RESCALED ADDITIVELY NON-IGNORABLE (RAN) MODEL OF ATTRITION

A CONVENIENT SEMI-PARAMETRIC BIAS CORRECTION FRAMEWORK FOR DATA WITH A SHORT PANEL COMPONENT*

İnsan TUNALI§, Emre EKİNCİ† and Berk YAVUZOĞLU‡

First version, May 2011

Revised, September 2011

This version, February 2012

1 Introduction

Attrition has been a major concern in applied research based on panel data. The study by Hausman and Wise (1979) constitutes an early attempt to model attrition as the outcome of rational economic behavior that can systematically bias the findings based on the balanced panel (subsample of non-attritors). As such the attritition problem is intimately related to the class of problems collected under the title of selectivity (Heckman, 1987). The subject has also drawn the attention of survey researchers (Madow et al., 1983). Formalizations by Little and Rubin (1987) have paved the way for establishing common terminology such as missing at random (which describes situations where non-attrittors constitute a random subsample of the full sample) and ignorable attrition (when attrition does not impart bias on the outcome under study).

Our paper builds on an important contribution by Hirano et al. (2001). This paper approaches the attrition issue as an identification problem that amounts to recovering the joint distribution of interest for the full population, when all we have is a subsample subjected to non-ignorable attrition. They work with a discrete joint distribution that characterizes the finite outcomes of interest, express the attrition probability as a function of the set of outcomes before and after attrition, and

0*Tunalı would like to acknowledge discussions with Geert Ridder that prompted this line of research. Funding was provided by grant no. 109K504 by TUBITAK, The Scientific and Technological Research Council of Turkey. We are grateful to Hüseyin İkizler, Bengi İlhan Yanık and Hayriye Özgül Özkan for research assistance. The first version was presented as a Keynote Lecture by Tunalı during the 12th International Symposium on Econometrics, Operations Research and Statistics held at Pamukkale University, Denizli, 26-29 May 2011.

0§Corresponding author, Department of Economics, Koç University; e-mail: itunali@ku.edu.tr

0†Department of Economics, Cornell University; e-mail: ee94@cornell.edu

0‡Department of Economics, University of Wisconsin, Madison; e-mail: yavuzoglu@wisc.edu

establish that identification can be achieved when unbiased estimates of the marginal distributions are available. While the typical panel data collection effort yields an unbiased estimator of the first round marginal distribution, attrition renders subsequent round marginals suspect. Hirano et al. exploit an independently conducted cross-section survey (so-called refreshment sample) to provide an unbiased estimator for the second round marginal distribution. As usual adjustment of the balanced sample proceeds by using the inverted attrition (selection) probabilities as weights. By equating the row and column sums of the reweighted balanced panel cell counts (fractions) to the respective marginals, a just-identified system of equations that yields the parameter estimates of the weighting function is obtained. Since the weighting function only allows for main effects and rules out interactions, Hirano et al. name this model Additively Non-ignorable (AN) model of attrition. They show that both the popular missing at random formulation and the Hausman-Wise model are nested within the AN model. As such, their model not only offers a more realistic correction for attrition, but it also affords tests of widely used models.

In this paper we modify the AN model so that it is suitable for data collection efforts that have a short panel component. Many household surveys (CPS in the U.S.A., the Canadian Labor Force Survey, and the Household Labor Force Surveys (HLFS) in Turkey, to name a few) have a rotating sample frame which calls for repeated visits to the same household. Often the sampling frame is chosen so that the units that are rotated in constitute a random sample from the population. If the data collection agency provides the weights needed for rendering the subset of new household nationally representative, this amounts to having a refreshment sample, as in Hirano et al. Our approach does not require a refreshment sample.

While CPS is cleary intended for longitudinal use (BLS, 2002), in some cases the data collection agency prefers to treat each round of the data as an independent cross-section. This treatment is especially common in household surveys based on an address-based rotating sample frame, such as HLFS Turkey (TURKSTAT, 2001). Residential addresses are kept in the sample frame for a certain time and visited according to the rotation schedule whether or not any respondents are found. Standard non-response adjustments (based on demographics) are used to obtain unbiased marginal distributions, which in turn serve as the source of published official statistics. Since a subset of the households are surveyed in two adjoining periods, such surveys lend themselves for dynamic analyses. However finding suitable weights is a challenge.

The problem is attributable to the fact that such data not only suffer from attrition (response followed by non-response) but also from substituition (non-response followed by response). As we show below, in these cases a key parameter of the AN model is not identified. However, a correction

scheme which renders the dynamic estimates consistent with the official cross-sectional statistics can still be found. Since this amounts to treating the unidentified probability as a nuisance parameter, we term the new model Rescaled Additively Non-ignorable (RAN) model of attrition. We show that the model can be estimated with semi-parametric methods which are computationally simpler than the EM-algorithm based imputation methods used in Hirano et al.

We begin our formal treatment by introducing our model and relate it to the AN Model. We then discuss our estimation and inference methodolgy. Next, we turn to examples that illustrate the utility and potential limitations of the proposed approach. We complement this section with a short discussion on the lessons learned from our broader investigation. We conclude the paper with a brief summary of the key aspects of our model and the advantages it offers.

2 RAN Model

Consider data collection efforts directed to households which utilize a rotational design, whereby each household remains in the sample frame for a predetermined number of periods. Two advantages are underscored: Firstly, by limiting the number of revisits, the cost of the data collection effort is balanced against the response burden imposed on the households. Secondly, by including a fresh subsample every period, the sample is kept up to date. Although the rotational design yields a short panel, this component is often not exploited for want of weighting schemes consistent with those used in obtaining the cross-sectional estimates.

Without loss of generality we refer to the equally spaced rounds of data collection as the first period and the second period. We distinguish between the complete panel (CP), which includes all subjects intended for repeat visits, and the balanced panel (BP), which only includes subjects who have been succesfully interviewed. We also keep track of households which are rotated out of the sample after period 1, and households which are rotated in at period 2. We introduce three random variables and associated parameters:

$$D = \begin{cases} 1 \text{ if designated for the Complete Panel (w/prob. } = \delta) \\ 0 \text{ if not (w/prob. } = 1 - \delta) \end{cases}, \tag{1}$$

$$CP = \begin{cases} 1 \text{ if observed in the } 1^{st} \text{ period only (w/prob.} = \gamma_1) \\ 2 \text{ if observed in the } 2^{nd} \text{ period only (w/prob.} = \gamma_2) \\ 3 \text{ if observed in both periods (w/prob.} = \gamma_3 = 1 - \gamma_1 - \gamma_2) \end{cases} \text{, given } D = 1; \quad (2)$$

$$R = \begin{cases} 1 \text{ if observed in the } 1^{st} \text{ period for the last time (w/prob } = \phi) \\ 2 \text{ if observed in the } 2^{nd} \text{ period for the first time (w/prob} = 1 - \phi) \end{cases} \text{, given } D = 0. \quad (3)$$

Although D and R are usually predetermined as part of the sampling frame, it is useful to treat them as random variables because of practical issues such as encountering an establishment rather than a household at the address, and non-response by households. We shift the focus to individuals, so that $D = 1$ indicates that the individual is designated for the CP. For such individuals, there are 3 possibilities ($CP = 1, 2, 3$). While $CP = 3$ denotes individuals observed in both periods (i.e. those in the balanced panel), $CP = 1$ denotes attritors and $CP = 2$ denotes substitutes. If an individual is not designated for the CP ($D = 0$), then she either rotates out ($R = 1$) or rotates in ($R = 2$).

Let y and x denote random variables which are objects of the data collection effort. We distinguish between endogenous outcomes (y) and exogenous covariates (x). Some of the exogenous covariates may serve as objects of stratification. Others may identify subpopulations of interest. The primary objective of the statistical agency is to produce period-specific statistical indicators based on y, conditional on x. In what follows we use subscripts to denote period-specific values of y, and for notational simplicity treat x as time invariant. The joint distribution of interest is $f(y_1, y_2|x)$. In the typical application this is a discrete distribution which classifes individuals of a given type according to a pair of outcomes (y_1, y_2). We supress the conditioning on x for brevity, and use (1) to express the joint distribution as:

$$f(y_1, y_2) = f(y_1, y_2, D = 0) + f(y_1, y_2, D = 1). \quad (4)$$

We then use (2)-(3) and break down the components further as

$$f(y_1, y_2) = f(y_1, y_2, D = 0, R = 1) + f(y_1, y_2, D = 0, R = 2) + f(y_1, y_2, D = 1, CP = 1) \\ + f(y_1, y_2, D = 1, CP = 2) + f(y_1, y_2, D = 1, CP = 3). \tag{5}$$

We examine each of the five components in turn. We begin with the terms for individuals who are not designated for the CP. Repeated use of Bayes' Theorem yields

$$\begin{aligned} f(y_1, y_2, D = 0, R = 1) &= \Pr(R = 1|y_1, y_2, D = 0) f(y_1, y_2, D = 0) \\ &= \Pr(R = 1|y_1, y_2, D = 0) \Pr(D = 0|y_1, y_2) f(y_1, y_2) \\ &= \Pr(R = 1|D = 0) \Pr(D = 0) f(y_1, y_2) \\ &= \phi(1 - \delta) f(y_1, y_2), \end{aligned} \tag{6}$$

where we used the fact that designation of an individual for rotation, or for the CP is done independently of (y_1, y_2). Likewise,

$$f(y_1, y_2, D = 0, R = 2) = (1 - \phi)(1 - \delta) f(y_1, y_2). \tag{7}$$

Equations (6) and (7) show that the contribution to the joint distribution of interest by individuals who are not designated for the CP is a fixed fraction of that distribution. Next, we turn to the terms in (**??**) for individuals designated for the CP and establish that this is no longer the case. For attritors we have

$$\begin{aligned} f(y_1, y_2, D = 1, CP = 1) &= \Pr(CP = 1|y_1, y_2, D = 1) f(y_1, y_2, D = 1) \\ &= \Pr(CP = 1|y_1, y_2, D = 1) \Pr(D = 1|y_1, y_2) f(y_1, y_2) \\ &= \Pr(CP = 1|y_1, y_2, D = 1) \Pr(D = 1) f(y_1, y_2), \\ &= \Pr(CP = 1|y_1, y_2, D = 1) \delta f(y_1, y_2), \end{aligned} \tag{8}$$

where we used the fact that designation of an individual for the CP is done independently of (y_1, y_2). Note that in equation (8) the probability of attrition is expressed as a function of (y_1, y_2). Likewise for substitutes we have

$$f(y_1, y_2, D = 1, CP = 2) = \Pr(CP = 2|y_1, y_2, D = 1) \delta f(y_1, y_2). \tag{9}$$

In equation (9) the probability of substitution is expressed as a function of (y_1, y_2). The remaining term in (5) may be expressed as:

$$\begin{aligned} f(y_1, y_2, D = 1, CP = 3) &= f(y_1, y_2 | D = 1, CP = 3) \Pr(D = 1, CP = 3) \\ &= f(y_1, y_2 | D = 1, CP = 3) \Pr(CP = 3 | D = 1) \Pr(D = 1) \\ &= f(y_1, y_2 | D = 1, CP = 3) \gamma_3 \delta. \end{aligned} \tag{10}$$

It is straightforward to see that $f(y_1, y_2 | D = 1, BP = 3)$ can be identified non-parametrically from the balanced panel. However, since the balanced panel consists of individuals who have not been subjected to attrition or substitution, in general $f(y_1, y_2 | D = 1, BP = 3) \neq f(y_1, y_2)$.

Substitution of the terms on the right hand sides of (6)-(10) in (5) yields:

$$\begin{aligned} f(y_1, y_2) = \phi(1 - \delta) f(y_1, y_2) + (1 - \phi)(1 - \delta) f(y_1, y_2) + \Pr(CP = 1 | y_1, y_2, D = 1) \delta f(y_1, y_2) \\ + \Pr(CP = 2 | y_1, y_2, D = 1) \delta f(y_1, y_2) + f(y_1, y_2 | D = 1, BP = 3) \gamma_3 \delta. \end{aligned}$$

Upon collecting terms, simplifying and rerarranging we get

$$f(y_1, y_2) = \frac{f(y_1, y_2 | D = 1, CP = 3) \gamma_3}{[1 - \Pr(CP = 1 | y_1, y_2, D = 1) - \Pr(CP = 2 | y_1, y_2, D = 1)]}. \tag{11}$$

Finally, using the fact that $\sum_{m=1}^{3} \Pr(CP = m | y_1, y_2, D = 1) = 1$,we get

$$f(y_1, y_2) = \frac{f(y_1, y_2 | D = 1, CP = 3) \gamma_3}{\Pr(CP = 3 | y_1, y_2, D = 1)}. \tag{12}$$

The last equation is equivalent to the key equation of the AN Model of Hirano et al. (2001: 1647). Hirano et al. specify the probability in the denominator as a parametric function of (y_1, y_2) and establish the conditions under which it can be identified. Recall that the case they study involves a two period panel, and attrition impacts the second period only. In our case we have an address based rotating sample design, whereby attrition and substitution are potentially present in both periods. This poses additional challenges for the identification of $\gamma_3 = \Pr(CP = 3 | D = 1)$.We therefore treat it as a nuisance parameter. Thus our version of (12) is:

$$f(y_1, y_2) = w(y_1, y_2) f(y_1, y_2 | D = 1, CP = 3), \tag{13}$$

where $w(y_1, y_2) = \gamma_3 / \Pr(CP = 3 | y_1, y_2, D = 1) > 0$ by construction. Additional restrictions on $w(y_1, y_2)$ are needed for identification.

In our case identifying information comes from the marginal distributions which are the (properly weighted) cross-sectional statistics published by the data collection agency. Restoring the conditioning on covariates x, the equations of interest are:

$$\sum_{y_2} f(y_1, y_2|x) = \sum_{y_2} w(y_1, y_2|x) f(y_1, y_2|D = 1, CP = 3, x) = f_1(y_1|x), \tag{14}$$

$$\sum_{y_1} f(y_1, y_2|x) = \sum_{y_1} w(y_1, y_2|x) f(y_1, y_2|D = 1, CP = 3, x) = f_2(y_2|x). \tag{15}$$

Suppose y has k distinct values so that $f(y_1, y_2|x)$ can be viewed as a $k \times k$ table. Equations (14)-(15) provide the restrictions that must be satisfied by the reflated balanced panel fractions where $w(y_1, y_2)$ serve as the reflation factors. Since $\sum_{y_1} \sum_{y_2} f(y_1, y_2|x) = 1$, for $k \geq 2$ the marginals provide $k(k-1)$ pieces of independent information. Thus the reflation factors viewed as functions of (y_1, y_2) can have at most $k(k-1)$ uknown parameters. Equation (13) has a form which is familiar to survey data users. Once the function $w(y_1, y_2)$ is estimated, it can be used to inflate/deflate the cells of the balanced panel so that the object of interest $f(y_1, y_2|x)$ can be recovered. To assess the role of parameteric assumptions, we follow Chen (2001) and entertain three different specifications for this function, respectively linear, convex and concave.

It is straightforward to establish that RAN model has all the features that render the AN model attractive. Firstly, since RAN model preserves the additivity restriction of the AN model, identification proof in Hirano et al. (2001) applies.[1] Secondly, it nests the popular models of attrition. If attrition is ignorable, $w(y_1, y_2) = 1$ for all (y_1, y_2) combinations. This is the case dubbed as Missing Completely at Random (MCAR) by Little (1986). If attrition is a function of the first period outcomes only, $w(y_1, y_2) = w(y_1)$. Little and others – for example Fitzgerald et al. (1998), Hirano et al. (2001) – call this case Missing at Random because it is straightforward to adjust the balanced panel fractions using probability weights expressed as a function of observables in the first period. Note that in the present case we are dealing with substitution as well as attrition. Since substitution implies that first period outcomes are unobserved, Little's designation is not appropriate. However we use it anyway, to preserve the convention. Finally, if attrition is a function of second period outcomes only, $w(y_1, y_2) = w(y_2)$. Hirano et al. (2001) call this the Hausman and Wise (HW) model because the case was first studied by Hausman and Wise (1979).[2]

[1] For a simpler proof see Bhattacharya (2004).

[2] Moffitt, Fitzgerald and Gootschalk (1998) also study this model and use the more popular "selection on unobservables" terminology to distinguish it from the MAR case where the selection is on the first period observables.

3 Estimation and Inference in RAN model

For a given set of observed fractions $f(y_1, y_2|D = 1, CP = 3, x)$ obtained from the balanced panel, and cross-section estimates $f_1(y_1|x)$ and $f_2(y_2|x)$ obtained from official statistics, estimation boils down to solving a system of k equations in k unknowns. In our empirical work we relied on MATLAB's predefined function $fsolve(.)$ to find the solution to this system.[3] We impose a functional form for $w(.)$, get the parameter estimates $\widehat{\underline{\theta}}$, and compute the joint probabilities of interest (reflated panel) as a product of the observed fractions and the estimated reflation factors:

$$f(y_1, y_2|x) = w(\widehat{\underline{\theta}}' \underline{z}|x) f(y_1, y_2|D = 1, CP = 3, x) \tag{16}$$

For inference, we rely on standard Bootstrap methodology (Efron, 1979). Each of the random components $f(y_1, y_2|D = 1, CP = 3, x)$, $f_1(y_1|x)$ and $f_2(y_2|x)$ need to be bootstrapped. Technically speaking the joint distribution for the balanced panel is extracted from the same data set that yields the marginals. That is, all three distributions are functions of the data that have been collected during the two periods under study. These functions involve predetermined features, such as censoring due to the rotation design. They also involve the unknown attrititon/substitution process. The function that maps the cross-section data into official statistics includes the weights used by the statistical agency. Thus, a joint boostrap scheme is elusive.

In the case of HLFS, the rotation feature of the sample frame ensures that at most about half of the addresses overlap ($D = 1$ in the set-up of section 2). Consequently two distinct groups of individuals who are not designated for the complete panel also contribute to the raw marginals. Futhermore TURKSTAT manipulates the raw marginals using period specific weights based on demographic characteristics of individuals (namely age, sex and geographic location). This adjustment aims to bring the distribution of demographic attributes in line with those obtained from independent population projections. The corrected marginals are reported as official statistics.[4] With these features in mind, we propose drawing three independent bootstrap samples that have the same sample size as in the raw data. We resample from the actual balanced panel that yields $f(y_1, y_2|D = 1, CP = 3, x)$ and two artificially created marginal samples which yield the fractions $f_t(y_t|x)$ published by TURKSTAT. Using these three independent bootstrap samples, we can use

[3] MATLAB routines we used are available to the research community. In fact EXCEL's predefined function 'solve' is also capable of handling the computations.

[4] Clearly the weights do some correction for attrition and substititution, but whether this is adequate can be debated in light of the evidence in Tunalı (2009). Since this methodology is sanctioned by Eurostat, we do not question it here.

MATLAB to calculate a new $\widehat{\underline{\theta}}$. After conducting a suitable number of replications (we used 100), we can obtain bootstrap means, standard errors and estimated variance-covariance matrix for $\widehat{\underline{\theta}}$. These statistics can be used along with standard asymptotic theory to test the statistical significance of the parameters of the RAN model, and hypotheses concerning the nature of the attrition process.

Apart from choice of the functional form for $w(.)$, our procedure is fully non-parametric. We propose treating each distinct x as a separate stratum, and repeating the estimation/inference exercise. Clearly there are some practical limits to this fully non-parametric procedure; we will return to this issue below, when we briefly discuss the lessons learned from our broader empirical investigation.

4 Examples

We illustrate the utility of the RAN model by applying it to a case where y indicates labor market status and takes one of three values (0 = non-participant, 1 = employed, 2 = unemployed). In this case the system (14)-(15) yields five independent equations, so we can estimate up to 5 parameters. We express $w(y_1, y_2|x)$ as function of a linear index in (y_1, y_2) and use indicators for distinct labor market states. We take the individuals who are not in the labor force in both periods $(y_1 = 0, y_2 = 0)$ as our reference category. The other distinct categories y_t have their own parameters in each time period. For period $t(= 1, 2)$, the indicators may be defined as:

$$z_{t1} = \begin{cases} 1 & \text{if employed } (y_t = 1) \\ 0 & \text{otherwise} \end{cases} ; \tag{17}$$

$$z_{t2} = \begin{cases} 1 & \text{if unemployed } (y_t = 2) \\ 0 & \text{otherwise} \end{cases} .$$

Let $\underline{z}' = (1\ \underline{z}_1'\ \underline{z}_2') = [1\ z_{11}\ z_{12}\ z_{21}\ z_{22}]$, $\underline{\theta}' = [\theta_{00}\ \theta_{11}\ \theta_{12}\ \theta_{21}\ \theta_{22}]$, and define the linear index:

$$i(y_1, y_2) = i(\underline{\theta}'\underline{z}|x) = \theta_{00} + \theta_{11}z_{11} + \theta_{12}z_{12} + \theta_{21}z_{21} + \theta_{22}z_{22} \tag{18}$$

This function is additive in the unknown θ's which capture the dependency on the labor market

states (y_1, y_2) via $\underline{z}_1$ and $\underline{z}_2$.As in Hirano et al. (2001), we rule out interactions and focus on the main effects of the labor market states. In obtaining the reflation factors, we use three parametric forms: (i) linear: $w_1(y_1, y_2|x) = i(\underline{\theta}'\underline{z}|x)$, (ii) convex: $w_1(y_1, y_2|x) = \exp\left\{i(\underline{\theta}'\underline{z}|x)\right\}$, and (iii) concave: $w_1(y_1, y_2|x) = 2 - \exp\left\{i(\underline{\theta}'\underline{z}|x)\right\}$. Note that $w(y_1, y_2) = 1\ iff\ \theta_{00} = 1, \theta_{11} = \theta_{12} = \theta_{21} = \theta_{22} = 0$ in the linear case. In the nonlinear cases, $w(y_1, y_2) = 1\ iff\ \theta_{00} = \theta_{11} = \theta_{12} = \theta_{21} = \theta_{22} = 0$.

For the linear case, the restrictions implied by (14)-(15) can be represented as in Table 1, where

we set $p_{y_1 y_2} = f(y_1, y_2|D = 1, CP = 3, x)$ for brevity. To recapitulate, the task amounts to finding the reflation factors (functions of θ's) which would bring the adjusted cell probabilites in line with the marginals reported by the data collection agency.

Table 1. A 3x3 Linear RAN Model

	$y_2 = 0$	$y_2 = 1$	$y_2 = 2$	Row sum
$y_1 = 0$	$\theta_{00}p_{00}$	$(\theta_{00} + \theta_{21})p_{01}$	$(\theta_{00} + \theta_{22})p_{02}$	$f_1(0)$
$y_1 = 1$	$(\theta_{00} + \theta_{11})p_{10}$	$(\theta_{00} + \theta_{11} + \theta_{21})p_{11}$	$(\theta_{00} + \theta_{11} + \theta_{22})p_{11}$	$f_1(1)$
$y_1 = 2$	$(\theta_{00} + \theta_{12})p_{21}$	$(\theta_{00} + \theta_{12} + \theta_{21})p_{11}$	$(\theta_{00} + \theta_{12} + \theta_{22})p_{11}$	$f_1(2)$
Col. sum	$f_2(0)$	$f_2(1)$	$f_2(2)$	1

For the linear case, the system of equations has the observationally equivalent representation given below:

$$\begin{bmatrix} \sum_{j=0}^{2} p_{0j} & 0 & 0 & p_{01} & p_{02} \\ \sum_{j=0}^{2} p_{1j} & \sum_{j=0}^{2} p_{1j} & 0 & p_{11} & p_{12} \\ \sum_{j=0}^{2} p_{2j} & 0 & \sum_{j=0}^{2} p_{2j} & p_{21} & p_{22} \\ \sum_{k=0}^{2} p_{k0} & p_{10} & p_{20} & 0 & 0 \\ \sum_{k=0}^{2} p_{k1} & p_{11} & p_{21} & \sum_{k=0}^{2} p_{k1} & 0 \\ \sum_{k=0}^{2} p_{k2} & p_{12} & p_{22} & 0 & \sum_{k=0}^{2} p_{k2} \end{bmatrix} \begin{bmatrix} \theta_{00} \\ \theta_{11} \\ \theta_{12} \\ \theta_{21} \\ \theta_{22} \end{bmatrix} = \begin{bmatrix} f_1(0) \\ f_1(1) \\ f_1(2) \\ f_2(0) \\ f_2(1) \\ f_2(2) \end{bmatrix} \tag{19}$$

Inspection reveals that the system is of the form $A\underline{\theta} = \underline{b}$ where A is of rank = 5. One of the constraints is redundant, in the sense that it will be automatically met once the solution to the reduced system is found. We prove this in the appendix by starting with a particular system of five equations in five unknowns, and showing that any other representation can be transformed to

the one we started with by a simple pivoting operation. Consequently, the solution to the reduced system is unique, and does not depend on which constraint is left out.

Table 2. A 3x3 RAN Model – Parameter Estimates

Annual Transitions between 2001-Q1 and 2002-Q1

x = age 15 and over

Parameter	θ_{00}	θ_{11}	θ_{12}	θ_{21}	θ_{22}
(i) $w(.)$ linear:					
Estimate	0.8987	0.0956	0.2524	0.1315	0.1779
Bootstrap mean	0.8994	0.0999	0.2423	0.1263	0.1755
Bootstrap std. error	0.0063	0.0282	0.0509	0.0290	0.0507
(ii) $w(.)$ convex:					
Estimate	−.1057	0.0957	0.2306	0.1293	0.1703
Bootstrap mean	−.1050	0.0999	0.2221	0.1243	0.1672
Bootstrap std. error	0.0070	0.0283	0.0440	0.0288	0.0462
(iii) $w(.)$ concave:					
Estimate	−.0975	0.0960	0.2848	0.1349	0.1885
Bootstrap mean	−.0968	0.1007	0.2725	0.1295	0.1875
Bootstrap std. error	0.0060	0.0298	0.0656	0.0308	0.0602
Sample sizes:					
Balanced panel			21,731		
First period cross-section			52,389		
Second period cross-section			53,810		

Data Source: Household Labor Force Survey, TURKSTAT.

In Table 2 we compiled a set of parameter estimates from a 3x3 RAN model for annual transitions on data from the Household Labor Force Survey (HLFS) in Turkey, together with bootstrap means and standard errors based on 100 replications. In this case x denotes the entire working age population, ages 15 and over. The balanced panel contained over 20 thousand observations. The first and second period marginals in the raw data contained over 52 thousand observations. Thus it is not surprising that all RAN model parameters are estimated extremely precisely.

As we noted earlier, HLFS sample frame ensures that about half of the addresses visited in a given period are also visited the next period. Taking the sample sizes we reported above, we see

that the balanced panel sample amounted to about 0.40% of the respective marginals. The fact that this fraction is considerably lower than 0.5 can be taken as a rough statistic that warns us about the potential severity of the attrition/substituition problem.[5] What matters, of course, is whether the process that excludes individuals designated for the complete panel from the balanced panel is ignorable. Given the evidence from the bootstrap exercise, we not expect this to be the case. In fact Wald tests provide overwhelming evidence that the attrition and substituition process is non-ignorable. Furthermore, alternatives to RAN model are deemed inadequate for capturing the selectivity (all $p-$values are practically zero). The key insight from labor economics, that attrition and substituition behavior is intimately connected with labor market behavior, is vindicated.

Table 3. A 3x3 RAN Model – Reflation Factors

Annual Transitions between 2001-Q1 and 2002-Q1

x = age 15 and over

	$y_2 = 0$	$y_2 = 1$	$y_2 = 2$	Row sum
$y_1 = 0$	$\left\{\begin{array}{l}0.8986\\0.8997\\0.8976\end{array}\right\}$ 0.5052	$\left\{\begin{array}{l}1.0302\\1.0238\\1.0366\end{array}\right\}$ 0.0566	$\left\{\begin{array}{l}1.0766\\1.0667\\1.0870\end{array}\right\}$ 0.0159	$f_1(0)$
$y_1 = 1$	$\left\{\begin{array}{l}0.9943\\0.9901\\0.9985\end{array}\right\}$ 0.0740	$\left\{\begin{array}{l}1.1258\\1.1267\\1.1248\end{array}\right\}$ 0.2952	$\left\{\begin{array}{l}1.1722\\1.1739\\1.1706\end{array}\right\}$ 0.0209	$f_1(1)$
$y_1 = 2$	$\left\{\begin{array}{l}1.1511\\1.1330\\1.1708\end{array}\right\}$ 0.0113	$\left\{\begin{array}{l}1.2826\\1.2894\\1.2754\end{array}\right\}$ 0.0122	$\left\{\begin{array}{l}1.3290\\1.3433\\1.3133\end{array}\right\}$ 0.0085	$f_1(2)$
Col. sum	$f_2(0)$	$f_2(1)$	$f_2(2)$	1

In Table 3 we compiled the set of reflation factor estimates we obtained from the RAN model parameter estimates reported in Table 2. For brevity we excluded the numbers for the margins. The numbers reported in each cell are of the form given on the right hand side of equation 16. For each cell we report the estimates of the reflation factors $w(.)$ associated with all three functional forms (respectively linear, convex, concave; shown inside braces) followed by the fraction obtained from the balanced panel. Reflation factors below (above) one mark labor market

[5] The realized magnitudes of attrition and substituition in the HLFS over the period 2000-2002 are reported in Tunalı (2009).

states which are overrepresented (underrepresented) in the balanced panel. Note that for some states the bias induced by attrition/substituition is practically zero [see $(y_1 = 1, y_2 = 0)$] but for others it is substantial [e.g. $(y_1 = 2, y_2 = 2)$]. The findings from our sensitivity analysis are typical, in that functional form does not make much of a difference.In Table 4 we report the unadjusted joint probabilities and marginals obtained from the balanced panel (shown in brackets) along with the adjusted versions obtained from the linear RAN model. The magnitudes of the biases in the balanced panel [discrepencies between $f(y_1, y_2|D = 1, CP = 3, x)$ and $f(y_1, y_2|x)$] range between -24% and 11%. Six of the 9 cells have biases of 10% or more in absolute value.

Table 4. A 3x3 RAN Model –
Adjusted and [Unadjusted] Joint and Marginal Probabilities
Annual Transitions between 2001-Q1 and 2002-Q1
x = age 15 and over, $w(.)$ linear

	$y_2 = 0$	$y_2 = 1$	$y_2 = 2$	Row sum
$y_1 = 0$	0.4540 [0.5052]	0.0584 [0.0566]	0.0172 [0.0160]	0.5296 [0.5778]
$y_1 = 1$	0.0736 [0.0740]	0.3323 [0.2952]	0.0245 [0.0209]	0.4305 [0.3902]
$y_1 = 2$	0.0130 [0.0113]	0.0156 [0.0122]	0.0113 [0.0085]	0.0399 [0.0320]
Col. sum	0.5406 [0.5905]	0.4063 [0.3640]	0.0530 [0.0454]	1

In Table 5 the associated forward transition probabilites are shown. As in the previous table, the numbers in brackets are the unadjusted ones. Almost surely someone who views the evidence will argue that the differences between unadjusted and adjusted magnitudes are not large enough to warrant correction. It is worth noting that even though the picture of labor dynamics that emerges might not be different by some measure of closeness, the correction is still warranted because it produces a version which is fully consistent with the cross-section estimates. This capabilitiy of RAN model is especially important in the case of statistical agencies like TURKSTAT, who refuse to exploit the short panel dimension of the HLFS on the grounds that there is no weighting method that can reconcile dynamic and static estimates.

Table 5. A 3x3 RAN Model – Adjusted and [Unadjusted] Transition Probabilities

Annual Forward Transition Matrix between 2001-Q1 and 2002-Q1

x = age 15 and over, $w(.)$ linear

	$y_2 = 0$	$y_2 = 1$	$y_2 = 2$	Row sum
$y_1 = 0$	0.8573	0.1102	0.0325	1
	[0.8744]	[0.0980]	[0.0276]	[1]
$y_1 = 1$	0.1710	0.7720	0.0570	1
	[0.1898]	[0.7565]	[0.0537]	[1]
$y_1 = 2$	0.3250	0.3917	0.2833	1
	[0.3525]	[0.3813]	[0.2662]	[1]

As we argued above, the non-parametric feature of RAN model is attractive, but it has the usual shortcomings that data based methods have. To illustrate the possible pitfalls, we consider another example, where x denotes males aged 35-54 who have high school education and reside in urban areas of Turkey. RAN model estimates for this partititon of the sample are reported in Table 6. In this case the statistical evidence favors the hypothesis that attrition/substitution is ignorable. Note that the sample sizes are small, and consequently bootstrapped standard errors are large. In fact in some cases the bootrapped means are very different from the estimated parameter value (see θ_{21} and θ_{22} for the concave case). This finding exposes the fragility of the bootstrap method, and serves as a call for caution when sample sizes are too small.

Table 6. Another 3x3 RAN Model – Parameter Estimates

Annual Transitions between 2001-Q1 and 2002-Q1

x = male, ages 35-54, high school education, residing in urban areas

Parameter	θ_{00}	θ_{11}	θ_{12}	θ_{21}	θ_{22}
(i) $w(.)$ linear:					
Estimate	0.9472	−.0234	0.1348	0.0688	0.3507
Bootstrap mean	0.9465	−.0003	0.2731	0.0524	0.3638
Bootstrap std. error	0.1271	0.2530	0.5869	0.2220	0.4227
(ii) $w(.)$ convex:					
Estimate	−.0540	−.0231	0.1191	0.0697	0.3127
Bootstrap mean	−.0699	0.0028	0.1813	0.0646	0.2934
Bootstrap std. error	0.1345	0.2620	0.4179	0.2281	0.3471
(iii) $w(.)$ concave:					
Estimate	−.0518	−.0239	0.1560	0.0681	0.4101
Bootstrap mean	−.0378	−.0037	2.6577	0.0415	2.6068
Bootstrap std. error	0.1340	0.2570	9.0427	0.2262	8.5999
Sample sizes:					
Balanced panel			460		
First period cross-section			1,416		
Second period cross-section			1,440		

Data Source: Household Labor Force Survey, TURKSTAT.

If the objective is to produce dynamic statistics consistent with the cross-section statistics, the correction can proceed despite our cautionary remark. In fact, the reflation factors for the subsample under examination reported in Table 7 point to a surprisingly consistent picture regardless of choice of functional form. Interestingly, small cell sizes that produced the fragility in the bootstrap stage rescues the reflation stage: when p_{jk} is small ($< .01$), the differences by functional form reflected in the second digit after the decimal point do not translate to comparable differences in the magnitudes of the adjusted fraction.

Table 7. Another 3x3 RAN Model – Reflation Factors

Annual Transitions between 2001-Q1 and 2002-Q1

x = male, ages 35-54, high school education, residing in urban areas

	$y_2 = 0$	$y_2 = 1$	$y_2 = 2$	Row sum
$y_1 = 0$	$\left\{\begin{matrix} 0.9471 \\ 0.9474 \\ 0.9468 \end{matrix}\right\}$ 0.0978	$\left\{\begin{matrix} 1.0160 \\ 1.0158 \\ 1.0162 \end{matrix}\right\}$ 0.0196	$\left\{\begin{matrix} 1.2978 \\ 1.2952 \\ 1.3011 \end{matrix}\right\}$ 0.0087	$f_1(0)$
$y_1 = 1$	$\left\{\begin{matrix} 0.9237 \\ 0.9258 \\ 0.9214 \end{matrix}\right\}$ 0.0652	$\left\{\begin{matrix} 0.9926 \\ 0.9927 \\ 0.9924 \end{matrix}\right\}$ 0.7543	$\left\{\begin{matrix} 1.2744 \\ 1.2657 \\ 1.2843 \end{matrix}\right\}$ 0.0239	$f_1(1)$
$y_1 = 2$	$\left\{\begin{matrix} 1.0819 \\ 1.0672 \\ 1.0989 \end{matrix}\right\}$ 0.0109	$\left\{\begin{matrix} 1.1508 \\ 1.1443 \\ 1.1583 \end{matrix}\right\}$ 0.0109	$\left\{\begin{matrix} 1.4326 \\ 1.4591 \\ 1.4021 \end{matrix}\right\}$ 0.0087	$f_1(2)$
Col. sum	$f_2(0)$	$f_2(1)$	$f_2(2)$	1

5 Findings from a Broader Investigation

As can be inferred from our second example, in our broader empirical investigation we exposed the parametric features of RAN model to a torture test by choosing x to identify smaller and smaller segments of the population. This exercise is warranted, because statistical agencies often publish official statistics broken down by a high dimensional x.The question is whether RAN model can rise to the challenge of yielding their dynamic counterparts.

The covariates we studied included sex (male, female), location (urban, rural), education (4 categories) and age (5 groups). Notably RAN model yielded extremely robust results as long as cell counts in the balanced panel remained within acceptable ranges for the sample sizes under investigation. In extreme cases when cell sizes were extremely small, we ran into occasional convergence problems during the bootstrap stage. This problem was attributable to the fact that some bootrapped samples yielded zero cell counts, in which case correction could not proceeed. Clearly zeroes encountered during bootstrapping are random as opposed to structural zeroes. We were able to fix the problem by adding an observation to the empty cell and adjusting the sample size accordingly. However, in light of the fragility exposed in Table 6, this computational fix should not be seen as a panacea.

Empirical findings regarding the nature of attrition/substitution can, and do vary, from one time

period to the other, and with choice of x. RAN model is useful for shedding light on the patterns. There are valid reasons for proceding with the correction whether or not attrition/substitution is ignorable. Overall our non-parametric approach with respect to x worked extremely well. In our systematic examination of annual and quarterly transitions over the 2000-2002 period, we discovered that RAN model produced very reasonable estimates of transition rates for commonly used partitions of the full sample (jointly by sex and location, by education, by broad age groups). Even further partioning of the subsamples identified by sex-location pairs either by education, or by broad age groups, yielded reasonable results, worthy of adoption for statistical and policy purposes.

6 Conclusion

In this paper we tackle a generalized version of the attirition problem, typically associated with longitudinal data. The motivaton for the generalization comes from the observation that many sustained large scale data collection efforts (CPS, SILC being some well-known examples) involve multiple visits to the same address/household over a short period of time and therefore offer a longitudinal component. Another feature of these efforts is the use a rotational design whereby a fresh set of addresses/households are systematically added to, and excluded from, the sample frame according to a predetermined schedule. These data sets have a short panel component that can support dynamic analyses. What stands in the way is the concern that the balanced panel which can be used for tracking the dynamics may not be representative of the population at a given point of time. The generalization we offer recognizes that proper use of such short panels requires corrections for non-response after initial response (attrititon) as well as response after initial non-response (substitution). Furthermore, attrition/substitution behavior is allowed to be endogenous to the outcomes of interest.

In our empirical example outcomes are labor market states occupied by an individual. Endogeneity implies that particular outcome combinations could make individuals more or less prone to exclusion from the balanced panel. The model we use exploits the set-up and insights in Hirano et al. (2001) but departs from it in its computational simplicity, especially when the linear version is adopted. The correction amounts to reflating the balanced panel fractions (cell means) by factors expressed as a parametric function of the states under examination. Our empirical investigation of annual transition data from the Household Labor Force Survey in Turkey showed that attrition/substitution is a serious concern when the full working age population is brought under focus.

The exercise demonstrated the superiority of the RAN model over popular models employed by researchers.

Based on our systematic empirical investigation, results did not display sensitivity to the parametric features of RAN model. Thus the linear version – which is extremely simple to implement – appears suitable for empirical work. Another attractive feature of RAN model is the non-parametric treatment of covariates (such as sex, location, age groups, etc.). That is, each distinct covariate combination is associated with its own set of parameters and reflation factors. Finally, since RAN model produces dynamic estimates which are consistent with cross-section statistics, it is likely to gain the approval of official statistical agencies.

References

Bhattacharya, D. (2004) "Semiparametric Inference in Panel data Models under Attrititon Caused by Unobservables." *Mimeo*, Department of Economics, Darthmouth College.

BLS (Bureau of Labor Statistics) (2002) *Design and Methodology: Current Population Survey.* Technical Paper 63 RV, U.S. Department of Labor and U.S. Department of Commerce.

Chen, K. (2001) "Parametric Models for Response-Biased Sampling." *Journal of Royal Statistical Society, B*, v. 63, Part 4, 775-789.

Efron, B. (1979) "Bootstrap Methods: Another Look at the Jackknife." *Annals of Statistics,* 7: 1, 1-26.

Fitzgerald, J., P. Gottschalk, ve R. Moffitt (1998) "An Analysis of Sample Attrition in Panel Data." *Journal of Human Resources*, 33:2, 251–299.

Hausman, J. A. ve D. A. Wise (1979) "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment." *Econometrica*, 47:2, 455–473.

Heckman, J. (1987) "Selection Bias and Self-selection." In J. Eatwell, M. Milgate, ve P. Newman (Ed.), The New Palgrave: A Dictionary of Economics, Vol. IV. London: McMillan.

Hirano, K., G. W. Imbens, G. Ridder, and D. B. Rubin (2001) "Combining Panel Data Sets with Attrition and Refreshment Samples." *Econometrica*, 69: 6, 1645-1660.

Little, R. ve D. Rubin (1987) *Statistical Analysis with Missing Data.* New York: Wiley.

Madow, W., I. Olkin, ve D. Rubin (Eds.) (1983) *Incomplete Data in Sample Surveys* (3 volumes). New York: Academic Press.

Rubin, D. (1976) "Inference and Missing Data." *Biometrika*, 63: 581–592.

Tunalı, I. (2009) "Analysis of Attrition Patterns in the Turkish Household Labor Force Survey, 2000-2002." Ch. 6 in Labor Markets and Economic Development, edited by R. Kanbur and J. Svejnar, 110-136. London and New York: Routledge.

TURKSTAT (Turkish Statistical Institute) (2001) *Household Labor Force Survey: Concepts and Methods.* Ankara: State Institute of Statistics.

Appendix

Let A_j denote the 5x5 partition of the A matrix defined implicitly by 19 with the jth row removed, and let $\underline{b}_j$ denote the 5x1 partition of vector $\underline{b}$ with the jth row removed, $j = 1, 2, ..6$. With this notation, the system with the 6th equation removed can be expressed as $A_6\underline{\theta} = \underline{b}_6$ and has the explicit form given below:

$$\begin{bmatrix} \sum_{j=0}^{2} p_{0j} & 0 & 0 & p_{01} & p_{02} \\ \sum_{j=0}^{2} p_{1j} & \sum_{j=0}^{2} p_{1j} & 0 & p_{11} & p_{12} \\ \sum_{j=0}^{2} p_{2j} & 0 & \sum_{j=0}^{2} p_{2j} & p_{21} & p_{22} \\ \sum_{k=0}^{2} p_{k0} & p_{10} & p_{20} & 0 & 0 \\ \sum_{k=0}^{2} p_{k1} & p_{11} & p_{21} & \sum_{k=0}^{2} p_{k1} & 0 \end{bmatrix} \begin{bmatrix} \theta_{00} \\ \theta_{11} \\ \theta_{12} \\ \theta_{21} \\ \theta_{22} \end{bmatrix} = \begin{bmatrix} f_1(0) \\ f_1(1) \\ f_1(2) \\ f_2(0) \\ f_2(1) \end{bmatrix}.$$

The solution to this system is unique and is given by $\widehat{\underline{\theta}} = A_6^{-1}\underline{b}_6$. Next, we define the following 5x5 pivot matrices:

$$E_1 = \begin{bmatrix} -1 & -1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, E_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -1 & -1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$E_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ -1 & -1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, E_4 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$E_5 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & -1 & -1 \end{bmatrix}.$$

It is straightforward to show that for $j = 1, 2, .., 5$, $E_j A_j = A_6$,and $E_j\underline{b}_j = \underline{b}_6$.Since the pivot matrices are of full rank, this proves that all six sytems are equivalent, and yield the same unique solution $\widehat{\underline{\theta}}$.