SHORT ANALYSIS: CAUSAL EFFECT OF ENGLISH LANGUAGE PRE-TRAININGS ON EMPLOYMENT OF REFUGEES IN THE USA

Causal Data Analysis

Abstract

The purpose of this work is to measure the effect of pre-trainings in English on the employability of refugees in the USA using methods of Causal Data Analysis. The obtained result is an additional 6.7 percentage points in the probability of employment for refugees who attend such trainings as compared to those who don't.

Stephen Kaspan and Mehdi Jaddour

Corvinus University of Budapest, Fall 2022

Introduction:

According to the 1951 Refugee Convention, the definition of a refugee is: “someone who is unable or unwilling to return to their country of origin owing to a well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group, or political opinion.”. Because of increasing tensions around the world, this term has become trending, and more people are presently bearing the name. The United States of America, besides the recent troubles in its resettlement system, has always been recognized as one of the biggest hosts of refugees. The act of hosting refugees is not limited to opening the country’s boarders in the face of these newcomers, it also comprises integrating these people in their new environment. One of the predominant figures of integration is employment, the US government grants work permission to refugees, they are, furthermore, assisted by the relevant organizations to find appropriate job opportunities. However, many other factors might facilitate or hinder employment for refugees, for instance, the personal attributes or skills of the refugee; It is plausible that one who is in good health conditions would easily find a job relatively to a refugee with disabilities. In this study, we orient our attention towards one of these factors – the ability to speak English. The aim of this work is to make use of Causal Data Analysis methods, principally Matching methods, in order to uncover the effect of English language pre training programs on the employability of refugees in the United States of America. The first part of the analysis is dedicated for methods specification, the second for exhibiting the results of the study (obtained using the software Stata), and we conclude with the discussion of these results.

Methods:

Hypothesis and mechanism:

Before we go further, we clearly state our causal question as follows: What is the effect of a pre training in English language on the employability of a refugee in the US labor market? We hypothesize that the pre training in English is likely to increase a refugee’s chances to find employment in the United States of America for two main expectations about the relevant mechanism. The first is that, for some kinds of jobs, applicants are required to have a minimum level of English language proficiency, training programs are likely to endow refugees with this ability and provide certified proofs to be used during applications. The second is that the English language training will enable the refugee to manage an appropriate job search procedure (creation of CV, filling information, interviews…).

Data:

We obtained our data from the Office of Refugee Resettlement (ORR) which conducts Annual Survey of Refugees (ASR). ASR collects information on refugees during their first five years after arrival in the U.S. It also gives information on progress that refugee families made towards learning English, participating in the workforce, and establishing permanent residence. We used the 2019 ASR public data file that is shared by the UN Refugee Agency on its Microdata library through an external link sending to the OPENICPSR website, where the Urban Institute has published the project "2019 Annual Survey of Refugees". The population of interest for the 2019 ASR is a sample of 4905 refugees entering the U.S. between 2014 and 2018 (inclusive), aged above 16 years at the time of survey. The five distinct fiscal years of refugee entrants were collapsed into three cohorts depending on year of arrival to the U.S.

- Endogeneity issues and Proxies: Since our work is based on observational data we had to consider the potential sources of endogeneity that might bias the results of the analysis. We restate that the causal variable is the taking of a pre training program in English language (variable in the model: english_train), and the output variable is the finding of an employment in the USA (variable in the model: work). The major sources of endogeneity we considered in our study are as follows (between parenthesis is the name of the related proxy)
  - Common causes: After brainstorming we judged that these variables might be correlated to both the treatment and the output.
    - Disabilities (disability): binary variable for the presence of a health disability
    - Previous qualifications: binary variables of employment and obtention of a degree before arrival to the USA (employed_bef, school_bef)
    - Learning methods other than training programs: the only proxy available is the measure of the level of English at arrival, but this is also a metric related to the mechanism, and thus after testing we decided to drop it from the model
    - Gender: binary variable for gender (female)
    - Previous experience or studies in American organizations (unavailable proxy)
    - Previous relationships with people from the USA (unavailable proxy)
    - The year of arrival to the USA (qn1year)

- Goals other than work: binary variable that tells whether the refugee is studying at the USA (school)
- Age: Age category of the refugee at the time of arrival to the USA (ui_agect_arrival)

- Unwanted mechanism:
  - Learning English during the stay at the USA: refugees who did not have access to pre-training programs in English language will be encouraged to put in more effort during their stay in the USA to learn the language. This will increase their chances of finding employment. On the other hand, refugees who have received pre-training are better equipped for self-directed English learning, and their proficiency in English can improve more rapidly after arrival, which in turn facilitates their employment.

Models:

- Simple linear regression:

  The first method we apply is a simple linear regression. Our objective is to visualize whether there is a correlation between the treatment and the output, and to create a baseline model that would serve as a reference for comparison with the subsequent model. We expect that the causal estimate in this phase will be biased because no action is taken to remedy the endogeneity issues. The simple regression equation is the following:

$$work = \alpha + \beta.english\_train$$

- P-score matching:

  The amelioration we brought is to control for the preselected endogenous variables. We employed P-score matching (1NN without caliber) in order to compare between subjects in the treatment and control groups that bear a similar degree of endogeneity as determined by the P-score.

$$Pscore = LOGIT(english\_train|\ disability, employed\_bef,$$
$$school\_bef, female, qn1year, school, ui\_agect\_arrival, eng\_interview\ )$$

Results:

Descriptive statistics:

	1		2		3	
	10	11	20	21	30	31
Prop_out	.576	.840	.782	.387	.827	.583
Prop_treat	.279	.327	.341	.146	.314	.291

Table 1: Proportions of employed refugees and treated refugees by employment before arrival, presence of disabilities, and gender.

Legend:

Prop_out: Proportion who got a job in the USA.

Prop_treat: Proportion of treated.

1: Employment before arriving to the USA. (10: wasn't employed, 11: was employed)

2- Disabilities. (20: without disabilities, 21: with disability)

3- Gender. (30: Male, 31: Female)

Simple linear regression:

VARIABLES	Without covariates
english_train	0.16**
	(0.02)
Constant	0.65**
	(0.01)
Observations	2,591
R-squared	0.03

Standard errors in parentheses
** p<0.01, * p<0.05

Table 2: Simple regression results.

The estimated coefficient for the treatment effect (0.16) and the constant are significant at 1%.

Propensity Score Matching

```
Treatment-effects estimation                Number of obs      =      2,591
Estimator      : propensity-score matching  Matches: requested =          1
Outcome model  : matching                               min =          1
Treatment model: logit                                  max =         28
```

work	Coef.	AI Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATE						
english_train (Treated vs Not treated)	.0672126	.0233229	2.88	0.004	.0215007	.1129246

Table 3: Propensity score matching results.

The coefficient of the treatment variable (ATE) is 0.07. The 95% confidence interval lies between 0.02 and 0.11, inclusive.

Discussion:

In table 1 we try to gain insights about the distribution of both the treatment and the output between the subcategories of three variables representing potential sources of endogeneity. Clearly, the results hint the presence of positive bias caused by all variables, for instance, the proportion of positive output and the proportion of treated subjects are both higher for people without disability as compared to those with disability (the same can be deduced from analyzing the correlation matrix of the three variables with the output and treatment variables, for example, disability is negatively correlated with both the output and the treatment). This remark is confirmed by the decrease in the causal estimate between linear regression (0.160 as shown on table 2) and P-s matching (0.067 as shown on table 3). Therefore, according to our identification assumption, the P-s matching method has removed more positive bias than negative from the causal effect estimate. Our final result indicates that people who attend a pre training program in English language have a 6.7 percentage points higher probability to be employed as compared to those who don't. However, before we base any proposal for interventions on this result, we need to question its internal validity. The first limit is that the treatment is not the same for all treated subjects; treated subjects might had done different pre trainings whose efficiency is unlikely the same. Second, around half the population of subjects' data had been provided by household representatives and not the person in question, this leaves more room for uncertainty about the exactitude of the information. Third, we couldn't find a

proxy for three of the potential sources of endogeneity. Finally, we believe that more insights about the characteristics of the USA labor market and the process of refugee integration could result in the determination of other sources of endogeneity. With this in mind, we can say that the US government should decide on whether to make English training programs compulsory during the first days of arrival of refugees, and so free of charge, depending on the cost of supplying these trainings and the cost of unemployment per refugee. For example, supplying 100 additional trainings would increase, on average, the number of employed refugees by six. Therefore, the return generated is six times the unemployment cost per refugee. As a conclusion, we share some thoughts about the external validity of our study. Whether this result is valid for other countries respective to their national languages depends on many factors such as the characteristics of the labor market. For instance, if the labor supply is low, then the absence of competition would result in less causal effect of language pre trainings, other factors might be the criteria of selection of refugees and the way they are managed by the national authorities.

References:

Source of the data: UNHCR Microdata library
https://microdata.unhcr.org/index.php/catalog/681/get-microdata

1951 Refugee Convention: UNHCR website https://www.unhcr.org/3b66c2aa10.html

Békés, Gábor, and Kézdi, Gábor. Data Analysis For Business, Economics, And Policy. Cambridge University Press, 2021.