Causal Effect of Puerto Rican Emigration on United States Worker Income

(STAT 556 Final Project)

Qi Suo
Evans School of Public Policy
University of Washington

Kate Louie-Hess
Department of Applied Mathematics
University of Washington

March 14, 2018

Abstract

This paper estimates the causal effect of Puerto Rican emigration on United States workers' personal income with the data from 2016 American Community Survy (ACS). The study uses a propensity score matching method to reduce the difference of the covariates between the treatment and control groups, then estimates the average causal effect (ATE) for the whole population and the average causal effect on the treatment group (ATT).

1 Introduction

Every introductory course on statistics quickly introduces the concept that "correlation does not imply causation." However, many of the social sciences are interested specifically on the causal impact and relationships with current real-world outcomes. Through application careful statistical testing, we can in fact deduce some casual effects from empirical surveys. More specifically, we can also control for the impact of multiple covariates upon estimated effect via a process of matching data points that differ only in treatment status.

In this paper, we consider trait matching and causal effect estimation in the context of the relationship between personal worker income in the United States, and the birthplace of that worker. While there have been previous explorations of the causal effect of foreign-born worker income in the United States, U.S. workers born in U.S. territories present an interesting grey area. Specifically, we are interested in those mainland workers who were born in Puerto Rico, which represents a relatively large portion of the territory-born population. A conclusive causal effect on income would be interested in the context of possible discrimination or adverse effects stemming from being born on what is, technically, United States soil.

We use the data from the 2016 American Community Survey (ACS). The ACS is conducted every year by the U.S. Census Bureau as a means to both gather demographic data and evaluate the efficacy of government-funded quality-of-life actions.[1]

Our metrics of concern are primarily the average treatment effect (ATE) and average treatment effect on the treated (ATT). Our matching process employs propensity score matching, in both 1-1 and more relaxed variants.

2 Methodology and Theoretical Background

2.1 Causal Effects

Causal effect is the difference in outcome under treatment and control conditions.[2] It assumes that every individual in the sample has one potential outcome under treatment and one under control. We can simply use a vector to represent it: (Y_{0i}, Y_{1i}). The individual's observed outcome is denoted as Y_i. We represent treatment status as $D_i = 1$ if the individual is treated, $D_i = 0$ if untreated. In strictly observational studies, both average treatment effect (ATE) and average treatment effect on the treated (ATT) are relevant metrics.

$$\begin{aligned} ATE &= E(Y_{1i} - Y_{0i}) \\ ATT &= E[(Y_{1i} - Y_{0i})|D_i = 1] \end{aligned}$$

The ATT refers to the mean difference in potential outcomes of those in the treated population. The ATE refers to the mean difference in potential outcomes among the population as a whole.[2] In this study, we estimate ATE and ATT using a propensity score matching technique, which is outlined below.

The naive estimator for the ATT based on the observed data is[3]:

$$E[Y_i|D_i = 1] = E[Y_i|D_i = 0]$$

In a randomized study, the potential outcome is independent from the treatment, so the naive estimator is an unbiased estimator for the ATT. However, due to lack of randomization in a purely observational study, the treatment and control populations may have inherently different distributions of some outcome covariate variables. Then the ATT is redefined as:

$$\begin{aligned} E[Y_i|D_i = 1] &= E[Y_i|D_i = 0] \\ &= E[Y_{1i}|D_i = 1] = E[Y_{0i}|D_i = 0] \\ &= \{E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0]\} \\ &\qquad + \{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]\} \\ &= ATT + \text{selection bias} \end{aligned}$$

where selection bias is the difference in Y_{0i} in the treatment and control conditions. It is non-zero if potential outcome distributions vary by treatment condition.[2]

2.2 Propensity Score Matching[4]

In order to adjust the bias in the selection, we decide to use statistical adjustment, in particular matching. Matching is useful when the treatment and control have different distribution of all the covariates in the model, especially when having a large sample size in the control group and a small sample size in the treatment group. With the matching technique, we try to remove the difference in the distribution of the two groups and unify the two distinctive population in to one, as did in the randomized trial. Specifically, we match each unit with another unit in the other group:

$$m(i) = \arg \min_{j:D_j \neq D_i} \|X_j - X_i\|$$

where $\|X_j - X_i\|$ is a measure of distance between the two covariate matrices:

$$\|X_j X_i\| = (X_j X_i)' W (X_j X_i)$$

and W is a weighing matrix, generally defined as the inverse variance matrix,

$$diag\{\hat{\sigma}_1^{-2}, ..., \hat{\sigma}_k^{-2}\}.$$

Propensity score matching is the most common method in this situation. Propensity score is an model showing that how likely an individual is to be in the treatment group:

$$P(D_i = 1|X_i).$$

Recall the definition of uncofoundedness (that, conditional on the covariates, potential outcome is independent of treatment assignment):

$$(Y_{0i}, Y_{1i}) \perp D_i|X_i$$

As proven by Rosenbaum and Rubin in 1983[5]:

$$(Y_{0i}, Y_{1i}) \perp D_i|X_i \implies (Y_{0i}, Y_{1i}) \perp D|p(X_i), p(X_i) = P(D_i = 1|X_i)$$

That is, conditional on the propensity score, treatment assignment is essentially randomized.[4] Then for individuals with the same propensity score, regardless of distance between their covariate vectors, the difference between the treatment and control units actually gives an accurate estimate of the conditional ATE, $E[Y_{1i} - Y_{0i}|p(X_i)]$.

In application, the propensity score can only be estimated from the dataset, commonly with a logistic regression model. The regression model should consist of all the covariates that have association with both the treatment condition and the potential outcome. However, in many cases, there are usually some hidden variables that are not considered in the model.

The propensity score matching model must to be tested on whether it effectively reduces the distribution difference of the covariates for both treatment and control groups, i.e.,

whether it makes the two distinctive sub-populations resemble one unified population. The question of whether to keep unmatched individuals (who have no other-group counterpart) in the calculations of the treatment effect is also significant – though in our case, the focus on ATT suggests keeping all treatment datapoints regardless of matching status.

2.2.1 Bias in Matching

A certain amount of matching difference will exist in a dataset, which gives a slight bias to an estimator calculated via matching. At unit level for a treated unit, this bias is:

$$E[Y_0|X = X_i] - E[Y_0|X = X_{m(i)}].$$

A common way to counteract this unit-bias is to assume the function $E[Y_0|X]$ is of linear form, and approximately calculate it using the slope of an ancillary regression and the inner product of the matching discrepancy.[4]

3 Application

3.1 About the Dataset[1]

We use the data from the 2016 American Community Survey (ACS). The ACS is conducted every year by the U.S. Census Bureau as a means to both gather demographic data and evaluate the efficacy of government-funded quality-of-life actions. The ACS gathers data on both a household and individual level of granularity. The 2016 individual survey is comprised of approximately 3 million individuals, randomly sampled across the 50 U.S. states.

Of the approximately 3 million individuals sampled across the United States by the ACS in 2016, approximately 11,000 (or 1%) of the respondents were born in Puerto Rico. Among these people, approximately 1.9 million reported some level of income in the previous year, were aged 15 or older, and were born in either the United States or Puerto Rico. Approximately 1100 (or 0.5%) of this filtered group were born in Puerto Rico. We consider Puerto Rico birth our treatment, and US-birth our control.

In order to achieve reasonable data processing times, the U.S-born group was down-sampled (sampled without replacement) to match the treatment group, resulting in a total dataset size of approximately 22000 people.

3.2 Covariates and Descriptive Statistics

Some of the significant covariates with income are detailed in Figure 1.

The noted coefficients are congruent with common assumptions about the variable's relationship with income.

Figure 1: Covariate Coefficients

	coef	std err	t	P>|t|	[0.025	0.975]
age	194.6458	1.927	101.020	0.000	190.869	198.422
female	-2.044e+04	72.397	-282.367	0.000	-2.06e+04	-2.03e+04
hispanic	-2487.8288	141.715	-17.555	0.000	-2765.585	-2210.073
edu	3747.0132	8.510	440.313	0.000	3730.334	3763.692
disab	-1.712e+04	99.765	-171.588	0.000	-1.73e+04	-1.69e+04
english	-2.157e+04	160.546	-134.382	0.000	-2.19e+04	-2.13e+04

1. `age`: Age in years
2. `female`: Binary where female = 1
3. `hispanic`: Binary where Hispanic = 1
4. `edu`: Roughly the number of years of education, with some non-linearity involving graduate degrees
5. `disab`: Binary where disabled status = 1
6. `english`: Discrete cardinal variable from 1 to 4 indicating English-speaking ability, where 1 is perfect English, and 4 is no English

Table 1: Initial Covariate Distributions

	Controls ($N_c = 11033$)		*Treated* ($N_t = 11033$)		
Variable	**Mean**	**S.d.**	**Mean**	**S.d.**	**Raw-diff**
Income	42422.710	55298.620	31065.515	40150.126	-11357.195
Log-Income	10.070	1.226	9.817	1.112	-0.253

	Controls ($N_c = 11033$)		*Treated* ($N_t = 11033$)		
Variable	**Mean**	**S.d.**	**Mean**	**S.d.**	**Nor-diff(z)**
Age	50.752	19.560	52.519	18.431	0.093
Female	0.500	0.500	0.530	0.499	0.062
Hispanic	0.071	0.256	0.970	0.171	4.126
Education	18.270	3.221	16.331	4.804	-0.474
Disabled	0.194	0.396	0.272	0.445	0.185
English (4-1)	1.012	0.137	1.649	0.913	0.975

There are some distinct differences in control and treatment population distributions of some of our chosen variables. The mean age of both groups are around 50 years, with similar distribution graph. As for the total personal income, we find out that although the

two groups have similar distribution, but there is a distinct mean difference. The skewness of the two groups is also different. There are also some outliers in both groups. (See Figure 2 in appendix)

The years-of-education variable also indicates that there is a distinct difference between the treatment group and the control group. On average, the treatment group is 3 years less educated than the control group. Both groups have outliers showing that there are people having less than 9 years education. (See Figure 3 in appendix)

In Figure 4, the gender split is about 50-50 in both groups, though the treatment group has a slightly higher female balance on average. However, the most considerable subpopulation variable difference is that of Hispanic origin: approximately 93% of the control is non-Hispanic, and approximately 97% of the treatment is Hispanic. Considering the inverse correlation between Hispanic background and income, this distribution difference is striking.

3.3 Propensity Score Development

Throughout this project, the Python package `causalinference` was used, as it has a quite robust set of features for causal modeling and matching strategies.

Because of the differences in distributions of some covariates between the treatment and control groups, we implemented propensity score matching in an attempt to control for those covariates and attain a more accurate measure for ATE and ATT.

In developing the propensity score (the likelihood that an individual is in the treatment group, based on their covariate values), a logistic regression model was created. The construction of the model used a chain of likelihood ratio (LR) tests to determine the inclusion of covariates and whether they should be linear or quadratic in nature.[7] Variable selection is done algorithmically, as below:

1. Run a logistic regression of Y on a core set of covariates X_c
2. Introduce an additional variable X_t and rerun the regression. Find the LR test statistics for H_0 (X_t coeff $= 0$)
3. Repeat the previous step for each covariate. If the greatest LR statistic is above some threshold, the covariate should be included in X_c, and Step 2 should be repeated. If not, the covariate should be discarded.
4. Repeat Steps 2 and 3 for second-order components of the final X_c, using a different threshold for the quadratic terms

The specifications for the final logistic propensity model can be seen in Table 2. As expected, both Hispanic heritage and less-fluent English skills are quite strong predictors of having been born in Puerto Rico instead of the United States, both of which are characteristics generally considered more common for native Puerto Ricans. (Interestingly, being both Hispanic **and** having relatively poor English skills is a mild predictor for having been born in the United States!)

Table 2: Propensity Score Logistic Model

	Coef.	**S.e.**	**z**	$P > \|z\|$	**[95% Conf. int.]**	
Intercept	-7.92	0.557	-14.224	0	-9.012	-6.829
Hispanic	5.39	0.297	18.159	0	4.808	5.972
English	3.598	0.49	7.342	0	2.638	4.559
Age	0.054	0.01	5.169	0	0.033	0.074
Education	-0.059	0.036	-1.654	0.098	-0.129	0.011
Disability	0.276	0.092	3.002	0.003	0.096	0.456
Age * Hispanic	0.028	0.004	6.845	0	0.02	0.035
Age * Age	-0.001	0	-5.478	0	-0.001	0
Hispanic * English	-0.863	0.204	-4.24	0	-1.262	-0.464
English * English	-0.347	0.114	-3.035	0.002	-0.571	-0.123
Education * Education	0.003	0.001	2.41	0.016	0.001	0.005

3.4 Matching Results

After finding the propensity score for each individual, the dataset was trimmed of "extreme" propensity scores. These are data points that are strongly likely to be of a particular class, and as such are unlikely to have a close propensity score match (and attempting to include them in the matched set would introduce a considerable amount of unnecessary bias to the final estimators). We used benchmark values of 0.1 and 0.9 as our cutoff thresholds, leaving approximately 10,000 control samples and 8,000 treatment samples in the final set.

Matching was done on both a 1-1 and a slightly relaxed basis, wherein a given individual could have up to 3 matches in the other set. While this method does have the possibility to introduce some bias due to imperfect matching, it also lowers variance (due to the counterfactual estimates depending less on any given single unit). The covariate distributions of the 1-1, treatment-matched dataset can be seen in Table 3.

Table 3: Treatment-Matching Covariate Distributions

	Controls		Treatment	
Variable	**Mean**	**S.d.**	**Mean**	**S.d.**
Income	41232.4255	48562.12	38903.23	42343.98
Log-Income	10.0869	1.2750	9.991	1.112
Age	49.143	18.120	47.991	17.981
Female	0.5270	0.4992	0.516	0.500
Hispanic	0.156	0.56	0.912	0.234
Education	18.277	3.1098	17.65	3.235
Disabled	0.199	0.3998	0.195	0.396
English (4-1)	1.09	0.215	1.103	0.342

As seen in the table, the covariate distributions of the treatment and control groups in the matched dataset have become more similar (to some degree). While covariates like age and education are nearly identical in the new dataset, one significant exception is in

the distribution of the Hispanic binary variable. While the Hispanic mean has increased considerably in the control group, the distributions in the original datasets were dissimilar enough that even a good matching technique wouldn't bring them much closer together.

3.5 Treatment Effect Results

Table 4: Treatment Effects

Match		**Est.**	**S.e.**	$P > \|z\|$	**Est. Adj**	**S.e. adj**	$P > \|z\|$ **adj**
1-1	**ATE**	1338.659	2680.301	0.617	1252.275	2676.944	0.64
	ATT	-1491.905	3537.212	0.673	-1489.588	3532.612	0.673
1-3	**ATE**	640.51	1979.041	0.746	446.016	1980.339	0.822
	ATT	-2344.51	2552.084	0.358	-2216.394	2550.155	0.385

Regardless of matching procedure or estimate adjustment, the p-value of all treatment effect estimates (both ATT and ATE) is large and therefore none of the metrics are statistically significant. The estimates using log-Income are similarly inconclusive (See Figure 5.) We therefore cannot reject the null hypothesis that there is no average treatment effect of either variety (ATT or ATE).

4 Conclusion and Discussion

There is no identifiable causal effect of emigration from Puerto Rico on the resultant income in the US, versus being born in the United States. At this stage in analysis, it would seem that there is no evident bias against Puerto Rican-born workers as compared to US-born workers in the United States.

Possible extensions to this analysis include normalizing income by geographic cost of living, and examining the causal effect of birth in other territories (i.e. Guam or the Virgin Islands).

It is also important to note that our matching process hasn't fully resolved the covariate imbalance issues inherent to the dataset (especially in regard to the Hispanic population balance), and as such some bias likely still exists in our estimators. There are certainly other, unobserved covariates that are affecting the relative subpopulations in differing ways.

References

[1] US Census Bureau. (2016). *2016 American Community Survey Public Use Microdata Sample* [Data files and data dictionary]. Retrieved from https://www.census.gov/programs-surveys/acs/data/pums.html.

[2] Patrik Karlsson, Anders Bergmark. *Compared with what? An analysis of control-group types in Cochrane and Campbell reviews of psychosocial treatment efficacy with substance use disorders.* 2015.

[3] Stephen L. Morgan, Christopher Winship. *Counterfactuals and Causal Inference.*. 2015.

[4] Wong, Laurence. *Causal Inference in Python.laurence-wong.com/software.* 2015.

[5] Rosenbaum, P. & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70, 41-55.

[6] Joshua D. Angrist, Jrn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion.* 2008.

[7] Imbens, G. & Rubin, D. *Causal inference in statistics, social, and biomedical sciences: An introduction. Cambridge University Press.* 2015.

A Graphics

Figure 2: Age and Log-Income Distributions

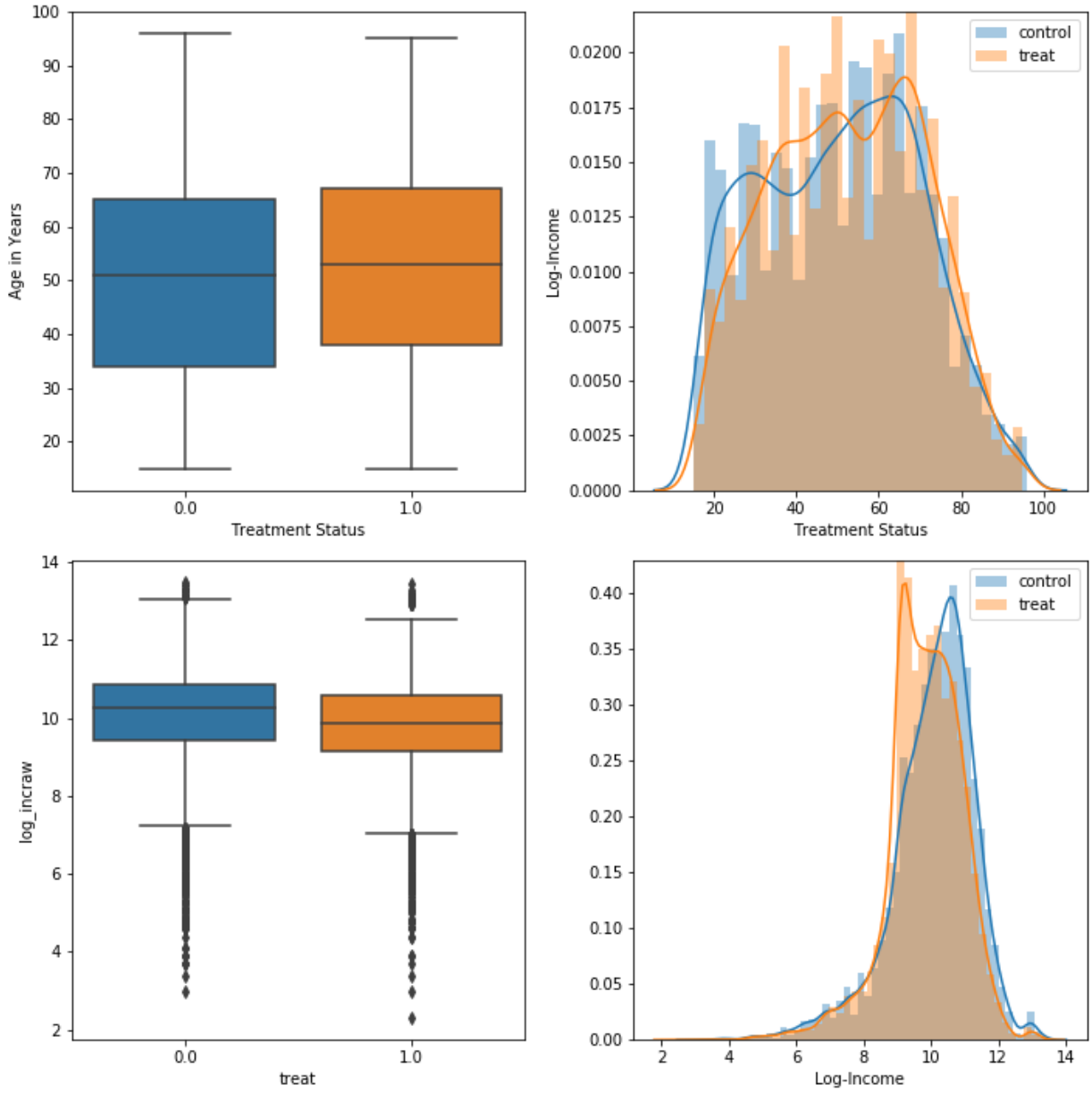

Figure 3: Years of Education Distribution

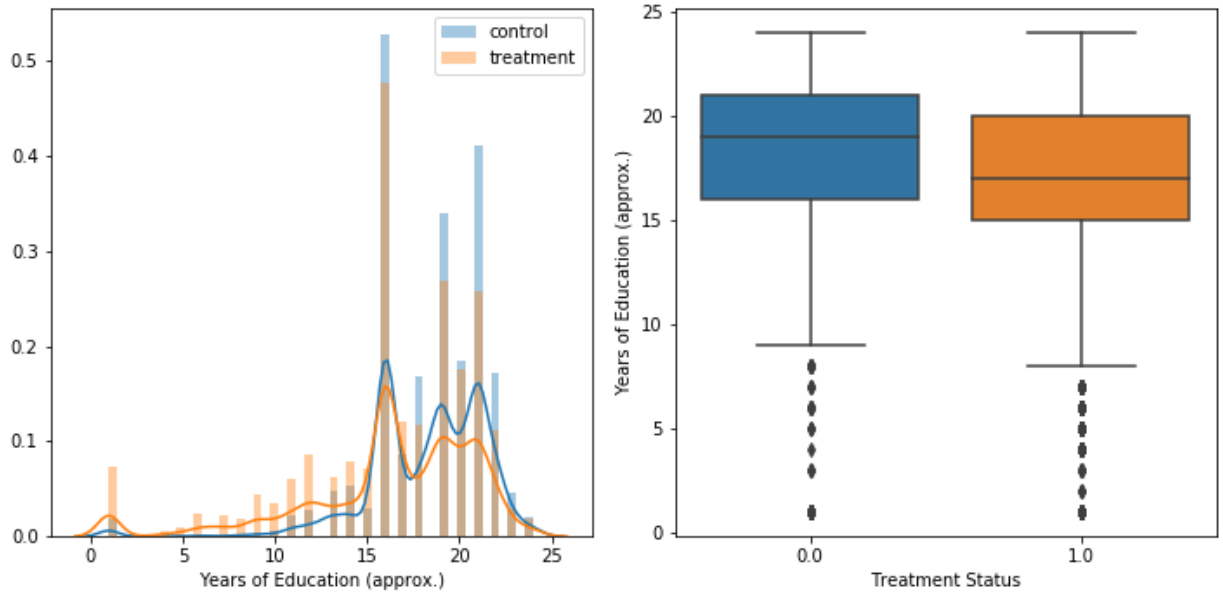

Figure 4: Gender and Hispanic Distributions

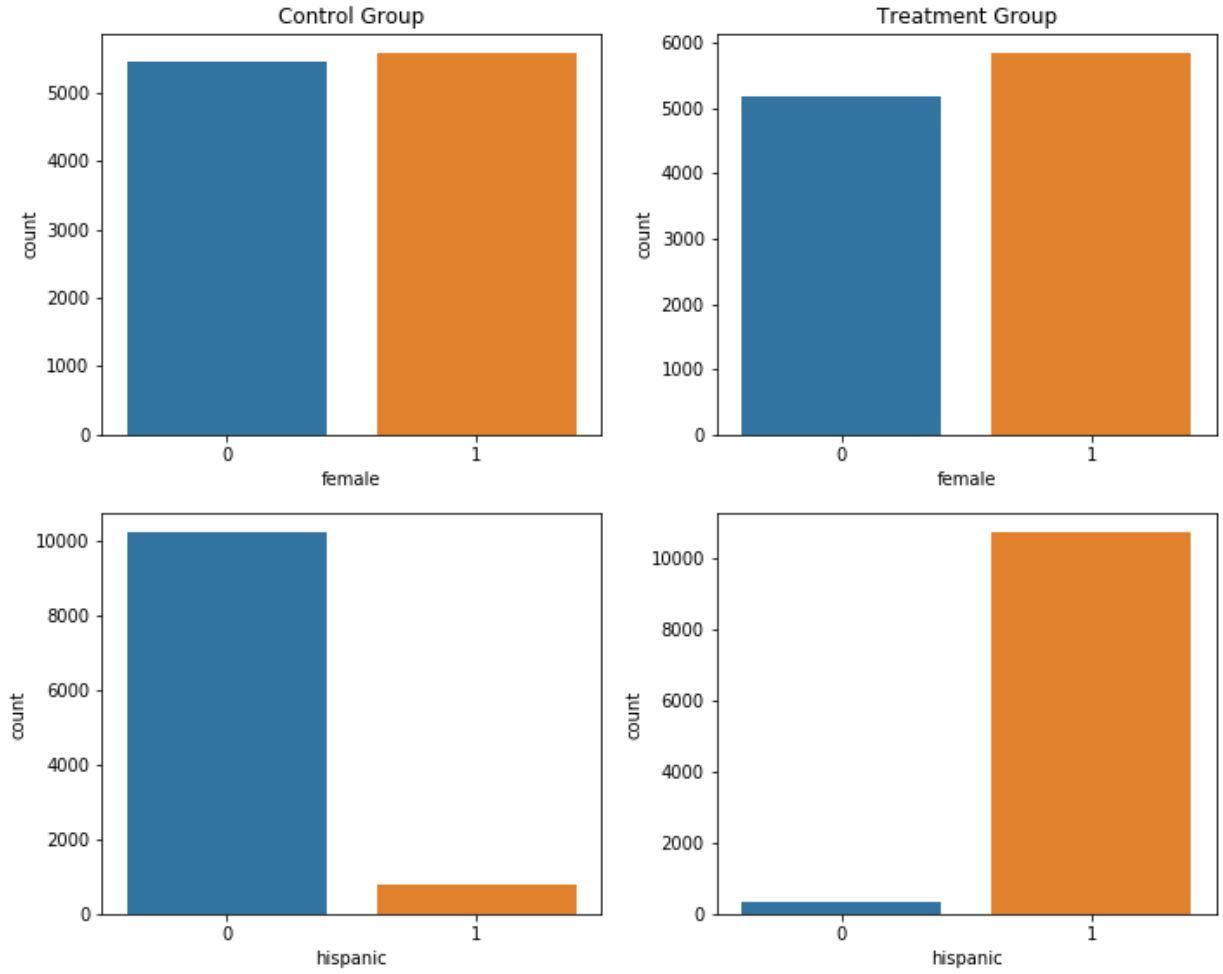

Figure 5: Treatment Effect Estimates for Log-Income

Treatment Effect Estimates: Matching (1:1)

	Est.	S.e.	z	P>|z|	[95% Conf. int.]	
ATE	0.008	0.067	0.122	0.903	-0.124	0.140
ATC	0.037	0.098	0.383	0.702	-0.154	0.229
ATT	-0.032	0.080	-0.404	0.687	-0.188	0.124

Treatment Effect Estimates: Matching (1:1, adj)

	Est.	S.e.	z	P>|z|	[95% Conf. int.]	
ATE	0.006	0.067	0.087	0.931	-0.126	0.138
ATC	0.035	0.097	0.355	0.723	-0.156	0.226
ATT	-0.034	0.079	-0.426	0.670	-0.189	0.122

Treatment Effect Estimates: Matching (1:3)

	Est.	S.e.	z	P>|z|	[95% Conf. int.]	
ATE	0.015	0.049	0.306	0.759	-0.081	0.111
ATC	0.062	0.072	0.860	0.390	-0.079	0.203
ATT	-0.050	0.055	-0.904	0.366	-0.157	0.058

Treatment Effect Estimates: Matching (1:3, adj)

	Est.	S.e.	z	P>|z|	[95% Conf. int.]	
ATE	0.010	0.049	0.201	0.841	-0.087	0.106
ATC	0.053	0.072	0.732	0.464	-0.089	0.194
ATT	-0.049	0.055	-0.904	0.366	-0.156	0.058

B Code

```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
import causalinference as ci
from causalinference import CausalModel
start = time.time()
data_a = pd.read_csv('data/ss16pusa.csv')
data_b = pd.read_csv('data/ss16pusb.csv')
end = time.time()
print("{} seconds".format(end-start))
data = pd.concat([data_a, data_b])
def pr_treat(col):
    if col <= 56:
        return 0
    elif col == 72:
        return 1
    else:
        return np.nan
dc = pd.DataFrame()
dc['id'] = data.SERIALNO
dc['puma'] = data.PUMA
dc['age'] = data.AGEP
dc['female'] = (data.SEX > 1).astype(int)
dc['hispanic'] = (data.HISP > 1).astype(int)
dc['edu'] = data.SCHL
dc['treat'] = data.POBP.apply(pr_treat)
```

```
dc['income'] = data.PINCP
dc['disab'] = (data.DIS == 1).astype(int)
dc['english'] = data.ENG.fillna(1)
dc['posinc'] = dc.income
dc.loc[(dc.posinc < 0), 'posinc'] = dc.loc[(dc.posinc < 0), 'posinc'] + 20001
dc.loc[dc.posinc == 0, 'posinc'] = 1
dc['loginc'] = np.log(dc.posinc)
dc['inc_raw'] = (data.OIP + data.PAP + data.RETP + data.SSIP + data.SSP + data.WAGP).ast
dc['inc_raw'] = dc.inc_raw.replace(np.inf, np.nan).replace(0, np.nan)
dc['log_incraw'] = np.log(dc.inc_raw)
print(dc.info(verbose=True, null_counts=True))
dc.dropna(inplace=True)
print(dc.info(verbose=True, null_counts=True))
samp = pd.concat([dc[dc.treat == 0].sample(11033), dc[dc.treat == 1]])
Y = samp.inc_raw.as_matrix()
Ylog = samp.log_incraw.as_matrix()
D = samp.treat.as_matrix()
X = samp[['age', 'female', 'hispanic', 'edu', 'disab', 'english']].as_matrix()
causal_lin = CausalModel(Y, D, X)
causal_log = CausalModel(Ylog, D, X)
causal_lin.est_propensity_s()
causal_log.est_propensity_s()
causal_lin.trim_s()
causal_log.trim_s()
causal_lin.est_via_matching(matches=1)
causal_log.est_via_matching(matches=1)
causal_lin.est_via_matching(matches=1, adj_bias=True)
causal_log.est_via_matching(matches=1, adj_bias=True)
causal_lin.est_via_matching(matches=3)
causal_log.est_via_matching(matches=3)
causal_lin.est_via_matching(matches=3, adj_bias=True)
causal_log.est_via_matching(matches=3, adj_bias=True)
```