

Sample Attrition in the presence of Population Attrition

*Seik Kim*¹

Department of Economics

University of Washington

seikkim@u.washington.edu

<http://faculty.washington.edu/seikkim/>

January 15, 2009

Abstract

This paper develops a method that accounts for non-ignorable sample attrition in the presence of population attrition for use with a non-representative panel sample. The method is applied to obtain attrition-correcting weights for the native and immigrant samples in the matched Current Population Survey (CPS). Of the two samples, the immigrant sample suffers from sample attrition due to changes in residence as well as population attrition caused by selective return migration. When there is population attrition, the second period cross-section is not representative of the first period population. Therefore, the existing sample attrition-correcting method developed by Hirano, Imbens, Ridder, and Rubin (2001) and Bhattacharya (2008) cannot be applied. We resolve this problem by generating a counterfactual, but representative cross-section prior to applying their method. The counterfactual sample can be obtained by weighting the second period cross-section by one minus the probability of population attrition. We show that the sample attrition and the population attrition processes are separately identified. This is useful because samples usually do not indicate which missing observations are due to sample attrition and which are due to population attrition. The attrition-correcting weights, once obtained, can be used in various studies of immigration using the CPS.

Keywords: Immigration, Population Attrition, Sample Attrition

JEL Classification Codes: C23, C81, J61

¹This paper is based on chapter one of my dissertation. I am grateful to my advisors, Joseph Altonji, Yuichi Kitamura, Fabian Lange, and Mark Rosenzweig. I have also benefited from helpful comments made by Donald Andrews, Debopam Bhattacharya, Keisuke Hirano, Thomas Lemieux, Taisuke Otsu, Peter Phillips, and seminar participants at Purdue University, Seoul National University, University of Washington, Vanderbilt University, and Yale University.

1 Introduction

The first wave of a longitudinal sample is usually designed to represent a target population. In consecutive waves, however, the sample tends to lose its representativeness due to nonrandom attrition.² Attrition is the result of many factors. One kind of attrition, which we call sample attrition, occurs when a respondent is not interviewed while he or she is in the population. A simple example is temporary absence. Another kind of attrition, which we call population attrition, occurs when a respondent drops out of the sample because he or she drops out of the population. An example is decease. Population attrition is often very small and is ignored in analyses. In some cases, however, the distortion that stems from population attrition can be large, and therefore, one may want to control for this particular type of attrition.

In an open economy, where international migration is possible, not being able to locate a respondent does not necessarily mean sample attrition. For example, consider a longitudinal sample on native-born and foreign-born populations in the United States. On one hand, when a native-born respondent is not traced in the second period, it would be natural to presume that the person is still somewhere in the United States.³ This is sample attrition. A cross-section of the U.S. population in the second period will select this missing person as well as all the other U.S. residents with an equal probability.⁴

On the other hand, when a foreign-born respondent is missing in the second period, it is difficult to conclude whether the person is in the United States or has gone back to his or her home country. If the person is still in the United States, this person will get an equal probability of being selected in the cross-section as all other U.S. residents. This is sample attrition. However, if the person has emigrated from the United States, this person has no chance of being selected in the cross-section. This is population attrition. When there is population attrition, the second period population becomes a nonrandom subset of the first period population.⁵ In consequence, the second period cross-section is not representative of the first period population.

The distinction between sample attrition and population attrition is important because additional information from “representative” cross-sections can be useful in accounting for attrition in longitudinal studies. A recently developed method by Hirano, Imbens, Ridder, and Rubin (2001) and Bhattacharya (2008) uses the availability of representative cross-sections as the basis for weighting the persons in a balanced part of

²This paper addresses attrition due to complete nonresponse, but the discussion can be extended to attrition due to item nonresponse. Complete nonresponse refers to the failure to collect any information from a respondent. Item nonresponse occurs when a respondent fails to answer certain items.

³This person might be missing because of decease, emigration, or other reasons, but these possibilities (especially emigration) are relatively low and negligible compared to the foreign-born population in the United States.

⁴Throughout this paper, we assume that there is no nonresponse problem in the first wave of a sample. It also means that there is no nonresponse problem in a cross-section sample.

⁵The foreign-born population in the United States is not stationary over time due to return migrants and new immigrants. New immigrants, however, are not of interest because we control for the year of entry.

the panel. Their method is more general than existing methods such as the missing at random or the sample selection models, because it allows non-ignorable sample attrition. More precisely, Hirano et al. show that the attrition process, as a function of both past and current variables, can be identified under fairly flexible assumptions up to a known link function such as the logit or probit.⁶ The identification strategy relies on the availability of representative cross-section samples of the target population throughout the entire sampling period of the panel sample. The attrition-correcting weighting function is given by the inverse of one minus the probability of sample attrition. When there is attrition in the population of interest, however, there are no representative cross-sections for consecutive periods, and the existing method should not be applied.

This paper develops a method that accounts for sample attrition in the presence of population attrition for use with panel data models where at least one cross-section is representative of the target population while the panel and other cross-section samples are not. The method separately identifies sample attrition and population attrition. This is useful because samples usually do not indicate which missing observations are due to sample attrition and which are due to population attrition. For simplicity, we consider a two-period panel sample.⁷ Assume that the first period cross-section is representative, but the second is not. The key estimation strategy is generating a representative counterfactual second period cross-section prior to applying the existing sample attrition-correcting method. The counterfactual sample can be obtained by weighting the second period cross-section by one minus the probability of population attrition. We prove that the population attrition function can be identified without knowing the type of attrition when the function is determined by variables of known transition probability.⁸ These variables include deterministic variables such as age or birth country. Then we prove that the population attrition-adjusted counterfactual cross-section can be used with Bhattacharya’s estimation procedure.

We apply the technique to a longitudinal study of the foreign-born population in the United States. In principle, longitudinal data on native-born and foreign-born populations, by tracking specific individuals over time, offers the huge advantage of permitting one to control for fixed unobserved heterogeneity. In practice, longitudinal analysis of U.S. immigrants has been limited by two key factors. First, sample sizes of immigrants in U.S. panels such as the Panel Study of Income Dynamics (PSID) or the National Longitudinal Survey of Youth 1979 (NLSY79) are too small. Second, the use of panel data gives rise to nonrandom sample attrition, not to mention population attrition caused by selective return migration.⁹ Given that an ideal sample is not

⁶We use the terms “attrition process” and “attrition function” interchangeably.

⁷Extension to longer panels is analogous to Bhattacharya (2008).

⁸Of course more restrictive assumptions are needed about the factors that drive it than is necessary to handle sample attrition.

⁹For example, if persons with negative wage shocks are more likely to drop out of the sample, panel data estimates will overstate the growth of wages. In addition, if unsuccessful immigrants tend to return to their home country, stayers will on average earn higher wages than return migrants. Consequently, estimates using only stayers will tend to overstate relative labor market

available, it is desirable to have a data set which enables us to control for both sample attrition and population attrition. As we show, one may do so with the Current Population Survey (CPS). We address the sample size problem by using the Merged Outgoing Rotation Groups (MORG) of the CPS. The sample is also known as the matched CPS. It is a collection of two-year panels and has the crucial advantage of being much larger than alternative panel data sets. In the matched CPS, however, attrition is particularly severe as the survey does not follow households who change residences.

To address the sample attrition problem in the presence of population attrition, we exploit the cross-sectional feature of the CPS. Suppose that the two-year panel of 1994-1995 is of interest. The CPS provides cross-sections for 1994 and 1995. The 1995 cross-section is not representative of the 1994 population. First, we use the 1994 cross-section as the basis for generating a representative counterfactual 1995 cross-section. Then the 1994 and counterfactual 1995 cross-sections are used as the basis for estimating attrition-correcting weighting functions. Finally, we assign weights for the persons in the balanced part of the 1994-1995 panel. We estimate the weighting functions for the matched CPS between 1994-2004. We do it year by year because residential mobility and return migration may vary by year and across samples.

Empirical results suggest that sample attrition is lower among older and married individuals for both native and immigrant samples and is lower among those who have citizenship and who have stayed longer in the United States for foreign-born samples. We find that education is positively correlated with sample attrition for natives, but is negatively correlated for immigrants. In general, both first period and second period wages are negatively correlated with the sample attrition rate, and those who are not working are more likely to stay in the sample than those who are working. This association is less significant for immigrants. Finally, immigrants from Europe tend to stay in the sample more than the other immigrants. The estimates of the population attrition function suggest that education is positively correlated with the probability of staying in the United States. The estimation results do not support the hypothesis that the rates of return migration decline with time spent in the United States. From the estimated functions of sample attrition and population attrition, we calculate the attrition-correcting weights. These weights, once constructed, can be used in various studies of immigration such as those addressing assimilation of immigrants. An application of this method on economic assimilation of foreign-born workers in the United States can be found in Kim (2008).

The paper proceeds as follows. Section 2 introduces the data set and presents summary statistics. The section highlights the problems of sample attrition and population attrition in the CPS. Section 3 presents identification and estimation of a panel data model with sample attrition in the presence of population attrition.

performance of immigrants compared to natives.

It also reviews the existing sample attrition-correcting method when the target population is stationary. Section 4 presents estimation results and discusses empirical findings. Section 5 offers conclusions and potential extensions.

2 Data Description

In this section, we discuss how one uses the CPS sample to account for sample attrition in the presence of population attrition. The CPS is a random sample of addresses. In consequence, attrition is directly related to residential mobility within the U.S. and return migration. The rates of sample attrition and population attrition for 1994 to 2004 are about 22-40% and 3% per year, respectively. As we discuss later, sample attrition has a larger impact on estimation results than population attrition does.

2.1 The Current Population Survey

The CPS sample is a collection of representative cross-sections. The CPS collects a sample of approximately 56,000 housing units from 792 sample areas. Each month, data are collected from the sample housing units on demographic and labor force characteristics of the civilian non-institutional population 16 years of age and older. The design of the CPS is as follows. A housing unit is interviewed for 4 consecutive months, is dropped out of the sample for the next 8 months, is brought back in the following 4 months, and then is retired from the sample.¹⁰ If a household is included in either the first or the last 4 months of the interview periods, it is said that the household is in the rotation group. The pre-selected housing units are kept unchanged over the interview periods. If the occupants of a dwelling unit move, the new occupants of the unit are interviewed. Although the interviewees may be replaced by new occupants within the sampling periods, the CPS provides a representative cross-section of the target population because the random sample of housing units is kept fixed.

An interesting feature of the CPS sample is its rotation scheme. Selected questions on labor market information, such as usual weekly earnings and usual weekly hours worked, are asked only in the last interview of each 4-month rotation group. The sets of households in the fourth or eighth month are called the outgoing rotation groups. If records from the 4th and 8th interviews are appended, we get repeated observations on the same individuals. The appended sample is called the Merged Outgoing Rotation Group (MORG) data.

¹⁰About 3/4 of the first and fifth interviews are conducted by visiting. In other interview months, almost 90% of the interviews are conducted over the phone. The rotation scheme ensures that in any 1 month, one-eighth of the housing units are interviewed for the first time, another eighth is interviewed for the second time, and so on. That is, after the first month, 6 of the 8 rotation groups will have been in the survey for the previous month; there will always be a 75 percent month-to-month overlap. When the system has been in full operation for 1 year, 4 of the 8 rotation groups in any month will have been in the survey for the same month, 1 year ago; there will always be a 50 percent year-to-year overlap.

By construction, an individual appears only once in a year, but may reappear in the following year. Due to the 4-8-4 rotation scheme, the CPS MORG is a collection of two-year panels. The 1994-1995 panel, for instance, contains the individuals in the households which enter the survey scheme between October 1993 and September 1994. Similarly, the 1995-1996 panel contains the individuals in the households which enter the survey scheme between October 1994 and September 1995. As the sampling periods of two adjacent short panel data sets overlap, short panels may mimic a longitudinal sample if combined properly.

An overlapping rotating panel data set shares most of the advantages of usual panel data sets and is superior in some dimensions. First, the sample has a longitudinal feature. This means that usual panel data models, such as the first difference or the fixed effects models, can be used to control for individual specific permanent components. Second, the rotating panel that we use, the CPS MORG, is large, which makes it even more powerful than a usual panel, such as the PSID or the NLSY79. Sample sizes matter in immigration studies because foreign-born persons, after all, are minorities. Third, the sample serves as a representative cross-section of the target population for any given time period. This property is the key in identifying sample attrition and population attrition processes.

2.2 Sample Attrition and Population Attrition: Summary Statistics

Since 1994, the CPS includes information on international migration, such as year of entry to the United States and country of birth along with demographic and labor market information, such as age, schooling, marital status, earnings per hour or week, usual hours of work, and labor market status. The sample used in this analysis is drawn from the CPS MORG between 1994 and 2004. We take a sample of foreign-born and native-born men of ages 18-64.¹¹ We define an individual as matched if the individual appears twice in the CPS MORG. In order to examine differences based on ethnic origin, we divide the foreign sample into 4 groups: immigrants from Central and South America, from Europe (including Australia, New Zealand, and Canada), from Asia, and from other countries.¹² The group of the other countries consists of immigrants from Africa, Oceania, and unclassified ones. The last group is of little interest due to its small sample size and heterogeneity. Summary statistics are reported in Kim (2008).

¹¹The foreign sample includes foreign-born men who were not U.S. citizens at the time of birth. Following Warren and Peck (1980), our foreign sample consists of persons born outside the United States, the Commonwealth of Puerto Rico, and the outlying areas of the United States. Foreign-born persons may have acquired U.S. citizenship by naturalization or may be in illegal status. The reference group consists of native-born white men. The native sample includes persons born in the United States, but excludes persons born in the Puerto Rico and the outlying areas.

¹²We combine Australia, New Zealand, and Canada with Europe because of sample size considerations and so that immigrants from countries that are predominantly white and are at a similar stage of political and economic development are grouped together. We refer to the group as Europe. The data do not identify mother tongue. The impact of language proficiency has been studied in a large literature. LaLonde and Topel (1997) provide a survey.

Matching is directly related to residential mobility and return migration as the housing units in the sample are kept fixed over the interview periods, provided that the non-interview rate is low.¹³ Between 1994 and 2004, the attrition rates are 28-40% among the immigrant samples and 22-32% among the native samples. In practice, matching is not possible between June 1994 - August 1995 and June 1995 - August 1996 due to sample redesign. If samples in 1994-1995 and 1995-1996 are excluded, the attrition rates are 28-35% among the immigrant samples and 22-29% of the native samples. The gaps between the foreign and native attrition rates are stable in these periods ranging 6-8% points. A part of the gap in the attrition rates may be due to return migration. Attrition rates are reported in Table 2. Foreign-born persons from Central and South America tend to attrite more than those from Europe and Asia. The consequence of nonrandom attrition, however, has not been addressed in immigration studies using the matched CPS.¹⁴ We find substantial sample attrition bias.

The United States stopped collecting information on return migrants in 1957. To estimate the rates of return migration, we exploit the structure of the CPS MORG. As housing units in the sample are kept fixed over the sampling period, the decrease in the sample size of immigrants will imply return migration. Using the panels prior to trimming individuals with extreme wages or negative experience, Table 1A provides the ratios of persons staying in the United States (one minus the population attrition rates) by year of entry. For instance, the cell in the first row and first column indicates that in the 1st year of the 1994-1995 panel, there were 5329 foreign-born persons in the United States. Then we count the number of foreign-born persons in the 2nd year of the 1994-1995 panel, which is 5331. We take the ratio between these numbers and get 1.00 (=5331/5329). This roughly means that little outmigration occurred during this period. Similarly in 1995-1996, the numbers of the foreign-born persons in the first and the second years are 5417 and 4605, respectively. It implies that about 15% (=1-4605/5417) of the foreign-born population in 1995 left the United States in 1996.

Conceptually, it is impossible to have the stay rate exceed unity (or the outmigration rate below zero). Estimates above unity could arise from sampling error and/or if the reentering foreign-born persons report their previous entry years. In the sample, values greater than unity are observed frequently, implying that sampling errors and measurement errors are relatively large. Taking this into account, the last column reports

¹³The average yearly non-interview rates for the CPS in the early 1990's are as low as 4-7%. This non-interview rate is comparable with the initial non-response rate of the National Longitudinal Survey of Youth 1979 (NLSY79), which is 10%. The Census Bureau classifies the noninterviews into three types. Type A noninterviews are for household members that refuse, are absent during the interviewing period, or are unavailable for other reasons. Type B noninterviews include a vacant housing unit (either for sale or rent), a unit occupied entirely by individuals who are not eligible for a CPS labor force interview, or other reasons why a housing unit is temporarily not occupied. Type C noninterviews are for addresses that may have been converted to a permanent business, condemned or demolished, or fall outside the boundaries of the segment for which it was selected.

¹⁴While many papers have used the matched CPS, only two that we are aware of focus on immigration: Duleep and Regets (1997a) and Bratsberg, Barth, and Raaum (2006).

the stay probability over the entire sample period. The last column of the first row reports that 25.2% ($=1-0.768$) of the foreign-born population who arrived in the United States in 1994 or before left the country in 2004.¹⁵ On average, 2.6% ($=1-0.974$) of the foreign-born population emigrates from the United States. The stay probability by ethnic origin is reported in Table 1B.

3 Correcting for Attrition

This part consists of two sections. Section 3.1 introduces the methodology to correct for sample attrition when the population is stationary. It is based on Hirano, Imbens, Ridder, and Rubin (2001) and Bhattacharya (2008). The estimation strategy consists of two steps. In the first step, we estimate the sample attrition function and obtain the weights for individuals in the balanced part of the longitudinal sample. This balanced panel is also called the matched sample. In the second step, we estimate the main model using the matched sample along with the weights. Section 3.2 develops an estimation strategy when sample attrition and population attrition occur at the same time. The estimation strategy consists of three steps. In the first step, we estimate the population attrition function and weight the second period cross-section. Next, we apply the two-step sample attrition-correcting method introduced in Section 3.1. For presentation purposes these methods are presented in multiple steps, but all these steps can be done simultaneously.

3.1 Correcting for Sample Attrition when the Target Population is Stationary

Assume that there is no population attrition. Consider a two-period panel data set where all the interviewees respond in the first period but some do not respond in the second period. Denote $D_S = 1$ when an individual is in the sample (or responds) in the second period and $D_S = 0$ when an individual is not in the sample (or does not respond) in the second period. Now it is possible to construct a balanced longitudinal sample by collecting all the individuals with $D_S = 1$: we call the sample the matched sample.

Suppose the model of interest is identified by a conditional moment restriction

$$E[m(y_1, y_2, x_1, x_2, \theta) | x_1, x_2] = 0, \quad \text{w.p.1,} \quad (1)$$

uniquely when $\theta = \theta_0$, where y is the endogenous variable, x is a vector of exogenous variables, θ is a parameter vector, $m(\cdot)$ is a known function, and the subscripts denote the period. We do not observe the joint

¹⁵This estimate is consistent to other empirical findings. For instance, Warren and Peck (1980) estimate that more than 1/6 of total immigrants admitted during the 1960s emigrated by the end of the decade.

distribution of (y_1, y_2, x_1, x_2) due to nonresponse. Instead we observe the joint distribution of the matched sample, $(y_1, y_2, x_1, x_2) | D_S = 1$. However,

$$E [m (y_1, y_2, x_1, x_2, \theta_0) | x_1, x_2] \neq E [m (y_1, y_2, x_1, x_2, \theta_0) | x_1, x_2, D_S = 1]. \quad (2)$$

Therefore, just using the matched sample will result in an inconsistent estimator of θ_0 .

Now assume that in addition to the two period panel there is a representative cross-section available in the second period.¹⁶ This cross-section is called the refreshment sample. Hirano, Imbens, Ridder, and Rubin (2001) specify the attrition process using the second period cross-section allowing the attrition to depend on the endogenous variables in the second period. This is a substantially more general attrition-correcting method than standard attrition-correcting methods which only allow dependence on the variables in the first period. The native samples of the CPS, by construction, include representative cross-sections for both periods. This is because a new two-year panel is activated from the target population in each year.

In order to specify the attrition function, assume that attrition is a function of u_1 , u_2 , and v , where u_1 and u_2 are vectors of time-varying variables in periods 1 and 2, respectively, and v is a vector of time invariant variables. For instance, u_1 (or u_2) is a vector of the endogenous variable, y_1 (or y_2), and time-varying exogenous variables in x_1 (or x_2) and v is a vector of time-invariant exogenous variables.¹⁷ It is worth noticing that u_2 is observed because the second period cross-section is available. This fact is crucial to the method.

We need to know $f(u_1, u_2, v)$ to calculate the LHS of (2), but it is not observed when there is attrition. What we do know, however, is the joint density of non-attriting individuals, $f(u_1, u_2, v | D_S = 1)$, along with marginal densities, $f(u_1, v)$ and $f(u_2, v)$. Notice that $f(u_1, v)$ is known from the first period cross-section. Notice also that $f(u_2, v)$ is known from the second period cross-section. A fact that is crucial to the method is

$$f(u_1, u_2, v) = \frac{f(u_1, u_2, v | D_S = 1) \Pr(D_S = 1)}{\Pr(D_S = 1 | u_1, u_2, v)}.$$

So if we come up with a candidate for $\Pr(D_S = 1 | u_1, u_2, v)$, we can obtain a consistent estimator of θ_0 .

Hirano et al. prove that $\Pr(D_S = 1 | u_1, u_2, v)$ is nonparametrically just-identified up to a known link function, $g(\cdot)$, if it takes an additive non-ignorable form:

$$\Pr(D_S = 1 | U_1 = u_1, U_2 = u_2, V = v) = g(k_0(v) + k_1(u_1, v) + k_2(u_2, v)), \quad (3)$$

¹⁶It is implicitly assumed that the first wave of the longitudinal sample is representative of the target population. The first wave sample serves as a representative cross-section sample.

¹⁷The attrition function does not have to be determined by the same variables in the main model (1). The variables in (u_1, u_2, v) may or may not include the variables in (y_1, y_2, x_1, x_2) .

where $k(\cdot)$ are unknown functions with the normalization of $k_1(0, v) = k_2(0, v) = 0$ and the known link function $g(\cdot)$ is a bounded strictly increasing function such that $\lim_{r \rightarrow -\infty} g(r) = 0$ and $\lim_{r \rightarrow \infty} g(r) = 1$. Identification stems from the fact that we observe two marginal densities, $f(u_1, v)$ from the year-one cross-section and $f(u_2, v)$ from the year-two cross-section, and $f(u_1, v)$ and $f(u_2, v)$ obey

$$\begin{aligned} f(u_1, v) &= \int \frac{\Pr(D_S = 1)}{g(k_0(v) + k_1(u_1, v) + k_2(u_2, v))} f(u_1, u_2, v | D_S = 1) du_2, \\ f(u_2, v) &= \int \frac{\Pr(D_S = 1)}{g(k_0(v) + k_1(u_1, v) + k_2(u_2, v))} f(u_1, u_2, v | D_S = 1) du_1, \end{aligned} \quad (4)$$

for almost all (u_1, u_2, v) .

In estimation of (4), the standard semiparametric methods cannot be applied because the attrition function is defined implicitly by nonlinear integral equations. Bhattacharya (2008) shows that the identification conditions in (4) can be transformed into conditional moment restrictions:

$$\begin{aligned} 1 &= E \left[\frac{D_S}{g(k_0(v) + k_1(u_1, v) + k_2(u_2, v))} \middle| u_1, v \right] \quad \text{w.p.1,} \\ 1 &= E \left[\frac{D_S}{g(k_0(v) + k_1(u_1, v) + k_2(u_2, v))} \middle| u_2, v \right] \quad \text{w.p.1.} \end{aligned} \quad (5)$$

The transformed identification conditions in (5) can be estimated, for instance, by the sieve minimum distance (SMD) developed by Ai and Chen (2003). When we specify a parametric attrition process, one can use the smoothed empirical log-likelihood (SEL) developed by Kitamura, Tripathi, and Ahn (2004). As $g(k_0(v) + k_1(u_1, v) + k_2(u_2, v))$ and $\Pr(D_S = 1)$ are estimable, we can construct the attrition-correcting weighting function

$$C(u_1, u_2, v) = \frac{\Pr(D_S = 1)}{g(k_0(v) + k_1(u_1, v) + k_2(u_2, v))}. \quad (6)$$

Then, we weight the matched sample by (6) and estimate

$$E[m(y_1, y_2, x_1, x_2, \theta_0) \cdot C(u_1, u_2, v) | x_1, x_2, D_S = 1] = 0, \quad \text{w.p.1,} \quad (7)$$

to obtain a consistent estimator of θ_0 . In sum, the model with attrition can be estimated consistently by assigning attrition-correcting weights to the individuals in the matched sample. The weighting function is proportional to the inverse of one minus the probability of attrition. Intuitively, the LHS of (5) is equivalent to weighting the sample with the inverse of one minus the probability of attrition, $1/g(k_0(v) + k_1(u_1, v) + k_2(u_2, v))$.

The attrition-correcting method has at least four attractive features. First, the sample attrition function

for a longitudinal sample is identified nonparametrically under relatively weak conditions. In particular, it is identified provided the attrition function is additive non-ignorable with a known link function such as the logit or probit and representative cross-sections are available. The constraint of additive non-ignorable assumption reduces the dimension of the attrition function of our interest.¹⁸ Second, different from the Heckman’s self-selection model, no exclusion restriction is needed to estimate the attrition function. Heckman’s solution requires at least one exogenous variable affecting selection that does not appear in the structural equation. The key to the approach used here is the availability of additional information than the self-selection setup. In consequence, there is no need of making assumptions on unobservables.

Third, the correction is robust to individual fixed effects. This is because each individual gets his or her unique weight which is a function of the characteristics in the first and second periods. Therefore, the usual fixed effects strategies for panel data models can be used to control individual heterogeneity. Fourth, the weighting function estimates do not have to be interpreted as causal effects. They simply describe the state. For instance, the wage may affect residential mobility, but the latter may affect the former, too. Therefore, even if there is reverse causality problem from mobility to wages, the weighting function estimates successfully describe the attrition function in a statistical sense.

3.2 Correcting for Sample Attrition in the presence of Population Attrition

When the target population is nonstationary and the model of interest involves a counterfactual situation of what if the population had remained stationary, the attrition-correcting technique has to be modified. Consider a pair of representative cross-section data sets where some of the interviewees drop out of the population in the second period. Denote $D_P = 1$ when an individual is in the population (or stays in the United States) in the second period and $D_P = 0$ when an individual is not in the population (or leaves the United States) in the second period. Now an individual is in the matched sample if $D_P = 1$ and $D_S = 1$. Similarly, an individual stays in the U.S. but does not respond in the second period if $D_P = 1$ and $D_S = 0$. An individual who leaves the U.S. in the second period is denoted by $D_P = 0$. A combination of $D_P = 0$ and $D_S = 1$, where an individual leaves the country and responds in the second period, is not possible. As a result, being in the matched sample, $D_S = 1$, also implies residing in the U.S. at the same time, $D_P \cdot D_S = 1$.

Again, the model of interest is identified by a conditional moment restriction (1). An available data set is

¹⁸As an additive non-ignorable attrition model includes the first and the second period variables, it nests models of the selection on observables and the selection on unobservables. Therefore, two models can be distinguished by use of the second period cross-section. The data provides testable restrictions on those models. An additive non-ignorable model, however, rules out interactions between the variables in the first and the second periods. For instance, consider $wage_1$ and $wage_2$. Panel attrition can depend on $\log wage_2 - \log wage_1$ but not on $(wage_2 - wage_1) / wage_1$, although both measure wage growth. In the Appendix, models of the selection on observables and the selection on unobservables are introduced.

a matched sample of $(y_1, y_2, x_1, x_2) | D_P \cdot D_S = 1$. Similar to (2), simply using the balanced part will lead to an inconsistent estimator. Specify the sample attrition function by

$$\Pr(D_S = 1 | U_1 = u_1, U_2 = u_2, V = v) = g(k'_0(v) + k'_1(u_1, v) + k'_2(u_2, v)),$$

where $k'(\cdot)$ and $g(\cdot)$ are defined as before. In the presence of population attrition, the LHS of the second identification condition in (4), $f(u_2, v)$, is unobservable: we observe $f(u_2, v | D_P = 1)$ instead from the second period cross-section. But, we know that

$$f(u_2, v) = \frac{f(u_2, v | D_P = 1) \Pr(D_P = 1)}{\Pr(D_P = 1 | u_2, v)}.$$

Therefore, the identification condition becomes

$$\begin{aligned} f(u_1, v) &= \int \frac{\Pr(D_S = 1)}{g(k'_0(v) + k'_1(u_1, v) + k'_2(u_2, v))} f(u_1, u_2, v | D_S = 1) du_2, \\ \frac{f(u_2, v | D_P = 1) \Pr(D_P = 1)}{\Pr(D_P = 1 | u_2, v)} &= \int \frac{\Pr(D_S = 1)}{g(k'_0(v) + k'_1(u_1, v) + k'_2(u_2, v))} f(u_1, u_2, v | D_S = 1) du_1, \end{aligned} \quad (8)$$

for almost all (u_1, u_2, v) . So if we come up with a candidate for $\Pr(D_P = 1 | u_2, v)$, the LHS of the second equation in (8) is known, and we can obtain a consistent estimator of θ_0 by the sample attrition-correcting technique in the previous section.

Now a remaining question is how we specify $\Pr(D_P = 1 | u_2, v)$. When $\Pr(D_P = 1 | u_2, v)$ is a function of variables of known transition probability, it can be nonparametrically identified when repeated cross-sections are available. Assume that the transition probability is given by $P(Z_2 = z_2 | Z_1 = z_1)$, where z is a vector of variables of known transition probability.¹⁹ For instance, if z is year of entry, the transition probability is given by $P(z_2 | z_1) = 1(z_2 = z_1)$, where $1(\cdot)$ is the indicator function. If z is age, the transition probability is given by $P(z_2 | z_1) = 1(z_2 = z_1 + 1)$. The assumption requires that the population attrition is solely determined by variables of known transition probability.²⁰ Selection on variables of known transition probability implies that one minus the population attrition probability is given by

$$\begin{aligned} \Pr(D_P = 1 | u_2, v) &= \Pr(D_P = 1 | z_2) \\ &\equiv k(z_2), \end{aligned} \quad (9)$$

¹⁹The variables in z_2 must be included in (u_2, v) .

²⁰This assumption is strong but necessary because we do not know who emigrated from the United States.

where $k(\cdot)$ is some unknown function. The population attrition process, $k(z_2)$, is nonparametrically identified from

$$\begin{aligned} k(z_2) &= \frac{f(z_2|D_P=1)\Pr(D_P=1)}{f_2(z_2)} \\ &= \frac{f(z_2|D_P=1)\Pr(D_P=1)}{\int f_1(z_1)p(z_2|z_1)dz_1}. \end{aligned}$$

The key estimation strategy can be described in two steps. First, the second period cross-section with population attrition is identical to the counterfactual second period distribution where there is no population attrition adjusted by $k(z_2)$. Second, counterfactual second period distribution where there is no population attrition is available from the first period cross-section and the known transition probability. These steps can be written in the following way. For simplicity, assume that $k(z_2)$ is given by a parametric form, $k(z_2) = k(z_2'\psi)$. Consider the following:

$$\begin{aligned} \Pr(D_P=1)E_{Z_2}[Z_2|D_P=1] &= E_{Z_2}[k(Z_2'\psi)Z_2] \\ &= E_{Z_1}[\int k(z'\psi)zP(dz|Z_1)], \end{aligned}$$

which is implied by the two steps. Now, we can apply a GMM type estimation by employing the sample analog of these equations.²¹ The LHS is the average over the variables in the second period population (after population attrition has taken place) adjusted by the probability of population attrition. The RHS is the average over the variables in the first period population (prior to population attrition) transformed into the second period variables by the transition probability. Therefore, the sample analog is given by

$$\begin{aligned} \frac{1}{n_2}\Pr(D_P=1)\sum_{j=1}^{n_2}z_{2j} &= \frac{1}{n_1}\sum_{i=1}^{n_1}[\int k(z'\psi)zP(dz|z_{1i})] \\ &= \frac{1}{n_1}\sum_{i=1}^{n_1}\sum_{z\in S_2}k(z'\psi)z\Pr(z|z_{1i}), \end{aligned}$$

where n_1 and n_2 are the sample sizes of the first and the second period cross-sections. The second equation holds if z is a vector of discrete variables, where S_2 is the support of Z_2 . In the Appendix, we illustrate the estimation strategy in the analysis.

In the Appendix, we show that the identification conditions in (8) under assumption (9) can be transformed

²¹Technically, this method is similar to the method developed by Guell and Hu (2006). Both methods require cross-sections for two periods and use individual level information, but their method only allows time-invariant variables to enter the process. The two methods are developed for conceptually different purposes. Our method targets the attrition in the population or the duration of staying in the United States, whereas their method focuses on the duration of unemployment.

into conditional moment restrictions given by

$$\begin{aligned} 1 &= E \left[\frac{D_S}{g(k'_0(v) + k'_1(u_1, v) + k'_2(u_2, v))} \middle| u_1, v \right] \quad \text{w.p.1,} \\ \frac{1}{k(z_2)} &= E \left[\frac{D_S}{g(k'_0(v) + k'_1(u_1, v) + k'_2(u_2, v))} \middle| u_2, v, D_P = 1 \right] \quad \text{w.p.1.} \end{aligned} \quad (10)$$

In sum, once the attrition-correcting weighting function

$$C(u_1, u_2, v) = \frac{\Pr(D_S = 1)}{g(k'_0(v) + k'_1(u_1, v) + k'_2(u_2, v))} \quad (11)$$

is estimated, we weight the matched sample by (11) and estimate (7) to obtain a consistent estimator of θ_0 . Intuitively, the RHS in the second period is equivalent to weighting the individuals in the population (or more precisely the cross-section) with the inverse of one minus the probability of population attrition, $1/k(z_2)$, and the LHS of (10) is equivalent to weighting the individuals in the matched sample with the inverse of one minus the probability of sample attrition, $1/g(k'_0(v) + k'_1(u_1, v) + k'_2(u_2, v))$.

In practice, the vector z_t includes age, years since migration, education (assuming that no additional schooling is obtained), country of origin, and year of entry. These variables have deterministic time paths and satisfy the known transition probability assumption. The assumption, however, is more restrictive than the sample selection model, for instance, because observable variables with unknown transition probability, such as the wage, cannot enter in the selection function. The assumption can be problematic as the transition probabilities of labor market performance variables are usually not known. Intuitively labor market performance will affect population attrition decision. If the assumption is indeed a serious problem in practice, it is required to develop an alternative way of handling population attrition.

Despite its limitation, the attrition-correcting method has at least four advantages. First, the population attrition function is identified nonparametrically under selection on variables of known transition probability when repeated cross-sections are available. It allows stochastic transition, which is given by transition probability, so it is more flexible than assuming a deterministic mapping from one period to the other. Second, the method identifies the sample attrition and the population attrition processes separately. This is a very useful result because we are using a data set which does not provide information on who left the country. Notice that Heckman's self-selection correction cannot be applied, as it is not possible to distinguish those who migrate internally and who emigrate. Third, the method is robust against fixed effects. Finally, the weighting function estimates need not have a causal interpretation.

4 Applications: Estimation of Attrition Functions

The empirical specification of the attrition-correcting weighting function is as follows. We parameterize (3) by

$$\Pr(D_S = 1 | U_1 = u_1, U_2 = u_2, V = v) = g(v'\phi_0 + u_1'\phi_1 + u_2'\phi_2),$$

where v is a vector of a constant, age, education, and dummy variables (marital status, years in the United States, citizenship status, country of birth), u_1 and u_2 are vectors of logged hourly real dollar wages and indicators of “not usually working”, and $g(r) = e^r / (1 + e^r)$. Assume that the population attrition function (9) is parameterized by

$$\Pr(D_P = 1 | u_2, v) = k(z_2'\psi),$$

where $k(r) = e^r$, and z_2 is a vector of age, years since migration, education (assuming that no additional schooling is obtained), country of origin, and year of entry. In the Appendix, we show that ψ can be estimated by a logit model when $k(r) = e^r$.

Tables 2A’s and 2B’s report the ϕ coefficient estimates for the native and the foreign samples under the assumption that population attrition is negligible.²² Tables 2A-1 and 2B-1 use the entire sample. Tables 2A-2 and 2B-2 drop allocated wages and use individuals with reported wages only. Positive ϕ coefficient estimates imply that the variables are positively correlated with the matching rate or negatively correlated with residential mobility. The estimates for the 1994-1995 and 1995-1996 samples are less stable than those for other samples because of their smaller sample sizes. In general, natives tend to have higher matching rates than immigrants.

For the native samples over the matching period from 1996-1997 through 2003-2004, matching is positively correlated with age and marriage and is negatively correlated with education. Among those who usually work, both first period and second period wages are positively correlated with the matching rate, although the first period estimates are less stable. In addition, those who are not working are more likely to stay in the same address than those who are working except for a few first period estimates. When we drop individuals with allocated wages, education becomes less significant. Other estimates do not change much.

For the foreign samples during the same period, matching is positively correlated with age and years in the United States. Those who are married or are citizens have higher matching rates. The key difference from the native sample is education. Different from the native estimates, education is not a significant factor for

²²The coefficient estimates do not necessarily have causal interpretation. For instance, labor market outcome and residential mobility may affect each other.

matching immigrants and is rather positively correlated. Matching is positively correlated with the second period wage and the second period indicator of not working, which is similar to the native samples. The corresponding first period variables are neither very significant nor stable across years. Finally, immigrants from Europe tend to move less than the other immigrants. Similar to the native samples, dropping individuals with allocated wages makes education less significant.

Table 2C reports the ψ estimates, where a positive coefficient implies that the probability of staying in the United States is positively correlated with the variable. The population attrition functions are rather poorly estimated. The only coefficient estimates that are stable over the matching years is education. More educated foreign-born persons have higher probabilities of staying than less educated ones. The other variables including age, years since migration, country of origin, and the arrival year are not significant, and their coefficient estimates are not stable over the matching years. The estimation results do not support the hypothesis that the rates of return migration decline with time spent in the United States. However, this may not be very surprising because the annual population attrition rate is very small. Finally, Tables 2D-1 and 2D-2 report the (sample and population) attrition-correcting weighting function estimates. The results are not very different from those in Tables 2B-1 and 2B-2, respectively.

The coefficient estimates in Tables 2A's and 2D's are used to calculate attrition-correcting weights for all the individuals in the matched CPS. Using these coefficient estimates, we obtain an estimate of the $C(u_1, u_2, v, \phi, \psi)$ function, say $C(u_1, u_2, v, \hat{\phi}, \hat{\psi})$. If a model is given by conditional moment restrictions (1), one can obtain an estimator based on $E \left[m(y_1, y_2, x_1, x_2, \theta_0) \cdot C(u_1, u_2, v, \hat{\phi}, \hat{\psi}) \mid x_1, x_2, D_S = 1 \right] = 0$, w.p.1. If a model is given by regression, an estimator can be obtained by weighted least squares, where the weights are the attrition-correcting weights, $C(u_1, u_2, v, \hat{\phi}, \hat{\psi})$. An application of this method on economic assimilation of foreign-born workers can be found in Kim (2008).

5 Concluding Remarks

This paper develops a method that accounts for sample attrition in the presence of population attrition for use with two-period panel data models where the first period cross-section sample is representative of the target population while the panel and the second period cross-section samples are not. The method separately identifies sample attrition and population attrition when sample attrition is non-ignorable and population attrition is determined by variables of known transition probability. This is useful because samples usually do not indicate which missing observations are due to sample attrition and which are due to population

attrition. The attrition-correcting method is computationally straightforward because it is given by models of conditional moment restrictions. It generates a counterfactual, but representative cross-section by weighting the second period cross-section by one minus the probability of population attrition. Then, the method applies the existing sample attrition-correcting method, which uses the representative cross-sections as the basis for weighting the persons in a balanced part of the panel.

The method is applied to a longitudinal study of the foreign-born population in the United States. We obtain attrition-correcting weights for the native and immigrant samples in the matched CPS for 1994-2004. Of the two samples, the immigrant sample suffers from sample attrition due to changes in residence as well as population attrition caused by selective return migration. The native sample suffers from sample attrition only. Empirical results suggest that older or married individuals tend to live longer in the same residence for both the native and immigrant samples. More educated natives tend to move more, while the opposite is true for immigrants. Immigrants who have stayed longer in the United States tend to move less. We also find that both the first and second labor market outcomes affect sample attrition. From the population attrition function estimates we learn that more educated foreign-born persons have higher probabilities of staying than less educated ones. The other variables including age, years since migration, country of origin, and the arrival year are not significant.

The attrition-correcting technique can be generalized to longer panels and can be applied to applications other than immigration studies. If a panel that has more than two periods, the method requires that there exists at least one cross-section that is representative of the target population. The representative cross-section can be used as the basis for weighting the other non-representative cross-sections. Furthermore, it is possible to apply the method where the target population is not stationary over time, which is more general than population attrition. A longitudinal analysis of working population would be an example of it. Finally, the method is applicable to various topics in development economics, industrial organization, and labor economics. Examples of population attrition include seasonal migration in developing countries and entry and exit of firms in a market.

6 References

- Ai, Chunrong and Xiaohong Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions containing Unknown Functions," *Econometrica*, 71 (6), 1795-1843.
- Bhattacharya, Debopam (2008): "Inference in Panel Data Models under Attrition Caused by Unobservables," *Journal of Econometrics*, 144 (2), 430-446.
- Borjas, George J. (1999): "The Economic Analysis of Immigration," in Ashenfelter, Orley C. and David Card, eds., *Handbook of Labor Economics*, Vol 2A, Ch28.
- Borjas, George J. and Bernt Bratsberg (1996): "Who Leaves? The Outmigration of the Foreign-Born," *Review of Economics and Statistics*, 78, 165-176.
- Bratsberg, Bernt, Erling Barth, and Oddbjorn Raaum (2006): "Local Unemployment and the Relative Wages of Immigrants: Evidence from the Current Population Surveys," *Review of Economics and Statistics*, 88 (2), 243-263.
- Chen, Xiaohong, Han Hong, and Elie Tamer (2005): "Measurement Error Models with Auxiliary Data," *Review of Economic Studies*, 72, 343-366.
- Dieleman, Frans M. (2001): "Modelling Residential Mobility; A Review of Recent Trends in Research," *Journal of Housing and the Built Environment*, 16, 249-265.
- Duleep, Harriet O. and Mark C. Regets (1997): "Measuring Immigrant Wage Growth using Matched CPS Files," *Demography*, 34, 239-249.
- Guell, Maia and Luoqia Hu (2006): "Estimating the Probability of Leaving Unemployment using Uncompleted Spells from Repeated Cross-Section Data," *Journal of Econometrics*, 133 (1), 307-341.
- Hirano, Keisuke, Guido W. Imbens, Geert Ridder, and Donald B. Rubin (2001): "Combining Panel Data Sets with Attrition and Refreshment Samples," *Econometrica*, 69, 1645-1659.
- Jasso, Guillermina and Mark R. Rosenzweig (1982): "Estimating the Emigration Rates of Legal Immigrants using Administrative and Survey Data: The 1971 Cohort of Immigrants to the United States," *Demography*, 19 (3), 279-290.
- Jasso, Guillermina and Mark R. Rosenzweig (1988): "How Well do U.S. Immigrants do? Vintage Effects, Emigration Selectivity, and Occupational Mobility," in T. Paul Schultz, ed., *Research in Population Economics*, Vol 6. 229-253.

- Jasso, Guillermina and Mark R. Rosenzweig (1990): *The New Chosen People: Immigrants in the United States*, New York: Russell Sage Foundation.
- Kim, Seik (2008): “Economic Assimilation of Foreign-Born Workers in the United States: An Overlapping Rotating Panel Analysis,” University of Washington Working Paper, UWEC-2008-19.
- Kitamura, Yuichi, Gautam Tripathi, and Hyungtaik Ahn (2004): “Empirical Likelihood-Based Inference in Conditional Moment Restriction Models,” *Econometrica*, 72, 1667-1714.
- LaLonde, Robert J. and Robert H. Topel (1997): “Economic Impact of International Migration and The Economic Performance of Migrants,” in Mark R. Rosenzweig and Oded Stark, eds., *Handbook of Population and Family Economics*, Vol 3B, Ch 14.
- Lemieux, Thomas (2006): “Increasing Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill?” *American Economic Review*, 96 (3), 461-498.
- Little, Roderick J. A., and Donald B. Rubin (2002): *Statistical Analysis with Missing Data*, Wiley Series in Probability and Statistics.
- Long, Larry (1988): *Migration and Residential Mobility in the United States*, Russell Sage Foundation.
- Peracchi, Franco and Finis Welch (1995): “How Representative are Matched Cross-Sections? Evidence from the Current Population Survey,” *Journal of Econometrics*, 68 (1), 153-179.
- Ridder, Geert, and Robert Moffitt (2003): “The Econometrics of Data Combination,” Working Paper.
- Schmitt, John (2003): “Creating a consistent hourly wage series from the Current Population Survey’s Outgoing Rotation Group, 1979-2002,” Working Paper.
- U.S. Census Bureau (2002): *Technical Paper 63RV: Current Population Survey - Design and Methodology*.
- Warren, Robert and Jennifer M. Peck (1980): “Foreign-Born Emigration from the United States: 1960 to 1970,” *Demography*, 17, 71-84.
- Wooldridge, Jeffrey M. (2002): “Inverse Probability Weighted M-Estimators for Sample Selection, Attrition, and Stratification,” *Portuguese Economic Journal*, 1, 117-139.

7 Appendix

7.1 Equivalence of Identification Conditions

Equivalence of the identification conditions (4) and the conditional moment restrictions (5) when the population is stationary is proved by Bhattacharya (2008). We show equivalence of (8) and (10) with population attrition. The first identification condition in (8) is

$$f(u_1, v) = \int \frac{\Pr(D_S = 1)}{\Pr(D_S = 1|u_1, u_2, v)} f(u_1, u_2, v|D_S = 1) du_2.$$

Dividing both sides with $f(u_1, v)$, we have

$$\begin{aligned} 1 &= \int \frac{\Pr(D_S = 1) f(u_1, u_2, v|D_S = 1)}{g(k'_0(v) + k'_1(u_1, v) + k'_2(u_2, v)) f(u_1, v)} du_2 \\ &= \int \frac{P(u_2, D_S = 1|u_1, v)}{g(k'_0(v) + k'_1(u_1, v) + k'_2(u_2, v))} du_2 \\ &= \sum_{s=0,1} \int \frac{s \cdot P(u_2, D_S = s|u_1, v)}{g(k'_0(v) + k'_1(u_1, v) + k'_2(u_2, v))} du_2 \\ &= E \left[\frac{D_S}{g(k'_0(v) + k'_1(u_1, v) + k'_2(u_2, v))} | u_1, v \right] \quad \text{for all } u_1, v, \end{aligned}$$

which is the first condition in (10). The second identification condition in (8) is

$$\begin{aligned} f(u_2, v) &= \frac{f(u_2, v|D_P = 1) \Pr(D_P = 1)}{\Pr(D_P = 1|u_2, v)} \\ &= \int \frac{\Pr(D_S = 1)}{\Pr(D_S = 1|u_1, u_2, v)} f(u_1, u_2, v|D_S = 1) du_1. \end{aligned}$$

Notice that, we do not observe $f(u_2, v)$, but do $f(u_2, v|D_P = 1)$. Thus, dividing both sides with $f(u_2, v)$, we have

$$\begin{aligned}
1 &= \int \frac{\Pr(D_S = 1) f(u_1, u_2, v|D_S = 1)}{g(k'_0(v) + k'_1(u_1, v) + k'_2(u_2, v)) f(u_2, v)} du_1 \\
&= \int \frac{\Pr(D_S = 1) f(u_1, u_2, v|D_S = 1)}{g(k'_0(v) + k'_1(u_1, v) + k'_2(u_2, v)) f(u_2, v|D_P = 1)} \cdot \frac{\Pr(D_P = 1|u_2, v)}{\Pr(D_P = 1)} du_1 \\
&= \int \frac{\Pr(D_S = 1|D_P = 1) f(u_1, u_2, v|D_S \cdot D_P = 1)}{g(k'_0(v) + k'_1(u_1, v) + k'_2(u_2, v)) f(u_2, v|D_P = 1)} \cdot \Pr(D_P = 1|u_2, v) du_1 \\
&= \int \frac{P(u_1, u_2, v, D_S = 1|D_P = 1)}{g(k'_0(v) + k'_1(u_1, v) + k'_2(u_2, v)) f(u_2, v|D_P = 1)} \cdot \Pr(D_P = 1|u_2, v) du_1 \\
&= \int \frac{P(u_1, D_S = 1|u_2, v, D_P = 1)}{g(k'_0(v) + k'_1(u_1, v) + k'_2(u_2, v))} \cdot \Pr(D_P = 1|u_2, v) du_1 \\
&= \sum_{s=0,1} \int \frac{s \cdot P(u_1, D_S = s|u_2, v, D_P = 1)}{g(k'_0(v) + k'_1(u_1, v) + k'_2(u_2, v))} \cdot \Pr(D_P = 1|u_2, v) du_1 \\
&= E \left[\frac{D_S \cdot \Pr(D_P = 1|u_2, v)}{g(k'_0(v) + k'_1(u_1, v) + k'_2(u_2, v))} \Big| u_2, v, D_P = 1 \right] \quad \text{for all } u_2, v,
\end{aligned}$$

where the second equation uses

$$f(u_2, v) = \frac{f(u_2, v|D_P = 1) \Pr(D_P = 1)}{\Pr(D_P = 1|u_2, v)}$$

and the third equation uses

$$\begin{aligned}
\Pr(D_S = 1) &= \Pr(D_S = 1|D_P = 1) \cdot \frac{\Pr(D_P = 1)}{\Pr(D_P = 1|D_S = 1)} \\
&= \Pr(D_S = 1|D_P = 1) \cdot \Pr(D_P = 1),
\end{aligned}$$

and

$$f(u_1, u_2, v|D_S = 1) = f(u_1, u_2, v|D_S \cdot D_P = 1),$$

as $D_S = 1$ implies $D_S \cdot D_P = 1$. Finally, by the assumption (9), we have

$$\frac{1}{\bar{k}(z_2)} = E \left[\frac{D_S}{g(k'_0(v) + k'_1(u_1, v) + k'_2(u_2, v))} \Big| u_2, v, D_P = 1 \right] \quad \text{for all } u_2, v.$$

This is the second condition in (10).

7.2 Estimation of Weighting Functions

This section discusses technical details in the weighting function estimation. The identification conditions in (10) can be transformed to

$$\begin{aligned} E \left[\frac{D_S \cdot a(u_1, v)}{g(v' \phi_0 + u_1' \phi_1 + u_2' \phi_2)} \right] &= E[a(u_1, v)], \\ E \left[\frac{D_S \cdot a(u_2, v)}{g(v' \phi_0 + u_1' \phi_1 + u_2' \phi_2)} \right] &= E \left[\frac{a(u_2, v)}{k(z_2)} \right], \end{aligned} \quad (12)$$

for an arbitrary function $a(\cdot)$. Let n be the sample size of the incoming sample and n_m be the sample size of the matched sample. In addition, let n_1 and n_2 be the sample sizes of the representative cross-section samples in the incoming and the outgoing years. For simplicity, write

$$g(u_1, u_2, v, \phi) \equiv g(v' \phi_0 + u_1' \phi_1 + u_2' \phi_2).$$

Then, the LHS of (12) can be estimated by

$$\begin{aligned} \frac{1}{n} \sum_{m=1}^n \frac{D_{Sm} \cdot a(u_{tm}, v_m)}{g(u_{1m}, u_{2m}, v_m, \theta)} &= \frac{1}{n} \sum_{l=1}^{n_m} \frac{1 \cdot a(u_{tm}, v_m)}{g(u_{1m}, u_{2m}, v_m, \theta)} + \frac{1}{n} \sum_{m=n_m+1}^n \frac{0 \cdot a(u_{tm}, v_m)}{g(u_{1m}, u_{2m}, v_m, \theta)} \\ &= \frac{1}{n} \sum_{l=1}^{n_m} \frac{a(u_{tm}, v_m)}{g(u_{1m}, u_{2m}, v_m, \theta)}, \end{aligned}$$

for $t = 1, 2$, and the RHS of (12) can be estimated by

$$\frac{1}{n_1} \sum_{i=1}^{n_1} a(u_{1i}, v_i) = 0,$$

for $t = 1$ and

$$\frac{1}{n_2} \sum_{i=1}^{n_2} \frac{a(u_{2i}, v_i)}{k(z_2)} = 0,$$

for $t = 2$. In estimation, the LHS uses the matched longitudinal sample and the RHS uses the representative cross-sections. The GMM can be used, where function $a(\cdot)$ is a vector of age , age^2 , age^3 , $educ$, $educ^2$, $educ^3$, a marital status dummy, $\log wage$, $\log wage^2$, $\log wage^3$, and a dummy of not working for period $t = 1, 2$. For the foreign sample, we add ysm , ysm^2 , ysm^3 , a citizenship dummy, and continent of origin (Europe, Asia, and Africa-Oceania) dummies.

The population attrition process, $k(z_2)$, can be estimated as follows. Notice that all the variables in z_1

have deterministic time paths and map to z_2 one-to-one. Without loss of generality we replace (9) with

$$\Pr(D_P = 1|Z_1 = z_1) = k(z_1).$$

When we have z_1 , we do not need to worry about the transition probability $P(Z_2 = z_2|Z_1 = z_1)$. Assume that $k(z_1)$ is given by a parametric form: $k(z_1'\psi)$. Consider the following transformation:

$$p(z_1'\psi) \equiv \frac{k(z_1'\psi)}{1 + k(z_1'\psi)}.$$

The estimation strategy is estimate $p(z_1'\psi)$ and transform it to $k(z_1'\psi)$. We use discrete choice model instead of applying GMM. Generate an indicator variable that is set to unity for observations in the second period cross-section. To fix idea, suppose there is no population attrition and assume that the sample sizes are the same. Then there will equal number of 0's and 1's. So $p(z_1'\psi) = 1/2$ for all z_1 . If population attrition occurs to individuals with $z_1 = \tilde{z}_1$, we expect $p(\tilde{z}_1'\psi) < 1/2$. We use a logit model

$$p(z_1'\psi) = \frac{e^{z_1'\psi}}{1 + e^{z_1'\psi}}$$

and obtain $p(z_1'\hat{\psi})$.

7.3 Comparison with Other Sample Attrition-Correction Approaches

There are two general approaches are often used to deal with attrition in panel data sets: selection on observables and selection on unobservables. Models of selection on observables make the assumption that the probability of attrition depends only on U_1 and V , which are observed. In this case, U_2 is missing at random in the panel, i.e.,

$$D_S \perp U_2 | (U_1, V),$$

and the attrition probability can be written as

$$\Pr(D_S = 1|U_1, U_2, V) = \Pr(D_S = 1|U_1, V).$$

With this structure $\Pr(D_S = 1|u_1, u_2, v) = \Pr(D_S = 1|u_1, v)$ can be observed and therefore $f(u_1, u_2, v)$ is identified. An example where the assumption of selection on observables fails is when an individual does not respond in the second period if the individual experiences an unexpected negative wage shock in the second

period. Under selection on observable assumption, Jeffrey Wooldridge (2002) proposes an inverse probability weighted (IPW) M-estimator for two-period panel data models. It is known that if selection is ignorable, an inverse probability weighting scheme generally identifies the population parameters.

Models of selection on unobservables make the assumption that the probability of attrition depends on U_2 , which may not be observed. In this case, the attrition process does not depend on the first period variables U_1 , i.e.,

$$D_S \perp U_1 | (U_2, V),$$

and the attrition probability can be written as

$$\Pr(D_S = 1 | U_1, U_2, V) = \Pr(D_S = 1 | U_2, V).$$

This assumption fails if individuals do not respond in the second period if there was a negative wage shock in the first period. For estimation, the probability of attrition can be specified to depend on arbitrary functions of U_2 . A special case of the method of selection on unobservables is the standard sample selection model by Heckman (1976, 1979). Heckman's solution requires at least one exogenous variable affecting selection that does not appear in the structural equation.

Table 1A. Stay Probability (One minus the Outmigration Rate) by Arrival Year

| | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 |
|---|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| | -1995 | -1996 | -1997 | -1998 | -1999 | -2000 | -2001 | -2002 | -2003 | -2004 |
| all foreign persons | | | | | | | | | | |
| # in 2nd year | 5331 | 4605 | 5011 | 5070 | 5398 | 5578 | 6299 | 6293 | 6831 | 6090 |
| # in 1st year | 5329 | 5417 | 5121 | 5220 | 5527 | 5435 | 6060 | 6021 | 7001 | 6811 |
| stay probability | 1.000 | 0.850 | 0.979 | 0.971 | 0.977 | 1.026 | 1.039 | 1.045 | 0.976 | 0.894 |
| before 1980 arrivals | | | | | | | | | | |
| # in 1st year | 2524 | 2417 | 2158 | 2078 | 2090 | 1936 | 1893 | 1793 | 1904 | 1745 |
| stay probability | 0.998 | 0.826 | 0.966 | 0.949 | 0.944 | 0.999 | 1.031 | 1.022 | 0.957 | 0.896 |
| 1980-1981 arrivals | | | | | | | | | | |
| # in 1st year | 517 | 615 | 511 | 520 | 467 | 474 | 458 | 534 | 457 | 483 |
| stay probability | 0.965 | 0.862 | 0.971 | 0.952 | 1.000 | 1.108 | 1.083 | 1.060 | 1.039 | 0.917 |
| 1982-1983 arrivals | | | | | | | | | | |
| # in 1st year | 323 | 343 | 282 | 317 | 329 | 294 | 321 | 313 | 349 | 338 |
| stay probability | 0.947 | 0.930 | 1.035 | 0.987 | 0.936 | 0.959 | 1.078 | 1.099 | 1.003 | 0.879 |
| 1984-1985 arrivals | | | | | | | | | | |
| # in 1st year | 456 | 521 | 411 | 451 | 401 | 389 | 429 | 395 | 444 | 447 |
| stay probability | 1.042 | 0.904 | 1.010 | 0.940 | 0.983 | 1.103 | 1.061 | 1.104 | 0.977 | 0.940 |
| 1986-1987 arrivals | | | | | | | | | | |
| # in 1st year | 400 | 433 | 421 | 405 | 353 | 357 | 375 | 353 | 426 | 409 |
| stay probability | 1.055 | 0.885 | 0.964 | 1.007 | 1.057 | 1.050 | 1.053 | 1.125 | 1.035 | 0.861 |
| 1988-1989 arrivals | | | | | | | | | | |
| # in 1st year | 567 | 545 | 473 | 529 | 528 | 596 | 502 | 497 | 498 | 527 |
| stay probability | 0.984 | 0.809 | 0.981 | 1.000 | 0.992 | 0.938 | 0.982 | 1.012 | 1.044 | 0.890 |
| 1990-1991 arrivals | | | | | | | | | | |
| # in 1st year | 542 | 543 | 491 | 437 | 478 | 476 | 536 | 587 | 588 | 542 |
| stay probability | 1.018 | 0.855 | 0.994 | 1.078 | 0.912 | 1.053 | 1.076 | 1.019 | 0.927 | 0.910 |
| 1992-1993 arrivals | | | | | | | | | | |
| # in 1st year | | | 374 | 483 | 424 | 442 | 458 | 450 | 481 | 477 |
| stay probability | | | 0.976 | 0.948 | 1.087 | 1.068 | 1.020 | 1.096 | 0.977 | 0.876 |
| 1994-1995 arrivals | | | | | | | | | | |
| # in 1st year | | | | | 457 | 471 | 542 | 572 | 520 | 488 |
| stay probability | | | | | 1.011 | 1.064 | 1.068 | 1.038 | 0.981 | 0.986 |
| 1996-1997 arrivals | | | | | | | | | | |
| # in 1st year | | | | | | | 546 | 527 | 566 | 575 |
| stay probability | | | | | | | 0.987 | 1.004 | 0.952 | 0.878 |
| 1998-1999 arrivals | | | | | | | | | | |
| # in 1st year | | | | | | | | | 768 | 780 |
| stay probability | | | | | | | | | 0.944 | 0.832 |
| # in 1st (2nd) year: the number of foreign-born persons in the 1st (2nd) year | | | | | | | | | | |
| stay probability: the ratio between the numbers of foreign-born persons in the 2nd and in the 1st years | | | | | | | | | | |

Table 1B. Stay Probability (One minus the Outmigration Rate) by Ethnic Origin

| arrival years | | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 |
|--------------------|-------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| | | -1995 | -1996 | -1997 | -1998 | -1999 | -2000 | -2001 | -2002 | -2003 | -2004 |
| <u>before 1980</u> | C.S.America | 1016 | 1014 | 1019 | 964 | 979 | 891 | 880 | 798 | 845 | 812 |
| | | 1.024 | 0.858 | 0.952 | 0.979 | 0.941 | 1.027 | 1.007 | 1.014 | 0.998 | 0.929 |
| | Europe | 611 | 674 | 537 | 540 | 520 | 461 | 486 | 445 | 498 | 424 |
| | | 1.005 | 0.785 | 0.944 | 0.930 | 0.940 | 0.978 | 0.998 | 1.054 | 0.940 | 0.854 |
| | Asia | 518 | 575 | 495 | 502 | 510 | 488 | 430 | 455 | 464 | 430 |
| | 1.041 | 0.854 | 1.032 | 0.896 | 0.961 | 0.945 | 1.100 | 0.991 | 0.905 | 0.884 | |
| <u>1980-1991</u> | C.S.America | 1399 | 1439 | 1396 | 1448 | 1456 | 1498 | 1490 | 1484 | 1597 | 1541 |
| | | 1.065 | 0.941 | 0.983 | 0.966 | 0.955 | 1.012 | 1.060 | 1.065 | 1.009 | 0.921 |
| | Europe | 279 | 385 | 264 | 295 | 268 | 247 | 241 | 307 | 289 | 330 |
| | | 1.018 | 0.875 | 1.038 | 0.980 | 1.045 | 1.117 | 1.083 | 1.010 | 0.972 | 0.855 |
| | Asia | 680 | 965 | 805 | 829 | 720 | 969 | 706 | 712 | 703 | 717 |
| | 1.059 | 0.802 | 0.970 | 1.037 | 1.001 | 0.752 | 1.067 | 1.098 | 0.980 | 0.881 | |
| <u>1992-1993</u> | C.S.America | | | 173 | 237 | 200 | 209 | 253 | 218 | 265 | 266 |
| | | | | 1.000 | 0.920 | 1.130 | 1.139 | 0.953 | 1.110 | 0.940 | 0.876 |
| | Europe | | | 63 | 73 | 89 | 83 | 58 | 63 | 64 | 56 |
| | | | | 0.937 | 0.973 | 1.101 | 0.916 | 1.000 | 1.032 | 1.016 | 0.911 |
| | Asia | | | 117 | 152 | 120 | 114 | 119 | 125 | 122 | 124 |
| | | | 0.957 | 0.967 | 0.925 | 1.088 | 1.092 | 1.168 | 1.008 | 0.863 | |
| <u>1994-1995</u> | C.S.America | | | | | 218 | 253 | 302 | 320 | 309 | 277 |
| | | | | | | 1.055 | 1.059 | 1.017 | 1.072 | 0.961 | 0.982 |
| | Europe | | | | | 78 | 69 | 80 | 59 | 69 | 69 |
| | | | | | | 0.962 | 1.072 | 1.050 | 1.017 | 1.043 | 1.072 |
| | Asia | | | | | 141 | 111 | 126 | 143 | 102 | 114 |
| | | | | | 0.894 | 1.117 | 1.111 | 1.014 | 1.010 | 0.939 | |
| <u>1996-1997</u> | C.S.America | | | | | | | 251 | 287 | 301 | 308 |
| | | | | | | | | 1.068 | 0.913 | 0.960 | 0.844 |
| | Europe | | | | | | | 87 | 58 | 85 | 75 |
| | | | | | | | | 0.908 | 1.207 | 0.918 | 0.973 |
| | Asia | | | | | | | 152 | 127 | 138 | 129 |
| | | | | | | | 0.875 | 1.150 | 0.855 | 0.938 | |
| <u>1998-1999</u> | C.S.America | | | | | | | | | 411 | 462 |
| | | | | | | | | | | 0.973 | 0.827 |
| | Europe | | | | | | | | | 118 | 99 |
| | | | | | | | | | | 0.941 | 0.828 |
| | Asia | | | | | | | | | 158 | 154 |
| | | | | | | | | | 0.949 | 0.812 | |

Table 2A-1. (Sample) Attrition-Correcting Weighting Function Estimates (Natives): All Wages

| | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 |
|------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| | -1995 | -1996 | -1997 | -1998 | -1999 | -2000 | -2001 | -2002 | -2003 | -2004 |
| Age | 0.027 | 0.045 | 0.054 | 0.052 | 0.057 | 0.053 | 0.054 | 0.056 | 0.049 | 0.039 |
| | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) |
| Education | 0.024 | 0.002 | -0.019 | -0.031 | -0.033 | -0.013 | -0.027 | -0.031 | -0.031 | -0.015 |
| | (0.005) | (0.006) | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) |
| Mari.Stat. | 0.404 | 0.536 | 0.576 | 0.611 | 0.467 | 0.615 | 0.666 | 0.548 | 0.577 | 0.503 |
| | (0.027) | (0.030) | (0.016) | (0.016) | (0.016) | (0.016) | (0.016) | (0.016) | (0.015) | (0.016) |
| LogWage1 | 0.372 | -0.283 | 0.109 | -0.015 | 0.196 | 0.148 | 0.027 | -0.046 | 0.174 | 0.057 |
| | (0.029) | (0.034) | (0.019) | (0.018) | (0.020) | (0.019) | (0.019) | (0.018) | (0.017) | (0.020) |
| LogWage2 | 0.084 | 0.499 | 0.277 | 0.252 | 0.094 | 0.167 | 0.068 | 0.306 | 0.221 | 0.226 |
| | (0.030) | (0.034) | (0.018) | (0.020) | (0.019) | (0.019) | (0.019) | (0.018) | (0.018) | (0.020) |
| NoWork1 | 0.960 | -0.621 | 0.310 | -0.026 | 0.392 | 0.459 | 0.057 | -0.134 | 0.523 | 0.253 |
| | (0.082) | (0.094) | (0.052) | (0.051) | (0.055) | (0.054) | (0.055) | (0.052) | (0.049) | (0.056) |
| NoWork2 | 0.160 | 1.059 | 0.465 | 0.573 | 0.159 | 0.363 | 0.055 | 0.562 | 0.314 | 0.391 |
| | (0.084) | (0.095) | (0.050) | (0.055) | (0.055) | (0.054) | (0.054) | (0.052) | (0.052) | (0.055) |
| Constant | -2.021 | -1.706 | -1.742 | -1.299 | -1.473 | -1.817 | -1.014 | -1.398 | -1.582 | -1.463 |
| | (0.085) | (0.096) | (0.052) | (0.054) | (0.055) | (0.055) | (0.056) | (0.055) | (0.053) | (0.057) |
| N | 17929 | 13691 | 36928 | 37178 | 37176 | 37194 | 35586 | 38265 | 42469 | 42259 |
| Mat.Rate | 68.0% | 70.3% | 78.1% | 77.1% | 77.5% | 77.9% | 78.8% | 78.3% | 77.2% | 71.2% |

Standard errors are reported in parentheses. N: sample size, Mat.Rate: matching rate

The LHS variable is the odds of staying in the same address.

Mari.Stat.: 1 if married; LogWage: log of hourly rate of pay (yrs 1&2); NoWork: no reported wage (yrs 1&2)

Table 2A-2. (Sample) Attrition-Correcting Weighting Function Estimates (Natives): Reported Wages Only

| | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 |
|------------|-------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| | -1995 | -1996 | -1997 | -1998 | -1999 | -2000 | -2001 | -2002 | -2003 | -2004 |
| Age | N/A | 0.036 (0.001) | 0.041 (0.001) | 0.040 (0.001) | 0.041 (0.001) | 0.040 (0.001) | 0.039 (0.001) | 0.040 (0.001) | 0.038 (0.001) | 0.031 (0.001) |
| Education | | 0.006 (0.006) | -0.002 (0.004) | -0.011 (0.004) | -0.006 (0.004) | 0.005 (0.004) | -0.004 (0.004) | -0.003 (0.004) | -0.013 (0.004) | -0.004 (0.004) |
| Mari.Stat. | | 0.574 (0.037) | 0.538 (0.020) | 0.511 (0.20) | 0.405 (0.021) | 0.493 (0.021) | 0.488 (0.021) | 0.455 (0.021) | 0.437 (0.020) | 0.422 (0.021) |
| LogWage1 | | -0.409 (0.052) | 0.107 (0.029) | -0.159 (0.028) | 0.049 (0.032) | -0.169 (0.034) | 0.069 (0.036) | -0.142 (0.032) | 0.134 (0.030) | 0.095 (0.036) |
| LogWage2 | | 0.538 (0.051) | 0.146 (0.028) | 0.272 (0.031) | 0.095 (0.032) | 0.306 (0.035) | -0.062 (0.034) | 0.308 (0.033) | 0.109 (0.033) | 0.101 (0.036) |
| NoWork1 | | -0.741 (0.138) | 0.386 (0.077) | -0.164 (0.075) | 0.318 (0.088) | -0.016 (0.090) | 0.441 (0.096) | -0.218 (0.088) | 0.659 (0.082) | 0.421 (0.099) |
| NoWork2 | | 1.244 (0.137) | 0.351 (0.075) | 0.681 (0.082) | 0.205 (0.089) | 0.735 (0.093) | -0.143 (0.094) | 0.846 (0.090) | 0.142 (0.090) | 0.307 (0.098) |
| Constant | | -1.680 (0.120) | -1.684 (0.065) | -1.323 (0.068) | -1.495 (0.072) | -1.727 (0.074) | -1.215 (0.075) | -1.627 (0.075) | -1.636 (0.072) | -1.610 (0.078) |
| N | | 11258 | 30454 | 30743 | 30081 | 29236 | 27240 | 29307 | 32934 | 32323 |
| Mat.Rate | | 71.0% | 78.4% | 77.6% | 77.8% | 78.3% | 79.2% | 78.4% | 77.6% | 72.0% |

Standard errors are reported in parentheses. N: sample size, Mat.Rate: matching rate

The LHS variable is the odds of staying in the same address.

Mari.Stat.: 1 if married; LogWage: log of hourly rate of pay (yrs 1&2); NoWork: no reported wage (yrs 1&2)

Table 2B-1. Sample Attrition-Correcting Weighting Function Estimates (Immigrants): All Wages

| | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 |
|------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| | -1995 | -1996 | -1997 | -1998 | -1999 | -2000 | -2001 | -2002 | -2003 | -2004 |
| Age | 0.032 | 0.036 | 0.021 | 0.037 | 0.027 | 0.026 | 0.031 | 0.026 | 0.026 | 0.030 |
| | (0.004) | (0.005) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) |
| Education | 0.021 | 0.019 | 0.008 | 0.015 | 0.003 | 0.043 | -0.037 | 0.031 | 0.001 | -0.003 |
| | (0.010) | (0.013) | (0.006) | (0.006) | (0.006) | (0.006) | (0.006) | (0.006) | (0.005) | (0.006) |
| Mari.Stat. | 0.227 | 0.360 | 0.497 | 0.344 | 0.620 | 0.457 | 0.119 | 0.605 | 0.313 | 0.257 |
| | (0.089) | (0.110) | (0.052) | (0.054) | (0.050) | (0.050) | (0.045) | (0.047) | (0.042) | (0.048) |
| LogWage1 | -0.018 | 0.267 | -0.341 | -0.027 | -0.022 | 0.238 | 0.120 | -0.081 | -0.108 | 0.222 |
| | (0.097) | (0.124) | (0.055) | (0.054) | (0.053) | (0.056) | (0.048) | (0.049) | (0.047) | (0.053) |
| LogWage2 | 0.054 | -0.008 | 0.452 | 0.208 | 0.033 | -0.062 | 0.180 | 0.280 | 0.062 | 0.103 |
| | (0.089) | (0.108) | (0.057) | (0.061) | (0.055) | (0.056) | (0.051) | (0.050) | (0.048) | (0.052) |
| NoWork1 | 0.190 | 0.443 | -0.774 | -0.329 | 0.049 | 0.573 | 0.138 | -0.064 | -0.336 | 0.356 |
| | (0.238) | (0.303) | (0.141) | (0.142) | (0.136) | (0.145) | (0.127) | (0.129) | (0.123) | (0.141) |
| NoWork2 | -0.221 | 0.224 | 1.177 | 0.736 | -0.266 | 0.052 | 0.433 | 0.643 | 0.223 | 0.238 |
| | (0.229) | (0.276) | (0.147) | (0.157) | (0.144) | (0.144) | (0.133) | (0.133) | (0.126) | (0.139) |
| Ysm | 0.052 | 0.054 | 0.048 | 0.050 | 0.028 | 0.092 | 0.023 | 0.076 | 0.033 | 0.037 |
| | (0.004) | (0.005) | (0.003) | (0.003) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) |
| Citizen | -0.421 | -0.086 | 0.045 | 0.250 | 0.152 | -0.341 | 0.154 | 0.134 | 0.166 | 0.236 |
| | (0.090) | (0.109) | (0.048) | (0.050) | (0.048) | (0.048) | (0.044) | (0.044) | (0.042) | (0.046) |
| Europe | 0.391 | 0.849 | 0.097 | 0.313 | 0.346 | -0.025 | 0.260 | 0.010 | 0.285 | 0.024 |
| | (0.102) | (0.120) | (0.051) | (0.063) | (0.059) | (0.062) | (0.059) | (0.060) | (0.054) | (0.061) |
| Asia | 0.362 | -0.178 | 0.059 | 0.195 | 0.105 | 0.002 | 0.202 | -0.016 | -0.154 | -0.024 |
| | (0.100) | (0.013) | (0.063) | (0.057) | (0.054) | (0.056) | (0.052) | (0.053) | (0.050) | (0.055) |
| Africa | 2.107 | -0.857 | 0.077 | -0.858 | -0.242 | 0.071 | 0.062 | -0.121 | -0.148 | -0.244 |
| | (0.107) | (0.218) | (0.058) | (0.133) | (0.109) | (0.092) | (0.082) | (0.082) | (0.078) | (0.089) |
| Constant | -1.896 | -2.581 | -1.420 | -2.171 | -1.100 | -2.417 | -1.093 | -2.265 | -0.802 | -1.950 |
| | (0.213) | (0.276) | (0.129) | (0.138) | (0.130) | (0.135) | (0.124) | (0.127) | (0.120) | (0.130) |
| N | 2159 | 1714 | 4965 | 5021 | 5339 | 5284 | 5885 | 5825 | 6771 | 6617 |
| Mat.Rate | 66.3% | 60.3% | 70.1% | 68.7% | 70.1% | 70.8% | 71.4% | 71.6% | 70.1% | 65.0% |

Standard errors are reported in parentheses. N: sample size, Mat.Rate: matching rate

The LHS variable is the odds of staying in the same address.

Mari.Stat.: 1 if married; LogWage: log of hourly rate of pay (yrs 1&2); NoWork: no reported wage (yrs 1&2)

Ysm: years since migration; Citizen: 1 if U.S. citizen; Constant: immigrants from Central & South America

Table 2B-2. Sample Attrition-Correcting Weighting Function Estimates (Immigrants): Reported Wages Only

| | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 |
|------------|-------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| | -1995 | -1996 | -1997 | -1998 | -1999 | -2000 | -2001 | -2002 | -2003 | -2004 |
| Age | N/A | 0.021 | 0.012 | 0.029 | 0.019 | 0.013 | 0.021 | 0.017 | 0.019 | 0.020 |
| | | (0.006) | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) |
| Education | | 0.022 | 0.002 | -0.009 | -0.007 | 0.013 | -0.027 | 0.017 | -0.007 | -0.004 |
| | | (0.016) | (0.008) | (0.007) | (0.007) | (0.008) | (0.008) | (0.008) | (0.007) | (0.008) |
| Mari.Stat. | | 0.581 | 0.547 | 0.461 | 0.519 | 0.460 | 0.210 | 0.480 | 0.244 | 0.244 |
| | | (0.140) | (0.069) | (0.069) | (0.069) | (0.067) | (0.065) | (0.066) | (0.067) | (0.067) |
| LogWage1 | | 0.301 | -0.240 | -0.002 | -0.121 | 0.364 | 0.054 | 0.230 | -0.119 | 0.191 |
| | | (0.216) | (0.095) | (0.103) | (0.100) | (0.110) | (0.094) | (0.089) | (0.089) | (0.106) |
| LogWage2 | | -0.182 | 0.433 | 0.217 | 0.152 | -0.090 | 0.179 | -0.092 | 0.134 | 0.055 |
| | | (0.196) | (0.098) | (0.117) | (0.103) | (0.111) | (0.099) | (0.089) | (0.094) | (0.103) |
| NoWork1 | | 0.542 | -0.335 | -0.085 | 0.112 | 0.888 | 0.174 | 0.439 | -0.088 | 0.386 |
| | | (0.485) | (0.225) | (0.239) | (0.238) | (0.258) | (0.231) | (0.223) | (0.215) | (0.262) |
| NoWork2 | | -0.052 | 1.196 | 0.773 | 0.079 | 0.083 | 0.510 | 0.059 | 0.349 | 0.172 |
| | | (0.455) | (0.233) | (0.274) | (0.250) | (0.262) | (0.243) | (0.225) | (0.228) | (0.257) |
| Ysm | | 0.025 | 0.038 | 0.198 | 0.022 | 0.048 | 0.012 | 0.046 | 0.017 | 0.025 |
| | | (0.007) | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) |
| Citizen | | -0.187 | -0.066 | 0.149 | 0.051 | -0.182 | 0.009 | 0.076 | 0.066 | 0.211 |
| | | (0.141) | (0.068) | (0.065) | (0.067) | (0.065) | (0.064) | (0.062) | (0.061) | (0.064) |
| Europe | | 0.664 | 0.055 | 0.197 | 0.151 | -0.021 | 0.055 | 0.076 | 0.203 | 0.029 |
| | | (0.157) | (0.084) | (0.082) | (0.083) | (0.084) | (0.086) | (0.085) | (0.077) | (0.086) |
| Asia | | 0.041 | 0.099 | 0.060 | 0.073 | -0.018 | 0.016 | -0.002 | -0.067 | 0.029 |
| | | (0.159) | (0.076) | (0.076) | (0.075) | (0.077) | (0.076) | (0.076) | (0.071) | (0.086) |
| Africa | | -0.860 | 0.245 | -0.726 | -0.289 | -0.158 | 0.050 | -0.151 | -0.143 | 0.018 |
| | | (0.305) | (0.136) | (0.177) | (0.157) | (0.131) | (0.118) | (0.117) | (0.114) | (0.078) |
| Constant | | -1.983 | -1.639 | -1.812 | -1.193 | -1.869 | -1.179 | -1.856 | -0.949 | -1.750 |
| | | (0.359) | (0.173) | (0.183) | (0.181) | (0.180) | (0.183) | (0.180) | (0.173) | (0.186) |
| N | | 1352 | 4005 | 4009 | 4211 | 3999 | 4372 | 4248 | 5015 | 4881 |
| Mat.Rate | | 60.9% | 70.4% | 69.7% | 70.1% | 72.1% | 71.5% | 72.0% | 71.4% | 65.8% |

Standard errors are reported in parentheses. N: sample size, Mat.Rate: matching rate

The LHS variable is the odds of staying in the same address.

Mari.Stat.: 1 if married; LogWage: log of hourly rate of pay (yrs 1&2); NoWork: no reported wage (yrs 1&2)

Ysm: years since migration; Citizen: 1 if U.S. citizen; Constant: immigrants from Central & South America

Table 2C. Population Attrition Process Estimates

| | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 |
|-----------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| | -1995 | -1996 | -1997 | -1998 | -1999 | -2000 | -2001 | -2002 | -2003 | -2004 |
| Age/10 | 0.017 | -0.003 | 0.023 | 0.010 | -0.002 | 0.016 | -0.009 | 0.001 | -0.009 | -0.001 |
| | (0.020) | (0.020) | (0.020) | (0.020) | (0.020) | (0.019) | (0.018) | (0.019) | (0.018) | (0.018) |
| Ysm/10 | 0.008 | -0.001 | -0.010 | -0.004 | -0.011 | -0.007 | 0.022 | 0.008 | 0.024 | 0.024 |
| | (0.022) | (0.023) | (0.023) | (0.023) | (0.021) | (0.022) | (0.020) | (0.021) | (0.018) | (0.019) |
| Education | 0.004 | -0.006 | 0.004 | 0.001 | 0.007 | 0.003 | 0.007 | 0.003 | 0.008 | 0.004 |
| | (0.005) | (0.005) | (0.005) | (0.005) | (0.005) | (0.005) | (0.005) | (0.005) | (0.005) | (0.005) |
| Europe | -0.042 | -0.063 | -0.001 | -0.016 | 0.014 | -0.009 | -0.030 | -0.006 | -0.045 | -0.035 |
| | (0.061) | (0.060) | (0.061) | (0.060) | (0.058) | (0.059) | (0.057) | (0.057) | (0.053) | (0.056) |
| Asia | -0.014 | -0.071 | 0.001 | 0.037 | -0.014 | -0.024 | 0.000 | 0.023 | -0.063 | -0.029 |
| | (0.055) | (0.053) | (0.053) | (0.051) | (0.051) | (0.051) | (0.049) | (0.049) | (0.047) | (0.049) |
| Others | -0.309 | -0.240 | 0.115 | 0.071 | 0.065 | -0.002 | 0.002 | -0.065 | -0.004 | -0.036 |
| | (0.064) | (0.091) | (0.097) | (0.110) | (0.099) | (0.085) | (0.076) | (0.076) | (0.072) | (0.078) |
| Constant | -0.073 | -0.031 | -0.149 | -0.083 | -0.081 | -0.055 | -0.049 | -0.011 | -0.106 | -0.188 |
| | (0.088) | (0.091) | (0.090) | (0.089) | (0.083) | (0.086) | (0.082) | (0.083) | (0.077) | (0.082) |
| N | 10534 | 9920 | 10010 | 10184 | 10801 | 10892 | 12212 | 12186 | 13681 | 12749 |

Standard errors are reported in parentheses. N: sample size

The LHS variable is the odds of staying in the United States.

Ysm: years since migration

Constant: immigrants from Central & South America; Continent Dummies are Deviations from the Constant:

Europe: Europe, Australia, New Zealand, and Canada; Africa: Africa and other countries

Table 2D-1. Attrition-Correcting Weighting Function Estimates (Immigrants): All Wages

| | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 |
|------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| | -1995 | -1996 | -1997 | -1998 | -1999 | -2000 | -2001 | -2002 | -2003 | -2004 |
| Age | 0.034 | 0.032 | 0.024 | 0.036 | 0.026 | 0.029 | 0.031 | 0.028 | 0.024 | 0.028 |
| | (0.004) | (0.005) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) |
| Education | 0.024 | 0.010 | 0.014 | 0.016 | 0.013 | 0.050 | -0.024 | 0.039 | 0.013 | 0.001 |
| | (0.010) | (0.012) | (0.006) | (0.006) | (0.006) | (0.006) | (0.006) | (0.006) | (0.005) | (0.006) |
| Mari.Stat. | 0.220 | 0.368 | 0.480 | 0.339 | 0.608 | 0.466 | 0.115 | 0.618 | 0.307 | 0.255 |
| | (0.089) | (0.108) | (0.052) | (0.053) | (0.050) | (0.050) | (0.045) | (0.047) | (0.042) | (0.047) |
| LogWage1 | -0.011 | 0.243 | -0.327 | -0.027 | -0.019 | 0.230 | 0.123 | -0.120 | -0.103 | 0.195 |
| | (0.096) | (0.121) | (0.055) | (0.054) | (0.053) | (0.057) | (0.048) | (0.049) | (0.047) | (0.053) |
| LogWage2 | 0.059 | -0.038 | 0.431 | 0.200 | 0.037 | -0.057 | 0.205 | 0.330 | 0.066 | 0.105 |
| | (0.089) | (0.106) | (0.057) | (0.061) | (0.055) | (0.056) | (0.051) | (0.050) | (0.048) | (0.052) |
| NoWork1 | 0.211 | 0.402 | -0.736 | -0.312 | 0.053 | 0.571 | 0.139 | -0.127 | -0.320 | 0.300 |
| | (0.236) | (0.299) | (0.141) | (0.141) | (0.136) | (0.145) | (0.127) | (0.130) | (0.123) | (0.141) |
| NoWork2 | -0.201 | 0.093 | 1.122 | 0.699 | -0.248 | 0.064 | 0.485 | 0.747 | 0.227 | 0.251 |
| | (0.229) | (0.272) | (0.146) | (0.156) | (0.144) | (0.144) | (0.133) | (0.133) | (0.126) | (0.140) |
| Ysm | 0.052 | 0.250 | 0.045 | 0.044 | 0.024 | 0.097 | 0.030 | 0.094 | 0.035 | 0.029 |
| | (0.004) | (0.005) | (0.003) | (0.003) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) |
| Citizen | -0.405 | 0.048 | 0.108 | 0.248 | 0.157 | -0.361 | 0.151 | 0.142 | 0.172 | 0.242 |
| | (0.089) | (0.107) | (0.051) | (0.049) | (0.048) | (0.048) | (0.044) | (0.044) | (0.042) | (0.046) |
| Europe | 0.350 | 0.497 | 0.048 | 0.255 | 0.363 | -0.034 | 0.218 | 0.014 | 0.205 | -0.037 |
| | (0.102) | (0.120) | (0.063) | (0.063) | (0.059) | (0.062) | (0.059) | (0.060) | (0.054) | (0.061) |
| Asia | 0.348 | -0.226 | 0.074 | 0.194 | 0.078 | -0.041 | 0.211 | 0.045 | -0.251 | -0.044 |
| | (0.100) | (0.121) | (0.058) | (0.057) | (0.054) | (0.056) | (0.052) | (0.052) | (0.050) | (0.054) |
| Africa | 0.900 | -1.105 | 0.385 | -0.752 | -0.172 | 0.060 | 0.068 | -0.241 | -0.157 | -0.269 |
| | (0.107) | (0.214) | (0.100) | (0.131) | (0.104) | (0.092) | (0.082) | (0.082) | (0.078) | (0.089) |
| Constant | -1.995 | -2.040 | -1.562 | -2.141 | -1.175 | -2.592 | -1.320 | -2.561 | -0.938 | -1.861 |
| | (0.212) | (0.270) | (0.129) | (0.138) | (0.130) | (0.135) | (0.124) | (0.128) | (0.120) | (0.129) |
| N | 2159 | 1714 | 4965 | 5021 | 5339 | 5284 | 5885 | 5825 | 6771 | 6617 |
| Mat.Rate | 66.3% | 60.3% | 70.1% | 68.7% | 70.1% | 70.8% | 71.4% | 71.6% | 70.1% | 65.0% |

Standard errors are reported in parentheses. N: sample size, Mat.Rate: matching rate

The LHS variable is the odds of staying in the same address.

Mari.Stat.: 1 if married; LogWage: log of hourly rate of pay (yrs 1&2); NoWork: no reported wage (yrs 1&2)

Ysm: years since migration; Citizen: 1 if U.S. citizen; Constant: immigrants from Central & South America

Table 2D-2. Attrition-Correcting Weighting Function Estimates (Immigrants): Reported Wages Only

| | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 |
|------------|-------|--------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| | -1995 | -1996 | -1997 | -1998 | -1999 | -2000 | -2001 | -2002 | -2003 | -2004 |
| Age | N/A | 0.018 (0.006) | 0.015 (0.003) | 0.029 (0.003) | 0.018 (0.003) | 0.015 (0.003) | 0.020 (0.003) | 0.018 (0.003) | 0.018 (0.003) | 0.019 (0.003) |
| Education | | 0.010 (0.016) | 0.007 (0.008) | -0.007 (0.008) | 0.001 (0.008) | 0.017 (0.008) | -0.018 (0.008) | 0.023 (0.008) | 0.002 (0.007) | -0.001 (0.008) |
| Mari.Stat. | | 0.561 (0.139) | 0.535 (0.069) | 0.453 (0.069) | 0.513 (0.068) | 0.469 (0.067) | 0.210 (0.065) | 0.480 (0.067) | 0.281 (0.060) | 0.241 (0.067) |
| LogWage1 | | 0.271 (0.203) | -0.226 (0.095) | -0.005 (0.103) | -0.129 (0.099) | 0.368 (0.111) | 0.051 (0.094) | 0.244 (0.089) | -0.119 (0.088) | 0.140 (0.106) |
| LogWage2 | | -0.153 (0.189) | 0.410 (0.097) | 0.214 (0.117) | 0.160 (0.103) | -0.091 (0.111) | 0.191 (0.099) | -0.112 (0.089) | 0.136 (0.093) | 0.085 (0.104) |
| NoWork1 | | 0.511 (0.465) | -0.304 (0.224) | -0.085 (0.238) | 0.087 (0.237) | 0.903 (0.259) | 0.165 (0.231) | 0.461 (0.225) | -0.087 (0.215) | 0.281 (0.264) |
| NoWork2 | | -0.057 (0.443) | 1.138 (0.232) | 0.756 (0.274) | 0.103 (0.250) | 0.088 (0.262) | 0.542 (0.242) | 0.033 (0.225) | 0.348 (0.227) | 0.232 (0.260) |
| Ysm | | 0.014 (0.007) | 0.035 (0.003) | 0.018 (0.003) | 0.020 (0.003) | 0.051 (0.003) | 0.155 (0.003) | 0.055 (0.003) | 0.019 (0.003) | 0.021 (0.003) |
| Citizen | | -0.095 (0.139) | -0.053 (0.068) | 0.151 (0.064) | 0.057 (0.067) | -0.193 (0.065) | 0.088 (0.064) | 0.071 (0.062) | 0.071 (0.061) | 0.219 (0.064) |
| Europe | | 0.496 (0.156) | 0.050 (0.084) | 0.164 (0.082) | 0.167 (0.083) | -0.027 (0.084) | 0.018 (0.086) | 0.085 (0.085) | 0.151 (0.077) | 0.000 (0.086) |
| Asia | | 0.011 (0.158) | 0.097 (0.076) | 0.061 (0.075) | 0.055 (0.075) | -0.049 (0.077) | 0.013 (0.076) | 0.038 (0.077) | -0.139 (0.071) | -0.005 (0.078) |
| Africa | | -1.080 (0.300) | 0.255 (0.136) | -0.645 (0.176) | -0.237 (0.156) | -0.174 (0.132) | 0.050 (0.118) | -0.242 (0.118) | -0.151 (0.114) | -0.472 (0.137) |
| Constant | | -1.701 (0.0357) | -1.752 (0.173) | -1.836 (0.183) | -1.251 (0.181) | -1.994 (0.180) | -1.289 (0.183) | -1.997 (0.181) | -1.056 (0.173) | -1.760 (0.185) |
| N | | 1352 | 4005 | 4009 | 4211 | 3999 | 4372 | 4248 | 5015 | 4881 |
| Mat.Rate | | 60.9% | 70.4% | 69.7% | 70.1% | 72.1% | 71.5% | 72.0% | 71.4% | 65.8% |

Standard errors are reported in parentheses. N: sample size, Mat.Rate: matching rate

The LHS variable is the odds of staying in the same address.

Mari.Stat.: 1 if married; LogWage: log of hourly rate of pay (yrs 1&2); NoWork: no reported wage (yrs 1&2)

Ysm: years since migration; Citizen: 1 if U.S. citizen; Constant: immigrants from Central & South America