Sample Attrition in the Presence of Population Attrition

Seik Kim

Department of Economics

University of Washington

seikkim@u.washington.edu

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

November 18, 2012

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. When there is population attrition, refreshment samples are 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. This paper shows that the problem can be resolved by generating a counterfactual, but representative cross-section prior to applying their procedure. The proposed method is used to obtain attrition-correcting weights for the native and immigrant panel samples in the Current Population Survey.

Keywords: Immigration, Population Attrition, Sample Attrition

JEL Classification Codes: C23, C81, J61

[0] 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, Taisuke Otsu, Peter Phillips, and seminar participants at Purdue University, Seoul National University, University of Carlifornia-Davis, 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. One kind of attrition, which we call sample attrition, occurs when a respondent is not interviewed while he or she remains in the population. A simple example of sample attrition 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 of population attrition is decease. Population attrition is often very small and is ignored in analyses. In some cases, however, 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 result in sample attrition. For example, consider a two-year 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.[1] 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 other U.S. residents with an equal probability.

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 have an equal probability of being selected in a 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 conditional on the time of last entry to the United States.[2] Therefore, the second period cross-section is not

[1] This person might be missing because of decease, emigration, or other reasons, but these possibilities for working age persons are relatively low and negligible compared to return migration of the foreign-born population in the United States.

[2] Unconditional on the time of the most recent arrival to the United States, the population is non-stationary over time because of new immigrants. This is more general than population attrition and includes population

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), Nevo (2003), and Bhattacharya (2008) uses the availability of representative cross-sections as the basis for weighting the persons in a balanced panel. Without loss of generality, assume that the first period population is the population of interest. The attrition-correcting weighting function is given by the inverse of one minus the probability of sample attrition. The identification strategy requires that the auxiliary samples are representative cross-sections of the target population throughout the entire sampling period of the panel sample. When there is attrition in the population of interest, however, refreshment samples are not representative, 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, usually the first period cross-section, is representative of the target population, while the balanced panel and the other cross-sections are not. Section 2 presents identification and estimation of a two-period panel data model with sample attrition in the presence of population attrition, where 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. Once the counterfactual sample is produced, the remainder of identification and estimation strategies is identical to Bhattacharya (2008).

The representative counterfactual sample can be obtained by weighting the second period cross-section by one minus the probability of population attrition. This paper shows that the population attrition function can be identified when the function is determined by variables of known transition probability. These variables, for example, include deterministic variables such as year of entry or age. The proposed method separately identifies sample attrition and population attrition processes. This is useful because samples usually do not indicate which

attrition as a special case. The method developed in this paper can be applied to non-stationary population cases.

missing observations are due to sample attrition and which are due to population attrition. The method is applicable to the analysis of questions where population attrition is a potential problem such as seasonal migration in developing countries and entry and exit of firms in a market. The method can be also used to properly weight a non-representative panel when administrative cross-sections are available.

Section 3 applies the outlined technique to obtain attrition-correcting weights for the native-born and foreign-born panel samples in the Current Population Survey (CPS). To analyze the economic performance of immigrants in the United States, a sufficiently large longitudinal sample is desirable since immigrants are minorities and unobserved individual heterogeneity needs to be controlled for. The CPS satisfies these criteria. It is a collection of two-year panels and has the crucial advantage of being much larger than alternative panel data sets. In the CPS, however, attrition is particularly severe as the survey does not follow households who change residences. Moreover, the immigrant sample suffers from population attrition caused by selective return migration as well as sample attrition due to changes in residence. Not accounting for population attrition will overstate the economic performance of immigrants if foreign-born individuals emigrate because they were unsuccessful in the labor market.[3]

To address these attrition problems, this paper exploits the cross-sectional structure 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. These weights, once constructed, can be used in various studies of immigration using the CPS.

[3] One needs to account for return migrants when the labor market performance of immigrants is of interest. One does not need to, however, when the impact of immigrant workers on the U.S. economy is of interest.

2 Correcting for Attrition

2.1 Previous Literature on Attrition and Refereshment Samples

Suppose 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.

Following Bhattacharya (2008), suppose the model of interest is given by a conditional moment restriction

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

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 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,

$$E\left[m\left(y_1, y_2, x_1, x_2, \theta_0\right)|x_1, x_2\right] \neq E\left[m\left(y_1, y_2, x_1, x_2, \theta_0\right)|x_1, x_2, D_S = 1\right]. \tag{2}$$

Therefore, simply using the matched sample will result in an inconsistent estimator of θ.

Now assume that in addition to the panel data there is a representative cross-section available in the second period.[4] This second period cross-section is called the refreshment sample. Suppose 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 example, 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). v is a vector of time-invariant exogenous variables in x_1. The attrition function

[4] The first wave of the longitudinal sample serves as a representative cross-section sample since it is representative of the target population. In some cases, an auxiliary cross-section sample is available for the first period as well as the second period. The CPS is one such case.

does not have to be determined by the same variables in the main model (1). The variables in (u_1, u_2, v) are a subset of those in (y_1, y_2, x_1, x_2).

To obtain the LHS of (2) we need to learn about the joint density, $f(u_1, u_2, v)$. We assume that the conditional probability of responding in the second period, $\Pr(D_S = 1|u_1, u_2, v)$, is strictly positive almost everywhere. Then due to the following identity,

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

identification of the unconditional joint density, $f(u_1, u_2, v)$, is implied by identification of the response probability, $\Pr(D_S = 1|u_1, u_2, v)$. This result is because $f(u_1, u_2, v|D_S = 1)$ and $\Pr(D_S = 1)$ can be directly estimated from the balanced panel and the full panel, respectively.

Hirano, Imbens, Ridder, and Rubin (2001) prove that $\Pr(D_S = 1|u_1, u_2, v)$ is nonparametrically just-identified up to a known link function, $g(\cdot)$, if its argument takes an additive non-ignorable form:

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

where $k_{\cdot}(\cdot)$ 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 $\lim_{r\to-\infty} g(r) = 0$ and $\lim_{r\to\infty} g(r) = 1$. It is non-ignorable in the sense that attrition determined by the first period variables only is called ignorable attrition. Identification results from the fact that two marginal densities, $f(u_1, v)$ and $f(u_2, v)$ are observed from the year one and the year two cross-sections, and $f(u_1, v)$ and $f(u_2, v)$ obey

$$\begin{aligned} f(u_1, v) &= \int \frac{\Pr(D_S = 1)}{\Pr(D_S = 1|u_1, u_2, v)} f(u_1, u_2, v|D_S = 1)\, du_2, \\ f(u_2, v) &= \int \frac{\Pr(D_S = 1)}{\Pr(D_S = 1|u_1, u_2, v)} f(u_1, u_2, v|D_S = 1)\, du_1, \end{aligned} \tag{4}$$

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 equations in (4) can be transformed into conditional moment restrictions:

$$\begin{aligned} 1 &= E\left[\frac{D_S}{\Pr\left(D_S = 1|u_1, u_2, v\right)}|u_1, v\right] \quad \text{w.p.1,} \\ 1 &= E\left[\frac{D_S}{\Pr\left(D_S = 1|u_1, u_2, v\right)}|u_2, v\right] \quad \text{w.p.1.} \end{aligned} \tag{5}$$

The transformed identification equations in (5) can be estimated, for example, by the sieve minimum distance developed by Ai and Chen (2003). It can be estimated by the smoothed empirical log-likelihood developed by Kitamura, Tripathi, and Ahn (2004) when a parametric attrition process is specified.

Once $k_0(v) + k_1(u_1, v) + k_2(u_2, v)$ and $\Pr(D_S = 1)$ are estimated, it is possible to construct the attrition-correcting weighting function

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

The weighting function is proportional to, $1/g\left(k_0(v) + k_1(u_1, v) + k_2(u_2, v)\right)$, the inverse of one minus the probability of attrition. Then, we weight the matched sample by (6) and estimate

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

to obtain a consistent estimator of θ. In sum, the model with attrition can be estimated consistently by assigning attrition-correcting weights to the individuals in the matched sample.[5]

The attrition-correcting method has several attractive features. First, the sample attrition function for a longitudinal sample is identified nonparametrically under relatively weak conditions. The link function can be logit or probit. The additive non-ignorable assumption for the model reduces the dimension of the attrition function of our interest.[6] Second, the correction

[5] The weights and the parameter in the main model, θ, can be estimated jointly. See Bhattacharya (2008) for details.

[6] As an additive non-ignorable attrition model includes the first and the second period variables, but not interactions between the variables in the first and the second periods. For example, sample attrition can depend on $\log wage_2 - \log wage_1$ but not on $(wage_2 - wage_1)/wage_1$, although both measure wage growth.

is robust to individual fixed effects. This is the case because each individual receives a unique weight, and individual fixed effects will be differenced out by the usual fixed effects strategies for panel data models.

2.2 Identification in the Presence of Population Attrition

When there is attrition in the target population and the model of interest involves a counterfactual situation of a stationary population, the existing 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. An individual is in the matched sample if $D_P = 1$ and $D_S = 1$. Similarly, an individual stays in the United States but does not respond in the second period if $D_P = 1$ and $D_S = 0$. An individual who leaves the United States 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 United States at the same time so that $D_P = 1$ and $D_S = 1$.

Again, the model of interest is given by a conditional moment restriction (1). We observe the joint distribution of the matched sample, $(y_1, y_2, x_1, x_2) \mid (D_P = 1, D_S = 1)$.[7] Similar to (2), simply using the balanced panel will lead to an inconsistent estimator. In the presence of population attrition, the LHS of the second condition in (4), $f(u_2, v)$, is not directly estimable. Instead, we observe $f(u_2, v|D_P = 1)$ from the second period cross-section. Using the following identity,

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

[7] In the presence of population attrition, $u_2 = (y_2, x_2)$ is the potential outcome of the individual corresponding to the situation where he or she remains in the population. For instance, in the application considered in this paper, y_2 corresponds to the potential wage the individual would obtain when staying in the United States.

(4) becomes

$$\begin{aligned} f\left(u_1, v\right) &= \int \frac{\Pr\left(D_S = 1\right)}{\Pr\left(D_S = 1 | u_1, u_2, v\right)} f\left(u_1, u_2, v | D_S = 1\right) du_2, \\ \frac{f\left(u_2, v | D_P = 1\right) \Pr\left(D_P = 1\right)}{\Pr\left(D_P = 1 | u_2, v\right)} &= \int \frac{\Pr\left(D_S = 1\right)}{\Pr\left(D_S = 1 | u_1, u_2, v\right)} f\left(u_1, u_2, v | D_S = 1\right) du_1, \end{aligned} \tag{8}$$

for almost all (u_1, u_2, v). Since the standard semiparametric methods cannot be applied to estimate (8), we transform it into conditional moment restrictions.

Proposition 1. The equations in (8) can be transformed into conditional moment restrictions given by

$$\begin{aligned} 1 &= E\left[\frac{D_S}{\Pr\left(D_S = 1 | u_1, u_2, v\right)} | u_1, v\right] \quad \text{w.p.1}, \\ \frac{1}{\Pr\left(D_P = 1 | u_2, v\right)} &= E\left[\frac{D_S}{\Pr\left(D_S = 1 | u_1, u_2, v\right)} | u_2, v, D_P = 1\right] \quad \text{w.p.1}. \end{aligned} \tag{9}$$

Proof. The equivalence of the first equation in (8) and the first conditional moment restriction in (9) is shown by Bhattacharya (2008). We show equivalence of the second equation in (8) and the second conditional moment restriction in (9). Divide both sides of the second condition in (8) by $f\left(u_2, v | D_P = 1\right) \Pr\left(D_P = 1\right)$, and we have

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

where the second equation uses

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

and

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

as $D_S=1$ implies $D_P=1$ and $D_S=1$. □

In the first equation of (9), the LHS is unity and the RHS is equivalent to weighting the individuals in the matched sample with the inverse of one minus the probability of sample attrition, $1/\Pr\left(D_S=1|u_1,u_2,v\right)$. In the second period, population attrition occurs and the LHS needs to be adjusted. Intuitively, the LHS of the second equation 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, $1/\Pr\left(D_P=1|u_2,v\right)$.

The next step is to find a candidate for $\Pr\left(D_P=1|u_2,v\right)$. When $\Pr\left(D_P=1|u_2,v\right)$ 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 $P\left(Z_2=z_2|Z_1=z_1\right)$, where z is a vector of variables of known transition probability. For example, if an element of z is year of entry, the transition probability is given by $P\left(z_2|z_1\right)=1\left(z_2=z_1\right)$, where $1\left(\cdot\right)$ is the indicator function. If an element of z is age, the transition probability is given by $P\left(z_2|z_1\right)=1\left(z_2=z_1+1\right)$.

Proposition 2. The population attrition process, $\Pr\left(D_P=1|u_2,v\right)$, is nonparametrically identified when the population attrition is solely determined by variables of known transition probability, z_2, where the variables in z_2 must be included in $\left(u_2,v\right)$.

Proof.

$$\begin{aligned}\Pr\left(D_P = 1 | u_2, v\right) &= \Pr\left(D_P = 1 | z_2\right) \\ &= \frac{f\left(z_2 | D_P = 1\right)\Pr\left(D_P = 1\right)}{f\left(z_2\right)} \\ &= \frac{f\left(z_2 | D_P = 1\right)\Pr\left(D_P = 1\right)}{\int f\left(z_1\right) p\left(z_2 | z_1\right) dz_1}.\end{aligned}$$

The last equation uses the fact the transition probability from $Z_1 = z_1$ to $Z_2 = z_2$ is known. Note that $f\left(z_1\right)$ and $f\left(z_2 | D_P = 1\right)$ are known from the first period and second period cross-sections. To estimate $\Pr\left(D_P = 1\right)$ by comparing the two cross-sections, new immigrants arriving between the two cross-section years must be dropped from the second period cross-section sample. □

Proposition 2 implies that one minus the population attrition probability under selection on variables of known transition probability is given by

$$\begin{aligned}\Pr\left(D_P = 1 | u_2, v\right) &= \Pr\left(D_P = 1 | z_2\right) \\ &\equiv k\left(z_2\right), \qquad (10)\end{aligned}$$

where $k\left(\cdot\right)$ is some unknown function using two cross-section samples. It can be estimated nonparametrically as well as parametrically when $k\left(z_2\right)$ is given by a parametric form. This method works under weaker data requirements than models that assume selection on observables or missing at random because (10) is identified from two cross-sections and does not require panel data. Once $\Pr\left(D_P = 1 | u_2, v\right)$ is known, identification and estimation of $\Pr\left(D_S = 1 | u_1, u_2, v\right)$ is identical to Bhattacharya (2008). Hence, if we specify the sample attrition function by

$$\Pr\left(D_S = 1 | U_1 = u_1, U_2 = u_2, V = v\right) = g\left(k'_0\left(v\right) + k'_1\left(u_1, v\right) + k'_2\left(u_2, v\right)\right), \qquad (11)$$

where $k'_{\cdot}\left(\cdot\right)$ and $g\left(\cdot\right)$ are defined as before, the $k'_{\cdot}\left(\cdot\right)$ functions are uniquely determined.

Proposition 3. (Identification) If

(i) conditional on each value v in the support of V, the support $\mathcal{U}_1\left(v\right) \times \mathcal{U}_2\left(v\right)$ of U_1, U_2 is not a lower-dimensional subspace of $R^{2\times\dim(Z)}$,

(ii) (10) and (11) are substituted into equations in (9), i.e.,

$$\begin{aligned}
1 &= E\left[\frac{D_S}{g\left(k_0'\left(v\right)+k_1'\left(u_1,v\right)+k_2'\left(u_2,v\right)\right)}|u_1,v\right] \quad \text{w.p.1,}\\
\frac{1}{k\left(z_2\right)} &= E\left[\frac{D_S}{g\left(k_0'\left(v\right)+k_1'\left(u_1,v\right)+k_2'\left(u_2,v\right)\right)}|u_2,v,D_P=1\right] \quad \text{w.p.1,}
\end{aligned}$$

(iii) $g\left(\cdot\right)$ is a strictly increasing function such that $\lim_{r\to-\infty} g\left(r\right)=0$ and $\lim_{r\to\infty} g\left(r\right)=1$,

(iv) for each v, there exists $\overline{u}_1\left(v\right)\in\mathcal{U}_1\left(v\right)$ and $\overline{u}_2\left(v\right)\in\mathcal{U}_2\left(v\right)$ such that $k_1\left(\overline{u}_1\left(v\right),v\right)=k_2\left(\overline{u}_2\left(v\right),v\right)=0$,

then $k_0'\left(v\right)+k_1'\left(u_1,v\right)+k_2'\left(u_2,v\right)$ is uniquely determined w.p.1.

Proof. The only difference between (9) and (5) is the fact that the LHS of the second equation of the former is $1/k\left(z_2\right)$, while the LHS of the second equation of the latter is unity. Since $k\left(\cdot\right)$ is identified from Proposition 2, the proof for Proposition 3 is identical to Bhattacharya (2008). □

Once the attrition-correcting weighting function

$$C\left(u_1,u_2,v\right)=\frac{\Pr\left(D_S=1\right)}{g\left(k_0'\left(v\right)+k_1'\left(u_1,v\right)+k_2'\left(u_2,v\right)\right)} \tag{12}$$

is estimated, we weight the matched sample by (12) and estimate (7) to obtain a consistent estimator of θ. The assumption of selection on variables of known transition probability in (10) is a strong, but necessary assumption because we do not know who emigrated from the United States. If one has prior knowledge about the dynamics of some stochastic variables, these variables can be used as an element of the z_2 vector. For example, one may have several possible forecasts for annual wage growth rates in the absence of population attrition. Since each of these forecasts will imply a specific transition probability, one can use this information to get a range of estimates under different scenarios. Another example would be the case in which one has additional sources of information from other data sources that allow observation of return migration.

Despite its limitations, the attrition-correcting method has several advantages. First, the population attrition function is identified nonparametrically under selection on variables of known transition probability when repeated cross-sections are available. 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 useful result because data sets do not provide information on who left the population and who left the sample without leaving the population. Finally, the method is robust to fixed effects.

The attrition-correcting technique can be generalized to longer panels and can be applied to applications other than immigration studies. If a panel 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. One such example would be a longitudinal analysis of the working population. In that case, the numerator of the second equation of (8) and (9) should be adjusted to incorporate the movement into and out of the labor force.[8] 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. In labor economics, the method can be applied to properly weight a non-representative panel using administrative cross-section samples.

3 Application

3.1 The Current Population Survey

The matched CPS sample or the CPS Merged Outgoing Rotation Group (MORG) is a collection of panel data sets two years in length initiated every year. As of July 2001, the CPS collects a

[8] The application section of this paper presents an empirical strategy that accounts for population attrition as well as non-stationary population.

sample of approximately 56,000 housing units from 792 sample areas on demographic and labor force characteristics of the civilian non-institutional population 16 years of age and older. When a housing unit is selected, each individual in the unit is asked twice with a one year interval about their economic activities, such as usual weekly earnings and usual weekly hours worked. As the sampling periods of two adjacent two-year panel data sets overlap, short panels may mimic a longer longitudinal sample if combined properly. We call this type of multiple short panels overlapping rotating panel data.

The CPS also serves to provide representative cross-sections. As part of the survey, addresses are selected randomly. These pre-selected housing units are kept unchanged over the interview periods. If the occupants of a selected dwelling unit move, it is the new occupants of the unit who are interviewed. By construction, an individual appears only once in a year, but may not reappear in the following year. Although the interviewees may be replaced by new occupants within the sampling periods, the CPS provides a representative cross-section of each year's population because the random sample of housing units remains fixed. As a result, attrition is directly related to residential mobility within the United States as well as return migration.

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, a rotating panel, such as the CPS, is likely to be larger than a usual panel, such as the Panel Study of Income Dynamics (PSID) or the National Longitudinal Survey of Youth 1979 (NLSY79), because tracking interviewees is less costly. 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 population for any given time period. This feature results because a new two-year panel is initiated from the population in each year. This property is the key in identifying sample attrition and population attrition processes.

3.2 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.[9] The sample used in this analysis is drawn from the matched CPS between 1994 and 2004. Our sample is comprised of foreign-born and native-born men of ages 18-64.[10] We define an individual as matched if the individual appears twice in the matched CPS. 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.[11] The group of the other countries consists of immigrants from Africa, Oceania, and unclassified ones. Due to its small sample size, the data will not be very informative about this group.

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.[12] Between 1994 and 2004, 28-40% of the immigrant samples and 22-32% of the native samples drop out of the sample. In practice, matching is not possible between June 1994 - August 1995 and June 1995 - August 1996 due to sample redesign. If the 1994-1995 and 1995-

[9] Prior to 1994, CPS supplements on immigration were administered to all households participating in the survey in November 1979, April 1983, June 1986, June 1988, and June 1991.

[10] The 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 Unites States, but excludes persons born in Puerto Rico and the outlying areas.

[11] We 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.

[12] The 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 NLSY79, which is 10%. The Census Bureau classifies the noninterviews into three types. Type A noninterviews indicate 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 indicate addresses that may have been converted to permanent businesses, condemned or demolished, or fall outside the boundaries of the segment for which they were selected.

1996 samples are excluded, the attrition rates are 28-35% of 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 percentage points. A part of the gap in the attrition rates may be due to return migration. 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.[13]

The United States stopped collecting information on return migrants in 1957. To estimate the rates of return migration, we exploit the structure of the matched CPS. As housing units in the sample are kept fixed over the sampling period, the relative 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 1 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 first year of the 1994-1995 panel, there were 5,329 foreign-born persons in the United States. In the second year of the 1994-1995 panel, residents at the same addresses are interviewed, but the panel is unbalanced because at some residences, previous residents have moved out and new resients have moved in. Among the new residents, it is possible that some are recent immigrants who would have never appeared in the first year panel. These new individuals are excluded from the analysis so that immigrants in each year's panel have the same range of arrival years.[14] We then count the number of foreign-born persons in the second year of the 1994-1995 panel, which is 5,331. The ratio between these numbers is 1.00 (=5,331/5,329). This suggests that only a very small amount of outmigration occurred during this period. Similarly in 1995-1996, the numbers of the foreign-born persons in the first and the second years are 5,417 and 4,605, respectively. This implies that about 15%

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

[14] In practice, this is complicated with the CPS since the year of arrival information is coded in an inconsistent way for the most recent entrants. For instance, the arrival year code 13 in the 1994 sample includes the 1992-1994 arrivals, the code 13 in the 1995 sample includes the 1992-1995 arrivals, and the code 13 in the 1996 sample and afterwards includes the 1992-1993 arrivals. Therefore foreign-born persons who arrived in the United States in 1992-1993 and are in the 1994-1995 or the 1995-1996 panels cannot be matched. In consequence, we drop immigrants with the arrival year code 13 in the 1994-1995 or the 1995-1996 panels. Thus, the most recent immigrants in the 1994-1995 and the 1995-1996 panels are those who entered the United States in 1990-1991 with the arrival year code of 12.

(=1–4,605/5,417) of the foreign-born population in 1995 left the United States in 1996.

Table 1. Stay Probability (One Minus the Population Attrition Rate)

	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	1994
	–1995	–1996	–1997	–1998	–1999	–2000	–2001	–2002	–2003	–2004	–2004
All Immig.											
# in Yr. 2	5331	4605	5011	5070	5398	5578	6299	6293	6831	6090	
# in Yr. 1	5329	5417	5121	5220	5527	5435	6060	6021	7001	6811	
Stay Prob.	1.000	0.850	0.979	0.971	0.977	1.026	1.039	1.045	0.976	0.894	0.768
										annual average:	[0.974]
C.S.America											
# in Yr. 2	2530	2224	2515	2561	2768	2937	3281	3237	3690	3320	
# in Yr. 1	2415	2453	2588	2649	2853	2851	3176	3107	3728	3666	
Stay Prob.	1.048	0.907	0.972	0.967	0.970	1.030	1.033	1.042	0.990	0.906	0.860
										annual average:	[0.985]
Europe											
# in Yr. 2	898	866	840	862	942	877	967	974	1075	924	
# in Yr. 1	890	1059	864	908	955	860	952	932	1123	1053	
Stay Prob.	1.009	0.818	0.972	0.949	0.986	1.020	1.016	1.045	0.957	0.877	0.683
										annual average:	[0.963]
Asia											
# in Yr. 2	1259	1265	1404	1457	1448	1438	1629	1670	1603	1472	
# in Yr. 1	1198	1540	1417	1483	1491	1409	1533	1562	1687	1668	
Stay Prob.	1.051	0.821	0.991	0.982	0.971	1.021	1.063	1.069	0.950	0.882	0.793
										annual average:	[0.977]
Others											
# in Yr. 2	644	250	252	190	240	326	422	412	463	374	
# in Yr. 1	826	365	252	180	228	315	399	420	463	424	
Stay Prob.	0.780	0.685	1.000	1.056	1.053	1.035	1.058	0.981	1.000	0.882	0.562
										annual average:	[0.944]

# in Yr. 2 (or Yr. 1): the number of foreign-born persons in the 1st (2nd) year

Stay Prob.: the ratio between the numbers of foreign-born persons in the 2nd and in the 1st years

Conceptually, it is not possible to have the stay rate exceed unity (or the outmigration rate be below zero). In the sample, however, values greater than unity are observed frequently, implying that measurement error is relatively large.[15] Taking this into account, the last column reports the stay probabilities over the entire sample period. For example, 0.768 is obtained by multiplying ten annual stay probabilities over 1994-2004. It suggests that 25.2% (=1–0.768) of the foreign-born population who arrived in the United States in 1994 or before left the country by 2004.[16] The numbers in square brackets in the last column show the geometric means of the 1994-2004 estimates. 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 the lower panel of Table 1. Foreign-born persons from Central and South America are the most likely to stay in the United States among the immigrant groups, followed by those from Asia, from Europe, and from other countries.

3.3 Estimation of Attrition-Correcting Weights

The estimation strategy consists of three steps. In the first step, we estimate the population attrition function and weight the second period cross-section. In the second step, we estimate the sample attrition function and obtain the weights for individuals in the balanced longitudinal sample. Finally, we estimate the main model using the matched sample along with the attrition-correcting weights. For expositional purposes, this method is presented in multiple steps using parametric specifications, but all of these steps can be done simultaneously and nonparametrically.

The identity in Proposition 2 implies

$$\begin{aligned} \Pr\left(D_P=1\right) E_{Z_2}\left[Z_2|D_P=1\right] &= E_{Z_2}\left[k\left(Z_2\right)Z_2\right] \\ &= E_{Z_1}\left[\textstyle\int k\left(z\right) z P\left(dz|Z_1\right)\right]. \qquad (13) \end{aligned}$$

[15] Borjas and Bratsberg (1996) also find negative outmigration rates for some groups of immigrants using the 1980 Census and administrative data from the Immigration and Naturalization Services.

[16] This estimate is consistent with other empirical findings. For instance, Warren and Peck (1980) estimate that more than one-sixth of total immigrants admitted during the 1960s emigrated by the end of the decade.

The first equation represents that the product of the probability of population attrition and the expectation of Z_2 in the presence of population attrition is identical to the expectation of $\Pr(D_P = 1|Z_2) \times Z_2$ in the absence of population attrition. The second equation replaces Z_2 with Z_1 using the known transition probability.

The sample analog of (13) is given by

$$\begin{aligned} \frac{1}{n_2}\Pr(D_P = 1)\textstyle\sum_{j=1}^{n_2} z_{2j} &= \frac{1}{n_1}\textstyle\sum_{i=1}^{n_1}\left[\int k(z)\, zP(dz|z_{1i})\right] \\ &= \frac{1}{n_1}\textstyle\sum_{i=1}^{n_1}\sum_{z\in S_2} k(z)\, z\Pr(z|z_{1i}), \end{aligned} \tag{14}$$

where n_1 and n_2 are the sample sizes of the first and the second period cross-sections, respectively.[17] The second equation holds if z is a vector of discrete variables, where S_2 is the support of Z_2. 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.

In this application, we specify the population attrition function (10) 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 as well as a constant.[18] In this case, all the variables in z_1 have deterministic time paths and map to z_2 one-to-one.[19]

[17] Technically, this part of the 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.

[18] Since $k(r)$ is allowed to be above 1 in this specification, this function can also be used when the target population is not stationary over time, which includes population attrition as a special case. If the year of entry is not observable, one cannot exclude from the year 2 sample in-migrants who enter after year 1, and the population becomes non-stationary.

[19] This choice of variables is restrictive, however, since the transition probabilities of other variables, such as labor market performance variables, are usually unknown.

Therefore, without loss of generality we estimate $k(z_1'\psi)$.

In principle, the population attrition process can be estimated by applying the generalized method of moments to (14), but in this analysis, we demonstrate a simpler method. Consider the following transformation:

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

We estimate $q(z_1'\psi)$ and then transform it to $k(z_1'\psi)$. Since the population attrition function can be identified from two cross-sections, we generate an indicator variable that is set to zero for individuals in the first period cross-section and unity for individuals in the second period cross-section, where new immigrants between the two cross-section years are excluded from the second period sample. Then, we can use a logit model

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

and obtain $q\left(z_1'\widehat{\psi}\right)$. To make this procedure more specific, suppose there is no population attrition and assume that the sample sizes are the same. Then there will be an approximately equal number of 0's and 1's, so it follows that $q(z_1'\psi) = 1/2$ for all z_1. If population attrition occurs to individuals with $z_1 = \widetilde{z}_1$, we expect $q(\widetilde{z}_1'\psi) < 1/2$.

The sample attrition functions in (3) and (11) are parameterized by

$$\begin{aligned} \Pr(D_S = 1 | U_1 = u_1, U_2 = u_2, V = v) &= g(v'\phi_0 + u_1'\phi_1 + u_2'\phi_2) \\ &\equiv g(u_1, u_2, v, \phi), \end{aligned}$$

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 $g(r) = e^r / (1 + e^r)$.

The conditional moment restrictions in (9) can be transformed to the following unconditional moment restrictions:

$$\begin{aligned} E\left[\frac{D_S \cdot a\left(u_1, v\right)}{g\left(u_1, u_2, v, \phi\right)}\right] &= E\left[a\left(u_1, v\right)\right], \\ E\left[\frac{D_S \cdot a\left(u_2, v\right)}{g\left(u_1, u_2, v, \phi\right)}\right] &= E\left[\frac{a\left(u_2, v\right)}{k\left(z_2\right)}\right], \end{aligned} \tag{15}$$

for an arbitrary function $a(\cdot)$. Let n be the sample size of the full panel 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, respectively. The distinction between n and n_1 is useful because the CPS provides auxiliary cross-sections for the first and the second periods. However, in the case that the first period of panel sample serves as the representative cross-section, n is equal to n_1.

The sample versions of the LHS of (15) are

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \frac{D_{Si} \cdot a\left(u_{ti}, v_i\right)}{g\left(u_{1i}, u_{2i}, v_i, \theta\right)} &= \frac{1}{n}\sum_{i=1}^{n_m} \frac{1 \cdot a\left(u_{ti}, v_i\right)}{g\left(u_{1i}, u_{2i}, v_i, \theta\right)} + \frac{1}{n}\sum_{i=n_m+1}^{n} \frac{0 \cdot a\left(u_{ti}, v_i\right)}{g\left(u_{1i}, u_{2i}, v_i, \theta\right)} \\ &= \frac{1}{n}\sum_{i=1}^{n_m} \frac{a\left(u_{ti}, v_i\right)}{g\left(u_{1i}, u_{2i}, v_i, \theta\right)}, \qquad \text{for } t = 1, 2, \end{aligned}$$

and those of the RHS of (15) are

$$\begin{aligned} \frac{1}{n_1}\sum_{i=1}^{n_1} a\left(u_{1i}, v_i\right), \qquad \text{for } t &= 1, \\ \frac{1}{n_2}\sum_{i=1}^{n_2} \frac{a\left(u_{2i}, v_i\right)}{k\left(z_2\right)}, \qquad \text{for } t &= 2. \end{aligned}$$

In estimation, the LHS uses the matched longitudinal sample and the RHS uses the representative cross-sections, where the 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 for 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, where ysm represents years since migration.

Table 2. 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 probability 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; Others: Africa and other countries

We estimate the attrition function coefficients, ψ and ϕ, for the matched CPS between 1994-2004 year by year. We do it for each year because residential mobility and return migration may vary by year and across samples. Table 2 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 poorly estimated since the population attrition is very small. The only coefficient estimate that is stable over the matching years is education. Foreign-born persons with more education have higher probabilities of staying in the United States than less educated foreign-born persons. 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. Population attrition is of concern because, for example, if persons with negative wage shocks are more likely to return to their home country, stayers will on average earn higher wages than return migrants, and estimates using only stayers will tend to overstate relative labor market performance of immigrants compared to natives. In the CPS, the bias due to return migration is not large.

Tables 3 and 4 report the ϕ coefficient estimates for the native and the foreign samples under the assumption that population attrition is negligible. Positive ϕ coefficient estimates imply that the variables are positively correlated with the matching rate or negatively correlated with residential mobility.[20] 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.

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

Table 3. (Sample) Attrition-Correcting Weighting Function Estimates (Natives)

	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 probability 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 4. Attrition-Correcting Weighting Function Estimates (Immigrants)

	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)
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 probability 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

Dummy variables for Europe, Asia, and Others are included, but are not reported.

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.

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 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 other immigrants.

Using the coefficient estimates in Tables 3 and 4, it is possible to calculate attrition-correcting weights, say $C\left(u_1, u_2, v, \widehat{\phi}, \widehat{\psi}\right)$, for all the individuals in the matched CPS. These weights, once constructed, can be used in various studies. If a model is given by conditional moment restrictions (1), we can obtain an estimator based on $E\left[m\left(y_1, y_2, x_1, x_2, \theta\right) \cdot C\left(u_1, u_2, v, \phi, \psi\right) | x_1, x_2, D_S = 1\right] = 0$ 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. As an application of this method, the next subsection estimates the economic performance of foreign-born workers in the United States as compared to native-born workers.[21]

3.4 An Application of Attrition-Correcting Weights

This section demonstrates how attrition-correcting weights can be used in empirical applications by presenting Kim (2012). The objective of the study is to find evidence on whether foreign-born workers assimilate, which we define as the degree to which the wages of foreign-born workers

[21] See Kim (2012, 2013) for details.

approach those of comparable native-born workers with additional time spent in the United States. Following the convention of the literature, economic performance of a worker i, either immigrant or native, in calendar year t can be specified by

$$\begin{aligned} y_{it}^{imm} &= (\alpha_{nat} + \alpha)\, age_{it} + \delta ysm_{it} + (\beta_{nat} + \beta)\, edu_i + \mu_i + \gamma_t + \varepsilon_{it}, & (16) \\ y_{it}^{nat} &= \alpha_{nat} age_{it} + \beta_{nat} edu_i + \mu_i + \gamma_t + \varepsilon_{it}, & (17) \end{aligned}$$

where y is the logarithm of the hourly wage, age is the worker's age, ysm is the number of years since migration, edu is the number of years of education, μ reflects ability or skill endowment, γ represents business cycles, and ε captures idiosyncratic errors. The economic performance of a foreign-born worker relative to a native-born worker at time t can be measured by

$$EA\,(age, ysm; t) = \frac{d}{dt} E\left[y_{it}^{imm} | age, ysm, t\right] - \frac{d}{dt} E\left[y_{it}^{nat} | age, t\right] \qquad (18)$$

This measure reflects the rate of convergence in wages between foreign-born and native-born workers. In the literature, wage convergence from below toward the higher native mean, $EA\,(age, ysm; t) > 0$, is indicative of economic assimilation. For example, consider a 30 year-old foreign-born worker who has lived in the United States for 5 years. Suppose that his wage grows faster than the wage of a 30 year-old native-born worker. This represents economic assimilation because the wage gap between these two individuals will narrow in the following year.

In addition to the benchmark models of (16) and (17), models that include age_{it}^2 and ysm_{it}^2 will be also considered. These equations are estimated by (a) accounting for sample attrition and outmigration and (b) accounting for sample attrition only as well as (c) without adjustment.[22]

[22] The main (wage) equations use the matched longitudinal sample of workers with positive wages. In this step, we exclude individuals with too high or too low wages and negative potential experience. In estimation of the matching functions, we use the matched longitudinal sample of individuals and cross-sections of all individuals including those not working, but we exclude extreme wage observations. Not-working individuals are included in this step in order to reflect market level changes, such as in the composition of natives, between consecutive years. In estimation of outmigration, we use the (unbalanced) panel of all individuals including extreme wage observations. In estimation of the outmigration process, labor market outcomes are not used as the variables must have known transition probabilities. To make sure that the foreign sample is large enough, we keep the largest available sample.

In the benchmark models of (16) and (17), the measure of assimilation in (18) simplifies to

$$
\begin{aligned}
EA\left(age, ysm; t\right) &= \left(\alpha_{nat} + \alpha + \delta + \frac{d}{dt}\gamma_t\right) - \left(\alpha_{nat} + \frac{d}{dt}\gamma_t\right) \\
&= \alpha + \delta.
\end{aligned}
$$

However, when age and years since migration enter as polynomials, it is difficult to interpret the implications of the coefficient estimates. Therefore, we present the regression results by predicting the wage path of a foreign-born worker who arrives in the United States at age 20 as many other studies do. This is a reasonable assumption since in the data the average age is about 40 and the average years since migration is about 20.

Table 5 reports the economic assimilation estimates, $EA\left(age, ysm\right)$, evaluated at $\left(age, ysm\right) = (24, 4)$, $(32, 12)$, $(40, 20)$, and $(48, 28)$. Positive estimates suggest assimilation, and negative estimates imply that immigrant and native wages diverge. The estimates are reported in percentage points. For example, –0.06 in the first line of the first column is interpreted as each additional year in the United States immigrant wages grow at a slower rate than native wages by 0.06 percentage points when sample attrition and population attrition are accounted for. This estimate is derived from the observation that immigrant wages grow annually by 2.75% and native wages by 2.81% under the assumption that year fixed effects on the level of wages are constant between two adjacent years (not reported in the paper). The difference is –0.06 percentage points and is not statistically different from zero.[23]

The first two columns of Table 5 present the estimates of economic assimilation that accounts for sample attrition and population attrition. The sample attrition-population attrition-adjusted estimates from the quadratic specification suggest that wages of foreign-born workers grow slower than those of native-born workers by 1.33 percentage points per year at age 24. When they become 32, the speed of divergence slows down, but immigrant wages still grow slower than native wages by 0.73 percentage points per year. These assimilation estimates are statistically different from zero. The nonlinear specification results reveal that young foreign-born workers

[23] To be precise, one-sided test should be used instead of a two-sided test, as the alternative hypothesis is given by either $EA\left(age, ysm\right) > 0$ or $EA\left(age, ysm\right) < 0$.

fall behind rather than catch up.

Table 5. Economic Assimilation Estimates in Percentage Points

	SP-Adjusted		S-Adjusted		Not Adjusted	
	linear	quadratic	linear	quadratic	linear	quadratic
age=24, ysm=4	–0.06	–1.33**	–0.03	–1.23**	0.20	–0.96
	(0.34)	(0.59)	(0.34)	(0.59)	(0.34)	(0.59)
age=32, ysm=12		–0.73*		–0.68*		–0.50
		(0.38)		(0.38)		(0.39)
age=40, ysm=20		–0.14		–0.13		–0.04
		(0.36)		(0.36)		(0.35)
age=48, ysm=28		0.45		0.42		0.42
		(0.53)		(0.53)		(0.54)

Standard errors are reported in parentheses. Confidence levels: 99% (***), 95% (**), 90% (*).
SP-Adjusted: Sample Attrition-Population Attrition-Adjusted; S-Adjusted: Sample Attrition-Adjusted
Estimates represent immigrants' annual percentage wage growth
relative to the natives' percentage wage growth.

The next two columns report sample attrition-adjusted estimates. These estimates are not very different from the sample attrition-population attrition-adjusted estimates. It suggests that the effect of population attrition is minor because the population attrition is not large between two adjacent years. The last two columns report unadjusted estimates. In general, the unadjusted estimates are greater than the sample attrition-adjusted ones, which implies that immigrants with slower wage growth are less likely to be observed in the second year panel than natives with slower wage growth. Since the signs of estimated assimilation measures do not change, there is little evidence of assimilation whether or not attrition is corrected for.

4 Concluding Remarks

This paper develops a method that accounts for sample attrition in the presence of population attrition for use with a two-period panel data model. 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. The attrition-correcting method is computationally straightforward because it is given by models based on 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 the balanced part of the panel.

The method is applied to a longitudinal sample 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.

5 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.

Duleep, Harriet O. and Mark C. Regets (1997): "Measuring Immigrant Wage Growth using Matched CPS Files," *Demography*, 34, 239-249.

Guell, Maia and Luojia 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.

Kim, Seik (2012): "Economic Assimilation of Foreign-Born Workers in the United States: An Overlapping Rotating Panel Analysis," University of Washington Working Paper.

Kim, Seik (2013): "Wage Mobility of Foreign-Born Workers in the United States," *Journal of Human Resources*, forthcoming.

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.

Nevo, Aviv (2003): "Using Weights to Adjust for Sample Selection When Auxiliary Information Is Available" *Journal of Business and Economic Statistics*, 21 (1), 43-52.

Warren, Robert and Jennifer M. Peck (1980): "Foreign-Born Emigration from the United States: 1960 to 1970," *Demography*, 17, 71-84.