RESCALED ADDITIVELY NON-IGNORABLE (RAN) MODEL OF GENERALIZED ATTRITION*

İnsan Tunalı†, Emre Ekinci‡, Berk Yavuzoğlu§

Revised (under a new title), June 2016

*This is a revised version of an earlier paper that circulated under different titles ("Rescaled Additively Non-Ignorable Model of Attrition: A Convenient Semi-Parametric Bias Correction Framework for Data with a Short Panel Component" and "Rescaled Additively Non-Ignorable Model of Attrition and Substitution"). Tunali 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 Huseyin Ikizler, Bengi Ilhan Yanik and Hayriye Ozgul Ozkan for research assistance. An earlier version was presented as a Keynote Lecture by Tunali during the 12th International Symposium on Econometrics, Operations Research and Statistics held at Pamukkale University, Denizli, 26-29 May 2011. Comments from seminar and workshop participants at Bilkent, Koç and LSE, Thierry Magnac, Christopher Taber, James Walker, and detailed feedback from John Kennan are gratefully acknowledged.

†Corresponding author, Department of Economics, Koç University, İstanbul and Economic Research Forum, Cairo; phone: +90-212-3381425; fax: +90-212-338-1653; e-mail: itunali@ku.edu.tr

‡Department of Business Administration, Universidad Carlos III de Madrid: e-mail: emre.ekinci@uc3m.es

§Department of Economics, Nazarbayev University: e-mail: berk.yavuzoglu@nu.edu.kz

Abstract

We augment the Additively Non-ignorable (AN) model of Hirano et. al. (2001) so that it is suitable for data collection efforts that have a short panel component. Our modification yields a convenient semi-parametric bias correction framework for handling selective non-response that can emerge when multiple visits to the same unit are planned. In such surveys, selective non-response can be due to attrition, when initial response is followed by nonresponse (the commonly studied case), as well as a phenomenon we term "reverse attrition", when initial nonresponse is followed by response. Accounting for reverse attrition creates an additional identification problem, which we circumvent by rescaling. We apply our methodology to data from the Household Labor Force Survey (HLFS) in Turkey, which shares a key design feature (namely a rotating sample frame) of popular surveys such as the Current Population Survey and the European Union Labor Force Survey. The correction amounts to adjusting the observed joint distribution over the state space (non-participant, employed, unemployed in our example) using reflation factors expressed as parametric functions of the states occupied in the initial and subsequent rounds. Our method produces a unique set of corrected joint probabilites that are consistent with unbiased marginal distributions (in our case published cross-section statistics). The linear additive version has a closed form solution, a feature which renders our method computationally attractive. Our empirical results show that selective attrition/reverse attrition in HLFS-Turkey is a statistically and substantially important concern.

Keywords: generalized attrition; selective nonresponse; selection on observables; selection on unobservables; short panel; rotating sample frame; rotating panel; address-based sampling; labor force survey; M-estimation.

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 attrition problem is intimately related to the class of problems collected under the title of selectivity (Heckman, 1987). Arguably the simplest diagnosis of the problem at hand is provided by Ridder and Moffit (2007), who define a sample in which the probability of observation depends on the outcome variable(s) of interest as a "biased sample" (p.5525). The preoccupation with attrition has a long history among survey researchers (Madow et al., 1983). Formalizations by Rubin (1976), and Little (1982) (collected in Little and Rubin, 1987) have paved the way for establishing common terminology such as missing completely 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). Fitzgerald et al. (1998) situated these important ideas within a modelling framework familiar to economists, by distinguishing between selection on observables (first round outcomes) and selection on unobservables (such as second period outcomes).

Our paper builds on an important contribution by Hirano, Imbens, Ridder and Rubin (2001), henceforth HIRR. They approach the challenge posed by attrition as an identification problem that amounts to recovering the joint distribution of interest for the full population, when the balanced panel suffers from potentially non-ignorable attrition. When the attrition probability is expressed as an additive function of the potentially endogenous outcomes before and after attrition, HIRR 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. HIRR exploit an independently conducted cross-section survey (what has been termed refreshment sample by Ridder, 1992) to provide an unbiased estimator for the second round marginal distribution. Adjustment of the balanced sample proceeds by using the inverted attrition (selection) probabilities as weights. Equating the row and column sums of the reweighted balanced panel cell counts (or 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, HIRR name this model Additively Non-ignorable (AN) model of attrition. They show that two earlier and popular formulations by Little-Rubin and Hausman-Wise are nested within the AN model. Thus the AN model not only offers a theoretically appealing correction for attrition, but it also affords tests of widely used models which have behavioral implications.

In this paper we establish that the key ideas embedded in the AN model can be used for correction of a broader class of non-response problems. In particular we modify the AN model so that it is suitable for data collection efforts that have a short panel component. Household surveys that have this feature – such as the Household Labor Force Surveys (HLFS) in Turkey we use below, as well as popular data sets such as the Current Population Survey (CPS) in the U.S., and most country surveys included in the European Union Labor Force Survey (EULFS) – call for repeat visits to the same household according to a pre-determined schedule but limit the maximum number of visits. The schedule is supported by a rotatating sample frame that ensures nationwide representation as well as regular updating. At each round a predetermined set of households visited for the last time are dropped and replaced by a set of new households. Since the data collection agency typically provides the weights needed for rendering each cross section nationally representative, this amounts to having unbiased marginals for both periods, as assumed in HIRR. However, as we show below, a parameter of the AN model becomes unidentified.

Surveys that rely on a rotational design typically have an address- or dwelling-based sample frame. In some cases a longitudinal view is adopted, so that households (or individuals) that enter the sample frame are followed even when they leave the original address (such as the CPS, see BLS, 2002). In other cases the data collection agency prefers to treat each round of the data as an independent cross-section (such as some country components of the EULFS, see EUROSTAT, 2007). The data set we work with, HLFS-Turkey, is a typical example of the latter (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 were found in the previous visit. Standard cross-section 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 (so-called balanced panel) are surveyed in two adjoining periods, such surveys also lend themselves for dynamic analyses. However finding suitable weights for rendering the balanced panel

representative is a challenge. Reconciliation of the joint distribution estimated from the balanced panel with the period-specific marginals is another challenge.

The problem is attributable to the fact that such data not only suffer from attrition (response followed by non-response) but also from what we term "reverse attrition" (non-response followed by response).[1] Reverse attrition occurs because visits to each address continue according to a predetermined schedule, and this sets the stage for the prospect of encountering new units (in place of the old ones), or new individuals in old (previousy visited) units. When both types of attrition are present, the conventional distinction between selection on observables and unobservables drawn in Fitzgerald et al. (1998) and invoked in HIRR disappears. Furthermore the difference between the first and the second period sample sizes does not contain useful information about the attrition process. Consequently, a parameter of the AN model of HIRR (namely the unconditional probability of retention in the panel) is not identified. However, a correction scheme which preserves the key ideas in HIRR can still be found. Since this amounts to treating the unidentified probability as a nuisance parameter and rescaling the weights used for the purposes of adjustment, we term the new model *Rescaled* Additively Non-ignorable (RAN) model of attrition. As in HIRR, we are able to test whether conventional correction schemes for attrition are valid.

Notably our weighting scheme produces dynamic estimates consistent with the cross-sectional statistics. We illustrate this in the context of a labor market application. Our objective is estimation of transition rates between labor market states that are consistet with the official cross-section statistics. We show that the parameters of the RAN model can be estimated and inference can be carried out using standard methods.[2] Furthermore the data requirements for the implementation of our methodology are extremely minimal: namely, the joint frequency distribution obtained from the balanced panel, and the marginal frequency distributions obtained from an external data source that does not have the representation problems. The findings establish that both attrition and

[1] In earlier versions we used the term "substitution" in place of what we now term "reverse attrition". Subsequently we realized that the literature used the term substitution to decribe use of other (suitable) individuals in place of those who could not be observed. An example of this approach is Dorsett (2010). He exploits administrative (unemployment registry) data and uses Propensity Score Matching to find substitutes for the attrited individuals in the second period. The matching is done on first period observables. Since our approach is very different, we have decided to change our terminology.

[2] The current version of our paper differs from Tunalı, Ekinci and Yavuzoğlu (2012) in two major ways. In the earlier paper we estimated the parameters by solving the system of (possibly non-linear equations) given by equations (19)-(29) numerically. Since we did not have a standard estimation frawrok, we relied on a bootstrapped covariance matrix for inference.

reverse attrition are non-ignorable, a result that substantiates the utility of our methodology in analyses of labor market dynamics.

The idea of reconciling observed flow data between states with the cross sectional stocks via probabilistic adjustments expressed as a function of the states predates HIRR. Abowd and Zellner (1985) and Stasny (1986, 1988) work with counts obtained from short panels, and focus on adjusting the flow data so that they are consistent with the properly weighted margins that represent the target population. The contrasts between their approaches and RAN model will be taken up below. Although it is cast in an entirely different setting, the adjustment methodology discussed by Golan *et al.* (1994) closely resembles our approach under the linear parameterization of the weighting functions. In fact all these approaches can be situated within a broader statistical framework directed at reconciling key statistical features of incomplete survey data with what is known about the population (Little, 1993). Notably the model based adjustment of Little and Wu (1991) echoes the fundamental ideas exploited in AN and RAN models.

We begin our formal treatment in Section 2 by introducing our conceptualization of the attrition process and derive the RAN model. We then establish its links with the AN model. In Section 3 we discuss our semi-parametric estimation and inference methodology in the context of a three-state labor market transition study. We then relate our approach to others developed in the statistics literature. Section 4 contains examples that illustrate the utility and potential limitations of the proposed approach. Section 5 offers a short compilation of the lessons learned from a broader investigation. We conclude the paper with a brief summary of the key aspects of our model and potential uses. The Appendix illustrates the uniqueness of the solution to the linear parameterization of the 3x3 RAN model used in our labor market application.

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. Several advantages are apparent: Firstly, a repeated visit to the same household ushers in some savings in the data collection effort and allows tracking dynamics. Secondly, by limiting the number of revisits, a better balance between the cost of the data collection effort and the response burden imposed on

the households can be achieved. Thirdly, by including a fresh subsample every period, the sample is kept up to date. As a result rotating panel designs have emerged as a useful compromise between longitudinal and repeated cross-section designs. However use of the short panel component ushers in new challenges when drawing inferences about the population. In fact it is often not fully 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 successfully interviewed in both periods. We also keep track of households which are rotated out of the sample after period 1, and households which are rotated in during period 2. Finally, for the sake of completeness we allow for "nonparticipants," units who have been selected for inclusion in the sample frame, but who never participate in the survey.

Let y and x denote random variables which are the main 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 (by location, for example). Others may identify subpopulations of interest (sex, age, education, etc.). 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 treat x as time invariant. The joint distribution of interest is $f(y_1, y_2|x)$. The endogenous outcome variables may be continuous, or discrete. In the latter case the distribution classifies individuals of a given type (x) according to a pair of discrete outcomes (y_1, y_2).

Assumptions:

Our first task is statement of the assumptions that RAN model rests on. Towards that end we introduce several random variables to keep track of the observation status of the unit within the interval under study. Some of these are predetermined in the sense that they are known before the survey reaches the field. Nonetheless we treat them as random variables, associate parameters with the outcomes and state the independence assumptions that enable us to examine the impact of selective nonreponse formally. The first random variable captures the assignment status of the

address to the complete panel (CP):

$$D_1 = \begin{cases} 1 \text{ if designated for the Complete Panel (w/prob. } = \delta_1) \\ 0 \text{ if not (w/prob.} = 1 - \delta_1) \end{cases} . \tag{1}$$

Assumption 1: $D_1 \perp (y_1, y_2)$.

This assumption is non-controversial: Since assignment status is predetermined, it is exogenous (indepent of/orthogonal to the endogenous outcomes).

The second random variable captures whether an intended interview took place during the observation window:

$$D_2 = \begin{cases} 1 \text{ if at least one interview took place (w/prob. } = \delta_2) \\ 0 \text{ if not (w/prob.} = 1 - \delta_2) \end{cases} . \tag{2}$$

The outcome of random variable $D_1$is observed before the survey is fielded. By contrast, D_2 is revealed when the survey reaches the field. Random variable D_2 keeps track of practical survey implementation problems. These typically include (i) encountering the wrong unit (for example, an establishment rather than a household) at the address, (ii) inability to contact the unit in any round, and (iii) refusal of participation in the survey by the unit.[3]

.*Assumption 2:* $D_2 \perp D_1$.

By ruling out dependence between D_2 and assignment status D_1, we disallow selective participation in the survey because of the differential burden involved.

Assumption 3: $D_2 \perp (y_1, y_2)$.

Assumption 3 rules out selective nonparticipation in the survey. Stated explicity, it enables us to draw the important distinction between *selective non-response that we model* (by virtue of having seen the unit at least once), and *non-response we do not model* (because we have never seen the unit). Reevaluation of the reasons listed above suggests that only (ii) and (iii) pose a potential threat to our objective of identifying is $f(y_1, y_2|x)$.[4] As we show below, Assumption 3 does not rule

[3]Based on information obtained from the data collection agency, non-response due to refusal to participate is uncommon in HLFS-Turkey. This is attributable to the fact that the Law obliges participation in official surveys. Non-response due to (i) and (ii) are more common. If (i) occurs during the initial visit, the address is simply dropped from the sample frame. When either (ii) or (iii) occur, a non-response form is filled. The most frequently recorded reason for non-response is "the household no longer resides at this address."

[4]Since we never observe the characteristics of nonparticipants, it would seem that the best we can do is to assume

out non-ignorable attrition in the balanced panel providing an interview took place ($D_2 = 1$).

While D_1 is an *ex ante* construct that captures intended use of a unit in the sample frame, D_2 is an *ex post* construct that indicates actual participation in the data collection effort. Clearly only units with $D_1 = D_2 = 1$ have the potential to contribute to the identification of the joint distribution, $f(y_1, y_2|x)$. Occurence of the phenomena that are of primary interest – attrition and reverse attrition – are revealed after both visits to the address are completed, according to the *ex post* construct:

$$C = \begin{cases} 1 \text{ if observed in the } 1^{st} \text{ period only (attrited w/prob.} = \gamma_1) \\ 2 \text{ if observed in the } 2^{nd} \text{ period only (reverse attrited w/prob.} = \gamma_2) \\ 3 \text{ if observed in both periods (w/prob.} = \gamma_3 = 1 - \gamma_1 - \gamma_2) \end{cases} \text{, given } D_1 = D_2 = 1. \tag{3}$$

Random variable C captures the possibly selective response status of participants among the $D_1 = D_2 = 1$ group, during the observation window. The subsample with $C = 3$ constitutes the balanced panel (BP). By virtue of being present either in the first or the second period, those with $C = 1$ (attrited unit) or $C = 2$ (reverse attrited unit) make contributions to the marginal distribution of that period. Additional contributions to the marginals come from partipants not designated for the complete panel ($D_1 = 0$) with whom the intended interview took place ($D_2 = 1$) . We classify these using the *ex post* construct:

$$R = \begin{cases} 1 \text{ if observed (in the } 1^{st} \text{ period) for the last time (w/prob } = \phi) \\ 2 \text{ if observed (in the } 2^{nd} \text{ period) for the first time (w/prob} = 1 - \phi) \end{cases} \text{, given } D_1 = 0, D_2 = 1. \tag{4}$$

Random variable R reflects the unit's rotation designation. Technically rotation designation is an *ex ante* construct (like D_1). By defining R conditional on $D_2 = 1$ we turn it into an *ex post* construct (like C). That is, we take into account the fact that it may not be possible to conduct an interview with all units designated for rotation. Once again, Assumptions 1 and 2, together with the definition given in 4 imply:

---

that the units are missing completely at random (MCAR), as defined in Little and Rubin (1987). This is not correct. Inclusion in the sample frame provides some information about nonparticipants, such as geographic location. At a minimum it would be possible to reconcile the sampled population with the target population using weights that depend on x. This is equivalent to the missing at random (MAR) assumption of Little and Rubin (1987). Since our focus is on non-ignorable attrition, we do not pursue this route.

Assumption 4: $R \perp (y_1, y_2)$.[5]

To set the stage for the analysis of likelihood contributions, we review what has been achieved. $D_1 = 1$ indicates that the unit was designated for the CP, and $D_2 = 1$ indicates that the unit was actually interviewed. For such individuals, there are 3 possibilities: $C = 1$ denotes attritors (present in the first, absent in the second period), $C = 2$ denotes reverse attritors (absent in the first, present in the second period), and $C = 3$ denotes units observed in both periods (those in the balanced panel). $D_1 = 0$ indicates that the unit has not been designated for the CP and $D_2 = 1$ indicates that the unit was actually interviewed. Such units are either rotated out ($R = 1$) or rotated in ($R = 2$) during the period under study. As such they contribute to the respective marginals. Individuals with $D_2 = 0$ do not participate in the survey, hence do not make contributions to either the marginals, or the joint distribution.

Indentification problem:

In what follows we suppress the conditioning on x for brevity, use the definitions in equations (1) and (2) to express the joint distribution as

$$f(y_1, y_2) = \Sigma_{D_1} \Sigma_{D_2} f(y_1, y_2, D_1, D_2), \tag{5}$$

and analyze the components one by one. We use Bayes' Theorem to extract the joint distribution function of interest and simplify the expressions by imposing our independence assumptions. We begin with the terms for nonparticipants, $D_2 = 0$, and simplify the expression by invoking, in turn, the implications of Assumptions 1, 3 and 2 to arrive at the final line.

$$f(y_1, y_2, D_1 = 0, D_2 = 0) \quad = \quad \Pr(D_1 = 0, D_2 = 0 | y_1, y_2) f(y_1, y_2)$$

[5]This assumption may seem controversial in a short panel effort on the grounds that a unit designated for rotation may have participated in earlier rounds (before period 1 in our observation window), or may return to the sample in later rounds (after period 2 in our observation window). A simple defense of our stance might be that the way we apply it, participation status is determined based on what happened during the observation window, not before, or after it. A more formal argument would start with the *ex ante* designation

$$R^* = \begin{cases} 1 \text{ if designated for out-rotation (not included in the sample frame in period 2)} \\ 2 \text{ if designated for in- rotation (not included in the sample frame in period 1)} \end{cases}, \text{ given } D_1 = 0.$$

Since participation status is random, the version defined in 4 will be determined via $R = R^* D_2$ given $D_2 = 1$. Since R^* (like D_1) is determined before the survey is fielded, Assumption 4 is no more controversial than Assumption 3.

$$
\begin{aligned}
&= \Pr(D_1 = 0, D_2 = 0) f(y_1, y_2) \\
&= \Pr(D_1 = 0) \Pr(D_2 = 0) f(y_1, y_2) \\
&= (1 - \delta_1)(1 - \delta_2) f(y_1, y_2). \qquad (6)
\end{aligned}
$$

By a similar reasoning,

$$
f(y_1, y_2, D_1 = 1, D_2 = 0) = \delta_1 (1 - \delta_2) f(y_1, y_2). \qquad (7)
$$

We turn to participants ($D_2 = 1$) next, and examine those who were not designated for the complete panel ($D_1 = 0$). These units consist of those who were rotated out, and those who were rotated in:

$$
f(y_1, y_2, D_1 = 0, D_2 = 1) = \Sigma_R f(y_1, y_2, D_1 = 0, D_2 = 1, R). \qquad (8)
$$

We examine the terms on the right hand side in turn, starting with those who were rotated out after the first period. We use Bayes' Theorem repeatedly to isolate the joint distribution of interest, and then use the fact that designation of a participating individual for rotation, and designation of the rotation status of that individual, are done independently of (y_1, y_2), via (1)-(3):

$$
\begin{aligned}
f(y_1, y_2, D_1 = 0, D_2 = 1, R = 1) &= \Pr(R = 1 | y_1, y_2, D_1 = 0, D_2 = 1) f(y_1, y_2, D_1 = 0, D_2 = 1) \\
&= \Pr(R = 1 | y_1, y_2, D_1 = 0, D_2 = 1) \Pr(D_1 = 0, D_2 = 1 | y_1, y_2) f(y_1, y_2) \\
&= \Pr(R = 1 | D_1 = 0, D_2 = 1) \Pr(D_1 = 0, D_2 = 1) f(y_1, y_2) \\
&= \phi (1 - \delta_1) \delta_2 f(y_1, y_2). \qquad (9)
\end{aligned}
$$

For those who were rotated in in the second period, we get:

$$
f(y_1, y_2, D_1 = 0, D_2 = 1, R = 2) = (1 - \phi)(1 - \delta_1) \delta_2 f(y_1, y_2). \qquad (10)
$$

The individuals who were designated for the complete panel and were interviewed consist of three subgroups:

$$
f(y_1, y_2, D_1 = 1, D_2 = 1) = \Sigma_C f(y_1, y_2, D_1 = 1, D_2 = 1, C) \qquad (11)
$$

Starting with the $C = 1$ subgroup, we invoke Bayes' Theorem and Assumptions 1 and 2 to get:

$$\begin{aligned} f(y_1, y_2, D_1 = 1, D_2 = 1, C = 1) &= \Pr(C = 1|y_1, y_2, D_1 = 1, D_2 = 1) f(y_1, y_2, D_1 = 1, D_2 = 1) \\ &= \Pr(C = 1|y_1, y_2, D_1 = 1, D_2 = 1) \Pr(D_1 = 1, D_2 = 1|y_1, y_2) f(y_1, y_2) \\ &= \Pr(C = 1|y_1, y_2, D_1 = 1, D_2 = 1) \Pr(D_1 = 1, D_2 = 1) f(y_1, y_2) \\ &= \Pr(C = 1|y_1, y_2, D_1 = 1, D_2 = 1) \Pr(D_1 = 1) \Pr(D_2 = 1) f(y_1, y_2) \\ &= \Pr(C = 1|y_1, y_2, D_1 = 1, D_2 = 1) \delta_1 \delta_2 f(y_1, y_2). \end{aligned} \tag{12}$$

The case of the $C = 2$ subgroup is similar:

$$f(y_1, y_2, D_1 = 1, D_2 = 1, C = 2) = \Pr(C = 2|y_1, y_2, D_1 = 1, D_2 = 1) \delta_1 \delta_2 f(y_1, y_2). \tag{13}$$

Turning to the $C = 3$ subgroup, we proceed in similar fashion, albeit with a slightly different focus:

$$\begin{aligned} f(y_1, y_2, D = 1, D_2 = 1, C = 3) &= f(y_1, y_2|D_1 = 1, D_2 = 1, C = 3) \Pr(D_1 = 1, D_2 = 1, C = 3) \\ &= f(y_1, y_2|D_1 = 1, D_2 = 1, C = 3) \Pr(C = 3|D_1 = 1, D_2 = 1) \Pr(D_1 = D_2 = 1) \\ &= f(y_1, y_2|D_1 = 1, D_2 = 1, C = 3) \gamma_3 \delta_1 \delta_2. \end{aligned} \tag{14}$$

It is straightforward to see that $f(y_1, y_2|D = 1, C = 3)$ 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 reverse attrition, in general $f(y_1, y_2|D_1 = 1, D_2 = 1, C = 3) \neq f(y_1, y_2)$.

Substitution of the terms we derived – namely (6)-(7), (9)-(10) and (12)-(14) – for those on the right hand side of equation (5) yields:

$$\begin{aligned} f(y_1,y_2) &= (1-\delta_1)(1-\delta_2)f(y_1,y_2)+\delta_1(1-\delta_2)f(y_1,y_2)+\phi(1-\delta_1)\delta_2 f(y_1,y_2)+(1-\phi)(1-\delta_1)\delta_2 f(y_1,y_2) \\ &\quad +Pr(C=1|y_1,y_2,D_1=1,D_2=1)\delta_1\delta_2 f(y_1,y_2)+\Pr(C=2|y_1,y_2,D_1=1,D_2=1)\delta_1\delta_2 f(y_1,y_2) \\ &\quad +f(y_1,y_2|D_1=1,D_2=1,C=3)\gamma_3\delta_1\delta_2. \end{aligned} \tag{15}$$

Upon collecting terms, simplifying and rearranging we get

$$f(y_1,y_2)=\frac{f(y_1,y_2|D_1=1,D_2=1,C=3)\gamma_3}{[1-\Pr(C=1|y_1,y_2,D_1=1,D_2=1)-\Pr(C=2|y_1,y_2,D_1=1,D_2=1)]}. \tag{16}$$

Finally, using the fact that $\sum_{c=1}^{3}\Pr(C=c|y_1,y_2,D_1=1,D_2=1)=1$, we get

$$f(y_1,y_2)=\frac{f(y_1,y_2|D_1=1,D_2=1,C=3)\gamma_3}{\Pr(C=3|y_1,y_2,D_1=1,D_2=1)}. \tag{17}$$

The last equation is equivalent to the key equation of the AN Model of Hirano et al. (2001: 1647, top). Recall that the case they study involves a two period panel, where the only concern is non-ignorable non-response in the second period (attrition). Thus $\gamma_3=\Pr(C=3|D_1=1,D_2=1)$, the fraction of retained individuals, is non-parametrically identified. They specify the probability in the denominator of (17) as a parametric function of (y_1,y_2), and investigate the conditions under which it can be identified. In our case the sampling design involves rotation, whereby non-ignorable non-response may occur either in period 1 (reverse attrition) or period 2 (attrition). This poses additional challenges for the identification of $\gamma_3=\Pr(C=3|D_1=1,D_2=1)$. The problem is attributable to the ambiguity regarding the cardinality of the set $\{D_1=1,D_2=1\}$. Given our objectives, we sidetrack this issue and treat γ_3 as a nuisance parameter. Thus our version of (17) is:

$$f(y_1,y_2)=w(y_1,y_2)f(y_1,y_2|D_1=1,D_2=1,C=3), \tag{18}$$

where $w(y_1,y_2)=\gamma_3/\Pr(C=3|y_1,y_2,D_1=1,D_2=1)>0$ by construction. This is equivalent to *rescaling* the probability in the denominator of (17). Additional restrictions on $w(y_1,y_2)$ are needed for identification.

Identification using external data:

As in HIRR, we use the links between the joint distribution and the marginals. In the AN model, the original sample yields the unbiased marginal distribution for the first period, and the refreshment sample provides an independent estimate of the unbiased marginal distribution for the second period.[6] In our case additional information for identifying the marginal distributions come from the subsamples of the original data set that yields the balanced panel, namely the subsamples of units which have been rotated in, and out. The marginal distributions we rely on (denoted by asterisk) are the (properly weighted) cross-sectional statistics published by the data collection agency.

Until now our treatment of $f(y_1, y_2)$ has been general, as in HIRR. Since our substantive application involves discrete outcomes, we supply the details for that case. Restoring the conditioning on covariates x, for the discrete case the equations of interest are:[7]

$$\sum_{y_2} f(y_1, y_2|x) = \sum_{y_2} w(y_1, y_2|x) f(y_1, y_2|D = 1, C = 3, x) = f_1^*(y_1|x), \tag{19}$$

$$\sum_{y_1} f(y_1, y_2|x) = \sum_{y_1} w(y_1, y_2|x) f(y_1, y_2|D = 1, C = 3, x) = f_2^*(y_2|x). \tag{20}$$

Equation (18) has a form which is familiar to survey data users. Once the function $w(y_1, y_2)$ is estimated (for a given x), it can be used to inflate/deflate (i.e. reflate) the cells of the balanced panel so that the object of interest $f(y_1, y_2|x)$ can be recovered. To set the stage for our substantive example, suppose y has k distinct values so that $f(y_1, y_2|x)$ can be viewed as a $k \times k$ table. Equations (19)-(20) provide the restrictions that must be satisfied by the reflated balanced panel fractions where $w(y_1, y_2|x)'s$ 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 $2k - 1$ pieces of independent information. Thus the k^2 reflation factors viewed as functions of (y_1, y_2) can have at most $2k - 1$ unknown parameters. We mimic the approach in HIRR and impose additivity. That is, for a given x we express $w(y_1, y_2|x)$ as an additive parametric function whereby only main effects of the endogenouos outcomes (y_1, y_2) are allowed. To assess the role of parametric assumptions, we follow Chen (2001) and entertain three different specifications

[6] To study the consequences of attrition in the standard panel context, Fitzgerald at al. (1998) and MaCurdy et al. (1998) rely on comparisons of later wave distributions with independent samples but do not propose a formal model of correction for attrition.

[7] Clearly the constraints for the continuous case have the same adding up feature. In the typical application the continuous distribution will be approximated by a discrete distribution, resulting in the version we have.

for this function, respectively linear, convex and concave. Details will emerge in the next section.

It is straightforward to establish that the RAN model has all the features that render the AN model attractive. Firstly, since the RAN model preserves the additivity restriction of the AN model, identification proof in HIRR still applies.[8] Secondly, it nests the popular models of attrition. These put restrictions on the function $w(y_1, y_2)$:

(i) If non-response 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 Rubin (1976).

(ii) If non-response is a function of the first period outcomes only, $w(y_1, y_2) = w(y_1)$. Little and Rubin (1987), and others – for example Fitzgerald et al. (1998), HIRR – call this case Missing at Random (MAR) because in a regular panel it is straightforward to adjust the balanced panel fractions using probability weights expressed as a function of observables in the first period.

(iii) Finally, if non-response is a function of second period outcomes only, $w(y_1, y_2) = w(y_2)$. HIRR call this the Hausman and Wise (HW) model because the case was first studied by Hausman and Wise (1979).[9]

Note that in the present case we are dealing with reverse attrition as well as regular attrition. When reverse attrition occurs, a unit that does not respond initially is observed to respond in the subsequent period. Since the first period outcomes are unobserved for a subset of the units designated for the panel, it is not possible to carry out the adjustment based on period 1 information alone, as in conventional MAR. We use the naming convention anyway, to convey the fact that the reflation factors are expressed as a function of first period outcomes only (even though they may be unobserved for a subset of the sample). Given the underlying identifying assumptions, we prefer to abbreviate the MAR model as MAR2 (because selection is ignorable with respect to period 2 outcomes), and the HW model as MAR1 (because selection is ignorable with respect to period 1 outcomes). This convention reflects the symmetry between the effects of attrition and reverse attrition and underscores a key difference between a regular panel, and short panel based on a rotating sample frame.

[8] For a simpler proof see Bhattacharya (2008). We provide another proof below, in the context of our application.

[9] Strictly speaking Hausman and Wise (1979) do not address the problem of identification of the joint distribution. They are interested in identifying the treatment effect in an experiment in which there is non-response in the second period. Fitzgerald et al. (1998) also study this model and contrast it with MAR using popular selection terminology. They point out that while selection in the MAR model is on (first period) observables, selection in the HW model is on unobservables (unobserved second period outcomes). The observable/unobservable distinction is not useful for characterizing the attrition/reverse attrition encountered in a short panel obtained from a rotating sample frame.

Before we proceed with a detailed examination of our estimation procedure, we return to our derivations and offer some observations about the role of our independence assumptions. Our derivations reveal a remarkable difference between the handling of units designed for the complete panel and the rest. While the terms that rescale $f(y_1, y_2)$ in equations 6-7 and 9-10 are exogenous probabilites, in equations (12) and (13) endogenous probabilities are present. In view of equation 3, the statement that attrition is ignorable amounts to $\Pr(C = 1|y_1, y_2, D_1 = 1, D_2 = 1) = \gamma_1$. Likewise the statement that reverse attrition is ignorable amounts to $\Pr(C = 2|y_1, y_2, D_1 = 1, D_2 = 1) = \gamma_2$. If we were to adopt this language for the other designations, we have essentially assumed that assignment status to the complete panel 1, interview status (survey nonparticipation) 2 and rotation status 4, are ignorable. Arguably the only potentially controversial one is ignorability of nonparticipation (Assumption 3). Note, however, that $\delta_2 = Pr(D_2 = 0)$ cancels out during our algebraic manipulations. Even if we were to assume non-ignorable non-participation, i.e. start with $Pr(D_2 = 0|y_1, y_2) = \delta_2(y_1, y_2)$, this term will drop out as we move from equation 15 to 16. Unlike attritors and reverse attritors, survey nonparticipants do not make any contribution whatsoever to the data collection effort – either in the first period, or in the second period. As such, survey nonparticipants do not have the same potential to distort the balanced panel. This line of thinking suggests that ignorability is a reasonable assumption in the case of nonparticipation.

3 Estimation and Inference in RAN Model

We illustrate the utility of the RAN model by applying it to a case where y is a multiple valued random variable that captures labor market status and takes one of three values (0 = non-participant, 1 = employed, 2 = unemployed). In this case the equation system (19)-(20) 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 and define the linear index as

$$i(y_1, y_2|x) = \mu + \rho_1 I(y_1 = 1) + \rho_2 I(y_1 = 2) + \kappa_1 I(y_2 = 1) + \kappa_2 I(y_2 = 1) \equiv i(\underline{\beta}|y_1, y_2, x) \qquad (21)$$

where $I(.)$ denotes the indicator function and $\underline{\beta} = [\mu\ \rho_1\ \rho_2\ \kappa_1\ \kappa_2]$'. This additive function of the unknown parameters captures the dependency of attrition and reverse attrition on the labor market states (y_1, y_2). As in HIRR, we rule out interactions and focus on the main effects of the labor market states. In our empirical work, we use three parametric forms for the reflation factor: (i) linear: $w_L(y_1, y_2|x) = i(\underline{\beta}|y_1, y_2, x)$, (ii) convex: $w_X(y_1, y_2|x) = \exp\left\{i(\underline{\beta}|y_1, y_2, x)\right\}$, and (iii) concave: $w_E(y_1, y_2|x) = 2 - \exp\left\{i(\underline{\beta}|y_1, y_2, x)\right\}$. Note that $w(y_1, y_2) = 1\ iff\ \mu = 1, \rho_1 = \rho_2 = \kappa_1 = \kappa_2 = 0$ in the linear case. In the nonlinear cases, $w(y_1, y_2) = 1\ iff\ \mu = \rho_1 = \rho_2 = \kappa_1 = \kappa_2 = 0$.

For the linear case, the restrictions imposed via equations (19)-(20) can be represented as in Table 1. The task amounts to finding the reflation factors (functions of $\underline{\beta}$) that adjust the balanced panel fractions – so that the adjusted cell probabilities are in line with the marginals reported by the data collection agency.[10]

Table 1: A 3x3 Linear RAN Model

	$y_2 = 0$	$y_2 = 1$	$y_2 = 2$	Row sum
$y_1 = 0$	μp_{00}	$(\mu + \kappa_1)p_{01}$	$(\mu + \kappa_2)p_{02}$	$f_1^*(0)$
$y_1 = 1$	$(\mu + \rho_1)p_{10}$	$(\mu + \rho_1 + \kappa_1)p_{11}$	$(\mu + \rho_1 + \kappa_2)p_{12}$	$f_1^*(1)$
$y_1 = 2$	$(\mu + \rho_2)p_{20}$	$(\mu + \rho_2 + \kappa_1)p_{21}$	$(\mu + \rho_2 + \kappa_2)p_{22}$	$f_1^*(2)$
Col. sum	$f_2^*(0)$	$f_2^*(1)$	$f_2^*(2)$	1

For the linear case, it can be shown that the system of equations has the observationally equivalent representation given below:

$$\begin{bmatrix} \sum_{j=0}^{2} p_{0j} & 0 & 0 & p_{01} & p_{02} \\ \sum_{j=0}^{2} p_{1j} & \sum_{j=0}^{2} p_{1j} & 0 & p_{11} & p_{12} \\ \sum_{j=0}^{2} p_{2j} & 0 & \sum_{j=0}^{2} p_{2j} & p_{21} & p_{22} \\ \sum_{k=0}^{2} p_{k0} & p_{10} & p_{20} & 0 & 0 \\ \sum_{k=0}^{2} p_{k1} & p_{11} & p_{21} & \sum_{k=0}^{2} p_{k1} & 0 \\ \sum_{k=0}^{2} p_{k2} & p_{12} & p_{22} & 0 & \sum_{k=0}^{2} p_{k2} \end{bmatrix} \begin{bmatrix} \mu \\ \rho_1 \\ \rho_2 \\ \kappa_1 \\ \kappa_2 \end{bmatrix} = \begin{bmatrix} f_1^*(0) \\ f_1^*(1) \\ f_1^*(2) \\ f_2^*(0) \\ f_2^*(1) \\ f_2^*(2) \end{bmatrix} \tag{22}$$

Inspection reveals that this six-equation system is of the form $A\underline{\beta} = \underline{b}$ where A is of rank= 5. One of the constraints is redundant, in the sense that it will be automatically met once the solution

[10] In Tunalı and Ekinci (2007) we used the subsamples from the two cross-sections that were not subjected to attrition, namely the units that were rotated in, and out. These constitute approximately 25% of the cross-section sample. In the current version we use the official statistics (reported by TURKSTAT) which rely on the full cross-section sample. The marginal distributions reported by TURKSTAT are obtained by a MAR type weighting scheme.

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. Consequently, the solution to the reduced five-equation system is unique, and does not depend on which constraint is left out. Without loss of generality we exclude the last constraint, and obtain the five-equation system

$$\begin{bmatrix} \sum_{j=0}^{2} p_{0j} & 0 & 0 & p_{01} & p_{02} \\ \sum_{j=0}^{2} p_{1j} & \sum_{j=0}^{2} p_{1j} & 0 & p_{11} & p_{12} \\ \sum_{j=0}^{2} p_{2j} & 0 & \sum_{j=0}^{2} p_{2j} & p_{21} & p_{22} \\ \sum_{k=0}^{2} p_{k0} & p_{10} & p_{20} & 0 & 0 \\ \sum_{k=0}^{2} p_{k1} & p_{11} & p_{21} & \sum_{k=0}^{2} p_{k1} & 0 \end{bmatrix} \begin{bmatrix} \mu \\ \rho_1 \\ \rho_2 \\ \kappa_1 \\ \kappa_2 \end{bmatrix} = \begin{bmatrix} f_1^*(0) \\ f_1^*(1) \\ f_1^*(2) \\ f_2^*(0) \\ f_2^*(1) \end{bmatrix}. \tag{23}$$

Here we left out row 6 of A and b, which can be represented in matrix notation as $A_6\underline{\beta} = \underline{b}_6$. The unique solution to our just-identified system is $\hat{\underline{\beta}} = A_6^{-1}\underline{b}_6$.

While a closed form solution is available for the linear version, this is not the case when non-linear rescaling functions are used. However it is possible to obtain numerical solutions to the non-linear version using widely available algorithms for solving systems of equations. In our empirical work we have pursued this route and established that the solution is robust to the alternative parameterizations decribed earlier.[11] Presently we turn to alternate characterizations that restate the problem in hand as optimization problems.

Maximum Likelihood

Although we established that the linear RAN model has an exact solution, derivation of the asymptotic covariance matrix of the estimated parameters requires additional work.[12] Since the ML approach has the advantage of producing a consistent estimate of this matrix, we go over it in the context of our example and relate it to our earlier discussion.

Given the nature of the outcome variable, the distribution in Table 1 can be characterized via a multinomial p.m.f. Towards that end, we first reparameterize the cell probabilities as shown in

[11] See Tunalı *et al.* (2012). Similar conclusions have been drawn in RAN model applications with more states – see İkizler and Tunalı (2012), Gökçe and Tunalı (2014), Özkan and Tunalı (2014).

[12] The papers cited in the previous footnote rely on Bootstrap methods for doing inference.

Table 2:

Table 2: Reparameterized 3x3 Linear RAN Model

	$y_2 = 0$	$y_2 = 1$	$y_2 = 2$	Row sum
$y_1 = 0$	θ_{00}	θ_{01}	θ_{02}	$f^*_{0\bullet}$
$y_1 = 1$	θ_{10}	θ_{11}	θ_{12}	$f^*_{1\bullet}$
$y_1 = 2$	θ_{20}	θ_{21}	θ_{22}	$f^*_{2\bullet}$
Col. sum	$f^*_{\bullet 0}$	$f^*_{\bullet 1}$	$f^*_{\bullet 2}$	1

Next, we let n_{jk} denote the number of observations in cell (j,k) of the balanced panel, $j,k = 0,1,2$. These are related to $p'_{jk}s$ via $p_{jk} = n_{jk}/N$, where N denotes the number of observations in the balanced panel. Using the reparameterized cell probabilities, the likelihood function for the linear version may be expressed as:

$$\mathcal{L}(\underline{\theta}) = \{\theta_{00}\}^{n_{00}} \{\theta_{01}\}^{n_{01}} \{\theta_{02}\}^{n_{02}} \{\theta_{10}\}^{n_{10}} \{\theta_{11}\}^{n_{11}} \{\theta_{12}\}^{n_{12}} \{\theta_{20}\}^{n_{20}} \{\theta_{21}\}^{n_{21}} \{\theta_{22}\}^{n_{22}} . \tag{24}$$

Maximization will be done subject to the adding up constraints, (19)-(20), which together imply (23). The standard approach would entail embedding an appropriate lagrangian function in the likelihood function. Using the correspondence between Tables 1 and 2, the log-likelihood function that incorporates the constraints may be expressed as:

$$ln\mathcal{L}(\underline{\beta},\underline{\lambda}) = n_{00} ln\{\theta_{00}\} + n_{01} ln\{\theta_{01}\} + n_{02} ln\{\theta_{02}\} + n_{10} ln\{\theta_{10}\} + n_{11} ln\{\theta_{11}\} + n_{12} ln\{\theta_{12}\} + n_{20} ln\{\theta_{20}\}$$

$$+ n_{21} ln\{\theta_{21}\} + n_{22} ln\{\theta_{22}\} - \lambda_1 [(\theta_{00} + \theta_{01} + \theta_{02}) - f^*_{0\bullet}] - \lambda_2 [(\theta_{10} + \theta_{11} + \theta_{12}) - f^*_{1\bullet}]$$

$$- \lambda_3 [(\theta_{20} + \theta_{21} + \theta_{22}) - f^*_{2\bullet}] - \lambda_4 [(\theta_{00} + \theta_{10} + \theta_{20}) - f^*_{\bullet 0}] - \lambda_5 [(\theta_{01} + \theta_{11} + \theta_{21}) - f^*_{\bullet 1}] . \tag{25}$$

It can be shown that the F.O.C.'s with respect to the parameter vector $\underline{\beta}' = [\mu\ \rho_1\ \rho_2\ \kappa_1\ \kappa_2]$ yield the following system of equations:

$$B\underline{\lambda} = C\underline{d}(\underline{\beta}), \tag{26}$$

where $\underline{\lambda}$ denotes the 5×1 vector of lagrange multipliers,

$$B = \begin{bmatrix} \sum_{j=0}^{2} p_{0j} & \sum_{j=0}^{2} p_{1j} & \sum_{j=0}^{2} p_{2j} & \sum_{k=0}^{2} p_{k0} & \sum_{k=0}^{2} p_{k1} \\ 0 & \sum_{j=0}^{2} p_{1j} & 0 & p_{10} & p_{11} \\ 0 & 0 & \sum_{j=0}^{2} p_{2j} & p_{20} & p_{21} \\ p_{01} & p_{11} & p_{21} & 0 & \sum_{k=0}^{2} p_{k1} \\ p_{02} & p_{12} & p_{22} & 0 & 0 \end{bmatrix}, \tag{27}$$

$$C = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}, \tag{28}$$

and

$$\underline{d}(\underline{\beta}) = \begin{bmatrix} \mu^{-1} n_{00} \\ (\mu+\kappa_1)^{-1} n_{01} \\ (\mu+\kappa_2)^{-1} n_{02} \\ (\mu+\rho_1)^{-1} n_{10} \\ (\mu+\rho_1+\kappa_1)^{-1} n_{11} \\ (\mu+\rho_1+\kappa_2)^{-1} n_{12} \\ (\mu+\rho_2)^{-1} n_{20} \\ (\mu+\rho_2+\kappa_1)^{-1} n_{21} \\ (\mu+\rho_2+\kappa_2)^{-1} n_{22} \end{bmatrix} \equiv \begin{bmatrix} \theta_{00} \\ \theta_{01} \\ \theta_{02} \\ \theta_{10} \\ \theta_{11} \\ \theta_{12} \\ \theta_{20} \\ \theta_{21} \\ \theta_{22} \end{bmatrix}. \tag{29}$$

..

Let $p_{k\bullet} = \sum_{j=0}^{2} p_{kj},\ k = 0,1,2;\ p_{\bullet j} = \sum_{k=0}^{2} p_{kj},\ j = 0,1,2$. With this reparametrization our auxiliary system of equations given by 26 becomes:

$$\begin{bmatrix} p_{0\bullet} & p_{1\bullet} & p_{2\bullet} & p_{\bullet 0} & p_{\bullet 1} \\ 0 & p_{1\bullet} & 0 & p_{10} & p_{11} \\ 0 & 0 & p_{2\bullet} & p_{20} & p_{21} \\ p_{01} & p_{11} & p_{21} & 0 & p_{\bullet 1} \\ p_{02} & p_{12} & p_{22} & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu^{-1} n_{00} \\ (\mu + \kappa_1)^{-1} n_{01} \\ (\mu + \kappa_2)^{-1} n_{02} \\ (\mu + \rho_1)^{-1} n_{10} \\ (\mu + \rho_1 + \kappa_1)^{-1} n_{11} \\ (\mu + \rho_1 + \kappa_2)^{-1} n_{12} \\ (\mu + \rho_2)^{-1} n_{20} \\ (\mu + \rho_2 + \kappa_1)^{-1} n_{21} \\ (\mu + \rho_2 + \kappa_2)^{-1} n_{22} \end{bmatrix}$$

Equation 29 serves to illustrate the key implication of our *identifying assumption:* the 9×1 vector $\underline{\theta}$ of adjusted multinomial probabilities, is a function of the 5×1 unknown parameter vector$\underline{\beta}$. It is straightforward to show that the 5×5 matrix B – which happens to be a function of the balanced panel cell fractions – is invertible, by virtue of the fact that it is equal to the transpose of matrix A_6 defined above, as the square matrix in the reduced equation system 23. This establishes that the log-likelihood function can be concentrated using

$$\underline{\lambda} = B^{-1} C \underline{d}(\underline{\beta}). \tag{30}$$

Thus the 10 parameter constrained optimization problem (in terms of unknowns $\underline{\beta}, \underline{\lambda}$) is equivalent to an appropriately transformed 5 parameter optimization problem, where the lagrange multipliers are expressed as explicit functions of the 5×1 unknown parameter vector, $\underline{\beta}$. This establishes that we have a standard ML estimation problem at hand, which can be tackled by standard software.

M-estimation

It is straightforward to establish that the problem we solved is equivalent to a minimization problem involving the scalar function $q(\underline{\beta}) = \underline{\lambda}' m(\underline{b}_6 - A_6 \underline{\beta})$, where A_6 and $\underline{b}_6$ have been defined earlier, $\underline{\lambda}$ denotes a vector with positive elements, $m(.)$ denotes an appropriate distance metric such as

quadratic, or absolute value. For a given $\underline{\lambda}$, the solution can be expressed as

$$\underline{\hat{\beta}}(\underline{\lambda}) = argmin\underline{\lambda}'m\left(\underline{b}_6 - A_6\underline{\beta}\right).$$

As in the ML case, the asymptotic theory for M-estimation is well-established (see, for example, Wooldridge, 2009).

Generalization

Generalization to k categorical outcomes is possible. Let $f_{ij} = f(y_1 = i, y_2 = j)$, $p_{ij} = f(y_1 = i, y_2 = j|D = 1, C = 3, x)$ and $w(y_1, y_2) = w_{ij}$ with $i = 1, ..., k$ and $j = 1, ..., k$. Equation (18) may be rewritten as:

$$\frac{f_{ij}}{p_{ij}} = w_{ij} \tag{31}$$

With some abuse of notation, express the linear case as:

$$\frac{f_{ij}}{p_{ij}} = i(\underline{\beta}|y_1, y_2, x) \tag{32}$$

The convex version may be written as:

$$ln\left(\frac{f_{ij}}{p_{ij}}\right) = i(\underline{\beta}|y_1, y_2, x) \tag{33}$$

The concave version may be written as:

$$ln\left(2 - \frac{f_{ij}}{p_{ij}}\right) = i(\underline{\beta}|y_1, y_2, x) \tag{34}$$

The general form of the $(2k)\times(2k-1)$ matrix A, the $(2k-1)\times 1)$ vector $\underline{\beta}$, and the $(2k) \times 1)$ vector $\underline{b}$ are easily discerned. We know that the linear case is additive in the unknown parameters, in this case identification is straightforward. In the non-linear versions, a known function $h(.)$ of the known ratio $\frac{f_{ij}}{p_{ij}}$ is additive in the unknown parameters. This establishes identification of the non-linear cases. Clearly systems 32,33 and 34 will yield different estimates of the unknown $\underline{\beta}'s$. Note that ultimately the quantities of interest are not the $\underline{\beta}'s$ but the weights used in rescaling, defined by $w_L(y_1, y_2|x) =$

$i(\underline{\beta}|y_1, y_2, x)$, $w_X(y_1, y_2|x) = \exp\left\{i(\underline{\beta}|y_1, y_2, x)\right\}$, and $w_E(y_1, y_2|x) = 2 - \exp\left\{i(\underline{\beta}|y_1, y_2, x)\right\}$. Thus investigation of the sensitivity of RAN model estimates to the parameteric assumptions hinges on comparison of the $w_S(j, k|x), j, k = 0, 1, .., k-1$ for $S = L, X, E$.

Discussion

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 in the next section, when we discuss the lessons learned from a broader empirical investigation.

At this point it is appropriate to provide a brief account of how our adjustment procedure relates to/differs from existing methods proposed in papers we view as being "close" to ours. As mentioned earlier, Abowd and Zellner (1985) and Stasny (1986, 1988) deal with the same substantive issues in the short panel context, but work with counts. Unlike Hirano et. al. (2001) that guided us, these papers do not offer a formal model of the selective nonresponse process. The goal is stated as adjustment of the gross flow data so that it can be reconciled with the marginals. Abowd and Zellner (1985) use a multiplicative model to distribute those who are not observed in both periods to the appropriate margins (original set of states plus two others, respectively attritor/reverse attritor and rotated in/out). Like us (see Section 4) they study three states (nine cells in the flow matrix), but estimate 18 unknown parameters subject to six restrictions coming from the margins. Thus, they not only allow interaction effects, but they also distinguish between attrition and reverse attrition parameters. Clearly this overparametrized model cannot be used to implement separate adjustments for each period pair. They assume stationarity and use multiple rounds of CPS data to estimate "average" parameters by minimizing the weighted squared deviation of the adjusted gross flow margins from the observed population margins.

Stasny (1986, 1988) uses an additive model that resembles ours. In her model an individual designated for the panel can lose either its row or column designation, with different probabilities. In the richest models (A and D in Stasny, 1988) she expresses one of these probabilities as a function of states occupied in the first period, the other as a function of states occupied in the second period. Thus, the probability that someone designated for the complete panel ends up in one of the margins, is a function of the states in both periods. This treatment is equivalent to

ours. Arguably our version is more attractive, as underscored by the representation given in the subsection titled Generalization. In her examples there are three states and six free parameters for each period pair, which can be estimated subject to the six restrictions on the margins. She is able to identify an extra parameter because she uses count data, while we work with shares. She uses maximum likelihood estimation on multiple rounds of data from CPS and Canadian LFS.

There is a well-established line of research in the statistical literature which is directed at the important distinction between the sampled and the target population, and on methods used in reconciling them (Madow et al., 1983). Little (1993) refers to adjustments of data obtained from surveys (i.e. sampled population) using aggregate data on the (target) population obtained from other sources as "post-stratification." The bulk of this paper is concerned with the case when the population joint distribution of the post-stratification variables is known. Little briefly discusses a case which is of special interest for us: only the marginal population distributions of the post-stratification variables are known. When non-response is present, the joint distribution of the post-stratification variables in the sample is not adequate for estimation (unless MCAR or MAR is assumed). This case is covered at length in Little and Wu (1991) where a formal model for nonresponse is given. Notably they address the identification issue and show that a model in which the response probability is expressed as a product of row and column effects is just identified. They propose an iterative method (raking) for estimation of this model. This version of the post-stratification exercise is intimately connected with the AN/RAN approach. Instead of the additive model that drives the correction is AN/RAN models, Little and Wu (1991) have a multiplicative model.

In AN model applications reported in HIRR, imputation (via a MCMC procedure) of the missing outcomes precedes the estimation of the joint distribution of interest. This amounts to adopting the predictive modeling perspective of Little (1991). In our application of the RAN model we proceed with the estimation of the reflation factors and the adjusted cell probabilities without engaging in computationally costly imputation.[13] Recent papers framed within the AN attrition-refreshment sample framework include Nevo (2003), Bhattacharya (2008) and Deng, Hillygus, Reiter and Zheng

[13]Evidently the idea of using reflation factors to bring a possibly biased joint distribution in line with marginals that can be trusted is an old one, discovered by researchers who work with cross-section data. An early example of this is Golan, Judge and Robinson (1994). Their objective is to recover the elements of expenditure, trade or income flows from limited or incomplete multisectoral economic data. They use a similar set of adding up restrictions on columns and raws of the data as we do. We thank Touhami Abdelhalek for bringing this paper to our attention.

(2013). Nevo (2003) and Bhattacharya (2008) cast the estimation problem in a familiar panel data framework where the object of interest is a conditional expectation function (CEF) rather than the joint distribution of outcomes. Nevo (2003) adopts a GMM procedure for estimation of the attrition function and the unknown parameters of the CEF. Apart from providing a simpler identification proof for AN model of HIRR, Bhattacaharya (2008) proposes a sieve-based estimation method and establishes the asymptotic properties of the estimator. Deng et al. (2013) examine the utility of refreshment samples in a multi period panel context. They characterize the data generation process using a Bayesian approach and apply it to a panel with three waves. From our point of view, the extension to multiple waves is useful in characterizing the links between adjacent two-period observation windows. However no useful insights emerge for relaxing the key AN identifying assumption, namely ruling out interaction effects. In the concluding section the authors offer a discussion on initial non-response, what we have termed non-participation to distinguish it from attrition an reverse attrition. They argue that the ignorability assumption may be too strong, and view this as a gap in the literature. We believe that our discussion of Assumption 3 sheds further light on the problem by separating what can, and cannot be modelled in a rotating panel context.

4 Examples

Our example is a familar one from Labor Economics: correction of transition rates obtained from balanced panels of the Household Labor Force Survey in Turkey (HFLST). In Table 3, we compiled a set of parameter estimates from a 3x3 RAN model for annual transitions together with bootstrap means and standard errors based on 100 replications. For the linear parameterization of the RAN model we also reported ML estimates of the standerd errors. In this example x denotes the entire working age population, ages 15 and over. The balanced panel contained over $20,000$ observations. The first and second period marginals in the raw data contained over $52,000$ observations. Thus it is not surprising that all RAN model parameters are estimated extremely precisely.

As we noted earlier, HLFS-Turkey 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 40 percent of the respective marginals. The fact that this fraction is considerably lower than the expected 0.5 can be taken as a rough

Table 3: 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
ML std. error	0.0075	0.0115	0.0182	0.0117	0.0163
(ii) $w(.)$ convex:					
Estimate	−0.1057	0.0957	0.2306	0.1293	0.1703
Bootstrap mean	−0.1050	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	−0.0975	0.0960	0.2848	0.1349	0.1885
Bootstrap mean	−0.0968	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.

statistic that warns us about the magnitude of the attrition/reverse attrition problem. In fact attrition in HLFS-Turkey is quite severe.[14] 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 do not expect this to be the case. In fact, Wald tests provide overwhelming evidence that the attrition and reverse attrition 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 reverse attrition behavior is intimately connected with labor market behavior, is vindicated.

In Table 4, we compiled the set of reflation factor estimates we obtained from the RAN model parameter estimates reported in Table 3. For brevity we excluded the numbers for the margins. The numbers reported in each cell are of the form given in Table 1: reflation factor, times the balanced

[14] A detailed analysis of the realized magnitudes of attrition in the HLFS-Turkey over the period 2000-2002 are reported in Tunalı (2009). There were a total of 66,467 households (headed by someone of age 15 or older) and 184,339 individuals (of age 15 or older), but not all were subjected to the full rotation plan. About 26% of eligible households and 31.6% of eligible individuals attrited sometime during the observation window. For the subset of (23,790) households headed by prime-age (20–54-years-old) individuals which were designated for four interviews, the cumulative probability of attrition was 8% by 3 months, 18.3% by 12 months, and 24.7% by 15 months.

Table 4: 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
$y_1 = 0$	0.8986 0.8997 0.8976	0.5052	1.0302 1.0238 1.0366	0.0566	1.0766 1.0667 1.0870	0.0159	$f_1(0)$
$y_1 = 1$	0.9943 0.9901 0.9985	0.0740	1.1258 1.1267 1.1248	0.2952	1.1722 1.1739 1.1706	0.0209	$f_1(1)$
$y_1 = 2$	1.1511 1.1330 1.1708	0.0113	1.2826 1.2894 1.2754	0.0122	1.3290 1.3433 1.3133	0.0085	$f_1(2)$
Col. sum	$f_2(0)$		$f_2(1)$		$f_2(2)$		1

panel fraction. For each cell we report the estimates of the reflation factors $w(.)$ associated with all three functional forms (respectively linear, convex, concave) inside braces. Reflation factors below (above) one mark labor market states which are overrepresented (underrepresented) in the balanced panel. Note that for some states the bias induced by attrition/reverse attrition 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.

Table 5: A 3x3 RAN Model - Adjusted and [Unadjusted] Joint and Marginal Probabilities

Annual Transitions Between 2001-Q1 and 2002-Q1 x = age 15 and over				
	$y_2 = 0$	$y_2 = 1$	$y_2 = 2$	Row sum
$y_1 = 0$	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

Table 45 provides 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 [discrepancies between $f(y_1, y_2|D = 1, C = 3, x)$ and $f(y_1, y_2|x)$] range between -24 and 11 percent. Six of the 9 cells have biases of 10% or more in

Table 6: 3x3 RAN Model - Adjusted and [Unadjusted] Transition Probabilities

Annual Transitions Between 2001-Q1 and 2002-Q1 x = age 15 and over				
	$y_2 = 0$	$y_2 = 1$	$y_2 = 2$	Row sum
$y_1 = 0$	0.8573 [0.8744]	0.1102 [0.0980]	0.0325 [0.0276]	1 [1]
$y_1 = 1$	0.1710 [0.1898]	0.7720 [0.7565]	0.0570 [0.0537]	1 [1]
$y_1 = 2$	0.3250 [0.3525]	0.3917 [0.3813]	0.2833 [0.2662]	1 [1]

absolute value.

In Table 6, the associated forward transition probabilities 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 capability of the RAN model is likely to be especially important in the case of statistical agencies like TURKSTAT, where experts refuse to exploit the short panel dimension of the HLFS-Turkey on the grounds that there is no weighting method that can reconcile dynamic and static estimates.

As we argued above, the non-parametric feature of the 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 partition of the sample are reported in Table 7. In this case the statistical evidence favors the hypothesis that attrition/reverse attrition is ignorable. Note that the sample sizes are small, and consequently bootstrapped standard errors based on 100 replications are large. In fact in some cases the bootstrapped means are very different from the estimated parameter value (see θ_{12} and θ_{22} for the concave case). This finding exposes the well-known fragility of the bootstrap method when resampling is done from small samples. In such cases it would be advisable to use ML estimation.

When the narrower objective of producing dynamic statistics consistent with the cross-section statistics is adopted, the correction can proceed despite our cautionary remark. In fact, the reflation

Table 7: 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	−0.0234	0.1348	0.0688	0.3507
Bootstrap mean	0.9465	−0.0003	0.2731	0.0524	0.3638
Bootstrap std. error	0.1271	0.2530	0.5869	0.2220	0.4227
(ii) $w(.)$ convex:					
Estimate	−0.0540	−0.0231	0.1191	0.0697	0.3127
Bootstrap mean	−0.0699	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	−0.0518	−0.0239	0.1560	0.0681	0.4101
Bootstrap mean	−0.0378	−0.0037	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				

factors for the subsample under examination reported in Table8 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. For example, consider cell $(y_1 = 2, y_2 = 2)$ for which substantial differences are observed across parametric forms (see the second digits after the decimal point reported inside braces). In the balanced panel there are only 4 individuals in this cell. Under the alternative parametric assumptions the adjusted fractions for this cell are respectively 0.012464, 0.012694 and 0.012198. Thus, as long as small cell sizes yield a small magnitude for p_{jk} (< 0.01, say), the differences in RAN model reflection factor estimates by functional form do not translate to comparable differences in the magnitudes of the adjusted fraction.

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

Table 8: Another 3×3 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
$y_1 = 0$	0.9471 0.9474 0.9468	0.0978		1.0160 1.0158 1.0162	0.0196		1.2978 1.2952 1.3011	0.0087	$f_1(0)$
$y_1 = 1$	0.9237 0.9258 0.9214	0.0652		0.9926 0.9927 0.9924	0.7543		1.2744 1.2657 1.2843	0.0239	$f_1(1)$
$y_1 = 2$	1.0819 1.0672 1.0989	0.0109		1.1508 1.1443 1.1583	0.0109		1.4326 1.4591 1.4021	0.0087	$f_1(2)$
Col. sum	$f_2(0)$			$f_2(1)$			$f_2(2)$		1

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 bootstrapped samples yielded zero cell counts, in which case correction could not proceed. Clearly zeros encountered during bootstrapping are random as opposed to structural zeros. We were able to fix the problem by adding an observation to the empty cell and adjusting the sample size accordingly.

The fix we developed was also useful in higher dimensional RAN models (up to 5×5) we experimented with. Clearly empirical findings regarding the nature of attrition/reverse attrition can, and do vary, from one time period to the other, and with choice of x. With sufficient data, a second stage analysis can be performed to shed light on the patterns (see Ikizler and Tunali, 2012, and Gokce and Tunali, 2012 for substantive examples from 4×4 RAN models). The fragility exposed in Table 7 suggests that the number of bootstrap replications we used (100) may not be adequate for credibly testing whether attrition/reverse attrition is non ignorable. Nonetheless there are valid reasons for proceeding with the correction whether or not attrition/reverse attrition 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 the RAN model

produced estimates of transition rates for commonly used partitions of the full sample (jointly by sex and location, by education, by broad age groups) that are robust to choice of functional form. Even further partitioning of the subsamples identified by sex-location pairs either by education, or by broad age groups, proved to be feasible. Thus our method is worthy of adoption for statistical and policy analysis purposes.

6 Conclusion

In this paper we tackle a generalized version of the attrition problem, typically associated with longitudinal data. The motivation for taking a fresh look comes from the observation that many sustained large scale data collection efforts (CPS, the EU-LFS and EU-SILC being some well-known examples) involve multiple visits to the same address/household over a short period (up to four years). A shared feature of these efforts is the use of 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. Notably 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 any given point of time. The generalization we offer recognizes that proper use of such short panels may require corrections for non-response after initial response (attrition) as well as response after initial non-response (reverse attrition). Furthermore, attrition behavior we envision is allowed to be very general as well, in that it can depend on endogenous outcomes in either period .

In our empirical example outcomes are labor market states occupied by an individual. Endogeneity implies that particular labor market outcome combinations could make individuals more or less prone to exclusion from the balanced panel. The model we use exploits the key insights in HIRR (Hirano, Imbens, Ridder and Rubin, 2001) and shares some features of their AN model, but departs from it in other respects. When both attrition and reverse attrition are at work, a parameter that can be non-parametrically identified in the AN model becomes unidentified. We show that the information loss can be sidestepped by rescaling, and work with the Rescaled AN (RAN) model. As in the AN model, correction in the RAN model amounts to reflating the balanced panel fractions (cell means) by factors expressed as a parametric function of the states under examination. The

parameters of the parametric reflation function are identified by exploiting the adding up constraints that the marginals impose on the joint distribution. Both models require additional data to remedy the losses from attriton. In the AN model, additional data take the form of a so-called refreshment sample, an independently collected cross-section. In the RAN model the additional data happen to be cross-section data collected along with the short panel, from units designated for rotation. As a result while the RAN model operates within the constraints of the original data collection effort, the AN model requires additional effort, or external data. Another attractive feature of the RAN model is its computational simplicity, especially when the linear version is adopted.

Our empirical investigation of annual transition data from the Household Labor Force Survey in Turkey showed that attrition is a serious concern, in the sense that transition rates obtained from the balanced panel are systematically distorted. RAN model based adjustment not only corrects these distortions, but it also reveals the attrition patterns. Based on our systematic empirical investigation, results did not display sensitivity to the parametric features of the RAN model. Thus the linear version – which is extremely simple to implement – appears suitable for empirical work. Yet another attractive feature of the 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. In a nutshell, RAN model is designed to produce estimates of transition rates which are consistent with cross-section statistics, conditional on covariates of interest. As such it is likely to gain the approval of official statistical agencies. Furthermore, estimation does not require micro data. To implement the adjustments, it is sufficient to have the joint distribution obtained from the balanced panel that links the two legs of the short panel along with the marginal distributions obtained from representative data collected at each leg. Since all of this information is readily available from statistical agencies in tabular form, the proposed methodology should appeal to a very broad audience.

7 References

Abowd, J. M. and A. Zellner (1985) "Estimating Gross Labor-Force Flows." *Journal of Business and Economic Statistics,* 3:3, 254-283.

Bhattacharya, D. (2008) "Inference In Panel Data Models Under Attrition Caused by Unobservables." *Journal of Econometrics*, 144: 430-446.

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.

EUROSTAT (2007) *Labor Force Survey in the EU, Candidate and EFTA Countries; 2007 Edition.* Office for Official Publications of the European Communities: Luxembourg.

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

Deng, Y., D. S. Hillygus, J. P. Reiter, Y. Si and S. Zheng (2913) "Handling Attrition in Longitudinal Studies: The Case for Refreshment Samples." *Statistical Science*, 28:2, 238–256.

Gökçe, O. Z and İ. Tunalı (2014) "Informality and Labor Market Mobility in Turkey: Evidence from Micro Data, 2000-2002." Paper presented at the 20th Annual Conference of the *Economic Research Forum*, Cairo.

Golan, A, G. Judge, and S. Robinson (1994) "Recovering Information from Incomplete or Partial Multisectoral Economic Data." *Review of Economics and Statistics*, 76: 541-549.

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.

İkizler, H. and İ. Tunalı (2012) "Agricultural Transformation and Labor Mobility During The ARIP Period in Turkey: Evidence From Micro-Data, 2000-2002." Paper presented at the 18th Annual Conference of the *Economic Research Forum*, Cairo.

Little, R. J. A. (1982) "Models for Nonresponse in Sample Surveys." *Journal of the American Statistical Association*, v. 77, no. 378, 237-250.

Little, R. J. A. (1993) "Post-stratification: A Modeller's Perspective." *Journal of the American Statistical Association*, v. 88, no. 423, 1001-1012.

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

Little, R. J. A. and M. Wu (1991) "Models for Contingency Tables With Known Margins When Target and Sampled Populations Differ." *Journal of the American Statistical Association*, v. 86, no. 413, 87-95.

MaCurdy, T., T. Mroz, and R. M. Gritz (1998) "An Evaluation of the National Longitudinal Survey on Youth." *The Journal of Human Resources*, v. 33, no. 2, 345-436.

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

Nevo, A. (2003) "Using weights to adjust for sample selection when auxiliary information is available." *Journal of Business and Economic Statistics*, 21: 1, 43–52.

Ridder, G. (1992) "An Empirical Evaluation of Some Models for Non-Random Attrition in Panel Data." *Structural Change and Economic Dynamics*, 3: 337-355.

Ridder, G. and R. Moffitt (2007) "The Econometrics of Data Combination." Ch. 75 in J.J. Heckman and E.E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B. Amsterdam: Elsevier B.V..

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

Stasny, E. A. (1986) "Estimating Gross Flows Using Panel Data With Nonresponse: An Example from the Canadian Labor Force Survey." *Journal of the American Statistical Association*, v. 81, no. 393, 42-47.

Stasny, E. A. (1988) "Modeling Nonignorable Nonresponse in Categorical Panel Data With an Example in Estimating Gross Labor-Force Flows." *Journal of Journal of Business and Economic Statistics,* 6:2, 207-219.

Tunalı, İ. (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.

Tunalı, İ. and E. Ekinci (2007) "Dealing with Attrition When Refreshment Samples are Available: An Application to the Turkish Household Labor Force Survey." Mimeo.

Tunali, İ., E. Ekinci ve B. Yavuzoğlu (2012) "Rescaled Additively Nonignorable Model of Attrition: A Convenient Semi-Parametric Bias-Correction Framework for Data with a Short Panel Component." Mimeo.

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 equation (22) 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 $A_6\underline{\beta} = \underline{b}_6$ and has the explicit form given below:

$$\begin{bmatrix} \sum_{j=0}^{2} p_{0j} & 0 & 0 & p_{01} & p_{02} \\ \sum_{j=0}^{2} p_{1j} & \sum_{j=0}^{2} p_{1j} & 0 & p_{11} & p_{12} \\ \sum_{j=0}^{2} p_{2j} & 0 & \sum_{j=0}^{2} p_{2j} & p_{21} & p_{22} \\ \sum_{k=0}^{2} p_{k0} & p_{10} & p_{20} & 0 & 0 \\ \sum_{k=0}^{2} p_{k1} & p_{11} & p_{21} & \sum_{k=0}^{2} p_{k1} & 0 \end{bmatrix} \begin{bmatrix} \mu \\ \rho_1 \\ \rho_2 \\ \kappa_1 \\ \kappa_2 \end{bmatrix} = \begin{bmatrix} f_1(0) \\ f_1(1) \\ f_1(2) \\ f_2(0) \\ f_2(1) \end{bmatrix}.$$

It is straightforward to establish that $\text{rank}(A_6) = 5$. Thus the solution to the reduced system of equations is unique and is given by $\underline{\widehat{\beta}} = A_6^{-1}\underline{b}_6$. 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 also 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 $\underline{\widehat{\beta}}$.