Earnings Distribution of Cuban Immigrants in the U.S.

Evidence from Quantile Regression with Sample Selection

Aleida Cobas Valdés
Institute of Public Economics. University of the Basque Country(UPV/EHU)
aleida.cobas@ehu.eus

Abstract

We analyze the conditional earnings distribution for Cuban immigrants in the U.S. considering Buchinsky (1998, 2002) sample selection in a quantile regression model and the test proposed by Huber and Melly (2015) to test the independence between error terms and regressors conditional on the selection probability. This is the first attempt in the migration literature to use quantile regression with sample selection taking into account the independence of errors test. The data used in the study come from the Census of Population and Housing in the U.S. provided by IPUMS (2013). The results show that increments in earnings associated with various socioeconomic characteristics, such as gender, marital status, ethnicity, proficiency in English and years of education, vary across the whole earnings distribution as well as between the cohorts considered and the hypothesis of conditional independence is not rejected.

Introduction

In terms of educational level, Cuban immigrants have positively self-selected in their migration decision to move to the U.S., that is, people with the highest levels of education migrate (Cobas and Fernández, 2014). In countries with low returns to skill and low wage dispersion, there will be positive selection of immigrants (Borjas, 1987). The U.S. has been the main destination for migrants from Cuba and other Latin American countries over the last century. Some recent research into the earnings distribution of Latino immigrants in the U.S. reveal considerable disadvantages with respect to native-born people. Since it is the most highly skilled people who migrate, it is undoubtedly of interest to describe the whole distribution of earnings of people in the host country and not only the mean and to explore differences between different periods comparing the role of socioeconomic characteristics across all quantiles of this earnings distribution. In the analysis of earnings, however, we can only observe wages when individuals are in work. This problem is known as the sample selection problem, that is, the variables of interest are only observed for a non-random subsample of the population.

Main Objectives

1. To describe the earnings distribution of Cuban immigrants in the U.S and compare the socioeconomics characteristics of three cohorts: people who migrated to the U.S. arriving in the 1980s, the 1990s and the 2000s.
2. To use the sample selection correction for quantile regression proposed by Buchinsky (1998, 2002) to estimate the effects of different socioeconomic characteristics on the conditional probability distribution of the earnings of Cuban immigrants in the U.S.
3. To test the conditional independence assumption using the recent tests proposed by Huber and Melly (2015), to our knowledge, the first time this approach has been used in immigration research.

The estimated model

We first estimate the traditional equation (Mincer, 1974) for Cuban immigrants in the USA considering the sample selection quantile regression model proposed by (Buchinsky, 1998, 2002). After that, we check the assumption of independence using the test proposed by Huber and Melly (2015).

Let y_i be the (log) gross hourly earnings for individual i, using the total pretax wage and salary income (expressed in contemporary US dollars), $\boldsymbol{x_i}$ a vector of $(k\times 1)$ socioeconomic characteristics of Cuban immigrant in the USA and $Q_\tau\left(y_i|\boldsymbol{x_i}\right)$ is the conditional τ-quantile of the distribution of y_i given $\boldsymbol{x_i}$.

Then, we define the wage offer equation or outcome equation in terms of a latent variable, y_i^*, which depends linearly on a vector of characteristics $\boldsymbol{x_{2i}}$:

$$y_i^* = \boldsymbol{x_{2i}'\beta_0} + u_i\,. \tag{1}$$

We do not observe the latent variable y_i^* but only y_i, that for people who are working, and hence, we have

$$y_i = D\cdot y_i^* = D\cdot\left(\boldsymbol{x_{2i}'\beta_0} + u_i\right) \tag{2}$$

where $D \equiv I\left(\boldsymbol{x_{1i}'\alpha_0} + v_i \geq 0\right)$. $I(\cdot)$ being the usual indicator function, defined as

$$I\left(\boldsymbol{x_{1i}'\alpha_0} + v_i \geq 0\right) = \begin{cases} 1 \text{ if } & \boldsymbol{x_{1i}'\alpha_0} + v_i \geq 0\,, \\ 0 \text{ if } & \boldsymbol{x_{1i}'\alpha_0} + v_i < 0\,. \end{cases} \tag{3}$$

Rewriting Equation (1) considering quantile regression, we have

$$y_i^* = \boldsymbol{x_{2i}'\beta_\tau} + u_{\tau i}, \qquad 0\leq\tau\leq 1 \tag{4}$$

and then we have to estimate the following quantile regression model

$$Q_y\left(\tau\,|\,\boldsymbol{x_1}\,, D=1\right) \equiv \boldsymbol{x_2'\beta_0} + h_\tau\left(\boldsymbol{x_1'\alpha_0}\right) \tag{5}$$

where $h_\tau\left(\boldsymbol{x_1'\alpha_0}\right) \equiv Q_u\left(\tau\,|\,\boldsymbol{x_1'\alpha_0}\,, D=1\right)$. Huber and Melly (2015) propose testing the assumption of conditional independence between the error term and the vector of covariates. The testing approach is based on the Kolmogorov-Smirnov and Cramér-von-Mises statistics.

The testing problem is given by

$$\begin{aligned} &H_0 : \beta\left(\tau\right) = \beta(0.5)\,, \quad \forall\,\tau\in\mathcal{T}\,, \\ &H_1 : \beta\left(\tau\right) \neq \beta\left(0.5\right)\,, \quad \text{for some} \quad \tau\in\mathcal{T} \end{aligned} \tag{6}$$

where $\mathcal{T}$ is a closed subset of $\left[e, 1-e\right]$, $0<e<1$ and $\beta\left(\tau\right)$ denotes the true τ quantile regression coefficient defined in our framework as

$$\beta\left(\tau\right) = \underset{\beta}{\operatorname{argmin}}\, E\left[\rho_\tau\left(y_i - \boldsymbol{x_{2i}'\beta} - h_\tau\left(\boldsymbol{x_{1i}'\alpha_0}\right)\right)\right]. \tag{7}$$

Data and Results

We use repeated cross-sections between 2000-2007 dividing the sample in three cohorts: people who migrated to the USA arriving in the 1980s, the 1990s and the 2000s. We restrict our sample to individuals who were between 25 and 55 years old, worked 60 hours or less weekly during the year preceding the census and entered the USA when they were between 17 and 49 years of age.

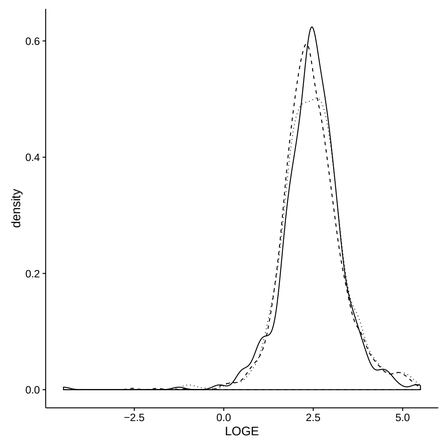

Figure 1: Kernel Density estimates for (log) hourly earnings

The vector of regressors $\boldsymbol{x_2}$ contains indicators for being black, being a citizen of the United States of America, proficiency in English, years of education, years of education squared, potential experience. The vector $\boldsymbol{x_1}$ contains the same variables as $\boldsymbol{x_2}$ plus an indicator for being a woman or married, an interaction variable that reflects the individual being married and a woman, and age at time of migration.

We have estimated the outcome equation including only individuals who were working at the time of the census. We considered quantiles from 0.05 to 0.95.

The conditional quantiles change differently across the three cohorts considered. In all the cohorts, being black has a negative effect on wages for people at the bottom of the earnings distribution.

For proficiency in English, the curve follows a U-shape for migrants in the 2000s, meaning that the effect of this variable on hourly earnings is similar at the extremes of the distribution.

For 2000s migrants, (log) earnings were clearly negatively related to being an American citizen across the wage distribution, while for workers who migrated in earlier decades, the contribution tended to be positive.

The null hypothesis of independence is not rejected in any cases and all variables contribute to the non-rejection of this hypothesis. This implies that we do not reject random selection or independence between the error terms and covariates in any quantiles of the distribution, and that the coefficients are consistently estimated by the Buchinsky (1998, 2002) method.

	Kolmogorov-Smirnov			Cramér-Von Mises		
	$1980s$	$1990s$	$2000s$	$1980s$	$1990s$	$2000s$
Number of observations	1137	2559	543	1137	2559	543
All variables	0.621	0.789	0.562	0.710	0.892	0.825
Years of education	0.974	0.595	0.788	0.967	0.778	0.863
Years of education squared	0.997	0.784	0.929	1.000	0.903	0.879
Experience	0.919	0.022	0.198	0.765	0.207	0.170
Experience squared	0.941	0.020	0.180	0.735	0.170	0.190
Is black	0.474	0.462	0.933	0.318	0.368	0.826
Is an American citizen	0.973	0.924	0.495	0.803	0.896	0.687
English Proficiency	0.147	0.922	0.688	0.090	0.926	0.710

Note: The figures shown are the p-values of the corresponding tests.

Table 1: Independence Test Results.

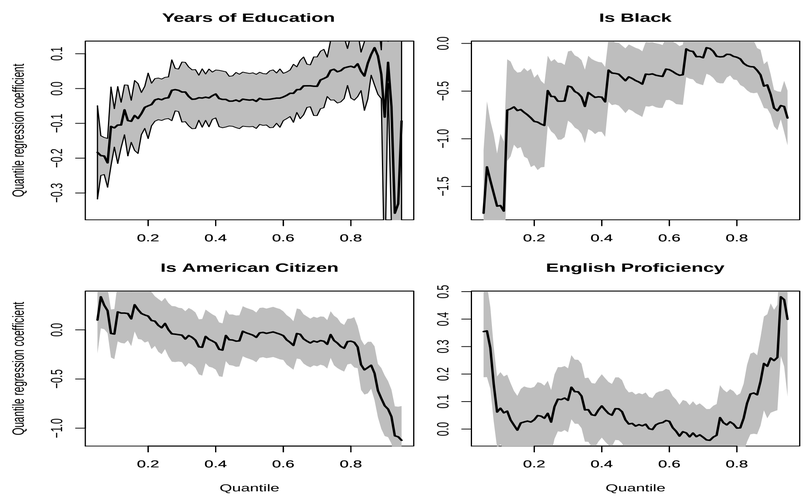

Figure 2: Quantile Regression Estimation for 2000s

Conclusions

- The results show that the role of the socioeconomic variables considered vary across the cohorts analyzed and we not reject the hypothesis of independence; therefore, the assumption of random selection or independence between the error terms and covariates is not rejected at any quantiles of the distribution.
- The main finding of our study is that there are differences not only in the socioeconomic characteristics of people grouped by time of arrival but also in their influence on the earnings of Cuban immigrants across the quantiles of the earnings distribution.
- The decline in returns from education for recent cohorts may be a sign that a high level of education no longer provides a competitive advantage, since Cubans with higher levels of education are those who migrate and once in the USA not all work in jobs commensurate with their level of education.
- The fact of being black is associated with significantly lower earnings across the working population, regardless of individuals' position in the earnings distribution.

Acknowledgements

The author thank Manuel Arellano, Alejandro Badel and James Albrecht for particulary helpful comments and Blaise Melly for his help in developing the R code used for independence testing.