A SEMI-PARAMETRIC MODEL OF NON-IGNORABLE GENERALIZED ATTRITION*

İnsan Tunalı[†] Berk Yavuzoğlu[‡] Emre Ekinci[§]

April 5, 2021

Abstract

Repeated surveys that rely on a rotating sample frame contain a short panel component, which is prone to potentially non-ignorable non-response in the form of attrition and reverse attrition (initial non-response followed by response). We start with a characterization of the data generation process, remaining loyal to the rotation schedule, and establish how the joint distribution of the multi-valued discrete outcome of interest is affected by generalized attrition. We then propose a semi-parametric correction model where weights expressed as suitable functions of outcomes in both periods facilitate adjustments to the joint distribution, so that it is consistent with marginals obtained from unbiased cross-section data. We offer an empirical likelihood formulation that permits estimation of the weights and tests of ignorability. The linear version of our model has a closed form solution, a feature which renders our method computationally attractive. We illustrate the utility of our model using a labor market example, where the goal is estimation of transition probabilities between states (inactive, employed, unemployed) that are consistent with published official cross-section statistics.

Keywords: Attrition; reverse attrition; selective non-response; rotating panel; forward-looking panel; labor force survey.

*The funding for this project was provided by grant no. 109K504 by TUBITAK, The Scientific and Technological Research Council of Turkey. We are indebted to Hüseyin İkizler and Hayriye Özgül Özkan for research assistance. Discussions with Geert Ridder that prompted this line of research, feedback on earlier versions from John Kennan, Thierry Magnac, David Margolis, Christopher Taber and James Walker, comments of seminar and workshop participants at Bilkent, Koç, and Nazarbayev Universities, London School of Economics, Paris School of Economics are gratefully acknowledged.

[†]Corresponding author, Department of Economics, Koç and Boğaziçi University, Istanbul, and Economic Research Forum, Cairo; phone: +90-212-3381427; fax: +90-212-338-1653; e-mail: itunali@ku.edu.tr.

[‡]Department of Economics, Nazarbayev University, Nur-Sultan; e-mail: berk.yavuzoglu@nu.edu.kz.

[§]Department of Economics, Koç University, Istanbul; e-mail: emekinci@ku.edu.tr.

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 attrition problem is provided by Ridder and Moffitt (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., 1993). 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). Fitzgerald et al. (1998) situate these important ideas within a regression framework familiar to economists, by distinguishing between selection on observables and selection on unobservables, and clarifying the independence assumptions needed for identification.

In this paper, we tackle non-response problems that arise in the short-panel components of surveys relying on a rotating sample frame (also known as "rotating" panels, see Cantwell, 2008). Surveys with this feature – such as the Household Labor Force Survey (HLFS) in Turkey we work with, as well as widely used data sets such as the Current Population Survey (CPS) in the U.S., most country surveys included in the European Union Labor Force Survey (EU-LFS) and the European Union Statistics on Income and Living Conditions (EU-SILC) – call for repeat visits to the same address or dwelling according to a pre-determined schedule, but limit the maximum number of visits.[1] Residential addresses are visited according to the rotation schedule whether or not any respondents were found in the previous visit. Consequently, a rotating panel suffers not only from attrition (response followed by non-response) but also from "reverse attrition" (non-response followed by response). Reverse attrition oc-

[1]This yields a dynamic sample frame whereby a given round of the survey contains units that are at various points of their visit schedule. In each round a predetermined set of units visited for the last time are dropped (rotated "out") and replaced by a set of new randomly chosen units (rotated "in"), ensuring nationwide representation as well as regular updating. Standard cross-section non-response adjustments (based on demographics) are used to yield period specific marginal distributions, which may serve as the source of published official statistics.

curs because a subsequent visit sets the stage for the prospect of encountering returning attrited units, new units (in place of the old ones), or new individuals at a previously visited unit.

In the HLFS-Turkey 2000-2002 data we utilize, average quarterly attrition and reverse attrition rates are respectively 10.4 (max 14.5) and 11.5 (max 15.8) percent. Annually, while attrition rates average to 17.3 percent (max 19.7), reverse attrition rates average to 15.5 percent (max 16.2). These figures are by no means aberrations. Our calculations on the 2016-7 segment of the CPS reveal that attrition rates for visits that are three months apart average to 10.2 percent (max 11.6), while reverse attrition rates average to 13.1 percent (max 14.9). For visits that are 12 months apart, both attrition and reverse attrition rates are 23.5 percent on average, with maximums of around 24.5 percent. Nicoletti and Peracchi (2005) provide the survey participation patterns in the first five waves of the European Community Household Panel across seven countries. Among those who participated in at least one wave of the survey, the percent who reverse attrited at some point varies from 15.3 percent in Ireland to 21.7 percent in Portugal.

Our semi-parametric "Non-ignorable Generalized Atttrition" (NGA) model adjusts the balanced panel to account for the presence of "generalized attrition," be it in the conventional form of attrition, or reverse attrition, and contains improvements over existing approaches. In a rotating panel initial period outcomes are observed for attritors (but not for reverse attritors), whereas the subsequent period outcomes are observed for reverse attritors (but not for attritors). This requires a symmetric treatment of the periods and alters the perception of the nature of non-response present in forward looking panels that suffer only from attrition. Using the distinction drawn in Fitzgerald et al. (1998), in the context of a forward-looking panel it is possible to differentiate between selection on observables (non-ignorable non-response attributable to known pre-attrition outcomes) versus unobservables (non-ignorable non-response attributable to unknown pre-attrition outcomes). This distinction is apparently lost in a rotating panel. Our formulation allows us to test whether generalized attrition is ignorable with respect to the initial, or the subsequent period outcomes, or both. Put differently, using the established terminology of Little and Rubin, we can test whether a "missing at random" (MAR) assumption is realistic with respect to both observed and unobserved outcomes in either adjoining period. Since the tests are based on the weights attached to balanced panel cells, what the adjustments achieve and how the unadjusted versions mislead us, become transparent. Our methodology resolves the ambiguity over which set of weights are suitable for adjusting a balanced panel subjected to generalized

attrition.

NGA model shares some of the key ideas in Hirano et al. (2001) which addresses attrition problems in forward-looking panels. To address ignorability, Hirano et al. (2001) express the probability of attrition as an additive function of outcomes in both periods, ruling out their interactions, and use the inverted parametric attrition probabilities as weights. By equating the row and column sums of the reweighed balanced panel cell counts (or fractions) to the respective marginals, a just-identified system of equations is obtained. In a forward-looking panel, data collected in the initial period are not subjected to reverse attrition, so an unbiased estimate of the first round marginal is available. The unbiased marginal for the follow-up period comes from an independently conducted cross-section survey (what has been termed a "refreshment sample" by Ridder, 1992; an example of "external data" as defined by Ridder and Moffitt, 2007). Our correction scheme for generalized attrition preserves the reweighing logic in their paper, but differs in many other ways. To begin with, instead of specifying the functional form of the generalized attrition probability and inverting it to obtain the weights, we specify the weighting function directly, as a "suitable" one-to-one transformation of an additive index function of the outcomes in the initial and subsequent periods. Secondly, our estimation approach is distribution-free and does not require imputations of unobserved outcomes as in Hirano et al. (2001). Thirdly, we begin with a broader characterization of non-response (appropriate for rotating panels) before we discuss the independence assumptions needed for identification. Finally, we situate the estimation problem in an empirical likelihood framework and establish that conventional methods of inference can be used for tests of the ignorability of generalized attrition.

In the linear version of the NGA model the parameters of the weighting function can be estimated via a simple matrix manipulation. Since our treatment of exogenous variables is completely non-parametric, our estimation approach is well-suited for tackling the need to obtain dynamic estimates (joint or transition probabilities) consistent with the static cross-section estimates (marginal probabilities). The data requirements for the implementation of our methodology are extremely minimal: the joint frequency distribution obtained from the balanced panel along with the marginal frequency distributions obtained from the representative cross-sectional data in both periods. Naturally, our ability to compute proper adjustment weights hinges on the availability of unbiased estimates of the marginals. Since the data collection agency that conducts the repeated survey offers weights to render each cross-section nationally representative, this amounts to having unbiased estimates of the marginals for both periods. In our labor market application, these marginal distributions

are used for computing the official published statistics.

The existing literature that deals with reverse attrition is sparse. Das (2004) builds on Hausman and Wise (1979), and proposes a non-parametric selection model that allows attrition in one period to be followed by reappearance later, in a multi-period panel set-up. The connection between the outcome of interest and the attrition indicator is captured by correlated random effects included in the error terms of the two equation model. Under the assumption that probability of non-response depends solely on the outcome in that period, Fitzmaurice et al. (2005) develop a pseudo-likelihood estimator for binary response models for panel data that are subject to non-monotone missingness (what we term generalized attrition). Deng et al. (2013) build on Hirano et al. (2001) and take into account what they call non-terminal attrition, non-response that occurs in the second period of a three-wave forward-looking panel followed by a response in the third period.[2]

A third type of non-response can occur when a unit designated for the rotating sample frame is unobserved in both periods. While survey statisticians painstakingly differentiate between different versions (Clarke and Tate, 1999) and try to document them (BLS, 2019, Ch. 3-2; Cantwell, 2008), the usual practice is to treat *non-participation* (in the survey) as being ignorable. Although we stick to this practice, we clarify the assumptions that justify this approach and address the consequences of its violation. We remain loyal to the logic of data collection subject to a rotating sampling frame, namely use the rotation schedule to distinguish between intended and unintended non-response, explicitly state the independence assumptions that are needed for identification, and address their suitability.

We illustrate the NGA model in the context of a labor market application, where the objective is the estimation of transition rates between labor market states (employment, unemployment, non-participation) that are consistent with official labor market statistics. The findings from our empirical work establish that both attrition and reverse attrition are non-ignorable. Simpler models nested under ours, whether attrition is ignorable with respect to the first or the second period outcomes, are handily rejected. The magnitudes

[2]To the best of our knowledge, the term *reverse attrition* was first used by Gruber (1997, S93, fn. 23) to acknowledge a potential limitation of his analysis due to omission of newly-formed firms in the second period. In a similar vein, Alderman et al. (2001, 116, note 2; 118, note 10) used it to refer to the respondents who were present in the second round but not the first round of a one-time survey. The terminology gained recognition over time (e.g., Chetty and Saez, 2004; Kazianga et al., 2014; Bigelow et al., 2017; Bigelow and Plantinga, 2017). Some recent papers use alternate terminology to describe the same phenomenon: Xie and Qian (2012) use "intermittent attrition," Hawkes and Plewis (2006) use "wave non-response," Longford et al. (2006), and Chaudhuri and Guilkey (2016) use "non-monotone non-response."

of the estimated adjustment weights underscore the perils of working with the unadjusted balanced panel. Notably the adjustments to the balanced panel are found to be robust to alternate transformations of the linear index function proposed by Chen (2001). This establishes that the linear NGA model is adequate for rendering the short panel dimension of widely available survey data usable.

The idea of reconciling observed flow data between states with the cross sectional stocks via probabilistic adjustments expressed as a function of the states is present in earlier papers. Abowd and Zellner (1985) and Stasny (1986, 1988) work with counts obtained from short panels, and focus on estimating the sizes of gross flows from one period to another. The contrasts between their approaches and our NGA model will be taken up in Section 3.3. 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. The idea of imposing moment conditions obtained from external data to render attrition-prone panel data usable is also present in Hellerstein and Imbens (1999). In fact all these approaches can be situated within a broader framework directed at reconciling key statistical features of incomplete survey data with what is known about the population (Little, 1993). The model based adjustment proposed in Little and Wu (1991) echoes the fundamental ideas exploited in our NGA model.

We begin our formal treatment in Section 2 by introducing our conceptualization of the generalized attrition process and derive the NGA model that renders the balanced panel usable. 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 empirical work on labor market dynamics that illustrates the utility and simplicity of the proposed approach. We conclude the paper with a brief summary of the key aspects of our model and its potential uses.

2 NGA Model

The context of our model is a repeated survey directed to units (typically households) which utilizes a rotational design, whereby each unit remains in the sample frame for a predetermined number of periods. Survey statisticians underscore several advantages: Firstly, a repeated visit to the same household allows tracking of 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 (Cantwell, 2008). Given these advantages, 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 the short panel embedded in a repeated survey is often not fully exploited for want of weighting schemes consistent with marginals 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 units intended for repeat visits, and the balanced panel (BP), which only includes units who have been successfully interviewed in both periods. We also keep track of units which are rotated out of the sample after period 1, and units which are rotated in during period 2. Finally, for the sake of completeness we allow for *non-participants*, defined as units which have been selected for inclusion in the sample frame, but never participated in the survey.

The objects of the data collection effort may be classified as 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 endogenous outcome variables may be discrete, or continuous. Our substantive application involves discrete outcomes, in which case the joint distribution classifies individuals of a given type (x) according to a pair of multi-valued discrete outcomes (y_1, y_2). In our example, y denotes labor market states. The primary objective of the statistical agency is to produce period-specific statistical indicators based on y – such as labor force participation rate, unemployment rate, etc. – conditional on x. Since y and x serve as adequate identifiers of differences across individuals, in what follows we suppress the observation subscript. We use subscripts to denote period-specific values of y, and treat x as time invariant. The joint distribution of interest is $f(y_1, y_2|x)$.

2.1 Assumptions

Our first task is to offer a characterization of the data generation process that exposes how generalized attrition affects the balanced panel, the short panel component that captures dynamics. As we proceed we also state and clarify the assumptions that the NGA model rests on. We begin by defining 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 probabilities with the outcomes, and state the independence assumptions that enable us to examine the impact of generalized attrition formally. The first random variable captures the rotation status of the address:

$$R = \begin{cases} 1 \text{ if designated for out-rotation (w/prob.} = \delta_1) \\ 2 \text{ if designated for in-rotation (w/prob.} = \delta_2) \\ 3 \text{ if designated for the CP (w/prob.} = \delta_3 = 1 - \delta_1 - \delta_2) \end{cases} . \tag{2.1}$$

Units with $R = 1$ are those designated for their last visit in period 1 while those with $R = 2$ are designated for their first visit in period 2. Units with $R = 3$ are those who have been assigned to the complete panel (CP). This last group is the target of dynamic analyses.

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

$$S = \begin{cases} 1 \text{ if at least one interview took place (w/prob.} = \phi) \\ 0 \text{ if not (w/prob.} = 1 - \phi) \end{cases} . \tag{2.2}$$

Assumption 1: R, (y_1, y_2), and S are mutually independent.

This assumption implies pairwise independence, i.e. (i) $R \perp (y_1, y_2)$, (ii) $S \perp R$, (iii) $S \perp (y_1, y_2)$. Viewing the components in turn: (i) Since rotation status is predetermined, the independence assumption between R and (y_1, y_2) is non-controversial. (ii) By ruling out dependence between S and rotation status R, we disallow selective participation in the survey because of the differential interview burden involved. (iii) The independence assumption between S and (y_1, y_2) underscores the distinction between selective non-response that we have to acknowledge (by virtue of having seen the unit at least once) and *non-participation* we treat as being ignorable, the usual practice in the survey literature. In Section 2.5 we show that this assumption is innocuous, in that it does not rule out non-ignorable attrition in the balanced panel providing an interview took place ($S = 1$).

While R is an *ex ante* construct that captures the assigned participation status of a unit in the sample frame, S is an *ex post* construct that indicates actual participation in the data collection effort. Clearly only units with $R = 3, S = 1$ have the potential to contribute to the identification of the joint distribution, $f(y_1, y_2|x)$. Occurrence 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:

$$A = \begin{cases} 1 \text{ if observed in the } 1^{st} \text{ period only (attrited w/prob.} = \gamma_1) \\ 2 \text{ if observed in the } 2^{nd} \text{ period only (reverse attrited w/prob.} = \gamma_2) \\ 3 \text{ if observed in both periods (w/prob.} = \gamma_3 = 1 - \gamma_1 - \gamma_2) \end{cases}, \text{given } R = 3, S = 1. \tag{2.3}$$

Random variable A captures the possibly selective response status of participants among the $R = 3, S = 1$ group during the observation window. The subsample with $A = 3$ constitutes the balanced panel (BP). By virtue of being present either in the first or the second period, those with $A = 1$ (attrited unit) or $A = 2$ (reverse attrited unit) make contributions to the marginal distribution of that period. The presence of the $A = 2$ group is the distinguishing feature of rotating panels. Additional contributions to the marginals come from participants not designated for the complete panel ($R = 1$ or 2) with whom the intended interview took place ($S = 1$).

Assumption 2: External data on unbiased marginal distributions of interest, $f_1^*(y_1|x)$ and $f_2^*(y_2|x)$, are available.

This is a key assumption in recovering the joint distribution of interest $f(y_1, y_2|x)$ from the possibly biased estimates $f(y_1, y_2, A = 3|x)$ obtained from the BP. In Section 2.3, we discuss different practical ways of using the data collected in a typical rotating panel to produce the marginals.

2.2 Identification problem

In this subsection we suppress the conditioning on x for brevity, and use the random variables R and S to index the respective supports. With these simplifications at hand, we express the joint distribution of interest as

$$f(y_1, y_2) = \Sigma_R \Sigma_S f(y_1, y_2, R, S), \tag{2.4}$$

and analyze the components one by one. We begin with non-participants, $S = 0$. Here and below we use Bayes' Theorem to isolate the joint distribution function of interest and then

simplify the expressions step by step, by imposing Assumption 1.

$$\begin{aligned} f(y_1, y_2, R = r, S = 0) &= \Pr(R = r, S = 0|y_1, y_2) f(y_1, y_2) \\ &= \Pr(R = r, S = 0) f(y_1, y_2) \\ &= \Pr(R = r)\Pr(S = 0) f(y_1, y_2) \\ &= \delta_r (1 - \phi) f(y_1, y_2), r = 1, 2, 3. \end{aligned} \tag{2.5}$$

We turn to participants ($S = 1$) next, and examine those who were not designated for the complete panel ($R = 1$ or 2). These units consist of those who were rotated out, and those who were rotated in.

$$\begin{aligned} f(y_1, y_2, R = r, S = 1) &= \Pr(R = r|y_1, y_2, S = 1) f(y_1, y_2, S = 1) \\ &= \Pr(R = r|y_1, y_2, S = 1) \Pr(S = 1|y_1, y_2) f(y_1, y_2) \\ &= \Pr(R = r) \Pr(S = 1) f(y_1, y_2) \\ &= \delta_r \phi f(y_1, y_2), r = 1, 2. \end{aligned} \tag{2.6}$$

To obtain the final form of the expressions given in equations (2.5) and (2.6), we used the notation adopted in equations (2.1) and (2.2).

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

$$f(y_1, y_2, R = 3, S = 1) = \Sigma_A f(y_1, y_2, R = 3, S = 1, A) \tag{2.7}$$

For subgroups $A = 1, 2$ we get:

$$\begin{aligned} f(y_1, y_2, R = 3, S = 1, A = a) &= \Pr(A = a|y_1, y_2, R = 3, S = 1) f(y_1, y_2, R = 3, S = 1) \\ &= \Pr(A = a|y_1, y_2, R = 3, S = 1) \Pr(R = 3, S = 1|y_1, y_2) f(y_1, y_2) \\ &= \Pr(A = a|y_1, y_2, R = 3, S = 1) \Pr(R = 3, S = 1) f(y_1, y_2) \\ &= \Pr(A = a|y_1, y_2, R = 3, S = 1) \Pr(R = 3) \Pr(S = 1) f(y_1, y_2) \\ &= \Pr(A = a|y_1, y_2) \delta_3 \phi f(y_1, y_2), a = 1, 2. \end{aligned} \tag{2.8}$$

To obtain the last line in equation (2.8), we used the notation introduced in equations (2.1) and (2.2) together with the fact that A denotes mutually exclusive subsets of $\{R = 3, S = 1\}$. Note that $\Pr(A = 2|y_1, y_2)$ captures the influence of non-ignorable non-response in the first period (reverse attrition) and as such plays a key role for the identification of the NGA model compared to model based approaches that focus on attrition in forward looking panels.

Turning to the $A = 3$ subgroup, we proceed in similar fashion, albeit with a different set of conditioning arguments:

$$\begin{aligned} f(y_1, y_2, R = 3, S = 1, A = 3) &= f(y_1, y_2 | R = 3, S = 1, A = 3) \Pr(R = 3, S = 1, A = 3) \\ &= f(y_1, y_2 | R = 3, S = 1, A = 3) \Pr(A = 3 | R = 3, S = 1) \\ &\quad \times \Pr(R = 3, S = 1) \\ &= f(y_1, y_2 | A = 3) \gamma_3 \delta_3 \phi. \end{aligned} \tag{2.9}$$

It is straightforward to see that $f(y_1, y_2|A = 3)$ can be identified non-parametrically from the balanced panel. Since the balanced panel consists of the subset of individuals who have been subjected to attrition or reverse attrition, in general $f(y_1, y_2|A = 3) \neq f(y_1, y_2)$.

Substitution of the terms we derived – via the manipulations in equations (2.5), (2.6), (2.8), and (2.9) – for the components on the right hand side of equation (2.4) yields:

$$\begin{aligned} f(y_1, y_2) &= \delta_1(1-\phi) f(y_1, y_2) + \delta_2(1-\phi) f(y_1, y_2) + \delta_3(1-\phi) f(y_1, y_2) \\ &\quad + \delta_1 \phi f(y_1, y_2) + \delta_2 \phi f(y_1, y_2) + Pr(A = 1|y_1, y_2) \delta_3 \phi f(y_1, y_2) \\ &\quad + Pr(A = 2|y_1, y_2) \delta_3 \phi f(y_1, y_2) + f(y_1, y_2|A = 3) \gamma_3 \delta_3 \phi. \end{aligned} \tag{2.10}$$

Upon collecting terms, simplifying and rearranging we get

$$f(y_1, y_2) = \frac{f(y_1, y_2|A = 3)\gamma_3}{[1 - \Pr(A = 1|y_1, y_2) - \Pr(A = 2|y_1, y_2)]}. \tag{2.11}$$

Finally, using the fact that $\sum_A \Pr(A|y_1, y_2) = 1$, we get

$$f(y_1, y_2) = \frac{f(y_1, y_2|A = 3)\gamma_3}{\Pr(A = 3|y_1, y_2)}. \tag{2.12}$$

The last equation looks like the key equation of the AN model of Hirano et al. (2001, p.1647), except our balanced panel also suffers from non-ignorable non-response in the first period (reverse attrition). Hirano et al. (2001) non-parametrically identify their version of γ_3, the fraction of retained individuals when the sample is only subject to attrition, then specify the probability in the denominator of their version of the equation (2.12) as a parametric function of (y_1, y_2). We follow a different strategy and treat γ_3 as a nuisance parameter.[3] We then *rescale* the probability in the denominator of the equation (2.12) and obtain:

$$f(y_1, y_2) = w(y_1, y_2) f(y_1, y_2|A = 3), \tag{2.13}$$

[3] As we establish in our empirical likelihood framework, the component that contains $\gamma_3 = \Pr(A = 3|R = 3, S = 1)$ becomes separable.

where $w(y_1, y_2) = \gamma_3 / \Pr(A = 3|y_1, y_2) > 0$ by construction. Balanced panel fractions non-parametrically identify $f(y_1, y_2|A = 3)$. NGA model will emerge fully in the next section when we parameterize $w(y_1, y_2)$ and establish the identification of $f(y_1, y_2)$. Additional restrictions on $w(y_1, y_2)$ needed for practical implementation are taken up in the empirical section.

2.3 Identification using external data

Equation (2.13) 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 pave the way for estimation, we mimic the approach in Hirano et al. (2001). For a given x, we express $w(y_1, y_2|x)$ as a suitable transformation of an additive linear index function of the endogenous outcomes in both periods, $i\{\beta|y_1, y_2, x\}$, where β denotes the unknown parameter vector.

Until now our treatment of $f(y_1, y_2)$ and derivation of the key equation (2.13) has been general. Since our substantive application involves discrete outcomes, we supply the details for that case. Clearly continuous variables can be subsumed within our framework by breaking them into mutually exclusive ranges, then assigning discrete labels to them. In fact, an unknown continuous distribution will be approximated by a discrete distribution in a typical application. Suppose y has K distinct values (or ranges, if continuous). Exploiting the constraints that link the balanced panel with externally obtained marginals, $f_1^*(y_1|x)$ and $f_2^*(y_2|x)$, and restoring the conditioning on covariates x, we obtain:

$$\sum_{y_2} f(y_1, y_2|x) = \sum_{y_2} w(i\{\beta|y_1, y_2, x\}) f(y_1, y_2|A = 3, x) = f_1^*(y_1|x), \tag{2.14}$$

$$\sum_{y_1} f(y_1, y_2|x) = \sum_{y_1} w(i\{\beta|y_1, y_2, x\}) f(y_1, y_2|A = 3, x) = f_2^*(y_2|x). \tag{2.15}$$

Equations (2.14) and (2.15) provide the restrictions that must be satisfied by the reflated balanced panel fractions where $w(i\{\beta|y_1, y_2, x\})'s$ 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 can have at most $2K - 1$ unknown parameters. We take one (y_1, y_2) combination as the reference group and express $i\{\beta|y_1, y_2, x\}$ as a linear function of $2K - 1$ main effects, so that $\dim(\beta) = 2K - 1$. To assess the role of our parametric assumptions further, we follow Chen (2001) and entertain three different one-to-one transforms of the index function, respectively the identity function (which we term the linear

NGA model), plus suitable convex and concave functions. Details will emerge in Section 3. Since additivity in main effects is assumed, the identification proof in Hirano et al. (2001), as well as the simpler version in in Bhattacharya (2008) still apply. Nonetheless, in the Appendix, we provide another proof in the context of the linear NGA model.

In a short panel data collection effort that relies on a rotating sample frame, both margins have to be estimated with the help of external data. Conveniently the rotating sample frame that supports the short panel also provides additional information on the marginal distributions. These come from two sources: units which are rotated out, and units which are rotated in. In the HLFS-Turkey – which calls for four visits to an address over a period of 18 months – units subjected to rotation constitute about one-half of all units interviewed in a given cross-section. Technically units rotated in for the first time (about a quarter of the full sample) constitute a refreshment sample, so unbiased estimates of the period-specific marginals can be obtained (Ekinci, 2007). Since our ultimate objective is to produce transition estimates consistent with the published labor market statistics (namely the period specific labor force participation rate, employment and unemployment rates), we do not pursue that route. Indeed, data collection agencies (BLS, EUROSTAT, in our case TURKSTAT) use all the cross-section data to arrive at the official statistics. Thus in our labor market example the marginal distributions we rely on are the (properly weighted) cross-sectional statistics published by TURKSTAT. The point of Assumption 2 is to emphasize the need to use data other than what is available in the balanced panel.

2.4 Tests of ignorability of attrition

In a forward-looking panel which is subjected to attrition in the second round, first period outcomes are always observed while second period outcomes are unobserved for attritors. As a result, one can conveniently test whether attrition behavior can be viewed as selection on observables, or unobservables (using the distinction drawn in Fitzgerald et al., 1998), rather than both. This mapping between periods and observation status does not apply to the short panel component of a survey that relies on a rotational design, the case we study. In our more general setting first period outcomes which are observed for attritors are unobserved for reverse attritors, whereas second period outcomes which are unobserved for attritors are observed for reverse attritors. In the NGA model it is straightforward to test whether attrition is ignorable with respect to the first, or the second period outcomes, but the test outcomes do not provide information on whether observables or unobservables are at work.

For brevity we suppress the conditioning on x, and examine in turn the restrictions on the NGA model weights $w(y_1, y_2)$ that produce special models of attrition and reverse attrition. We also link these with earlier models.

(a) 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).

(b) If non-response is a function of observed outcomes only, $w(y_1, y_2) = w(y_1)$ for attritors, and $w(y_1, y_2) = w(y_2)$ for reverse attritors. Using the partition given in equation (2.3), we can express the restriction as $w(y_1, y_2) = \alpha * w(y_1) + (1 - \alpha) * w(y_2)$, where α denotes the share of attritors among the set of attritors ($A = 1$) and reverse-attritors ($A = 2$).

In the context of a regular panel that is subjected to attrition but not reverse attrition, $\alpha = 1$ and $w(y_1, y_2) = w(y_1)$. That is, weights are expressed solely as a function of observed first period outcomes. This is the case popularized by Little and Rubin (1987), and has been dubbed Missing at Random (MAR). In a short panel context it would seem that a similar logic can be applied to reverse-attritors, using the observed outcomes for the second period. Unfortunately in a K state NGA model, this would imply $2K$ parameters, one more than what can be identified.

(c) If non-response is a function of unobserved outcomes only, $w(y_1, y_2) = w(y_2)$ for attritors, and $w(y_1, y_2) = w(y_1)$ for reverse-attritors. Using the notation in (b), we can express the restriction as $w(y_1, y_2) = (1 - \alpha) * w(y_1) + \alpha * w(y_2)$. In a regular panel without reverse attrition, $w(y_1, y_2) = w(y_2)$. That is, weights are a function of unobserved second period outcomes only. Hirano et al. (2001) call this the Hausman and Wise (HW) model because a correction based on the unobserved second period outcomes was first proposed by Hausman and Wise (1979), in the context of a two-period forward-looking panel. In a short panel context, reverse-attritors are unobserved in the first period. As in case (b) the correction logic can be extended to reverse attritors, but the model yields one more parameter than what can be identified using the $2K - 1$ available restrictions.

Strictly speaking neither Hausman and Wise (1979) nor Little and Rubin (1987) address the problem of identification of the joint distribution. Fitzgerald et al. (1998) contrast the two approaches (HW and MAR) using selection terminology popular among economists. They point out that while selection in the MAR model is on (first period) observables, selection in the HW model is on unobservables that include second period outcomes. As our discussion under (b) and (c) shows, if the observable/unobservable distinction is applied to the characterization of attrition and reverse attrition behavior encountered in a short panel context, the rigid mapping between periods and observability, present in HW, MAR and

consequently in Hirano et al. (2001), cannot be sustained. Furthermore it is not feasible to estimate these models.

Upon dropping the observable/unobservable distinction, we obtain two models nested under the NGA model that can be estimated:

(d) If non-response is a function of first period outcomes only, $w(y_1, y_2) = w(y_1)$ for both attritors and reverse attritors.

(e) If non-response is a function of second period outcomes only, $w(y_1, y_2) = w(y_2)$ for both attritors and reverse attritors.

What remains to be done is to attach labels to approaches (d) and (e). Taking cue from Little and Rubin (1987), these respectively assume that non-response is ignorable with respect to period 2 and period 1 outcomes. We therefore call them MAR2 and MAR1, using the convention that selection is assumed to be ignorable with respect to the period indexed by the suffix.

2.5 $S \perp (y_1, y_2)$ revisited

Before we proceed with a detailed examination of our estimation procedure, we return to our derivations and offer some observations about the role of the independence assumption between S and (y_1, y_2), implied by Assumption 1. The derivations in Section 2.2 reveal a remarkable difference in the handling of units designed for the complete panel and the rest. While the terms that rescale $f(y_1, y_2)$ in equations (2.5) and (2.6) are exogenous probabilities, in equation (2.8) endogenous probabilities are present. Given the partition in equation (2.3), the statement that attrition is ignorable amounts to $\Pr(A = 1|y_1, y_2) = \gamma_1$, a constant. Likewise the statement that reverse attrition is ignorable amounts to $\Pr(A = 2|y_1, y_2) = \gamma_2$. If we were to apply this language for the other designations, we have essentially assumed that rotation status (given in equation (2.1)) and interview status (survey non-participation, given by equation (2.2)) are ignorable. Arguably the only potentially controversial assumption we make – which is also the usual practice in the attrition correction literature – is the ignorability of survey non-participation (units that were selected for inclusion in the rotating sample frame, but did not participate in both periods). Note, however, that $\phi = Pr(S = 1)$ cancels out during the algebraic manipulations that led to equation (2.12). Even if we were to assume non-ignorable non-participation, that is let $Pr(S = 1|y_1, y_2) = \phi(y_1, y_2)$ in equation (2.6), this term would drop out as we move from equation (2.10) to (2.11). Unlike attritors and reverse attritors, survey non-participants do not make any contribution whatsoever to the data collection effort – either in the first period, or in the second period.

As such, survey non-participants 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 non-participation.

From a practical point of view, random variable S keeps track of practical survey implementation problems. These typically include (a) encountering the wrong unit (for example, an establishment rather than a household) at the address, (b) inability to contact the unit in any round, and (c) refusal of participation in the survey by the unit (Clarke and Tate, 1999). Based on information obtained from the data collection agency, non-response of type (c), refusal to participate, is uncommon in the HLFS-Turkey, attributable to the the law that obliges participation in official surveys. Non-response due to (a) and (b) are more common. If (a) occurs during the initial visit, the address is simply dropped from the sample frame. The most frequently recorded reason for type (b) non-response is "the household no longer resides at this address."

3 Estimation and Inference in NGA Model

We illustrate the utility of the NGA 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 (2.14)-(2.15) yields five independent equations, so we can estimate up to 5 parameters. We express $w(y_1, y_2|x)$ as a 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:

$$i(y_1, y_2|x) = \mu + \rho_1 I(y_1 = 1) + \rho_2 I(y_1 = 2) + \kappa_1 I(y_2 = 1) + \kappa_2 I(y_2 = 2) \equiv i(\boldsymbol{\beta}|y_1, y_2, x), \quad (3.1)$$

where $I(.)$ denotes the indicator function and $\boldsymbol{\beta} = [\mu\ \rho_1\ \rho_2\ \kappa_1\ \kappa_2]'$. This additive function of the unknown parameters captures the dependency of non-response attributable to attrition and reverse attrition on the labor market states (y_1, y_2). As in Hirano et al. (2001), 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: (a) linear: $w_L(y_1, y_2|x) = i(\boldsymbol{\beta}|y_1, y_2, x)$, (b) convex: $w_X(y_1, y_2|x) = \exp\{i(\boldsymbol{\beta}|y_1, y_2, x)\}$, and (c) concave: $w_E(y_1, y_2|x) = 2 - \exp\{i(\boldsymbol{\beta}|y_1, y_2, x)\}$. 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 non-linear 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 (2.14)-(2.15) can be represented as in Table 1. Here $p_{jk} = f(y_1 = j, y_2 = k|C = 3, x),\ j, k = 0, 1, 2$. The task amounts to

finding the reflation factors (functions of $\boldsymbol{\beta}$) that adjust the balanced panel fractions – so that the adjusted cell probabilities are in line with the marginals obtained externally. Ekinci (2007) used the subsamples from the two cross-sections, namely the units that were rotated in. 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 where the marginal distributions are obtained by a MAR type weighting scheme.

Table 1: A 3×3 Linear NGA 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

Let $p_{j\bullet} = \sum_{k=0}^{2} p_{jk}$, $j = 0, 1, 2$ and $p_{\bullet k} = \sum_{j=0}^{2} p_{jk}$, $k = 0, 1, 2$. It can be shown that this system of equations is observationally equivalent to the representation given below:

$$\begin{bmatrix} p_{0\bullet} & 0 & 0 & p_{01} & p_{02} \\ p_{1\bullet} & p_{1\bullet} & 0 & p_{11} & p_{12} \\ p_{2\bullet} & 0 & p_{2\bullet} & p_{21} & p_{22} \\ p_{\bullet 0} & p_{10} & p_{20} & 0 & 0 \\ p_{\bullet 1} & p_{11} & p_{21} & p_{\bullet 1} & 0 \\ p_{\bullet 2} & p_{12} & p_{22} & 0 & p_{\bullet 2} \end{bmatrix} \begin{bmatrix} \mu \\ \rho_1 \\ \rho_2 \\ \kappa_1 \\ \kappa_2 \end{bmatrix} = \begin{bmatrix} f_1^*(0) \\ f_1^*(1) \\ f_1^*(2) \\ f_2^*(0) \\ f_2^*(1) \\ f_2^*(2) \end{bmatrix} \tag{3.2}$$

Inspection reveals that this six-equation system is of the form $\boldsymbol{A\beta} = \boldsymbol{b}$ where $rank(\boldsymbol{A} : \boldsymbol{b}) = 5$. One of the constraints is redundant, in the sense that it is 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 start with by a simple pivoting operation. Consequently, the solution to the reduced five-equation system is unique and does not depend on which constraint is left out. If we were to exclude the last constraint, we would obtain the five-equation system which can be represented in matrix notation as $\boldsymbol{A_6\beta} = \boldsymbol{b_6}$, where subscripts denote the fact that the 6th constraint has been excluded. Written explicitly we

get equation (3.3).

$$\begin{bmatrix} p_{0\bullet} & 0 & 0 & p_{01} & p_{02} \\ p_{1\bullet} & p_{1\bullet} & 0 & p_{11} & p_{12} \\ p_{2\bullet} & 0 & p_{2\bullet} & p_{21} & p_{22} \\ p_{\bullet 0} & p_{10} & p_{20} & 0 & 0 \\ p_{\bullet 1} & p_{11} & p_{21} & p_{\bullet 1} & 0 \end{bmatrix} \begin{bmatrix} \mu \\ \rho_1 \\ \rho_2 \\ \kappa_1 \\ \kappa_2 \end{bmatrix} = \begin{bmatrix} f_1^*(0) \\ f_1^*(1) \\ f_1^*(2) \\ f_2^*(0) \\ f_2^*(1) \end{bmatrix}. \tag{3.3}$$

The unique solution to this just-identified system is $\hat{\boldsymbol{\beta}} = \boldsymbol{A_6^{-1}} \boldsymbol{b_6}$. While a closed form solution is available for the linear version, this is not the case when non-linear transforms of the index function are used. It is possible to obtain numerical solutions as long as the transform is one-to-one.

3.1 Empirical Likelihood

Although we established that the linear NGA model has an exact solution, derivation of the asymptotic covariance matrix of the estimated parameters requires additional work. Since the maximum likelihood (ML) approach has the advantage of producing a consistent estimate of this matrix, we employ 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 an empirical probability mass function. Towards that end, we first reparameterize the cell probabilities as shown in Table 2.

Table 2: Reparameterized 3×3 Linear NGA Model

	$y_2 = 0$	$y_2 = 1$	$y_2 = 2$	Row sum
$y_1 = 0$	θ_{00}	θ_{01}	θ_{02}	$f_1^*(0)$
$y_1 = 1$	θ_{10}	θ_{11}	θ_{12}	$f_1^*(1)$
$y_1 = 2$	θ_{20}	θ_{21}	θ_{22}	$f_1^*(2)$
Col. sum	$f_2^*(0)$	$f_2^*(1)$	$f_2^*(2)$	1

Next, 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 empirical likelihood function for the linear version may be expressed as:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i,j=0,1,2} \{\theta_{ij}\}^{n_{ij}}. \tag{3.4}$$

Maximization will be done subject to the adding up constraints, equations (2.14)-(2.15), which together with equation (3.1) imply equation (3.3). The standard approach would entail embedding an appropriate Lagrangian function in the log-likelihood function which may be expressed as:

$$\begin{aligned} ln\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\lambda}) &= \sum_{i,j=0,1,2} n_{ij} ln\{\theta_{ij}\} - \lambda_1\left[(\theta_{00}+\theta_{01}+\theta_{02}) - f_1^*(0)\right] \\ &\quad -\lambda_2\left[(\theta_{10}+\theta_{11}+\theta_{12}) - f_1^*(1)\right] - \lambda_3\left[(\theta_{20}+\theta_{21}+\theta_{22}) - f_1^*(2)\right] \\ &\quad -\lambda_4\left[(\theta_{00}+\theta_{10}+\theta_{20}) - f_2^*(0)\right] - \lambda_5\left[(\theta_{01}+\theta_{11}+\theta_{21}) - f_2^*(1)\right]. \end{aligned} \tag{3.5}$$

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

$$\boldsymbol{B\lambda} = \boldsymbol{Cd}(\boldsymbol{\beta}), \tag{3.6}$$

where $\boldsymbol{\lambda}$ denotes the 5×1 vector of Lagrange multipliers,

$$\boldsymbol{B} = \begin{bmatrix} p_{0\bullet} & p_{1\bullet} & p_{2\bullet} & p_{\bullet 0} & p_{\bullet 1} \\ 0 & p_{1\bullet} & 0 & p_{10} & p_{11} \\ 0 & 0 & p_{2\bullet} & p_{20} & p_{21} \\ p_{01} & p_{11} & p_{21} & 0 & p_{\bullet 1} \\ p_{02} & p_{12} & p_{22} & 0 & 0 \end{bmatrix}, \tag{3.7}$$

$$\boldsymbol{C} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}, \tag{3.8}$$

and

$$\boldsymbol{d}(\boldsymbol{\beta}) = \begin{bmatrix} \mu^{-1}n_{00} \\ (\mu+\kappa_1)^{-1}n_{01} \\ (\mu+\kappa_2)^{-1}n_{02} \\ (\mu+\rho_1)^{-1}n_{10} \\ (\mu+\rho_1+\kappa_1)^{-1}n_{11} \\ (\mu+\rho_1+\kappa_2)^{-1}n_{12} \\ (\mu+\rho_2)^{-1}n_{20} \\ (\mu+\rho_2+\kappa_1)^{-1}n_{21} \\ (\mu+\rho_2+\kappa_2)^{-1}n_{22} \end{bmatrix}. \tag{3.9}$$

Equation (3.9) serves to illustrate the key implication of our *identifying assumption*: the 9×1 vector $\boldsymbol{\theta}$ that adjusts the cell probabilities is a function of the 5×1 unknown parameter vector $\boldsymbol{\beta}$. It is straightforward to show that the 5×5 matrix $\boldsymbol{B}$ – 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 $\boldsymbol{A_6}$ defined above, the square matrix in the reduced equation system (3.3). This establishes that the log-likelihood function can be concentrated using

$$\boldsymbol{\lambda} = \boldsymbol{B^{-1}Cd(\beta)}. \tag{3.10}$$

Thus the 10 parameter constrained optimization problem (in terms of unknowns $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$) 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, $\boldsymbol{\beta}$.

3.2 Generalization

Generalization to K categorical outcomes is straightforward. Let $f_{jk} = f(y_1 = j, y_2 = k)$, $p_{jk} = f(y_1 = j, y_2 = k|A = 3, x)$, and $w(y_1, y_2) = w_{jk}$ with $j, k = 1, ..., K$. Equation (2.13) may be rewritten as:

$$\frac{f_{jk}}{p_{jk}} = w_{jk}. \tag{3.11}$$

Using the definition in equation (3.1), we may express the linear case as:

$$\frac{f_{jk}}{p_{jk}} = i(\boldsymbol{\beta}|y_1, y_2, x). \tag{3.12}$$

The convex version may be written as:

$$ln\left(\frac{f_{jk}}{p_{jk}}\right) = i(\boldsymbol{\beta}|y_1, y_2, x). \tag{3.13}$$

The concave version may be written as:

$$ln\left(2 - \frac{f_{jk}}{p_{jk}}\right) = i(\boldsymbol{\beta}|y_1, y_2, x). \tag{3.14}$$

The general form of the $2K \times (2K - 1)$ matrix $\boldsymbol{A}$, the $(2K - 1) \times 1$ vector $\boldsymbol{\beta}$, and the $2K \times 1$ vector $\boldsymbol{b}$ are easily discerned. We know that the linear case is additive in the unknown parameters, so the extension of the uniqueness proof given above for $K = 3$ is straightforward. In the non-linear versions, a known invertible monotonic function $h(.)$ of the ratio $\frac{f_{jk}}{p_{jk}}$ (where p_{jk} is known) is additive in the unknown parameters. This link between

the non-linear cases and the linear case suggests that the solutions to the non-linear cases are also unique. Clearly, systems (3.12), (3.13), and (3.14) will yield different estimates of the unknown $\boldsymbol{\beta}$. Note that ultimately the quantities of interest are not the $\boldsymbol{\beta}$'s but the weights used in rescaling, defined by $w_L(y_1, y_2|x) = i(\boldsymbol{\beta}|y_1, y_2, x)$, $w_X(y_1, y_2|x) = \exp\{i(\boldsymbol{\beta}|y_1, y_2, x)\}$, and $w_E(y_1, y_2|x) = 2 - \exp\{i(\boldsymbol{\beta}|y_1, y_2, x)\}$. Thus investigation of the sensitivity of the NGA model estimates to the parametric assumptions hinges on comparison of the $w_S(j, k|x)$, $j, k = 0, 1, .., K-1$ for $S = L,\ X,\ E$. It is straightforward to modify equation (3.5) and pursue maximum empirical likelihood estimation of the non-linear cases.[4]

3.3 Discussion

Apart from the 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.

At this point it is appropriate to provide an 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. These papers do not offer a formal model of the possibly non-ignorable non-response process. The goal is stated as estimating period-to-period gross flows – p_{jk}'s in our model. Abowd and Zellner (1985) use a multiplicative model to inflate the unadjusted proportions. The idea is that unmatched individuals who show up in one of the margins have some probability of being in a given cell of the joint distribution. The easiest way to relate their model to ours is to focus on equation (3.13) above. They essentially express the natural logarithm of the reflation factor defined in equation (3.11) above as a linear function of the natural logarithms of the counts of unmatched individuals. Using our own language to establish the links, unmatched individuals can either be attritors (observed only in the first period) or reverse attritors (observed only in the second period), plus an approximation error, ensuring that the adjusted cell proportions sum to one. Like us (see Section 4 below) they study three states (nine cells in the flow matrix), but estimate 18 unknown parameters subject to six adding up restrictions that link (the rows and columns of the matrix of) proportions with the respective margins. Thus, they not only allow interaction effects but also distinguish between attrition and reverse attrition parameters. Leaving the difference introduced by the use of counts aside, this would be equivalent to using an index function that exhausts all $K \times K$ cells via an indicator function

[4]The MATLAB code used in our empirical work is available from the authors upon request.

$I(y_1, y_2)$ and has separate parameters for attrition (ξ_a^{jk}) and reverse attrition (ξ_r^{jk}) on the right hand side of equation (3.13) in place of our own (3.1):

$$i(y_1, y_2|x) = \sum_{j=1}^{K}\sum_{k=1}^{K}(\xi_a^{jk} + \xi_r^{jk})I(y_1 = j,\ y_2 = k) \equiv i(\boldsymbol{\xi}|y_1, y_2, x). \tag{3.15}$$

Clearly this over-parameterized model cannot be used to implement separate adjustments for each period pair. Abowd and Zellner (1985) assume stationarity and use multiple rounds of the monthly CPS data to estimate "average" values of the parameters by minimizing the weighted squared deviation of the adjusted gross flow margins from the period specific CPS data.

Stasny (1986, 1988) has a similar set-up, except she uses additive models to implement the adjustment. According to her conceptualization, an observation designated for the two-period panel can lose either its column or row designation, with different probabilities, and show up in one of the margins. In terms of the distinction we draw, these are respectively attrition and reverse attrition probabilities. Thus Stasny's approach is equivalent to use of a more complicated linear function of the labor market states on the right hand side of equation (3.12). In fact her unconstrained model has the same number of free parameters as Abowd and Zellner (1985), so equation (3.15) can be used to capture the link with the NGA model. Unlike Abowd and Zellner (1985), Stasny (1986) shies away from a stationarity assumption and estimates different constrained models that can be identified with the available adding up constraints. In particular (like us) she sets the interaction effects to zero ($\xi_a^{jk} = \xi_r^{jk} = 0$ if $j \neq k$). In her richest (just identified) models she expresses the attrition and reverse attrition probabilities as functions of either the observed, or unobserved states. Since attritors are observed in the first, and reverse attritors are observed in the second period, in a given panel the sum of the two probabilities is able to capture dependence on states in both periods. In terms of the nesting designations given in Section 2.4, these are similar to models (b) and (c) which cannot be identified in the NGA model. Stasny is able to identify her version because she uses counts, and there are K restrictions to work with. Although the treatment of non-response in her just identified models has the same flavor as our NGA model, her models allow dependence either on observed, or on unobserved states; not both. Thus MAR1-MAR2 distinction cannot be drawn. Stasny estimates many two-period models on multiple rounds of data from the Canadian LFS and the CPS. Her empirical findings provide ample evidence against the stationarity assumption of Abowd and Zellner (1985).

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 meth-

ods used in reconciling them (Madow et al., 1993). 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 his paper is concerned with the case when the population joint distribution of the post-stratification variables is known. He 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 non-response 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/NGA approach. Instead of the additive model that drives the correction in AN/NGA models, Little and Wu (1991) have a multiplicative model.

In AN model applications reported in Hirano et al. (2001), 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 and Wu (1991). In our application of the NGA model, we proceed with the estimation of the reflation factors and the adjusted cell probabilities without engaging in computationally costly imputation. 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 et al. (1994). Their objective is to recover the elements of expenditure, trade, or income flows from limited or incomplete multisectoral economic data using a similar set of adding up restrictions. Recent papers framed within the attrition-refreshment sample framework include Nevo (2003), Bhattacharya (2008), Deng et al. (2013), and Hoonhout and Ridder (2019). 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 the AN model Hirano et al. (2001), Bhattacharya (2008) proposes a sieve-based estimation method and establishes the asymptotic properties of the estimator. As we noted earlier, Deng et al. (2013) extend the AN model to three wave forward-looking

panels with two refreshment samples. In the concluding section the authors offer a discussion on initial non-response, what we have termed non-participation to distinguish it from attrition and 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 in Section 2.5 sheds further light on the problem by separating what can, and cannot be modeled in a rotating panel context. Hoonhout and Ridder (2019) extend the AN model to multi period panels that suffer only from attrition. No useful insights emerge for relaxing the key AN/NGA identifying assumption, namely ruling out interaction effects.

4 A Labor Economics Application

Our example is a familiar one from Labor Economics: correction of transition rates obtained from balanced panels of the Household Labor Force Survey in Turkey (HLFS-Turkey). In Table 3, we compiled a set of ML parameter estimates from a $3x3$ NGA model for annual transitions. In this example, x denotes the entire working age population, aged 15 and over. The balanced panel contains over $20,000$ observations. The first and second period marginals in the raw data contain over $52,000$observations. Thus it is not surprising that all NGA model parameters are estimated extremely precisely.

As we noted earlier, the 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

Table 3: A 3×3 NGA Model - Parameter Estimates

Annual Transitions Between 2001-Q1 and 2002-Q1 x = age 15 and over					
Parameter	μ	ρ_1	ρ_2	κ_1	κ_2
$w(.)$ linear:					
Estimate	0.8987	0.0955	0.2510	0.1315	0.1794
Std. error	0.0084	0.0196	0.0490	0.0202	0.0391
$w(.)$ convex:					
Estimate	−0.1057	0.0956	0.2294	0.1293	0.1716
Std. error	0.0092	0.0192	0.0404	0.0195	0.0346
$w(.)$ concave:					
Estimate	0.0975	−0.0959	−0.2830	−0.1349	0.1902
Std. error	0.0076	0.0203	0.0635	0.0213	0.0456

we reported above, we see that the balanced panel sample amounted to about 40% of the respective marginals. The fact that this fraction is considerably lower than the expected 0.5 can be taken as a rough statistic that warns us about the magnitude of the attrition/reverse attrition problem. In fact attrition in the HLFS-Turkey is quite severe as documented by Tunalı (2009): Around 26% of eligible households and 32% of eligible individuals attrited sometime during the observation window over the period 2000-2002. For the subset of 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. What matters, of course, is whether the process that excludes individuals designated for the complete panel from the balanced panel is ignorable. However, Wald tests provide overwhelming evidence that the attrition and reverse attrition process is non-ignorable. Furthermore, alternatives to NGA model (MAR1 and MAR2) 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 utilizing the NGA 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 panel fraction. For each cell we report the estimates of the reflation

Table 4: A 3×3 NGA 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
$y_1 = 0$	0.8987 0.8997 0.8976	0.5052		1.0302 1.0239 1.0367	0.0566		1.0780 1.0681 1.0886	0.0159	$f_1^*(0)$
$y_1 = 1$	0.9942 0.9900 0.9984	0.0740		1.1257 1.1267 1.1248	0.2952		1.1736 1.1753 1.1719	0.0209	$f_1^*(1)$
$y_1 = 2$	1.1497 1.1316 1.1693	0.0113		1.2812 1.2879 1.2741	0.0122		1.3290 1.3434 1.3132	0.0085	$f_1^*(2)$
Col. sum	$f_2^*(0)$			$f_2^*(1)$			$f_2^*(2)$		1

factors, $w(.)$, associated with all three functional forms (respectively linear, convex, concave) inside braces. The findings from our sensitivity analysis are typical, in that functional form does not make much of a difference. The reflation factor estimates give us information about the direction of the bias in a given cell of the unadjusted balanced panel. For generalized attrition to be ignorable for a cell, its reflation factor estimate must be equal to 1. We find that the null is rejected for six of the nine estimated reflation factors, at the 5 percent level of significance. In the absence of correction, the non-participant to non-participant transition is overestimated, while the participant to the participant transitions (in particular, outcomes that involve unemployment) are underestimated. This is not surprising as it is more likely to find non-participants than employed, and employed than unemployed in their old addresses in subsequent visits.[5]

Table 5 provides the unadjusted joint probabilities and marginals obtained from the balanced panel (shown in brackets) along with the adjusted versions obtained from the linear NGA model. The magnitudes of the biases in the balanced panel [discrepancies between $f(y_1, y_2|A = 3, x)$ and $f(y_1, y_2|x)$] range between -25% and 11% percent. Six of the nice cells have biases of 10% or more in absolute value.

Table 5: A 3×3 NGA 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]	0.0246 [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]	0.0531 [0.0454]	1

[5]In our broader investigation we applied the three versions of the NGA model on estimated quarterly and annual transition data from the HLFS-Turkey and found similar patterns. We will be happy to share them with interested parties.

Table 6: A 3×3 NGA Model - Adjusted and [Unadjusted] Forward 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]	0.7719 [0.7565]	0.0571 [0.0537]	1 [1]
$y_1 = 2$	0.3249 [0.3525]	0.3916 [0.3813]	0.2836 [0.2662]	1 [1]

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 NGA model is likely to be especially important in the case of statistical agencies. The official position of TURKSTAT appears to be total neglect of the short panel dimension of the HLFS-Turkey micro data on the grounds that there is no weighting method that can reconcile dynamic and static estimates.

5 Conclusion

In this paper we tackle a generalized version of the attrition problem, typically associated with data from rotating panels. The motivation for taking a fresh look comes from the observation that many sustained large scale data collection efforts (the CPS, the EU-LFS, and the EU-SILC being some well-known examples) involve multiple visits to the same address/household over a fixed period (8 months in case of the CPS, up to four years in case of the EU-SILC). 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. Consequently these data sets have short panel components that support dynamic analyses. What stands in the way is the concern that the

balanced panel used for tracking the dynamics may not be representative of the population at any given point of time because of non-response after initial response (attrition) and response after initial non-response (reverse attrition). In fact, reverse attrition is observed to be as sizable as attrition in both the Household Labor Force Survey in Turkey we utilize, and the CPS.

We propose a bias correction framework for rotating panels and adopt an empirical likelihood approach that allows standard methods of inference. Our approach has several attractive features. Firstly, it generates transition estimates consistent with the period-specific marginals. Secondly, attrition behavior is allowed to be non-ignorable, in that it can depend on endogenous outcomes in either period. Thirdly, correction in our model resembles commonly used weighting methods in the literature on survey statistics. As such, it amounts to reflating the balanced panel fractions (cell means) by factors expressed as parametric functions of the states under examination. Fourthly, correction can be implemented conditional on exogenous observables, without imposing additional parametric assumptions. Finally, the parametric functions that link the weights with the states allow tests of simpler characterizations of generalized attrition.

In our empirical example, outcomes are labor market states occupied by an individual. Endogeneity arises because particular labor market outcome combinations could make individuals more or less prone to exclusion from the balanced panel. Our empirical investigation of annual transition data from the Household Labor Force Survey in Turkey shows that generalized attrition is a serious concern, in the sense that transition rates obtained from the balanced panel are systematically distorted. Simpler models nested under ours that assume attrition is completely ignorable, or ignorable either with respect to the first or the second period outcomes are handily rejected. Based on our systematic empirical investigation, results did not display sensitivity to the parametric features of the NGA model. Thus the linear version – which has a closed form solution and is extremely simple to implement – appears suitable for empirical work. Yet another attractive feature of the NGA model is the non-parametric treatment of covariates (such as sex, location, age groups, etc.). Each distinct covariate combination is associated with its own set of parameters and reflation factors.

In a nutshell, NGA 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 frequency

distribution obtained from the balanced panel that links the two legs of the short panel along with the marginal frequency 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.

References

Abowd, J. M. and A. Zellner (1985). Estimating gross labor-force flows. *Journal of Business and Economic Statistics 3*, 254–283.

Alderman, H., J. R. Behrman, H.-P. Kohler, J. A. Maluccio, and C. S. Watkins (2001). Attrition in longitudinal household survey data: Some tests for three developing country samples. *Demographic Research 5*, 79–124.

Bhattacharya, D. (2008). Inference in panel data models under attrition caused by unobservables. *Journal of Econometrics 144*, 430–446.

Bigelow, D. P. and A. J. Plantinga (2017). Town mouse and country mouse: Effects of urban growth controls on equilibrium sorting and land prices. *Regional Science and Urban Economics 65*, 104–115.

Bigelow, D. P., A. J. Plantinga, D. J. Lewis, and C. Langpap (2017). How does urbanization affect water withdrawals?: Insights from an econometric-based landscape simulation. *Land Economics 93*(3), 413–436.

BLS (2019). Current population survey: Design and methodology. Technical Paper 77, U.S. Bureau of Labor Statistics and U.S. Census Bureau, Washington, DC.

Cantwell, P. J. (2008). Rotating panel design. In P. J. Lavrakas (Ed.), *Encyclopedia of Survey Research Methods*, pp. 772–775. Thousand Oaks, CA: SAGE Publications.

Chaudhuri, S. and D. K. Guilkey (2016). GMM with multiple missing variables. *Journal of Applied Econometrics 31*(4), 678–706.

Chen, K. (2001). Parametric models for response-biased sampling. *Journal of the Royal Statistical Society, Series B 63*, 775–789.

Chetty, R. and E. Saez (2004). Do dividend payments respond to taxes?: Preliminary evidence from the 2003 dividend tax cut. NBER Working Paper Series 10572, National Bureau of Economic Research, Cambridge, MA.

Clarke, P. S. and P. F. Tate (1999). Methodological issues in the production and analysis of longitudinal data from the labour force survey. Government Statistical Service Methodology Series 17, Office for National Statistics, London.

Das, M. (2004). Simple estimators for nonparametric panel data models with sample attrition. *Journal of Econometrics 102*(1), 159–180.

Deng, Y., D. S. Hillygus, J. P. Reiter, Y. Si, and S. Zheng (2013). Handling attrition in longitudinal studies: The case for refreshment samples. *Statistical Science 28*, 238–256.

Ekinci, E. (2007). Dealing with attrition when refreshment samples are available: An application to the Turkish Household Labor Force Survey. Master's thesis, Koc University, Istanbul, Turkey.

EUROSTAT (2019). Labor force survey in the EU, Candidate and EFTA countries: Main characteristics of the national surveys, 2018. Statistical report, Publications Office of the European Union, Luxembourg.

Fitzgerald, J., P. Gottschalk, and R. Moffitt (1998). An analysis of sample attrition in panel data: The Michigan Panel Study of Income Dynamics. *Journal of Human Resources 33*, 251–299.

Fitzmaurice, G. M., S. R. Lipsitz, G. Molenberghs, and J. G. Ibrahim (2005). A protective estimator for longitudinal binary data subject to non-ignorable non-monotone missingness. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 168*(4), 723–735.

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.

Gruber, J. (1997). The incidence of payroll taxation: Evidence from Chile. *Journal of Labor Economics 15*(S3), S72–S101.

Hausman, J. A. and D. A. Wise (1979). Attrition bias in experimental and panel data: The Gary Income Maintenance Experiment. *Econometrica 47*, 455–473.

Hawkes, D. and I. Plewis (2006). Modelling non-response in the National Child Development Study. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 169*(3), 479–491.

Heckman, J. J. (1987). Selection bias and self-selection. In J. Eatwell, M. Milgate, and P. Newman (Eds.), *The New Palgrave: A Dictionary of Economics*, pp. 287–297. Basingstoke: Palgrave Macmillan.

Hellerstein, J. K. and G. W. Imbens (1999). Imposing moment restrictions from auxiliary data by weighting. *Review of Economics and Statistics 81*(1), 1–14.

Hirano, K., G. W. Imbens, G. Ridder, and D. B. Rubin (2001). Combining panel data sets with attrition and refreshment samples. *Econometrica 69*(6), 1645–1659.

Hoonhout, P. and G. Ridder (2019). Nonignorable attrition in multi-period panels with refreshment samples. *Journal of Business and Economic Statistics 37*(3), 377–390.

Kazianga, H., W. A. Masters, and M. S. McMillan (2014). Disease control, demographic change and institutional development in africa. *Journal of Development Economics 110*(6), 313–326.

Little, R. J. A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association 77*(378), 237–250.

Little, R. J. A. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association 88*(423), 1001–1012.

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

Little, R. J. A. and M.-M. Wu (1991). Models for contingency tables with known margins when target and sampled populations differ. *Journal of the American Statistical Association 86*(413), 87–95.

Longford, N. T., P. Tyrer, U. A. M. Nur, and H. Seivewright (2006). Analysis of a long-term study of neurotic disorder, with insights into the process of non-response. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 169*(3), 507–523.

Madow, W. G., H. Nisselson, I. Olkin, and D. Rubin (eds.) (1993). *Incomplete Data in Sample Surveys*. Volumes 1-3. New York, NY: Academic Press.

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

Nicoletti, C. and F. Peracchi (2005). Survey response and survey characteristics: Microlevel evidence from the European Community Household Panel. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 168*(4), 763–781.

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. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics, Volume 6, Part B*, pp. 5469–547. Amsterdam: Elsevier.

Rubin, D. B. (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 Labour Force Survey. *Journal of the American Statistical Association 81*, 42–47.

Stasny, E. A. (1988). Modeling nonignorable nonresponse in categorical panel data with an example in estimating gross labor-force flows. *Journal of Business and Economic Statistics 6*, 207–219.

Tunalı, I. (2009). Analysis of attrition patterns in the Turkish Household Labor Force Survey, 2000-2002. In R. Kanbur and J. Svejnar (Eds.), *Labour Markets and Development*, pp. 110–136. London and New York: Routledge.

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

Xie, H. and Y. Qian (2012). Measuring the impact of nonignorability in panel data with non-monotone nonresponse. *Journal of Applied Econometrics 27*(1), 129–159.

Appendix

In this Appendix, we provide a proof of the uniqueness of the solution to the linear parameterization of the 3×3 NGA model used in our labor market application.

Let $\boldsymbol{A_j}$ denote the 5×5 partition of the $\boldsymbol{A}$ matrix defined implicitly by equation (3.2) with the jth row removed, and let $\boldsymbol{b_j}$ denote the 5×1 partition of vector $\boldsymbol{b}$ with the jth row removed, $j = 1, 2, ..6$. With this notation, the system with the 6th equation removed can be expressed as $\boldsymbol{A_6\beta = b_6}$ and has the explicit form given below:

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

It is straightforward to establish that $rank(\boldsymbol{A_6}) = 5$. Thus the solution to the reduced system of equations is unique and is given by $\widehat{\boldsymbol{\beta}} = \boldsymbol{A_6^{-1} b_6}$. Next, we define the following 5×5 pivot matrices:

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

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

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

It is also straightforward to show that for $j = 1, 2, .., 5$, $\boldsymbol{E_j A_j} = \boldsymbol{A_6}$, and $\boldsymbol{E_j b_j} = \boldsymbol{b_6}$. Since the pivot matrices are of full rank, this proves that all six systems are equivalent, and yield the same unique solution $\widehat{\boldsymbol{\beta}}$.