For high school equivalent workers, the elasticity of substitution between immigrants and born citizens is ∞
H40: For naturalized citizen workers, the elasticity of substitution between below college and college equivalent education is ∞
H41: For college equivalent workers, the elasticity of substitution between immigrants and born citizens is ∞
H42: For born citizen workers, the elasticity of substitution between below college and college equivalent education is ∞
H43: For below college workers, the elasticity of substitution between immigrants and born citizens is ∞
H44: For immigrant workers, the elasticity of substitution between below college and college equivalent education is ∞
H45: For high school or more workers, the elasticity of substitution between immigrants and born citizens is ∞
H46: For all citizen workers, the elasticity of substitution between below college and college equivalent education is ∞
H47: For all workers, the elasticity of substitution between immigrants and born citizens is ∞
H48: For all workers, the elasticity of substitution between below college and college equivalent education is ∞
H49: For below high school workers, the elasticity of substitution between non-citizens and citizens is ∞
H50: For non-citizen workers, the elasticity of substitution between below and above high school education is ∞
H51: For high school equivalent workers, the elasticity of substitution between non-citizens and citizens is ∞
H52: For naturalized citizen workers, the elasticity of substitution between below and above high school education is ∞
H53: For college equivalent workers, the elasticity of substitution between non-citizens and citizens is ∞
H54: For born citizen workers, the elasticity of substitution between below and above high school education is ∞
H55: For below college workers, the elasticity of substitution between non-citizens and citizens is ∞
H56: For immigrant workers, the elasticity of substitution between below and above high school education is ∞
H57: For high school or more workers, the elasticity of substitution between non-citizens and citizens is ∞
H58: For all citizen workers, the elasticity of substitution between below and above high school education is ∞
H59: For all workers, the elasticity of substitution between non-citizens and citizens is ∞
H60: For all workers, the elasticity of substitution between below and at or above high school education is ∞