Essays in
Unauthorized Immigration and Migration

by

Bilesha B. Weeraratne

A dissertation submitted to the Graduate Faculty in Economics
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy,
The City University of New York

2012

This manuscript has been read and accepted for the
Graduate Faculty in Economics in satisfaction of the
dissertation requirement for the degree of Doctor of Philosophy.

Dr. David Jaeger

Date	Chair of Examining Committee

Dr. Merih Uctum

Date	Executive Officer

Dr. David Jaeger

Dr. Wim Vijverberg

Dr. Michael Grossman

Dr. Partha Deb

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

Essays in Unauthorized Immigration and Migration

by

Bilesha B. Weeraratne

Adviser: Professor David Jaeger

This dissertation consists of three essays. The first essay develops a new methodology defined as the microdata-based methodology to identify unauthorized immigrants in the US. This identification is based on a discrete choice model on observed authorized and unauthorized immigrant data in 1986 with adjustments for time dynamics between 1986 and 2010 and for calibration in new data. This method produces an algorithm to identify unauthorized immigrants in the American Community Survey (ACS) 2010, and estimates that there were 7,700,869 adult unauthorized immigrants in the US on January 1, 2011. The essay includes detailed descriptive analyses of unauthorized immigrants and their children.

In the second essay, I evaluate the impact of parents' legal status on the high school drop out probability of 16-18 year old children. The analysis uses data from the ACS 2010, and the methodology is a linear probability model. The study finds that when controlling for other covariates, a child of unauthorized immigrants has a lower drop out probability than a similar child of authorized immigrant parents, and among unauthorized immigrant parents, a non-citizen child has a higher drop out probability than a similar native-born child. In mixed families, among unauthorized immigrant mothers, a non-citizen child has a higher drop out probability than an otherwise similar native-born child, and children with the parent combination of highly unauthorized and highly legal immigrants have a higher high school drop out probability than those with highly unauthorized and unsure legal immigrant parents.

In the third essay, I estimate a multinomial logit model to analyze the contextual determinants of labor migration in Sri Lanka, using data from the Consumer Finance and Socio Economic Survey 2003/4 of the Central Bank of Sri Lanka. The study finds higher internal migration probabilities for residents of rural areas, districts with a lower degree of structural transformation, and districts with a larger share of population in 19-34 years. Internal migration probabilities are lower for residents of districts with large labor forces.

Acknowledgements

I am thankful to Dr. David Jaeger for his guidance and advice during my doctoral studies. I am especially thankful to him for believing in my idea for the first essay and for showing me the research potential in the area of children of unauthorized immigrants, which formulated into my second essay.

Additionally, I am grateful to all members of my dissertation supervisory committee for their advice and suggestions. I am especially thankful to Dr. Wim Vijverberg for all his guidance before and during the dissertation phase of my doctoral studies and with the third essay. I am grateful to Dr. Partha Deb and Dr. Michael Grossman for all their advice and support.

I am especially thankful to Takuya Hasebe for all his support in numerous ways since the first day of my doctoral studies. I have learned a lot from all my discussions with Takuya, and his thoughts and comments have been helpful in improving my research. I am also thankful to Catherine Lau for finding time to help me with editing my manuscripts, and for commenting on my work.

This dissertation has greatly benefited from my entire family. I am thankful to all of them for believing in me and for being patient with me. I am especially thankful to my father for introducing me to Economics, and to my mother for all her support, without which I would not have been successful in managing the demands of a family and a doctoral degree. I am indebted to her for encouraging my studies and for being willing to visit me in the US on multiple occasions. I am also thankful to my brother, sister and nephew for all their dedication in coordinating my access to data in Sri Lanka, and to my sister-in-law and niece for their moral support.

I am especially thankful to my twin sons for all their innocent support since infancy. They have done a tremendous task in helping me accomplish this dissertation by understanding that sometimes their 'mommy needs to study'.

Finally, I am most grateful to my husband for all his patience, understanding and support during my doctoral studies. My research has greatly benefited from his ideas, and especially from his technical infrastructure and support at home, which enabled me to include time intensive computer modeling techniques in my dissertation. I am also indebted to him for standing by me and encouraging me during the most stressful last two months of completing this dissertation.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Migration is an important function in a labor market. Migration can take place across geographic borders from one country to another, which is referred to as immigration, or within geographic borders of a country, which is referred to as internal migration.

Immigration is often associated with strict laws and regulations that prohibit unlawful presence in the host country. The presence of such stringent immigration laws is inevitably associated with a segment of immigrants choosing to unlawfully reside in the host country. These unauthorized immigrants could have entered the host country either lawfully or unlawfully. For example, some immigrants enter the host country with all required authorization, but do not leave at the end of their authorized stay. Such unauthorized immigrants are identified as overstayers. Other unauthorized immigrants unlawfully enter the host country either by crossing geographic boundaries without inspection – who are identified as Entry Without Inspection (EWI) unauthorized immigrants, or by producing forged documentation at inspection at the borders. The presence of any type of unauthorized immigrants has many implications for the host country.

The reasons for such unlawful presence are often better economic and labor market conditions in the host country compared to home country. On the other hand, in most countries internal migration is relatively less regulated: a resident of a country can choose where he would live within that country. Similar to immigration, these internal migration flows are also often instigated by disparities in labor market conditions, but here the disparities are between regions of the same country.

This dissertation focuses on unauthorized immigration and internal migration, in two different

geographic areas of the world. Specifically, in Chapter 2 and 3, I focus on unauthorized immigration to the US, and in Chapter 4, I focus on internal migration in Sri Lanka.

In the US unauthorized immigration is an important concern of the policy makers due to the impact of such unauthorized immigration on the labor market and fiscal costs, to name a few. Nonetheless, the hidden nature of unauthorized immigrants results in the absence of data on this population for extensive analysis. In this context, Chapter 2 makes an important contribution to the analysis of unauthorized immigration by developing a methodology to identify unauthorized immigrants in large cross-sectional datasets of US, which would pave the path for rigorous analysis.

This methodology shifts away from the existing "residual method", which is based on aggregate estimates, and produces an algorithm to identify unauthorized immigrants using cross sectional data. The development of this new methodology is based on three data sources – the Legalized Population Survey (LPS), Immigration and Naturalization Services (INS) and the American Community Survey (ACS). Using these three data sources I develop a legal status prediction model that is valid for 2011, and I estimate that there were 7,700,869 adult unauthorized immigrants in the US on January 1, 2011.

In addition to the methodological development of this approach, Chapter 2 includes examples on the application of the new methodology, and includes a detailed description of unauthorized immigrants present in the United States on January 1, 2011. Moreover, this chapter also includes an in-depth descriptive analysis of children of unauthorized immigrants. However, the potential of the microdata generated from this methodology is not limited to descriptive statistics.

The descriptive analysis in Chapter 2 shows that there are over 6 million children of unauthorized immigrants. However, the near absence of any econometric analysis of this child population is appalling. In order to address this vacuum, the broad analysis of children of unauthorized immigrants started in Chapter 2 is developed further in Chapter 3, to an econometric analysis of the impact of parents' legal status on the high school drop out probability of a child. Chapter 3 uses data from the ACS 2010 and focuses on children in the ages of 16 to 18 years whose parents are non-naturalized citizens. Non-naturalized citizens include: legal permanent residents, legally residing non-immigrants, and unauthorized immigrants. As such, this analysis uses the methodology developed in Chapter 2 to predict parents' legal status in the US, and a child's high school drop out probability is predicted with a linear probability model. These high school dropout probabilities

are estimated on multiple sub-samples of the data divided as per the types of parents – both highly legal, both highly unauthorized, and mixed.

The findings of the study show that, when controlling for other covariates, a child of two unauthorized immigrant is less likely to drop out of high school than a similar child of two authorized immigrant parents, and among two unauthorized immigrant parents, a non-citizen child is more likely to drop out of high school than a similar native-born child. Additionally, the study also finds that in mixed families a non-naturalized citizen child of an unauthorized immigrant mother is more likely to drop out of high school than an otherwise similar native-born child, and among both types of children (native and non-naturalized citizen), the parent combination of highly likely unauthorized and highly legal results in a higher high school drop out probability than the parent combination of highly unauthorized and unsure legal immigrants.

The chapter discusses the relevance and plausibility of the findings as well as their policy implications in the context of the proposed Development, Relief and Education for Alien Minors (DREAM) Act. The DREAM Act proposes conditional permanent residency for unauthorized immigrants of good moral character, who graduate from US high schools, arrived in the US as minors, and lived in the country continuously for at least five years prior to the enactment of the bill.

In Chapters 2 and 3, migration flows are analyzed from the point of view of labor economics. In contrast, Chapter 4 views migration in Sri Lanka from the point of view of economic development. Sri Lanka has performed favorably in reducing the significance of agriculture both as a component of GDP and of employment. Nonetheless, changes in the location of economic activity and living quarters brought about by internal migration has not caught up with these two aspects of structural transformation. Despite the need to focus on internal migration, the focus of labor migration in Sri Lanka is polarized on international migration. This polarization is evident by the absence of any reference to internal migration in the most recently launched 'National Labour Migration Policy for Sri Lanka' (Ministry for Foreign Employment Promotion and Welfare, 2008). In this context, the analysis in Chapter 4 makes an important contribution to growth and development literature on Sr Lanka by providing a rigorous analysis of internal migration. Among the many unexplored research questions in terms of internal migration in Sri Lanka, I focus on exploring the factors in sending ares that impact internal labor migration.

Specifically, in Chapter 4, I explore internal migration in Sri Lanka, to identify contextual determinants of internal labor migration. The implications of contextual variables are important from a policy perspective. Policy makers often focus on contextual variables such as the labor force and income inequality to improve the conditions in different areas in the country. This study shows the implications of policy changes on the internal migration front. This analysis focuses on district level disaggregation of contextual variables using data from the Consumer Finance and Socioeconomic Survey conducted by the Central Bank of Sri Lanka. The analysis finds that residents of rural areas tend to migrate in order to overcome the unfavorable socioeconomic conditions in these areas, districts lagging behind in structural transformation are more likely to experience large outflows of internal migrants, and the availability of agglomeration production externalities in districts with large labor forces will result in smaller outflows of internal migrants, while districts with a larger share of population between 19-34 years of age would experience large outflows of internal migrants. The analysis shows that these findings are applicable in the post conflict growth and development policy formulation framework in Sri Lanka.

Chapter 2

A Microdata-Based Approach to Identify Unauthorized Immigrants in the US

2.1 Motivation

"illegal immigration costs U.S. taxpayers
about $113 billion a year at the federal,
state and local level."

Martin & Ruark (2010/2011)

The current debates on unauthorized immigrants range from the "DREAM Act"[1] which proposes to provide conditional permanent resident status to certain unauthorized aliens, to passing of stringent laws against unauthorized immigrants in some states. These debates are fueled by two perspectives on unauthorized immigration. One postulates that unauthorized immigration increases the domestic supply of low-skilled labor and puts downward pressure on US wages, which results in redistributing income from low-skilled native workers to employers and thereby creating a net gain in national income by allowing employers to use their resources more productively. This

[1]Development, Relief and Education for Alien Minors Act. This act proposes conditional permanent residency to unauthorized immigrants of good moral character who graduate from US high schools, arrived in the US as minors, and lived in the country continuously for at least five years prior to the enactment of the bill.

is referred to as the immigration surplus. The other is the net fiscal impact - which is the difference between the various local, state, and federal levels taxes paid by unauthorized immigrants and amount of government expenditure allocated to unauthorized immigrants. Existing estimates show that immigration surplus is essentially set off by the net fiscal impact leaving a near zero impact on the US national accounts (Hanson, 2009). As such, those focusing on the near zero overall impact of unauthorized immigration on the US national accounts are somewhat indifferent to the presence of unauthorized immigrants. Those focusing on certain negative aspects such as the losses to native low skilled workers, or increases in government expenditure are against unauthorized immigration, while those focusing on positive implications such as productivity gains, ability of unauthorized immigrant labor force to better respond to market conditions and increase in government revenue are sympathetic to unauthorized immigration. Thus, the issue about unauthorized immigration has remained controversial for many years.

In order to streamline polices on unauthorized immigration, the US needs in-depth analysis of unauthorized immigrants both from micro and macro economic perspectives. However, the existing data sources and methodologies are primarily macroeconomic analyses. Two institutions that regularly produce analyses and estimates of unauthorized immigrants are the Department of Homeland Security (DHS) and the Pew Hispanic Center (PHC). Both these institutions derive their data and estimates of unauthorized immigrants using the residual methodology, which is an aggregate level method where 'the final estimates are the result of subtracting one set of data from another' (Warren & Passel, 1987). For 2010 the DHS estimates 10.8 million unauthorized immigrants while the PHC estimates 11.2 million within a range of +/- 0.5 million (Hoeffer *et al.*, 2011; Passel & Cohn, 2011). However, as highlighted by Warren & Passel (1987) such [residual] techniques can obviously be sensitive to errors in underlying figures. Moreover, disaggregating residual estimates by characteristics of unauthorized immigrants is tedious, since the entire estimation process has to be repeated for each characteristic (Passel & Cohn, 2011).

Against this backdrop, this paper develops an alternative to the existing methods of analyzing and estimating the number of unauthorized immigrants by predicting the probability that a foreign born person is an unauthorized immigrant, using recent large cross-sectional datasets. Specifically, I predict unauthorized immigrants in the American Community Survey (ACS) 2010, based on a prediction model developed with data on a sample of *actual* authorized and unauthorized immi-

grants during the Immigration Reform Control Act (IRCA) 1986, which is subsequently adjusted for calibration and time dynamics between 1986 and 2010. This methodology can be outlined as follows.

i. Develop a "derivation" data set by appending unauthorized immigrant data from Legalized Population Survey (LPS) with authorized immigrant data from Immigration and Naturalization Services (INS) database.

ii. Estimate a model to predict legal status using derivation data, defined as the "derivation" model.

iii. Adjust the derivation model for (a) extrapolation to the ACS 2010; (b) time dynamics between 1986 and 2010; and (c) for the inclusion of new variables.

iv. Use the adjusted model to predict the probability of being legal for foreign born persons enumerated in the ACS 2010.

v. Sum the product of each individual's predicted probability to be unauthorized and the number of persons represented by each individual to estimate the aggregate number of unauthorized immigrants.

vi. Determine a cutoff probability, and identifying those below the cutoff probability as predicted unauthorized immigrants. The resulting microdata on unauthorized immigrants provide individual level data on this population.

Based on this methodology, I estimate that there were 7,700,869 adult unauthorized immigrants in the US on January 1, 2011. The paper makes seven significant contributions to the methodology of estimating unauthorized immigrants. First, unlike the existing estimates, this methodology predicts the legal status of an individual based on a model estimated for actual unauthorized immigrants (from LPS) and actual authorized immigrants (from the INS). All existing estimates either do not base their estimates on actual unauthorized immigrants, or if they do, they are not supplemented with concurrent actual authorized immigrants.[2] Second, contrary to existing estimates,

[2]Passel (2007) employs a similar algorithm in the context of residual methodology, to arrive at a "more detailed tabulations of family, social and economic characteristic of unauthorized immigrants". Here "Individuals that are definitely authorized and those that are ***potentially*** unauthorized are identified in the CPS ...' A notable limitation

which assume that the characteristics of unauthorized immigrants have remained constant since 1986, the prediction model developed here is adjusted for changes in the effect of each predictor on legal status of immigrants. This methodology also adjusts for the validity of the prediction model for extrapolation on the new dataset (ACS 2010). The third contribution of this methodology is the use of available information on legal status in the ACS to arrive at a more recent data set with partially observed outcomes of legal status of foreign born persons, to adjust the derivation model. Fourth, instead of a macro level adjustment for the undercount of unauthorized immigrants in the ACS at the end of the estimation process, this methodology performs micro level adjustments prior to arriving at the final estimates. Fifth, this methodology has the flexibility to be performed on ACS, CPS and the decinial census of any year, and sixth, facilitating further research this methodology produces a large cross sectional dataset of unauthorized immigrants. Finally, this methodology shows the possibility for cross fertilization between medical statistics and econometrics by borrowing adjustment techniques developed in the context of medical science.

The remainder of this paper is organized as follows. Section 2.2 reviews the residual methodology adopted by the DHS and the PHC. Section 2.3 introduces the three data sources used in the development of this methodology as well as develop the derivation dataset by merging LPS data with INS data. Section 2.4 is devoted to the development of the new microdata-based methodology. As such, Section 2.4.1 estimates the derivation model, Section 2.4.2 shows the need for adjusting the derivation model and Section 2.4.3 introduces the methodology to derive new data for the subsequent adjustment procedure. The adjustments for calibration and time dynamics is performed in Section 2.4.4, and the final adjusted coefficients are presented in Section 2.4.5. The microdata-based methodology is capable of producing two types of aggregate estimates – sample weight based estimates and assignment based estimates. Section 2.4.6 describes the procedure to obtain sample weight based estimates, while Section 2.4.7 shows the the procedure to obtain assignment based estimates as well as the cross-sectional data of this population. Section 2.7 presents the characteristics of unauthorized immigrants followed by the characteristics of their children in Section 2.8. The discussion of the policy relevance is presented in Section 2.9, followed by Section 2.10 with final remarks and future directions of the methodology.

of this version is the use of 'potentially' unauthorized immigrants as the basis for estimation. For, given that the objective of the exercise is to estimate characteristics of unauthorized immigrants, the certainty of the initial sample being unauthorized immigrants is critical.

2.2 Review of existing methodologies

2.2.1 Residual method by the Department of Homeland Security (DHS)

The DHS estimated that there were 10.79 million unauthorized immigrants in the US on January 1, 2010 based on their version of the residual method (Hoeffer *et al.*, 2011). This version of the residual method requires the estimation of the following two populations to arrive at the unauthorized population estimate: (i) the total-foreign born population living in the US on January 1, 2010 , and (ii) the legally resident population on the same date. The unauthorized population is then found by subtracting the legally resident population from the total-foreign born population.

To estimate the total-foreign born population living in the United States on January 1, 2010, the total foreign-born population that entered between 1980-2009 is estimated based on ACS data. The year 1980 serves as the starting year for this estimate on the assumption that "foreign-born residents who had entered the United States prior to 1980 were authorized residents since most were eligible for authorized permanent resident status", either under the Registry Provision of the Immigration and Nationality Act (INA) of 1965 or the Immigration Reform and Control Act (IRCA) of 1986. This foreign born population represented by the ACS is then adjusted for three sources of undercount, namely (i) foreign born non-immigrants; (ii) Lawful Permanent Residents[3] (LPRs), refugees, and asylees; and (iii) unauthorized immigrants. Then it is adjusted for the shifts in the dates of enumeration of ACS across different years. The estimated legally resident immigrant population on January 1, 2010 is arrived at by adding the emigration and mortality adjusted LPR, refugee, and asylee flow that entered between 1980-2009 to Non immigrant population on January 1, 2010.

Hence, data on the legally resident population comes from various administrative sources. For instance, data on Legal Permanent Residents (LPR) is sourced from US Citizenship and Immigration Services (USCIS), refugees from Department of State, persons granted asylum from USCIS and Executive Office of Immigration Review of the Department of Justice, and non-immigrants from TECS system of the US Customs and Border Protection.

In order to explore the characteristics of unauthorized immigrants, "estimates of the unauthorized population were generated for the ten leading countries of birth and states of residence, age,

[3]Also referred to as Legal Permanent Residents.

and gender". This is a significant limitation of the residual method, where each characteristic has to be estimated separately for the entire dataset, and replicated several times to arrive at several characteristics. Moreover, the necessity to gather data from many administrative sources limits the usage of this version by researchers without access to administrative data. The proposed new methodology has the capacity to overcome both the above limitations.

Additionally, this variation of residual methodology suffers from the following limitations as noted in Hoeffer *et al.* (2011): (i) the estimates are sensitive to the assumptions about undercount of the foreign-born population in the ACS and rates of emigration; (ii) limited validity and reliability of Census survey data on the year of entry question, "When did this person come to live in the United States?"; (iii) errors in converting DHS administrative dates for authorized resident immigrants to year of entry dates; (iv) the non-immigrant population estimate are based on admission dates and length of visit by class of admission and not actual population counts;[4] (v) the 2009 ACS data are based on a sample of the US population. Thus the estimates of the total foreign-born population that moved to the United States in the 1980-2009 period are subject to sampling variability; and (vi) the state of residence for legally resident 1980-2009 entrants is assumed to be the state of residence on the date the most recent status (e.g., refugee, LPR, or naturalized citizen) was obtained; however, the accuracy of the estimates may be affected by state-to-state migration that occurred between the date of the status change and January 1, 2010.

2.2.2 Residual method by the Pew Hispanic Center (PHC)

The PHC estimated that 11.2 million unauthorized immigrants were in the US on March 2010 (Passel & Cohn, 2011), which is 3.7 percent higher than the DHS estimate. All recent reports by the PHC (Passel, 2005a,b, 2006; Passel & Cohn, 2008, 2009, 2010) adopt a residual methodology originated by Passel & Clark (1998), which is slightly different from that of the DHS. The PHC version of the residual method relies on Current Population Survey (CPS) data. Similar to the DHS version, the number of unauthorized immigrants included in the survey is achieved by subtracting of the number of legally residing immigrants from the total number of immigrants in the survey. The authorized resident immigrant population is estimated by applying demographic methods to

[4]Length of visit, which is calculated by matching arrival and departure records, is subject to more error than admissions data

counts of authorized admissions covering the period from 1980 to the present, which are obtained from the Department of Homeland Security's Office of Immigration Statistics. In terms of the characteristic estimates of unauthorized immigrants, the PHC version of the residual method is more flexible than the DHS version for it is based on a derived dataset of assigned unauthorized immigrants, and does not require repeated calculations, as in the DHS method. In order to arrive at the characteristics of unauthorized immigrants, the PHC methodology proceeds as follows.

First, individual foreign-born respondents in the survey are assigned either an authorized or unauthorized immigrant status, based on the individual's demographic, social, economic, geographic and family characteristics. The exact methodology for assignment of individual "legal status" proceeds in the following way. First, immigrants entering the United States prior to 1980 are assigned a status of "authorized" immigrant. Then, the CPS data are corrected for known over-reporting of naturalized citizenship for recently arrived immigrants and all remaining naturalized citizens are assigned as authorized, except for those from Mexico and Central American countries. Second, those entering the US as refugees are identified on the basis of country of birth and year of immigration to align with known admissions of refugees and asylees. Similarly, individuals holding certain kinds of temporary visas (including students, diplomats and high-tech guest workers) are identified in the survey, and each is assigned authorized temporary migrant status using information on country of birth, date of entry, occupation, education and certain family characteristics. Additionally, some individuals are assigned as legal immigrants because they are in certain occupations (e.g., police officer, lawyer, military occupation, federal job) that require legal status or because they are receiving public benefits that are limited to authorized immigrants. The completion of these initial assignments as "definitely legal" immigrants, leaves a pool of "potentially unauthorized" immigrants, which is felt to typically overestimate the target residual estimates by 20-35 percent. The third step assigns these "potentially" unauthorized immigrants an authorized or unauthorized immigrant status. In my opinion this is the critical step of the methodology because all subsequent estimates are based on this assignment. Hence, it requires a clear explanation for the validity of the estimates. A noteworthy fact here is that Passel & Cohn (2011) cite Passel & Bean (2004) for the methodology and in turn Passel & Bean (2004) refer back to Passel & Clark (1998). As such, the assignment of unauthorized immigrant status in Passel & Cohn (2011) is based on Passel & Clark (1998), which focused on the state of New York since they used LPS data on the occupa-

tional structure of amnesty applicants who lived in New York State at the time they legalized to assign legal status to aliens in the 1995 March CPS. Their procedure assigns a probability of being unauthorized to individuals in the CPS on the basis of three variables – age, sex, and occupation. It is unclear if the assignment mechanism is extended to include all states or if it is based only on characteristics of New Yorkers. Subsequent to assigning unauthorized immigrant status this methodology returns to the main steps.

The second step adjusts the estimates of authorized and unauthorized immigrants counted in the survey for omissions. The basic information on coverage is drawn principally from comparisons with Mexican data, US mortality data and specialized surveys conducted at the time of the 2000 Census (Bean & Woodrow-Lafield, 1998; Capps *et al.*, 2002; Marcelli & Ong, 2002). These adjustments increase the estimate of the authorized foreign-born population, generally by 1-3 percent and the unauthorized immigrant population by 10-15 per cent. Finally, the individual survey weights are adjusted to account for immigrants missing from the survey.

The methodology adopted by the PHC can be criticized for four notable limitations in the assignment of authorized versus unauthorized immigrant status. First, in order to assign a probability, the outcome variable should be a dichotomous (authorized versus unauthorized) variable. But the authors do not explain their data source for observed authorized immigrants in the outcome variable. Second, this assignment mechanism is based on only three variables, none of which is country of origin. As will be discussed in subsequent sections, my analyzes show that 82 percent of unauthorized immigrants can be accurately assigned as unauthorized immigrants with country of birth as the only predictor. Third, Passel & Clark (1998) assume that between 1987-88–the years in which IRCA aliens applied to become legal US residents, and the mid-1990s, the occupational structure of illegal aliens did not change significantly. Moreover, estimates for subsequent years have extended this assumption upto 2009. This is too restrictive and assumes away all time dynamics. My methodology tests this assumption in Section 2.4.4 and finds that there were changes in the effect of occupational structure in the probability of being an unauthorized alien. Finally, the assignment of legal status to the pool of unauthorized immigrants is unconvincing, for Passel and Clark (1998) repeats this process until assigned number of unauthorized immigrants matches the residual estimate arrived separately. Moreover, as identified in Hoeffer *et al.* (2011) the March CPS has a smaller sample size (100,000 households) compared to the ACS (3 million households),

which has an impact on aggregate level estimates.

In this context, the microdata-based methodology I develop is different from both the existing methodologies. First, the microdata-based methodology is based on parameter estimates produced by cross-sectional models rather than relying on the difference between two aggregate level estimates. Second, unlike the PHC methodology, I rely on the larger of the two national data sources with foreign born persons (the ACS), and unlike the DHS method I do not use restricted administrative data along with the ACS data. My methodology only uses publicly-available data sources, which allows the application of this methodology by those without access to administrative data. Third, unlike the DHS version of the residual method which calls for repeated calculations for each characteristic, the my methodology aligns more with the PHC method with greater flexibility, but the new methodology overcomes limitations in the PHC method by : developing a model based on observed authorized and unauthorized immigrants; assigning predicted legal status based on a larger number of control variables; relaxing the assumption of constant occupational structure of unauthorized immigrants; and the new methodology does not dictate that number of foreign born persons assigned a predicted legal status should match a predetermined number. Fourth, the proposed methodology facilitates rigorous econometric analysis of unauthorized immigrants beyond the descriptive analysis accomplished by using the residual method.

2.3 Data

The proposed methodology employs three distinct data sources: the Legalized Population Survey (LPS) of 1987/88, the Immigration and Naturalization Services (INS) data from 1972 to 1986, and the American Community Survey (ACS) of 2010.

2.3.1 Legalized Population Survey (LPS)

The base data for this study is from the LPS, which contains data of unauthorized immigrants who were granted amnesty under the IRCA 1986. The IRCA consisted of two programs - Legally Authorized Workers (LAW) and Seasonal Agriculture Worker (SAW). The LPS consists of a nationally representative sample of 6,193 out of 1,491,897 applicants under the LAW component–the larger of the two programs– who were continuous unauthorized residents in the US since the end

of 1981 and were successful in their application for LPR status in the United States.[5] Survey data were collected from the entire group in 1989 and the data used in this paper are those as of 1986 to reflect their characteristics prior to becoming legalized immigrants, such as the occupation as an unauthorized immigrant.

However, literature criticizes the LPS data for the alleged inclusion of fraudulent amnesty applicants, and the exclusion of short-term and seasonal migrants (Borjas & Tienda, 1993; Orrenius & Zavodny, 2003). Moreover, as noted by literature approximately 6 percent of applications were rejected and such rejected applicants are not reflected in the LPS (Cooper & ONeil, 2005). Given that the application for the amnesty had the possibility of being rejected, some unauthorized immigrants would have selected not to apply due to the fear of being denied and subsequently having to leave the US after being exposed as an unauthorized immigrant. The LPS is not representative of this segment of unauthorized immigrants as well.

Nonetheless, as succinctly noted by Kossoudji & Cobb-Clark (2002, p. 602) despite all these limitations "legalized population [survey] represents an important component of the unauthorized population, and its members' experiences while unauthorized, although not representative of all unauthorized workers, $\cdots$". Moreover, given that the LPS is the only data source with a representative sample *ever* of the unauthorized immigrants population in the US that originated from a broad cross section of countries, I use the LPS data despite its limitations.

2.3.2 Immigration and Naturalization Services (INS)

Data on authorized immigrants are obtained from the INS data set. The INS data capture immigrants obtaining Legal Permanent Resident (LPR) status from 1972 to 2000 under two types - Adjustment of status and New Admissions. Those adjusting their status to LPRs had been in the US as non-immigrants, while those entering as new admissions have not been residing in US immediately prior to becoming an LPR. I use INS data only up to 1986 in order to keep LPS and INS data comparable.

The overlapping years in the LPS and the INS data enable the development of a combined dataset, which I call the "derivation data", that consist of actually observed unauthorized and

[5]The other program was the Season Agriculture Worker (SAW), which granted an additional 1.3 million farm workers LPR status. However, there are no similar data for the SAW population.

authorized immigrants for the estimation of the initial probabilistic model to predict legal status.

2.3.3 American Community Survey (ACS)

The 2010 ACS is the most recent data source that gives a comprehensive insight into the US population. Data for ACS 2010 was collected between January 1, 2010 and December 31, 2010 and ACS data of a given year are considered to reflect the US population for January 1 of the following year (Hoeffer *et al.*, 2011). This paper uses the 1 percent Public Use Microdata Sample (PUMS) of the 2010 ACS, made available by Ruggles *et al.* (2010), for two purposes. For the purpose of predicting legal status, a sub-sample of the ACS is derived as follows. First, following the tradition of existing estimates, all those who entered the US prior to 1980 are excluded on the assumption that IRCA and the INA eliminated all unauthorized immigrants up to 1980. Second, the sample excludes natives persons born in Puerto Rico, Guam, US Virgin Islands, and Northern Marianas and persons born abroad to American parents, for all such persons are US citizens "by birth". Additionally, foreign-born persons who are already "naturalized citizens" and specific "not-naturalized citizens" are excluded for prediction purpose, since their legal status in US is not questionable. These "not-naturalized citizens" excluded here are either employed in occupations in which one must be an authorized immigrant, or are receiving certain public benefits for which one must be an authorized immigrants (see Section 2.4.1 for details). Hence, the sub-sample used for predicting of legal status consists of "non-naturalized citizens"over the age of 18 years who entered the US after 1979 with ambiguous legal status reported in the ACS. This group includes LPRs, legally residing Non-Immigrants (NIM's) such as temporary workers, students, extended stay business visitors, refugees and asylees, as well as the much sought-after unauthorized immigrants[6].

For the intermediate step of adjusting the prediction model, all foreign born persons who entered the US after 1979 (both citizen and not naturalized citizens) are used.

2.3.4 Derivation data

In order to predict legal status of foreign born persons the initial probabilistic model needs to represent all components of the foreign born persons with uncertain legal status in the ACS 2010. The

[6]Despite the certainty of legal status of LPRs, they are included in the sub-sample for prediction of legal status because the ACS data does not distinguish LPR holders among the "not citizen" foreign born persons.

LPS represents unauthorized immigrants and the INS data is used to reflect the authorized immigrants. Despite the fact that INS data does not directly represent NIM status holding authorized immigrants, it can be used to approximate this population, as follows.

The LPRs in the INS who are adjusting their status from NIM are already in the US, and they have remained in the US by lawfully extending their NIM status. As such, the "Adjusting LPRs" in the INS are a subset of the larger population of NIMs extending their status. In this setting, I make the following assumption.

Assumption 1 *: All extending NIM status holders ("extending NIMs") can become adjusting LPRs.*

With Assumption 1 the INS adjustees can be considered to represent "extending NIMs". In the case of newly entering LPRs, it is apparent that they did not enter the US unlawfully. Moreover, given the scrutiny that the new entrant LPRs undergo is more intense than the scrutiny a new entrant NIM would undergo, I make Assumption 2 as follows.

Assumption 2 *: New entrant LPRs are a subset of new entrant NIMs.*

With Assumption 2, I consider new entrant LPRs to represent new entrant NIMs. In the absence of self-selection into LPR as opposed to NIMs, Assumptions 1 and 2 shows that LPRs are representative of NIMs. Anecdotal evidence indicates that majority of immigrants who enter the US as a NIM eventually opt to remain in the US to become a LPR. This shows that the self-selection into LPR is not very different from self-selection into NIM. The only notable difference between Adjusting LPRs and NIMs is the longer duration in the US of the former. However, a longer duration in US does not affect the probability of a foreign born person being an authorized immigrant as opposed to an unauthorized immigrant.

In this context, in the absence of an alternative source of data on authorized immigrants without the risk of being contaminated by unauthorized immigrants claiming to be authorized, the INS is the best source for the requirement at hand. Moreover, the difference between adjusting LPRs and NIMs in terms of duration of stay in US does not have an impact on being authorized because if one overstays and becomes an unauthorized immigrant he cannot reclaim authorized NIM status until a 10 year period is lapsed since unauthorized presence in US.

The INS data in the derivation data set therefore represents LPRs and NIMs, and appending the LPS data to the INS data ensures that the derivation dataset represents both authorized and unauthorized immigrants.

2.4 Methodology

2.4.1 Estimation of the derivation model

The derivation dataset (LPS+INS) is used to estimate a discrete choice model (a logit model) to predict the probability of observations in the derivation dataset to be authorized immigrants[7] (follow the flow diagram in the Appendix 2.C for clarity at each step). The dependent variable – L (legal), is defined as $L = 1$ if an observation originated from the INS dataset and $L = 0$ if an observation originated from the LPS dataset.

$$L = \begin{cases} 1 & \text{if INS observation} \\ 0 & \text{if LPS observation} \end{cases} \tag{2.1}$$

Observations originating from the INS are weighted as 1 since INS data are the population of legal immigrants, and the observations originating from LPS data are weighted accordingly to represent 1,442,122 unauthorized immigrants represented in the LPS. The dependent variable is regressed on a vector of six explanatory variables X, using a weighted logit model as depicted in Equation (2.2).

$$P(L = 1) = \Lambda(\alpha_{derivation} + X'\beta_{derivation}) \tag{2.2}$$

where $\Lambda(z)$ is a logistic function: $\exp(z)$ / $(1+\exp(z))$. The determinants of legal status are limited to six characteristics–country of origin, age, occupation, marital status, sex, and state of residency. This is necessary because of the limited scope of variables captured in INS data. However, given that many of these variables are categorical and are split up into series of dummy variables, tis model consists of 131 explanatory variables. This model is referred to as the derivation model. Table 2.1 depicts the parameter estimates of the derivation model.

[7]The logit model is estimated on all observations with overlapping states of residence in INS and LPS.

Variable	Coeff.	S.E.	Variable	Coeff.	S.E
Age	0.054	0.002	**Country of birth**		
Female	-0.27	0.035	Algeria	0.684	0.834
Married	-0.019	0.042	Antigua-Barbuda	-0.666	0.781
Widowed	-1.827	0.155	Argentina	0.113	0.701
Divorced/separated	-2.354	0.094	Australia	1.011	0.787
State of residence			Bahamas	-0.524	0.758
Arizona	-0.644	1.138	Bangladesh	-0.661	0.759
Arkansas	0.312	1.473	Barbados	2.439	0.993
California	-1.187	1.132	Belize	-1.225	0.693
Colorado	2.839	1.489	Bolivia	-0.279	0.743
Connecticut	3.445	1.519	Brazil	0.108	0.731
D.C.	0.021	1.154	Canada	0.168	0.682
Florida	-0.858	1.135	Chile	-0.072	0.702
Georgia	1.425	1.456	China	1.476	0.687
Illinois	-0.892	1.133	Colombia	-0.548	0.675
Indiana	1.084	1.291	Costa Rica	0.180	0.732
Iowa	1.768	a	Cuba	4.040	0.814
Louisiana	-1.601	1.141	Cyprus	0.593	0.952
Maryland	-0.842	1.139	Dominican Republic	0.389	0.679
Massachusetts	3.666	1.515	Ecuador	-0.339	0.681
Mississippi	-0.991	1.271	Egypt	1.033	0.763
Nevada	2.902	1.564	El Salvador	-2.246	0.671
New Jersey	-0.209	1.135	Ethiopia	-1.549	0.707
New Mexico	-2.15	1.138	Fiji	2.771	1.253
New York	-0.455	1.133	France	1.071	0.775
North Carolina	2.222	1.381	Germany	1.910	0.768
Oklahoma	-2.403	1.141	Ghana	-1.466	0.695
Oregon	2.869	1.543	Greece	2.382	0.809
Rhode Island	1.913	1.572	Grenada	0.660	0.816
Texas	-0.437	1.132	Guatemala	-1.707	0.673
Virginia	-0.661	1.142	Guyana	0.969	0.705
Washington	-1.59	1.139	Haiti	-0.387	0.679

Notes:
(a): Standard error for Iowa is not estimated because except for one observation, all others came from the INS dataset. Nonetheless the indicator for Iowa is retained in the model because this single observation from the LPS represents 128 unauthorized immigrants.

Table 2.1: Derivation logit model

Variable	Coeff.	S.E	Variable	Coeff.	S.E
Honduras	-0.960	0.681	Paraguay	0.385	1.382
Hong Kong	1.209	0.722	Peru	-0.444	0.681
Hungary	2.585	1.243	Philippines	1.176	0.676
Iceland	0.111	1.209	Poland	-0.781	0.679
India	1.486	0.697	Portugal	1.996	0.788
Indonesia	0.548	0.838	Senegal	-2.365	1.087
Iran	-0.451	0.678	Sierra Leone	-1.997	0.752
Iraq	1.373	0.833	Singapore	0.946	1.348
Ireland	0.857	0.797	Somalia	0.101	0.929
Israel	1.173	0.745	South Africa	0.886	0.858
Italy	2.131	0.762	Spain	1.209	0.790
Jamaica	0.785	0.681	Sri Lanka	0.877	1.196
Japan	1.529	0.750	St. Lucia	0.270	0.940
Jordan	1.228	0.769	St. Vincent-Grenadines	1.224	0.880
Kenya	0.047	0.790	Sweden	1.224	0.976
Korea	2.055	0.700	Switzerland	2.129	1.028
Lebanon	0.976	0.741	Syria	0.815	0.812
Liberia	-1.086	0.754	Taiwan	0.872	0.737
Malaysia	0.395	0.886	Thailand	0.351	0.690
Mali	-2.126	1.057	Tonga	-0.555	0.779
Mexico	-1.763	0.670	Trinidad-Tobago	0.870	0.703
Morocco	1.662	1.301	Turkey	0.795	0.788
Netherlands	0.963	0.810	Uganda	-0.168	0.808
New Zealand	1.753	1.295	United Kingdom	0.793	0.685
Nicaragua	-1.089	0.683	Uruguay	-0.429	0.746
Nigeria	-1.587	0.688	Venezuela	-0.475	0.741
Norway	1.270	1.206	Vietnam	4.245	0.838
Pakistan	-0.227	0.689	Western Samoa	-1.677	0.767
Panama	0.918	0.735	Yugoslavia	0.940	0.725

Table 2.2: Derivation logit model (cont.)

Variable	Coeff.	S.E.
Occupations		
Lawyers, medical doctors, engineers, –surveyors, scientists & urban planners	1.420	0.156
Other health hssessment, treating, –health technologists and technicians	0.274	0.291
Teachers, postsecondary, secondary, – councellors, librarians, archivists, & curators	1.604	0.275
Social, rec., and religious workers	1.433	0.485
Writers, artists, entertainers, and athletes	0.456	0.186
Technologists & technicians, except health	0.659	0.174
Sales occupations	-0.493	0.120
Administrative support occupations	0.223	0.095
Service occupations	-0.565	0.080
Farming, forestry,& fishing occupations	0.538	0.117
Precision production, craft, repair occupations –operators, fabricators, and laborers	-0.076	0.077
Homemakers ,unemployed or retired etc.	1.924	0.084
Constant	-0.067	1.316
Log pseudo likelihood		-1086697.1
Number of observations		3,447,126
Pseudo R square		0.4806

Table 2.3: Derivation logit model (cont.)

As shown in Table 2.4 the derivation model includes 3,440,956 authorized immigrants, which is the population of LPRs and 6,170 unauthorized immigrants representing 1,442,122 unauthorized immigrants. When the cut off point for assigning predicted unauthorized immigrant status is arbitrarily set at the probability of 0.5, the accurate prediction rate for unauthorized immigrants is 75 percent,[8] while for authorized immigrants the rate of accurate predictions is 90 percent.

In addition to estimating the prediction model with all 6 variables, each variable is considered one at a time to discern which has the greatest predictive power to determine legal status. Interestingly, this step revealed that country of birth alone is capable of predicting 82 percent of the actual unauthorized immigrants as predicted unauthorized immigrants (82 percent is the weighted

[8]The 75 percent rate of accuracy in prediction is arrived using person weights. Based on the 4,241 observations predicted as unauthorized immigrants the corresponding rate is 69 percent.

rate, while 71 percent is the unweighted rate). This proves that country of birth is an important determinant of legal status of foreign born persons which ought to be an independent variable on any model to predict the legal status. As noted before, the methodology employed by Passel & Cohn (2011) fails to include country of birth in their mechanism to assign unauthorized immigrant status, nor do they provide a reason for the omission. Hence, the microdata-based methodology developed here overcomes the omitted variable bias in the PHC version of the residual method by including country of origin as an independent variable.

Based on the derivation model a naive estimate of unauthorized immigrants for January 2011 (based on the ACS 2010) is 6,413,914. This naive estimate is arrived by first eliminating not-naturalized immigrants who are very likely to be authorized immigrants from the pool of potential unauthorized immigrants. This elimination assumes that those employed in US government jobs, those receiving social security income or welfare income, those subscribing to government supported health insurance plans and those who are veterans, are authorized immigrants in the US.[9] Second, following the tradition of existing methodologies (Hoeffer *et al.*, 2011), the person weights of observations with a predicted probability less than 0.5 is multiplied by 1.1 to account for undercount of unauthorized immigrants in the ACS. Here the cutoff probability of 0.5 is coincidentally the same as the cutoff probability for predicting legal versus unauthorized immigrant status. Finally, the naive estimate is arrived by summing up these adjusted person weights.

Hence, this naive estimate does not have any adjustments for calibration nor for time dynamics, but includes an adjustment for unrepresented states in the derivation model, since the derivation model is estimated on 25 overlapping states and D.C. This adjustment account for the remaining state coefficients as follows. First, the 50 states and D.C. are divided in to 9 census divisions following the Census Bureau definition. Second, the average of the state coefficients of each division estimated by the derivation model is assigned as the state coefficient for each missing states in the same census division[10].

[9]The government jobs considered here are postal service jobs, legislatures, judges, military and police/detective jobs. The 1990 occupation codes in ACS 2010 for these jobs are 003, 016, 179, 354, 423 and 905. The public insurance plans are Medicare, Medicaid, Veterans Affairs insurance and Tricare.

[10]1) New England : Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island and Vermont. 2) Mid Atlantic : New Jersey, New York and Pennsylvania. 3) East North Central : Illinois, Indiana, Michigan, Ohio and Wisconsin. 4) West North Central : Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota and South Dakota. 5) South Atlantic : Delaware, D.C., Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia and West Virginia. T 6) East South Central : Alabama, Kentucky, Mississippi and Tennessee. 7) West South Central : Arkansas, Louisiana, Oklahoma and Texas. 8) Mountain : Arizona, Colorado, Idaho, Montana, Nevada, New Mexico,

Legal immigrants (represents 3,440,956)	3,440,956
Unauthorized immigrant (represents 1,442,122)	6,170
Cut off probability	0.5
Observed authorized and predicted authorized	3,103,749
Percentage of accurate predicted authorized	90 %
Observed unauthorized predicted unauthorized (weighted)	$1,076,855^a$
Percentage of accurate predicted unauthorized (weighted)	$75\%^b$

Notes:
a. unweighted number is 4,175
b. unweighted percentage is 69 %

Table 2.4: Performance of the derivation model

As elaborated in the next section, this naive estimate needs to be refined further to arrive at the final estimate of the count of unauthorized immigrants.

2.4.2 Need for adjustments to derivation model

The derivation model reflects the predictive capacity of the independent variables as observed in 1986. The prediction of legal status based on this derivation model implies that the effect of each predictor remained constant for *nearly a quarter of a century*. For instance, this implies that the percentage of Mexicans among unauthorized immigrants is the same in 1986 and 2010, and that the proportion of unauthorized immigrants in each state remained constant since 1986. Estimates based on such a restrictive assumption would be similar to the characteristics of unauthorized immigrants produced by the Pew Hispanic Center discussed in section 2.2. Moreover, the definition of legal immigrants in the derivation data and ACS data are slighly different. For both these reasons the derivation model needs to be adjusted.

However, methodologies for updating prediction models are scarce in the economic and demographic literature. Nonetheless, such techniques are often used in literature of medical statistics (Janssen *et al.*, 2008; Steyerberg *et al.*, 2004). As such, this paper adopts the adjustment techniques developed by Steyerberg *et al.* (2004) in the context of clinical statistics to improve calibration for out-of-sample predictions, and to account for time dynamics of predictors when data for out of

Utah and Wyoming. 9) Pacific : Alaska, California, Hawaii, Oregon, and Washington.

sample prediction and data used for estimation of the model are from two disparate time periods.[11] In order to carry out these adjustments, recent data on observed legal status and other relevant explanatory variables of individuals are required. However, as noted earlier there is no recent dataset that is representative of unauthorized immigrants from a broad cross-section of countries. Had there been such new data on unauthorized immigrants, there would be no necessity for adjustments, for the derivation model can be developed on such new data. In absence of new data on unauthorized immigrants, I exploit the available information on authorized immigrants in the ACS. However, the reader should bear in mind that this procedure serves only as an intermediate step which facilitates generating new data on observed authorized immigrants, and thus should not be confused with the ultimate objective of predicting legal status on ACS 2010.

The five step approach to generate data on recent unauthorized immigrants is outlined below.

2.4.3 New data on unauthorized immigrants

First, a dichotomous variable named *pseudo legal status* is generated and all naturalized citizens in the ACS 2010 are assigned a value of 1, for naturalized citizens are by default authorized immigrants. Given that naturalized citizens are the closest counterparts to legally residing non-immigrant foreign born persons in available data sources, naturalized citizens are considered to reflect the characteristics of those who were originally authorized non-immigrant status holders based on Assumption 3.

Assumption 3 *: All LPRs will be naturalized.*

Assumption 3 is an extension of Assumption 2. This extension of the previous assumption is realistic, given the similarity in characteristics of LPRs and LPRs eligible for naturalization presented by Rytina (2011).[12]

Second, as in the case of the naive estimation, certain not naturalized citizens are assigned pseudo legal status=1 based on their occupations, receipt of social security income and/or welfare income, subscription to public health insurance and veteran status.

[11] Which may result in changing the effect of predictors on the outcome variable. An empirical evaluation of these techniques are done by Janssen *et al.* (2008).

[12] It is reasonable to assume that LPRs eligible to naturalization will become naturalized. It is also assumed that there is no reverse causality from dynamics in one's legal status (from authorized non-immigrant to LPR to naturalized citizen) towards the six predictors in the derivation model, since all three immigration statuses are equally stable relative to being an unauthorized immigrant.

Third, I predict the legal status of not-naturalized citizen in ACS 2010, excluding those who are assumed to be authorized immigrants using the derivation model developed in Section 2.4.1.[13]

Fourth, non-citizen foreign born persons with a predicted probability of greater than 0.5 are assigned a pseudo legal status = 1 while those with predicted probability less than 0.5 are assigned a pseudo legal status = 0 (assignment of pseudo unauthorized immigrant). There were no 'not naturalized' foreign born persons with a predicted probability of 0.5.[14] As per the derivation model and the cutoff probability of 0.5, 66 percent of Naturalized citizens[15] in the ACS 2010 are predicted authorized with the derivation model. Those assigned pseudo authorized=1 (naturalized citizens, assumed authorized immigrants and prediction based authorized immigrants) are identified as 'enhanced authorized immigrants' from here onwards.

The assignment of pseudo legal status is summarized in Equation 2.3 below.

$$\text{Pseudo legal} = \begin{cases} 1 & \text{if naturalized citizen in ACS 2010} \\ 1 & \text{if non-naturalized citizen in ACS 2010 \& assumed legal} \\ 1 & \text{if non-naturalized citizen in ACS 2010 \& } P(L)_{naive} \geq 0.5 \\ 0 & \text{if non-naturalized citizen in ACS 2010 \& } P(L)_{naive} < 0.5 \end{cases} \tag{2.3}$$

Fifth, for observations with assigned pseudo legal=0, the associated person weights are multiplied by 1.1 to account for undercount of unauthorized immigrants in ACS, while the person weights of pseudo legal=1 observations are unchanged. This adjustment for undercount of unauthorized immigrants ensures that the subsequent parameter estimates are not affected by the undercount of unauthorized immigrants in the ACS 2010. This newly generated data set is referred to as the "pseudo legal" data; it constitutes of 243,269 observations as shown in Table 2.5. This concludes the intermediate step of generating new data.

A possible limitation of this pseudo legal dataset is that it only contains new information on authorized immigrants. Hence, adjusting the prediction model using pseudo legal dataset may lead

[13] See footnote 9 for details of this assumption. There are 21,793 such assumed authorized immigrants.

[14] Another version was considered in estimation where adjustments were carried out on a 30 per cent random sample and the remaining 70 -holdout sample was used to check the performance of the adjustment model. The use of the entire updating dataset is favored over a 30 percent sample because the former is able to add more countries and other indicators at the adjustment stage, which results in greater variables in the final adjusted prediction model, and thus better precision in final estimates.

[15] Arrived after 1979 and over 17 years of age.

	Number	Pseudo legal status assignment
Naturalized citizens	97,196	1
Not naturalized citizens	146,073	
Assumed[1]	22,898	1
$P(L)_{naive} \geq 0.5$	81,156	1
$P(L)_{naive} < 0.5$	42,019	0

1. Assumed based on govt. occupation, subscription to public health insurance plans, receipt of public assistance and veteran status.

Table 2.5: Composition of the pseudo legal data set

predictions to be biased towards assigning a foreign born person an 'authorized' outcome. However, this bias can be a virtue by itself, for greater accuracy in predicting authorized immigrants implies a lower rate of false authorized prediction for those who are in fact unauthorized immigrants. All adjustments are carried out using the pseudo legal dataset with the intuition to adjust the prediction model to ensure that enhanced authorized immigrants are predicted as pseudo legal immigrants. The exact procedure is as follows.

2.4.4 Adjusting methodology

Step 1

Compute $(\hat{\alpha}_{derivation} + X'\hat{\beta}_{derivation})$ with the parameters of the derivation model[16] applied to explanatory variables for each observation in the pseudo legal data, and refer to this as the linear predictor.

[16]These parameters include state coefficients for the non-overlapping states, as per footnote 10.

Variable	Coefficient	Robust S.E.
Linear Predictor	1.4135	0.0097
Constant	1.0873	0.0075
Number of obs	243,269	
Wald chi2(1)	21056.07	
Prob > chi2	0.00000	
Pseudo R2	0.5116	
Log pseudo likelihood	-61833.182	

Table 2.6: Calibration model

Step 2

Estimate a logistic regression of the pseudo legal status on the linear predictor ($\hat{\alpha}_{derivation} + X'\hat{\beta}_{derivation}$ as the only covariate as seen in equation (2.4),

$$P(pseudo\ legal\ status) = \Lambda(\alpha_{calibration} + \beta_{calibration} \times (\hat{\alpha}_{derivation} + X'\hat{\beta}_{derivation})) \tag{2.4}$$

where $\Lambda(z)$ is the logistic function: $\exp(z)$ / $(1+\exp(z))$. The intuition behind this step is to adjust the derivation model to ensure that enhanced authorized immigrants are predicted to be pseudo *legal*. Technically, the purpose of this step is to adjust the derivation model for calibration in the context of ACS data.

A calibration slope equal to one and a calibration constant equal to zero means that derivation model is perfectly applicable in the context of new data and the derivation coefficients and the constant do not need adjustments for calibration.

As seen in Table 2.6, empirical estimation of this step on the pseudo legal data estimates a statistically significant calibration slope of 1.4135, which is different from 1, and a calibration constant of 1.0873, which is different from zero. This implies that the calibration power of the

derivation model requires adjustment. The adjustment to the intercept is shown in equation 2.5.[17]

$$\alpha_{adjusted} = \alpha_{calibration} + \hat{\alpha}_{derivation} \times \beta_{calibration} \tag{2.5}$$

Adjustments to derivation coefficients require further refinement for time dynamics as shown in the subsequent steps.

Step 3

The intuition behind the third step is to check if the effect of each variable on predicting the probability of being legal changed over the years, even after adjustments for calibration are made as in Step 2 above. The econometric procedure used here is model extension by stepwise forward selection and the linear predictor ($X'\hat{\beta}_{derivation}$) is not subjected to the selection criteria.[18] Since the linear predictor ($X'\hat{\beta}_{derivation}$) is already in the model, a maximum of $K-1$ predictors may be included at this step where K is the number of predictors in derivation model.

As such, I define this subset of the original vector of (X) consisting of $K-n$ elements as Z_1, while γ is the vector of $K-n$ coefficients reflecting the time dynamic adjustment for each covariate in the derivation model. γ not equal to zero indicates that the effect of the predictor is still different in the pseudo legal dataset even after recalibration with equation (2.4).

$$P(pseudo\ legal\ status) = \Lambda(\alpha_{calibration} + \beta_{calibration} \times (\hat{\alpha} + X'\hat{\beta}_{derivation} + Z_1'\gamma) \tag{2.6}$$

where $\Lambda(z)$ is once again a logistic function $\exp(z)$ / $(1+\exp(z))$.

The empirical estimation of this step resulted in 66 predictors[19] being selected for inclusion as shown in Table 2.9. This implies that 66 variables have a different effect on the legal status in recent data compared to 1986 data. The inclusion of 7 out of 13 occupational cluster dummies at this stage shows that occupational structure of unauthorized immigrants has changed since 1986.

[17]The methodology developed by Steyerberg *et al.* (2004) does not multiply the the calibration constant by the derivation slope. However, such a multiplication is required since the linear predictor includes the derivation constant.

[18]Each predictor is considered for model extension at a time and the test of term significance is based on the likelihood-ratio test, where the level of significance for adding and removing terms are set at 0.05 and 0.051, respectively.

[19]The predictors include 35 country dummies, 20 state dummies, 7 occupation cluster dummies, 3 marital status dummies, and male.

Moreover, three of these coefficients on the occupational clusters are statistically significant at the 5 percent significance level, which serves as a formal test of the critical assumption made by the residual methodology namely, that the occupational structure of unauthorized immigrants remained constant since 1986. Similarly, the other variables listed in Table 2.9 also require adjusting for time dynamics.

Among the imputed state coefficients (those that did not overlap between the LPS and INS in the derivation dataset), 5 were picked up in the time dynamic model indicating these coefficients required adjustment for time dynamics. As such Step 3 also serves the purpose of showing that many of the imputed state coefficients were appropriate.

Step 4

As identified by above Steyerberg *et al.* (2004) and Janssen *et al.* (2008), the newly estimated coefficients can suffer from overfitting, which requires refitting towards the recalibrated values, as specified by equation 2.7 below. The intuitive idea of this step is to balance the adjusted coefficients between changes for time dynamic adjustments and calibration adjustments.

$$\beta_{adjusted} = \beta_{calibration} \times \beta_{derivation} + \text{refitting factor} \times \gamma \tag{2.7}$$

$\beta_{calibrated}$ is from equation (2.4), and γ is from equation (2.6). The refitting factor is calculated as follows.

$$\text{refitting factor} = \frac{model\chi^2_{updated-calibration} - ddf}{model\chi^2_{updated-calibration}} \tag{2.8}$$

where $\chi^2_{updated-calibration}$ is the negative value of twice the difference in the log likelihood of a model with estimated deviations of individual regression coefficients (time dynamic model) and the log likelihood of the calibration model. *ddf* refers to the difference in degrees of freedom between these two models. This produces a refitting factor of 1.004382.

2.4.5 Adjusted coefficients

Finally, the adjusted coefficients can be summarized as follows (see Table 2.10 for the full listing of adjusted coefficients):

Variable	Coefficient	SE (robust)	dy/dx	SE (delta)
Linear predictor	1.5667	0.0134	0.1186	0.0008
Male	-0.4931	0.0224	-0.0373	0.0017
Single	-1.1979	0.0976	-0.0907	0.0074
Married	-0.9504	0.0973	-0.0719	0.0074
Divorced /Separated	0.4578	0.1051	0.0347	0.0080
Occupations				
Other health assessment, treating	0.3973	0.1649	0.0301	0.0125
– health technologists and technicians				
Writers, artists, entertainers, and athletes	-0.3085	0.1015	-0.0234	0.0077
Administrative support, including clerical	-0.1224	0.0456	-0.0093	0.0035
Service occupations	-0.3250	0.0339	-0.0246	0.0026
Farming, forestry, and fishing occupations	-1.0637	0.0415	-0.0805	0.0031
Precision production, craft, repair,	-0.7265	0.0319	-0.0550	0.0024
– operators, fabricators, and laborers				
Homemakers ,unemployed or retired etc.	-0.3303	0.0477	-0.0250	0.0036
State of residence				
Alabama	-0.4744	0.1144	-0.0359	0.0087
Arizona	0.2610	0.0551	0.0198	0.0042
Arkansas	-0.2956	0.1225	-0.0224	0.0093
California	0.5776	0.0299	0.0437	0.0023
Colorado	-0.4942	0.1926	-0.0374	0.0146
Florida	-0.1440	0.0380	-0.0109	0.0029
Georgia	0.3463	0.0914	0.0262	0.0069
Hawaii	1.3313	0.4515	0.1008	0.0342
Idaho	0.6810	0.2221	0.0515	0.0168
Indiana	0.6112	0.1425	0.0463	0.0108
Maryland	-0.1844	0.0674	-0.0140	0.0051
Michigan	0.2508	0.1166	0.0190	0.0088
New Jersey	-0.2499	0.0499	-0.0189	0.0038
New Mexico	1.0176	0.1687	0.0770	0.0128
Oklahoma	0.6173	0.1763	0.0467	0.0134
South Carolina	-0.3600	0.0901	-0.0272	0.0068
Texas	-0.2118	0.0303	-0.0160	0.0023
Virginia	-0.3177	0.0661	-0.0240	0.0050
Washington	0.6921	0.0818	0.0524	0.0062
West Virginia	-0.9747	0.3742	-0.0738	0.0283

Table 2.7: Time dynamic model

Variable	Coefficient	SE (robust)	dy/dx	SE (delta)
Country of birth				
Albania	1.8156	0.8109	0.1374	0.0614
Argentina	-0.6793	0.1292	-0.0514	0.0098
Australia	-1.0002	0.2906	-0.0757	0.0220
Bangladesh	0.4392	0.1555	0.0332	0.0118
Bermuda	-1.4182	0.6109	-0.1073	0.0462
Bolivia	-0.3455	0.1738	-0.0261	0.0132
Brazil	-0.5240	0.1035	-0.0397	0.0078
Bulgaria	-0.5005	0.2071	-0.0379	0.0157
Canada	-0.6297	0.1078	-0.0477	0.0082
Colombia	-0.2633	0.0677	-0.0199	0.0051
Egypt	2.6017	1.0031	0.1969	0.0759
Ethiopia	0.8110	0.1412	0.0614	0.0107
France	-0.8183	0.3350	-0.0619	0.0254
Ghana	0.3468	0.1466	0.0263	0.0111
Guatemala	-0.5919	0.0502	-0.0448	0.0038
Honduras	-0.6877	0.0552	-0.0521	0.0042
India	0.6939	0.2626	0.0525	0.0199
Iran	0.2717	0.1338	0.0206	0.0101
Israel	-0.9011	0.3799	-0.0682	0.0288
Jamaica	-0.3194	0.1319	-0.0242	0.0100
Japan	-0.7152	0.2587	-0.0541	0.0196
Mexico	-0.5776	0.0290	-0.0437	0.0022
Morocco	-0.9718	0.4663	-0.0736	0.0353
Netherlands	-1.2840	0.3928	-0.0972	0.0297
Nigeria	0.4454	0.1336	0.0337	0.0101
Pakistan	0.3915	0.1419	0.0296	0.0107
Peru	-0.2155	0.0852	-0.0163	0.0065
Sierra Leone	1.0593	0.3048	0.0802	0.0231
Somalia	0.7893	0.3400	0.0597	0.0257
Tanzania	-0.8287	0.3833	-0.0627	0.0290
Thailand	0.4443	0.1904	0.0336	0.0144
Tonga	-1.0031	0.4769	-0.0759	0.0361
United Kingdom	-0.7554	0.1316	-0.0572	0.0099
Uruguay	-0.8995	0.1885	-0.0681	0.0143
Venezuela	-0.5516	0.1122	-0.0418	0.0085
Constant	2.9937	0.1018		

Table 2.8: Time dynamic model (cont.)

Statistic	Value
Number of obs	225,489
Wald chi2(67)	29724.7
Prob > chi2	0.00000
Pseudo R2	0.5536
Log pseudo likelihood	-54303.142

Table 2.9: Model statistics of time dynamic model

- Predictors which have a different effect on the outcome variable between the derivation and pseudo legal datasets (66):

$$\beta_{adjusted} = \beta_{calibration} \times \beta_{derivation} + \text{refitting factor} \times \gamma \tag{2.9}$$

- Predictors which do not have a different effect on the outcome variable between the derivation and pseudo legal datasets (97 variables):

$$\beta_{adjusted} = \beta_{calibration} \times \beta_{derivation} \tag{2.10}$$

- Another 7 new predictors were added to the model at the time dynamic model stage.[20] These new predictors are also refitted as per equation (2.8).

The methodology developed by Steyerberg *et al.* (2004) does not have a methodology to calculate standard errors for the adjusted coefficients. However, given that the adjusted coefficients are a linear combination of coefficients of three models, – derivation, calibration and time dynamic models the intuitive standard errors can be approximated as follows. First, for adjusted coefficient k , I find the minimum values that derivation and time dynamic coefficients as well as calibration slope ($\beta_{deri,k}$, β_{cali} and γ_k) can take by subtracting their respective standard errors from the coefficient estimates. Similarly, the maximum values that $\beta_{deri,k}$, β_{cali} and γ_k can take are calculated by adding their respective standard errors to the coefficient estimates. Finally, the intuitive standard

[20]These were 4 countries (Tanzania, Albania, Bermuda and Bulgaria) and base categories (married, male; and the state Alabama) in the derivation model.

errors can be arrived by taking the absolute value of difference between the linear combinations of the maximum and minimum values that the adjusted coefficient of variable k can take and dividing same by two.

$$\begin{aligned} k_{min} &= (\beta_{cali} - SE_{\beta\ cali}) \times (\beta_{deri,k} - SE_{\beta\ deri,k}) + \text{refitting factor}(\gamma_k - SE_{\gamma,k}) \\ k_{max} &= (\beta_{cali} + SE_{\beta\ cali}) \times (\beta_{deri,k} + SE_{\beta\ deri,k}) + \text{refitting factor}(\gamma_k + SE_{\gamma,k}) \end{aligned} \tag{2.11}$$

$$SE\ \beta_{adjusted_k} = |k_{min} - k_{max}|/2 \tag{2.12}$$

2.4.6 Prediction and aggregate estimate of unauthorized immigrants

In order to predict legal status, the adjusted coefficients are multiplied by their respective variables ($X'\hat{\beta}_{adjusted}$) for each foreign born person in the 2010 ACS, excluding those assumed legal and those who entered the US prior to 1980, and a probability between 0 and 1 is arrived by the following transformation.

$$P(y_i = 1|X) = \frac{\exp(X'\hat{\beta}_{adjusted})}{1 + \exp(X'\hat{\beta}_{adjusted})} \tag{2.13}$$

The aggregate estimate of unauthorized immigrants (T) for January 1, 2011 is arrived by taking the summation of the product of each individual's predicted probability to be *unauthorized* and the number of persons represented by this individual. This estimate is then adjusted for undercount of unauthorized immigrants by multiplying by 1.1, as in equation (2.14).

$$T = \left\{ \sum_{i}^{n} (1 - P_i) \times w_i \right\} \times 1.1 \tag{2.14}$$

where P_i is from equation (2.13) and w_i is the sample weight.

In order to distinguish from the alternative estimates produced in subsequent sections, this estimate is defined the sample weight based estimate of unauthorized immigrants. In this context, the sample weight based aggregate estimate shows that there were 7,700,869 adult unauthorized immigrants in the US on January 1, 2011. Aggregate level characteristic estimates of unauthorized immigrants can also be estimated by this sample weight based approach.

Variable	Adjusted Coeff.	SE	Variable	Adjusted Coeff.	SE
Age	0.077	0.004	Michigan	0.387	2.093
Male	-0.495	0.023	Minnesota	2.499	0.017
Female	-0.381	0.047	Mississippi	-1.401	1.787
Single	-1.203	0.098	Missouri	2.499	0.017
Marries	-0.981	0.157	Montana	1.041	3.597
Widow	-2.582	0.201	Nebraska	2.499	0.017
Divorce/separated	-2.867	0.216	Nevada	4.102	2.238
State of residence			New Hampshire	4.252	1.378
Alabama	-0.477	0.115	New Jersey	-0.546	1.653
Alaska	0.043	3.486	New Mexico	-2.018	1.757
Arizona	-0.648	1.658	New York	-0.643	1.597
Arkansas	0.144	2.208	North Carolina	3.140	1.974
California	-1.097	1.618	North Dakota	2.499	0.017
Colorado	3.516	2.325	Ohio	0.135	1.976
Connecticut	4.870	2.181	Oklahoma	-2.776	1.766
D.C.	0.029	1.631	Oregon	4.055	2.209
Delaware	0.308	1.852	Pennsylvania	-0.469	0.243
Florida	-1.357	1.634	Rhode Island	2.705	2.240
Georgia	2.361	2.164	South Carolina	-0.054	1.942
Hawaii	1.380	3.940	South Dakota	2.499	0.017
Idaho	1.725	3.820	Tennessee	-1.401	0.010
Illinois	-1.261	1.593	Texas	-0.831	1.626
Indiana	2.145	1.978	Utah	1.041	3.597
Iowa	2.499	0.017	Vermont	4.252	1.378
Kansas	2.499	0.017	Virginia	-1.253	1.674
Kentucky	-1.401	0.010	Washington	-1.553	1.677
Louisiana	-2.263	1.597	West Virginia	-0.671	2.228
Maine	4.252	1.378	Wisconsin	0.135	1.976
Maryland	-1.375	1.669	Wyoming	1.041	3.597
Massachusetts	5.181	2.178			

Note: Standard errors as per Equations (2.11) and (2.12).

Table 2.10: Adjusted model

Variable	Adjusted Coeff.	SE	Variable	Adjusted Coeff.	SE
Country of birth					
Albania	1.824	0.814	Grenada	0.933	1.160
Algeria	0.967	1.185	Guatemala	-3.007	0.984
Antigua-Barbuda	-0.941	1.098	Guyana	1.369	1.006
Argentina	-0.522	1.122	Haiti	-0.547	0.956
Australia	0.425	1.415	Honduras	-2.048	1.008
Bahamas	-0.741	1.067	Hong Kong	1.709	1.033
Bangladesh	-0.493	1.222	Hungary	3.654	1.783
Barbados	3.448	1.427	Iceland	0.157	1.710
Belize	-1.732	0.968	India	2.798	1.264
Bermuda	-1.424	0.614	Indonesia	0.775	1.190
Bolivia	-0.742	1.222	Iran	-0.364	1.088
Brazil	-0.374	1.138	Iraq	1.941	1.191
Bulgaria	-0.503	0.208	Ireland	1.212	1.135
Canada	-0.396	1.074	Israel	0.752	1.445
Chile	-0.102	0.992	Italy	3.012	1.098
China	2.087	0.986	Jamaica	0.789	1.102
Colombia	-1.040	1.017	Japan	1.442	1.334
Costa Rica	0.254	1.037	Jordan	1.736	1.099
Cuba	5.711	1.191	Kenya	0.066	1.117
Cyprus	0.838	1.352	Korea	2.904	1.009
Dominican Rep.	0.550	0.964	Lebanon	1.379	1.057
Ecuador	-0.479	0.959	Liberia	-1.535	1.056
Egypt	4.073	2.095	Malaysia	0.558	1.256
El Salvador	-3.174	0.926	Mali	-3.005	1.473
Ethiopia	-1.376	1.126	Mexico	-3.072	0.959
Fiji	3.916	1.799	Morocco	1.373	2.323
France	0.692	1.443	Netherlands	0.072	1.549
Germany	2.700	1.104	New Zealand	2.478	1.848
Ghana	-1.724	1.115	Nicaragua	-1.539	0.954
Greece	3.367	1.167	Nigeria	-1.796	1.092

Note: Standard errors as per Equations (2.11) and (2.12).

Table 2.11: Adjusted model (cont.)

Variable	Adjusted Coeff.	SE	Variable	Adjusted Coeff.	SE
Norway	1.795	1.717	Sweden	1.731	1.392
Pakistan	0.072	1.114	Switzerland	3.010	1.474
Panama	1.298	1.048	Syria	1.152	1.156
Paraguay	0.544	1.957	Taiwan	1.232	1.050
Peru	-0.844	1.044	Tanzania	-0.832	0.385
Philippines	1.663	0.967	Thailand	0.942	1.170
Poland	-1.103	0.952	Tonga	-1.792	1.575
Portugal	2.822	1.134	Trinidad-Tobago	1.229	1.002
Senegal	-3.342	1.514	Turkey	1.123	1.122
Sierra Leone	-1.758	1.349	Uganda	-0.238	1.140
Singapore	1.338	1.914	United Kingdom	0.362	1.109
Somalia	0.935	1.655	Uruguay	-1.509	1.239
South Africa	1.253	1.222	Venezuela	-1.225	1.155
Spain	1.708	1.129	Vietnam	6.000	1.226
Sri Lanka	1.239	1.700	Western Samoa	-2.370	1.068
St. Lucia	0.381	1.331	Yugoslavia	1.328	1.034
St. Vincent-Grenadines	1.730	1.255			
Occupations					
Lawyers, medical doctors, engineers, – surveyors, scientists and urban planners				2.007	0.235
Other health assessment, treating occupations, – health technologists and technicians				0.787	0.580
Teachers, postsecondary, secondary, councellors – edu. and vocational, librarians, archivists, and curators				2.267	0.404
Social, recreation, and religious workers				2.026	0.699
Writers, artists, entertainers, and athletes				0.335	0.370
Technologists and technicians, except health				0.931	0.253
Sales occupations				-0.697	0.165
Administrative support occupations, including clerical				0.192	0.182
Service occupations				-1.126	0.142
Farming, forestry, and fishing occupations				-0.309	0.212
Precision production, craft, repair occupations –operators, fabricators, and laborers				-0.837	0.140
Homemakers, unemployed, students or retired etc.				2.387	0.185
Constant				0.992	1.438

Note: Standard errors as per Equations (2.11) and (2.12).

Table 2.12: Adjusted model (cont.)

2.4.7 Assignment of unauthorized immigrant status

In addition to arriving at the aggregate estimate of unauthorized immigrants, this methodology can also be used to derive a cross sectional data set of unauthorized immigrants. Such a cross sectional data set is based on assignment of legal versus unauthorized immigrant status after determining a cutoff probability. Upon identification of such a critical probability or a cutoff point each observation is assigned a predicted authorized or unauthorized status. An observation is considered a predicted unauthorized immigrant if its predicted probability is *below* the identified critical probability.[21] This completes the development of the microdata-based methodology.

Similar to the existing methodologies, this micro data based methodology too has its limitations. They are the sensitivity of estimates to i) the Assumptions 1 to 3; ii) sample selection in the ACS PUMS; iii) choice of cutoff probability in deriving pseudo legal data; iv) choice of cutoff probability for assignment of predicted unauthorized immigrant status; and (v) choice of the rate of adjustment for undercount of unauthorized immigrants. Moreover, given that the LPS does not represent unauthorized immigrants that have been living in US for less than five years, the estimates produced by this methodology should be qualified as lower bound estimates.

2.5 Assignment based cross sectional data of unauthorized immigrants

The ACS PUMS captures a total of 178,932 not naturalized foreign born persons[22]. Out of this number, 123,175 are considered for the prediction of legal status, after the 55,757 are excluded based on year of entry, age and assumed authorized immigrant status. The component of 'not naturalized citizens' considered for the prediction of legal status includes both unauthorized and authorized immigrants. Figure 2.1 depicts the predicted legal probabilities for this pool of non-citizen immigrants for whom legal status is predicted. The bimodal distribution of predicted probabilities for this pool shows that there are two distinct groups at the two extremes and the hard to distinguish immigrants in between.

As such, the assignment of predicted authorized and unauthorized immigrant status is affected

[21] Since the derivation model estimated the probability of being *legal*.

[22] These not naturalized foreign born persons represents 22,460,564. These numbers include those arriving on or before 1979 as well as minors.

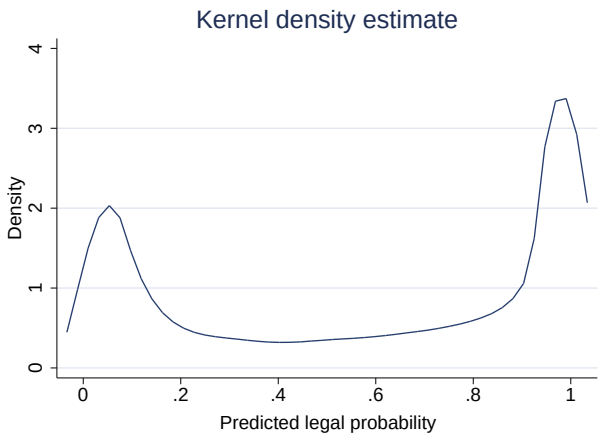

Figure 2.1: Distribution of Predicted Probabilities

by the choice of the cutoff point. Hence, the selection of this cutoff probability cannot be arbitrary, but must be well founded on criteria such as (i) an accurate reflection of actual unauthorized immigrants; and (ii) efficiency in terms of minimum number of mis-classification and maximum number of accurate classification. A critical value based on observed unauthorized immigrants is the ideal candidate for this purpose. In the absence of recent data with observed outcomes of unauthorized immigrants, the second best alternative is a critical value based on enhanced legal immigrants.

The predicted probability of being legal for enhanced authorized immigrants ranged between .0009664 and 1, with the mean at 0.7592231. This mean probability of 0.7592231, corresponds to the 31.9th percentile of enhanced authorized immigrants. As such, by setting the cutoff at the mean, I would be able to accurately assign a predicted legal status of 'authorized immigrant' for 69.1 percent of enhanced authorized immigrants. When the cutoff is set at this point, the microdata-based method estimates 9,603,458 adult unauthorized immigrants. Meanwhile when the cutoff probability is at 0.5, more than 78 percent of enhanced authorized immigrants are assigned an authorized immigrants status, and only 22 percent of enhanced authorized immigrants are mis-classified as unauthorized immigrants.

Column (i) in Table 2.13 depicts percentiles of enhanced legal immigrants, when their legal probability is predicted using the microdata-based method, while column (ii) shows the corresponding probabilities for the percentiles. Column (iii) shows the estimated number of unauthorized immi-

grants [23] using the probabilities in column (ii) as the cutoff probabilities. As such, column (i) can be interpreted as the margin of error in mis-classification. For instance, if I were to predict legal status of the enhanced immigrants using the cutoff of 0.0538, a 5 percent of enhanced legal would be mis-classified as unauthorized immigrants. Estimates in column (iii) excludes all enhanced authorized immigrants. As such, the margin of error implies that 5 percent of the true legal immigrants in the pool of non-naturalized immigrants may be mis-classified.

In order to compare estimates from microdata-based methodology with naive estimates, the last three columns in Table 2.13 correspond to the naive estimates. As such, column (iv) shows the percentiles of enhanced immigrants, when their legal probability is predicted using the naive methodology. Similarly, column (v) reports the corresponding probabilities, while column (vi) reports the estimated number of unauthorized immigrants among the pool of non-immigrants, when probability in column (v) is used as the cutoff.

As evident in Table 2.13, the estimated number of unauthorized immigrants is sensitive to the choice of cutoff probability as well the adjustment rate for undercount of unauthorized immigrants. When the cutoff probability is closer to 1 the estimated count and the mis-classification of enhanced legal immigrants (indicated by percentile) increase.

The two different estimates (columns (iii) and (vi)) depicted in Table 2.13 can be compared in two ways. On the one hand, the estimates can be compared for a specific percentile or misclassification rate along a given row, while the cutoff probabilities vary. On the other hand, the estimates can be compared for a given cutoff probability when the misclassification rate varies.

As evident in the entire table, across rows the naive estimates are much smaller than the microdata-based estimate. At lower margins of error the naive estimate is much smaller than the microdata-based estimate. For instance, up to the 22nd percentile the naive estimate underestimates the microdata-based estimate by over 1 million. As seen in the table, when the margin of error increases, the gap between the two estimates decreases. For instance at the 95^{th} percentile, where almost all enhanced authorized immigrants are misclassified as unauthorized immigrants under both methodologies, the naive estimate is only 42,504 smaller than the microdata-based estimate.

[23] Among the pool of non-immigrants that includes unauthorized immigrants. This pool is different from enhanced legal immigrants.

As depicted in the left panel in in Figure 2.2, the naive estimates are always to the right of microdata-based estimate, reflecting that for any estimate the naive estimate corresponds to a higher margin of error. Given that the naive method is comparable to existing methodologies due to the absence of the two adjustments for calibration and time dynamics, the comparison of it with microdata-based estimates gives an idea about the error evident in existing estimates. As such, until the development of the new microdata-based methodology, there was no knowledge about the degree of accuracy in the assignment of legal status in the existing estimates.

The right panel in Figure 2.2 depicts estimates by the two methodologies while holding the cutoff probabilities constant and varying the margin of error. Here it is evident that up to the cutoff probability of approximately 0.8155 the naive methodology underestimates unauthorized immigrants and thereafter the naive methodology overestimates the microdata-based estimates.

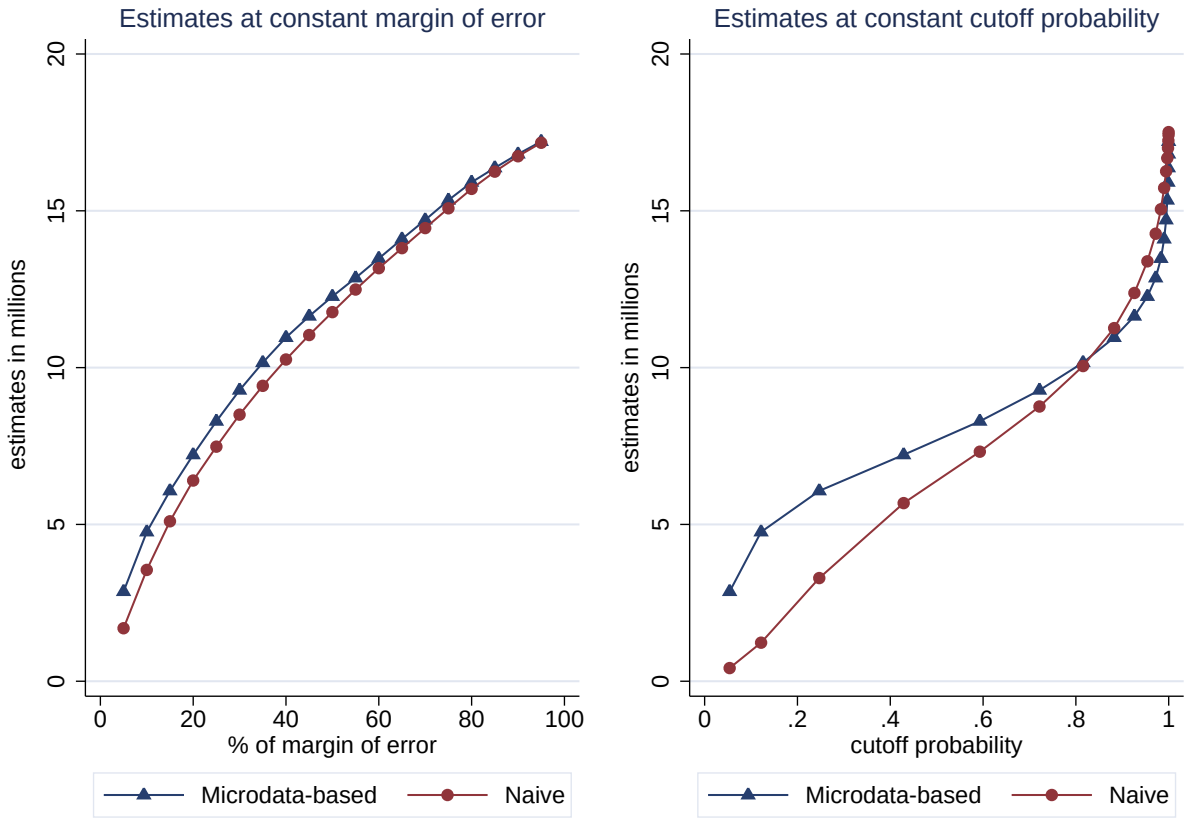

Figure 2.2: Comparison of microdata-based estimate with naive estimates of unauthorized immigrants

Microdata Based			Naive		
Percentile (i)	Cutoff probability (ii)	Estimate (iii)	Percentile (iv)	Cutoff probability (v)	Estimate (vi)
5.00	0.0538924	2,855,267	5.00	0.1502683	1,689,155
10.00	0.1217110	4,764,940	10.00	0.2633987	3,550,071
15.00	0.2469559	6,068,766	15.00	0.3770186	5,097,655
20.00	0.4289307	7,221,710	20.00	0.4991578	6,401,924
22.00	0.5000000	7,661,917	20.05	0.5000000	6,413,914
25.00	0.5929732	8,292,323	25.00	0.6076100	7,479,169
30.00	0.7215978	9,281,653	30.00	0.6981044	8,498,251
31.85	0.7592231[a]	9,603,458	34.35	0.7636631[b]	9,300,661
35.00	0.8155083	10,157,616	35.00	0.7724156	9,423,192
40.00	0.8825435	10,955,512	40.00	0.8285199	10,261,754
45.00	0.9261333	11,637,901	45.00	0.8715953	11,035,984
50.00	0.9541220	12,272,877	50.00	0.9047912	11,774,554
55.00	0.9720300	12,859,203	55.00	0.9295163	12,485,874
60.00	0.9833262	13,476,915	60.00	0.9487190	13,170,080
65.00	0.9902495	14,099,286	65.00	0.9634084	13,812,174
70.00	0.9944584	14,713,459	70.00	0.9749316	14,453,937
75.00	0.9970374	15,340,174	75.00	0.9836746	15,081,059
80.00	0.9984754	15,906,021	80.00	0.9899695	15,696,617
85.00	0.9992943	16,373,175	85.00	0.9943818	16,248,287
90.00	0.9997437	16,807,680	90.00	0.9973314	16,743,203
95.00	0.9999484	17,212,498	95.00	0.9990648	17,169,994

a: Mean of predicted probability of enhanced authorized immigrants
– using Microdata based method
b: Mean of predicted probability of enhanced authorized immigrants
– using naive method

Table 2.13: Estimates of the number of adult unauthorized immigrants

2.6 Application of microdata-based methodology

The microdata-based methodology can be applied in four distinct manners. First, a sample weight based aggregate estimates of the population of adult unauthorized immigrants can be arrived as per foregone equation (2.14) in Section 2.4.6. Sample weight based characteristic estimates also can be arrived in the similar fashion.

Second, for a given estimated count, this method can be used to reserve engineer and find the cutoff probability and thereby assign predicted authorized and unauthorized immigrant status and derive a cross sectional dataset. As an example of the second application of the microdata-based approach, I replicated the age adjusted DHS estimate of 9.56 million adult unauthorized immigrants for January 1, 2010 using data from the ACS 2009.[24] This replication exercise revealed that the margin of error associated with the age adjusted DHS estimate is 33.7 percent.

Third, the microdata-based method can be used with a predetermined cutoff probability (or a margin of error) to assign predicted authorized and unauthorized status and forward engineer to derive a cross sectional dataset with assigned legal status. As an example of the forward engineered application, I estimate the number of adult unauthorized immigrants with a predetermined margin of error of 33.7 percent on the ACS 2010.

In order to distinguish from the sample weight based estimate produced as per equation (2.14) in Section 2.4.6, I define the new estimate as the assignment based estimate of adult unauthorized immigrants. The assignment based estimate of adult unauthorized immigrants for January 1, 2011 is 9,920,602. This margin of error of 33.7 percent is chosen to coincide with the existing estimate by DHS for January 1, 2010 (DHS estimate for January 1, 2011 is not yet released). The characteristics of unauthorized immigrants analyzed in the next section is based on this forward engineered application, thus they are also assignment based estimates.

Among the two alternative approaches to aggregate estimates (including characteristic estimates), sample weight based estimates are superior to assignment based estimates, due to the absence of a cutoff point and the related sensitivity and mis-classification issues. Nonetheless, the

[24] ACS captures data for the entire 2009 year, as such the data reflects a snapshot of the US population as of January 1, 2010. The DHS estimate for January 1, 2010 is 10.79 million, which includes 1.23 million unauthorized immigrants under 18 years of age. As such, the age adjusted DHS estimates is 9.56 million. In order to arrive at these estimates for January 1, 2010 the entire microdata-based methodology was replicated using ACS 2009 data. As such, the derivation model, calibration model, time dynamic model and adjusted coefficients applicable the estimates for January 1, 2010 are different from those presented in Section 2.4.

analysis in Section 2.7 next is based on assignment based estimates, with the interest of comparing the performance of the microdata-based methodology relative to other existing methodologies.

This second and third applications discussed above also serves as a robustness test for the methodology. Specifically, compared to the existing estimate for January 1, 2010, the estimate for January 1, 2011 (at the same margin of error) is higher by 358,832 unauthorized immigrants. This increase is consistent with the overall trend of the existing estimates of unauthorized immigrants for recent years by the DHS and the PHC.

Fourth, the predicted probability of being unauthorized can be used as a continuous variable in econometric applications.

2.7 Characteristics of predicted unauthorized immigrants: January 2011

In this section I perform a descriptive analysis of unauthorized immigrants based on the *assignment* based estimate of 9,920,602 adult unauthorized immigrants on January 1, 2011.[25] Table 2.14 depicts predicted unauthorized immigrants by period of entry. Nearly 50 percent of predicted unauthorized immigrants have arrived in US during the last decade. The entry period 2000-2004 accounts for the largest share of arrivals among the predicted unauthorized immigrants for January 2011. The furthest entry cohorts 1980-1984 records the smallest share of predicted unauthorized immigrants for January 2011. This may be due to either or both of the following reasons; (i) fewer unauthorized immigrant entered the US in the first half of 1980's, and (ii) most unauthorized immigrants who entered in the early years have left the US. An accurate idea of departure of unauthorized immigrants from US can be gained once the microdata-based estimate is replicated for several years using a single cutoff probability which facilitates comparison across years.

Table 2.15 shows the distribution of predicted unauthorized immigrants by state of residence for January 1, 2011. As seen in the table nearly 50 percent of predicted unauthorized immigrants reside in 2 states, namely, California and Texas, and the top 8 states account for nearly 75 percent of predicted unauthorized immigrants. As noted previously, existing estimates by the DHS assume

[25]This assignment based aggregate estimate is different from sample weight based aggregate estimate of 7,700,869 introduced in Section 2.4.6. The former is chosen for here to facilitate comparison between existing and new estimates.

Entry cohort	Estimated number	%
2005-2010	2,209,866	22
2000-2004	2,685,708	27
1995-1999	1,977,077	20
1990-1994	1,464,397	15
1985-1989	1,031,712	10
1980-1984	551,841	6
Total	9,920,602	100

Table 2.14: Period of entry of predicted unauthorized immigrants : January 2011

that immigrants remain in the same state that they initially settled after arriving in US. On the contrary, estimates by the microdata-based methodology do not assume away internal migration. As such, the estimates presented in Table 2.15 is a true reflection of the state of residence of predicted unauthorized immigrants for January 1, 2011.

Table 2.16 shows the top 25 countries of origin for predicted unauthorized immigrants, and Mexico accounted for 63 percent. Such a dominance of Mexico among predicted unauthorized immigrants is consistent with the existing estimates. For instance, DHS estimates for the past seven years show that on average 59 percent of unauthorized immigrants originated from Mexico (Hoeffer *et al.*, 2011). Similarly, PHC estimates for the same period ranged between 6 to 7 million unauthorized immigrants originating from Mexican (Passel & Cohn, 2011).

State	Estimate	%	State	Estimate	%
California	2,828,516	28.51	Kentucky	49,401	0.50
Texas	1,755,441	17.69	Kansas	44,700	0.45
Florida	861,860	8.69	Minnesota	36,183	0.36
New York	705,729	7.11	Missouri	24,908	0.25
Illinois	590,391	5.95	Mississippi	24,061	0.24
New Jersey	362,518	3.65	Nevada	23,679	0.24
Arizona	299,218	3.02	Idaho	23,300	0.23
Virginia	266,267	2.68	Iowa	22,950	0.23
Maryland	235,098	2.37	Nebraska	21,095	0.21
Washington	227,294	2.29	Oregon	16,160	0.16
Georgia	213,693	2.15	District of Columbia	13,813	0.14
North Carolina	132,226	1.33	Delaware	13,528	0.14
Tennessee	120,272	1.21	Rhode Island	9,160	0.09
Pennsylvania	108,331	1.09	Alaska	6,455	0.07
Oklahoma	104,632	1.05	West Virginia	3,975	0.04
New Mexico	88,420	0.89	Hawaii	3,924	0.04
South Carolina	87,903	0.89	Wyoming	3,596	0.04
Alabama	77,205	0.78	Connecticut	1,234	0.01
Utah	75,712	0.76	South Dakota	1,085	0.01
Louisiana	67,156	0.68	Montana	833	0.01
Michigan	65,469	0.66	Massachusetts	361	0.00
Ohio	64,523	0.65	North Dakota	353	0.00
Wisconsin	62,962	0.63	New Hampshire	-	0.00
Indiana	60,711	0.61	Vermont	-	0.00
Arkansas	57,712	0.58	Maine	-	0.00
Colorado	56,593	0.5705	TOTAL	9,920,602	100.00

Table 2.15: State of residence of predicted unauthorized immigrants : January 2011

Country	Estimate	%	Country	Estimate	%
Mexico	6,274,956	63.25	Venezuela	69,922	0.70
El Salvador	681,666	6.87	Jamaica	52,227	0.53
Guatemala	468,898	4.73	Philippines	46,332	0.47
Honduras	305,845	3.08	Argentina	46,299	0.47
Colombia	172,473	1.74	Nigeria	43,930	0.44
Ecuador	147,524	1.49	United Kingdom	43,841	0.44
Peru	124,494	1.25	Ethiopia	34,958	0.35
Haiti	120,038	1.21	Ukraine	33,898	0.34
Poland	102,656	1.03	Ghana	33,776	0.34
Dominican Republic	91,054	0.92	Trinidad and Tobago	31,481	0.32
Nicaragua	86,233	0.87	Other USSR	30,609	0.31
Brazil	84,949	0.86	Slovenia	30,062	0.30
Canada	78,860	0.79	TOTAL	9,920,602	100

Table 2.16: Country of origin of predicted unauthorized immigrants : January 2011

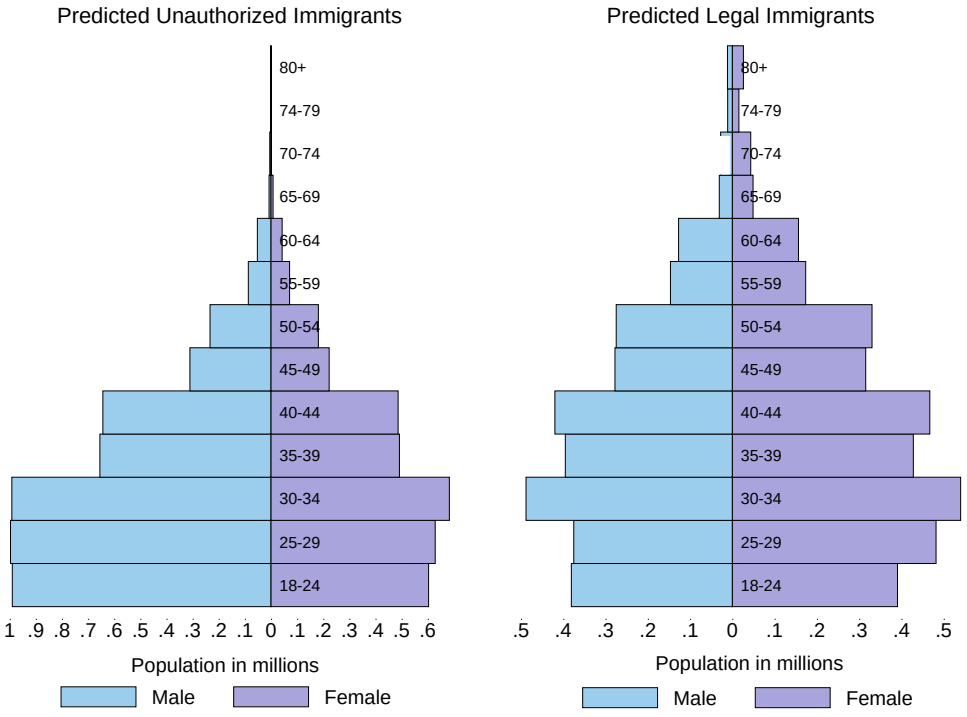

Figure 2.3: Population pyramids of predicted legal and predicted unauthorized immigrants : January 2011

Figure 2.3 depicts population pyramids for predicted legal and unauthorized immigrants. As

evident in the figure the two pyramids are different. The predicted legal immigrants' pyramid is almost symmetric with slightly more females at all ages. On the contrary, the population pyramid of predicted unauthorized immigrants is visibly asymmetric with disproportionate amounts of males between 18 and 54 years of age. This concentration of males is most prominent at younger ages indicating that majority of predicted unauthorized immigrants are working age men. Similar features of high concentration of working age males is consistent with the PHC estimates for 2009 (Passel & Cohn, 2009). This implies that a majority of unauthorized immigrants in the US may be employment related unauthorized immigrants. As such, next I explores employment/occupational characteristics of unauthorized immigrants.

Table 2.17 depicts the distribution of annual wages of predicted unauthorized immigrants as of January 1, 2011. The mean of the annual wage income among the unauthorized immigrants was USD 16,973.20, which is USD 9,897 lower than the mean for authorized immigrants. At the mean the predicted unauthorized immigrants have a smaller standard deviation than the predicted authorized immigrants. Upto the 10^{th} percentile both predicted authorized and unauthorized immigrants had zero income, while at the 25^{th} percentile predicted unauthorized immigrants earned upto UDS 700 per year, while predicted unauthorized immigrants faired worse with zero income. Similarly, at the 50^{th} percentile unauthorized immigrants earned up to USD 14,000 per year, which is USD 4,000 greater than the wage income of the 50^{th} percentile of predicted authorized immigrants. However, at higher percentiles predicted authorized immigrants over take unauthorized immigrants. For instance, when 75 percent of unauthorized immigrants earned less than USD 24,000 per year, the bottom 75^{th} percent of predicted authorized immigrants earned up to USD 35,900, which is nearly 50 percent higher that the predicted unauthorized immigrants' wage income. Moreover, 90 percent of predicted unauthorized immigrants earned less than USD 36,400, which is closer to the income earned by the 75^{th} percentile of predicted authorized immigrants. This wage gap is further accentuated at higher percentiles, where wages of 99 percent of predicted unauthorized immigrants was less than USD 90,000, when authorized immigrants' wages were less than USD 200,000. As such, at the 95^{th} and 99^{th} percentile the unauthorized immigrants earned less than half of the respective income earned by their predicted authorized counterparts.

In order to further explore if unauthorized immigrants are paid less than authorized immigrants Table 2.18 depicts mean annual wage income for the top 5 occupational categories among unau-

Annual Wage Income	Predicted unauthorized	Predicted authorized
Mean	16,973	26,870
Std. Dev.	22,531	46,265
Percentile		
1%	0	0
5%	0	0
10%	0	0
25%	700	0
50%	14,000	10,000
75%	24,000	35,900
90%	36,400	76,000
95%	49,000	100,000
99%	90,000	200,000
Observations	9,923,758	6,915,039

Table 2.17: Distribution of annual wage income of predicted unauthorized immigrants : January 2011

thorized immigrants (see Table 2.19 for the top 25 occupations among unauthorized immigrants). For all five occupations reported in Table 2.18 the mean wages and the standard deviations for predicted unauthorized immigrants are lower than that of predicted authorized immigrants. Among those employed as cooks, at the 25^{th} percentile, the wages of predicted unauthorized immigrants was higher than the 25th percentile of predicted authorized immigrants, but this phenomenon is reversed thereafter. A similar relationship is seen for unauthorized immigrants occupied as farmers. In the case of housekeepers, maids, butlers, stewards etc., there is a gap between the two immigrant groups at all percentiles of wages, where predicted unauthorized immigrants earned less than their authorized counterparts and both groups reported zero income below the 50^{th} percentile. In the case of construction laborers an interesting scenario is depicted where upto the 50^{th} percentile predicted unauthorized immigrants earned more than predicted authorized immigrants, and at the 50^{th} percentile both groups earned equal income. Thereafter, the authorized immigrants' wages surpass that of predicted unauthorized immigrants. Similar to housekeepers etc., among gardeners the predicted unauthorized immigrants earned less annual wages than predicted authorized immi-

Occupation	Mean	Std.Dev	Obs
Predicted unauthorized			
Cooks, variously defined	17,072.47	12,166.95	604,585
Housekeepers, maids, butlers, stewards,	8,364.93	10,123.02	506,592
Construction laborers	18,176.93	16,570.27	486,705
Gardeners and groundskeepers	14,092.75	13,300.82	397,657
Farm workers	14,578.68	10,757.38	385,630
Janitor	15,127.35	12,018.33	370,007
Predicted authorized			
Cooks, variously defined	18743.80	15,497.76	157,690
Housekeepers, maids, butlers, stewards,	12189.89	13,279.60	124,735
Construction laborers	19854.78	20,647.72	78,602
Gardeners and groundskeepers	15879.25	12,882.31	79,323
Farm workers	15455.97	11,706.71	51,872
Janitor	17368.03	14,381.02	101,905

Table 2.18: Distribution of annual wages for top 6 occupations for predicted unauthorized immigrants : January 2011

grants. The unauthorized immigrants occupied as janitors earned more than predicted authorized immigrants at the 25^{th} percentile and beyond the 50^{th} percentile predicted unauthorized immigrants earned less than the predicted authorized immigrants.

Table 2.20 reflects annual family income for non-citizen immigrant families with at least one predicted unauthorized immigrant. For comparison, the table also reports the annual family income of families without any predicted unauthorized immigrants. The mean annual family income among households with at least one predicted unauthorized immigrant is USD 39,741.82, which is USD 29,525.21 less than that of households with no predicted unauthorized immigrants. At very low levels of family income there are interesting differences between families with and without predicted unauthorized immigrants. Among the families with predicted unauthorized immigrants the bottom 5 percent earned an annual family income of less than USD 1,300, and families of predicted authorized immigrants faired worse with the bottom 5 percent earning no income. At the 10^{th} percentile families with predicted unauthorized immigrants earned less than USD 7,400 a

Occupation 1990 basis	Predicted unauthorized			Predicted authorized		
	All	Male	Female	All	Male	Female
Cooks, variously defined	6.09	6.52	5.47	2.07	2.96	1.30
Housekeepers, maids, butlers, stewards,	5.10	0.78	11.39	1.64	0.56	2.58
Construction laborers	4.90	8.15	0.20	1.03	2.20	0.02
Gardeners and groundskeepers	4.01	6.56	0.29	1.04	2.17	0.05
Farm workers	3.89	5.08	2.15	0.68	1.13	0.29
Janitors	3.73	3.44	4.14	1.34	1.74	0.98
Cashiers	2.72	1.52	4.46	1.65	1.30	1.94
Truck, delivery, and tractor drivers	2.65	4.32	0.22	1.21	2.46	0.11
Carpenters	2.61	4.36	0.07	0.66	1.39	0.02
Waiter/waitress	2.24	1.64	3.11	0.91	0.59	1.19
Misc food prep workers	2.19	2.17	2.21	0.85	1.00	0.71
Laborers outside construction	2.04	2.76	1.00	0.78	1.30	0.33
Nursing aides, orderlies, and attendant	2.00	0.48	4.20	1.81	0.70	2.78
Painters, construction and maintenance	1.94	3.20	0.10	0.49	1.01	0.03
Machine operators, n.e.c.	1.74	1.98	1.39	0.72	0.93	0.53
Retail sales clerks	1.60	1.21	2.16	1.12	1.03	1.20
Child care workers	1.42	0.07	3.39	0.59	0.06	1.05
Packers and packagers by hand	1.32	0.80	2.09	0.50	0.46	0.54
Supervisors and proprietors of sales jobs	1.30	1.33	1.26	1.48	1.98	1.03
Assemblers of electrical equipment	1.19	1.18	1.21	0.65	0.72	0.58
Waiter's assistant	1.14	1.31	0.90	0.42	0.40	0.43
Managers and administrators, n.e.c.	1.10	1.38	0.69	2.97	4.79	1.38
Masons, tilers, and carpet installers	1.02	1.72	0.01	0.27	0.56	0.01
Roofers and slaters	0.94	1.58	0.02	0.14	0.30	
Stock and inventory clerks	0.90	0.92	0.87	0.76	1.02	0.53

Note: All numbers in percentages

Table 2.19: Top 25 occupations for predicted unauthorized immigrants : January 2011.

year, which is USD 2,400 better than the family income of the bottom 10 percent of predicted authorized immigrants. Given that unauthorized immigrants are a hidden population often operating in the underground economy, this may be an indication that access to the underground economy provides these low income earning predicted unauthorized immigrants to have better income earning opportunities than the legally residing immigrant population. However, at higher percentiles, as expected families of predicted authorized immigrants are considerably better off than the other group. At the 25^{th} percentile family income of households with predicted authorized immigrants was a 42 percent greater than those with at least one predicted unauthorized immigrant in the household. This gap is progressively increased at all subsequent percentiles, where at the 99^{th} percentile, families of predicted authorized immigrants earned upto USD 382,000 while families of predicted unauthorized immigrant reported annual family income less than USD 187,500, which is approximately half of the former. This family level income gap between families with and without predicted unauthorized immigrants have far reaching implications spreading beyond the unauthorized immigrants themselves and affecting health and wellbeing of all family members including children of unauthorized immigrants.

In terms of health insurance coverage, 73 percent of unauthorized immigrants do not have any insurance coverage, where as among the predicted authorized immigrants only 41 percent did not have insurance coverage. Such a vast difference between the two groups indicate the high possibility that unauthorized immigrants are trapped in a vicious cycle of poor health and poverty, where low income deprives them from access to health insurance, and in-access to health insurance leads to higher health care costs, and resulting in poor health, which leads to lower productivity and/or fewer days to earn income, which completes the vicious cycle. Hence, families of unauthorized immigrants have very few opportunities to break away from the continuation of this vicious cycle.

Among the 2,705,421 predicted unauthorized immigrants with insurance coverage, 86 percent had health insurance through employer/union, while 11 percent had purchased health insurance directly, and 3 percent had both these types of insurance coverage. As a percentage of total predicted unauthorized immigrants, those with health insurance through employer/union accounted for 23 percent, while those with health insurance purchased directly accounted for 3 percent and another 1 percent accounted for health insurance coverage through both the above channels. On the contrary, among the predicted authorized immigrants 47 percent had health insurance through

Annual Family Income	Predicted unauthorized	Predicted authorized
Mean	39,741.82	69,267.03
Std. Dev.	41,544.80	76,316.40
No. of Households	4,162,345	5,746,576
Percentiles		
5	1,300	0
10	7,400	5,000
25	15,400	22,000
50	29,000	50,000
75	51,000	91,000
90	81,800	149,520
95	107,800	198,040
99	187,500	382,000

Note: Number of households arrived using household weights.

Table 2.20: Annual family income of families with and without predicted unauthorized immigrants : January 2011

employer/union, which is more than double the share of the predicted unauthorized immigrants.

Among the predicted unauthorized immigrants 74 percent (7,392,630) were employed (see Table 2.22), and among the employed only 31 percent had health insurance through employer/union. On the contrary, among the predicted authorized immigrants, a smaller share was employed (58 percent), but a greater proportion of those employed, specifically 88 percent, were covered with health insurance provided through union or employer. The unfavorable differences in the employment context experienced by predicted unauthorized immigrants, such as the lower share of health insurance provided through employment and lower average annual wages suggests that the types or conditions of employment contracted by the two groups are different. Perhaps most of the employment opportunities secured by predicted unauthorized immigrants may be in the underground economy, with little or no bargaining power to the employee.enced by predicted unauthorized immigrants, such as the lower share of health insurance provided through employment and lower average annual wages suggests that the types or conditions of employment contracted by the two groups are different. Perhaps most of the employment opportunities secured by predicted unauthorized immigrants may be in the underground economy, with little or no bargaining power to the employee.

Heath Insurance Coverage	Number	%
Predicted unauthorized immigrants		
No health insurance coverage	7,218,337	73
Has health insurance	2,705,421	27
Total	9,923,758	100
–Has health insurance purchased privately	286,781	11
–Has health insurance through employer/union	2,331,349	86
–Has both -through employer and purchased directly	87,291	3
Predicted authorized immigrants		
No health insurance coverage	3,140,624	41
Has health insurance coverage	4,468,875	59
Total	7,609,499	100
–Has health insurance purchased privately	744,580	17
–Has health insurance through employer/union	3,553,634	80
–Has both -through employer and purchased directly	170,661	4

Table 2.21: Insurance coverage of predicted legal and predicted unauthorized immigrants : January 2011

Employment status	Predicted unauthorized		Predicted authorized	
	Number	%	Number	%
Employed	7,392,630	74	4,437,798	58
Unemployed	824,543	8	536,559	7
Not in labor force	1,706,585	17	2,635,142	35
–attending school	235,076		585,995	
–not attending school	1,471,509		2,049,147	
Total	9,923,758		7,609,499	

Note: Percentages may not add to zero due to rounding

Table 2.22: Labor market status of predicted legal and predicted unauthorized immigrants : January 2011

Education	Freq.	Percent
Grade 5, 6, 7 and 8	6,432	2.74
Grade 9	6,170	2.62
Grade 10	8,397	3.57
Grade 11	37,573	15.98
Grade 12	56,236	23.92
1 year of college	78,967	33.59
2 years of college	14,352	6.11
4 years of college	21,166	9.00
5+ years of college	5,783	2.46
Total	235,076	100

Table 2.23: Educational attainment of predicted unauthorized immigrants (not in the labor force) : January 2011

As seen in Table 2.22, 17 percent of predicted unauthorized immigrants were not in the labor force. Out of this 1,706,585 a 14 percent was attending school and their educational attainments are presented in Table 2.23. Among the predicted unauthorized immigrants attending school 51 percent were in college, despite their predicted unauthorized status. Among the predicted unauthorized immigrants attending school, 90 percent were under the age of 32 years, and 12 percent was 24 years or younger. Perhaps some of these younger cohorts of predicted unauthorized immigrants were children of predicted unauthorized immigrants.

2.8 Characteristics of children of unauthorized immigrants: January 2011

The microdata-based methodology is applicable only in the context of individuals 18 years or older. However, children of unauthorized immigrants can be identified by linkings children to their parents. The microdata-based methodology estimates 6,079,997 children (under 18 years of age) with at least one unauthorized immigrant parent for January 1, 2011.[26]

[26]

i. This section is based on assignment based characteristic estimates which is different from sample weight based estimates. The former is chosen for comparison between existing and new estimates.

ii. An adult individual is considered an unauthorized immigrant parent when an appropriate adult in the same

Citizenship Status	Place of Birth	Number	Number (using rounded person weights)
Citizens by birth	US	5,094,974	5,096,741
Citizens by birth	Abroad-parent US citi.	34,649	34,661
Naturalized citizen	Foreign born	34,156	34,174
Not naturalized citizens	Foreign born	916,219	916,541
Total		5,966,273	6,082,117

Table 2.24: Children of predicted unauthorized immigrants : January 2011

Out of this 6,079,997 children 5,094,974 children were born in US (see Table 2.24), who are US citizens by birth. A small share of 0.6 percent (34,649 children) were born abroad to an American parents which in turn makes these children also American. Another 34,156 were naturalized citizens with at least one predicted unauthorized immigrant parent, while the remaining 916,219 children were foreign born and not naturalized. This group of 916,219 children are most likely to be unauthorized immigrants themselves. The proportion of US born children of unauthorized immigrant parents[27] is 84 percent, which is comparable to the PHC estimate of 82 percent for 2010 (Passel & Cohn, 2011). The slight differences in microdata-based estimate and the PHC estimate can be attributed to differences in the methodologies, the differences in data used (ACS in microdata-based estimates and CPS data used by the PHC methodologies), the difference in reference periods, and the choice of cutoff probability to assign unauthorized immigrant status to parents.

The last column in Table 2.24 indicates the estimates arrived by rounding off the person weights of the children of unauthorized immigrants to the nearest integer. The subsequent characteristics of children of unauthorized immigrants –except for Table 2.25 and related discussion, are based on the numbers in the last column in Table 2.24 .

As shown in Table 2.25 among the US born children of unauthorized immigrants 32 percent had both parents unauthorized, while another 32 percent had an unauthorized immigrant mother and 36 percent had an unauthorized immigrant father. Among the foreign born non-citizen children

family unit within the household can be unambiguously linked to a child as a parent, and has a predicted probability less than .7929.

[27]Children of unuthorized immigrants refers to children with at least one unauthorized immigrant parent.

Status	US born children	FB children	Total
At least 1 unauthorized immigrant parent	5,096,741	916,541	6,013,282
Mother unauthorized immigrant	3,260,707	688,152	3,948,859
Father unauthorized immigrant	3,471,743	632,222	4,103,965
Both parents unauthorized immigrants	1,637,477	404,155	2,041,632

Table 2.25: Predicted unauthorized immigrant parents : January 2011

of unauthorized immigrants 31 percent had an unauthorized immigrant mother and 25 percent had an unauthorized immigrant father, while a considerably larger share (44 percent) were of two unauthorized immigrant parents.

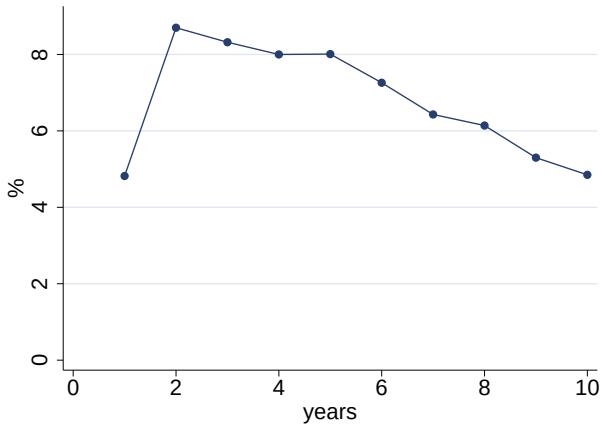

Figure 2.4: Mothers' years in the US before their child's birth in US : January 2011

Out of the 3,261,862 US born children of unauthorized immigrant mothers[28], 4.82 percent had their mothers arrived in the US within one year prior to the birth of the child in the US. On the contrary, a significantly larger 8.70 percent mothers arrived 2 years before the birth of her child in the US. As seen in Figure 2.4 the peak occurred at 2 years before a child birth in the US. Mother's arrival before 2 years of child birth infers that the mother arrived approximately 1 year before the conception of the child. This is an indication that the birth of these children may be part of the unauthorized immigration plan and thus may have served as Anchor babies.[29] The relatively low

[28]This number is arrived by using person weights rounded to the nearest integer.

[29]The term "anchor babies" refers to children born in the United States to unauthorized immigrant parents. It is believed that unauthorized immigrants give birth to children in the US as "anchors" to prevent their own deportation.

share of 4.82 percent associated with 1 year before child birth reflects that only a smaller number of women choose to unlawfully immigrate to the US while pregnant with a child. Similarly, 4.85 percent associated with mothers who arrived 10 years ahead of child birth implies that the child birth was not part of the unauthorized immigration plan.

The school attendance among the children with and without predicted unauthorized immigrant parents show some interesting differences. For instance, the share of children not attending school was higher among children with at least one predicted unauthorized immigrant parent. Although this difference is less than one percentage point when all school age children are considered together, the difference is significant when children of 17 years of age are considered separately. As shown in Table 2.26 among the 17 year old children, 6.1 percent of children with at least one predicted unauthorized immigrant parent are not in school, while only 3.4 percent of children of predicted immigrant parents were not in school. This notable gap in school attendance indicates that there may be a greater high school drop-out rate among children of predicted unauthorized immigrant parents.

Moreover, among children with at least one unauthorized immigrant parent, there is a vast difference in school attendance at higher ages between the US and foreign born children. As seen in Table 2.27 the proportion of children not attending school is consistently higher among 15 to 17 year olds who are non-citizen and foreign-born. Among the 17 year old not citizen-foreign born children of predicted unauthorized immigrants, nearly 10 percent are out of school, which is more than double the rate for the US born children of predicted unauthorized immigrants. These not citizen-foreign born children of predicted unauthorized immigrants are most likely to be unauthorized immigrants themselves. As such, this indicates that a child's citizenship status and school attendance may be correlated, which would serve invaluable in evaluating the policy implications of the proposed DREAM Act, for this Act proposes to provide conditional permanent residency to certain unauthorized aliens of good moral character who graduate from US high schools, arrived in the US as minors, and lived in the US continuously for at least five years prior to the bill's enactment.

Table 2.28 compares educational attainments of 17 year old children of predicted unauthorized and predicted authorized immigrant parents. Ideally a 17 year old child should be in the 11th or 12th grade depending on his birthday. As evident in column (ii) in Table 2.28, when approximately

Age in yrs	In school	Not in school	Obs.
	Parents predicted unauthorized		
5 to 17 yrs	96.8	3.2	4,235,662
15	97.4	2.6	277,110
16	97.6	2.4	289,044
17	93.9	6.1	258,066
	Parents predicted authorized		
5 to 17 yrs	97.3	2.7	30,530,593
15	98.4	1.6	2,387,592
16	97.9	2.1	2,404,834
17	96.6	3.4	2,329,383

Table 2.26: School attendance by children of predicted unauthorized immigrants : January 2011

	US born children*			Foreign born children*		
Age in yrs	In school	Not in school	Obs.	In school	Not in school	Obs.
5 to 17	97.21	2.79	3,313,544	95.16	4.84	860,992
15	99.06	0.94	177,551	94.59	5.41	92,719
16	98.21	1.79	184,444	96.23	3.77	96,200
17	95.83	4.17	163,295	90.11	9.89	88,651

* All children with at least one unauthorized immigrant parent.

Table 2.27: School attendance by birth place : January 2011

Educational attainment	Parents predicted authorized	Parents predicted unauthorized		
		Total	US born	Foreign Born
	(i)	(ii)	(iii)	(iv)
N/A or no schooling	0.12	0.29	0.12	0.60
Nursery school to grade 4	0.00	0.03	0.06	0.00
Grade 5, 6, 7, or 8	0.84	2.01	1.34	3.29
Grade 9	3.06	8.51	6.55	12.45
Grade 10	27.53	30.88	30.52	31.89
Grade 11	58.58	48.49	51.69	41.96
Grade 12	9.47	9.23	9.11	9.49
1 year of college	0.56	0.62	0.43	0.33
2 years of college	0.04	0.00	0.00	0.00
Observation	2,557,761	256,119	164,060	85,097

Note: values in percentages.

Table 2.28: Educational attainments of 17 year old children : January 2011

80 percent of children of predicted authorized immigrants are in this ideal educational attainment threshold, only 73 percent of foreign born children of predicted unauthorized immigrants are in the 11th or 12th grade. At lower grades there is a consistent pattern of relatively larger shares of children of predicted unauthorized immigrant parents (column (ii)) than children of predicted authorized parents (column (i)) indicating that a larger share of 17 year old children of unauthorized immigrants are in lower than the ideal grade. However, when 0.6 percent of children of predicted authorized immigrants are in college, there is a slighly larger proportion of children of predicted unauthorized immigrants in college. Nonetheless, some of these students with predicted authorized immigrant parents are already in the 2nd year in college, while none of the children of predicted unauthorized immigrants had reached the 2nd year in college. This indicates that there may be disparities in terms of educational attainments among children of authorized and unauthorized immigrants.

In addition to education, another significant difference is seen in health insurance coverage between children with and without at least one unauthorized immigrant parent. As shown in Table 2.29 a disproportionate number (22.6 percent) of children with at least one unauthorized

Health insurance coverage	Parents Predicted authorized	Parents predicted unauthorized		
		Total	US born	Foreign born
	(i)	(ii)	(iii)	(iv)
Not covered	5.49	22.6	16.76	54.8
Covered	94.51	77.4	83.24	45.2
Observations	41,225,162	6,082,117	5,096,741	916,541

Table 2.29: Health insurance coverage of children of predicted unauthorized immigrants : January 2011

immigrant parent did not have any insurance coverage, while only 5.49 percent of children of predicted authorized immigrants did not have any health insurance.

The comparison between the US born and non-citizen-foreign-born children of predicted unauthorized immigrants in columns (iii) and (iv) of Table 2.29 show that 54.8 percent of the foreign born children did not have health insurance, while only a smaller share of 16.76 percent of US born children did not have health insurance.

The above analysis of the characteristics of children of unauthorized immigrants, implies that children of unauthorized immigrants are less likely to have access to health services, more likely to drop out of school and are worse off than children of legal immigrants. Moreover, among children of unauthorized immigrants there are noticeable differences in health insurance coverage and in educational attainment among native and foreign born children. Most likely these disparities will be carried over into their adulthood where they may experience unfavorable opportunities in the labor market as well.

2.9 Policy relevance

Policies concerning unauthorized immigrants can be separated into 2 segments; ex ante and ex post. Ex ante policies address unauthorized immigration before it takes place and ex post policies address unauthorized immigration that have already taken place. The microdata-based method-

ology can be used in ex ante policy formulation by analyzing the resulting cross sectional data to discern the characteristics of unauthorized immigrants in the US, and thereby collaborating with governments in high sending countries to improve the socioeconomic conditions of those with a high propensity to become an unauthorized immigrant in the US. Moreover, statistics indicates that unauthorized immigrants are commonly employed as unskilled labor. The main reason for these unskilled unauthorized immigrants to newly enter or remain in US is the high propensity to find employment in the US. This implies that the US employers continue to recruit unauthorized immigrants, due to cheaper labor costs. However, as noted by Hanson (2009, p. 8) "US visa programs are simply not designed to accommodate the changing demands of US industry. Under current policies, if businesses want to hire additional low-skilled foreign workers, their primary option is to employ unauthorized immigrants." Yet, the US immigration policy does not acknowledge this mismatch between demand and labor supply for unskilled labor. The microdata resulting from this methodology can used to analyze the skill sets of unauthorized immigrants, characteristics of their occupations as well as the characteristics of their employers at the national or state level to better understand this mismatch in the labor market and formulate more responsive labor market and immigration policies in the US to discourage unauthorized immigration.

In terms of ex post policies, the micro data from this methodology can be effective in designing policies and programs to mitigate the competition for low skilled jobs among natives and unauthorized immigrants, and to address possible exploitation of unauthorized immigrant labor. Moreover despite the fact that "most households headed by unauthorized immigrants are poor, they make minimal use of Temporary Assistance for Needy Families, Supplemental Security Income, energy assistance, housing subsidies, or other welfare programs" (Hanson, 2009, p. 11). It may be that unauthorized immigrants themselves contribute to government revenue more than they add a burden on government expenditure. However, unauthorized immigrants do draw on public expenditure in other ways, especially through their children, who may attend public schools and, if they are native-born, receive Medicaid and participate in school breakfast and lunch programs (Hanson, 2009). These fiscal implications can be studied using the microdata of this methodology by identifying families/households headed by an unauthorized immigrant. Additionally, from the perspective of children of unauthorized immigrants, the identification of parents who are predicted unauthorized immigrants can lead to analyzing many related aspects of this population such as

poverty among children of predicted unauthorized immigrants, differences in educational attainments among US born and foreign born children of predicted unauthorized immigrants, the role of anchor babies in the legal status dynamics of predicted unauthorized immigrant parents, and the benefit/cost analysis of the proposed DREAM Act. The policy relevance of this methodology is not limited to the few policy areas outlined as examples in this section. The cross sectional dataset resulting from the microdata-based methodology can be relevant in any policy setting where legal status of immigrants is involved.

2.10 Concluding remarks

As stated by Rivera-Batiz (1999) greatest barrier to the analysis of unauthorized immigrants is their hidden nature and the corresponding absence of new data. The approach taken in this microdata-based methodology to exploit available information on authorized immigrants compensates for this gap in data. As such, the methodology developed here can be summarized as a prediction model based on hard evidence of actual authorized and unauthorized immigrants in 1986 that is adjusted for calibration and time dynamics so that the results are valid for the current context. The most striking features of this methodology is the resultant cross-sectional dataset, which is based on parameter estimates. The cross sectional data has flexibility for the researcher to define the cutoff point and the margin of error as per the research question at hand. This methodology also proves the unrealistic nature of the fundamental assumption of existing estimates, namely that characteristics of unauthorized immigrants have remained constant since 1986.

The comparison of existing residual methodology with the new microdata-based methodology highlights that the former does not include adjustments for time dynamics nor calibration. Moreover, the residual method does not have any information of the margin of error in term of mis-classification, nor the ability to control this margin of error. The age adjusted DHS estimate for the previous year (2010) replicated using the microdata-based methodology shows that the margin of error in misclassification in existing estimates are beyond 37 percent.[30] A similar estimate produced by the microdata-based methodology limits the error component to 33.7 percent. Moreover, at any margin of error the naive estimate (which is comparable to existing estimates) underestimate

[30]The actual margin of error of this DHS estimate could be even higher, once the error corresponding to all ages is taken into account.

the microdata-based estimate. For January 1, 2011, I estimate an *adult* unauthorized immigrant population of 7,700,869. This sample weight based estimate is different from the assignment based estimate of 9,920,602 using the cutoff probability of 0.7929, which is associated with a possible mis-classification rate of 33.7 percent.

Finally, the microdata-based methodology developed in this paper has great potential to be applied in many different ways and contexts. In terms of future directions of the methodology, first I intend to produce estimates for the remaining years of the ACS. Such estimates for several years with a constant cutoff probability would enable me to analyze trends in predicted unauthorized immigrants. Second, I intend to apply this methodology in the context of CPS data, so that the micro data estimates can be directly compared against estimates produced by the PHC. Finally, I intend to do an in-depth analysis of children of unauthorized immigrants, which so far remains an under-researched area in the field of unauthorized immigration literature.

Appendices

2.A Making INS data comparable to LPS data

i. The LPS captures only those born before February 1, 1971 , which includes those born in January 1971. Since the INS data reflects only the year of birth, those born in January 1971 are dropped from the LPS sample to make LPS birth cohorts compatible with INS birth cohorts.

ii. In INS data, the year of entry corresponding to adjustments is the year of most recent non-immigrant admission prior to adjusting status into LPR. Hence, LPS is augmented with only those who were born before 1971 and arrived between 1972 and 1982 and adjusted their status between 1972-1986. In the case of new admissions year of entry is same as year of obtaining LPR, and the LPS is augmented with all new admissions who were born before 1971 and entered between 1972 and 1982.

2.B Variables in derivation model

- **Age** : Age in 1986.
- **Sex** : 1 = male and 2 = female.
- **Marital status** : 1 = Single, 2 = Married, 3 = Widowed, 4 = Divorced /Separated
- **State of residency (in 1986)**: The state of residency for observations from the LPS pertains to the state of residence at the time of applying for amnesty (on LAPS prior to LPS1). Information for this variable for observations from the INS is gathered from the 'INS office having jurisdiction over immigrants intended place of residency'. The rationale for this variable is that areas where authorized immigrants live differ from where unauthorized immigrants reside.
- **Occupation** : For observations from the LPS, the occupation corresponds to the occupation at the time of amnesty application. In the case of observations from INS data, occupation

related variable provides two types of information. For those obtaining LPR through employment based preferences, this variable refers the job they will perform in US, while for other, this refers to the last job in country of origin. Here I assume that the occupation in US will be in line with the occupation held at the country of origin. The occupation codes in INS and LPS datasets are made consistent and arrived at the 28 occupation codes as per the 1983 INS file. In estimation of the base model these 28 occupation codes are further condensed into 13 occupation cluster categories as follows:

- occupation 1 : Executive, administrative, and managerial occupations.
- occupation 2 : Lawyers, medical doctors,engineers, surveyors, mapping scientists, mathematical and computer scientists, natural scientists, social scientists and urban planners.
- occupation 3 : Other health assessment and treating occupations, health technologists and technicians.
- occupation 4 : Teachers, postsecondary and secondary, councellors educational and vocational, librarians, archivists, and curators.
- occupation 5 : Social, recreation, and religious workers.
- occupation 6 : Writers, artists, entertainers, and athletes.
- occupation 7 : Technologists and technicians, except health.
- occupation 8 : Sales occupations.
- occupation 9 : Administrative support occupations, including clerical.
- occupation 10 : Service occupations.
- occupation 11 : Farming, forestry, and fishing occupations.
- occupation 12 : Precision production, craft, and repair occupations, and operators, fabricators, and laborers.
- occupation 13 : Homemakers, unemployed or retired, students and/or children under age 16 , occupation not reported, not employed in home country.

2.C Summary of microdata-based approach

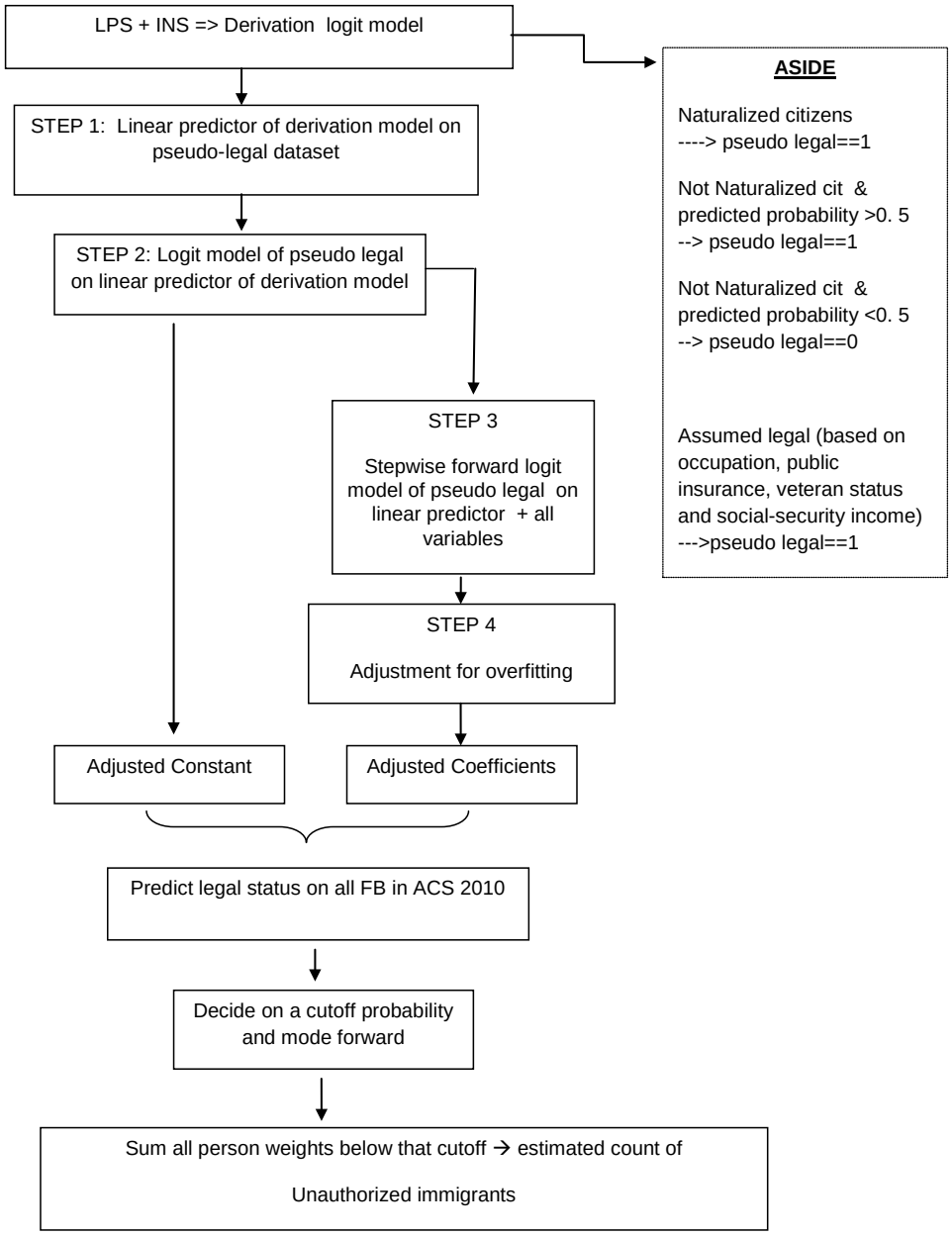

Figure 2.5: Summary of microdata-based approach

Chapter 3

Are Children of Unauthorized Immigrants More Likely to Drop out of High School?

3.1 Motivation

Children are an important component of the future of an economy. The US Census Bureau estimates show that 24 percent of the US population are under 18 years of age (2010 American Community Survey) and this consists of children of native and foreign born parents. Among the foreign born parents some are legally[1] residing immigrants and the others' presence in the United States is unauthorized[2]. Nonetheless, some children of unauthorized immigrants are born in the US and thus they are US citizens. Regardless of parents' legal status, the future of these children of immigrants and the future of the US economy are intertwined and thus remain the responsibility of the US. As such, their current educational performance and future labor market performance need to be steered in a manner that will benefit both the economy at the macro level and the individual child at the micro level.

The literature has documented that poverty rates among immigrant children, irrespective of parents' legal status, are far higher than those among children with non-immigrant parents (Rector,

[1] Also referred to as authorized or documented immigrants.
[2] Also referred to as undocumented or illegal immigrants.

2006). The risk of poverty is even higher among families of unauthorized immigrants. For instance, "a third of the children of unauthorized immigrants and a fifth of adult unauthorized immigrants live in poverty. This is nearly double the poverty rate for children of US-born parents (18%) or for US-born adults (10%)" (Passel & Cohn, 2009). Despite the presence of a clear understanding of the socio-economic vulnerability of children of unauthorized immigrants, there is a near vacuum of rigorous economic analysis focusing on this population. The little analysis available only touches the surface through mere descriptive statistics. In this context, this paper takes the first step in providing an econometric analysis of children of unauthorized immigrants. Specifically, this paper focuses on the nexus between mother's legal status and a child's likelihood of dropping out of high school. Card (1999, p. 1802) shows that there is a plethora of literature that "confirm that better-educated individuals earn higher wages, experience less unemployment, and work in more prestigious occupations than their less-educated counterparts". In this context, understanding the impact of parents' legal status on a child's educational outcomes will be important for understanding the future labor market performance of children of unauthorized immigrants.

The analysis in this paper uses data from the American Community Survey (ACS 2010) in conjunction with the "microdata-based methodology" to identify unauthorized immigrants developed Chapter 2. Using a linear probability model on multiple sub-samples, this study finds that when controlling for other covariates children of two unauthorized immigrant are less likely to drop out of high school than similar children of two authorized immigrant parents, and that among two unauthorized immigrant parents, non-citizen[3] children are more likely to drop out of high school than similar native-born children. Additionally, the study also finds that in mixed families with unauthorized immigrant mothers, non-naturalized citizen children are more likely to drop out of high school than otherwise similar native-born children, and that the parent combination of highly likely unauthorized and highly legal results in a higher high school drop out probability than the parent combination of highly unauthorized and unsure legal.

The remainder of the paper is organized as follows. Section 3.2 reviews relevant empirical literature, while Section 3.3 describes the data and constructs the variables. Section 3.4 formulates the model , while Section 3.5 presents the results. Section 3.6 puts the findings of the study into perspective, followed by concluding remarks in Section 3.7.

[3]Also referred to as non-naturalized citizen.

3.2 Literature Review

As mentioned, there is no economic literature that performs an econometric analysis of children of unauthorized immigrants. Nonetheless, there are three strands of literature that are tangentially related to the scope of this paper. First is the literature dealing with children of immigrants, with no distinction about their or their parents' legal status in the US, and second is the literature dealing with dropping out of high school without differentiating between immigrant status of these children or their parents. Third, there are a few studies that focus on unauthorized youth. Specifically, Jefferies (2008) discusses the educational opportunities for undocumented youth. This study does not involve any empirical analysis but limits its scope to a brief discussion. The author notes that undocumented children are guaranteed access to K-12 education as a result of the Supreme Court decision, but there is no such pathway for higher education, and that "society for them [undocumented students] is not a fair and equal playing ground" for "structural factors, such as social and cultural advantages, unequal educational opportunities, and discrimination in all of its forms, are barriers to this success" Jefferies (2008, p. 251). Shields & Behrman (2004) perform an analysis of children of immigrant families, again using a descriptive methodology. They find that some of the strengths of immigrants families are their health, intact families, and their strong work ethics and aspirations. On the other hand, the authors also note that these immigrant families are challenged with low parental educational attainment, low wage work with no benefits, language barriers, discrimination and racism, and poverty and lack of support. The authors make two important observations which are relevant to the current paper. First, US born children of undocumented immigrants are not getting government assistance despite their eligibility due to fears of adverse immigration consequences that may face their parents. Second, though "the vast majority of teens in immigrant families attend school, they are more likely than those in US-born families to be behind grade and not to graduate"(Shields & Behrman, 2004, p. 11).

Unlike the previously reviewed group of literature on unauthorized youth, the literature dealing with children of immigrants is methodologically richer. One such paper is by Driscoll (1999), where the relationship between immigrant generation and high school drop out rate among Hispanic students is investigated. This study uses data from National Education Longitudinal Study of 1998, to evaluate the odds of ever dropping out of high school and the odds of early and late

high school drop out. The study finds that, "while the odds of early high school drop out are uniformly high among all generations, net of individual and family resources second generation eighth graders are less likely to drop out any time, and first and second generation sophomores are more likely to complete high school. High educational expectations, family income and past academic performance protect against high school drop out" (Driscoll, 1999, p. 857). Moreover, Driscoll (1999) provides a detailed background about how different segments of the population invest in human capital.

Contrary to Driscoll (1999), Perreira *et al.* (2006) focus on the high school completion of immigrants and native youth using data from the National Longitudinal Study of Adolescent Health (Add Health). The authors find that first-generation youth of Hispanic, Asian, and African heritage obtain more education than their parents, but the second and subsequent generations lose ground. Differences in drop out rates by race-ethnicity and immigrant generation are driven by differences in human, cultural, and social capital. Low levels of family human capital, school social capital, and community social capital place the children of immigrants at risk of dropping out. The most important finding of this paper in the context of the current study is that cultural capital and immigrant optimism buffer first-generation Hispanic youth and the children of Asian immigrants from the risk of dropping out of high school.

Djajic (2003) investigates how the pace of assimilation of immigrants in various dimensions affect the rate of human capital accumulation among children of immigrants. He finds that rapid assimilation in certain dimensions results in increasing the rate of human capital accumulation of the second generation, while other dimensions may have the opposite effect. For instance, differences in customs, values and attitudes between natives and immigrants, such as the higher tendency for immigrant families stay intact without divorce or separation of parents "serves as a stabilizing element that contributes to a better academic performance and economic success of the second generation" (Djajic, 2003, p. 834). Another is the inclusion of grandparents in immigrant households, who "encourage children to spend more time doing homework, to develop strong work habits, and to appreciate more fully the role of their scholastic achievement in the family struggle to attain its social and economic objectives" (Djajic, 2003, p. 835). Dimensions that have a negative effect on human capital accumulation include the quality of housing and neighborhood, which affects educational outcome through inadequate space and noise; and the tightly knit communities

in which immigrants live. Such tight knit communities result in a slower assimilation of immigrants in host country language proficiency, etc.

The strand of literature that deals with dropping out of high school in general has several important works. One of the notable papers in this group is by Angrist & Krueger (1991) who evaluate the causal effect of compulsory schooling laws on schooling and earnings. The authors first establish that season of birth is related to educational attainment because of school start age policy and compulsory school attendance laws, and thus individuals born in the beginning of the year start school at an older age, and can therefore drop out after completing less schooling than individuals born near the end of the year. Angrist & Krueger (1991) use the quarter of birth as an instrument for compulsory schooling and find that approximately 25 percent of potential dropouts remain in school because of compulsory schooling laws. In response to this influential contribution by Angrist & Krueger (1991), Bound *et al.* (1995) argue that compulsory schooling laws are not the only correlation between quarter of birth and educational outcomes. They highlight that the quarter of birth may affect a student's school attendance rates, cause behavioral difficulties and requirements for mental health services and affect performance in reading, writing and arithmetic. The authors also argue that 'there are identifiable differences in physical and mental health of individuals born at different times of the year'. Moreover, the authors argue that that there are clear regional patterns in birth seasonality and the association between family income and quarter of birth. Bound *et al.* (1995) show that in the context of all these *other* channels that link quarter of birth with educational attainment, the findings by Angrist & Krueger (1991) cannot be seen as a causal effect of compulsory schooling laws on educational outcomes.

Eckstein & Wolpin (1999) investigate why youths drop out of high school by looking at how their preferences, opportunities and abilities impact high school drop out. The author uses the 1979 youth cohort of the National Longitudinal Surveys of Labor Market Experience (NLSY79) and employ a combined methodology of dynamic optimization with the maximization of a likelihood function that accounts jointly for annually observed work- schooling choices, wages, credits earned, and grades, to structurally estimate a sequential model of high school attendance and work decisions. The model's estimates imply that youths who drop out of high school have different traits than those who graduate. Specifically, those who drop out have lower school ability and/or motivation, they have lower expectations about the rewards from graduation, they have a comparative advantage

at jobs that are done by non graduates, and they place a higher value on leisure and have a lower consumption value of school attendance. The authors also found that working while in school reduces school performance. However, policy experiments based on the model's estimates indicate that even the most restrictive prohibition on working while attending high school would have only a limited impact on the high school graduation rates of white males.

Evans & Schwab (1995) investigate the impact of Catholic schools on finishing high school and starting college. The authors use data from High School and Beyond for 1980 which were followed up in 1982 and 1984, in a bivariate probit model. The use of a bivariate probit model enables the authors to rule out selection bias encountered in a single equation model and appropriately approach the dichotomous nature of both the outcome and the treatment variable. This study finds evidence to support the argument that Catholic schools are more effective than public schools using both single equation probit models as well as bivariate probit models. The most interesting finding of this study is the near absence of selection bias, which makes the bivariate probit estimates of the average treatment effect of Catholic schools on high school graduation very similar to single equation probit estimates.

Plug & Vijverberg (2005) evaluate how family income affect schooling outcomes. In their analysis they control for parental ability that is genetically transfered by focusing on adopted children. Moreover, the authors also control for biases arising from unobserved parenting qualities and parents' differentiation between their own and adopted children, and find a significant positive income effect. The authors use data from the Wisconsin Longitudinal Survey in a spline regression to address the curvature of the income effect.

However, as evident in the foregone selected literature review, there is no literature examining the educational outcomes of children of unauthorized immigrants. As such, this study is the first to perform any economic analysis beyond a descriptive analysis of children of unauthorized immigrants. In this context, this analysis provides valuable information about the correlation between educational outcomes of children and parents' legal status.

3.3 Data

This study uses data from American Community Survey (ACS) 2010. The sample is restricted to native and non-citizen children[4] in the ages of 16 to 18 years (both years inclusive) with non-citizen parents.

The outcome variable D, is a dichotomous variable which takes the value 1 if the child has not attended school or college at any time during the last three months and takes the value 0 otherwise.

$$\mathrm{D} = \begin{cases} 1 & \text{if dropped out from school} \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

The parents' legal probabilities are predicted as per the methodology developed in Chapter 2. Figure 3.1 depicts the distribution of predicted legal probabilities of non-citizen parents. As evident in the figure, both distributions take interesting shapes. The left panel in Figure 3.1 shows that the predicted legal probabilities of non-citizen mothers of 16 to 18 year old children is bimodal with one concentration around the predicted probability of 0.1 and another noticeable peak towards a probability of 1. Similarly, the graph of fathers too is bimodal, with a more noticeable valley between 0.2 and 0.8.

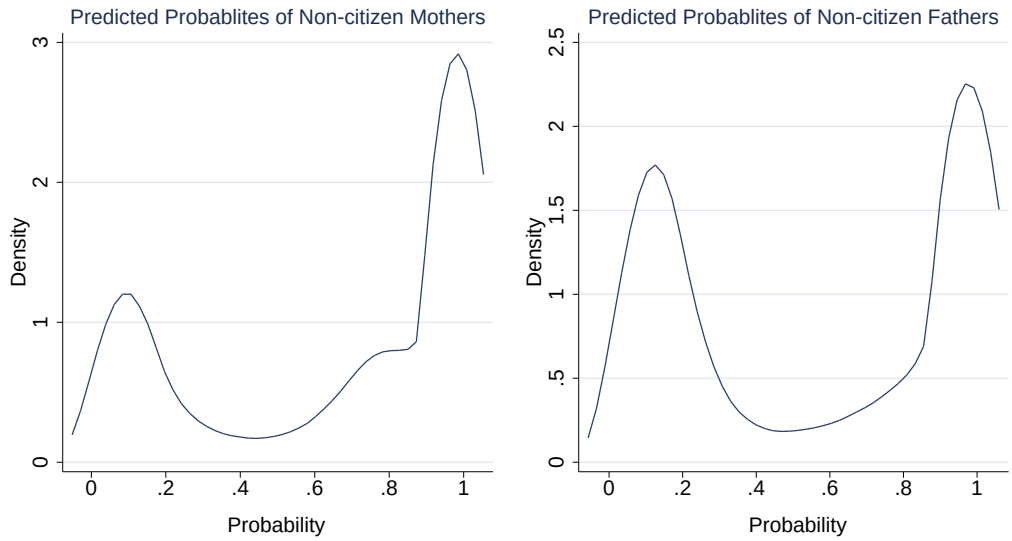

Figure 3.1: Predicted legal probability of non-citizen parents

[4]Other types are dropped due to the negligible share in the total number of children of unauthorized immigrants.

In addition to parents' legal status, this analysis also pays attention to the child's citizenship status in the US. However, among the four possible citizenship statuses, the study focuses only on native and non-naturalized citizen children. These two citizenship groups are chosen because they account for the majority of children of unauthorized immigrants. Moreover, on the one hand, non-citizen children of unauthorized immigrants are often also unauthorized (unauthorized minors in their case). On the other hand, native-born children of unauthorized immigrants are US citizens. As such, sometimes a family may consist both of these extremes in children's citizenship. However, despite similar socio-economic status in the family, one child may have a higher probability to drop out of high school simply due to his citizenship status. In order to capture the impact of child's citizenship status on his probability to drop out of high school, the indicator variable N is included in the model, where N takes the value 1 for native-born children and 0 for non-citizen children.

Studies have identified many important variables that influence the educational performance of children. Driscoll (1999) notes that the language spoken at home and race influence a child's educational performance, while Bound *et al.* (1995) elaborates the connection between quarter of birth and ability that influences educational attainments. In order to account for these other channels that influence dropping out of high school, the vector of child's characteristics includes age and indicator variables for sex, the language spoken at home, the quarter of birth, and race.

The language spoken at home takes the value 1 if the child speaks English at home, 2 if Spanish is spoken at home and 3 if any other language is spoken at home. The quarter of birth variable is self explanatory, while the race takes one of the following 9 categories; (i) white, (ii) black , (iii) American Indian or native Alaskan, (iv) Chinese, (v) Japanese,(vi) other Asian or pacific islander, (vii) other race, (viii) two major races, (ix) three major races.

In addition to child's own characteristics, studies have found that family characteristics, such as family income, family structure and family size, have an impact on children's educational outcomes (Black *et al.*, n.d.; Boggess, 1998; Chiswick *et al.*, 2005; Gennetian, 2005; Plug & Vijverberg, 2005). As such, in addition to child specific characteristics, I also includes the natural log of total family income, indicators variables for father's and mother's presence at home and the number of siblings to control for possible effect of family characteristics on the high school drop out probability. Moreover, in order to capture state level variation in high school drop out probabilities, indicators for the 50 states and Washington D.C. are also included. Other findings in the literature indicate

that parents' educational achievements, occupation and their cultural values and attitudes towards education have implications on children's educational performance. As such, each parent's vector of characteristics in this model includes indicator variables for parent's high school education, parents' occupation, and country of birth. Parents in different occupations have different influences on the child's high school educations. For instance, a child would get greater parental encouragement towards completing high school education if his parent is a high school teacher, than if his parent was a farmer. Similarly, immigrants from different countries have varied attitudes towards education. The inclusion of country of birth is intended to control for such differences in parental influence brought about by parents' country of birth.

3.4 Methodology

The methodology of this study is the linear probability model on multiple sub-samples of the initial sample. The nature of data and the research question together dictate such a multiple sample approach. Specifically, parents' legal probability is a predicted variable using six variables corresponding to parents,[5] many of which have an impact on the child's high school drop out probability. For instance, a parent who is a high school teacher might encourage a child to complete high school more than a parent who is a farmer. As such, some of the variables that are involved in parents' legal status prediction are also involved in the prediction of a child's high school drop out probability. As such, parents' legal status and its determinants cannot be included simultaneously as a predictors of child's high school drop out probability. In this case, I divide the sample into sub-samples based on combinations of parents' legal probabilities as follows: (i) both parents highly likely to be legal immigrants; (ii) both parents highly likely to unauthorized immigrants; and (iii) parents with a combination of the above two identified as "Mixed".

Such division into sub-samples facilitates the estimation of the same linear probability model on each sub-sample, where the difference in the predicted high school drop out probabilities between sub-samples (i) and (ii) for a representative child, allows to discern if children of unauthorized immigrants are more likely to drop out of high school.

The division into sub-samples is based on a critical legal probability P_l to distinguish between

[5]See Chapter 2 for details.

highly legal and highly unauthorized immigrants. Each parent with a predicted probability greater than the critical probability P_l is considered a "highly likely legal immigrant", while a parent below a probability of $(1 - P_l)$ is considered a "highly likely unauthorized immigrant".[6] The exact procedure for creating these sub-samples is outlined below.

First, I set the critical probability P_l at 0.5. Those with a legal probability greater than or equal to 0.5 are classified as legal immigrants, and unauthorized otherwise. At this critical probability where the all parents are classified either highly legal or highly unauthorized, there are 1,634 observations in the sub-sample of highly likely unauthorized parents. The sub-sample of highly likely legal parents is made up of 2,229 observations and 1,257 observations make up the sub-sample of mixed parents.

Second, for each sub-sample the following equation is estimated to find the probability of dropping out of high school.

$$P(D) = \beta_1 N + \mathbf{X}\theta \tag{3.2}$$

where X represents the vector of variables discussed in Section 3.3, except for the child's nativity. The dichotomous indicator (N) indicates the child's native status, which takes the value 1 if native-born and 0 if non-naturalized citizen. Only in the case of the sub-sample of mixed parents, additionally the dichotomous indicator for unauthorized immigrant mother (U) and the interaction between child's citizenship status and unauthorized immigrant mother $(N \times U)$ are included as follows.

$$P(D) = \beta_1 N + \beta_2 U + \beta_3 (N \times U) + \mathbf{X}\theta \tag{3.3}$$

where $U = 1$ if mother is classified as highly unauthorized and $U = 2$ if father is classified as highly unauthorized.

Alternatively, as a test for robustness, I repeat the entire process with the critical probability (P_l) set at 0.9. This is a more strict classification of legal status, for only parents with a very high predicted legal probability would be classified as highly likely to be legal. Similarly, only parents with a very low predicted probability (lower than 0.1) will be classified as highly likely to be unauthorized immigrants. Unlike before, this results in assigning parents with a predicted

[6] In this set up, it is required that $(1 - P_l) < P_l$.

legal probability between 0.1 and 0.9 as "unsure legal status". With this alternative cutoff point, if both parents' predicted legal probability falls between 0.1 and 0.9 such children are excluded from the sample, because they would not add any information to the research question. Under this robustness test, the groups of mixed families are broader than in the previous case with the cutoff at 0.5. For instance, contrary to before, now in addition to the two parents being in one of the two groups –highly likely legal or highly likely unauthorized, now one parent can be in the uncertain region of 0.1-0.9 probability. This is captured in the model by expanding the dichotomous indicator for unauthorized immigrant mother (U) into a polytomous indicator U which takes the value 0 if no parent is highly unauthorized, and as before $U = 1$ if mother is highly unauthorized and $U = 2$ if father is highly unauthorized. This mix group also includes some families with a legal parent, which is accounted for by including dichotomous indicator $L = 1$ if either parent is highly likely to be legal. This extended model is summarized in Equation (3.4) below.

$$P(D) = \beta_1 N + \beta_2 U + \beta_2 (N \times U) + \beta_3 (N \times L) + \mathbf{X}\theta \tag{3.4}$$

3.5 Empirical Results

Table 3.1 shows the results of the six models estimated. Panel 1 corresponds to the sub-samples derived by using the cutoff of $P_l = 0.5$, while Panel 2 corresponds to the more strict definition of authorized/unauthorized classification at $P_l = 0.9$. Columns (i) and (iv) are associated with the subsamples of two legal immigrant parents, while columns (ii) and (v) relate to the subsamples of unauthorized immigrant parents. Similarly, columns (iii) and (vi) show the subsamples of mixed parents. As shown in the bottom panel of Table 3.1 the models are based on a large number of parameters, due to the inclusion of indicators for the child's state of residence and each parent's country of origin. As such, this table only reports relevant coefficients. Moreover, all models are weighted by their sample weights, while the standard errors reported are robust standard errors.

Table 3.2 shows the predicted high school drop out probabilities for an otherwise average child when his parents' legal status and his native/non-citizen status take different combinations. Here Panel A and B correspond to Panels 1 and 2, respectively in Table 3.1 above, while columns in each panel refer to the same sub-samples of parent types discussed earlier. The results summarized

	Panel 1 ($P_l = 0.5$)			Panel 2 ($P_l = 0.9$))		
	Legal parents (i)	Unauthor--ized parents (ii)	Mixed parents (iii)	Legal parents (iv)	Unauthor--ized parents (v)	Mixed parents (vi)
Coefficients						
Native	-0.0328	-0.1018	-0.0768	-0.0143	-0.1157	0.0138
	0.0020	0.0033	0.0037	0.0025	0.0063	0.0140
Mother			-0.0603			0.0820
–unauthorized			0.0063			0.0116
Father						0.0260
–unauthorized						0.0115
At least 1						0.1582
–parent legal						0.0114
Native &			0.0912			-0.0934
–mother unauth.			0.0071			0.0137
Native &						0.0116
–father unauth.						0.0138
Native &						-0.1345
–1 parent legal						0.0135
Constant	-0.9516	-2.0372	-1.8253	-0.977	-0.892	-0.7781
	0.0215	0.0202	0.0334	0.020	0.046	0.0255
Sample size	2203	1,632	1251	1331	341	1326
Weighed sample	249,130	199,577	152,738	146,860	45,080	165,179
Drop out (N)	157	161	142	87	25	110
Native born (N)	635	1,018	752	333	192	691
Parameters (N)	187	53	68	160	32	108
R^2	0.1918	0.1466	0.2133	0.2452	0.3376	0.2397

Notes:
Robust standard errors reported below estimated parameters.
In column (vi) omitted category is No parent unauthorized.

Table 3.1: Selected results of Models 1 and 2

in Table 3.2 are as follows.

First, a child of two unauthorized immigrants has a lower high school drop out probability than a similar child of two legal immigrants. Specifically, as per the more lax distinction between legal and unauthorized immigrants depicted in Panel A, for a native-born child, having two unauthorized immigrant parents makes him 14 percent less likely to drop out of high school, than having two legal immigrant parents. Similarly, a non-citizen child also has a lower high school drop out probability if he has unauthorized immigrant parents. However, here the gap is smaller, where the non-citizen child of unauthorized immigrant parents is 7 percent less likely to drop out of high school than a similar child of legal immigrant parents. These two findings are robust to the strict definition of legal status depicted in Panel B. However, here the differences in predicted high school drop out probabilities are smaller than in Panel A. For instance, a native-born child with unauthorized immigrants is 12 percent less likely to drop out of high school than a similar child with legal immigrant parents. In the case of a non-naturalized citizen child, one with unauthorized immigrant parents is 2 percent less likely to drop out of high school than a similar child with legal immigrant parents.

Second, the study show that among children with unauthorized immigrant parents, a non-naturalized citizen child is more likely to drop out of high school than a similar native-born child. This finding is robust to both speficications of P_l. In the context of the more lax definition of parents's legal status in Panel A, a non-naturalized citizen child is 10 percent more likely to drop out of high school than a native-born child, while the findings pertaining to strict definition in Panel B show that a non-naturalized citizen child is 12 percent more likely to drop out of high school than a similar native-born child.

The findings in the context of mixed families in Panel A are not always robust to findings with the more strict definition of legal status depicted in Panel B. In Panel A, a non-naturalized citizen child with an unauthorized immigrant father is over 7 percent more likely to drop out of high school than a native-born child . This finding is only robust to the case of a child with a legal immigrant mother in Panel B (predicted probabilities of 0.0858 versus 0.1949).

Moreover, other findings for mixed parents shown in the last column in Panel B are as follows. A non-naturalized citizen child of an unauthorized immigrant mother is more likely to drop out of high school than a similar native-born child. For instance, a non-naturalized citizen child with a legal

immigrant father and a unauthorized immigrant mother has a 25 percent probability of dropping out of high school, while a similar native-born child's probability is only 4 percent. A similar pattern is seen when the father is not a legal immigrant. Specifically, such a non-naturalized citizen child's high school drop out probability is 9 per cent while a similar native-born child's probability is only 1 percent.

In the case of a non-naturalized citizen child with a mixed family, when one parent is highly likely to be an unauthorized immigrant, the high school drop out probability is higher if the other parents is highly likely to be legal (0.2509 and .1949 versus 0.0928 and 0.0368, respectively). Similarly, among native-born children also, the combination of a highly likely legal and unauthorized immigrant parents resulted in higher high school drop out probabilities for the child than the parent combination of an unsure and unauthorized immigrants. (0.0369 and 0.0858 versus 0.0132 and 0.0621, respectively).

3.6 Discussion

The empirical findings of this study can be summarized into four interesting results. First, when controlling for other covariates a child of two unauthorized immigrant is less likely to drop out of high school than a similar child of two authorized immigrant parents. Second, among two unauthorized immigrant parents, a non-citizen child is more likely to drop out of high school than a similar native-born child. Third, in the context of mixed families, among unauthorized immigrant mothers, a non-naturalized citizen child is more likely to drop out of high school than an otherwise similar native-born child. Fourth, among both native and non-naturalized citizen children, the parent combination of highly likely unauthorized and highly legal results in a higher high school drop out probability than the parent combination of highly unauthorized and unsure legal.

The first finding shows that a child of two unauthorized immigrant is less likely to drop out of high school than a similar child of two authorized immigrant parents, while the fourth finding can be seen as a weaker version of the first, because the later is a combination of highly legal and highly unauthorized immigrant parents. These two findings together may appear puzzling at first. Nonetheless, they are consistent with the hypothesis that children of immigrants (without distinguishing between legal status of parents) perform better than children of natives (Djajic, 2003;

Type of Child		Legal parents	Unauthorized parents	Mixed parents
Panel A (cutoffs at 0.5)				
Native	(N=1)	0.09524	-0.04753	
-Mother sure unauthorized	(U=1)			0.17272
-Father sure unauthorized	(U=2)			0.14178
Non naturalized	(N=0)	0.12808	0.05430	
-Mother sure unauthorized	(U=1)			0.15834
-Father sure unauthorized	(U=2)			0.21861
Panel B (cutoffs at 0.1 and 0.9)				
Native Born Child	(N=1)	0.1001	-0.0166	
At least one parent sure legal	(L=1)			
–No parent sure unauthorized	(U=0)			0.0483
–Mother sure unauthorized	(U=1)			0.0369
–Father sure unauthorized	(U=2)			0.0858
No parent sure legal	(L=0)			
–No parent sure unauthorized	(U=0)			0.0246
–Mother sure unauthorized	(U=1)			0.0132
–Father sure unauthorized	(U=2)			0.0621
Non-naturalized Child	(N=0)	0.1144	0.0990	
At least one parent sure legal	(L=1)			
–No parent sure unauthorized	(U=0)			0.1689
–Mother sure unauthorized	(U=1)			0.2509
–Father sure unauthorized	(U=2)			0.1949
No parent sure legal	(L=0)			
–No parent sure unauthorized	(U=0)			0.0108
–Mother sure unauthorized	(U=1)			0.0928
–Father sure unauthorized	(U=2)			0.0368

Table 3.2: Predicted high school drop out probabilities

Driscoll, 1999; Shields & Behrman, 2004; Wojtkiewicz & Danato, 1995). The proponents of this argument postulate that children of immigrants perform better due to strong ethics and aspirations. For instance, Shields & Behrman (2004, p. 5) notes that "[immigrant] parents are willing to work hard, and they expect their children to do the same" and "children of immigrants tend to have high educational aspirations and less likely than children of US-born families to engage in risky behaviors $\cdots$ they also tend to spend more time doing homework and that they do better in school". Similarly, Djajic (2003) shows that traditional family values such as inclusion of grandparents at home – who encourage children in educating themselves – and immigrants' emphasis on educational achievement contribute towards children of immigrants performing better than natives. Moreover, in addition to higher expectations for the educational attainment of their children, the parents' immigration decision itself is often based on the possibility of "the advancement and success of their children" (Driscoll, 1999, p. 861).

The literature also notes that the "the environment that immigrants encounter at destination plays a crucial role in determining their path of assimilation and the pace of human capital accumulation of their children" (Djajic, 2003, p. 833). Compared to legal immigrants, unauthorized immigrants are more likely to be confined to associate immigrants with similar circumstances and be shunned by natives, which results in inhibiting the assimilation process of unauthorized immigrants. Djajic (2003, p. 834) notes that "persistence of differences in customs, values and attitudes between immigrants and native may also have positive implications for the pace of human capital accumulation of immigrant children". For instance, divorce or separation rates among immigrants are lower than among native, and this tendency for "immigrant families to stick together" shows the incomplete assimilation which also "serves as a stabilizing element that contributes to a better academic performance and economic success of second generation". Moreover, "those who avoid assimilation by maintaining close ties with their ethnic community and subculture may in some cases have a better chance of achieving social, educational and economic mobility" Djajic (2003, p. 834, 835,841).

As such, among the two types of children considered in the first finding, in terms of assimilation, children of authorized immigrants are more analogous to children of natives in the division between native-born parents and immigrant parents. Similarly, children of unauthorized immigrants are more analogous to children of immigrants in the division between native and immigrant parents.

In that case, the finding that a child of two unauthorized immigrant parents is less likely to drop out of high school, than a similar child of two authorized immigrant parents is consistent with the hypothesis that children of immigrants perform better than children of natives.

In general, immigrants place a higher faith "in the role of their [children's] scholastic achievement in the family's struggle to attain its social and economic objectives" than natives (Djajic, 2003, p. 835). This would be even stronger among unauthorized immigrants, because the decision to be an unauthorized immigrant in the US involves an enormous risk. A rational individual would take this risk only if there is a possibility of an offsetting return to this risk. Often the return sought by unauthorized immigrants is in terms of socio-economic success. Hence, it is reasonable to consider that unauthorized immigrant parents would place a higher importance for education than an authorized immigrant parent would. For children of authorized immigrants, poor educational performance can be compensated by privileges associated with their authorized immigrant status. Some such privileges include: a larger pool of jobs to choose from, which expands beyond the relatively smaller informal sector to the much larger formal sector; the ability to apply for unemployment benefits if unemployed; the flexibility to go back to their country of origin and return to the US at a more opportune time at their own discretion, such as go back during an economic downturn in the US and return during an economic boom; and the possibility to stand up against discrimination. On the contrary, a child of two unauthorized immigrant parents would be cautious in applying to certain jobs, such as jobs that e-verify, or for unemployment benefits, due to possible betrayal of parents, and in some cases their own, unauthorized presence in the US. Moreover, for an unauthorized immigrant who has unlawfully resided in US for over one year, lawful re-entry to the US is barred for a minimum of 10 years.[7] This difference in the importance of education for socio-economic success may be the driving force behind the lower high school drop out probability among children of unauthorized immigrants, than among children of legal immigrants, as seen in the first and fourth findings.

The second finding shows that among unauthorized immigrant parents, non-naturalized citizen children are more likely to drop out of high school. The third finding, which shows that when only the mother is an unauthorized immigrant, a non-naturalized citizen child is more likely to drop out of high school than an otherwise similar native-born child, can be identified as a weaker

[7]For unlawful presence for less than 1 year, a minimum of three years have to be lapsed before lawful re-entry,

version of the second finding. These results are plausible given that a native-born child may have better access to welfare services than a non-naturalized citizen child. For example, Hanson (2009) notes that native-born children of unauthorized immigrants receive Medicaid and participate in school breakfast and lunch programs. Such better health and nutrition has the capacity to result in the lower high school drop out probability among native-born children. Such a difference in the high school drop out probabilities may occur for two children in the same household, when one is native-born and the other is non-naturalized citizen.

In this context, this finding is particularly important in the context of the proposed Development, Relief and Education for Alien Minors (DREAM) Act. The DREAM Act proposes conditional permanent residency to unauthorized immigrants of good moral character, who graduate from US high schools, arrived in the US as minors, and lived in the country continuously for at least five years prior to the enactment of the bill. The potential beneficiaries of the DREAM act are among the non-naturalized citizen children, because it is highly likely that non-naturalized children of unauthorized immigrants are unauthorized immigrants themselves. Given that these findings show that relative to a native, a non-naturalized citizen child of two unauthorized immigrants is more likely to drop out of high school, these minor unauthorized immigrants with a relatively higher probability to drop out of high school would have an incentive to complete high school education, if the DREAM Act is enacted. However, the relevance of the findings in the context of the proposed DREAM Act should be interpreted with caution. Specifically, as seen in both panels, the high school drop out probability of a non-naturalized citizen child of unauthorized immigrants parents is lower than predicted drop out probabilities of both native and non-naturalized citizen children of legal immigrants. As such, the general phenomenon of lower high school drop out probability among children of unauthorized immigrants should not be mis-understood as positive benefit of the DREAM Act, if the DREAM Act is enacted in future.

3.7 Concluding remarks

This study investigates the affect of parents' legal status on the probability of a child to drop out of high school, using a linear probability model in multiple sub-samples to control for parents legal status. When controlling for other covariates that determine high school drop out probability in the

sample of children between the ages 16-18 years, this econometric analysis finds four conclusions. First, a child of two unauthorized immigrant is less likely to drop out of high school than a similar child of two authorized immigrant parents. Second among two unauthorized immigrant parents, a non-citizen child is more likely to drop out of high school than a similar native-born child. Third, in the context of mixed families among unauthorized immigrant mothers, a non-naturalized citizen child is more likely to drop out of high school than an otherwise similar native-born child, and fourth, among both native and non-naturalized citizen children, the parent combination of highly likely unauthorized and highly legal results in a higher high school drop out probability than the parent combination of highly unauthorized and unsure legal.

These findings have important policy implications. Specifically, the first and fourth findings show that children of unauthorized immigrants are less likely to drop out of high school, while the second and third findings show that non-naturalized citizen children of unauthorized immigrants are more likely to drop out of high school. As such, the later show that the DREAM Act would encourage children that would have dropped out of high school to complete high school, due to the expectation of getting permanent resident status and subsequent citizenship status. But, on the other hand the other result shows that children of unauthorized immigrants in general have a lower high school drop out probability than children of legal immigrants. In this context, this study shows that the DREAM Act can be viewed to encourage a child to continue high school only if his high school drop out probability is sufficiently large to result in him dropping out in the absence of the DREAM Act.

Finally, the findings of this study are important on two dimensions. First, this study is among the first few studies (if not the first study) that explores the children of unauthorized immigrants, and second, the findings are consistent with the broader literature of children of immigrants.

Chapter 4

A Micro Analysis of Contextual Determinants of Labor Migration in Sri Lanka

4.1 Motivation

Migration from rural to urban areas is an important facet of structural transformation experienced during the development process of an economy. Sri Lanka is a developing country at the cross-roads of structural transformation. A notable characteristic of Sri Lanka's structural transformation experience is the decline in the share agriculture in the composition of GDP and in the composition of sector of employment in the labor force. This decline in agriculture has made way for the growth of the industrial and service sectors. Along with these structural dynamics the location of economic activities and location of living quarters of the population also have moved from rural to urban areas. However, this residential transformation is lagging behind the compositional transformation. For instance, compared to 1963, in 2008 the share of agriculture in GDP (at current factor cost prices) declined to 13.4 percent from 38 percent and employment in agriculture declined to 32.7 percent of labor force from 52.6 percent. In terms of urbanization, UNESCAP (2008) indicates that only 15 percent of the total population in Sri Lanka live in urban areas. Given that 42 percent of the population of China and 29 percent of the Indian population live in urban areas in that same

year, one could hypothesize that Sri Lanka is yet to face mass migration from rural to urban areas as part of the ongoing structural transformation process.

Despite this clear evidence of emerging urbanization in Sri Lanka, there exists a dearth of literature on internal migration related to the economic development process.[1] Against this backdrop, there is an abundance of research questions on development-related internal migration that can be addressed in the case of Sri Lanka, out of which, this study aims to identify determinants of migration with an emphasis on contextual variables. The study employs a discrete choice model using micro level data from the Consumer Finance and Socioeconomic Survey (CFS), conducted by the Central Bank of Sri Lanka in 2003/2004.

The remaining sections of this paper will be as follows. Section 4.2 sets the current paper in the context of existing literature. Section 4.3 introduces and describes the data while section 4.4 deals with the methodological approach and the selection of the appropriate model. Section 4.5 provides the empirical results in terms of coefficient estimates and marginal effects followed by concluding remarks in section 4.6.

4.2 Literature review

In terms of the rural to urban migration literature, a natural starting point is the Lewis model (Lewis, 1954), which explores the growth of a developing economy in terms of labor transition between two sectors–a traditional agricultural sector and a modern industrial sector. Other important work on the subject include the Todaro (1969) and the Harris & Todaro (1970) models, which essentially model determinants for internal migration and establish that the migration decision is based on expected income differentials between rural and urban areas, and not based on wage differentials.

Katz & Stark (1986) note that later empirical research on rural-to-urban migration shows that the expected income hypothesis of rural-to-urban migration associated most closely with Todaro (1969) does not perform well either in terms of the sign of the coefficients or their statistical significance. The authors also note that in many cases the expected income in the urban area is not larger than the expected income in the rural area. The combination of these observations

[1] However, literature on internal migration resulting as a result of the ethnic conflict is abundant. The determinants of these conflict related migration flows are different from development related ones.

result in an empirical paradox: "how is it that calculative behavior by rational people results in choice of an actuarially unfair risky prospect"(Katz & Stark, 1986, p. 135). In answering this paradox, Katz & Stark (1986), find that a small chance of reaping a high reward is sufficient to trigger rural-to-urban migration. Their result can be supported by either of their following two explanation. First, capital markets are incomplete in such a way that rural to urban migration may serve as an intermediary between capital and labor markets to partially correct the deficiency in the former. The second explanation provided by the authors rests upon the combination of wealth and rank (relative deprivation or satisfaction) considered by individuals. Here the authors say that individuals who are globally risk averse in wealth and rank might migrate to the urban area if there is a small prospect of greatly enhancing their rank, even in the presence of an actuarially unfair risks with respect to wealth. A similar hypothesis is the relative deprivation hypothesis by Stark (1984), which is essentially that one person who is poor relative to the reference group in his own village may choose to migrate to a urban setting to uplift his ranking relative to reference group, while another person who is even poorer, may choose to not move for a similar gain if he is well off compared to his own reference group.

Another branch of rural to the urban migration literature approaches migration as a response to risk or uncertainty. As noted by Stark & Levhari (1982) the early literature (Myrdal, 1968; Sahato, 1968) argued that the nexus between migration and risk was driven by individuals' "risk-taking or risk-loving properties"(Stark & Levhari, 1982, p. 196). A subsequent stream of literature viewed migration as a response to their risk avoiding properties. The main argument in this branch of literature (Katz & Stark, 1986; Stark & Levhari, 1982) was that internal migration leads to reducing total familial risk via diversification of earning sources. One such diversification strategy is the placement of a family member in a town or urban area, where the earnings at the source and destination are not highly positively correlated (Lucas & Stark, 1985; Stark & Lucas, 1988).

Another branch of literature on internal migration focuses on the connection between structural transformation and related migration. Literature in this area has followed several waves. The first wave focuses on the development process of the now-developed regions such as North America and Western Europe. Another wave deals with the Latin American region and a more recent wave deals with the experiences of present day developing countries led by China and India. Among these several aspects of migration literature, the characteristics of the Sri Lankan migration best fits that

of Latin America or India and other South Asian economies due to similarities in socio-economic characteristics.

In studying internal migration in Brazil, Sahato (1968) identifies three broad approaches to migration modeling. First, the Chicago School (Schultz, 1962; Sjaastad, 1961, 1962) puts internal migration in a framework of costs and returns of investment in human capital. The second approach refers to Simon Kuznets' work which delineates the relationship between internal migration and economic development in terms of the selectivity of people. The third approach deals with the 'push' and 'pull' factors associated with the work of Ravenstein (1885) and Redford (1926). Sahato (1968)'s study blends and tests all of these approaches by using a single-equation and a simultaneous equation model.

Lipton (1980) broadly argues that in the case of India, intra-rural inequality is a major cause of rural-urban migration, where better-off villagers tend to be 'pulled', and worse-off villagers 'pushed', from the same subset of relatively 'unequal' villages. The author also states that migration does not equilibrate between urban and rural sectors, largely because of externalities and compositional factors; but it does smooth itself, largely because individuals behave rationally and learn quickly. Moreover, Lipton (1980) also notes that the lesson for development studies is not that 'markets fail'. But, under conditions of both poverty and structural inequality, they function - but with generally unacceptable consequences.

Similar to Lipton (1980), in the context of Bangladesh, Hossain (2001, p. 2) states that "the propensity of migration is usually influenced by a combination of push-pull factors". He focuses on the differentials and determinants of migration, and identifies the factors influencing out-migration using household level microdata. The findings indicate that poverty, job searching and family influences were the main push factors for out-migration, while better opportunity, prior migrants and availability of jobs were the main pull factors behind migration. This study uses a multivariate logistic regression analysis.

Among contextual variables, Lucas (1997) notes that location of infrastructure such as public transportation has an impact on migration patterns both directly and indirectly. Another such contextual variable is the non-farm income in the rural areas. As noted by Foster & Rosenzweig (2008), non-farm income in rural areas has experienced a substantial growth in developing countries and thus for better analysis of internal migration in these countries the models should include non-

farm income.

As noted in the foregoing review, there is no literature dealing with internal migration in Sri Lanka. As such, this study is very timely given the present context of Sri Lanka, where the post-conflict economic development process will lead to waves of mass internal migration in the near future.

4.3 Data

The micro data used in this study come from the Consumer Finance and Socioeconomic Surveys, conducted by the Central Bank of Sri Lanka in 2003/2004. The data capture 11,722 households with 50,545 non-migrants and 1,545 migrants. Migrants captured in this survey are household members "who went abroad or elsewhere in the country during the last 24 months and presently living there". This dataset has two limitations pertaining to the design of the survey. First, the information gathered from migrants is a small component of the large survey. As such, though non-migrants information contains much detail on demographic features, housing, education, health conditions, economic activity, income, expenditure and savings, migrant data is limited to nine simple variables. They are relationship to head of household, sex, age, marital status, ethnicity, purpose of migration, type of employment (15 types), region migrated to (8 international regions and within Sri Lanka), and duration of migration (in months). Two notable variables are absent in these migrant data: any kind of indication about human capital variable, and for internal migrants, any kind of information about the destination district or province.[2] Second, given that the CFS collected data from the origin, those who migrated as a household are not captured in this dataset. Unfortunately, no correction can be made in this regard. As such, some of the findings should be interpreted with caution. The remainder of this section discusses some descriptive statistics of the CFS.

Among the 1,545 migrants captured in the CFS (03/04), 1,350 had migrated for employment or business purposes, which was almost equally distributed among internal and international migrants (see Table 4.1). The international migration cohort was dominated by females (see Table 4.2). The main driver of this female bias in international migration is the high outflow of Sri Lankan females

[2]Nonetheless the richness in data on non-migrants compensates for the absence of a human capital variable for migrants. This will be discussed subsequently.

Type	International	%	Internal	%	Total	%
Employment/ Business	672	94.6	678	81.2	1350	87.4
Education/training	25	3.5	109	13	134	8.7
Other	13	1.8	48	5.7	61	3.9
Total	710	100	835	100	1545	100

Table 4.1: Distribution of migrants by purpose of migration

to Middle Eastern countries as unskilled labor (Gamburd, 2004; SLBFE, 2003). The majority of the international migrants were married, while on the contrary internal migrants were mostly never married. Among ethnicity groups Sinhala had the highest number of internal migrants for both employment related migration and overall migration, while a large proportion of Sri Lankan Tamils migrated overseas. The mean age of employment related internal migrants was 29 years, while it was 33.4 years for international migrants. This indicates that the internal migrants were younger, with a relatively lower level of family commitment, while international migrants were mature with greater family commitment.

Table 4.3 above shows the ratio of internal to international migration. While the ratio equals approximately 1 at the national level, there is significant disparity among districts. As expected, districts with large metropolitan areas such as Colombo and Gampaha have a very low ratio. This indicates that individuals from these areas are better off by migrating overseas than by migrating internally. On the other hand, districts such as Ampara or Trincomalee reflect a high proportion of international migrants perhaps due to the ethnic conflict. The highest ratio of 6 is recorded by both Monaragala and Ratnapura, while Matara recorded 3.38 reflecting a greater outflow of internal migrants from these regions.

In addition to the use of microdata from the CFS(2003/4) this study enhances the CFS data with district level data from the Socio-economic indicators published by the CFS (2004).

	External	Internal	External labor	Internal labor
Male	332	630	314	539
Female	379	206	358	139
Never Married	229	541	203	411
Married	436	286	424	262
Widowed	25	5	24	3
Separated	18	4	18	2
Divorced	3	0	3	0
Total	711	836	672	678
Sinhala	469	641	442	535
Sri Lankan Tamil	75	109	67	72
Indian Tamil	18	37	17	35
Moor	137	48	135	35
Malay	6	0	6	0
Burgher	6	1	5	1

Table 4.2: Distribution of migrants by characteristics

District	Internal		International		Ratio*
	Employ. based	Non Employ. based	Employ. based	Non Employ. based	
Colombo	16	5	63	12	0.25
Gampaha	26	3	80	4	0.33
Kalutara	31	12	40	3	0.78
Kandy	36	1	44	2	0.82
Matale	20	2	18	2	1.11
Nuwara Eliya	24	3	8	0	3.00
Galle	51	14	31	0	1.65
Matara	27	11	8	0	3.38
Hambantota	31	16	12	1	2.58
Jaffna	27	7	18	3	1.50
Vavuniya	6	4	5	0	1.20
Batticaloa	34	23	28	2	1.21
Ampara	22	4	38	1	0.58
Trincomalee	20	5	30	1	0.67
Kurunegala	88	9	85	2	1.04
Puttalam	24	3	66	6	0.36
Anuradhapura	63	5	37	0	1.70
Polonnaruwa	6	4	15	0	0.40
Badulla	29	5	14	0	2.07
Monaragala	30	7	5	0	6.00
Ratnapura	24	7	4	0	6.00
Kegalle	43	8	23	0	1.87
Total	678	158	672	39	1.01

Notes:
Ratio of Internal employment based migrants to international employment based migrants.

Table 4.3: Internal and international migrants by district

District	Labor in agri (%)	Gini coeff.	Own house (% of hh)	19-34 yr pop. (%)	Borrowing (% of income)	Labor force pop. (%)
Colombo	3.1	0.45	87.0	25.4	17.4	47.1
Gampaha	7.0	0.43	91.5	25.7	17.6	46.7
Kalutara	20.2	0.36	95.2	25.9	28.6	48.3
Kandy	26.7	0.43	83.8	24.0	18.8	44.4
Matale	44.5	0.43	91.2	25.4	16.3	46.7
Nuwara Eliya	70.8	0.39	53.5	24.2	19.0	52.8
Galle	31.2	0.35	95.4	22.0	13.9	47.4
Matara	48.3	0.35	92.9	22.7	28.3	47.2
Hambantota	42.2	0.48	96.5	22.4	61.9	45.1
Jaffna	38.4	0.44	58.2	24.3	37.0	31.7
Vavuniya	26.9	0.41	81.3	24.8	47.4	37.7
Baticaloa	32.6	0.43	91.7	20.0	66.9	34.2
Ampara	45.4	0.56	94.3	26.1	37.0	40.4
Trincomalee	30.2	0.48	86.5	24.7	29.5	35.7
Kurunegala	28.9	0.39	96.4	24.3	24.6	46.6
Puttalam	31.7	0.46	92.4	25.6	17.9	45.0
Anuradhapura	57.8	0.46	98.1	25.5	12.7	48.5
Polonnaruwa	47.9	0.48	96.8	27.1	16.6	48.1
Badulla	68.0	0.44	79.9	20.4	19.0	48.4
Monaragala	67.3	0.40	95.4	22.6	25.9	47.5
Ratnapura	49.1	0.42	89.7	24.4	16.8	53.7
Kegalle	28.8	0.38	91.7	24.5	12.3	48.3

Table 4.4: District characteristics

4.4 Methodology

The econometric methodology employed in this study is a discrete choice model to investigate how contextual variables affect internal migration in Sri Lanka. As shown in the descriptive analysis above, the dataset has information pertaining to both internal and international migration. In order to improve the precision of the model, in addition to internal migrants, I also consider international migrants in the estimation. A such, I consider a multinomial logit model with three outcomes : internal migrants, international migrants and stayers which is also referred to as non-migrants. The explanatory variables considered in the model consists of three vectors: a vector of individual level variables, a vector of household level variables, and a vector of district level variables.

In the vector of individual level characteristics, a human capital variable is absent because the migrant module of the CFS did not gather such information. As such, for migrants, their years of schooling is imputed by fitting an OLS regression for each occupational group, on data pertaining to non-migrants and applying the estimated equation to migrants. In this imputation procedure some judgment calls were required due to the difference in coding occupations of migrants and stayers. The occupations of migrants were coded with a list of 15 broad occupational groups[3] while occupations of non-migrants were coded with the comprehensive list of occupation codes of the 2001 Census of Population and Housing. For instance, in imputing values for the broad category of migrant laborers, the sample used consists of only construction laborers based on anecdotal evidence of the high demand for construction labor in the survey years in urban areas. For the category of migrant occupation "other", the human capital variable is imputed based on the entire sample of working non-migrants. In the case of doctors, engineers and accountants the imputed value was set to 13, the maximum years of schooling, since passing the 13th year in school is a prerequisite for these professions.

Along with the imputed years of schooling for migrants and the survey recorded years of schooling for non-migrants (defined as years of schooling), other individual level variables used are age, age squared, gender, marital status and relationship to head of household, an indicator for residents of rural areas and an indicator variable for ethnicity. These individual level variables are identified

[3] 1. Mason/carpenter/plumber; 2. Driver; 3.Domestic aide; 4. Laborer; 5. Teacher; 6.Technician; 7. Nurse; 8. Doctor; 9. Engineer; 10. Accountant; 11. Administrative & clerical service; 12. Business; 13.Hotel services; 14. Armed forces police service; 15. Other

as vector X_1. These demographic characteristics are included in the model to find the impact of contextual variables net of these characteristics. The indicator variable for rural area is critical to the analysis of migration because, the socioeconomic conditions in rural areas is different from that of urban areas. For instance, for the same point of time Karunaratne (n.d.) shows that unemployment is higher in rural areas, while (Bandara, 1997) shows the high incidence of poverty in rural areas.

The vector of household level variables, X_2, consists of maximum years of schooling in the respondent's household, number of income receivers in the household and share of not employed persons in the household. Better educated persons are expected to have better information, particularly about employment opportunities in potential destinations. As such, the household level human capital variable is expected to control for information that a household would have. The direction of the relationship between the share of not employed persons in a household and the probability of migration indicates the structure a of household that is likely to have a migrant. For instance, a positive relationship implies that the larger share of not employed persons and the associated fewer family income compels one family member to migrate in search of higher income. However, in the presence of reverse causality, a positive coefficient would imply that higher household income from migration resulted in fewer persons needing to be employed. Meanwhile, a negative coefficient on this variable could imply that those not employed at home are elderly or sick persons who require the assistance of an additional family member at home.

The vector of district level variables are identified as X_3 and this vector consists of 6 variables. They are share of labor in agriculture, household level Gini coefficient, and the share of households that owned a house, size of the district labor force, share of population in ages 19-34 years and the share of borrowings as a percentage of income. The share of labor in agriculture reflects the degree of district level structural transformation in terms of sectoral composition of employment. As such this variable is important to this analysis which focuses on internal migration connected to structural transformation. The household level Gini coefficient for districts reflects income inequality within districts. Hence, this variable allows to test if the relative deprivation hypothesis is valid in the case of migration in Sri Lanka. The share of household level population that owned a house has multiple links to migration.

On the one hand, house ownership can be viewed as an asset that can be offered as collateral

Variable	Obs.	Mean	Std. Dev.
Internal migrant	20503	0.0398	0.1956
International migrant	20503	0.0344	0.1823
Stayer	20503	0.9257	0.2622
Age	20498	38.1447	13.4206
Age square	20498	1635.1220	1114.9940
Male	20503	0.6692	0.4705
Schooling (yrs)	20011	8.3419	3.7186
Married	20503	0.6734	0.4690
Daughter	20503	0.1020	0.3026
Son	20503	0.1914	0.3934
Tamil	20504	0.1255	0.3313
Burgher	20503	0.0024	0.0488
Malay/Moor	20504	0.0672	0.2504
Max. yrs of schooling in hh	20176	9.6811	3.1403
Income receivers in hh (no.)	20503	2.1660	1.0928
Share of unemployed in hh	20503	0.0512	0.1487
Rural	20504	0.8799	0.3251
Labor in agri (%)	20504	33.2422	19.9567
Labor force ('000)	20503	46.6549	4.1525
Gini coefficient (hh)	20503	0.4214	0.0444
Own house (% of hh)	20503	89.1558	10.0110
19-34 yr pop (%)	20503	24.3403	1.6065
Borrowing (% of income)	20503	22.7085	11.6841

Table 4.5: Summary statistics

for debt, while house ownership can also be interpreted as wealth. On the other hand, house ownership shows the degree of permanency in the choice of residential area. A more permanent resident population is less likely to migrate as part of their residential choice. The inclusion of this variable enables me to capture the effects of migration due to house ownership. A larger labor force has many externalities for its member as well as opportunity for specialization. Hence, this inclusion of the variable is intended to show's how the magnitude of the labor force affects migration. The share of population in ages 19-34 years shows if the district has a large concentration of young working aged persons who are more likely to migrate, while the share of borrowings as a percentage of income shows the availability of loanable funds at the district level. Table 4.5 shows the summary statistics of these variables included in the empirical estimation.

The above 18 regressors for the model were selected by a stepwise method. First, a preliminary model was estimated with individual level variables and three contextual variables namely the share of labor in agriculture, household level Gini coefficient for districts, and the share of households in district population that owned a house. These three variables were picked first because the share of labor in agriculture is central to the analysis of migration that is related to structural transformation, the Gini coefficient enables to test the relative deprivation hypothesis of Stark (1984) and Katz & Stark (1986), and the share of house ownership is relavent in discerning the impact of wealth on migration. Subsequently, combinations of other contextual variables were included to the model until the statistical significance of the initially selected three contextual variables started to decline.

Using the above covariates, I estimate a multinomial logit model (MNL) with three outcomes: stayer, internal migrant and international migrant. Alternatively, I also consider the model where I pool the internal and international migrants together and estimate a logit model. In order to test the appropriateness of a MNL versus a model with pooled outcomes, I follow the Likelihood Ratio test designed by Cramer & Ridder (1991), depicted in equation 4.3 to test the null hypothesis if regressors in the MNL and the pooled model share the same coefficients.

$$LR = 2(\ln L_U - \ln L_R) \tilde{} \chi^2(q) \tag{4.1}$$

Here q is the number of restrictions, $\ln L_U$ is the log likelihood of the full multinomial logit model with 3 outcomes (internal migrant, international migrants and non-migrants) and $\ln L_R$ is computed from the following expression:

$$\ln L_R = \ln L_R^* + n_1 \ln n_1 + n_2 \ln n_2 - (n_1 + n_2) \ln(n_1 + n_2) \tag{4.2}$$

where L_R^* is the log likelihood of the pooled model (internal and international migrants pooled as emigrants) while n_1 and n_2 are the number of observations that were international migrants and internal migrants, respectively. Given that there were 519 internal migrants and 512 international migrants in the sample used for estimation, and the log likelihood of the pooled model being -3263.63, the $\ln L_R$ is -3978.24. The $\ln L_U$ associated with the MNL model is -3727.11. This gives a χ^2 statistic of 502.26 with 19 degree of freedom for the LR test. As such, the test rejects the

Variable	Internal Coeff	SE(robust)		International Coeff	SE(robust)	
Age	-0.1128	0.0179	***	0.1398	0.0336	***
Age square	0.0010	0.0002	***	-0.0025	0.0005	***
Male	0.5116	0.2297	**	-2.1814	0.3008	***
Schooling (yrs)	-0.0697	0.0238	***	-0.2740	0.0366	***
Married	-0.2026	0.1977		0.4795	0.1593	***
Daughter	2.2390	0.2484	***	1.2558	0.1795	***
Son	1.8627	0.2098	***	2.3283	0.2215	***
Tamil	0.0387	0.2381		-0.0252	0.2003	
Burgher	0.0494	0.7482		0.1377	1.0993	
Malay/Moor	-0.6006	0.2502	**	1.0819	0.2131	***
Max. yrs of schooling in hh	0.0792	0.0299	***	0.2402	0.0233	***
Income receivers in hh (no.)	-1.0370	0.0641	***	-1.2374	0.1362	***
Share of unemployed in hh	-2.1854	0.3450	***	-1.3868	0.3025	***
Rural	0.5245	0.2531	**	0.1855	0.1589	
Labor in agri (%)	0.0291	0.0056	***	-0.0007	0.0031	
Labor force ('000)	-0.0806	0.0196	***	-0.1109	0.0249	***
Gini coefficient (hh)	-5.4355	1.9343	***	-2.4369	1.6986	
Own house (% of hh)	0.0242	0.0068	***	0.0284	0.0107	***
19-34 yr pop (%)	0.0285	0.0448		0.1493	0.0324	***
Borrowing (% of income)	-0.0046	0.0071		-0.0150	0.0053	***
Constant	1.4161	1.5965		-2.6882	1.2803	
Number of observations					19,953	
Pseudo R^2					0.2192	
Log pseudo likelihood					-3727.1103	

Table 4.6: Multinomial logit model

null hypothesis of slope coefficients in internal and international migration being equal and favors a MNL model with three outcomes. As such the I estimate the multinomial logit model as follows,

$$P(Y_i = j|x_i) = \frac{\exp(x_i\beta_j)}{1 + \sum_{j=0}^{2} \exp(x_i\beta_j)} \tag{4.3}$$

where Y_i is the outcome for the i^{th} individual and j indicates the three outcomes: $j = 0$ for stayers, $j = 1$ international migrants and $j = 2$ internal migrants.

4.5 Empirical findings and discussion

The coefficient estimates of the empirically estimated multinomial logit model for equation (4.3) is presented in Table 4.6. The model is based on 19,953 observations with 519 internal migrants and 512 international migrants. The pseudo R^2 value of this model is 0.2192, and the standard errors reported are clustered at the district level. Based on this model the predicted probabilities for internal migration (evaluated at the respective values of each observation) range between 0.0000067 and 0.5945519 while the mean is 0.026 with a standard deviation of 0.049. Similarly, predicted probabilities for international migration at the respective values of each observation range between 0.0000005 and 0.7167377 while the mean is 0.0256603 with a standard deviation of 0.0495259. In the case of the base outcome – stayers, the mean of the predicted probabilities is 0.9483286 with a standard deviation of 0.0789561. These predicted probabilities range between 0.2773248 and 0.9999914.

The coefficient estimates of the MNL model indicates the direction of the relationship of the regressors with the probability of each outcome. The magnitude of this directional relationship can be discerned through marginal effects. The marginal effects discussed here are calculated at the mean values of the rest of the variables when only the relevant variable is allowed to change. For simplicity, the remainder of the discussion on marginal effects may not explicitly state this condition, but it is upheld throughout. The discussion on empirical findings is limited to the marginal effects applicable to the outcome of internal migration due to the scope of the study.[4] Moreover, given the possible issue of the absence of information about migrants who left the sending areas as a household, these findings have to be interpreted with caution.

In terms of individual level covariates for internal migration, the marginal effects of age and age squared are negative and positive, respectively and both are statistically significant. This shows that the increase in age by 1 year is associated with a 0.1 percentage point drop in the predicted probability for internal migration, while this decrease is smaller at higher ages. Interestingly, this is the exact opposite of the marginal effects for international migration which shows that an increase in the age by one year results in a 0.1 percentage point increase in the probability for international migration, and that this increase is smaller at higher ages. The marginal effect of age in the case

[4]This discussion is limited to statistically significant marginal effects.

	Internal		International		Stayers	
	dy/dx		dy/dx		dy/dx	
Age	-0.1033	***	0.1052	***	-0.0020	
	0.0192		0.0279		0.0345	
Age square	0.0009	***	-0.0018	***	0.0010	**
	0.0002		0.0004		0.0004	
Male	0.4791	**	-1.6337	***	1.1556	***
	0.1899		0.1991		0.3039	
Schooling (yrs)	-0.0614	**	-0.2043	***	0.2658	***
	0.0238		0.0288		0.0364	
Married	-0.1871		0.3597	**	-0.1728	
	0.1723		0.1490		0.2189	
Daughter	2.0223	***	0.9230	***	-2.9470	***
	0.1835		0.2273		0.2802	
Son	1.6736	***	1.7271	***	-3.4027	***
	0.2288		0.3003		0.3820	
Tamil	0.0353		-0.0191		-0.0162	
	0.2161		0.1483		0.2724	
Burgher	0.0438		0.1025		-0.1464	
	0.6787		0.8253		1.1495	
Malay/Moor	-0.5522	**	0.8126	***	-0.2607	
	0.2177		0.2359		0.3168	

Notes:
All values in percentage points.
Standard errors calculated by delta method are reported below marginal effects.
* =p<0.1, **=p<0.05, ***=p<0.001

Table 4.7: Marginal effects for migration after multinomial logit model

	Internal		International		Stayers	
	dy/dx		dy/dx		dy/dx	
Max. yrs of schooling in hh	0.0702	**	0.1789	***	-0.2493	***
	0.0295		0.0308		0.0447	
Income receivers in hh (no.)	-0.9322	***	-0.9176	***	1.8508	***
	0.0936		0.1100		0.1540	
Share of not employed in hh	-1.9729	***	-1.0213	***	2.9959	***
	0.3343		0.2503		0.3399	
Rural	0.4745	**	0.1350		-0.6098	
	0.2397		0.1319		0.2699	
Labor in agri (%)	0.0264	***	-0.0007		-0.0257	***
	0.0053		0.0022		0.0064	
Labor force ('000)	-0.0724	***	-0.0824	***	0.1548	***
	0.0179		0.0222		0.0323	
Gini coefficient (hh)	-4.9138	***	-1.7836		6.7010	**
	1.8719		1.2512		2.7754	
Own house (% of hh)	0.0217	***	0.0211	**	-0.0428	***
	0.0063		0.0086		0.0119	
19-34 yr pop (%)	0.0248		0.1113	***	-0.1362	***
	0.0407		0.0279		0.0472	
Borrowing (% of income)	-0.0041		-0.0112	**	0.0153	*
	0.0064		0.0045		0.0080	

Notes:
All values in percentage points.
Standard errors calculated by delta method are reported below marginal effects.
* =p<0.1, **=p<0.05, ***=p<0.001

Table 4.8: Marginal effects for migration after multinomial logit model (cont.)

of international migration is consistent with the concept that younger individuals have a longer life horizon, which makes the present value of the income differential between home and the potential destination large, "offering an enticement to move which diminishes with age" (Lucas, 1997, p. 731). The departure of internal migration from this phenomenon is puzzling. Nonetheless, it implies that an individual is more likely to be an internal migrant if he is young.

Compared to females, males have a 0.5 percentage point higher probability for internal migration, while on the contrary, for international migration males have a 1.6 percentage point lower probability.[5] However, in the case of relationship to head of the household, the marginal effect on a daughter is higher than that of a son. Specifically, relative to other categories a daughter of the head of the household has a 2 percentage points higher probability for internal migration. Where relative to other categories, a son has a 1.68 percentage point higher probability for internal migration. As such, compared to sons, daughters are more likely to become internal migrants. The relatively higher marginal effect for daughters is consistent with the pattern of high concentration of young female workers in manufacturing plants in export processing zones, which is the most popular occupation of female internal migrants in Sri Lanka (Abeywardene *et al.*, 1994; Hancock *et al.*, 2009).

Among the three ethnicity groups included in the model only one (Malay/Moor) is statistically significant. Specifically, Malay/Moor have a 0.55 percentage point lower probability for internal migrants than other ethnicity groups. Interestingly, despite fact that data reflects the period before the cessation of hostilities that affected Tamils in Sri Lanka, the estimated marginal effects for the indicator for Tamil is not statistically significant. This reflects that, when controlled for other variables, relative to other ethnicities, being a Tamil does not affect one's probability for internal migration in a statistically significant way. This absence of statistical significance in the marginal effect of Tamil is consistent across all three outcomes. However, given that Tamils migrated from Northern and Eastern Provinces to the rest of the island during this period of hostilities, this finding is contrary to expected. This contradiction may have resulted due to the out migration of entire Tamil families, leaving no household member in the sending areas to provide information to the survey.

[5]This negative marginal effect for males in the context of international migration is consistent with literature (Gamburd, 2004; IOM, 2009; SLBFE, 2003).

In the case of years of schooling, an additional year of schooling completed by an individual results in decreasing his probability for internal migration by 0.06 percentage points. Contrary to the negative marginal effect of one's own years of schooling, an additional year of schooling by the highest educated person at home results in increasing the probability to internal migration by 0.07 percentage points. This can be interpreted as internal migrants originating from families with a higher level of education, however, one is more likely to be an internal migrant if *he* is less educated. The positive marginal effect on the household level human capital variable is consistent with the intuition that better educated person in the household would have access to information about employment opportunities in other locations (Lucas, 1997). On the other hand, when the better educated individual in the household is younger than the migrant and is schooling during the absence of the migrant at home, this can also be interpreted that the higher income from migration has enabled a better education for other household members.

Among the other household level variables, the number of income receivers in a household has a negative effect on internal migration. Specifically, an additional income receiver at home is associated with a 0.93 percentage point decrease in the probability for internal migration for a person from that household. As noted by Lucas (1997, p. 730) among many alternative reasons "migrants often move to gain access to a higher income stream". As such, the negative correlation between the number of income earners and the probability for migration confirms that internal migration in Sri Lanka is associated with the motive of earning higher income. However, the negative marginal effect on the share of not employed persons at home is counter intuitive to the previous finding. A one unit increase in share of not employed persons at home results in dropping the probability for internal migration for a person in that household by nearly 2 percentage points. Among the two possible channels of the link between this variable and internal migration, this finding on the one hand, suggests that "not employed" in the household are not in the labor market, such as the sick or the elderly. On the other hand, this result could also be interpreted as having an internal migrant requires fewer income earners at home because migrant's remittance is large enough to compensate for the larger share of not employed persons.

The positive marginal effect on the indicator for rural areas shows that relative to those in urban areas, those in rural areas have a 0.5 percentage point higher probability for internal migration.[6]

[6]Rural sector includes the estate sector.

When characteristics of the rural areas such as high underemployment and high incidence of poverty are taken into consideration (Bandara, 1997; Karunaratne, n.d.), this positive marginal effect confirms that internal migration flows from rural areas are originated due to adverse socio-economic conditions in rural areas.

Among the contextual variables, a 1 unit increase in the share of district labor employed in agriculture results in increasing the probability for internal migration by 0.03 percentage points for its residents. The statistically significant positive marginal effect on the share of labor in agriculture indicate that districts with a larger share of its workers in agriculture are more likely to migrate. This finding and the positive marginal effect on the indicator for rural areas are consistent with the Lewis model Lewis (1954) which shows that surplus labor from agriculture sector leads to the exudos of labor from rural ares. This finding also shows that, despite Sri Lanka's the notable structural transformation in terms of the sectoral composition of employment as discussed in Section 4.1, still there is scope to further lower the share of labor in agriculture. Moreover, this implies that districts lagging behind in the structural transformation in terms of composition of employment are most likely to experience large outflows of internal migrants.

In terms of the district level labor force, a one unit[7] increase is associated with a 0.07 percentage point decrease in the likelihood of internal migration. For this variable, a result consistent with the Lewis model would expect to have a positive marginal effect. As such, the negative marginal effect can be interpreted that districts with a larger labor forces do not consist of surplus labor from the agriculture sector but on the contrary, these labor forces are sufficiently specialized and industrialized with agglomeration production externalities, that diminishes the requirement to seek employment elsewhere (Shukla & Stark, 1985, 1990).

The marginal effect on the Gini coefficient shows that a one unit increase in the Gini coefficient is associated with decreasing the probability for internal migration by 4.9 percentage points. An increase in the Gini coefficient is considered a worsening of income inequality in a district.[8] As such, the negative marginal effect on the Gini coefficient indicates that a worsening of income inequality results in decreasing the probability for internal migration for residents of that district. In the context of the relative deprivation hypothesis of Stark (1984) and Katz & Stark (1986), this

[7] A one unit increase corresponds to the increase of the labor force by 1000.

[8] A Gini coefficient of 0 indicates perfect equality and 1 indicates complete inequality.

marginal effect implies that on the one hand, the change in income distribution does not make individuals sufficiently worse off to have a sentiment of being relatively poor. On the other hand, the change in income distribution may be brought about by an increase in the income to a few individuals who previously felt relatively worse off. As such, when income inequality is worsening to their benefit, these individuals will no longer want to migrate internally.

The positive marginal effect on the share of the district population with ownership of a house shows that a one unit increase in the share of the district population with ownership of a household is associated with a 0.02 percent increase in the probability for internal migration. This result can be interpreted in multiple ways. First, when house ownership is considered to indicate wealth or assets, the finding that internal migrants originate from wealthy areas is contrary to expectation that less wealthy may migrate to improve their wealth. However, the finding is valid when the financial costs of migration are taken into consideration. Lucas (1997, p. 747) shows that migration involves financial cost in terms of relocation and settling costs etc., which can be reduced with "contacts in town among the educated/wealthy". In addition to reducing these costs through networks that the wealthy has, the wealthy can also better afford such initial costs. Secondly, the ownership of a house opens the possibility to borrow against the collateral of this asset. As such, even if the house ownership does not coincide with availability of liquidity, the prospect for borrowing against this asset makes migration more feasible.

However, in the context of the data issue of missing migrant families, the last two findings should be interpreted with caution. Specifically, families without ownership of a house might have migrated as a family without leaving anyone in the sending area. In such a case, the positive coefficient on this variable, which is counter to the hypothesis that less wealthy societies are more likely to migrate, may be due to such less wealthy families have already left the sending areas. Similarly, the poorest families would have already left as a household, contributing to the component of missing migrant in the data. In such a case, the the worsening of the income inequality depicted by an increase in the Gini coefficient would not make the remaining population in the sending areas feel sufficiently relatively poor to consider migrating.

The positive marginal effect on the share of district population in the ages 19-34 years is not statistically significant at conventional significance levels. However, given that the marginal effects for the same variable on the equation for international migration is of the same sign and in the

stayer equation it is the opposite sign while both are highly statistically significant, I consider this to be economically significant. Thus, I interpret that a one unit increase in the share of district population in ages 19-34 increases the probability for internal migration by 0.2 percentage points.

.

The above findings are based on data pertaining to 2003/4. Nonetheless, the findings are applicable in the current economic context of Sri Lanka, because economic indicators relevant for structural transformation have not changed significantly between 2003 and 2011. For instance, as shown in Table 4.9 during the 8 year period, the share of agriculture in the GDP and in employment declined by only 1.8 and 1.3 percentage points, respectively. Most importantly, the share of the labor force in rural areas has declined by only 0.6 percentage points. This relatively small decline for a period of 8 years confirms that large scale internal migration flows in Sri Lanka have not yet taken place, and these findings will be applicable once it takes place.

Indicator	2003	2011
Population ('000)	19,252	20,653
Labor force ('000)	7,654	8,108
19-34 yr pop (%)	24.5	25.0
Labor force participation %	48.9	48.1
Unemployment, (% of the labor force)	8.4	4.9
Labor force in rural areas (%)	49.6	49.0
Employment in agriculture (%)	34.0	32.7
GDP by Sectors (%)		
–Agriculture	13.7	11.9
–Industry	27.7	28.7
–Services	58.6	59.3

Source: extracted from CBSL (2011)

Table 4.9: Comparison of socioeconomic indicators in Sri Lanka : 2003 and 2011

4.6 Concluding remarks

This study uses a multinomial logit model to identify the contextual variables that affect internal migration in Sri Lanka. The empirical results provide six policy relevant findings. The first finding

shows that residents of rural areas have a higher probability to migrate than residents of urban areas, while the second shows that districts with larger concentration of labor in agriculture, which is a characteristics of lagging behind in the structural transformation process, are more likely to experience large outflows of internal migration. The third finding implies that the availability of agglomeration production externalities in districts with large labor forces will result in smaller outflows of internal migrants from such areas. Similarly,the fourth finding shows that districts with a more unequal income distribution will experience smaller outflows of internal migrants. The fifth and sixth findings show that districts with a larger share of population with ownership of assets and districts with a larger share of population in 19-34 years, will experience large outflows of internal migrants.

Additionally the study also finds that when controlled for other covariates, Tamils do not have a statistically different probability for migration. This finding , as well as fourth and fifth findings have to be viewed with caution due to the possibility of results being driven by the issue of lack migrants' information in CFS when an entire family migrated.

As per theories on structural transformation, once the initial level of structural transformation is achieved the overall growth in the economy would go hand in hand with internal migration. This importance of internal migration in growth and development is already incorporated into the development policy framework of the government of Sri Lanka (Department of National Planning, 2010). For instance, the policy framework identifies that by 2020 approximately 60 percent of the Sri Lankan population will live in urban areas, thus policies are in place to gear urban areas for the impending migration flows. However, the policy focus on areas where migration flows will originate is inadequate. In this setting, the findings of this study from the perspective of sending areas is valuable for repositioning the policy balance between sending and receiving areas of internal migrants.

Recommendations for such repositioning the balance in policy focus are as follows. First, migrants leave rural areas due to the relatively unfavorable socio-economic conditions. As such, there is a possibility for these conditions to further deteriorate with the departure of large number of migrants, because the development issues in the locations where a majority of the population lives will be prioritized over issues of where a minority lives. Such deterioration of conditions in rural areas can be minimized by developing the rural infrastructure to efficiently receive remittances

and to improve investment of these remittances in areas such as agriculture technology. Second, macroeconomic indicators show that Sri Lanka has expererienced a greater degree of transformation in terms of the decrease in the share of agriculture in composition of GDP and employment, than in terms of internal migration. Nonetheless, the findings show that migrants are more likely to originate from agriculture regions, and that migrants are from the younger cohorts in these regions. Such further decrease in labor in agrarian regions might have the potential to challenge agriculture productivity. As such, policies should be formulated to ensure that agriculture productivity in these areas are not adversely effected due to internal migration. Third, the departure of large proportions of young cohorts will have adverse implications on the composition of the population left behind. Therefore, internal migration might result in a lack of young role models for the very young cohorts of the population, as well as fewer people to care for the elderly. Thus development policies should account for the potential for the demand for the care of elderly in rural areas, with better health and social services. Fourth, given that areas with larger labor forces as well as areas with greater income inequality will not lose their labor force due to internal migration, such *rural* areas should also be included in the urban development strategies. These areas would serve as ideal candidates to be developed in to metro cities as noted by the development policy framework Department of National Planning (2010).

Chapter 5

Conclusion

This dissertation consists of three essays. The first two essays in Chapters 2 and 3 deal with unauthorized immigration in the US, while the third essay in Chapter 4 focuses on internal migration in Sri Lanka.

In Chapter 2 the objective was to develop an alternative methodology to estimate the number of unauthorized immigrants in the US. This objective was successfully achieved by the development of the microdata-based methodology which can be summarized in six main steps. In the first step I developed the derivation dataset, which is the merged dataset of the Legalized Population Survey (LPS) and the Immigration and Naturalization Services (INS) data. This merged derivation dataset consisting of actual unauthorized and legal immigrants was used in the second step to estimate the derivation model. In the third step, I adjusted the derivation model for (i) calibration in the American Community Survey (ACS) 2010, (ii) time dynamics between 1986 and 2010; and (iii) the inclusion of new variables. In the fourth step, I used the adjusted model to predict legal status of foreign born persons enumerated in the ACS 2010, and fifth I sum the product of each individual's predicted probability to be unauthorized and the number of persons represented by each individual, to estimate the aggregate number of unauthorized immigrants. In the sixth step, I determined a cutoff probability to assign predicted unauthorized immigrant status and obtain a cross sectional data set of predicted unauthorized immigrants.

This newly developed methodology is different from existing methodologies in many dimensions. First, the microdata based methodology is based on parameter estimates produced by cross-sectional models rather than relying on the difference between two aggregate level estimates as in

the case of the methodologies adopted by the Department of Homeland Security (DHS) and the Pew Hispanic Center (PHC). Second, unlike the PHC methodology, I relied on the larger of the two national data source with foreign born persons – the ACS, and unlike the DHS method, I did not use restricted administrative data along with the ACS data. As such, my microdata-based methodology uses only publicly available data sources, which is important for the potential application of this methodology by other researchers. Third, unlike the DHS version of the residual method which calls for repeated calculations for each characteristic, my microdata methodology aligns more closely with the PHC method with its greater flexibility. Nonetheless, the new methodology overcame the limitations in the PHC methodology by (i) developing a model based on observed authorized and unauthorized immigrants, (ii) assigning predicted legal status based on a larger number of control variables, (iii) relaxing the assumption that the occupational structure of unauthorized immigrants did not change significantly since 1986, and (iv) by not dictating that the number of foreign born persons assigned a predicted legal status should match a predetermined number.

The microdata-based methodology developed in Chapter 2 can be used in four distinct approaches. First, a sample weight based aggregate estimate of the population of adult unauthorized immigrants can be arrived by taking the summation of the product of predicted probability and sample weight of each observation. The sample weight based estimate shows that there were 7,700,869 adult unauthorized immigrants in the US on January 1, 2011. Second, for an existing estimate of the number of unauthorized immigrants, the microdata-based method can be used to reverse engineer and find the cutoff probability and the related margin of error in mis-classifying legal immigrants as unauthorized immigrants. An example of this application revealed that the DHS estimate for January 1, 2010 is associated with a higher margin of error than that of a similar estimate under the microdata based methodology. The third application of the the microdata based methodology is to produce an estimate of unauthorized immigrants for a predetermined cutoff probability (and/or margin of error) to assign predicted legal and unauthorized immigrant status. As part of this application I estimate an assignment based adult unauthorized immigrant population of 9,920,602 for January 1, 2011 at the cutoff probability of 0.7929 and the margin of error of 33.7 percent. As part of this third application I also presented a descriptive analysis of unauthorized immigrants and their children. The fourth application of the microdata based methodology is to use the predicted legal probability as a continuous variable in econometric applications without

assignment of legal status.

The objective of Chapter 3 was to explore if children of unauthorized immigrants are more likely to drop out of high school. This analysis focused on children in the ages of 16-18 years (inclusive of both years) with parents who were reported to be non-naturalized citizens in the ACS 2010. This study first used the microdata-based methodology developed in Chapter 2 as the auxiliary methodology to predict the legal probability of parents of the children in the sample. The main methodology of the analysis in Chapter 3 is a linear probability model.

When controlling for other covariates the study discerns four interesting findings. First, a child of two unauthorized immigrants is less likely to drop out of high school than a similar child of two authorized immigrant parents. Second, among two unauthorized immigrant parents, a non-citizen child is more likely to drop out of high school than a similar native-born child. Third, in the context of mixed families, a non-naturalized citizen child of an unauthorized immigrant mothers is more likely to drop out of high school than an otherwise similar native-born child. Fourth, in mixed families, the parent combination of highly likely unauthorized and highly legal results in a higher high school drop out probability than the parent combination of highly unauthorized and unsure legal. The fourth finding can be considered a weaker version of the first finding, while the third is a weaker version of the second finding.

Initially, the first and fourth findings may appear puzzling. Nonetheless, when the relatively lower degree of assimilation of unauthorized immigrants is taken into consideration, these two findings are consistent with the literature (Djajic, 2003; Driscoll, 1999; Shields & Behrman, 2004; Wojtkiewicz & Danato, 1995), which postulates that children of immigrants (without distinguishing between authorized and unauthorized) perform better than natives. The second and third findings imply that native-born children of unauthorized immigrants are better off that non-naturalized citizens with similar parentage. These two findings are particularly important in the context of the current policy interest on unauthorized immigrants and the proposed Development, Relief and Education for Alien Minors (DREAM) Act. The potential beneficiaries of the DREAM Act are among the non-naturalized citizen children. Moreover, it is highly likely that non-naturalized children of unauthorized immigrants are unauthorized immigrants themselves. Therefore, if the DREAM Act is passed, these minor unauthorized immigrants who currently have a higher probability of dropping out of high school, would have an incentive to complete high school education. However, the

analysis in Chapter 3 cautions that, relatively lower high school drop out probabilities associated with children of unauthorized immigrants should be factored in if a benefit-cost analysis of the DREAM Act were to be done in the future, if and when the said act is passed.

Moving away from the theme of unauthorized immigration, Chapter 4 was devoted to the analysis of internal migration in Sri Lanka. The objective of Chapter 4 was to discern how changes in contextual variables affect internal migration in Sri Lanka. The study adopts a development economics point of view with emphasis on structural transformation. The analysis used the district level disaggregation for contextual variables in a multinomial logit model, and data for the analysis is from the Consumer Finance and Socioeconomic Survey conducted by the Central Bank of Sri Lanka in 2003/4.

The findings of this study showed that residents of rural areas tend to migrate in order to overcome the unfavorable socioeconomic conditions in these areas. Similarly, districts lagging behind in the structural transformation process are more likely to experience large outflows of internal migrants, while the availability of agglomeration production externalities in districts with large labor forces will result in smaller outflows of internal migrants from such districts. Similarly, districts with more unequal income distribution will experience smaller outflows of internal migrants, while districts with a larger share of population with ownership of assets and districts with a larger share of population between 19-34 years of age would experience large outflows of internal migrants.

The findings in Chapter 4 have important policy implications in the post-conflict economic development setting in Sri Lanka in balancing the development policy focus between sending and receiving areas of internal migrants.

Bibliography

Abeywardene, Janaki, de Alwis, Romayne, Jayasena, Asoka, Jayaweera, Swarna, & Sanmugam, Thana. 1994. *Export Processing Zones in Sri Lanka: Economic Impact and Social Issues*. Working Paper 69. International Labor Organization.

Angrist, Joshua D., & Krueger, Alan B. 1991. Does Compulsory School Attendance Affect Schooling and Earnings? *The Quarterly Journal of Economics*, **106**(4), 979–1014.

Bandara, A. 1997. *Rural Poverty in Sri Lanka.* Paper presented at UNESCAP Regional Expert Meeting on Capability-Building to Alleviate Rural Poverty, held in Beijing.

Bean, Frank D., R. Corona R. Tuirn, & Woodrow-Lafield, K. 1998. The Quantification of Migration Between Mexico and the United States. *Pages 1–90 of: Migration Between Mexico and the United States, Binational Study, Vol. 1.* Mexican Ministry of Foreign Affairs, Mexico City and U.S. Commission on Immigration Reform, Washington, DC.

Black, Sandra E., Devereux, Paul J., & Salvanes, Kjell G. The More the Merrier? The Effect of Family Size and Birth Order on Children's Education. *The Quarterly Journal of Economics*, **120**(2), 669–700.

Boggess, Scott. 1998. Family Structure, Economic Status, and Educational Attainment. *Journal of Population Economics*, **11**(2), 205–222.

Borjas, George J., & Tienda, Marta. 1993. The Employment and Wages of Legalized Immigrants. *International Migration Review*, **27**(4), 712–747.

Bound, J., Jaeger, David A., & Baker, Regina M. 1995. Problems with Instrumental Variables

Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak. *Journal of the American Statistical Association*, **90**(430), 443–450.

Capps, Randolph, Ku, Leighton, Fix, Michael, Furgiuele, Chris, Passel, Jeffrey, Ramchand, Rajeev, McNiven, Scott, Perez-Lopez, Dan, Fielder, Eve, Greenwell, Michael, & Hays, Tonya. 2002. *How Are Immigrants Faring After Welfare Reform? Preliminary Evidence from Los Angeles and New York City.* Urban Institute, Washington D.C.

Card, David. 1999. The Causal Effect of Education on Earnings. *Pages 1801–1863 of:* Ashenfelter, O, & Card, David. (eds), *Handbook of Labor Economics*, vol. 3.

CBSL. 2011. *Socio Economci Statistics of Sri Lanka 2011.* Central Bank of Sri Lanka.

CFS. 2004. *Consumer Finance Socioeconomic Survey.* Central Bank of Sri Lanka.

Chiswick, Barry R., Lee, Yew Liang, & Miller, Paul W. 2005. Family Matters: The Role of the Family in Immigrants' Destination Language Acquisition. *Journal of Population Economics*, **18**(4), 631–647.

Cooper, Betsy, & ONeil, Kevin. 2005. *Lessons From The Immigration Reform and Control Act of 1986.* Policy Brief, Migration Policy Institute.

Cramer, J.S., & Ridder, G. 1991. Pooling states in the multinomial logit model. *Journal of Econometrics*, **47**, pp.267–272.

Department of National Planning. 2010. *The Development Policy Framework - The Government of Sri Lanka.*

Djajic, Slobodan. 2003. Assimilation of Immigrants: Implications for Human Capital Accumulation of the Second Generation. *Journal of Population Economics*, **16**(4), 831–845.

Driscoll, Anne K. 1999. Risk of High School Dropout among Immigrant and Native Hispanic Youth. *International Migration Review*, **33**(4), 857–875.

Eckstein, Zvi, & Wolpin, Kenneth I. 1999. Why Youths Drop Out of High School: The Impact of Preferences, Opportunities, and Abilities. *Econometrica*, **67**(6), 1295–1339.

Evans, William N., & Schwab, Robert M. 1995. Finishing High School and Starting College: Do Catholic Schools Make a Difference? *The Quarterly Journal of Economics*, **110**(4), 941–974.

Foster, Andrew D, & Rosenzweig, Mark R. 2008. Economic Development And The Decline Of Agricultural Employment. *Handbook of Development Economics*, **4**, 3051–3083.

Gamburd, Michele Ruth. 2004. Money That Burns like Oil: A Sri Lankan Cultural Logic of Morality and Agency. *Ethnology*, **43**(2), 167–184.

Gennetian, Lisa A. 2005. One or Two Parents? Half or Step Siblings? The Effect of Family Structure on Young Children's Achievement. *Journal of Population Economics*, **18**(3), 415–436.

Hancock, Peter, Middleton, Sharon, & Moore, Jamie. 2009. Export Processing Zones (EPZs), globalisation, feminised labour markets and working conditions: A study of Sri Lankan EPZ workers. *Labour and Management in Development*, **10**.

Hanson, Gordon. 2009. *The Economics and Policy of Illegal Immigration in the United States.* Immigration Policy Institute.

Harris, John R., & Todaro, Michael P. 1970. Migration, Unemployment and Development: A Two-Sector Analysis. *American Economic Review*, **60**(1), 126–142.

Hoeffer, M., Rytina, N., & Baker, B. 2011. *Estimates of the Unauthorized Immigrant Population Residing in the United States: January 2010.* Department of Homeland Security.

Hossain, M. Z. 2001. *Rural-Urban Migration in Bangladesh: A Micro-Level Study.* Presentation in a Poster Session on Interal Migration at the Brazil IUSSP Conference during August 20-24, 2001.

IOM. 2009. *International Migration Outlook - Sri Lanka 2008.* International Organization for Migration.

Janssen, K.J.M., Moons, K.G.M., Kalkman, C.J., Grobbee, D.E., & Vergouwe, Y. 2008. Updating methods improved the performance of a clinical prediction model in new patients. *Journal of Clinical Epidemiology*, **61**(1), 76 – 86.

Jefferies, Julian. 2008. Do Undocumented Students "Play by the Rules"? *Journal of Adolescent & Adult Literacy*, **52**(3), 249–251.

Karunaratne, Hettige Don. *Structural Change and the State of the Labour Market in Sri Lanka.* http://archive.cmb.ac.lk:8080/research/bitstream/70130/2231/1/75-1karunaratne.pdf.

Katz, Eliakim, & Stark, Oded. 1986. Labor Migration and Risk Aversion in Less Developed Countries. *Journal of Labor Economics*, **4**(1), 134–149.

Kossoudji, Sherrie A., & Cobb-Clark, Deborah A. 2002. Coming out of the Shadows: Learning about Legal Status and Wages from the Legalized Population. *Journal of Labor Economics*, **20**(3), 598–628.

Lewis, Arthur. 1954. Economic Development with Unlimited Supplies of Labor. *The Manchester School*, 400–449.

Lipton, Michael. 1980. Migration from rural areas of poor countries: The impact on rural productivity and income distribution. *World Development*, **8**(1), 1–24.

Lucas, Robert. 1997. Internal Migration in Developing Countries. *Pages 721–798 of:* Rosenzweig, Mark, & Stark, Oded (eds), *Handbook of Population and Family Economics.*

Lucas, Robert, & Stark, Oded. 1985. Motivations to remit: evidence from Botswana. *Journal of Political Economy*, **93**, 901–918.

Marcelli, Enrico A., & Ong, Paul M. 2002. 2000 Census Coverage of Foreign-Born Mexicans in Los Angeles County: Implications for Demographic Analysis.

Martin, Jack, & Ruark, Eric. 2010/2011. *Fiscal Burden of Illegal Immigration on United States Taxpayers.* Federation for American Immigration Reform.

Myrdal, Gunnar. 1968. *Asian Drama: An Inquiry into the Poverty of Nations.* New York: Twentieth Century Fund.

Orrenius, Pia M., & Zavodny, Madeline. 2003. Do Amnesty Programs Reduce Undocumented Immigration? Evidence from IRCA. *Demography*, **40**(3), 437–450.

Passel, Jeffrey, and Jennifer Van Hook, & Bean, Frank D. 2004. *Estimates of Legal and Unauthorized Foreign Born Population for the United States and Selected States, Based on Census 2000.* Urban Institute, Washington DC.

Passel, Jaffrey. 2005a. *Estimates of the Size and Characteristics of the Undocumented Population.* Pew Hispanic Center.

Passel, Jeffrey. 2005b. *Unauthorized Migrants Numbers and Characteristics.* Pew Hispanic Center.

Passel, Jeffrey. 2006. *The Size and Characteristics of the Unauthorized Migrant Population in the U.S. Estimates Based on the March 2005 Current Population Survey.* Pew Hispanic Center.

Passel, Jeffrey. 2007. *Unauthorized Migrants in the United States: Estimates, Methods, and Characteristics.* OECD Social, Employment and Migration Working Papers.

Passel, Jeffrey, & Clark, Rebecca. 1998. *Immigrants in New York Their Legal Status, Income, and Taxes.* The Urban Institute, Washington D.C.

Passel, Jeffrey, & Cohn, D'Vera. 2008. *Trends in Unauthorized Immigration: Undocumented Inflow Now Trails Legal Inflow.* Pew Hispanic Center.

Passel, Jeffrey, & Cohn, D'Vera. 2009. *A Portrait of Unauthorized Immigrants in the United States.* Pew Hispanic Center.

Passel, Jeffrey, & Cohn, D'Vera. 2010. *U.S. Unauthorized Immigration Flows Are Down Sharply Since Mid-Decade.* Pew Hispanic Center.

Passel, Jeffrey, & Cohn, D'Vera. 2011. *Unauthorized Immigrant Population: National and State Trends, 2010.* Pew Hispanic Center.

Perreira, Krista M., Harris, Kathleen Mullan, & Lee, Dohoon. 2006. Making It in America: High School Completion by Immigrant and Native Youth. *Demography*, **43**(3), 511–536.

Plug, Erik, & Vijverberg, Wim. 2005. Does Family Income Matter for Schooling Outcomes? Using Adoptees as a Natural Experiment. *The Economic Journal*, **115**(506), 879–906.

Ravenstein, E. G. 1885. The Laws of Migration. *Journal of Royal Statistical Society*, **48**(2), 167–227.

Rector, Robert. 2006. *Importing Poverty: Immigration and Poverty in the United States: A Book of Charts.* Special Report No. 9.

Redford, Arthur. 1926. *Labor Migration in England, 1800-1850.* Manchester: Univ. Press.

Rivera-Batiz, Francisco L. 1999. Undocumented Workers in the Labor Market: An Analysis of the Earnings of Legal and Illegal Mexican Immigrants in the United States. *Journal of Population Economics*, **12**(1), 91–116.

Ruggles, J., Trent Alexander, and Katie Genadek, and Ronald Goeken and Matthew B. Schroeder, & Sobek, Matthew. 2010. *Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database].* Minneapolis: University of Minnesota.

Rytina, Nancy. 2011. *Estimates of the Legal Permanent Resident Population in 2010.* Department of Homeland Security.

Sahato, Gian S. 1968. An Economic Analysis of Internal Migration in Brazil. *he Journal of Political Economy*, **76**(2), 218–245.

Schultz, Theodore W. 1962. Reflections on Investment in Man. *Journal of Political Economy*, **70**(5), 1–9.

Shields, Margie, & Behrman, Richard. 2004. Children of Immigrant Families: Analysis and Recommendations. *The Future of Children*, **14**(2), 4–15.

Shukla, V., & Stark, O. 1985. On agglomeration economies and optional migration. *Economics Letters*, **18**, 297–300.

Shukla, V., & Stark, O. 1990. Policy comparisons with an agglomeration effect-augmented dual economy model. *Journal of Urban Economics*, **27**, 1–15.

Sjaastad, Larry A. 1961. *Income and Migration in the United States.* Ph.D. thesis, Univ. of Chicago.

Sjaastad, Larry A. 1962. The Costs and Returns of Human Migration. *Journal of Political Economy*, **70**(5).

SLBFE. 2003. *Statistical Handbook on Migration, 2002.* Sri Lanka Bureau of Foreign Employment, Colombo.

Stark, Oded. 1984. Rural-to-urban migration in LDCs: a relative deprivation approach. *Economic Development and Cultural Change*, **32**, 475–486.

Stark, Oded, & Levhari, David. 1982. On Migration and Risk in LDCs. *Economic Development and Cultural Change*, **31**(1), 191–196.

Stark, Oded, & Lucas, Robert. 1988. Migration, remittances and the family. *Economic Development and Cultural Change*, **36**, 465–481.

Steyerberg, Ewout W., Borsboom, Gerard J. J. M., van Houwelingen, Hans C., Eijkemans, Marinus J. C., & Habbema, J. Dik F. 2004. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Statistics in Medicine*, **23**(16), 2567–2586.

Todaro, Michael P. 1969. A Model for Labor Migration and Urban Unemployment in Less Developed Countries. *American Economic Review*, **59**(I), 138–48.

UNESCAP. 2008. *Statistical Yearbook for Asia and the Pacific 2008*. Economic and Social Commission for the Asia Pacific of the United Nations.

Warren, Robert, & Passel, Jeffrey. 1987. A Count of the Uncountable: Estimates of Undocumented Aliens Counted in the 1980 United States Census. *Demography*, **24**(3), 375–393.

Wojtkiewicz, R.A., & Danato, K.M. 1995. Hispanic Educational Attainment: The effect of Family Background and Nativity. *Social Forces*, **74**(2), 559–574.