Recovering the Counterfactual Wage Distribution with Selective Return Migration

Costanza Biavaschi*

This Draft: January 13, 2012

Abstract

This paper studies what the immigrant wage distribution would be in the absence of return migration. In particular, it recovers the counterfactual wage distribution if all Mexican immigrants were to stay in the United States and no out-migration of Mexican born workers occurred. Due to the presence of self-selection, the overarching problem addressed by this study is the development of a consistent estimator for the counterfactual density of interest. I propose a semi-parametric procedure that recovers this distribution. I find that Mexican returnees are middle to high wage earners, at all levels of human capital.

JEL Classification: J61, F22
Keywords: return migration, self-selection, assimilation

1 Introduction

For longer than a century the United States have been the primary destination for international migrants. Still in 2000, the U.S. was chosen as primary destination of world migrants and hosted a fifth of them (Özden, Parsons, Schiff and Walmsley, 2011). It is not surprising, then, that the immigration debate has often been pivotal in the political and academic arena of this country. Why do people migrate? How do migrants fare in the host country? And what is the impact of immigration on the native-born population of the host country? A vast and valuable literature has investigated these questions. Often, however, this research has failed to address the fact that some fraction of immigrants return to their home country. Taking into account such return migration compels us to re-think how we model the migration decision as well as how we measure the effects of migration on both immigrants and natives, on both sending and receiving regions. Immigrant performance and immigrant assimilation in the host country might have been over or underestimated depending on how returnees compare with the permanent settlers. The return flow might mitigate the effects that immigration has on the labor market outcomes of native workers, in particular for those groups of natives within the same skill group of the returnees. At last, the fiscal costs associated to immigration might depend on the composition of the return flow.

This paper combines data from the U.S. and Mexican Censuses of 2000 to estimate the wage distribution of Mexican immigrants in the U.S. both with and without return migration.

*Institute for the Study of Labor (IZA). Emai: biavaschi@iza.org.

Some important contributions have attempted to examine the average wage of immigrants (Hu, 2000; Lubotsky, 2007) or Mexican immigrants (Lindstrom and Massey, 1994; Reinhold and Thom, 2009; Lacuesta, 2010) taking into account return migration. The major challenge faced by all these studies is the absence of administrative data collected by immigration authorities. As a consequence, these studies either do not have information on return migrants per se but rather infer return migration from attrition from survey samples (Hu, 2000; Lubotsky, 2007), or do not use representative samples (Lindstrom and Massey, 1994; Reinhold and Thom, 2009; Lacuesta, 2010), or argue the unimportance of selectivity and compare Mexicans in Mexico with and without experience in the U.S. (Lacuesta, 2010).

This paper uses representative data, accounts for selection on both observable and unobservable attributes of the migrants, and focuses on their *wage distribution*. This not only sheds light on the earning opportunities of returnees in the host country, but also yields a more profound understanding of where in the distribution of wages selective return migration has its largest impact. Due to the presence of self-selection in the return choice, the overarching problem is to recover the counterfactual wage density in presence of selective return migration. The paper implements a semiparametric procedure that applies to density estimation the method in Andrews and Schafgans (1998), that complements in spirit the estimator used by Chiquiar and Hanson (2005) which accounted for selection on observable traits, and provides an alternative to the use of pre-migration earnings as a measure of selectivity (Kaestner and Malamud, 2010; Fernandez-Huertas Moraga, 2011; Ambrosini and Peri, 2011; McKenzie and Rapoport, 2010), when such information is not available to the researcher.

I find that, conditioning on observable characteristics, Mexican returnees are middle to high wage earners, consistent with models in which the decision to return hinges on reaching desired goals in the host country. Overall, the return flow has a small effect on immigrant wage inequality: the outflow of immigrants decreases dispersion in the lower part of the distribution and it increases it in the upper part. Selective return migration does not have a constant effect across educational levels: while it increases inequality at low levels of education, it decreases inequality for the highly skilled. These results suggest that in designing optimal migration policies policy makers should consider that selective outmigration might have a greater impact at high levels of human capital. At last, because at all levels of education the immigrants who leave are the high-wage earners, the immigrant-native wage gap would slightly close if there was no return migration.

The paper is organized as follows. Section 2 presents a brief overview of the literature. Section 3 describes the data. Section 4 presents the estimation technique and Section 5 and 6 the results. Some conclusions are drawn in Section 7.

2 Immigration, Return Migration and Self-Selection across the Mexican-U.S. border

2.1 Immigration and Self-Selection

The study of the type of migrants who decide to move has been an area of intensive debate since the early contribution by Borjas (1987). Although this paper focuses on the selection of the return migrants, it has its fundaments in the debate on the selection of immigrants. While the empirical evidence on the selection of the migrants at the international level seems to suggest that on average migrants are positively selected (Belot and Hatton, 2008; Grogger and Hanson, 2008; Rosenzweig,

2007), focusing only at the Mexican-U.S. migration, the evidence is mixed, and several recent contributions have expanded the debate in the area.

Combining the U.S. and the Mexican Census, Chiquiar and Hanson (2005) contradict Borjas' theoretical predictions showing intermediate to positive selection based on observable characteristics for Mexican emigrants. In contrast, using only the information provided in the Mexican Census about experience in the U.S., Ibarraran and Lubotsky (2007) find negative selection on observable skills. They identify various causes for the lack of consistent results. Of particular importance there are: the mistranslation or misunderstanding of the grade and degree choices in the U.S. Census that is possibly causing the misreporting of the education variable - the key factor in studying selection; and the undercount of the young and largely illegal Mexican immigrants in the U.S. Census. Fernandez-Huertas Moraga (2011), using the Encuesta Nacional de Empleo Trimestral, is able to quantify the source of the discrepancies between the previous two studies, showing that Chiquiar and Hanson's results are primarily driven by the undercount of unskilled immigrants in the U.S. and secondarily by the omission of unobservables in the estimation procedure.

A few other studies in the U.S. context have pointed to various other driving factors of self-selectivity in the migration decision. McKenzie and Rapoport (2010) use the Encuesta Nacional de la Dinámica Demográfica, and reconcile Borjas's theoretical framework with Hanson's empirical findings by suggesting that the results are driven by the differential impact of education in community with large and small networks. Kaestner and Malamud (2010) use the Mexican Family Life Survey and suggest that migrants are selected from the middle of the education distribution, and middle of the wage distribution, although after controlling for network effects, these results are not equally strong. Interestingly, the authors find little evidence of selection on unobservables. Using the same data source, Ambrosini and Peri (2011) confirm the findings in Borjas (1987) and Fernandez-Huertas Moraga (2011) on the negative selection of Mexican immigrants to the U.S., result primarily driven by differences in unobservable skills between migrants and non-migrants.

The current debate has brought to the attention of the researchers the importance of two key elements in the analysis of the selectivity of migrants. First, the necessity of using nationally representative data sources with a longitudinal component able to capture the pre-migration earnings of migrants and non-migrants (Kaestner and Malamud, 2010; Fernandez-Huertas Moraga, 2011; Ambrosini and Peri, 2011; McKenzie and Rapoport, 2010). Second, the importance of controlling for unobservable differences between migrants and non-migrants (Fernandez-Huertas Moraga, 2011; Ambrosini and Peri, 2011).

2.2 Return Migration and Self-Selection

While the literature on the selection of immigrants has been growing in the recent years, less has been done about the selection of the returnees and its impact on wage estimates in the sending or receiving economies. Yet, return migration is a predominant phenomenon: historic trends show that about 30% of the immigrants leave the host country (Jasso and Rosenzweig, 1982). Borjas and Bratsberg (1996) estimate that on average a third of immigrants left the U.S. between 1970 and 1980, with outmigration rates ranging from 3.5% of the Asian immigrants to 34.5% of South American immigrants.

The early contributions of Borjas (1989) and Chiswick (1986) draw attention on the importance of selective return migration when studying immigrant assimilation. Later a theoretical contribution by Borjas and Bratsberg (1996) showed that the average skill level of the returnees accentuates the original selection process of the immigrants: for example, if immigrants are the low skilled in their

origin country, returnees will be those with the lowest skills of this group; vice versa if immigrants are the high skilled in their origin country, returnees will be those with the highest skills of this group. Depending on the answer to who out-migrates, the consequences for the host economy will be different. If, for example, the immigrants with above average skills are those who leave, immigrant progress in the host country would be underestimated due to the presence of selective out-migration. On the other hand, if those who leave have below average skills, the destination country would be hosting only the 'best' immigrants.

The overall evidence for the U.S. economy suggests that returnees have below average skills, and therefore selective return migration has induced an overestimation of the economic progress of immigrants (Lubotsky, 2007): comparing longitudinal and cross-sectional data, Lubotsky finds that return migration by low-wage immigrants has systematically led past researchers to an overestimation of 10 to 15% of the wage progress of immigrants who remain in the United States. Likewise Hu (2000) shows lower immigrant wage growth once return migration is taken into account. Such results are however less strong in the subgroup of Hispanic workers and, due to data confidentiality, Lubotsky suggests that more analysis would be needed.

The literature that explicitly looks at the selection of Mexican returnees has a few contributions. Lindstrom and Massey (1994) combine U.S. Census data and Mexican survey data to study the selection of return migrants. They find that selection does not bias the cross-sectional wage estimates.[1] Lacuesta (2010) and Reinhold and Thom (2009) provide evidence of selection and skill upgrading for Mexican returnees in Mexico. Lacuesta (2010) uses the Mexican Census and with a pooled OLS regression of Mexican non-migrants and returnees estimates that temporary migration increases wages upon return by 7%-10%. However, the Mexican Census misses information on the Mexican migrants who stay in the U.S. Although the author provides suggestive evidence that return migrants are a close comparison group to non-returning migrants, it is still possible that much of this increase might be caused by the type of selection of the return migrants that was not accounted for in the analysis. Reinhold and Thom (2009) use the Mexican Migration Project, and estimate the return to U.S. labor market experience for a randomly selected migrant, correcting for the endogeneity of the migration decision. They provide unbiased estimates of the returns to foreign-experience, self-selection and skill upgrading, focusing also on differences across documentation statuses. They find that returnees are negatively selected in terms of unobservable traits, although selection is not significant in their analysis. At last, Ambrosini and Peri (2011) find preliminary evidence that returnees are positively selected compared to the non-migrants and to the permanent migrants.

The main challenge in studying return migration is the difficulty and the importance of obtaining comprehensive data sources on these flows. While great advances have been made in the study of Mexican-US migration thanks to the availability of new, longitudinal, representative dataset, in the study of return migration data sources are still lacking. Hu (2000) and Lubotsky (2007) provide interesting insights on the nature of return migration and its impact on the host economy, however in their longitudinal datasets returnees are not directly identified, and return migration cannot be separated by other sources of panel attrition.[2] Furthermore, the estimation techniques proposed in

[1] Massey (1987) and Reagan and Olsen (2000) analyze the selectivity of return migrants but do not study the implications in the estimation of their wage progress in the host country.

[2] In particular, these authors identify non employment, outmigration, employment in the informal sector, and non-match as possible causes of panel attrition.

these studies do not directly control for the possibility of non-random sample selection and the results are therefore valid only under the assumptions that selection operates purely through unobserved heterogeneity and, hence, that selection is a stable process over time. Massey (1987), Reagan and Olsen (2000) and Reinhold and Thom (2009) base their analysis on the Mexican Migration Project which focuses on rural areas in Mexico and might therefore not capture a representative sample. In Ambrosini and Peri (2011), as noted by the authors, the results on returnees self-selectivity are based on a very small sample (an average of 54 observations per cell). Lacuesta (2010) decides to overcome these limitations arguing and providing indicative evidence that self-selectivity should be negligible. [3]

This paper attempts to study the selection of Mexican returning migrants, comparing them to Mexican non-returning migrants in the U.S. Selection is studied in terms of observable and unobservable differences that might cause differences in earnings. In particular, the paper recovers the distribution of wages of Mexican workers that would be observed in the U.S. in the absence of return migration, in other words if all Mexican workers were to settle permanently in the country. As explained earlier, this question is important firstly to reach a comprehensive understanding of migration, and secondarily to understand whether the measures of immigrant performance in the U.S. have been misled by the presence of self-selectivity in return migration.[4]

The paper contributes to the literature that looks at the consequences of return migration from the host country perspective, and it relates to the immigrant assimilation literature.

Unlike Hu (2000) and Lubotsky (2007), the paper analyzes the actual return choice of Mexican migrants combining the U.S. and the Mexican Census data. This allows to distinguish return from panel attrition and treat all those forms of sample selection and heterogeneity that are not eliminated by fixed effects estimators.

The paper goes beyond point estimates and poses attention to the counterfactual distribution. The interest in wage distributions seem wide in the literature (Chiquiar and Hanson, 2005; Kaestner and Malamud, 2010; Fernandez-Huertas Moraga, 2011). Furthermore, in recovering the counterfactual of interest, it extends the estimator in Chiquiar and Hanson (2005)[5] to account also for selection on unobservables. The treatment of the unobservables, despite the use of a single cross-section and the absence of information on pre-return outcomes, allows to improve the work in Ambrosini et al. (2011).

At last, the technique developed in this study could complement the non-parametric analysis in studies who do have access to pre-migration outcomes (Kaestner and Malamud, 2010; Fernandez-Huertas Moraga, 2011; Ambrosini and Peri, 2011; McKenzie and Rapoport, 2010) and offer an alternative to ignoring selection on unobservables whenever such information is not available (Lacuesta, 2010; Ambrosini et al., 2011). In the first case, the short horizon of the panels often makes it impossible to distinguish between selection on unobservable and an Ashenfelter dip. Since the technique presented does not use information on returnees' earnings, it could be used whenever there is a concern of a change in outcome right before migration. In the second case, the paper could provide a check of whether assuming away selection on unobservables is a sensible strategy.

---

[3] Note that in studying return migration in Romania, Ambrosini et al. (2011) renounce to the study of selection on unobservable, as they do not have access to pre-return information on the returnees.

[4] Note that immigrant performance might be over or underestimated depending on how returnees compare with the immigrant stayers, but also the return flow might mitigate the effects that immigration has on the labor market outcomes of natives, in particular for those groups of native workers with the same skill mix of returnees. Understanding who returns could therefore provide a further explanation for the weak evidence of the labor market impact of immigration.

[5] This estimation is in turn based on DiNardo et al. (1996)

The estimation technique adopted uses an identification strategy proposed by Chamberlain (1986) and further advocated by Heckman (1990) (*identification at infinity*). Andrews and Schafgans (1998) use it to identify the constant in sample selection models, Klein et al. (2011) use it to identify the marginal effect in a sample selection model with endogenous treatment and binary outcomes. In terms of applications, Schafgans (1998) has applied the estimator in Andrews and Schafgans (1998), while Mulligan and Rubinstein (2008) use identification at infinity as proposed in Chamberlain (1986) to estimate a sample selection model without exclusion restrictions

3 Data

The analysis uses the Public Use Sample of the the U.S. Census Data and Public Use Sample of the Mexican Census data, both collected in 2000.[6] I define the *Mexican born immigrants* as the individuals born in Mexico appearing in the U.S. Census. I define the *Mexican born return migrants* as temporary migrants in the U.S. appearing in the Mexican Census, with returnees identified as those who report having been residing in the U.S. in the five years preceding the Mexican Census enumeration. As a comparison I also use data on a small random sample of U.S. native-born workers.

The use of different data sources to identify return migrants is not without limitations. As discussed in Chiquiar and Hanson (2005) and Ibarraran and Lubotsky (2007) the most important drawbacks are changes in education once in the U.S., misreporting of education in the U.S. Census, and illegal immigration. Since this study focuses on return migration, the possibility that Mexican immigrants have obtained additional schooling after arriving in the U.S. should not be as invalidating, since returnees could have made the same choice. There is a concern that Mexican migrants in the U.S. might overstate their educational attainment (Ibarraran and Lubotsky, 2007). If so, observed differences in educational attainment might in part be due to misreporting in the U.S. Census and not to the selection of the returnees. However, the pattern shown in the data below is found also in other studies that do not combine the two Censuses (Fernandez-Huertas Moraga, 2011; Ambrosini and Peri, 2011), and this could reduce such concerns. The undercount of illegal immigrants in the U.S. Census might indeed constitute a problem. I will return to this point in Section 6, when exploring the robustness of the results.

There is, however, a final worry specific to this study: the universe of returnees is much broader than the one captured by the Mexican Census. Since no further information is available about having been abroad, looking at the place of residence in 1995 is the best proxy for return status. If the Mexican workers who returned before 1995 systematically differ from those who returned between 1995 and 2000, the conclusions of this paper are not going to have external validity. I will assume throughout the analysis that this is not the case and consider the sample as representative of the full population of returnees.[7]

Table 1 reports the average characteristics for the natives, the U.S. stayers and the returnees that

[6] The U.S. Census is a 5% sample of the population, the Mexican Census is a 10.6% of the population.

[7] The 1990s have been a decade of radical transformation in the Mexican economy, with the signing of the NAFTA in 1994, the Mexican peso crisis and the subsequent period of macroeconomic growth. It is possible that changing macroeconomic conditions affected the return migration flow. However, given that Mexico experienced both a period of financial crisis and a period of growth in the five years of interest, it could be plausible to expect a small average effect of these conditions on return behavior. Finally, a quick parametric analysis of selection in 1990 shows a similar pattern of return migration. These results are available upon request.

are relevant for the analysis. The sample is restricted to men whose age is between 35 and 55 years, born in Mexico and currently working for wages. The total sample size is 67,381 men. Of these, there are 62,071 immigrants who stay in the U.S. and 5,310 are return migrants. Return migrants are therefore 7.8% of the population. There are four indicators for educational attainment (*Less than primary school completed, Primary school completed, Secondary school completed, College Degree*); socioeconomic characteristics are represented by an indicator for being married (*Married*), indicators for having children (*Child*) and indicators for having a U.S.-born spouse (*Spouse U.S. born*) or a U.S.-born child (*Child U.S. born*). Furthermore, experience in the U.S. is represented by indicators for length of stay between 0 and 5 years, 5 to 10 years, 10 to 20 years, 30 to 40 years and more than 40 years (*... Years in the U.S.*). Due to the limited information collected by the Mexican Census about returnees' experience abroad, it is unknown how long these workers had stayed in the U.S. before returning to Mexico. Regional labor market characteristics are represented by indicators of residence location in four regions: *West, Northeast, Midwest, South.* Fourteen industry variables are reported. The table also reports the average wage for the U.S. stayers.[8] The average wage for the returnees is unobserved and therefore not reported.

The decision to stay is modeled as a function of the educational variables and the indicators for being married, having children, having a U.S. born spouse and having a U.S. born child. The wage process is modeled according to various specifications. In the first set of regressions, the observable characteristics include those regressors used in previous analyses on Mexican-U.S. selection (Kaestner and Malamud, 2010; Fernandez-Huertas Moraga, 2011; Ambrosini and Peri, 2011; McKenzie and Rapoport, 2010; Lacuesta, 2010), i.e. education, age and family status. In a second set, adopted throughout the analysis, also having a U.S. born spouse is included, to capture attachment and networks which have been shown to be relevant in this type of work (McKenzie and Rapoport, 2010; Ambrosini and Peri, 2011). The exclusion of such variable might exacerbate the unobserved correlation between the wage and the return processes. In the last specifications all the variables mentioned above are included to reduce the unobserved variability, with the exclusion of having a U.S. born child (a discussion of identification follows in the next section).

Although only a careful modeling of the return decision and of the wage determination process will highlight where in the wage distribution returnees are likely to fall, it is now possible to examine whether return migrants and immigrants differ in terms of observable traits. Observable differences in characteristics valued in the labor market should translate into observable differences in wages and, therefore, in differences between the actual and the counterfactual wage distributions. The characteristics reported below can be subdivided into three different categories: differences in human capital and labor market activities, differences in labor market locations, and differences in socioeconomic characteristics. There is little difference in age, with returnees being slightly younger men. Returnees and stayers, however, greatly differ in terms of their educational attainment: returnees are 17 percentage points more likely than the U.S. stayers to have no education, and 21% less likely to have completed high school. Interestingly, they are however 1 percentage point more likely to have a college degree compared to the stayers. All the reported differences are statistically significant. In terms of socioeconomic characteristics these two groups are equally likely to be married and have children, however stayers are 10% more likely to have a U.S.-born spouse and 46% more likely to have a U.S.-

[8] The wage variable is constructed as wage and salary income divided by hours of work. To avoid division bias (Borjas, 1980), earnings were also used as the dependent variable, without changes in the conclusions of the paper.

born child. It should be noted that the majority of the stayers has been in the U.S. for long periods of time. In fact, 70% of the sample arrived between the 70s and the 90s. This is not surprising given the Bracero program and the Immigration Reform and Control Act of 1986. Most Mexicans in the U.S. work in agriculture, in the electrical industry, in hotels and restaurants, and in the wholesale and retail trade. Returnees have similar occupations in Mexico, although this information is reported here just as a comparison. It is not necessarily true, in fact, that occupations are transferable across borders. Finally, the average wage for the U.S. stayers is about 14 dollars per hour.

As a comparison, the table also reports information on the native born workers, and it shows the well-known differences between natives and immigrant in terms of labor market skills and education. Native born workers have higher levels of education (64% of the sample has an high school degree and 29% has a college degree) and earn about 10 dollars more per hour than the Mexican immigrant on average.

Given the lower labor market experience and human capital of returnees, the descriptive analysis indicates that, probably, had there be no return migration the wage distribution of Mexican immigrants in the U.S. would have more mass in its lower tail. In other words, it seems that returnees are negatively selected and therefore their wage should be below the U.S. stayers' wage. The next section explains how this counterfactual distribution can be estimated and the rest of the paper compares these results with the descriptive analysis just presented.

4 The Model and the Estimation Strategy

The research question requires the recovery of the wage distribution for all Mexican born men who have been in the U.S., even though wages are observed only for Mexican born immigrants who are currently residing in the United States. Let S_i be an indicator of whether or not individual i decides to stay in the U.S. In the following model this decision depends on the net benefits of staying, $(Z_i'\alpha_0 - \epsilon_i)$, being greater than zero.[9]

Let r be the number of returnees and n be the number of stayers. The decision to stay can be represented as:

$$S = \begin{cases} 1 & Z_i'\alpha_0 > \epsilon_i \\ 0 & Z_i'\alpha_0 \leq \epsilon_i \end{cases} \qquad \text{for} \quad i = 1, \ldots, r+n \tag{1}$$

Let the true wage determination process for a randomly selected Mexican immigrant present in the U.S. be:

$$Y_i^* = X_i'\beta_0 + c_0 + u_i^* \qquad i = 1, \ldots, r+n. \tag{2}$$

In the model, Y_i^* is the log of the hourly wage for Mexican immigrants, and X_i are the determinants of the log-wage process.

The wage is observed only for the immigrants who stay in the U.S., however. In other words, the

[9] I am not using a Roy-type model here as return migration can happen even in presence of persistently higher returns in the host country (Dustmann, 2003). Theoretical models that allow for return migration show that this choice can be rationalized by assuming that returnees have a preference for consumption in their own country, or by differences in the prices of skills between the home and the host economy (Dustmann et al., 2011). Furthermore, Stark and Bloom (1985) argue that nonpecuniary motives move individuals.

observed wage is:

$$Y_i = S_i Y_i^* \qquad i = 1, \ldots, r+n. \tag{3}$$

From the model in equation (1) and equation (2) it follows that (Y, S_i, X_i, Z_i) are observed random variables. Below I discuss the assumptions needed in the estimation procedure. These are:

Assumption 1. $(X_i, Z_i, u_i^*, \epsilon_i)$ *i.i.d.*

Assumption 2. $E(u_i^*) = 0$ *and* (u_i^*, ϵ_i) *independent from* (X_i, Z_i).

Figure 1 explains the structure of the model and why these assumptions are needed. Assume that only one exogenous variable X determines both the decision to stay in the U.S. and the wage process. In particular, assume that X is positively related to the log of the wage. Given the model in equation (1) and equation (2), individuals with 'high levels' of X will not only be earning higher wages in the market but also will be more likely to stay in the host country. Therefore, the x-axis represents both X and the probability of staying, while the y-axsis represents the wage process. Let us also assume for the moment that whoever earns below $log(w) = 0.75$ returns. The shaded area shows the fact that the wage is unobserved for some observations, while the individual characteristic X is always observed.

It is well known that focusing only on the selected sample would yield misleading conclusions about the distribution of the outcome in the population. For example, a simple OLS regression on the selected sample would yield biased estimates of the population regression function, represented by the $\widehat{\ln(Wage)}$-line. However, it should also be noted that at 'high levels' of X, i.e., whenever $P(S = 1|X)$ exceeds the threshold $\bar{p}_n$ (the dashed line), selection does not matter. In fact, the error distribution is not truncated and inference could be made about the distribution of the outcome. This intuition lays behind the estimation strategy proposed in the next section.

The possibility of focusing only on the observations for which $P(S = 1|X)$ is above $\bar{p}_n$ relies on the above assumptions. In fact, if X was endogenous (i.e. Assumption 2 not valid), selection on X would further exacerbate the selection problem. If the observations were not i.i.d. and, in particular, if the error was heteroskedastic (Assumption 1), then the distribution of the error for people with 'high' X would differ from the distribution of the error for people with 'low' X.

4.1 Potentials and Limitations of the Estimation Strategy

Assumption 1 and 2 are standard in the area. For example, any non-parametric analysis based on pre-migration earnings (e.g. Ambrosini and Peri (2011); Fernandez-Huertas Moraga (2011)) will recover selection only if the subdivision into cell is exogenous.[10] Similarly any analysis that hinges on estimating the choice of migrating and a wage equation relies on the exogeneity of the explanatory variables in the models (Kaestner and Malamud, 2010; Lacuesta, 2010; Reinhold and Thom, 2009).

Assumption 1 and 2 might be however questionable in empirical applications.

Assumption 2 is needed for consistency. In fact, the assumed exogeneity of the regressors Z in model (1) from u^* guarantees the randomness of this selection rule. Since the regressors used in the analysis are variables such as education and family characteristics, this assumption might be questioned in practice. A check on the validity of this assumption can be done by comparing the

[10] As already mentioned, the variables used here to identify individuals with high probability of staying are similar to those used to identify cell-probabilities in other studies.

estimated unconditional distribution of the error term, $\hat{f}(u^*)$, with the estimated distribution of the error conditioned on the index, $\hat{f}(u^*|Z_i'\hat{\alpha})$. If Z was endogenous these two estimated distributions would differ, and, in particular, the conditional expectation of u^* on $(Z_i'\hat{\alpha})$ would change at different values of $Z_i'\hat{\alpha}$. If Z_i could be treated as exogenous, the two estimated distributions would still differ slightly due to the inherent randomness of the estimation procedure, but would be relatively close. To sum up, if $(Z_i'\hat{\alpha})$ was exogenous, $\hat{f}(u^*)$ should stay 'close' to $\hat{f}(u^*|Z_i'\hat{\alpha})$.

Figure 3(a) shows how 'close' $\hat{f}(u^*)$ and $\hat{f}(u^*|Z_i'\hat{\alpha})$ are. Conditioning on different quantiles of the index $(Z_i'\hat{\alpha})$ does not induce a considerable change in the distribution of u^*. Specifically, it seems that the recovered distribution $f(u^*)$ is relatively conservative, as it shows higher variability than the conditional distributions. To further confirm this finding a Kolmogorov-Smirnov test can be used to test whether these distributions come from the same underlying density. Table 2 tabulates the D statistic at different deciles of the index. The test delivers the expected answer: in most cases the null of $\hat{f}(u^*)$ being close to $\hat{f}(u^*|Z_i'\hat{\alpha})$ cannot be rejected at 5% significance level.[11] This comparison hints that selecting on $(Z_i'\hat{\alpha})$ should not be a concern and the exogeneity of the selection rule seem verified in the data.

Turning to Assumption 1, the variance structure of the model could be extended to allow for an unknown form of heteroskedasticity, at the some cost of tractability.[12] In the current application, a score test for heteroskedasticity failed to reject the null hypothesis that the error is homoskedastic at any significance level. Given that there is no evidence of heteroskedasticity in the data, Assumption 1 is maintained throughout.

These assumptions guarantee that, correcting for observable differences, returnees and stayers have unobservables drawn from the same, unknown, probability distribution. Under such condition, the technique selects a subsample of the data where selection is negligible. Although this might be perceived as a particular group of the sample, such selection is random (based on the above indicative evidence). A selection of a sub-population based on random selection rule to extrapolate information on the population as a whole is common practice.

The intuition now presented allows to recover the distribution of interest despite the presence of sample selection. As mentioned previously, this technique is advantageous whenever the data do not provide enough information on returnees wages such as in Lacuesta (2010) and Ambrosini et al. (2011) or the sample size is too small to provide enough power to the analysis (Ambrosini and Peri, 2011). It should be noted that the estimation uses only information on the wages of the permanent

[11] Two observations needs to be done. First, it does not come as a surprise the fact that in the lowest deciles of the index the difference in the two distributions is larger. This is inherit to the fact that individuals with high probability of staying are identified out of high values of the index. It is reassuring, in particular, that at the median the two distributions are the same also at 10% significance level. Second, the procedure here proposed is based on a comparison of two estimated vectors. To my knowledge, the theory of the Kolmogorov-Smirnov test has not been developed in this context. Therefore, this test is currently only indicative.

[12] Suppose that the model is:

$$Y_i^* = X_i'\beta_0 + c_0 + e_i^*,$$

where there is heteroskedasticity in e^* of unknown form, i.e. $e_i^* = u^* k(X\delta_0)$. The observed model could be written as:

$$Y_i = X_i'\beta_0 + c_0 + G(Z_i'\alpha_0) + u^* k(X\delta_0),$$

where $G(\cdot)$ is the piece due to selection and $k(X\delta_0)$ is the piece due to heteroskedasticity. I will argue below that in a particular set, $G(Z_i'\alpha_0)$ is zero, i.e. sample selection disappears. In that set, then, it would be possible to recover $\hat{k}(\cdot)$ simply by estimating the conditional variance of the model. A simple GLS estimator would then recover the distribution of u^*.

immigrants. In this application this is necessary as no data are collected on returnees' earnings in the U.S. However, in cases where Y_i^* before return is actually observed (Kaestner and Malamud, 2010; Fernandez-Huertas Moraga, 2011; Ambrosini and Peri, 2011; McKenzie and Rapoport, 2010), the technique could still be applied on the sample of permanent migrant to judge whether migration was driven by unexpectedly low earnings in the period before migration (Ashenfelter dip).

4.2 Counterfactual Density Estimation

The aim of the paper is to obtain the distribution of Y_i^*, given that only Y_i is observed. Under a normality assumption of (ϵ_i, u_i), the estimation of the counterfactual of interest would simply require the estimation of the covariance structure of the error terms in the set where selection disappears at the limit, once consistent estimates of the parameters in the model have been obtained.[13] However, if the normality assumption is incorrect, the parametric procedure will yield inconsistent estimates of the parameters of interest and of the counterfactual distribution. For generality, the rest of the paper focuses on estimation techniques that are free of distributional assumptions, while a comparison with the parametric model is reported as a robustness check. The development of a consistent estimator for $f(u^*)$ is the main contribution of this paper, and its asymptotic properties only rely on the consistency of the parameters of the model. Using flexible estimators will be particularly important whenever the parametric assumptions are not satisfied.

Turn now to the estimation strategy. The distribution of Y_i^* in equation (3) corresponds to the distribution of u_i^* up to a location shift represented by the observable characteristics, $(X_i'\beta_0 + c_0)$. Most of the following discussion will therefore focus on recovering the distribution of u_i^*. Let $f(u_i^*)$ be the unknown distribution of u_i^*. By the Law of Total Probability, $f(u_i^*)$ can be written as a weighted sum of the distribution of the error terms in the subsamples of stayers and returnees with weights given by the probability of being in either subsample, i.e.:

$$f(u_i^*) = f(u_i^*|Z_i'\alpha_0) = f(u_i^*|S_i = 1, Z_i'\alpha_0)\Pr(S_i = 1|Z_i'\alpha_0) + f(u_i^*|S_i = 0, Z_i'\alpha_0)\Pr(S_i = 0|Z_i'\alpha_0).$$

The first equality is guaranteed by the independence of the error term from Z (Assumption 2). Second, note that this density cannot be directly estimated using the sample wage density, as the latter is observed only conditional on the decision of staying. In other words, it is not possible to directly obtain an estimate of $f(u_i^*)$ as no information can be directly extrapolated from the data about $f(u_i^*|S_i = 0, Z_i'\alpha_0)$. However, note that:

Result 1. *Whenever* $\Pr(S_i = 1|Z_i'\alpha_0)$ *is close to 1,* $f(u_i^*) \approx f(u_i^*|S_i = 1, Z_i'\alpha_0)$.

Result 1 can then be exploited to identify $f(u_i^*)$. Intuitively, selection disappears in the limit for individuals for which $\Pr(S = 1|Z_i'\alpha_0)$ is close to one, i.e. for observations for whom $\Pr(S_i = 1|Z_i'\alpha_0)$ exceeds a threshold $\bar{p}_n$, function of the sample size. In Figure 1 this threshold is represented by the dashed line. As introduced in the previous section, selection is negligible above this threshold. Let H_i be an indicator for being in this "high-probability" set, i.e. $H_i = 1\Big[\Pr(S_i = 1|Z_i'\alpha_0) > \bar{p}_n\Big]$. Recovering the distribution of u^* simplifies to estimating the distribution of the error term for those observation in the sample $(S_i = 1)$ and for which $H_i = 1$. In other words, the proposed estimator for

[13] A standard two-step Heckman estimation or a joint maximum likelihood estimation would do this.

$f(u_i^*)$ is:

$$\widehat{f(u_i^*)} = \frac{\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{u-u_i^*}{h}\right) S_i H_i}{\sum_{i=1}^{n} S_i H_i}, \tag{4}$$

where $K(\cdot)$ is a kernel density estimator and h is the bandwidth parameter. This estimator is simply a kernel density estimate of the random variable u^* over a fraction of observations for which the probability of being in the selected sample is close to one in the limit. This estimator is the density counterpart of the estimator proposed by Andrews and Schafgans (1998), based on identification at infinity noted by Chamberlain (1986) and further advocated by Heckman (1990). Schafgans (1998) has applied the estimator in Andrews and Schafgans (1998), Mulligan and Rubinstein (2008) use identification at infinity as proposed in Chamberlain (1986) to estimate a sample selection model without exclusion restrictions, Klein et al. (2011) use identification at infinity and an extension of Andrews and Schafgans (1998) to identify the marginal effect in a sample selection model with endogenous treatment and binary outcomes.

To get some sense of how well this method works, I conducted a small Monte Carlo experiment. The data generating process is the following:

$$S_i = \begin{cases} 1 & 1 + X_1 + 2X_2 \geq \epsilon \\ 0 & 1 + X_1 + 2X_2 < \epsilon \end{cases}$$

$$Y_i = 1 + X_1 + u_i \qquad \text{if} \quad S_i = 1$$

Here X_1, X_2, u and ϵ are standard normal random variables. For each iteration in the Monte Carlo experiment, I calculate the deciles of the distribution of u^*, estimated as explained above, and the deciles of the distribution of u^* for those observations for which $S_i = 1$, i.e. for the stayers, and for the observations in the high probability set. These represent the deciles of the two distributions of interest: the 'actual' distribution, $\hat{f}(u^*|S_i = 1)$, and the counterfactual distribution, $\hat{f}(u^*)$. Due to sample selection, the deciles of the actual distribution should be far from the deciles of the normally distributed random variable u^*, while, if the estimator proposed in equation (4) works, the deciles of the distribution in the high probability set should be close to the deciles of a normal distribution. I run this experiment for $N = 5,000$, $N = 10,000$ and $N = 60,000$ with 1,000 replications each. Table 3 reports the bias between each decile of $\widehat{f}(u^*|S_i = 1)$ or $\widehat{f}(u^*)$ and a normally distributed random variable. The first, third and fifth columns of the table shows how using the distribution of the error term in the selected sample does not recover the true distribution in the population: in fact, the estimation of each decile of the distribution is consistently biased. On the contrary, column two, four and six reports the deciles of the distribution estimated using (4). Across all sample sizes, the estimator performs very well and the bias is negligible. This suggests that the estimator in equation (4) is able to recover the true distribution in the presence of self-selection.

4.3 Parameter Estimation

To estimate the density in equation (4), unbiased estimates of the parameters in the model (α_0, β_0, c_0) must be obtained in order to construct the residuals of the model, $\hat{u}^*$. To study the S_i choice, I

estimate a semiparametric dichotomous choice model,[14] applying the estimation method developed by Klein and Spady (1993).

The recovery of $Z_i'\hat{\alpha}$ is useful for two reasons.[15]

First, it is now possible to select those individuals for whom $\Pr(S_i = \widehat{1|Z_i'\hat{\alpha}}) > \bar{p}_n$, i.e. to identify those observations in the high-probability set, for which selection can be ignored at the limit. I define individuals in the high probability set as those observations in the 95th percentile of $\Pr(\widehat{S_i = 1|Z_i'\hat{\alpha}})$. Although this cut point is arbitrary in the paper, results are stable when a different definition of the high probability set is used. Figure 2 shows the estimated counterfactual distribution when the high probability set is defined as individuals in the 95th, 97.5th and 90th percentile of the index in the selection equation. These alternative definitions yield a similar counterfactual density, so that results do not seem sensitive to different definitions.

Second, the estimation of $(Z_i'\hat{\alpha})$ allows us to obtain unbiased estimates of the outcome equation parameters. In the wage equation, I employ Robinson's differencing method (Robinson, 1988) to correct for sample selection and recover unbiased estimates of β_0, and the estimator in Andrews and Schafgans (1998) to recover c_0.

Before proceeding to the results, there is one identification issue that needs to be discussed. At least one variable is needed in the Z_i matrix that does not appear in the X_i matrix. The variable included in the selection process and excluded in the wage process is having a U.S. born child. Having a U.S. born spouse and having a U.S. born child are both proxies for social attachment to the destination country. Attachment to people and institutions in the destination country raises the opportunity cost of returning and should predict well this choice. On the other hand, while having a U.S. born spouse might have an effect on the wage process as it could measure networks and assortative mating, it is unlikely that the wage process would depend on the birthplace location of the individual's children. The effect of having a U.S.-born child should not predict the individual's wage, after controlling for attachment and network effects through the U.S.born spouse indicator and length of stay in the U.S. [16]

[14] On the contrary, DiNardo et al. (1996) choose to adopt a parametric specification of their "selection" probabilities, hence their approach is called 'semi-parametric'. For coherence, I estimate all the parts of the model without any distributional assumptions. In the Result section, however, I present also parametric estimates for comparison.

[15] Effectively, in the estimation of the index, the only identified parameters in terms of the original model are ratio of coefficient, i.e. α_j/α_1, with $j = 1 \dots k$ and where α_1 is the coefficient for the continuous variable, which is normalized to 1. In order to reduce the notational burden, I disregard this technicality in the rest of the discussion.

[16] The research of valid exclusion restrictions poses many challenges to the migration area. I propose three methods to check how sensitive the results are to the choice of the exclusion restriction. First, later on in the paper, I implement a parametric estimation of these counterfactual densities, which is shown to yield similar results to the semiparametric procedure. The parametric procedure has the advantage that identification can be reached through the non-linearities in the functional form of the selection term. Even when no variable is excluded from the model, the results presented in the paper are still obtained. Second, technically identification at infinity does not require the use of an exclusion restriction (Chamberlain, 1986). This strategy has not been developed theoretically to my knowledge, but has been implemented in Mulligan and Rubinstein (2008) to study the impact of changes in self-selectivity on the gender wage gap. Complementary models that do not use exclusion restrictions were run, and provided similar results to the one proposed here.[17] Third, if the excluded variable did enter the wage equation and hence was an invalid exclusion restriction, the estimated density for individuals in the high probability set would change at different quantiles of the estimated index $Z_i\hat{\alpha}$: in fact, an invalid exclusion restriction would cause spurious correlation between the error term in the wage equation and the observable characteristics, and such correlation would still be present in the high probability set. Figure 3(a)-3(b) and the Kolmogorov-Smirnov tests in Table 2 provided evidence that the distribution of the error term is not sensitive to changes in the index quantiles. This could be used as further confirmation of the validity of the exclusion restriction.

5 Results

Besides the interest in the counterfactual estimation, the data allow us to study different components of the return choice and the wage determination process for Mexican born immigrants in the U.S. The next subsection focuses on the study of these choices while the following subsection at last introduces the density estimation.

5.1 Parameter Estimates

Estimates of the marginal effects for the observable characteristics determining the decision to stay in the U.S. are presented in Table 4. The marginal effects are computed at the mean, so the first column of the table reports the average characteristics of the immigrant sample.

Each additional year of age has a small effect on the probability of staying, increasing it by 0.1% for each additional year of age. Compared to individuals without education, Mexicans who have completed primary school are about 0.6% more likely to stay; Mexicans who have completed secondary school are 3.3% more likely to stay than the average migrant, while Mexicans with a college degree are about 0.6% more likely to stay. Having a foreign-born spouse reduces the probability of staying by 3% while individuals with a U.S.-born spouse are about 4% more likely to stay compared to individuals with a foreign spouse. Having a foreign-born child reduces the probability of staying by about 2% while having a U.S.-born child increases the probability of staying by about 12%. It should be noted that the two variables indicating social attachment to the host country are strongly significant. Moreover, it seems that, based on observable characteristics, stayers are more likely to have better educational outcomes, as already highlighted by the descriptive analysis.

Table 5 reports the estimates for the wage equation for the Mexican born workers, while as a comparison Table 6 reports the same estimates for the native born workers. The first two columns report results for a parsimonious specification, where only the human capital variables, the labor market experience variables and the indicators for being married and having children are used. The last two columns better specify the wage equation adding indicators of U.S. length of stay, region and industry. Overall, the wage equation indicates a relatively low return to experience, proxied by *Age*, and high returns to human capital: 6% (3%) to a primary education degree in the parsimonious (full) model; between 15% and 20% increase in the wage for an high school diploma, and between 45% and 50% increase for a college degree. The returns to being married to a foreign person are about 7%, while individuals married to a U.S.-born person have an additional return of 10% compared to those married to a foreign-born spouse. Individuals with children earn about 10% more than individuals without children. Within the cross-section, the longer the individual has been in the U.S. the higher is his wage.

The same results for the natives suggest that the largest differences in wages are due to the difference in schooling.

A few observations should be made. Firstly, the results are reasonably stable across specifications whenever the place of birth of the spouse or the industry and regional indicators are added as a controls. Secondly, having a U.S.-born spouse seems to enter the model even after controlling for how long the individual has been in the U.S. This suggests that excluding this variable might cause a spurious correlation between the error terms not due to selection. Thirdly, it should be remembered that the indicators for length of stay in the U.S., industry and location within the States are unknown for the returnees. Therefore, in predicting the counterfactual wage distribution some imputation

would have to be made about these variables for the returnees. But then the gain in precision from using a better specified wage equation might be lost due to the use of imputation techniques. In estimating the desired density, therefore, I use the parsimonious specification in column (2), where all the characteristics for both stayers and returnees are known and no additional complications are introduced. I present the results from the other specifications as a robustness check.

5.2 Density Estimates

Three questions will now be answered: first, how different is the full immigrant population in terms of observable and unobservable traits compared to the population that stays in the U.S.; second, what would the distribution of wages be in the absence of return migration; third, how does this distribution change, conditional on educational characteristics.

How different is the immigrant population compared to the population of stayers in the U.S., in terms of observable and unobservable traits? To compare the two different groups of interest, I report in Table 7 the deciles of the predicted wage, the residuals, and the wage distributions that is observed and that would have been observed had there been no return migration. These quantities were calculated in the following manner. The first panel shows the predicted actual and counterfactual wage. They are both calculated as the product of the returns to skills reported in Table 5 and the immigrant stayers' (immigrant population) characteristics for the actual (counterfactual) predicted wage distribution, i.e. $\hat{c} + \hat{\beta} X_j$, where $j =$ only stayers, immigrant full population. Deciles of the predicted wage are reported as a summary measure.

In terms of observable characteristics, Mexican immigrants would on average be earning less, had there been no return migration. In fact, the log-difference across the different quantiles is always negative. This is in line with the descriptive analysis that found returnees as having below average skills. Likewise, it is consistent with the analysis of the decision to stay, which highlighted that stayers are more likely to have higher levels of experience and human capital. However, these differences are relatively small, reaching at most a few cents decrease (approximately 0.9% decrease) in the wage between the two scenarios.

The role of the unobservable traits is shown in the second panel of this table. The unobservables were calculated as the difference between the actual and the predicted wage for the stayers, and are directly estimated for the full population using the estimation technique described in Section 4. Positive differences between the counterfactual and the actual distribution are driven by dissimilarities in the unobservable traits. Had there be no return migration, the immigrant population would have been earning approximately 7.3% more (about 1 dollar, at the median) due to unobservable differences between stayers and returnees. The effect at the average is relatively small (2% change) and consistent with the relatively small effect of selection at the mean reported in other studies (Lindstrom and Massey, 1994; Ambrosini and Peri, 2011)

The evidence presented suggests that the immigrant stayers and the full population composed of stayers and returnees are somehow close in terms of observable traits while some differences arise in terms of unobservable traits. In particular, although in terms of observable traits returnees are a disadvantaged group in the labor market, their unobservable abilities seem to compensate for this lack of skills. It seems that returnees might have unobservable motives that push them to be more successful in the host country than the immigrants who stay. Although we cannot directly tackle the understanding of the motives behind the return, it is possible to conjecture, then, that these

immigrants might leave the host country upon reaching their savings or skill acquisition goals, and the more motivated immigrants are able to do so, despite their original disadvantage in the host country labor market.[18]

What would the wage distribution be in the absence of return migration? The overall impact of return migration is represented in the last panel of Table 7. This panel reports the deciles of the actual wage distribution for the stayers and of the counterfactual wage distribution that would have occurred in the absence of return migration. In practice this second distribution sums the observable (panel one) and unobservable components (panel two) for the immigrant population at each deciles. Almost at all deciles, the implied counterfactual distribution suggest that Mexican immigrants would be earning more had there been no return migration. The largest impact is at the median, with an increase of about one dollar (7% higher wage).

To better visualize the actual and the counterfactual distributions just described, Figure 4 represents them graphically. Although relatively close to each other, some differences in the two distributions appear from this figure. In the absence of return migration Mexican immigrants would be more in the upper tail of the distribution and the average wage in the population would increase. To better observe this point, Figure 5 represents the difference between the counterfactual and the actual distribution. Without return migration there would be more mass in the upper tail of the wage distribution, as the wage difference is first negative and then positive. Therefore, the disadvantage in terms of human capital skills that returnees face is balanced by the higher unobserved motivation and productivity that this group exhibits. This translates into an increase in the concentration of individuals at the middle-upper part of the wage distribution in the absence of return migration. A Kolmogorov-Smirnov test for the difference in these two distributions delivers a D statistic of 3.05, so that based on the standard critical values of such test, the actual and the counterfactual are different at any conventional significance level.

This effect is not the only insight of the analysis, however. The last panel of Table 7 shows how return migration affects also wage inequality, reporting the 90-10, 90-50 and 50-10 wage gaps for the actual and the counterfactual distributions. At the bottom of the distribution, in the absence of return migration there would be an increase in the difference between the 50th and the 10th percentile (roughly a 8%) increase. On the contrary, at the top of the distribution, there would be a reduction of dispersion. Overall, in the absence of return migration inequality within the Mexican population would increase slightly (1%). Therefore, because selective return migration induces the high wage earners to leave, it implies a reduction in inequality within the Mexican population in the U.S. If the returnees were to stay, the full wage distribution in the population would exhibit slightly higher dispersion compared to what it is currently observed.

How does the wage distribution change conditional on educational characteristics? Since the educational characteristics greatly affect both the decision to stay and the wage, the importance of selection might vary by educational levels. Table 8 reports the deciles of the predicted wage, of the unobservables and of the actual wage distributions for people with a primary school degree, with an high school degree and with a college degree. As before, the differences in observables are negligible across all educational groups, while unobservables drive the dissimilarities in the wage process.

[18] As mentioned, Yang (2006) provides a direct test for the reason to return for Philippine migrants.

However, while on average returnees with primary and secondary education have higher unobservable traits, the distribution of unobservables is quite different for workers with a college degree. To better visualize these differences, Figure 6 shows the actual and counterfactual distributions and their dissimilarity at different educational levels. Figure 6(a) and 6(b) show the distribution of log-wages for individuals who have completed primary school: returnees are again disproportionately drawn from the upper tail of the density; the same conclusion can be inferred by from Figure 6(c) and 6(d) which shows the same distribution for workers with a secondary degree. At last Figure 6(e) and 6(f) show what would have happened if all returnees with a college degree had stayed. In this case there would be a much larger mass of individuals in the center of the distribution.

It is possible to conclude this analysis with two remarks. First, returnees are not low-wage earners, across all educational groups. Although the descriptives highlighted huge educational differences between stayers and returnees, within each educational group returnees are the high-wage earners. Second, most of the action happens at the tails of the distribution: while almost no difference can be detected for individuals with a high school degree, selective return migration has a much larger impact on individuals with low or high education.

6 Robustness Checks

Various Specifications. Section 5 discussed the main results of the paper, based on the estimation of a parsimonious wage equation. This section briefly presents the results under different specifications of the model. In particular, results are presented on a fully specified model (column 4 of Table 5), I use parametric techniques and I report an analysis for individuals between 25 and 45 years old. Table 11 shows the deciles of the overall distribution for the three different robustness checks, and Figure 7 provides graphical evidence for the distributions conditional on educational attainment.

The previous discussion constructed the counterfactual and the actual distribution based on the estimation of the parsimonious wage equation reported in column 2 of Table 5. There could be some concern that a better specified model could change the results. As explained previously, the main problem of using a fully specified wage equation is that no information is present for the returnees on length of stay, the location and the industry in the U.S. Not to introduce extra uncertainty due to the imputation of these missing variables, I assumed that the returnees would present characteristics like the average non-returning migrant. Given the previous result of similarities of the quantiles of $X\hat{\beta}$, this assumption seems reasonable. All the conclusions explained above carry on for this further specification. The first three columns of Table 11 are consistent with the results shown before. Figures 7(a), 7(d), and 7(g) show the difference in the counterfactual and actual log-wage distributions at different levels of education, when a full specification of the wage equation is adopted.It is apparent from these figures that, as expected, a better specified model reduces the variance of the wage distribution. However, all the conclusions are equally valid in this setup.

A second possibility of extending the results is to focus on younger individuals who are well represented in the Mexican population. In principle, the selection process in different age groups might vary. For example, it could happen that younger Mexican who come back are the least successful in the host country while the older Mexican workers are those who have stayed long enough to acquire experience, skills and savings to bring back to Mexico. The last three column of Table 11 shows the overall actual and counterfactual distribution for individuals between 25 and 45 years old. The results are close to the one found before, although the overall impact of selection is relatively smaller for this

group. Figures 7(b), 7(e), and 7(h) do show the same pattern of selection found previously along the educational distribution, however. At different levels of education, even at younger age, the returnees are positively selected and in the absence of return migration there would be more workers in the middle-upper part of the wage distribution. However, it is true that selection is smaller for this group of workers. Still, in the absence of return migration there would be more middle-to-high wage earners in the U.S. than what we are currently observing.

Throughout the analysis a fully semi-parametric specification has been adopted to avoid inconsistency of the parameters if the normality assumption was violated in the data. However, the same technique presented for the recovery of the population distribution of the error term u^* could also be applied in a parametric setup. The reliability of these results will depend on the distributional assumptions of the model. Table 10 presents the estimates for the decision to stay and the wage equation when both models have been estimated parametrically. A probit model was implemented to estimate the binary model. The first column of the table reports the implied marginal effects. The parametric model consistently overestimates the effects of the characteristics on the decision to stay; the parametric marginal effects are two to three times larger than the semiparametric marginal effects. This indicates the importance of avoiding the normality assumption. The second column of Table 10 shows the results for the wage equation. Here the coefficients are generally close to the semiparametric results. Following the same logic used for the semi-parametric estimator, I then construct $\hat{u}^*$ as the vector of residuals for individuals in the top 95th-percentile of the probability of staying, now defined by the cumulative normal distribution evaluated at the index in the S_i decision. I compare the distribution of wages implied by this sample, where selection has been removed, to the distribution of wages in the selected sample. The central columns of Table 11 present the results for this specification. Figures 7(c), 7(f), and 7(i) shows the difference in the counterfactual and actual distributions of the residuals at different educational levels using the parametric procedure. The parametric results are very close to the semiparametric results. This is not completely surprising as the log-wage transformation has long been used to produce normality. This result is also reassuring, as it shows that the technique presented could be easily implemented in a parametric setup.

As an overall look at Table 11 and Figure **??** shows, across all the different specifications and techniques, the results are not overturned and remain relatively stable. Mexican returnees come from the middle-top part of the distribution so that in the absence of return migration there would be a larger mass of people with wages laying in the upper part of the wage distribution. This conclusion holds across all the educational levels, with a larger impact of selective return migration for individuals with either primary or college education.

Illegal Immigration. The problem of the undercount of illegal immigrants is often a concern when using Census data. The fear of deportation might induce the illegal immigrants not to fill the Census form. As a consequence, the Census sample might not represent the Mexican population in the U.S. For example, the result of positive selection in Chiquiar and Hanson (2005) seemed largely driven by the non-representativennes of the Mexican sample taken from the U.S. Census (Fernandez-Huertas Moraga, 2011).

Let us distinguish two types of concerns related to this issue.

The first concern is that the observable traits of the U.S. sample used in the analysis do not capture the characteristics of the Mexican population in the U.S. due to illegal immigration. Passel (2006) shows that illegal immigrants are young, low-educated, low wage workers. If this is the case,

the actual distribution of wages and the counterfactual should not change their position relative to each other as the introduction of the missing individuals would simply constitute a shift in means in both distributions.

The second - serious - worry comes from the non-randomness of the Census sample even after controlling for other characteristics. In the following discussion, I will argue that, under some conditions, the identification strategy that has been used in the analysis is robust to the non-randomness of the sample in the U.S.

To visualize the effects of an undercount of the Mexican immigrants in the U.S. Census, let $C_i = 1$ be an indicator that equals one if the respondent appears in the Census and zero otherwise:

$$C_i = \begin{cases} 1 & W_i'\gamma \geq \eta_i \\ 0 & W_i'\gamma < \eta_i. \end{cases}$$

Then the choice of staying in the U.S. is observed only if the individual does appear in the Census:

$$S_i = \mathbf{1}(Z_i\alpha \geq \epsilon_i) \qquad \text{if} \qquad C_i = 1.$$

The concern is that η_i and ϵ_i are correlated and, in particular, based on results Passel (2006), we expect them to be positively correlated: individuals who are more likely to appear in the sample are also those more likely to stay. If η_i and ϵ_i are correlated, then there might be a concern that the probability of staying $P(S_i = 1|Z_i'\alpha)$ has been misestimated. Using again the Law of Total Probability, in fact:

$$P(S_i = 1|Z_i'\alpha) = P(S_i = 1|Z_i'\alpha, C_i = 1)Pr(C_i = 1) + P(S_i = 1|Z_i'\alpha, C_i = 0)Pr(C_i = 0),$$

where the second part of the addition is missing. However, note that the high probability set was constructed by sending $P(S_i = 1|Z_i'\alpha, C_i = 1)$ to one. In doing so, implicitly individuals with large values of $Z_i'\alpha$ were selected. However, whenever $Z_i'\alpha$ is high, also $W_i'\gamma$ will be high. In fact, the main variable that can send that probability to one is age. Individuals who are older, are not only more likely to stay but they have been found to be more likely to be captured by the Census (Passel, 2006). Then, whenever $P(S_i = 1|Z_i'\alpha, C_i = 1)$ is high also $Pr(C_i = 1)$ will be high. This implies that in the high probability set the probability of staying is mostly determined by individuals who do appear in the sample, i.e.: $P(S_i = 1|Z_i'\alpha) \approx P(S_i = 1|Z_i'\alpha, C_i = 1)Pr(C_i = 1)$. As a consequence of using individuals in the high probability set, the distribution of the unobservables recovered should not be affected by illegal immigration. In other words, given the relation between S_i and C_i in this particular application, using the high probability set seem to marginalize the problems related to the censoring in the selection rule due to illegal immigration.

7 Discussion and Policy Implications

A few implications can be drawn from the previous results.

In the absence of return migration, there would be more Mexican immigrants in the upper tail of the wage distribution. The main conclusion from the presented results is that

the immigrants who decide to leave are the high wage earners. Without return migration, then, the average wage in the population would be higher. This is true not only overall, but also looking at education groups within the immigrant population.

Returnees are less skilled than the stayers, but have higher unobservable traits that make them more successful in the labor market. This implies that an analysis that simply controls for differences in observable characteristics might come to the misleading conclusions that returnees are those who fail in the host country. On the contrary, returnees are not immigrants who failed, but instead they are probably immigrants who have reached their goals in the host country, either in terms of savings or in terms of skill acquisition. This result contrasts with the findings by Lubotsky (2007) and Hu (2000). Although these studies show how the use of repeated cross-sections and of panel data can yield very different answers about immigrant wage progress, it is not clear whether the techniques they use are actually capturing the effect of self-selectivity in return migration. In particular, these analyses did not study directly the effect of selection. Their implicit assumption is that heterogeneity operates only through an uncorrelated individual fixed effect that is completely determined by the individual's observable characteristics. The above cross-section estimate, however, shows the importance of non-random selection in determining the decision to stay. Of course it would be of great interest in the future to shed further light on this puzzle by comparing results from cross-sectional data and panel data. As reviewed in Vella (1998), only certain forms of sample selection bias can be eliminated using fixed effects estimator, which is the technique adopted in the migration literature. It would be therefore important to compare cross-sectional results with panel data results, when both fixed effects and sample selection estimators for panel data are used. This results is however in line with the evidence presented in Ambrosini and Peri (2011). Despite the small sample size, the authors show that returnees are positively selected respect to immigrant stayers.

Return migration impacts immigrant inequality. Return migration decreases inequality at the bottom of the distribution and increases inequality at the top of the distribution. As a consequence, the 90-10 wage differential changes only slightly. These effects are similar even if only individuals with primary and secondary education are considered. The conclusion about the high skilled are different, however: return migration undoubtedly increases wage inequality within this group. Therefore, in terms of policy consequences, if policy makers are concerned with the low-earners, selective return migration seems to alleviate the dissimilarities in this population. However, if the goal of immigration reform were to increase the average skill level of the incoming alien population, it should be recognized that the top-earners of this group would be returning to their home country.

The immigrant-native born wage gap would slightly close in the absence of return migration. An implication of this paper can be drawn by comparing the counterfactual distribution of wages with the wage distribution of the native-born workers. All the figures presented above show also the native-born workers wage distribution. In addition, Table 7 and Table 9 report the deciles of this distribution, overall and at different educational levels. From Figure 4 it can be observed that in the absence of return migration the immigrant wage distribution would become closer to the native-born wage distribution. The most interesting comparison can be observed in Figure 6, where the wage distribution is represented at different educational levels. Across all levels of human capital there is a consistent earning gap between Mexican born and native-born workers. This gap would slightly close, however, at both very low levels of education and at very high levels of education if all immigrants

were to stay. The difference between the actual, the counterfactual and the native-born wage distributions is striking for individuals with a primary school degree or for individuals with a college degree. Two observations can be made. First, it is apparent from Figure 6 that selection in return migration is inducing the middle-top earners to leave the U.S. and therefore is biasing the picture we have in mind of Mexican performance, at both low and high levels of education. For example, in the absence of return migration, more of the top-earners among the low skilled would stay in the U.S. A similar conclusion holds also for the high skilled workers. As a consequences, a randomly selected Mexican immigrant would actually be doing better than what we observe. As an example, consider a migration policy that guarantees entry to the U.S. to individuals with high levels of education. This policy might still not fully benefit the U.S. as the middle-top wage earners - the most productive workers - would leave.[19]

8 Conclusions

The political discussion generated by Mexican migration flows to the U.S. has focused on understanding who decides to migrate and, until recently, the role of selective return migration in shaping our estimates of immigrant labor market outcomes has been ignored. Relatively little literature has examined how returnees compare to the stayers in the host country. This paper adds to the literature by analyzing this question through recovering a counterfactual wage distribution in the absence of return migration.

The estimation procedure extends the estimator in Andrews and Schafgans (1998) to its density counterpart and shows the overall distribution of wages that would be observed if all migrants were permanent, and such distribution conditional on educational attainment.

Results suggest that selective return migration improves the average performance of immigrants and causes a decrease in immigrant wage dispersion. Return migration has a greater impact in the tail of the wage distribution. In particular, in the absence of return migration the would be more mass in the upper part of the wage distribution at both very low educational and very high educational groups, implying up to a 7% increase of the median wage in the Mexican population. These results are stable across different wage specifications, different samples and using also parametric techniques. The impact at the mean is however relatively small and might be the reason of the findings sometime conflicting of the literature. The idea that we have on Mexican migration has been distorted by selective return migration. Furthermore, the presented results contrast with the general perception that those who return have 'failed' in the host country, and it contrast with the previous literature on the nature of return migration in the U.S.

[19] I am implicitly assuming that this policy would not change the selection process of immigrants with high levels of education from Mexico to the U.S.

A Tables

Table 1: Demographic and socio-economic characteristics, Native Born and Foreign Born Men, 35-55 Years Old

Variable	**Natives**	**All Mexican Born**	**Stayers**	**Returnees**
Age	44.26	42.43	42.48	41.80***
	(5.85)	(5.67)	(5.66)	(5.70)
Less than Primary School	0.00	0.22	0.21	0.38***
	(0.06)	(0.41)	(0.41)	(0.49)
Primary Education	0.07	0.44	0.44	0.46***
	(0.25)	(0.50)	(0.50)	(0.50)
Secondary Education	0.64	0.29	0.30	0.09***
	(0.48)	(0.45)	(0.46)	(0.29)
College Education	0.29	0.05	0.05	0.06***
	(0.45)	(0.22)	(0.21)	(0.24)
Married	0.88	0.89	0.89	0.89
	(0.32)	(0.31)	(0.31)	(0.31)
Spouse US born	0.72	0.11	0.11	0.01***
	(0.45)	(0.31)	(0.32)	(0.10)
Child	0.57	0.72	0.72	0.72
	(0.50)	(0.45)	(0.45)	(0.45)
Child US born	0.57	0.60	0.64	0.18***
	(0.50)	(0.49)	(0.48)	(0.38)
0-5 Years in U.S.	-	-	0.10	-
	-	-	(0.30)	-
5-10 Years in U.S.	-	-	0.09	-
	-	-	(0.28)	-
10-20 Years in U.S.	-	-	0.36	-
	-	-	(0.48)	-
20-30 Years in U.S.	-	-	0.35	-
	-	-	(0.48)	-
30-40 Years in U.S.	-	-	0.09	-
	-	-	(0.28)	-
>40 Years in U.S.	-	-	0.02	-
	-	-	(0.14)	-
Northeast Region	0.19	-	0.02	-
	(0.39)	-	(0.14)	-
Midwest	0.26	-	0.10	-
	(0.44)	-	(0.30)	-
South	0.35	-	0.28	-
	(0.48)	-	(0.45)	-

Continue to next page

Continued from previous page

Variable	Natives	All Mexican Born	Stayers	Returnees
West	0.20	-	0.60	-
	(0.40)	-	(0.49)	-
Agriculture, fishing, and forestry	0.02	0.15	0.14	0.25***
	(0.15)	(0.36)	(0.35)	(0.43)
Mining	0.01	0.01	0.01	0.01
	(0.10)	(0.08)	(0.08)	(0.07)
Manufacturing	0.23	0.23	0.25	0.09***
	(0.42)	(0.42)	(0.43)	(0.29)
Electricity, gas and water	0.02	0.01	0.01	0.00***
	(0.15)	(0.07)	(0.07)	(0.04)
Construction	0.11	0.18	0.18	0.12***
	(0.32)	(0.38)	(0.39)	(0.32)
Wholesale and retail trade	0.16	0.17	0.18	0.09***
	(0.36)	(0.38)	(0.38)	(0.28)
Hotels and restaurants	0.01	0.02	0.02	0.03***
	(0.08)	(0.13)	(0.13)	(0.17)
Transportation and Communications	0.08	0.04	0.04	0.04
	(0.27)	(0.20)	(0.20)	(0.20)
Financial services	0.03	0.01	0.01	0.00
	(0.18)	(0.07)	(0.08)	(0.06)
Public administration and defense	0.08	0.01	0.01	0.01
	(0.26)	(0.11)	(0.11)	(0.11)
Real estate and business services	0.08	0.07	0.07	0.02***
	(0.27)	(0.25)	(0.25)	(0.15)
Education	0.06	0.02	0.02	0.01***
	(0.23)	(0.13)	(0.14)	(0.10)
Heath and social work	0.05	0.01	0.02	0***
	(0.21)	(0.12)	(0.12)	(0.08)
Other services	0.07	0.05	0.05	0.05**
	(0.26)	(0.22)	(0.23)	(0.21)
Private household services	0.000	0.002	0.001	0.01
	(0.02)	(0.04)	(0.04)	(0.08)
Wage	23.46	14.39	14.39	-
	(18.60)	(12.36)	(12.36)	-
N	71,484	67,381	62,071	5,310

Standard deviations in parentheses
Significance levels: *: 10%, **: 5%, ***: 1% for a t-test for differences in means between Returnees and U.S. Stayers.

Table 2: Kolmogorov-Smirnov Test for Exogeneity of $(Z_i'\alpha)$.

	D Statistic
Decile 1	2.766
Decile 2	2.052
Decile 3	1.576***
Decile 4	1.313**
Decile 5	1.077*
Decile 6	0.812*
Decile 7	0.377*
Decile 8	0.416*
Decile 9	0.373*

Significance levels: *: 10%, **: 5%, ***: 1%.
Critical Values: 10%: 1.22; 5%: 1.36; 1%: 1.63;
The test was constructed comparing u^* with the u^* for individuals in each decile of $(Z_i'\alpha)$. Under the null hypothesis, these two vectors are assumed to be drawn from the same distribution.

Table 3: Comparison of the Deciles of $\hat{f}(u_i^*)$ and $\hat{f}(u_i^*|S_i = 1)$ with the Deciles of a Normal Random Variable.

	N = 5,000		N = 10,000		N = 60,000	
Decile	$f(u^*\|S=1)$	$f(u^*)$	$f(u^*\|S=1)$	$f(u^*)$	$f(u^*\|S=1)$	$f(u^*)$
1.0	-0.469	-0.018	-0.469	-0.012	-0.467	-0.009
2.0	-0.494	-0.017	-0.494	-0.014	-0.493	-0.011
3.0	-0.514	-0.015	-0.513	-0.012	-0.513	-0.012
4.0	-0.530	-0.019	-0.530	-0.017	-0.529	-0.014
5.0	-0.545	-0.025	-0.545	-0.020	-0.545	-0.017
6.0	-0.562	-0.028	-0.562	-0.026	-0.561	-0.022
7.0	-0.579	-0.036	-0.579	-0.033	-0.579	-0.029
8.0	-0.601	-0.049	-0.601	-0.044	-0.600	-0.040
9.0	-0.631	-0.080	-0.632	-0.071	-0.629	-0.069

Table 4: Marginal effects of variables on the Probability of Staying in the U.S., Mexican Born Men, 35-55 Years old

	Average Characteristics	Marginal Effects
Baseline	0.921	0.922
Age	42.43	0.001***
		(2.07E-04)
Primary	0.44	0.006***
		(0.001)
Secondary	0.29	0.033***
		(0.004)
College	0.05	0.006***
		(0.001)
Married	0.89	-0.030***
		(0.003)
US born spouse	0.11	0.039***
		(0.003)
Child	0.72	-0.019***
		(0.003)
US born child	0.60	0.118***
		(0.005)

Standard errors in parentheses.
Significance levels: *: 10%, **: 5%, ***: 1%.
The marginal effects are calculated at the average X.

Table 5: Wage Equation Estimates, Mexican Born Men working for wages, 35-55 Years old.

	(1)	(2)	(3)	(4)
Constant	1.672***	1.538***	1.764***	1.675***
	(0.534)	(0.530)	(0.510)	(0.508)
Age	0.027***	0.028***	0.011	0.012
	(0.007)	(0.007)	(0.007)	(0.007)
Age Sq	-3.E-04***	-3.E-04***	-1.E-04*	-2.E-04*
	(8.E-05)	(8.E-05)	(8.E-05)	(8.E-05)
Primary Education	0.058***	0.057***	0.034***	0.033***
	(0.006)	(0.006)	(0.006)	(0.006)
Secondary Education	0.177***	0.216***	0.134***	0.161***
	(0.008)	(0.009)	(0.008)	(0.009)
College Education	0.514***	0.506***	0.454***	0.450***
	(0.011)	(0.011)	(0.011)	(0.011)
Married	0.084***	0.071***	0.076***	0.068***
	(0.009)	(0.009)	(0.009)	(0.009)
Child	0.097***	0.102***	0.084***	0.087***
	(0.008)	(0.008)	(0.008)	(0.008)
Spouse U.S. born	-	0.099***	-	0.067***
	-	(0.011)	-	(0.010)
5-10 Years in U.S.	-	-	0.012	0.011
	-	-	(0.010)	(0.010)
10-20 Years in U.S.	-	-	0.101***	0.100***
	-	-	(0.008)	(0.008)
20-30 Years in U.S.	-	-	0.198***	0.195***
	-	-	(0.008)	(0.008)
30-40 Years in U.S.	-	-	0.281***	0.276***
	-	-	(0.011)	(0.011)
>40 Years in U.S.	-	-	0.351***	0.340***
	-	-	(0.017)	(0.017)
Industry indicators	No	No	Yes	Yes
Regional indicators	No	No	Yes	Yes
N	62,071	62,071	62,071	62,071

Standard errors in parentheses
Significance levels: *: 10%, **: 5%, ***: 1%.
The industry and regional indicators used in column (3) and (4) are the variables presented in the descriptive statistics.

Table 6: Wage Equation Estimates, Native Born Men working for wages, 35-55 Years old.

	(1)	(2)	(3)	(4)
Constant	1.914***	1.811***	1.761***	1.657***
	(0.030)	(0.030)	(0.030)	(0.030)
Age	0.010***	0.015***	0.008***	0.013***
	(0.001)	(0.001)	(0.001)	(0.001)
Age Sq	-2.E-05	-9.E-05***	-1.E-06	-7.E-05***
	(1.E-05)	(1.E-05)	(1.E-05)	(1.E-05)
Primary Education	0.066***	0.054***	0.061***	0.050***
	(0.008)	(0.008)	(0.008)	(0.008)
Secondary Education	0.339***	0.324***	0.307***	0.293***
	(0.008)	(0.008)	(0.008)	(0.008)
College Education	0.794***	0.777***	0.762***	0.745***
	(0.008)	(0.008)	(0.008)	(0.008)
Married	0.156***	0.101***	0.153***	0.099***
	(0.002)	(0.002)	(0.002)	(0.002)
Child	0.106***	0.072***	0.100***	0.067***
	(0.001)	(0.001)	(0.001)	(0.001)
Spouse US	-	0.110***	-	0.108***
	-	(0.001)	-	(0.001)
Industry indicators	No	No	Yes	Yes
Regional indicators	No	No	Yes	Yes
N	71,484	71,484	71,484	71,484

Standard errors in parentheses
Significance levels: *: 10%, **: 5%, ***: 1%.
The industry and regional indicators used in column (3) and (4) are the variables presented in the descriptive statistics.

Table 7: Deciles of $\hat{Y}_i$ and $\hat{u}_i$ and Y_i, Parsimonious Model, Native Born and Foreign Born Men 35-55 Years Old.

Decile	**Actual**	**Counterfactual**	**Log Difference**	**Natives**
		Observables		
1	2.332	2.332	0.000	2.638
2	2.391	2.389	-0.002	2.753
3	2.426	2.421	-0.005	2.842
4	2.437	2.436	-0.001	2.873
5	2.468	2.462	-0.006	2.899
6	2.491	2.488	-0.003	2.924
7	2.541	2.532	-0.009	2.981
8	2.608	2.601	-0.007	3.302
9	2.652	2.652	0.000	3.363
Average	2.494	2.490	-0.004	2.954
		Unobservables		
1	-0.672	-0.693	-0.020	-0.688
2	-0.486	-0.475	0.012	-0.434
3	-0.333	-0.289	0.044	-0.265
4	-0.197	-0.138	0.058	-0.126
5	-0.070	0.003	0.073	0.003
6	0.065	0.129	0.064	0.127
7	0.211	0.271	0.060	0.258
8	0.393	0.428	0.036	0.410
9	0.660	0.656	-0.004	0.644
Average	-0.021	0.000	0.021	0.000
		Log-Wage		
1	1.660	1.639	-0.020	1.951
2	1.905	1.915	0.010	2.319
3	2.093	2.132	0.039	2.577
4	2.240	2.298	0.057	2.748
5	2.398	2.465	0.067	2.902
6	2.556	2.617	0.061	3.051
7	2.752	2.803	0.051	3.239
8	3.001	3.029	0.028	3.713
9	3.312	3.308	-0.004	4.006
Average	2.473	2.490	0.017	2.954
		Inequality Measures		
10-90 Wage	1.652	1.669	0.017	2.056
10-50 Wage	0.738	0.826	0.087	0.826
50-90 Wage	0.914	0.843	-0.071	0.843
N	62,071	62,071		71,484

The first column (*Actual*) shows $\hat{Y}$ and $\hat{u}$ for the observed sample. The second column (*Counterfactual*) shows $\hat{Y}$ and $\hat{u}$ if all returnees had stayed. Therefore, the observable characteristics of the sample correspond to the observables for both stayers and returnees. The unobservables correspond to the predicted u^*.

Table 8: Deciles of $\hat{Y}_i$ and $\hat{u}_i$ and Y_i, Parsimonious Model, Mexican Born Men 35-55 Years Old.

Decile	**Act.**	**Counterfact.**	**Log Diff**	**Act.**	**Counterfact.**	**Log Diff**	**Act.**	**Counterfact.**	**Log Diff**
	Primary Education			Secondary Education			College Education		
				Observables					
1	2.324	2.324	0.000	2.474	2.474	0.000	2.763	2.763	0.000
2	2.375	2.375	0.000	2.519	2.519	0.000	2.814	2.814	0.000
3	2.417	2.417	0.000	2.576	2.576	0.000	2.866	2.866	0.000
4	2.434	2.434	0.000	2.585	2.585	0.000	2.883	2.883	0.000
5	2.449	2.449	0.000	2.608	2.608	0.000	2.904	2.898	-0.007
6	2.462	2.462	0.000	2.621	2.621	0.000	2.921	2.916	-0.005
7	2.477	2.477	0.000	2.640	2.640	0.000	2.935	2.933	-0.002
8	2.490	2.488	-0.002	2.651	2.651	0.000	2.942	2.942	0.000
9	2.494	2.494	0.000	2.692	2.692	0.000	3.010	3.004	-0.006
Average	2.434	2.433	-0.001	2.594	2.593	-0.001	2.890	2.887	-0.003
				Unobservables					
1	-0.655	-0.693	-0.038	-0.698	-0.693	0.006	-0.890	-0.693	0.197
2	-0.479	-0.475	0.005	-0.486	-0.475	0.011	-0.618	-0.475	0.143
3	-0.335	-0.289	0.046	-0.308	-0.289	0.019	-0.410	-0.289	0.122
4	-0.204	-0.138	0.066	-0.166	-0.138	0.028	-0.206	-0.138	0.068
5	-0.083	0.003	0.085	-0.029	0.003	0.032	-0.036	0.003	0.039
6	0.048	0.129	0.081	0.102	0.129	0.027	0.140	0.129	-0.011
7	0.189	0.271	0.082	0.254	0.271	0.017	0.308	0.271	-0.037
8	0.367	0.428	0.062	0.425	0.428	0.003	0.510	0.428	-0.081
9	0.642	0.656	0.014	0.662	0.656	-0.005	0.798	0.656	-0.142
Average	-0.025	0.000	0.025	-0.011	0.000	0.011	-0.027	0.000	0.027
				Log-Wage					
1	1.669	1.631	-0.038	1.775	1.781	0.006	1.874	2.071	0.197
2	1.896	1.900	0.005	2.033	2.044	0.011	2.196	2.339	0.143
3	2.082	2.128	0.046	2.268	2.287	0.019	2.455	2.577	0.122
4	2.230	2.296	0.066	2.419	2.447	0.028	2.677	2.745	0.068
5	2.367	2.452	0.085	2.579	2.611	0.032	2.868	2.901	0.033
6	2.510	2.591	0.081	2.723	2.750	0.027	3.061	3.045	-0.016
7	2.666	2.748	0.082	2.894	2.911	0.017	3.243	3.204	-0.039
8	2.857	2.916	0.059	3.076	3.079	0.003	3.452	3.370	-0.081
9	3.136	3.150	0.014	3.354	3.349	-0.005	3.808	3.660	-0.148
Average	2.410	2.433	0.023	2.583	2.593	0.010	2.863	2.887	0.024
				Inequality Measures					
10-90 Wage	1.468	1.519	0.051	1.579	1.568	-0.011	1.934	1.589	-0.345
10-50 Wage	0.698	0.821	0.123	0.804	0.830	0.026	0.994	0.830	-0.164
50-90 Wage	0.770	0.698	-0.072	0.775	0.738	-0.038	0.940	0.759	-0.181
N	27,403	27,403		18,811	18,811		3,001	3,001	

Act. shows $\hat{Y}$ and $\hat{u}$ for the observed sample. *Counterfact.* shows $\hat{Y}$ and $\hat{u}$ if all returnees had stayed. Therefore, the observable characteristics of the sample correspond to the observables for both stayers and returnees. The unobservables correspond to the predicted u^*.

Table 9: Deciles of $\hat{Y}_i$ and $\hat{u}_i$ and Y_i, by Education Level, Parsimonious Model, Native-Born Men 35-55 Years Old.

Decile	Primary Education	Secondary Education	College Education
	Observables		
1	2.390	2.673	3.130
2	2.455	2.755	3.234
3	2.513	2.806	3.301
4	2.559	2.856	3.319
5	2.591	2.873	3.338
6	2.605	2.888	3.355
7	2.627	2.902	3.370
8	2.641	2.921	3.384
9	2.670	2.945	3.413
	Unobservables		
1	-0.670	-0.662	-0.662
2	-0.445	-0.413	-0.413
3	-0.285	-0.251	-0.251
4	-0.142	-0.118	-0.118
5	-0.010	0.007	0.007
6	0.115	0.127	0.127
7	0.244	0.252	0.252
8	0.403	0.399	0.399
9	0.635	0.614	0.614
	Log-Wage		
1	1.720	2.011	2.468
2	2.010	2.341	2.821
3	2.228	2.554	3.050
4	2.418	2.738	3.202
5	2.581	2.880	3.345
6	2.720	3.015	3.483
7	2.871	3.154	3.622
8	3.044	3.320	3.783
9	3.305	3.559	4.026
	Inequality Measures		
10-90 Wage	1.585	1.548	1.558
10-50 Wage	0.861	0.870	0.877
50-90 Wage	0.724	0.678	0.681
N	4,897	45,754	20,577

Act. shows $\hat{Y}$ and $\hat{u}$ for the observed sample. *Counterfact.* shows $\hat{Y}$ and $\hat{u}$ if all returnees had stayed. Therefore, the observable characteristics of the sample correspond to the observables for both stayers and returnees. The unobservables correspond to the predicted u^*.

Table 10: Probit and Wage Equation Estimates, Parametric Model, Men working for wages, 35-55 Years old.

	Probit Marginal Effects, $S = 1$	Wage Equation
Constant	0.959	1.823***
		(0.147)
Age	0.003***	0.007
	(1.30E-04)	(0.006)
Age Sq	-	0.000
	-	(0.000)
Primary Education	0.025***	0.019***
	(0.002)	(0.006)
Secondary Education	0.057***	0.139***
	(0.002)	(0.007)
College Education	0.014***	0.440***
	(0.002)	(0.011)
Married	-1.58E-04	0.064***
	(0.002)	(0.008)
Spouse US born	0.039***	0.060***
	(0.002)	(0.007)
Child	-0.044***	0.108***
	(0.001)	(0.005)
Child US born	0.153***	-
	(0.003)	(-)
5-10 Years in U.S.		0.013
		(0.010)
10-20 Years in U.S.		0.104***
		(0.008)
20-30 Years in U.S.		0.202***
		(0.008)
30-40 Years in U.S.		0.284***
		(0.011)
>40 Years in U.S.		0.359***
Lambda		-0.140***
		(0.018)
Industry indicators	No	Yes
Regional indicators	No	Yes
N	67381	62071

Standard errors in parentheses
Significance levels: *: 10%, **: 5%, ***: 1%.
The industry and regional indicators used in column (3) and (4) are the variables presented in the descriptive statistics.

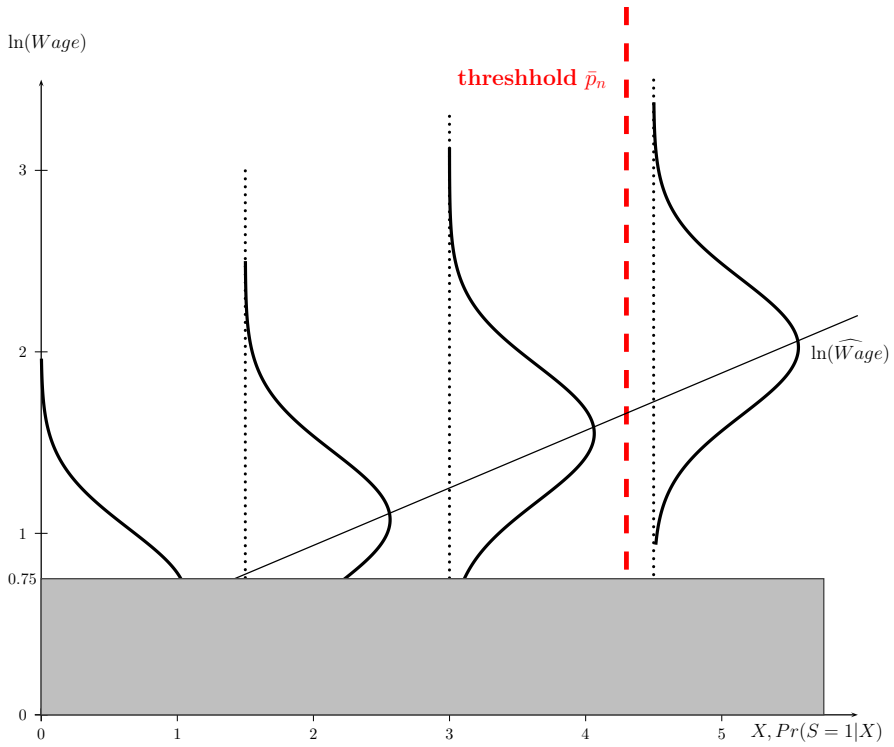
Table 11: Deciles of $\hat{Y}_i$ and $\hat{u}_i$ and Y_i, Various Specifications.

Decile	**Act.**	**Counterfact.**	**Log Diff**	**Act.**	**Counterfact.**	**Log Diff**	**Act.**	**Counterfact.**	**Log Diff**
	Full Specification			Parametric Model			25-45 Years		
					Observables				
1	2.214	2.209	-0.005	2.336	2.334	-0.002	2.219	2.210	-0.009
2	2.300	2.298	-0.003	2.428	2.428	0.000	2.279	2.277	-0.002
3	2.363	2.362	-0.001	2.466	2.461	-0.005	2.315	2.313	-0.002
4	2.422	2.415	-0.007	2.475	2.475	0.000	2.348	2.346	-0.002
5	2.473	2.463	-0.010	2.492	2.488	-0.004	2.387	2.378	-0.009
6	2.518	2.508	-0.010	2.506	2.504	-0.002	2.412	2.407	-0.005
7	2.572	2.560	-0.012	2.560	2.517	-0.043	2.443	2.433	-0.009
8	2.634	2.623	-0.011	2.625	2.625	0.000	2.487	2.477	-0.010
9	2.729	2.704	-0.025	2.656	2.656	-0.001	2.551	2.549	-0.001
Average	2.473	2.464	-0.009	2.514	2.511	-0.003	2.389	2.383	-0.006
					Unobservables				
1	-0.617	-0.636	-0.019	-0.689	-0.693	-0.004	-0.609	-0.638	-0.029
2	-0.431	-0.430	0.000	-0.505	-0.476	0.029	-0.434	-0.446	-0.012
3	-0.288	-0.272	0.016	-0.354	-0.298	0.056	-0.293	-0.293	0.000
4	-0.166	-0.134	0.031	-0.217	-0.150	0.067	-0.167	-0.158	0.008
5	-0.048	-0.012	0.036	-0.089	0.001	0.090	-0.046	-0.035	0.011
6	0.073	0.116	0.043	0.044	0.120	0.076	0.081	0.085	0.004
7	0.209	0.247	0.037	0.189	0.273	0.084	0.219	0.232	0.013
8	0.377	0.408	0.031	0.374	0.426	0.052	0.389	0.386	-0.003
9	0.638	0.612	-0.026	0.642	0.649	0.006	0.657	0.651	-0.006
Average	0.000	0.000	0.000	-0.041	-0.006	0.035	0.008	0.000	-0.008
					Log-Wage				
1	1.597	1.573	-0.024	1.647	1.641	-0.005	1.610	1.572	-0.038
2	1.870	1.867	-0.002	1.923	1.952	0.029	1.845	1.831	-0.015
3	2.075	2.090	0.015	2.112	2.163	0.051	2.022	2.020	-0.002
4	2.256	2.280	0.025	2.258	2.324	0.067	2.182	2.188	0.006
5	2.425	2.451	0.025	2.403	2.489	0.086	2.341	2.343	0.001
6	2.591	2.624	0.033	2.550	2.624	0.074	2.493	2.492	0.000
7	2.781	2.807	0.026	2.749	2.790	0.041	2.662	2.665	0.004
8	3.011	3.031	0.020	2.999	3.051	0.052	2.876	2.863	-0.013
9	3.367	3.317	-0.051	3.298	3.304	0.006	3.208	3.201	-0.007
Average	2.473	2.464	-0.009	2.473	2.505	0.032	2.397	2.383	-0.013
					Inequality Measures				
10-90 Wage	1.770	1.744	-0.027	1.652	1.663	0.011	1.598	1.629	0.031
10-50 Wage	0.828	0.878	0.050	0.756	0.848	0.091	0.732	0.771	0.039
50-90 Wage	0.942	0.866	-0.076	0.896	0.815	-0.080	0.866	0.858	-0.008
N	62071	62071		62071	62071		101819	101819	

Act. shows $\hat{Y}$ and $\hat{u}$ for the observed sample. *Counterfact.* shows $\hat{Y}$ and $\hat{u}$ if all returnees had stayed. Therefore, the observable characteristics of the sample correspond to the observables for both stayers and returnees. The unobservables correspond to the predicted u^*.
The Full Specification is based on the estimation in Table 5 column 4.
The Patametric Model is based on a fully parametric specification of the estimator presented in Section 4.
The last columns report the estimation of the same models presented in Table 5, column 2, for a sample of 25-45 Mexican Born Men.

B Figures

Figure 1: Intuitive Explanation of the Econometric Technique



1

Figure 2: Counterfactual Distributions under Different Definitions of High Probability Set

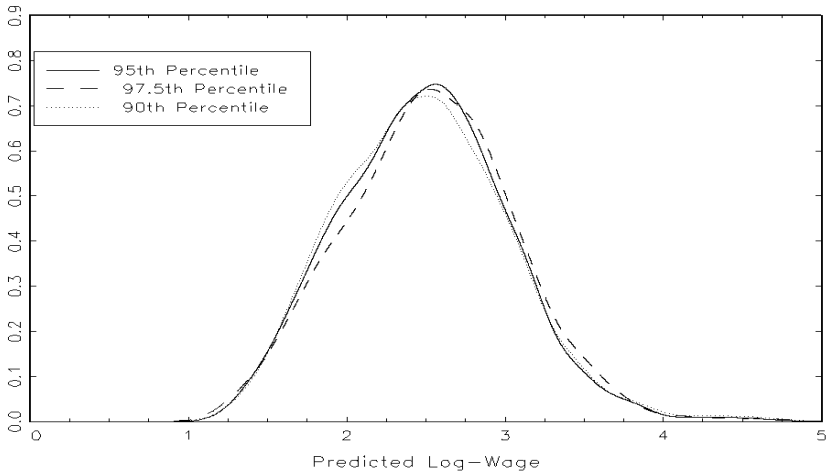
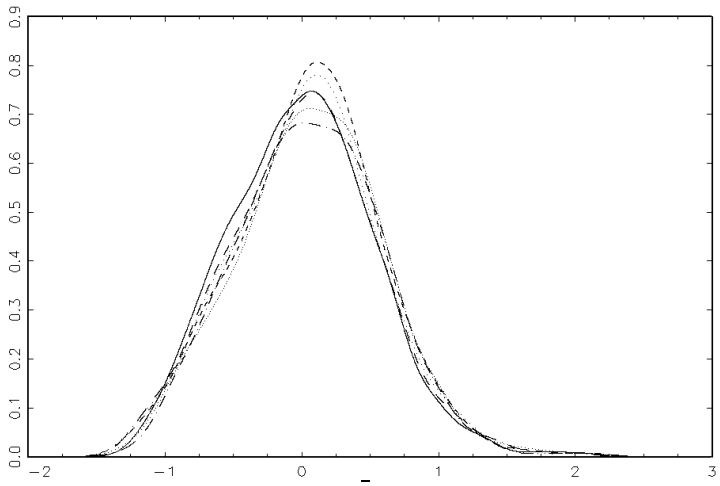
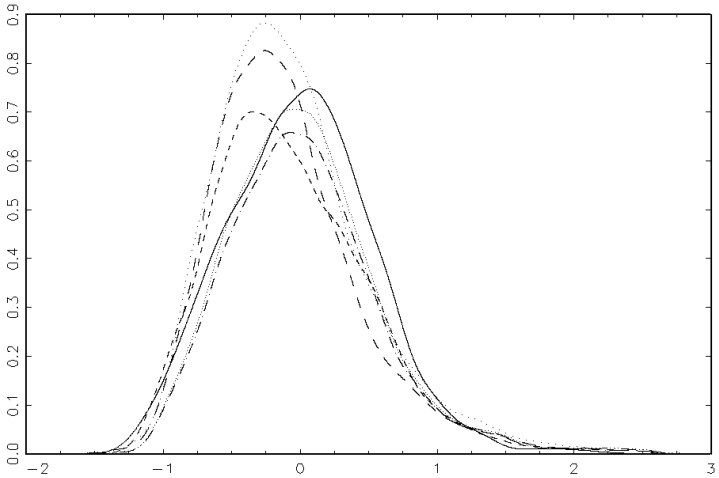


Figure 3: Estimated Unconditional (solid line) and Conditional Densities of the error term



(a) High Probability Set: $f(u^*)$ - solid line, and $f(u^*|Z_i'\hat{\alpha})$



(b) Stayers: $f(u^*)$ - solid line, and $f(u^*|Z_i'\hat{\alpha}, S=1)$

Figure 4: Actual (Solid Line) and Counterfactual (Dashed Line) Log-Wage Distributions, Parsimonious Model, Men, 35-55 Years Old.

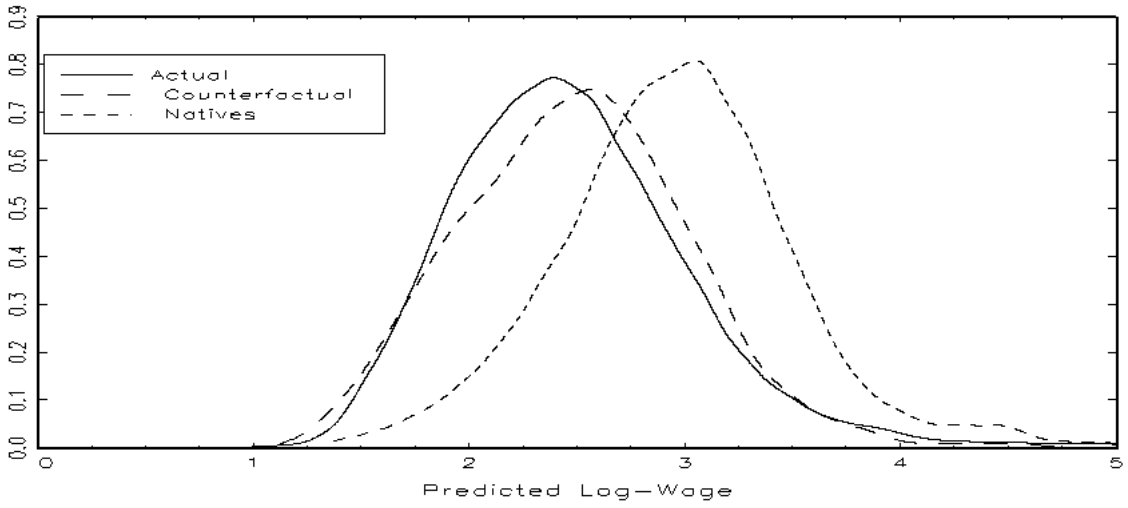


Figure 5: Difference in the Counterfactual and Actual Log-Wage Distributions, Parsimonious Model, Men 35-55 Years Old.

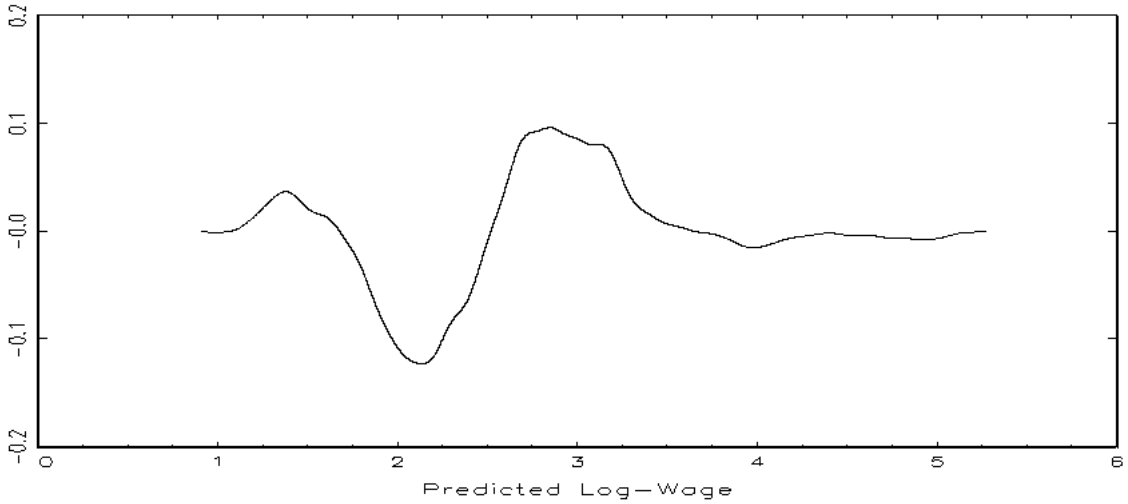
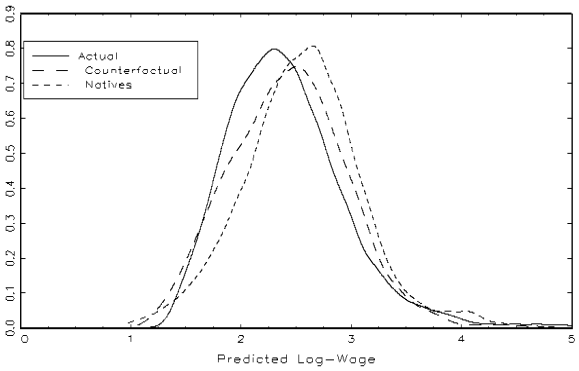
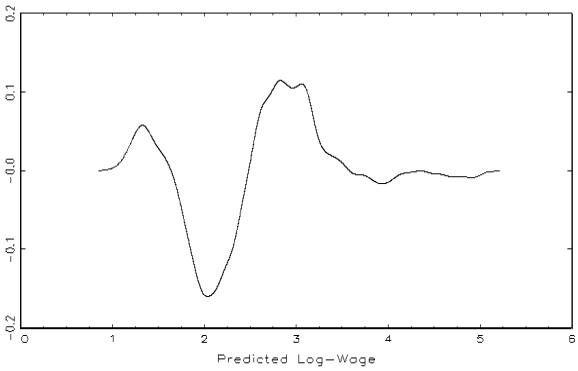


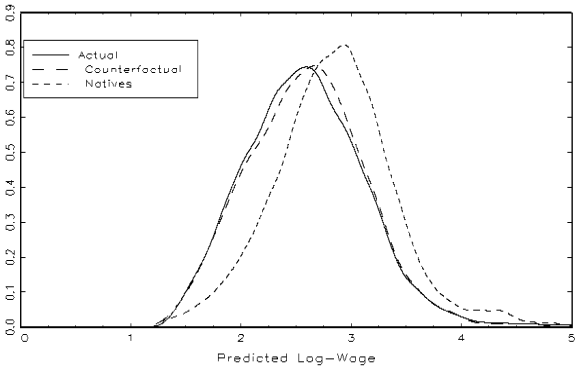
Figure 6: Estimated Actual and Counterfactual Log-Wage Densities for Mexican immigrants with Primary (a), Secondary (c), and College Education (e), Parsimonious Model, Men 35-55 Years Old.



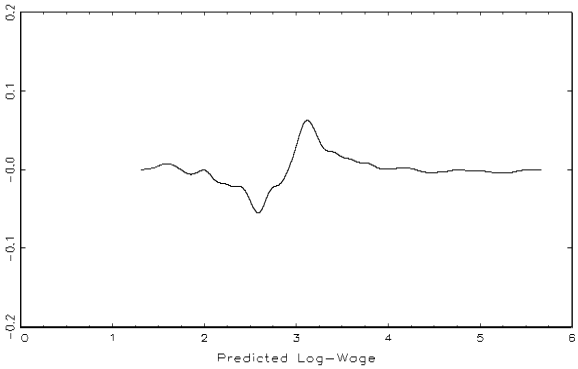
(a) Primary Education, $n = 27,403$



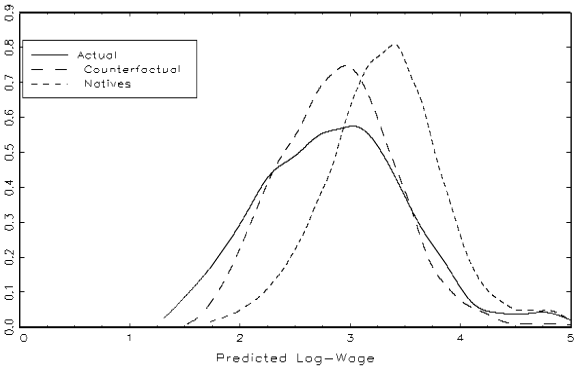
(b) Difference in Counterfactual and Actual Distributions, Primary Education



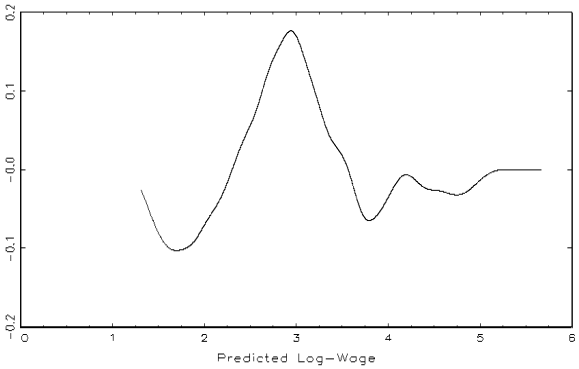
(c) Secondary Education, $n = 18,811$



(d) Difference in Counterfactual and Actual Distributions, Secondary Education

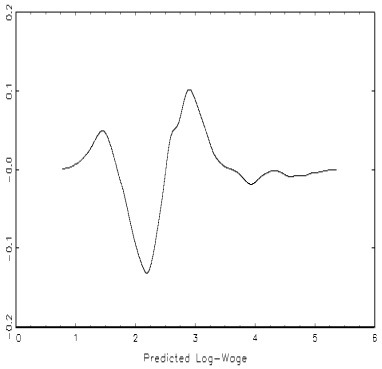


(e) College Education, $n = 3,001$

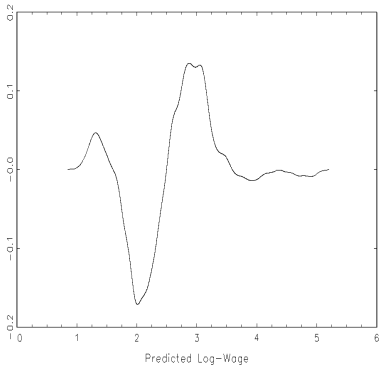


(f) Difference in Counterfactual and Actual Distributions, College Education

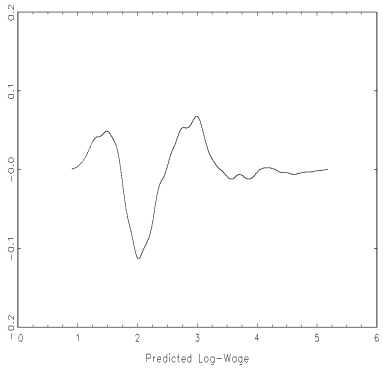
Figure 7: Difference in Counterfactual and Actual Log-Wage Densities for Mexican immigrants with Primary (a)-(c), Secondary (d)-(f), and College Education (g)-(i), Full Model, Parametric Model, Model with Men in Age 25-45.



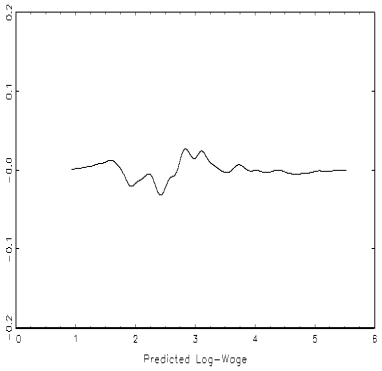
(a) Full Model, Primary Education



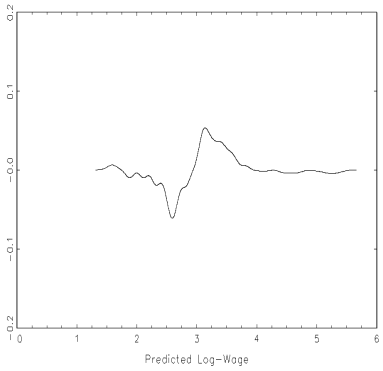
(b) Parametric Model, Primary Education



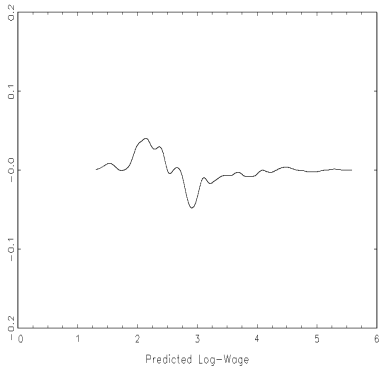
(c) Age 25-45, Primary Education



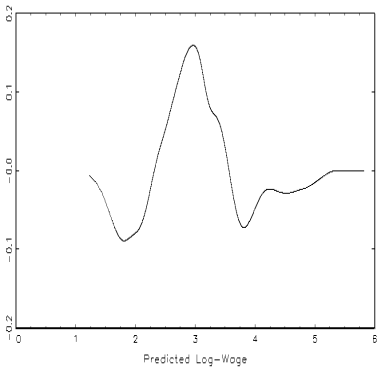
(d) Full Model, Secondary Education



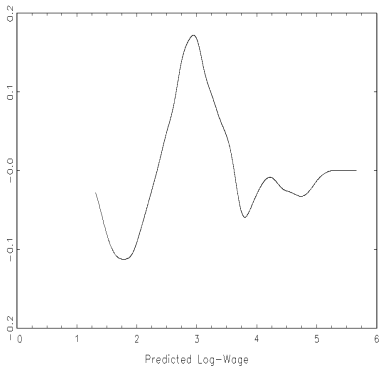
(e) Parametric Model, Secondary Education



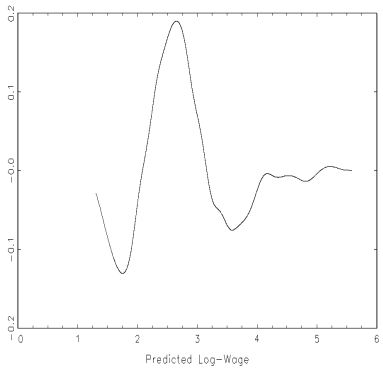
(f) Age 25-45, Secondary Education



(g) Full Model, College Education



(h) Parametric Model, College Education



(i) Age 25-45, College Education

References

Ambrosini, J. William and Giovanni Peri, "The Determinants and the Selection of Mexico-US Migrations," March 2011. Working Paper.

_ **, Karin Mayr, Giovanni Peri, and Dragos Radu**, "The Selection of Migrants and Returnees: Evidence from Romania and Implications," Working Paper 16912, National Bureau of Economic Research 2011.

Andrews, Donald W. K. and Marcia M.A. Schafgans, "Semiparametric Estimation of the Intercept of a Sample Selection Model," *The Review of Economic Studies*, 1998, *65* (3), 497–517.

Belot, Michele V.K. and Timothy J. Hatton, "Immigrant Selection in the OECD," Discussion Paper 571, CEPR February 2008.

Borjas, G. J., "Self-selection and the Earnings of Immigrants," *American Economic Review*, September 1987, *77*, 531–553.

_ , "Immigrant and Emigrants Earnings: A Longitudinal Study.," *Economic Inquiry*, 1989, *27* (1), 21–37.

Borjas, George J., "The Relationship between Wages and Weekly Hours of Work: the Role of Division Bias," *The Journal of Human Resources*, 1980, *15* (3), 409–423.

_ **and Bernt Bratsberg**, "Who Leaves? The Outmigration of the Foreign-Born," *The Review of Economics and Statistics*, February 1996, *78* (1), 165–176.
ΩÖzden et al.

Çağlar Özden, Christopher R. Parsons, Maurice Schiff, and Terrie L. Walmsley, "Where on Earth is Everybody? The Evolution of Global Bilateral Migration 1960-2000," *The World Bank Economic Review*, 2011, *25* (1), 12–56.

Chamberlain, Gary, "Asymptotic efficiency in semi-parametric models with censoring," *Journal of Econometrics*, 1986, *32* (2), 189 – 218.

Chiquiar, Daniel and Gordon H. Hanson, "International migration, self-selection, and the distribution of wages: evidence from Mexico and the United States," *Journal of Political Economy*, 2005, *113* (2), 239–278.

Chiswick, Barry R., "Human Capital and the Labor Market Adjustment of Immigrants: Testing Alternative Hypothesis," in Oded Stark, ed., *Research in Human Capital and Development*, Vol. 4, JAI Press, 1986, pp. 1–26.

DiNardo, John, Nicole M. Fortin, and Thomas Lemieux, "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, 1996, *64* (5), 1001–1044.

Dustmann, Christian, "Return migration, wage differentials and the optimal migration duration," *Europan Economic Review*, 2003, *47*, 353 – 369.

_ **, Itzhak Fadlon, and Yoram Weiss**, "Return Migration, Human Capital Accumulation and the Brain Drain," *Journal of Development Economics*, July 2011, *25* (1), 58–67. Working Paper.

Grogger, Jeffrey and Gordon H. Hanson, "Income Maximization and the Selection and Sorting of International Migrants," Working Paper 13821, National Bureau of Economic Research February 2008.

Heckman, James J., "Varieties of Selection Bias," *The American Economic Review*, 1990, *80* (2), 313–138, Papers and Proceedings.

Hu, Wei-Yin, "Immigrant earnings assimilation: estimates from Longitudinal data," *The American Economic Review*, 2000, *90* (2), 368–372.

Ibarraran, Pablo and Darren Lubotsky, "Mexican Immigration and Self-Selection: New Evidence from the 2000 Mexican Census," in "Mexican Immigration to the United States" NBER Chapters, National Bureau of Economic Research, Inc, 2007, pp. 159–192.

Jasso, Guillermina and Mark R. Rosenzweig, "Estimating the Emigration Rates of Legal Immigrants Using Administrative and Survey Data: The 1971 Cohort of Immigrants to the United States," *Demography*, 1982, *19* (3), 279–290.

Kaestner, Robert and Ofer Malamud, "Self-Selection and International Migration: New Evidence from Mexico," Working Paper 15765, National Bureau of Economic Research February 2010.

Klein, Roger W. and Richard H. Spady, "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica*, March 1993, *61* (2), 387–421.

_ , Chan Shen, and Francis Vella, "Semiparametric Selection Models with Binary Outcomes," Discussion Paper 6008, Institute for the Study of Labor (IZA) 2011.

Lacuesta, Aitor, "A Revision of the Self-Selection of Migrants Using Returning Migrants Earnings," *Annals of Economics and Statistics*, 2010, *97/98*, 235–259.

Lindstrom, D. P. and D. S. Massey, "Selective Emigration, Cohort Quality, and Models of Immigrant Assimilation," *Social Science Research*, 1994, *23* (4), 315 – 349.

Lubotsky, Darren, "Chutes or Ladders? A Longitudinal Analysis of Immigrant Earnings," *Journal of Political Economy*, 2007, *115* (5), 820–867.

Massey, Douglas S., "Understanding Mexican Migration to the United States," *The American Journal of Sociology*, 1987, *92* (6), pp. 1372–1403.

McKenzie, David and Hillel Rapoport, "Self-selection Patterns in Mexico-U.S. Migration: the Role of Migration Networks," *The Review of Economics and Statistics*, 2010, *92* (4), 811–821.

Moraga, Jesus Fernandez-Huertas, "New Evidence on Emigrant Selection," *The Review of Economics and Statistics*, 2011, *93* (1), 72–96.

Mulligan, Casey and Yona Rubinstein, "Selection, Investment, and Women's Relative Wages Over Time," *The Quarterly Journal of Economics*, 2008, *123* (3), 1061–1110.

Passel, Jeffrey S., "The Size and Characteristics of the Unauthorized Migrant Population in the U.S," Research Report, Pew Hispanic Center 2006.

Reagan, Patricia B. and Randall J. Olsen, "You Can Go Home Again: Evidence from Longitudinal Data," *Demography*, 2000, *37* (3), pp. 339–350.

Reinhold, Steffen and Kevin Thom, "Temporary Migration, Skill Upgrading, and Legal Status: Evidence from Mexican Migrants," *MEA DP*, 2009, *182.*

Robinson, P.M., "Root-N Consistent Semiparametric Regression," *Econometrica*, July 1988, *56* (4), 931–954.

Rosenzweig, Mark R., "Education and Migration: a Global Perspective," 2007. mimeo, Yale University.

Schafgans, Marcia M. A., "Ethnic Wage Differences in Malaysia: Parametric and Semiparametric Estimation of theChinese-Malay Wage Gap," *Journal of Applied Econometrics*, 1998, *13* (5), 481–504.

Stark, Oded and David E. Bloom, "The New Economics of Labor Migration," *The American Economic Review*, May 1985, *75* (2), 173–178.

Vella, Francis, "Estimating models with sample selection bias: a survey," *The Journal of Human Resources*, 1998, *33* (1), 127–169.

Yang, Dean, "Why Do Migrants Return to Poor Countries? Evidence from Philippine Migrants' Responses to Exchange Rate Shocks," *Review of Economic and Statistics*, November 2006, *88* (4), 715–735.