Kiel
Working Papers

Kiel Institute
for the World Economy

Empirical Characteristics of legal and Illegal Immigrants in the U.S.

by Vincenzo Caponi and Miana Plesca

No. 1835 | April 2013

Web: www.ifw-kiel.de

Kiel Working Paper No. 1835 | April 1835

Empirical Characteristics of Legal and Illegal Immigrants in the U.S.

Vincenzo Caponi and Miana Plesca

Abstract:
We combine the New Immigrant Survey (NIS), which contains information on US legal immigrants, with the American Community Survey (ACS), which contains information on legal and illegal immigrants to the U.S. Using econometric methodology proposed by Lancaster and Imbens (1996) we compute the probability for each observation in the ACS data to refer to an illegal immigrant, conditional on observed characteristics. The results for illegal versus legal immigrants are novel, since no other work has quantified the characteristics of illegal immigrants from a random sample. We find that, compared to legal immigrants, illegal immigrants are more likely to be less educated, males, and married with their spouse not present. These results are heterogeneous across education categories, country of origin (Mexico) and whether professional occupations are included or not in the analysis. Forecasts for the distribution of legal and illegal characteristics match aggregate imputations by the Department of Homeland Security. We find that, while illegal immigrants suffer a wage penalty compared to legal immigrants, returns to higher education remain large and positive.

Keywords: legal immigrants, illegal immigrants, contaminated controls

JEL classification: J15, F22

Vincenzo Caponi
Kiel Institute for the World Economy
24100 Kiel, Germany
Telephone:
E-mail: Vincenzo.caponi@ifw-kiel.de

Miana Plesca
University of Guelph
Guelph, Ontario
Canada
E-mail: miplesca@uguelph.ca

1 Introduction

In the recent history of immigration, a raising concern has been the increasing trend of illegal immigration, especially in developed countries. Most countries have in place immigration policies designed to welcome immigrants under terms deemed beneficial for the host country. Illegal immigration is often seen as problematic because, by its nature, it circumvents the control of policy makers. Immigrants eluding legal immigration channels are those who would not be accepted otherwise, because they are either in excess number or of different quality than desired by the destination country. Yet, every day, thousands of persons cross the border to start their new lives as illegal immigrants. In this paper, we provide answers to some questions about the illegal immigrants: how many they are, what are their characteristics, how do they differ from the legal immigrants, what determines their human capital, and how it is rewarded in the labor market.

We provide and apply methodology able to separate the legal from the illegal immigrants in a large U.S. national survey. Using information on immigrants in the U.S. from the American Community Survey (ACS) and the New Immigrant Survey (NIS), we are able to identify a set of conditional probability weights determining whether individuals are legal or illegal immigrants based on their observed characteristics. Since the ACS refers to the entire population of immigrants, legal and illegal, and the NIS refers to the sub-population of legal immigrants only, a difference between the two datasets can provide information on the characteristics and number of illegal immigrants. Given the conditional probability for a random non-native individual from the data to be illegal, we can compute statistics for the legal and illegal immigrants and their characteristics, thus getting a better understanding of who are the illegal immigrants and how do they compare with the legal ones.

There is a large related literature on the economic outcomes of immigrants to the U.S., and in particular of Mexican immigrants. Most of this literature uses the the Current Population Survey (CPS) or Census data to investigate economic outcomes for immigrants, sometimes in relation to those of the natives, without distinguishing between legal and illegal immigration status. Our paper is complementary to that literature because we focus primarily on a methodology to identify the illegal and legal immigrants in a large dataset comparable with

the U.S. Census.

A different approach has been to use information on legal and illegal migrants from Mexico using data available from the Mexican Migration Project (MMP), whose aim is to collect information on the legal and illegal Mexican migration to the U.S. The main drawback of the MMP is that it is not a random survey; instead, it selects certain rural communities from Mexico with high propensity migration rates. The MMP interviews households from these pre-specified communities where out-migration is more likely to occur, and asks questions about Mexican residents living in Mexico and in the US.[1] Because of the non-random design, any research conclusions using the MMP data cannot be generalized over all Mexican immigrants to the U.S., and even less so over the entire population of immigrants to the U.S.

Despite the fact that the MMP sample is not a representative random survey, it still provides some relevant information about the migration histories of the respondents, including whether they arrived in the US legally or illegally, and their socio-economic characteristics. From research using the MMP we have learned that most of the Mexican illegal migration is return migration, with about 85% of illegal immigrants returning home to Mexico; that the majority of illegal Mexcan migrants are young males working in farming; that the recent trends see a shift away from farming and construction into services; that women have very different migration patterns than men, in the sense that they are more likely to follow a spouse to the U.S. and are more likely to stay in the U.S. once they arrive, and that the traditional rural sources of Mexican immigration are shifting towards urban areas (Durand and Massey (2006)).

The other literature that provides estimates of the undocumented foreign born population in the US comes mostly from demographers who rely on the Residual Method. This method compares data from a representative survey, most often the U.S. Census or the CPS, with aggregate statistics on legal entrants provided by the Department of Homeland Security (DHS). The population survey is giving information on the size and the characteristics of the total population of foreign born residents in the U.S. at a given point in time. The DHS provides aggregate statistics for the inflows and outflows of individuals who are legally entitled to

[1]Since its inception in 1987, the MMP has surveyed every year between four to eight communities (during earlier survey years) and between two to five communities (during more recent survey years), to an overall total of eighty-one selected communities. For each household head and spouse, full migration and labour market histories are constructed from recall information; other household members are also interviewed about their first and their last trip to the U.S.

reside in the US. The aggregate measure of the unauthorized migrant population is given by the total foreign population minus the sum of all current and previous net flows of legal immigrants (also accounting for attrition through mortality). This measure should give a reasonably accurate count of the size of the illegal immigrant population (Passel (2006), Passel, Randolph, and Fix (2004)). Nevertheless, because of the aggregate nature of the DHS statistics, the Residual Method can only give the number of legal and illegal immigrants across broad aggregate dimensions, while important socio-economic characteristics such as age, education, or marital status remain missing from these analyses.[2]

Our proposed aproach is similar in spirit to the Residual Method, in that it compares two sets of information: one on legal immigrants (the NIS), the other on the general population of foreign born (the ACS).[3] The methodology we implement here was proposed by Lancaster and Imbens (1996) to deal with applications when the treated population can be identified form a sample of treated observations, or "cases", while the "control" population can not be immediately identified; instead, a mixed sample of case and control individuals is observed. In applying their methodology to our problem, the "cases" are the legal immigrants form the NIS survey and the mixed "case-control" sample are the immigrants from the ACS, where we cannot identify ex-ante who are legal immigrants − "cases"− and who are illegal immigrants −"controls". A random observation is drawn either from the case sample or from the mixed case-control sample, based on a Bernoulli process. Given covariates X, a likelihood function is written. Lancaster and Imbens (1996) provide moment conditions which are equivalent to maximizing the likelihood function. From the moment conditions we generate a set of legal and illegal immigration probabilities conditional on observed characteristics X.

[2]Related literature investigates the effect of U.S. border enforcement in stemming the flows of illegal Mexican immigration. This approach, referred to as the Apprehensions Method, does not provide an estimate of the number of illegal immigrants in the US, nor of their characteristics. However, it does provide information on the change in time of the inflow of legal and illegal immigrants (Rosenblum (2012)).

[3]In a spirit somewhat related to our approach, Burtless and Singer (2011) combine data from the MMP with CPS data to get a measure of how many illegal Mexicans contribute to Social Security (being illegal, they have no hope of withdrawing benefits, despite contributing to Social Security). Because they need to identify who in the CPS data is an undocumented immigrant, they use a matching algorithm which they call "cold decking" to infer who in the representative CPS data would be a legal or an illegal Mexican migrant based on the observed characteristics of legal and illegal migrants in the MMP data. While their approach has a very different context, it still suffers from the fact that the MMP is a non-random sample and the characteristics of legal and illegal migrants in MMS may be different from the overall characteristics of legal and illegal migrants in the Mexican migrant population.

Compared to existing studies, such as Passel (2006), is that we use representative microdata for both samples, which allows us to estimate not only the number of immigrants residing illegally, but also their personal characteristics, labor market performance, human capital determinants, along with any other information available in both surveys. The contribution of our paper relative to studies using the MMP is that we use representative random samples of legal and total population of immigrants, and, consequently, our results hold generally for all immigrants to the U.S., and not only for a selected subsample.

The paper proceeds as follows. Section 2 describes the two datasets used in the analysis, ACS and NIS, and section 3 describes how we adapt the contaminated-controls methodology proposed in Lancaster and Imbens (1996) to identify the propensity to be an illegal immigrant. Section 4 presents our main results and Section 5 provides some sensitivity checks and examples of the further analysis that can be pursued given our identification of the legal/illegal conditional probabilities. Sections A and B in the Appendix provide more sensitivity checks. Section 7 concludes.

2 Data and sample statistics

We describe here the two data sets we base our analysis on. The New Immigrant Survey (NIS) samples legal permanent residents (LPR) in the U.S. who acquired legal status, or green cards, in 2003. They are our sample of legal immigrants in the U.S. We compare them with a sample of all immigrants in the U.S., either legal or illegal, who are surveyed yearly in the American Community Survey (ACS). The difference in these two populations, all immigrants vs. legal immigrants, will give us a measure of the characteristics of legal and illegal immigrants to the U.S. In this section we detail the data filtering performed to ensure the two samples are comparable and representative of their underlying populations.

2.1 NIS

After a pilot project in 2001, the NIS started officially with its first wave in 2003. Within this flow of new legal residents, some individuals were already temporary residents of the US, while others entered the US for their first time only after having received their green card. Table 1

Table 1: NIS - Class of Admission - Adult Sample

Visa types in data	Unweighted	Weighted*
Spouse of U.S. Citizen	16.7%	34.2%
Spouse of Legal Permanent Resident	2.4%	2.4%
Parent of U.S. Citizen	11.6%	11.9%
Child of U.S. Citizen	3.3%	3.4%
Family Fourth Preference	6.2%	6.4%
Employment Preferences	19.5%	9.6%
Diversity Immigrants	16.9%	8.1%
Refugee	6.5%	6.6%
Legalization	7.7%	8.0%
Other	9.2%	9.4%
Total	100%	100%

*Note: Uses NIS survey weights.

reproduces the unweighted and weighted frequencies for each class of immigration as tabulated by the NIS.

Most of those who obtained their green card while already temporary residents qualified for the permanent status under the class of family reunification. The NIS under-samples the family reunification, and in particular those admitted to permanent residence because of marriage to US citizens − who are also more likely to have already been residing in the U.S. under legal visas. In contrast, people admitted through the visa lottery (Diversity Immigrants) and on the basis of arranged employment are over-sampled. To keep our experiment as clean as possible, we restrict our attention to the flow of new legal immigrants who *entered* the US in 2003, eliminating those who had been legally residing in the U.S. and changed their visa status to green card. This allows us to compare them with the overall flow of all new immigrants in the U.S. in 2003.

2.2 ACS

The ACS samples households randomly across the entire population of US residents, without distinction between citizens or aliens. For the foreign-born population, it makes no distinction between legal or illegal status.[4] Because of language barriers, immigrants are twice more likely

[4]The ACS, which has been piloted since 1996, is intended as a replacement for the Census Long form. While estimates from ACS are slightly less precise than those from the Census long form, a comparison of data from

to fail to complete the mail-in questionnaires. Consequently they are more likely to have their data collected during a second, in-person, phase of the interviewing process, resulting in data of better quality (albeit at the cost of higher standard errors, because only about a third are selected for in-person interviews).

In the ACS we restrict our attention to foreign born individuals who immigrated in the U.S. in 2003. Since our final analysis includes both data surveys, ACS and NIS, we need to be sure that the likelihood that each observation is drawn from the population is the same except for the legal/illegal status. That is, if we knew who was legal in the ACS data, we would like to make sure that this observation had exactly the same probability to be observed in the ACS as in the NIS.

Another issue related to weights refers to the possibility that non-response rates differ between legal and illegal immigrants. While we know little about the overall response rate of legal and illegal immigrants in the ACS survey, however, we do have some information about the response rates in the 2000 Census on which the ACS is based. The US Bureau of Census estimates (quote here) an overall non response rate of about 10% for the whole population, while among illegal immigrants this rate raises up to $15\% - 20\%$. To account for the differential non-response rates, we need to modify the survey weights to better reflect the under-counting of illegal immigrants relative to legals in the ACS. We deal with the theoretical implications in Section 3.1 where we show how to modify the survey weights such that they account for differences in response rates. We present in parallel results using two different set of weights, one corrected for differences in non-response, the other uncorrected.

Moreover, the NIS uses weights that represent the inverse of the probability for each observation to be randomly chosen, normalized such that the weights sum to the sample size. The ACS has a similar weighing scheme, except that the normalization is done such that the sum of all the weights reproduces the overall population in the US. We make all weights consistent across the surveys by normalizing the ACS weights, at the same time accounting for differences in the sampling frame.

---

the 2000 Census with data from the 1999-2001 ACS indicated that data quality from ACS was very close to the one in the Census (Camarota and Jeffrey (2004)). The obvious advantage of ACS over Census data is that it is a yearly survey, thus providing in a timely manner information on the characteristics of the foreign-born population. Compared to the CPS, estimates from ACS on the characteristics of the immigrant population are more precise.

2.3 Filtering the data: visa holders

One major concern in our analysis is the fact that the ACS includes not only legal and illegal immigrants, but also two other types of foreign born residents in the U.S. indistinguishable from the legal and illegal types: refugees and temporary visa holders. We are less concerned about refugees because they are few, but we are concerned about temporary visa holders as they represent a much larger number.

There are many different types of temporary visa that can be issued to temporary workers or other visitors to the US. In 2003 the US issued a total of 4,881,632 visas to foreign citizens to enter the US. Most of these visa, 3.64 million of them, were however issued to travelers for pleasure or business (B and C visa categories for tourists or crew members). Tourists are unlikely to be counted in the census or the ACS, and therefore we do not worry about them. We are left with a potential 1.23 million visa holders that could be drawn into our ACS sample. However, out of these 1.23 million about 520,000 are students (F and J visas); students can be easily identified in the ACS and the NIS and excluded from the analysis. The remaining visa holders, about 700,000 individuals, are more heterogeneous. If we exclude temporary residents belonging to military personnel (NATO visas) and foreign government personnel (A visas), we are left with about half a million visa holders. Most of the remaining visas are issued for occupations that require high skills (269,000 visas)[5] and to spouses and children of H visa holders (70,000) and of permanent residents (40,000). The remaining visas amount to about 110,000 H-2 visa holders, who are workers in the agricultural sector (30,000) and in other services (80,000).

In our analysis we can not distinguish who in the ACS belongs to this intermediate category, temporary visa holders. In order to minimize the extent of contamination, we exclude students from the analysis. Moreover, to identify the characteristics of legal and illegal immigrants we use data from the ACS 2007 (or, with similar results, from ACS 2006) restricted to foreign-born individuals who have immigrated in 2003. We do this because most of the temporary visas in the US have a period up to two years of validity, and therefore foreign-born workers who had

[5]About 110,000 visas were given for intracompany transferees and their spouses (L visas), about 30,000 to representatives and staff of international organizations, 10,000 to persons with extraordinary ability in the sciences, arts, education, business or athletics (O visas), 12,000 to representatives of foreign media (I visas) and 107,000 to workers of distinguished merit and ability (H-1 visas).

visa status in 2003 should no longer have the same visa status by 2007.[6]

However, visas can also be renewed. Most often a renewal occurs when there is some continuity in the occupation taken by the temporary resident, and often the temporary resident changes the status to a permanent one after one or few renewals. In any case, it is reasonable to believe that renewals happen mostly for highly skilled workers in high occupations. Therefore, as a further measure trying to exclude temporary visa holders from our analysis, in sensitivity analysis we exclude all persons from high occupations from the ACS and the NIS samples. Once we do that, our results do not change substantively, except the estimated fraction of legal immigrants increases slightly and moves closer to the DHS benchmark.

Finally, we use ACS 2007 compared to NIS to obtain the probabilities, conditional on observable characteristics, of each observation to belong to a legal or illegal immigrant. We use these probabilities as weights in three years of ACS data, 2005 to 2007, to predict the distribution of legal and illegal immigrant characteristics in the immigrant population, and to compare our result with some known statistics published by the DHS.

2.4 Other data manipulation

While the two data sets, NIS and ACS, are very comparable, we need to make sure that all variable definitions are consistent across the two sets.

In ACS we construct the hourly wage variable by dividing total yearly labour income by total hours worked in the year (hours per week times weeks a year). Note that yearly labour income will have more observations than hourly wages because some individuals report positive labour earnings but zero weeks or hours. In NIS wages/salaries are reported either as hourly wage, or as salaries which have attached a salary schedule; if the latter, we convert earnings into the hourly wage measure.

In both datasets we aggregate the information on education into the following categories: Less than 4 years ("None"), 5 to 8 years ("Elementary"), 9 to 12 years, no diploma ("Junior High"), High School diploma, College (Post-secondary education up to Bachelor's), and Higher

[6] In particular, we would expect that the 40,000 spouses or children of permanent residents from 2003 would have acquired their permanent resident status by 2007 and therefore should be no longer temporary residents. We would also expect that some of the 110,000 H-2 visas will have expired, since most of these visas are given to seasonal workers, or workers in the construction sector which usually does not guarantee permanent jobs.

Education (post-graduate, including Master's, Ph.D., and Professional degrees).

The ACS distinguishes between married individuals whose spouse is present or not. We reconstruct this information in NIS as well, so now the marital status categories are: "Married spouse present", "Married spouse absent", "Widowed", "Divorced", "Separated", "Never married". In the analysis we aggregate some of these categories (such as divorced, widowed or separated) into one.

Table 2: Summary statistics from ACS 2007 and NIS

	I ncluding professionals		E xcluding professionals	
	ACS	NIS	ACS	NIS
Gender (female)	0.490	0.554	0.374	0.397
Age	31.915	41.754	30.491	36.831
Married, spouse present	0.472	0.645	0.402	0.596
Married, no spouse present	0.161	0.095	0.179	0.094
Divorced/Widowed/Sep	0.089	0.079	0.081	0.058
Never married	0.278	0.180	0.337	0.253
None	0.077	0.103	0.081	0.034
Elementary	0.110	0.074	0.134	0.056
Junior High	0.184	0.202	0.217	0.251
High-School	0.290	0.245	0.337	0.310
College	0.244	0.292	0.200	0.297
Higher Education	0.094	0.085	0.031	0.052
Europe	0.083	0.137	0.066	0.161
Asia	0.240	0.408	0.150	0.347
America	0.254	0.204	0.292	0.268
Africa	0.053	0.144	0.051	0.153
Mexico	0.370	0.102	0.441	0.068
Sample Size	5335	4129	3244	1269

Note: Australia and Canada are excluded from the analysis.

3 Methodology

The methodology we apply is largely based on Lancaster and Imbens (1996) and Ridder and Moffitt (2007). Let s be a "stratum" indicator that takes value equal to one when an obser-

vation belongs to the NIS data set and zero when it belongs to the ACS. Let Y be a random variable for immigrant status which takes values 1 for legal immigrant and 0 for illegal; then, we know that if $s = 1$ then $Y = 1$ as well. However, when $s = 0$ we do not know what Y is since both legal and illegal immigrants are recorded in the ACS data. We further assume that the status of an immigrant is a choice variable determined by the following model:

$$Y = \begin{cases} 1 & \text{if} \quad Y^* > 0 \\ 0 & \text{if} \quad Y^* \leq 0 \end{cases}$$

where,

$$Y^* = x\beta + \epsilon \tag{1}$$

with $\epsilon \sim N(0, \sigma_\epsilon^2)$. The set of covariates x includes variables that are assumed to be relevant in determining the relative gain of immigrating legally in the US.

Since Y^* is the latent unobserved variable that determines the decision to be legal or illegal, we can write the probability of such choice as

$$P(Y = 1|x) = P(Y^* > 0|x) = P(\epsilon > -x\beta) = 1 - P(-x\beta) \tag{2}$$

and, assuming that $\epsilon \sim N(0, \sigma_\epsilon^2)$,

$$P(Y = 1|x) = 1 - P(-x\beta) = 1 - F(-x\beta) = 1 - \int_{-\infty}^{\frac{-x\beta}{\sigma_\epsilon}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = 1 - \Phi(-x\beta) = \Phi(x\beta) \tag{3}$$

The total number of legal immigrants in the population is therefore given by,

$$q = \int \Phi(x\beta) dF(x) \tag{4}$$

where $F(x)$ is the distribution function of the covariates x. We can also define the density function of the covariates x conditional on being observed in stratum 1 or 0. If they are observed in stratum 0, then their density function is the same as the unconditional density function, since stratum 0 randomly selects across all individuals,

$$p(x|s=0) = f(x) \tag{5}$$

if instead, the observation comes from stratum 1 then the density function is given by,

$$p(x|s=1) = \frac{\Phi(x\beta)f(x)}{q} \tag{6}$$

Accordingly with Lancaster and Imbens (1996), we assume that the pooled sample is determined by a sequence of Bernoully trials with parameter h unknown and independent on the other parameters of interest. Trials are repeated N times; each time if the trial results in a success we randomly sample from the subpopulation with $Y = 1$, if instead results in a failure we randomly select from the whole population. Therefore, in case of success the sampled observation enters stratum $s = 1$ ("legal"), and in case of failure the stratum $s = 0$. Then, h represents the probability that a randomly chosen observation from the pooled data belongs to stratum 1, while $1 - h$ to stratum 0.

Given these assumptions we can write the joint density of stratum and covariates as follows,

$$g(x,s) = \left(\frac{h\Phi(x\beta)f(x)}{q}\right)^{s} \cdot \Big((1-h)f(x)\Big)^{(1-s)} \tag{7}$$

The corresponding log likelihood function is therefore given by,

$$\ell(\beta,h,q) = \sum_{n=1}^{N} \left[s_n \log[\Phi(x_n\beta)f(x_n)/q] + (1-s_n)\log f(x_n)\right] + N_1 \log h + N_0 \log(1-h) \tag{8}$$

In the form derived in equation (8) to be operative the log likelihood needs that we know the density function $f(x)$. However, Lancaster and Imbens (1996) show that we can rewrite the likelihood greatly simplifying the problem and in particular eliminating the need to know the density of x. Define,

$$R_1(x;\beta,q,h) = \frac{\frac{h}{q}\Phi(x\beta)}{\frac{h}{q}\Phi(x\beta) + 1 - h} \tag{9}$$

$$g(x) = [\frac{h}{q}\Phi(x\beta) + 1 - h]f(x) \tag{10}$$

$$R_0(x;\beta,q,h) = 1 - R_1(x;\beta,q,h) \tag{11}$$

then we can re-write the log likelihood,

$$\ell(\beta, h, q, \pi) = \sum_{n=1}^{N} [s_n \log[R_{1n}(x; \beta, q, h)] + (1 - s_n) \log R_{0n}(x; \beta, q, h)] + \sum_{n=1}^{N} \log g(x_n; \pi) \quad (12)$$

Lancaster and Imbens (1996) show that it is sufficient to maximize the first part of the likelihood function in order to obtain the maximum of the whole function. That is because the the four variables in the likelihood function are related by a functional relationship which is implicitly imposed by the maximization of the first part.[7] However, ignoring the second part of the likelihood makes it impossible to perform any inference because we don't know the actual value the likelihood function takes. Lancaster and Imbens (1996) solve this problem by showing that the first order conditions for the maximization of the first part of the likelihood function can also be interpreted as a system of moments conditions leading to a Generalized Method of Moments (GMM) estimation procedure. As such, it is possible to proceed with the GMM estimation and with inference using the covariance matrix from the GMM.

3.1 Accounting for Different Response Rates

As the non-response rate is higher in the illegal immigrant population (20% relative to 10% in the legal immigrant population) we need to modify the model to reflect the under-counting of illegal immigrants relative to legals in the ACS. Because of under counting only a $\zeta_1 = 0.9$ portion of legal immigrants will enter the sample, while the portion of illegal will be only $\zeta_2 = 0.8$. Given these proportions, unconditionally on other characteristics, the probability that a randomly chosen observation from the ACS belongs to a legal immigrant is given by,

$$P(Y = 1|s = 0) = \frac{\zeta_1 P(Y = 1)}{\zeta_1 P(Y = 1) + \zeta_2(1 - P(Y = 1))} = \frac{\zeta_1 q}{\zeta_1 q + \zeta_2(1 - q)} \quad (13)$$

This implies that the density function of an observation from the ACS is given by

$$p(x|s = 0) = \frac{\zeta_1 q}{\zeta_1 q + \zeta_2(1 - q)} \frac{\Phi(x\beta)}{q} f(x) + \frac{\zeta_2(1 - q)}{\zeta_1 q + \zeta_2(1 - q)} \frac{1 - \Phi(x\beta)}{1 - q} f(x) = \quad (14)$$

$$[\xi_1(q)\Phi(x\beta) + \xi_2)(1 - \Phi(x\beta))]f(x) \quad (15)$$

[7] See Lancaster and Imbens (1996) pag. 149 for a discussion on this point.

Where $\xi_i(q) = \frac{\zeta_i}{\zeta_1 q + \zeta_2(1-q)} = \frac{\zeta_i}{\zeta_2 + (\zeta_1 - \zeta_2)q}$, or defining $\zeta = \zeta_1 - \zeta_2$, $\xi_i(q) = \frac{\zeta_i}{\zeta_2 + \zeta q}$. Accordingly, equations (9)-(11) rewrite,

$$R_1(x;\beta,q,h) = \frac{\frac{h}{q}\Phi(x\beta)}{\frac{h}{q}\Phi(x\beta) + (1-h)[\xi_1(q)\Phi(x\beta) + \xi_2(q)(1-\Phi(x\beta))]} \tag{16}$$

$$g(x) = [\frac{h}{q}\Phi(x\beta) + (1-h)[\xi_1(q)\Phi(x\beta) + \xi_2(q)(1-\Phi(x\beta))]]f(x) \tag{17}$$

$$R_0(x;\beta,q,h) = 1 - R_1(x;\beta,q,h) \tag{18}$$

Taking the derivative of the log likelihood function with respect to β we have, for the single observation,

$$\frac{\partial \ell(\beta,h,q)}{\partial \beta} = -R_1' \frac{s - R_1}{R_1(1-R_1)} \tag{19}$$

Defining N the numerator of R_1 and D the denominator, we have,

$$R_1' = \frac{N'}{D} - R_1\frac{D'}{D} = \phi(x\beta)'\frac{1}{D}\left\{\frac{h}{q} - R_1[\frac{h}{q} + (1-h)(\xi_1 - \xi_2)]\right\} \tag{20}$$

where $\phi(x\beta) = \frac{\partial \Phi(x\beta)}{\partial \beta}$ is a $1\times$ k vector, k being the dimension of the vector β. Therefore,

$$R_1' = \phi(x\beta)'\frac{1}{D}\left\{(1-R_1)[\frac{h}{q} - R_1(1-h)(\xi_1 - \xi_2)]\right\} \tag{21}$$

therefore,

$$\frac{\partial \ell(\beta,h,q)}{\partial \beta} = -\phi(x\beta)'(s - R_1)\left\{\frac{\frac{h}{q}}{DR_1} - \frac{(1-h)(\xi_1 - \xi_2)}{D(1-R_1)}\right\} \tag{22}$$

or,

$$\frac{\partial \ell(\beta,h,q)}{\partial \beta} = -\phi(x\beta)'(s - R_1)\left\{\frac{1}{\Phi(x\beta)} - \frac{(1-h)(\xi_1 - \xi_2)}{(1-h)[\xi_1(q)\Phi(x\beta) + \xi_2(q)(1-\Phi(x\beta))]}\right\} \tag{23}$$

3.2 Taking the derivative of the log likelihood function with respect to q we have, for the single observation

$$\frac{\partial \ell(\beta,h,q)}{\partial q} = -\frac{s - R_1}{1 - R_1}\left(\frac{1}{q} + \frac{D'}{D}\right) \tag{24}$$

where D is the denominator of R_1. Therefore, since,

$$D' = -\frac{1}{q}\frac{h}{q}\Phi(x\beta) + (1-h)[\xi_1'(q)\Phi(x\beta) + \xi_2'(q)(1-\Phi(x\beta))]] \tag{25}$$

where,

$$\xi_i' = -\frac{\zeta_i\zeta}{(\zeta_2+\zeta q)^2} = -\xi_i(q)\frac{\zeta}{\zeta_2+\zeta q} \tag{26}$$

therefore,

$$D' = -\frac{1}{q}\frac{h}{q}\Phi(x\beta) - (1-h)[\xi_1(q)\Phi(x\beta) + \xi_2(q)(1-\Phi(x\beta))]]\frac{\zeta}{\zeta_2+\zeta q} \tag{27}$$

so that,

$$\frac{D'}{D} = -\frac{1}{q}R_1 - (1-R_1)\frac{\zeta}{\zeta_2+\zeta q} \tag{28}$$

therefore,

$$\frac{\partial\ell(\beta,h,q)}{\partial q} = -\frac{s-R_1}{1-R_1}\left(\frac{1}{q}+\frac{D'}{D}\right) = -\frac{s-R_1}{1-R_1}\left(\frac{1}{q}-\frac{\zeta}{\zeta_2+\zeta q}\right)(1-R_1) \tag{29}$$

or,

$$\frac{\partial\ell(\beta,h,q)}{\partial q} = -\left(\frac{1}{q}-\frac{\zeta}{\zeta_2+\zeta q}\right)(s-R_1) \tag{30}$$

Equations (23) and (30) are the (single observation) equivalent of Equations (3.6) and (3.7) in Lancaster and Imbens (1996), Equation (3.8) remains exactly the same in the modified model. The GMM interpretation of Lancaster and Imbens represented by the system of equations in (3.9) of their article needs to be modified accordingly, and is given by,

$$\begin{aligned}
\psi_1(\beta,h,q,s,x) &= -\phi(x\beta)'(s-R_1)\left\{\frac{1}{\Phi(x\beta)} - \frac{(1-h)(\xi_1(q)-\xi_2(q))}{(1-h)[\xi_1(q)\Phi(x\beta)+\xi_2(q)(1-\Phi(x\beta))]}\right\} \\
\psi_2(\beta,h,q,s,x) &= -\left(\frac{1}{q}-\frac{\zeta}{\zeta_2+\zeta q}\right)(s-R_1) \\
\psi_3(\beta,h,q,s,x) &= h - R_1
\end{aligned} \tag{31}$$

4 Results

We start by presenting results for the probability of being legal or illegal in the U.S. conditional on observed characteristics X. As a first formal estimation of the impact of personal characteristics on the propensity of being a legal or illegal immigrant, these results represent the paper's main contribution.

4.1 Estimating the probability of being legal/illegal

Table 3 shows the results from the modified probit model described in the methodology section. The probability fitted by the model measures the likelihood that each observation belongs to a legal immigrant. We provide two sets of results, "adjusted" and "unadjusted"; in the former case the sampling weights do not account for different non-response rates between legal and illegal aliens in the U.S. (non-response rates being about 90% for legals and 80% for illegals), while in the latter they do. We focus our discussion on results that use the adjusted weights, but the two sets of results are very similar; in Appendix A we repeat the entire analysis using unadjusted sampling weights.

The numbers represent probit coefficients; marginal effects would be scaled down by a positive factor of $\varphi(X\beta)$. All independent variables are categorical dummies, except for age which is measured in years, and all coefficients with the exception of education beyond Bachelor's are statistically significant.

The positive coefficient on the gender variable indicates that women are less likely to be illegally present in the U.S. than men; women may be more risk averse than men, may have dependent children who make illegal immigration more costly, or the return from illegal immigration could be lower for women than for men. Relative to being single, being married with the spouse present increases the probability to be legally in the U.S., while being married and living without the spouse decreases it. Age and education have the expected effect on the probability to be legal: older immigrants and more educated immigrants are more likely to be legal. The only exception is for the highest category of education: all else equal, immigrants with Master's, Professional degrees, or Ph.D.'s seem to be less likely to be legal − although the coefficient is not statistically significant. The negative impact may be an indication that,

despite our best efforts, we may still mis-classify some visa holders as illegal immigrants – especially the professional H-type visas, the largest visa category not a-priori excluded from the analysis. With this in mind, we report in Section 5.1 sensitivity analysis from dropping professional occupations from the analysis.

Due to its proximity, Mexico is the biggest source of illegal immigration into the U.S., and the determinants of legal/illegal immigration may differ for Mexicans compared to other source countries. To this extent, we estimate a separate set of interaction dummies between schooling levels and Mexican origin. Indeed, we find that (i) at all levels of education, Mexicans are more likely to be illegal in the U.S. than immigrants from other countries, and (ii) in terms of education, for Mexicans the opposite holds true: the more educated an immigrant, the more likely s/he is to be undocumented in the U.S.[8]

Relative to baseline Europe, the continent of origin coefficients are significant and negative, except for Africa, indicating that African-born immigrants are more likely to be legally in the U.S. than Europeans, while Asians and Latin Americans are more likely to be illegally there.[9]

Lastly, q, the unconditional probability of being legal, seems to be slightly underestimated at .384 or .418, depending on the specification (correcting or not for different non-response rates). We can get slightly higher estimates of q, more in line with DHS projections, from the sensitivity analysis where we drop professionals and thus reduce some of the noise coming from visa observations.

4.2 Statistics on legal and illegal immigrants inferred from ACS

From the probit model discussed above we can compute for each observation the probability that an individual has legal or illegal status given her/his characteristics. We use these probabilities as weights to infer the distribution of certain characteristics in the legal (illegal) subpopulation of immigrants. We start by investigating discrete categorical variables Z, and

[8]For Mexicans the total effect of schooling on the propensity of legal immigration is the sum of the effect of Education (relative to Elementary) plus the effect of Mex∗Education (relative to Elementary). In our specification, it should be computed as the difference between the "Mex∗Education" interaction coefficient and the coefficient of "Mex∗ <Elementary (which is -.74 in the adjusted and -.75 in the unadjusted specifications). This overall effect is negative and small.

[9]Canadian observations are left out of the estimation, so America includes all of Central and South America except for Mexico, which is estimated alone.

we focus on the first moment, the mean. [10]

As a measure of how well our methodology performs, we compare the means predicted from our model for variables in the illegal population with benchmark statistics reported by the DHS Office of Immigration Statistics for 2007, as in Hoefer, Rytina, and Baker (2008). Means for education categories, marital status, country of origin, and U.S. state of current residence are reported in Table 4. The first column, (i) "All ACS", reports the means in the entire immigrant population using ACS samples from 2005 to 2007. The next two columns focus on the illegal subpopulation: (ii) "Illegal ACS" uses our methodology of computing means in the illegal subpopulation by applying the weights $1 - \omega_i(X_i)$, while (iii) "Illegal DHS" has benchmark estimates coming from the DHS benchmark statistics on undocumented immigrants. The last two columns provide the means in the legal immigrant subpopulation: (iv) "Legal ACS" computed by applying our legal probability weights ω_i to the ACS data, and (v) "NIS" which reports summary statistics from the NIS data on legal green card holders, and can thus be seen as our own benchmark for the legal immigrant subpopulation.

Note that we do not need that the observed characteristics X involved in the determination of the legal/illegal probabilities as reported in Table 3 match one-on-one with the variables whose mean we compute here. For instance, we did not use the state of residence variable when computing the legal/illegal weights because the state information was not reliable in the NIS data, where we know the state where the green card was mailed, which is not necessarily the state where the immigrant resides in 2007 (or even 2003 for that matter).

In terms of model fit, our statistics are a very reasonable match for the benchmark. For instance, there are 40% Mexican immigrants in the general population of immigrants (ACS); our procedure makes this fraction go up to to 48% in the illegal subpopulation, moving it closer to the 59% illegal Mexican immigrants reported by the DHS, Hoefer, Rytina, and Baker (2008). Our procedure estimates 43% women in the illegal subpopulation, compared to 44% in the DHS exstimates, and 54.6% women in the legal immigrant subpopulation, compared to 51.4% in the benchmark NIS. (In the general immigrant population in ACS the precentage of women is 45%). For Asians, the illegal percentage decreses from 23.1% in the overall immigrant

[10] The mean in the overall immigrant population is a weight between the legal and illegal means, with weights given by the unconditional probability of being legal (q) and illegal ($1 - q$).

population to 19.7% in the illegal population, whereas the DHS benchmark is 12%; the legal frequency increases to 36.6%, which is closer to the NIS 32.8% benchmark.

In the probit analysis, a more educated immigrant was more likely to be legal. We see some of the same effect here: post-secondary education is more prevalent (higher mean) in the legal subpopulation, while professional degrees, whose effect was statistically insignificant in the probit, have the same means in the legal and illegal subpopulations. Despite having positive coefficients in the probability model, we do not see lower levels of education (such as Junior High or High School) more frequently in the legal subpopulation. We believe this is due to Mexican immigrants for whom the probability to be legal is negatively related to education; their presence brings down the mean of education at lower levels of education within the legal subpopulation. Like in our previous discussion of probit coefficients, married individuals with spouse present are more frequent in the legal population, while single individuals or married with spouse absent are more frequent in the illegal population. As mentioned before, we did not use state of residence as a predictor for legal/illegal probabilities. In therms of forecast means, California, Texas and Arizona have more illegal immigrants than legal, New York has more legal immigrants, while Florida has about the same.

5 Sensitivity Analysis

5.1 Sensitivity to dropping professionals from ACS 2007

This section repeats the previous analysis with one difference: professionals are dropped from the estimation, to further minimize the chance of ACS including immigrants with legal temporary status. Table 5 reports the results of the probit estimation and Table 6 the corresponding means in the legal and illegal subpopulations, once professionals are dropped from the analysis. Here we report results from analysis which re-weights for non-response rates, while Appendix A contains sensitivity results when non-response rates are not accounted for.

The probit results are very much in line with those from the analysis on the entire sample which were reported in Table 3. One notable exception is the higher fraction of legal immigrants in the population, q, which is now .52 and respectively .49, depending on whether weights are adjusted or not for differential non-response rates in the surveys. This higher q

indicates that at least some of the professionals must have been identified as illegal previously. The two highest education categories are now grouped, because after dropping professional occupations their size has reduced considerably. At the same time, we see a much higher likelihood for an educated individual of being legal. In fact, compared to the base analysis, all education categories indicate a slightly higher probability of being legal relative to individuals with no education.Also, the coefficient on women has become negative, indicating that, once professionals are excluded, women are less likely to be legal than men.

From the distribution of covariates reported in Table 6 we can see that excluding professionals results in a smaller gap between the number of illegals forecast by our procedure and the benchmark DHS number, as well as for statistics for country of origin and U.S. state of residence. In particular, we predict a fraction of illegal Mexicans closer to the one reported by the DHS. We also get a better forecast fit for gender and some education categories such as high-school.

Further sensitivity analysis, using ACS 2006 for immigrants who respond having entered in 2003, is available from Appendix B.

Table 3: Probit Results: Conditional probability of being a legal immigrant from NIS and ACS 2007

	Unadjusted		Adjusted	
	Estimate	Std. Err.	Estimate	Std. Err.
Constant	-2.4244	0.1773	-2.4716	0.1713
Female	0.1215	0.0489	0.1190	0.0474
Age	0.0515	0.0041	0.0505	0.0040
Elementary	0.6491	0.1597	0.6368	0.1548
Junior High	1.0955	0.1731	1.0759	0.1674
High School	0.5520	0.1351	0.5414	0.1310
College	0.3450	0.1319	0.3386	0.1277
Higher Education	-0.0989	0.1444	-0.0972	0.1399
Married Spouse Present	0.1477	0.0554	0.1455	0.0538
Married Spouse not pres.	-0.5245	0.0866	-0.5144	0.0846
Mex*<Elementary	-0.7546	0.1948	-0.7394	0.1884
Mex*Elementary	-1.7213	0.1815	-1.6877	0.1771
Mex*Junior High	-2.1610	0.2083	-2.1205	0.2035
Mex*High School	-1.6654	0.1542	-1.6332	0.1510
Mex*College Degree	-1.6102	0.2121	-1.5794	0.2097
Mex*Higher Education	-1.0224	0.3891	-1.0020	0.3852
America	-0.8648	0.1095	-0.8473	0.1056
Africa	0.5357	0.1450	0.5273	0.1371
Asia	-0.3633	0.0942	-0.3551	0.0900
q	0.4175	0.0476	0.3839	0.0472
LogLik	-3574.2083		-3574.2285	
N.Obs	9464		9464	

Data: NIS 2003 and ACS 2007, excluding students, Canadians and Australians, and including professionals.
"Unadjusted" uses normalized sampling weights provided by each survey.
"Adjusted" rescales sampling weights to account for differences in non-response rates.
Reference category: Below elementary education (0 to 4 years of school), Single, European.
Education categories:

1. Below elementary = None to Grade 4 (base);
2. Elementary = Grades 5 to 8;
3. Junior High = Grades 9 to 12, no diploma;
4. High School diploma;
5. College = Post-secondary education up to Bachelor's;
6. Higher Education = Master's, Professional degrees, PH.D.

Table 4: Legal and Illegal Distributions from ACS 2005 - 2007 using immigrant weights from ACS 2007

	All ACS	Illegal		Legal	
		ACS	DHS	ACS	NIS
Below Elementary	0.0610	0.0672		0.0372	0.0845
Elementary	0.1606	0.1812		0.0811	0.1277
Junior	0.1399	0.1433		0.1265	0.1517
High School	0.2943	0.2930		0.2994	0.2705
College	0.2453	0.2174		0.3529	0.2634
Higher Education	0.0989	0.0979		0.1030	0.1023
Married with sp	0.4348	0.3862		0.6225	0.8063
Married no sp	0.1427	0.1597		0.0771	0.0475
Single	0.4221	0.4537		0.3003	0.1462
European	0.0892	0.0630	0.0200	0.1904	0.1427
Asian	0.2315	0.1966	0.1200	0.3665	0.3282
American	0.2321	0.2379	0.2400	0.2095	0.2548
African	0.0483	0.0226	0.0200	0.1477	0.1000
Mexican	0.3989	0.4799	0.5900	0.0859	0.1743
sex	0.4514	0.4267	0.4400	0.5465	0.5140
California	0.2038	0.2102	0.2400	0.1791	
Texas	0.1094	0.1177	0.1400	0.0776	
Florida	0.0911	0.0902	0.0800	0.0947	
Arizona	0.0330	0.0370	0.0500	0.0174	
New York	0.0890	0.0824	0.0500	0.1148	

Data: ACS 2005 to 2007, excluding students, Canadians and Australians, and including professionals.
Benchmarks: DHS = estimation on illegal demographics by the Department of Homeland Security; see Hoefer, Rytina, and Baker (2008). NIS = statistics on legal green card holders from the 2003 NIS (also used in computing the weights).
Estimation using the legal/illegal probability weights from ACS 2007, Table 3 (sampling weights rescaled to account for differences in non-response rates).

For a description of education categories see footnotes to Table 3.

Table 5: Probit Results: Conditional probability of being a legal immigrant from NIS and ACS 2007; **excluding professionals**

	Uncorrected		Corrected	
	Estimate	Std. Err.	Estimate	Std. Err.
Constant	-2.0166	0.3388	-2.0702	0.3265
Female	-0.0071	0.0941	-0.0070	0.0913
Age	0.0489	0.0071	0.0481	0.0070
Elementary	0.8234	0.2762	0.8087	0.2705
Junior High	1.3615	0.3076	1.3402	0.3003
High School	0.7078	0.2386	0.6955	0.2341
College and Higher	0.5987	0.2374	0.5880	0.2324
Married Spouse Present	0.1772	0.1030	0.1747	0.1002
Married Spouse not pres.	-0.6536	0.1572	-0.6426	0.1542
Mex*<Elementary	-1.4462	0.4188	-1.4201	0.4087
Mex*Elementary	-2.2500	0.3646	-2.2097	0.3538
Mex*Junior High	-2.5591	0.4047	-2.5162	0.3929
Mex*High School	-1.9233	0.3048	-1.8891	0.2945
Mex*College and Higher	-2.1489	0.4086	-2.1114	0.4021
America	-1.1044	0.2469	-1.0850	0.2348
Africa	0.4732	0.3023	0.4673	0.2819
Asia	-0.4731	0.2264	-0.4632	0.2126
q	0.5206	0.0866	0.4858	0.0866
LogLik	-1395.5067		-1395.5140	
N.Obs	4513		4513	

Data: NIS 2003 and ACS 2007, excluding students, Canadians and Australians, and excluding professionals.
"Unadjusted" uses normalized sampling weights provided by each survey.
"adjusted" rescales sampling weights to account for differences in non-response rates.
Reference category: Below elementary education (0 to 4 years of school), Single, European.

For a description of education categories see footnotes to Table 3.

Table 6: Legal and Illegal Distributions from ACS 2005-2007; **excluding professionals**. Weights from 2007 ACS.

	All ACS	Illegal		Legal	
		ACS	DHS	ACS	NIS
Below Elementary	0.0610	0.0761		0.0221	0.0845
Elementary	0.1606	0.1973		0.0663	0.1277
Junior	0.1399	0.1492		0.1158	0.1517
High School	0.2943	0.3025		0.2734	0.2705
College	0.2453	0.2027		0.3547	0.2634
Higher Education	0.0989	0.0722		0.1678	0.1023
Married with sp	0.4348	0.3614		0.6237	0.8063
Married no sp	0.1427	0.1713		0.0692	0.0475
Single	0.4221	0.4670		0.3070	0.1462
European	0.0892	0.0436	0.0200	0.2066	0.1427
Asian	0.2315	0.1670	0.1200	0.3975	0.3282
American	0.2321	0.2433	0.2400	0.2031	0.2548
African	0.0483	0.0160	0.0200	0.1314	0.1000
Mexican	0.3989	0.5301	0.5900	0.0615	0.1743
sex	0.4514	0.4262	0.4400	0.5160	0.5140
California	0.2038	0.2138	0.2400	0.1780	
Texas	0.1094	0.1233	0.1400	0.0738	
Florida	0.0911	0.0904	0.0800	0.0928	
Arizona	0.0330	0.0394	0.0500	0.0164	
New York	0.0890	0.0788	0.0500	0.1153	

Data: ACS 2005 to 2007, excluding students, Canadians and Australians, excluding professionals.
Benchmarks: DHS = estimation on illegal demographics by the Department of Homeland Security; see Hoefer, Rytina, and Baker (2008). NIS = statistics on legal green card holders from the 2003 NIS.
Estimation using the legal/illegal probability weights from ACS 2007 (no professionals), Table B3; sampling weights rescaled to account for differences in non-response rates.

For a description of education categories see footnotes to Table 3.

6 Immigrants' human capital

6.1 Returns to schooling and experience

Our methodology allows us to determine the conditional probability of each immigrant in the cross-section to be a legal or illegal resident. We use these probability weights in a Mincer wage regression to investigate the comparative returns to schooling and experience for legal and illegal immigrants. We consider in the wage regression all the immigrants from the 2005 to 2007 waves of the ACS who have immigrated since 2001. The conditional probability weights ofr being a legal or illegal immigrant are computed using ACS 2007 for immigrants who reported 2003 as their entry time in the US. The dependent variable is log wages, and thus the OLS coefficients can be approximated as percentage effects, except at larger values of the coefficients where the exact percentage values need to be computed as $exp(\beta) - 1$; we refer here to log points. While the findings are not surprising, this is a very relevant exercise because it quantifies the magnitude of human capital returns within the legal and illegal populations. We do the analysis separately by gender.

All else equal, being an illegal male immigrant brings a substantive wage penalty of 58 log points relative to a legal immigrant; for females, the penalty is 44 log points.Potential experience, which we construct as age-schooling-6, has the expected small positive effect on wages, at a decreasing rate. Wages grow between 4 to 6 percent each year, as indicated by survey year dummies.

Relative to elementary education, having some high-school but no diploma ("Junior High") seems to hurt legal immigrants, for whom the return is negative: a penalty of 19 log points for men and 12.7 log points for women. This is no longer true for illegal immigrants, especially for illegal male immigrants who get a positive return of about 5% from having more than elementary education. High-school diplomas seem to have no significant impact except for illegal male immigrants who get positive returns from having graduated high-school.

Having a college degree has large significant returns for both immigrant men and women relative to uneducated immigrants. The return is similar for legal and illegal women immigrants, but there is a penalty for illegal immigrant men. For them, the return to college, while still positive, is much smaller than for their legal immigrant counterparts: 43 log points

Table 7: Returns to Legal Status and Education from ACS 2005 - 2007

	Males		Females	
	Estimate	Std. Error	Estimate	Std. Error
Constant	2.3975	0.0582	2.1752	0.0603
Junior High	-0.1918	0.0637	-0.1269	0.0659
High School	-0.0064	0.0600	0.0746	0.0609
College	0.6661	0.0611	0.4042	0.0626
Higher Education	1.0167	0.0777	0.8202	0.0896
Illegal	-0.5826	0.0611	-0.4440	0.0660
Junior High/Illegal	0.2457	0.0689	0.1571	0.0779
High School/Illega	0.1510	0.0648	0.0343	0.0717
College/Illegal	-0.2303	0.0677	0.0737	0.0751
Higher Education/Illegal	-0.0540	0.0896	0.0880	0.1093
Experience	0.0243	0.0014	0.0186	0.0020
Experience^2	-0.0005	0.0000	-0.0005	0.0000
Survey Year 06	0.0352	0.0079	0.0360	0.0112
Survey Year 07	0.0591	0.0077	0.0750	0.0109
R^2	0.2438		0.1798	
N.Obs	43443		25057	

Data: ACS 2005-2007, individuals who have immigrated since 2001, excluding students, Canadians and Australians, and including professionals.
Estimation using the legal/illegal probability weights from ACS 2007, Table 3; sampling weights rescaled to account for differences in non-response rates.
"Illegal" defined as the conditional probability for each observation to be legal.

For a description of education categories see footnotes to Table 3.

compared to 66 log points. The same story holds for post-graduate education, where returns are very large for legal immigrants − almost double than for college −; the penalty for illegal immigrants in this case is smaller and not significant, for both men and women.

6.1.1 Sensitivity to using a different set of legal/illegal immigrant weights

For a sensitivity check, we repeat the wage regression analysis using a different set of legal/illegal conditional probabilities, obtained form ACS 2006. The results are reported in Table 8 and tell very similar story: there is an overall wage penalty from being an illegal immigrant relative to a legal one; on top of it, the penalty is even higher for college-educated men, but not for other education and demographic categories. The returns to college, and especially to post-graduate education are very large, and remain positive even for illegal immigrants. Relative to little or no education, all other levels of education receive a premium, except for

Table 8: Returns to Legal Status and Education from ACS 2005-2007 using illegal immigrant weights from ACS 2006

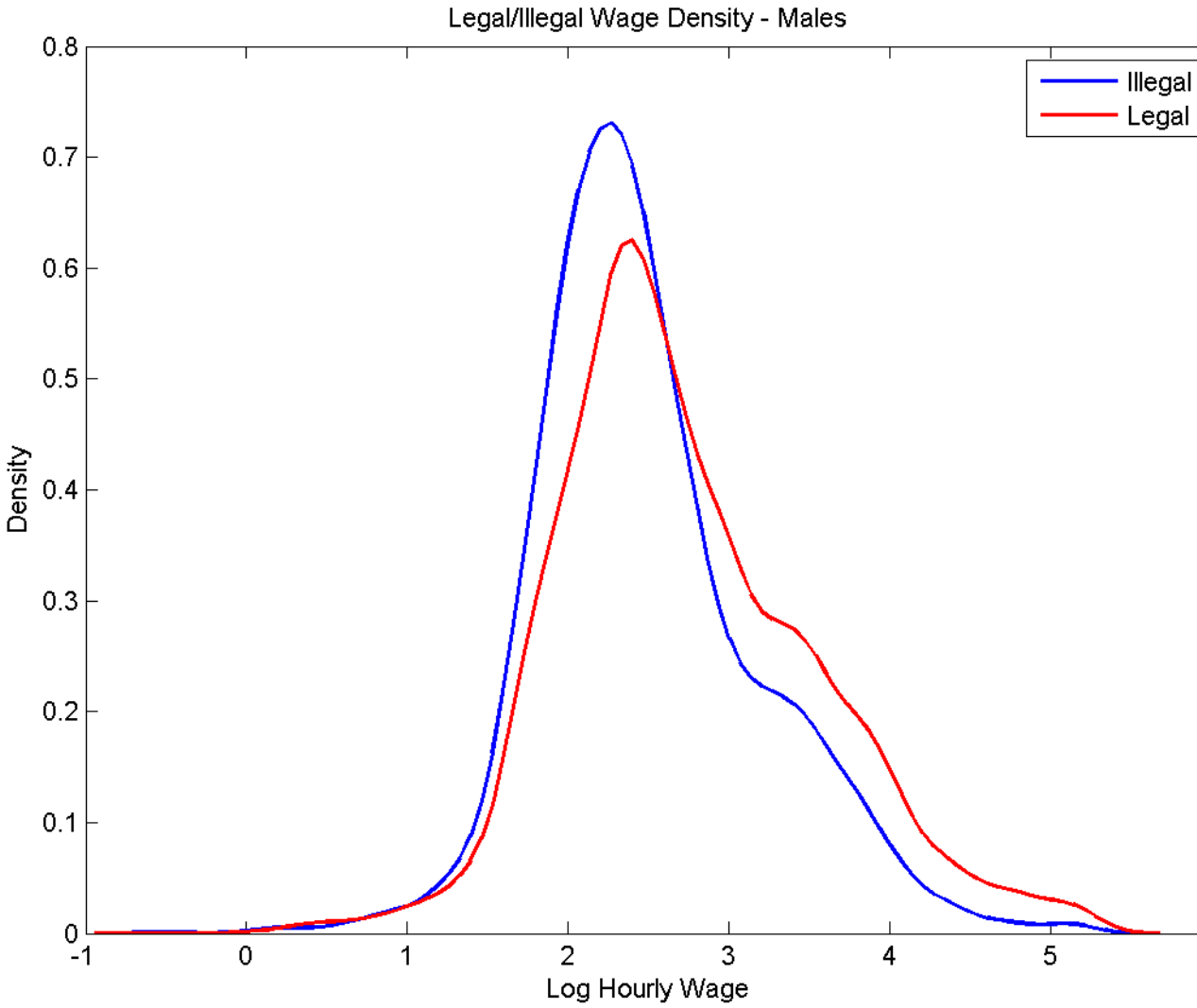
	Males		Females	
	Estimate	Std. Error	Estimate	Std. Error
Constant	2.4743	0.0789	2.2275	0.0753
Junior High	-0.1876	0.0888	-0.1318	0.0860
High School	0.0297	0.0839	0.1176	0.0801
College	0.7915	0.0852	0.4120	0.0821
Higher Education	1.1246	0.1188	0.8496	0.1323
Illegal	-0.6821	0.0826	-0.5145	0.0826
Junior High/Illegal	0.2507	0.0943	0.1735	0.0987
High School/Illega	0.1309	0.0892	0.0091	0.0917
College/Illegal	-0.3214	0.0922	0.0843	0.0955
Higher Education/Illegal	-0.1310	0.1312	0.0763	0.1531
Experience	0.0259	0.0014	0.0197	0.0020
Experience^2	-0.0005	0.0000	-0.0005	0.0000
Survey Year 06	0.0350	0.0079	0.0359	0.0112
Survey Year 07	0.0588	0.0077	0.0746	0.0109
R^2	0.2410		0.1787	
N.Obs	43443		25057	

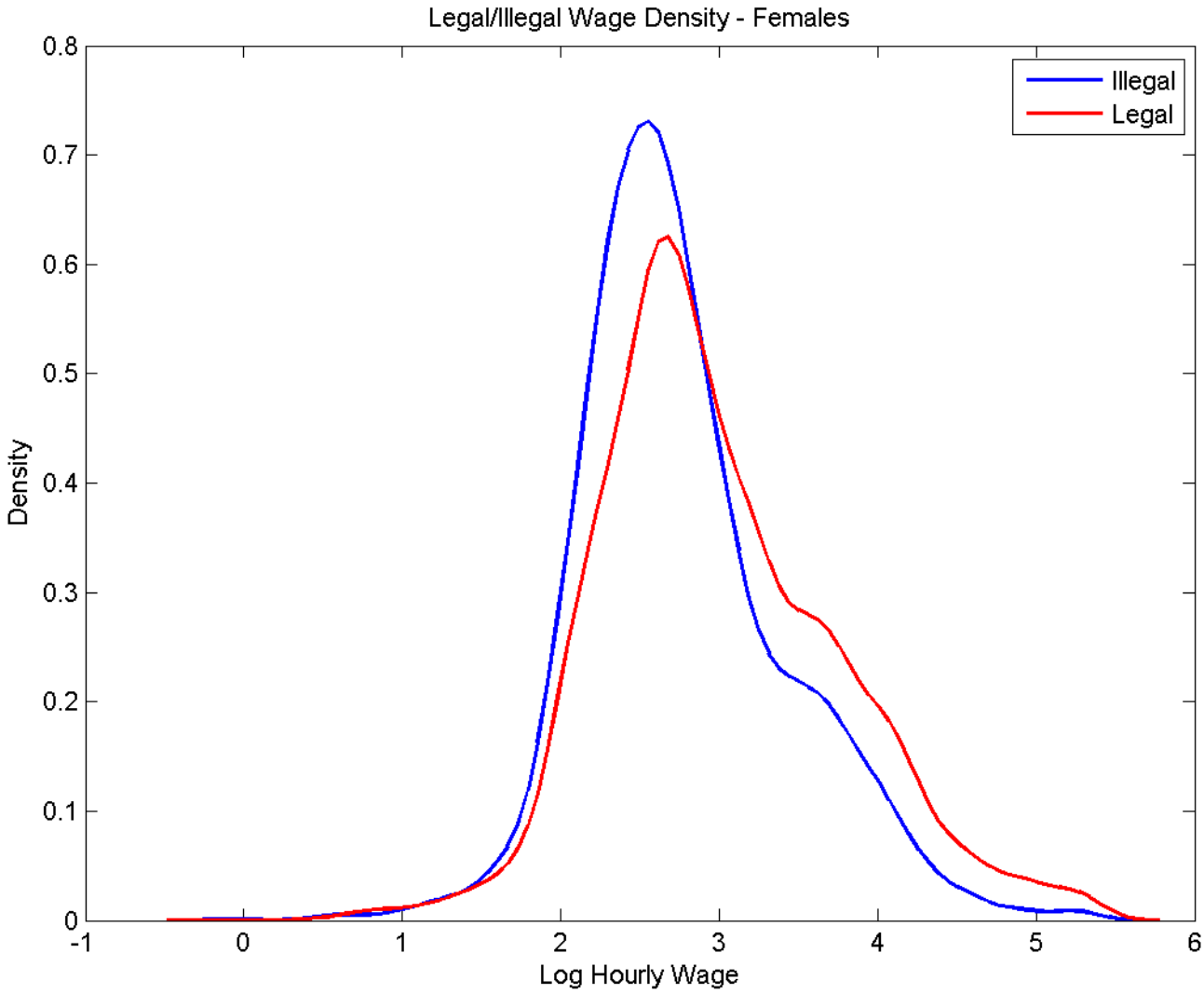
Data: ACS 2005-2007, individuals who have immigrated since 2001, excluding students, Canadians and Australians, and excluding professionals.
Estimation using the legal/illegal probability weights from ACS 2006, Table B3; sampling weights rescaled to account for differences in non-response rates.
"Illegal" defined as the conditional probability for each observation to be illegal, $1 - \omega_i(X_i)$.

For a description of education categories see footnotes to Table 3.

immigrants with junior high education, who fare worse on average. The illegal immigrant penalty varies by education categories, with a big difference between men and women: while for men the penalty seems to increase with higher education, the opposite is true for women. Put differently, highly educated immigrant women do not seem to be penalized too much for illegal status.

These results which illustrate that, as expected, illegal immigrants suffer a wage penalty due to their status, can be further seen from non-parametric wage distribution plots: the wage distribution for legal immigrants presents a higher mean and more skewness to the right compared to the wage distribution for illegal immigrant.


Legal/Illegal Wage Density - Males
Illegal
Legal
Density
Log Hourly Wage
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-1
0
1
2
3
4
5



While there certainly appears to be a penalty for illegal status, heterogeneous depending on eduction and gender, higher educated immigrants still get a substantive overall premium to their education. From a policy standpoint, this may warrant further thought into the welfare implications of a skill-selective immigration policy like the ones employed by Australia or Canada.

7 Conclusion

In this paper we have provided methodology to separate the legal and illegal immigrants from two random surveys in the U.S. Using information on all U.S. immigrants from ACS and information on legal U.S. immigrants from NIS, we were able to identify a set of probability weights which, conditional on observed characteristics, can determine the likelihood for each individual to be a legal or an illegal immigrant, based on the observed characteristics. From a substantive point of view, we wanted to use this methodology in investigating what are the characteristics of legal versus illegal immigrants and whether the legal status of an immigrant had an impact on their human capital, wages, and returns to human capital.

We have found that, compared to legal immigrants, illegal immigrants are more likely to be less educated, males, and married with their spouse not present. These results are heterogeneous across education categories, country of origin (Mexico) and whether professional occupations have been included in the analysis. While illegal immigrants experienced a wage penalty compared to legal immigrants, returns to higher education have remained large and positive. The penalty was found to be heterogeneous across education categories and gender, with women experiencing less penalty to illegal immigrant status compared to men at higher education levels. Further research can use the weights computed here to decompose wage differentials between legal and illegal immigrants at all quantiles of the earning distribution using the density re-weighting methodology from Fortin, Lemieux, and Firpo (2011).

Some caveats apply. We had to be extremely careful in how we treated immigrant visa holders, who we could not directly observe, and were concerned not to mis-identify as illegals. We believe that all the sensitivity analysis indicates that our approach was successful in that regard. Another caveat is we focus on 2003 flow data; as such, our methodology can generalize to other immigrant cohorts only to the extent that there have not been major demographic changes in the composition of legal versus illegal immigrant flows. If NIS releases subsequent waves of the survey, we can update the weights to reflect the experience of more recent immigrants.

We see as our main contribution the fact that we were able to use representative microdata to back out legal immigrant status out of personal characteristics, and then predict the relative

labor market performance of the two categories. Our methodology should be of interest to all researchers who need to make some inferences based on legal or illegal immigrant status.

References

Burtless, G., and A. Singer (2011): "The Earnings and Social Security Contributions of Documented and Undocumented Mexican Immigrants," No. 2 in Working Paper. Boston College Retirement Research Center.

Camarota, S., and C. Jeffrey (2004): "Assessing the Quality of Data Collected on the Foreign Born: An Evaluation of the American Comm unity Survey (ACS)," Methodology and data Quality. COPAFS The Council of Professional Associations on Federal Statistics.

Durand, J., and D. Massey (2006): *What We Learned from the Mexican Migration Project*vol. Crossing the Border: Research from the Mexican Migration Project. Russel Sage Foundation, New York.

Fortin, N., T. Lemieux, and S. Firpo (2011): *Decomposition Methods in Economics*vol. 4 of *Handbook of Labor Economics*, chap. 1, pp. 1–102. Elsevier.

Hoefer, M., N. Rytina, and B. C. Baker (2008): "Estimates of the Unauthorized Immigrant Population Residing in the United States: January 2007," *Population Estimates*, U.S. Department of Homeland Security, Office of Immigration Statistics(September).

Lancaster, T., and G. Imbens (1996): "Case-control studies with contaminated controls," *Journal of Econometrics,*, 71(1-2), 145–160.

Passel, J. (2006): "The Size and Characteristics of the Unauthorized Migrant Population in the U.S. Estimates Based on the March 2005 Current Population Survey," Research Report. PEW Hispanic Center.

Passel, J., C. Randolph, and M. Fix (2004): "Undocumented Immigrants: Facts and Figures," Immigration Studies Program. Urban Institute, Washington DC.

Ridder, G., and R. Moffitt (2007): *Chapter 75 The Econometrics of Data Combination*vol. Volume 6, Part 2 of *Handbook of Econometrics,*, pp. 5469–5547. Elsevier.

Rosenblum, M. (2012): "Border security: Immigration enforcement between ports of entry," No. 2 in Congressional Research Service. Washington, DC.

A Appendix A: Using 2007 ACS with sampling weights not adjusted for differences in non-response rates

Table A1: Legal and Illegal Distributions from ACS 2007, including professionals, survey sampling weights

	All ACS	Illegal		Legal	
		ACS	DHS	ACS	NIS
Below Elementary	0.0610	0.0672		0.0372	0.0845
Elementary	0.1606	0.1812		0.0811	0.1277
Junior	0.1399	0.1433		0.1265	0.1517
High School	0.2943	0.2930		0.2994	0.2705
College	0.2453	0.2174		0.3529	0.2634
Higher Education	0.0989	0.0979		0.1030	0.1023
Married with sp	0.4348	0.3862		0.6225	0.8063
Married no sp	0.1427	0.1597		0.0771	0.0475
Single	0.4221	0.4537		0.3003	0.1462
European	0.0892	0.0630	0.0200	0.1904	0.1427
Asian	0.2315	0.1966	0.1200	0.3665	0.3282
American	0.2321	0.2379	0.2400	0.2095	0.2548
African	0.0483	0.0226	0.0200	0.1477	0.1000
Mexican	0.3989	0.4799	0.5900	0.0859	0.1743
sex	0.4514	0.4267	0.4400	0.5465	0.5140
California	0.2038	0.2102	0.2400	0.1791	
Texas	0.1094	0.1177	0.1400	0.0776	
Florida	0.0911	0.0902	0.0800	0.0947	
Arizona	0.0330	0.0370	0.0500	0.0174	
New York	0.0890	0.0824	0.0500	0.1148	

Data: ACS 2005 to 2007, excluding students, Canadians and Australians, but including professionals.
Benchmarks: DHS = estimation on illegal demographics by the Department of Homeland Security; see Hoefer, Rytina, and Baker (2008). NIS = statistics on legal green card holders from the 2003 NIS.
Estimation using the legal/illegal probability weights from ACS 2007, Table 3. No correction for differences in non-response rates.
For a description of education categories see footnotes to Table 3.

B Appendix B: Analysis using ACS 2006

Table B2: Probit Results: Conditional probability of being a legal immigrant from NIS and ACS 2006.

	Unadjusted		Adjusted	
	Estimate	Std. Err.	Estimate	Std. Err.
Constant	-2.2946	0.1600	-2.3433	0.1561
Female	0.1699	0.0435	0.1667	0.0422
Age	0.0399	0.0038	0.0392	0.0038
Elementary	0.4092	0.1431	0.4017	0.1385
Junior High	0.6923	0.1482	0.6793	0.1435
High School	0.2239	0.1232	0.2199	0.1190
College Degree	0.1318	0.1234	0.1294	0.1192
Higher Education	-0.2923	0.1365	-0.2867	0.1321
Married Spouse Present	0.1804	0.0490	0.1769	0.0477
Married Spouse not pres.	-0.4227	0.0786	-0.4147	0.0770
Mex*¡Elementary	-0.5447	0.1735	-0.5352	0.1680
Mex*Elementary	-1.3791	0.1626	-1.3533	0.1600
Mex*Junior High	-1.7446	0.1850	-1.7120	0.1825
Mex*High School	-1.3186	0.1342	-1.2945	0.1326
Mex*College Degree	-1.3982	0.1979	-1.3726	0.1963
Mex*Higher Education	-0.8849	0.3683	-0.8690	0.3633
America	-0.5922	0.0894	-0.5812	0.0871
Africa	0.6292	0.1302	0.6178	0.1247
Asia	-0.1755	0.0726	-0.1723	0.0699
q	0.3124	0.0575	0.2837	0.0568
LogLik	-3654.9818		-3655.0017	
N.Obs	9647		9647	

Data: NIS 2003 and ACS 2006, excluding students, Canadians and Australians, and including professionals.
"Uncorrected" uses normalized sampling weights provided by each survey.
"Corrected" rescales sampling weights to account for differences in non-response rates.
Reference category: Below elementary education (0 to 4 years of school), Single, European.
For a description of education categories see footnotes to Table 3.

Table B3: Legal and Illegal Distributions from ACS 2005-2007 using immigrant weights from 2006 ACS.

	All ACS	Illegal		Legal	
		ACS	DHS	ACS	NIS
Below Elementary	0.0610	0.0633		0.0476	0.0845
Elementary	0.1606	0.1731		0.0866	0.1277
Junior	0.1399	0.1423		0.1257	0.1517
High School	0.2943	0.2962		0.2835	0.2705
Some College	0.2453	0.2255		0.3619	0.2634
Higher Education	0.0989	0.0997		0.0947	0.1023
Married with sp	0.4348	0.4005		0.6372	0.8063
Married no sp	0.1427	0.1543		0.0743	0.0475
Single	0.4221	0.4448		0.2884	0.1462
European	0.0892	0.0756	0.0200	0.1693	0.1427
Asian	0.2315	0.2076	0.1200	0.3728	0.3282
American	0.2321	0.2362	0.2400	0.2074	0.2548
African	0.0483	0.0287	0.0200	0.1640	0.1000
Mexican	0.3989	0.4518	0.5900	0.0865	0.1743
sex	0.4514	0.4312	0.4400	0.5699	0.5140
california	0.2038	0.2079	0.2400	0.1797	0.0000
texas	0.1094	0.1147	0.1400	0.0786	0.0000
florida	0.0911	0.0909	0.0800	0.0924	0.0000
arizona	0.0330	0.0356	0.0500	0.0174	0.0000
ny	0.0890	0.0849	0.0500	0.1135	0.0000

Data: ACS 2005 to 2007, excluding students, Canadians and Australians, and including professionals.

Benchmarks: DHS = estimation on illegal demographics by the Department of Homeland Security; see Hoefer, Rytina, and Baker (2008). NIS = statistics on legal green card holders from the 2003 NIS.

Estimation using the legal/illegal probability weights from ACS 2006, Table B2; sampling weights rescaled to account for differences in non-response rates.

For a description of education categories see footnotes to Table 3.

Table B4: Probit Results: Conditional probability of being a legal immigrant from NIS and ACS 2006; excluding professionals.

	Unadjusted		Adjusted	
	Estimate	Std. Err.	Estimate	Std. Err.
Constant	-1.6483	0.3368	-1.7085	0.3229
Female	-0.0671	0.0757	-0.0660	0.0735
Age	0.0472	0.0054	0.0466	0.0052
Elementary	0.5804	0.3883	0.5744	0.3678
Junior High	1.1105	0.4527	1.1068	0.4280
High School	0.1963	0.2995	0.1939	0.2850
College Degree	-0.2913	0.2949	-0.2869	0.2808
Higher Education	-0.9267	0.3059	-0.9136	0.2921
Married Spouse Present	-0.0168	0.0837	-0.0165	0.0816
Married Spouse not pres.	-0.2377	0.1484	-0.2335	0.1444
Mex*¡Elementary	-0.4564	0.4001	-0.4501	0.3802
Mex*Elementary	-1.7794	0.3346	-1.7565	0.3202
Mex*Junior High	-2.2937	0.4381	-2.2704	0.4194
Mex*High School	-1.4634	0.2093	-1.4406	0.2054
Mex*College Degree	-1.3312	0.2718	-1.3108	0.2698
Mex*Higher Education	-0.5005	0.4420	-0.4911	0.4382
America	-0.4571	0.1301	-0.4494	0.1260
Africa	1.0787	0.3164	1.0719	0.2981
Asia	-0.2520	0.1032	-0.2485	0.0999
q	0.5342	0.0634	0.5028	0.0631
LogLik	-2021.3182		-2021.2919	
N.Obs	5039		5039	

Data: NIS 2003 and ACS 2006, excluding students, Canadians and Australians, and excluding professionals.
"Uncorrected" uses normalized sampling weights provided by each survey. "Corrected" rescales sampling weights to account for differences in non-response rates.
Reference category: Below elementary education (0 to 4 years of school), Single, European.
For a description of education categories see footnotes to Table 3.

Table B5: Legal and Illegal Distributions from ACS 2005-2007 (excluding professionals) using immigrant weights from 2006 ACS.

	All ACS	Illegal		Legal	
		ACS	DHS	ACS	NIS
Below Elementary	0.0610	0.0565		0.0730	0.0845
Elementary	0.1606	0.1741		0.1245	0.1277
Junior	0.1399	0.1317		0.1618	0.1517
High School	0.2943	0.2849		0.3196	0.2705
Some College	0.2453	0.2384		0.2637	0.2634
Higher Education	0.0989	0.1145		0.0574	0.1023
Married with sp	0.4348	0.4123		0.4950	0.8063
Married no sp	0.1427	0.1472		0.1306	0.0475
Single	0.4221	0.4401		0.3742	0.1462
European	0.0892	0.0736	0.0200	0.1310	0.1427
Asian	0.2315	0.2129	0.1200	0.2815	0.3282
American	0.2321	0.2039	0.2400	0.3074	0.2548
African	0.0483	0.0130	0.0200	0.1428	0.1000
Mexican	0.3989	0.4967	0.5900	0.1373	0.1743
sex	0.4514	0.4435	0.4400	0.4723	0.5140
california	0.2038	0.2152	0.2400	0.1733	0.0000
texas	0.1094	0.1192	0.1400	0.0834	0.0000
florida	0.0911	0.0854	0.0800	0.1064	0.0000
arizona	0.0330	0.0379	0.0500	0.0198	0.0000
ny	0.0890	0.0786	0.0500	0.1169	0.0000

Data: ACS 2005 to 2007, excluding students, Canadians and Australians, and excluding professionals.

Benchmarks: DHS = estimation on illegal demographics by the Department of Homeland Security; see Hoefer, Rytina, and Baker (2008). NIS = statistics on legal green card holders from the 2003 NIS.

Estimation using the legal/illegal probability weights from ACS 2006 (no professionals), Table B4; sampling weights rescaled to account for differences in non-response rates.

For a description of education categories see footnotes to Table 3.