Identification and Estimation of Panel Data Models with Attrition Using Refreshment Samples

submitted in partial fulfilment of the requirements

for the degree of Doctor of Philosophy (PhD)

University College London (UCL)

Pierre Hoonhout

January 2011

Declaration

I, Pierre Hoonhout, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Pierre Hoonhout

Abstract

This thesis deals with attrition in panel data. The problem associated with attrition is that it can lead to estimation results that suffer from selection bias. This can be avoided by using attrition models that are sufficiently unrestrictive to allow for a wide range of potential selection. In chapter 2, I propose the Sequential Additively Nonignorable (SAN) attrition model. This model combines an Additive Nonignorability assumption with the Sequential Attrition assumption, to just-identify the joint population distribution in Panel data with any number of waves. The identification requires the availability of refreshment samples. Just-identification means that the SAN model has no testable implications. In other words, less restrictive identified models do not exist.

To estimate SAN models, I propose a weighted Generalized Method of Moments estimator, and derive its repeated sampling behaviour in large samples. This estimator is applied to the Dutch Transportation Panel and the English Longitudinal Study of Ageing. In chapter 4, a likelihood-based alternative estimation approach is proposed, by means of an EM algorithm. Maximum Likelihood estimates can be useful if it is hard to obtain an explicit expression for the score function implied by the likelihood. In that case, the weighted GMM approach is not applicable.

Contents

List of Figures

List of Tables

Dedication

I dedicate this thesis to my wife, Maria, and to our future together.

Acknowledgements

I must dearly thank the following people:

My advisor, Prof. Richard Blundell, for giving good advice at all stages of my research. Geert Ridder and Guido Imbens, for initiating much of the work which this thesis relates to, as well as builds upon. The Department of Economics at UCL has been an inspiring place to do research. I am particularly grateful to Adam Rosen for his support and enthousiasm. Finally, I would like to thank the many people that have contributed to making my stay in London as pleasurable and exciting as it has been.

Chapter 1

Introduction

This thesis examines the problem of attrition in panel data. The problem is described below, together with a short description of the solutions that I propose.

A panel dataset consists of a set of individuals, each of which is followed over time. Attrition occurs if not all these individuals continue to respond to the survey in all time-periods. The resulting missing data induces a problem of identification: more than one population distribution of responses is consistent with the partially observed information.

Two approaches can be distinguished to deal with this identification problem. The first approach aims at point-identification of the population distribution by adding information to the model. An example of this is the Missing At Random (MAR) attrition model, which postulates that subjects that leave the panel and subjects that stay in the panel, differ in terms of observables only. This thesis follows this first approach. The second approach studies what conclusions can still be drawn without making any identifying assumptions. Inference is hence based on a *set* of population distributions (or, more generally, parameters of interest), those that are consistent with the observed information. The parameters of interest are called set-identified in this case (see Manski (1995) and Manski (2003) for details). This approach is appealing, as it leads to a set of potential parameter values, containing the true value, that provides a range of agreement between researchers. However, in many cases the range of agreement is too wide to be informative. Indeed, as attrition typically occurs in each wave, it potentially contaminates the whole panel, with the possible exception of the first wave. A set identification approach to the problem of attrition would therefore by necessity disregard most of the observable information.

My opinion is that there is room for approaches that add information to achieve point identification using all the information observable from the panel. However, misspecification of the attrition model that point-identifies the population distribution leads to unreliable inference. It is therefore essential to only maintain those assumptions that are strictly necessary for identification. By exploiting the information contained in refreshment samples, the set of modelling assumptions can be reduced. In this thesis, I aim to show how this can be done.

Hirano et al. (2001) were the first to suggest the use of this type of auxiliary information. They show that their Additively Nonignorable attrition model just-identifies the population distribution for panels with two waves. In chapter 2, I generalize this attrition model to panels with any number of waves. The first main result in this thesis is that application of the identification strategy of Hirano et al. (2001) to multi-wave panels leads to an attrition model that has undesirable properties. The model is shown to be over-identified, fails to encompass MAR, and is time-inconsistent (time-consistent attrition models are defined in section 2.5). The second main result is that the Sequential Additively Nonignorable (SAN) attrition model, proposed in section 2.5, resolves these issues: the SAN model identifies the population distribution, has no testable implications, encompasses Missing At Random and is time-consistent. In section 2.6, I propose a weighted GMM approach to estimate parameters that solve a set of moment conditions, free of attrition bias under SAN, and derive its asymptotic properties. Both chapter 2 and chapter 3 contain an application of this estimator.

In chapter 3, I investigate the attrition problem in the English Longitudinal Study of Ageing (ELSA). As is the case with most panel studies, the ELSA panel suffers from attrition. As the ELSA panel collects refreshment samples, the weighted GMM estimator of chapter 2 can be used. I present estimates of the probability of transition into retirement (or more precisely: inactivity) that are free of attrition bias under SAN attrition. The estimation results are compared with those obtained from Missing Completely At Random (MCAR) and Missing At Random (MAR) attrition.

The Fourth chapter investigates an alternative estimation approach. In this chapter, an EM algorithm is formulated that estimates Sequential Additively Nonignorable attrition models. This approach can be useful if an explicit expression for the score function implied by the likelihood is hard to obtain. Almost all EM algorithms proposed in the missing data literature require Missing At Random. Moreover, usually the values taken by regressor variables are assumed to be constant over time. The algorithm proposed in chapter 4 requires neither of these assumptions. It is shown that estimation by direct maximization of the likelihood has three disadvantages. First, it requires the specification of $f(x_2|x_1, z; \pi)$ and estimation of its parameters π. Second, desirable properties of the population model likelihood are not necessarily retained in the incomplete panel likelihood. Third, maximization over the complete vector of parameters (β, α, π) is required. The latter is particularly inconvenient if the vector of nuisance parameters π is of high dimension. The EM algorithm solves these problems. Moreover, when the time-varying variables are discrete, it is shown that the nuisance parameters π can be estimated by simply calculating sample fractions.

Chapter 5 summarizes and concludes.

Chapter 2

Non-ignorable Attrition in Multi-Wave Panel Data with Refreshment Samples

2.1 Introduction

Panel data studies aim to collect responses from the same set of subjects repeatedly over time. These subjects can be individuals, households, firms, regions or countries. In what follows, I will refer to the subjects as individuals. Panel data are particularly useful for modelling dynamic responses and to control for unobserved heterogeneity. The main problem associated with panel data is attrition. Attrition occurs when some of the individuals that participated in the first wave of the panel study do not participate in the second wave. The resulting missing data leads to a problem of identification: more than one population distribution of the variables in both waves are consistent with the incompletely observed information. In later waves further drop-out can occur, each time reducing the number of individuals in the sample. The survey sampler can respond to this by collecting a new random sample[1] from the population each time attrition occurs. These random samples are called *refreshment samples*[2].

Most panel studies suffer from attrition. The implied identification problem requires researchers to add information to the model before point estimates of the parameters of interest can be obtained. The additional information can come in the form of restrictions imposed on the attrition model, restrictions imposed on the population distribution or it can come from auxiliary data. Misspecification of the attrition or population model leads to unreliable inference. It is therefore essential to only maintain those assumptions that are strictly necessary for identification. This chapter investigates which assumptions are needed. In particular, it investigates the benefits of exploiting the information contained in the refreshment samples to reduce the set of modelling assumptions. Hirano et al. (2001) were the first to suggest the use of this type of auxiliary information. They show that the Additively Nonignorable attrition model just-identifies the population distribution for panels with two waves. In this chapter, I generalize their result to panels with any number of waves. The first main result of this chapter is that application of their

[1]Stratified sampling is allowed as long as the variables that define the strata do not change over time.

[2]Sometimes, the individuals in the refreshment samples are also followed over time, giving rise to *refreshment panels* instead of refreshment samples. These panels are as likely to suffer from attrition as the original panel. Although the additional information available in refreshment panels could possibly be exploited as well, this chapter only discusses the use of refreshment samples.

identification strategy to multi-wave panels, used in Nevo (2003), leads to an attrition model has undesirable properties. This model is shown to be over-identified, fails to encompass MAR, and is time-inconsistent (time-consistent attrition models are defined in section 2.5). The second main result is that the Sequential Additively Nonignorable (SAN) attrition model, proposed in section 2.5, resolves these issues. The SAN model identifies the population, has no testable implications, encompasses Missing At Random and is time-consistent. In section 2.6, I propose a weighted GMM approach to estimate parameters that solve a set of moment conditions, free of attrition bias under SAN, and derive its asymptotic properties. Finally, I apply the weighted GMM estimator to the Dutch Transportation panel.

2.2 Missing Data Pattern

In this section I lay out the structure of the missing data problem in multi-wave panel data with attrition. Table 2.1 summarizes the pattern of observed and missing information for a panel with three waves, with attrition and refreshment samples. If individual i is observed in period t the observation indicator-variable D_{it} equals 1. Otherwise D_{it} equals zero. Sampling in the first wave is assumed to be unselective. For later reference we define the Balanced Panel Indicators d_{it} that equal 1 if individual i was part of the balanced panel after wave t became available. In other words, $d_{it} = 1$ if $\Pi_{s=1}^{t} D_{is} = 1$ and zero otherwise. Only variables that vary over time are affected by attrition, so Table 1 distinguishes between time-varying variables Z_t and time-constant variables X. We do not observe Z_t for individuals that drop out of the panel study in period t. If individuals that drop out do not return to the study later, Z will also be unobserved in later periods. For now, it is assumed that there is no return. I will refer to the combined set of observations from BP, IP3 and IP2 simply as "the panel."

The Balanced Panel consists of the individuals that participated in all waves. Incomplete observations are obtained for individuals in the incomplete panel sub-samples. Standard techniques for analyzing panel data estimate the parameters of interest using the balanced panel alone, thereby ignoring the problems associated with attrition. These estimates are likely to be inconsistent if the individuals in the balanced panel are different from individuals in the other sub-samples of the panel (IP2 and IP3) with respect to the variables of interest. The next sections examine methods that do not ignore the potentially biasing effects of attrition. Section 2.3 summarizes methods that have been proposed to deal with attrition in two-period panels. Later sections discuss generalizations of these methods to multi-wave panels.

2.3 Models for Attrition in Panel Data with Two Waves

In what follows, the dependence on X is suppressed. All statements continue to hold conditional on X.

Name of the sub-population	Observation Indicators			Balanced Panel Indicators			Individual Characteristics			
	D_1	D_2	D_3	d_1	d_2	d_3	Z_1	Z_2	Z_3	X
Balanced Panel (BP)	1	1	1	1	1	1	Obs	Obs	Obs	Obs
Incomplete Panel 3 (IP3)	1	1	0	1	1	0	Obs	Obs	.	Obs
Incomplete panel 2 (IP2)	1	0	0	1	0	0	Obs	.	.	Obs
Refreshment Sample 2 (RS2)	0	1	0	0	0	0	.	Obs	.	Obs
Refreshment Sample 3 (RS3)	0	0	1	0	0	0	.	.	Obs	Obs

Table 2.1: Missing data pattern in a three-wave panel data set with attrition and refreshment samples and no return. D_{it} equals 1 if individual i is observed in wave t. d_{it} equals 1 if individual i is part of the balanced panel after wave t has been observed. Z denotes the set of time-varying variables and X denotes the set of time-constant variables. "Obs" denotes "observed" and "." denotes "missing". The combined set of observations from the balanced panel (BP) and the incomplete panel (IP3 and IP2) is called "the panel."

2.3.1 Why aim to identify the joint population distribution?

Consider a panel data set with two waves. The data can be analyzed using a wide variety of models describing the population distribution. The choice of population-model is governed by the type of questions the panel study is meant to answer. For instance, some parameters of interest describe a linear panel data model, while others describe some transition or duration model. In the absence of attrition, any such parameter, if at all identified, can be deduced from the joint population distribution $f(Z_1, Z_2)$. If attrition occurs, an attrition model that identifies this distribution therefore ensures identification of any parameter that would be identified in the absence of attrition. This avoids the need for a separate identification analysis for e.g. linear models, transition models and duration models.

2.3.2 MAR, HW and AN models for Attrition

The presence of attrition in a panel with two waves implies that the distribution $f(Z_2|Z_1, D_2 = 0)$ is not observed. Attrition can be modelled by restricting the conditional probability of observation in both waves. This observation probability would be unrestricted if $P(D_2 = 1|Z_1, Z_2) = G(k(Z_1, Z_2))$, where G denotes some cdf function and k denotes the index function.[3] For example, in a logit model, G would be the cdf of the logistic distribution and k would be a linear function of the variables in Z_1 or Z_2. Any particular choice of the index function corresponds with a particular unobserved distribution

[3] As the individuals are fully observed in the first wave, $P(D_1 = 1) = 1$.

$f(Z_2|Z_1, D_2 = 0)$. Attrition models can hence be described by the restrictions they place on the index function k. What follows is a discussion of the four most commonly used attrition models for panels with two waves.

As mentioned earlier, one possible approach to the problem of attrition in panel data is to ignore incomplete observations and use only the observations from the balanced panel. This approach is valid only if the attrition is Missing Completely At Random (MCAR)[4]. Formally, MCAR attrition maintains that $P(D_2 = 1|Z_1, Z_2) = G(k_0)$, where k_0 is some constant. In other words, the probability of observation does not vary with Z_1 or Z_2. The population distribution *solution* implied by the assumption of MCAR attrition, $f_{MCAR}(Z_1, Z_2)$, then equals the balanced panel distribution $f(Z_1, Z_2|D_2 = 1)$. Any complete-cases analysis of panel data implicitly assumes MCAR.

Attrition is most often taken into account by assuming the Missing At Random (MAR) attrition model, $P(D_2 = 1|Z_1, Z_2) = G(k_0+k_1(Z_1))$. The observation probability is allowed to vary with Z_1 in arbitrary ways, via the unrestricted function k_1, but cannot depend on Z_2. The population distribution solution implied by MAR is $f_{MAR}(Z_1, Z_2) = f(Z_2|Z_1, D_2 = 1)f(Z_1)$. As Z_1 is observed for all individuals in the panel, MAR is sometimes referred to as *selection on observables* (Fitzgerald and Moffitt (1998)).

The HW model allows for selection on unobservables in that it has the observation probability depend on the partially observed Z_2, $P(D_2 = 1|Z_1, Z_2) = G(k_0 + k_2(Z_2))$. This model was suggested by Hausman and Wise (1979). Attrition models that depend on partially observed information are called Non-ignorable. Note that HW admits selection on unobservables but at the same

[4]The MCAR and MAR models for missing data are treated in Little and Rubin (1987).

time rules out selection on observables.

Both the MAR and HW attrition models are nonparametrically just-identified. They are identified and have no testable implications and are hence observationally equivalent. The consequences of this are not always well-understood. For example, it is not possible to test for selection on unobservables unless one relies on untestable functional form restrictions. Indeed, for any HW solution that would suggest selection on unobservables, there is an observationally equivalent MAR solution suggesting selection on observables.

Additional information from a refreshment sample can disentangle the two forms of selection. Hirano et al. (2001) show that, if a second wave refreshment sample is available, the Additively Non-ignorable (AN) attrition model identifies the population distribution, with observation probability $P(D_2 = 1|Z_1, Z_2) = G(k_0 + k_1(Z_1) + k_2(Z_2))$. The AN model admits Non-ignorable attrition and does not rule out selection on observables.

2.4 Sequential Attrition Models in Panel Data with More than Two Waves

We will now examine the multi-wave versions of the attrition models discussed in section 2.3.2. To simplify notation we again suppress the time-constant variables X and denote the event $Z_1 = z_1, Z_2 = z_2, \ldots, Z_t = z_t$ by $Z^t = z^t$. Similarly, $D^t = 1$ denotes the event $D_1 = 1, \ldots, D_t = 1$.

2.4.1 A Running Example with Binary Variables

In what follows, I will illustrate the development of attrition models by specializing the results we achieve for the general case to the simplest possible problem: a panel with three waves where Z_1, Z_2 and Z_3 are scalar random variables taking only the values 0 or 1. Due to attrition, the population distribution $f(Z^3)$ is not identified. This population distribution has 8 parameters, subject to an adding-up restriction. Consider the identity

$$f(Z^3) = \frac{P(D^3 = 1)}{P(D^3 = 1|Z^3)} f(Z^3|D^3 = 1). \tag{2.1}$$

As $f(Z^3|D^3 = 1)$ is in principle observable from the balanced panel, this identity shows that the population distribution can be recovered from the balanced panel if the *observation probability* $P(D^3 = 1|Z^3)$ can be found. In the binary example, this observation probability can be parameterized as

$$P(D^3 = 1|Z^3) = G(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \tag{2.2}$$

$$\beta_{12} Z_1 Z_2 + \beta_{13} Z_1 Z_3 + \beta_{23} Z_2 Z_3 + \beta_{123} Z_1 Z_2 Z_3) \tag{2.3}$$

This parameterization with 8 parameters imposes no restrictions on the observation probability due to the binary nature of Z_1, Z_2 and Z_3. Identification of $f(Z^3)$ can be achieved by either imposing restrictions on $f(Z^3)$ itself, or by imposing restrictions on the observation probability $P(D^3 = 1|Z^3)$, or both.

Another possibility to parameterize the observation probability is by means of its *observation hazards:*

$$P(D_2 = 1|Z^3) = G(\gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_2 + \gamma_3 Z_3 + \tag{2.4}$$

$$\gamma_{12} Z_1 Z_2 + \gamma_{13} Z_1 Z_3 + \gamma_{23} Z_2 Z_3 + \gamma_{123} Z_1 Z_2 Z_3) \tag{2.5}$$

$$P(D_3 = 1|D^2 = 1, Z^3) = G(\delta_0 + \delta_1 Z_1 + \delta_2 Z_2 + \delta_3 Z_3 +$$

$$\delta_{12} Z_1 Z_2 + \delta_{13} Z_1 Z_3 + \delta_{23} Z_2 Z_3 + \delta_{123} Z_1 Z_2 Z_3) \tag{2.6}$$

This formulation of the mechanism by which attrition occurs is more flexible as it requires 16 parameters.

2.4.2 Observationally Equivalent Population Distributions and Observation Probabilities

Attrition problems in panels with three waves can be modelled by specifying the joint population distribution $f(Z^3)$ and the observation probability $P(D^3 = 1|Z^3)$. In the binary case, this specification requires 16 parameters using (2.2) or 24 parameters using (2.4). Models that have no testable implications can be obtained by requiring that the pair $(f(Z^3), P(D^3 = 1|Z^3))$ be consistent with the observable distributions $f(Z^3|D^3 = 1)$ from the balanced panel, $f(Z^2|D_2 = 1, D_3 = 0)$ from IP3, $f(Z_1|D_2 = 0)$ from IP2, and the response fractions $P(D_2 = 1)$ and $P(D_3 = 1|D_2 = 1)$. Clearly, many of these pairs exist, as each particular set of unobserved distributions $f(Z_3|Z_1, Z_2, D_2 = 1, D_3 = 0)$, $f(Z_3|Z_1, Z_2, D_2 = 0)$, $f(Z_3|Z_1, Z_2, D_2 = 0)$ and $f(Z_2|Z_1, D_2 = 0)$ implies one. For general panels with three waves we have[5]

[5]Replace summations with integrals in the continuous case.

Definition. (Observational equivalence)

If there is no return and only the (incomplete) panel is observed, distinct pairs of population distributions and observation probabilities $(f(Z^3), P(D^3 = 1|Z^3))$ are called *observationally equivalent* if they are solutions to the following sets of equations:

$$P(D^3 = 1|Z^3)f(Z^3) = P(D^3 = 1)f(Z^3|D^3 = 1) \quad \textbf{(2.7)}$$

$$\sum_{Z_3} P(D_3 = 0|D_2 = 1, Z^3)P(D_2 = 1|Z^3)f(Z^3) = P(D_2 = 1, D_3 = 0) \quad \textbf{(2.8)}$$

$$f(Z^2|D_2 = 1, D_3 = 0)$$

$$\sum_{Z_2}\sum_{Z_3} P(D_2 = 0|Z^3)f(Z^3) = P(D_2 = 0)f(Z_1|D_2 = 0) \quad \textbf{(2.9)}$$

The left-hand sides of these three sets of equations represent the model and the right hand sides the observable distributions from BP, IP3 and IP2, respectively. In the binary case, these equations provide 14 restrictions. Models that identify the population distribution without having testable implications therefore require 2 additional restrictions under (2.2) and 10 additional restrictions under (2.4). Restrictions could be imposed on the joint population distribution $f(Z^3)$ (e.g. assuming stochastic independence of Z_1, Z_2 and Z_3 provides 5 restrictions). For reasons outlined in Section 2.3.1, I only consider restrictions on the attrition probabilities.

2.4.3 MCAR

The generalization of MCAR to multi-wave panels is immediate; it assumes that $P(D^T = 1|Z^T)$ does not vary with Z^T. This assumption is valid if the balanced panel can be considered a random sample from the population distribution of interest. The implied solution $f_{MCAR}(Z^T)$ equals the observed $f_{BP}(Z^T) = f(Z^T|D^T = 1)$. This solution is not consistent with all information available in the panel if $f(Z_1|D^T = 1) \neq f(Z_1)$. As both these distributions are in principle observable, this assertion is testable. As MCAR has testable implications, more general (less restrictive) attrition models can identify the population distribution. In the simple binary example, MCAR corresponds to imposing on (2.2) the seven restrictions $\beta_1 = \beta_2 = \beta_3 = \beta_{12} = \beta_{13} = \beta_{23} = \beta_{123} = 0$. It is clearly not necessarily consistent with all the information in the panel.

2.4.4 Sequential Attrition

To discuss generalizations of the other three attrition models, we will need to introduce the notion of sequential attrition.

Definition. (Sequential Attrition)

An attrition model is called *sequential* if

$$P(D_t = 1|D^{t-1} = 1, Z^T) = P(D_t = 1|D^{t-1} = 1, Z^t)\ for\ all\ t = 2, \ldots, T-1.$$

We can then write the observation probability as

$$P(D^T = 1|Z^T) = \prod_{t=2}^{T} P(D_t = 1|D^{t-1} = 1, Z^t).$$

Sequential attrition imposes restrictions on the sequence of observation hazards $P(D_t = 1|D^{t-1} = 1, Z^T)$. These observation hazards are allowed to depend on current and past values of Z, but not on future values of Z. In the binary example, Sequential Attrition restricts the second period observation hazard to not depend on Z_3 by imposing the four restrictions $\gamma_3 = \gamma_{13} = \gamma_{23} = \gamma_{123} = 0$, reducing the number of restrictions necessary for just-identification to 6. As the current values of Z are partially unobserved, sequential attrition admits selection on unobservables.

Proposition. *(SA in terms of population distributions)*

Sequential Attrition is equivalent to

$f(Z_t|Z^{t-1}) = f(Z_t|Z^{t-1}, D^{t-1} = 1)$ *for* $t = 1, 2, \ldots, T$.

Proof. The definition of sequential attrition implies that $P(D^{t-1} = 1|Z^t) = P(D^{t-1} = 1|Z^{t-1})$. As $f(Z_t|Z^{t-1}, D^{t-1} = 1) = \frac{P(D^{t-1}=1|Z^t)}{P(D^{t-1}=1|Z^{t-1})} f(Z_t|Z^{t-1})$, the result follows. □

The proposition phrases Sequential Attrition in terms of population distributions: the conditional distribution $f(Z_t|Z_{t-1})$ can be obtained by only considering the sub-population of individuals that are still in the panel in period $t-1$. Under sequential attrition, individuals that dropped out earlier are not informative for the purpose of recovering $f(Z_t|Z_{t-1})$.

Sequential Missing at Random

The MAR assumption for panel data is usually formulated as $P(D^T = 1|Z^T) = \prod_{t=2}^{T} P(D_t = 1|D^{t-1} = 1, Z^{t-1})$.[6] Note that MAR assumes Sequential Attrition *as well as* $P(D_t = 1|D^{t-1} = 1, Z^t) = P(D_t = 1|D^{t-1} = 1, Z^{t-1})$ for $t = 2, \ldots T$. It combines the SA assumption with a MAR assumption in each period. For that reason, I will refer to this assumption as Sequential MAR (SMAR). As all SMAR observation hazards are observable, the population distribution is clearly identified. The next proposition shows that SMAR just-identifies the population distribution, and as such has no testable implications.

Proposition. *SMAR just-identifies the population distribution $f(Z^T)$, i.e. SMAR satisfies the following two conditions:*

(i) Any population distribution $f_{SMAR}(Z^T)$ implied by the SMAR attrition model is consistent with all the information in the panel.

(ii) The SMAR attrition model implies a unique solution $f_{SMAR}(Z^T)$.

Proof. To show (i), note that from $P(D_1 = 1) = 1$ we have $f_{SMAR}(Z_1) = f(Z_1)$. We now show that all conditional distribution solutions $f(Z_t|Z^{t-1})$ implied by SMAR are consistent with the panel as well. Sequential attrition implies that $f_{SMAR}(Z_t|Z_{t-1}) = f(Z_t|Z_{t-1}, D^{t-1} = 1)$. From Bayes' rule, we obtain the identity

$$f(Z_t|Z^{t-1}, D^{t-1} = 1) = \frac{P(D_t = 1|D^{t-1} = 1, Z^{t-1})}{P(D_t = 1|D^{t-1} = 1, Z^t)} f(Z_t|Z^{t-1}, D^t = 1). \qquad (2.10)$$

[6]This definition does not correspond exactly to the MAR assumption in its original form (Rubin (1976)). If there is no return, however, they are equivalent (see Robins et al. (1995)).

Maintaining MAR in each period t therefore implies that $f(Z_t|Z^{t-1})$ coincide with the distributions $f(Z_t|Z^{t-1}, D^t = 1)$. As these latter distributions correspond to all that is observed in the panel beyond $f(Z_1)$, SMAR satisfies (i). Their uniqueness implies that SMAR satisfies (ii). □

It is instructive to understand the just-identification result in the binary example. The unknown distributions are $f(Z_2|Z_1, D_2 = 0)$, $f(Z_3|Z_2, Z_1, D_2 = 1, D_2 = 0)$ and $f(Z_3|Z_2, Z_1, D_2 = 0, D_2 = 0)$. Sequential Attrition implies that the latter distribution is uninformative for the identification of $f(Z_3|Z_{1,}Z_2)$ under SA, reducing the number of unknown parameters from 10 to 6. As mentioned above, this provides a second way of understanding the identifying power of the sequential attrition assumption, based on population distributions instead of observation hazards. From the definition of SA, the second period observation hazard becomes

$$P(D_2 = 1|D_1 = 1, Z_1, Z_2, Z_3) = G(\gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_2 + \gamma_{12} Z_1 Z_2). \quad (2.11)$$

The remaining 6 restrictions are obtained by imposing MAR: $\gamma_2 = \gamma_{12} = 0$ and $\delta_3 = \delta_{13} = \delta_{23} = \delta_{123} = 0$.

Sequential Hausman and Wise

Following the development of the SMAR model, the Sequential Hausman and Wise (SHW) model can be defined as $P(D^T = 1|Z^T) = \prod_{t=2}^{T} P(D_t = 1|D^{t-1} = 1, Z_t)$. It combines the sequential attrition assumption with a Hausman and Wise assumption in each period. This model is over-identified. To see this, it suffices to look at our binary example. As in the SMAR case, there are 6 unknowns after SA is imposed. In addition, SHW imposes $\gamma_1 = \gamma_{12} = 0$ and $\delta_1 = \delta_2 = \delta_{12} = \delta_{13} = \delta_{23} = \delta_{123} = 0$. These 8 restrictions for 6 unknowns imply over-identification.

2.5 Identification of the Sequential Additively Non-ignorable Attrition Model

Definition. (sampled population distribution)

The sampled population distribution is defined as

$$f_s(Z^T) = f(Z_1)\Pi_{t=2}^{T} f(Z_t|Z^{t-1}, D^{t-1} = 1).$$

The SMAR attrition model implies a solution $f_{SMAR}(Z^T)$ that is consistent with all the information in the panel. This is obvious as the SMAR solution equals the sampled population distribution $f_s(Z^T)$. Although the sampled population distribution may not equal the *target population* distribution $f_t(Z^T)$, they are observationally equivalent, so no testable implications arise. If, however, in addition to the observations available in the panel, refreshment samples are available, SMAR does have testable implications. For instance, the marginal distribution $f_{SMAR}(Z_2)$ may be different from $f_t(Z_2)$ obtained from the second period refreshment sample. Although SMAR is consistent with the information available from the panel, it is not generally consistent with the information in the refreshment samples.

Motivated by this conflict, I aim to develop attrition models that have three key properties. First, they are just-identified, meaning that they are identified (not unidentified) and are consistent with the information in the panel (not over-identified). The latter property was satisfied by $f_s(Z^T)$ by construction. The analysis in section 2.4.2 showed, however, that there exist other distributions that are observationally equivalent. Second, the implied population distribution is consistent with the refreshment samples. This requirement is natural, as the refreshment samples are random draws from the target population distribution, unaffected by selection due to attrition. Third, they have SMAR as a special case, meaning that they do not rule out selection on observables a priori.

2.5.1 Marginal AN Attrition

The AN model for two-period panels can be motivated as finding the population distribution as close as possible to the balanced panel, while being consistent with $f(Z_1)$ obtained from the panel and $f(Z_2)$ obtained from the second period refreshment sample (see Hirano et al. (2001)). This relates the AN model to the problem of estimating cell probabilities in contingency tables with known marginals (Little and Wu (1991), Haberman (1984), Ireland and Kullback (1968)). A multi-wave extension (it suffices to use three waves) that follows this motivation can be formalized by setting up the following optimization problem

Definition. (Marginal AN Attrition)

Let $\overline{f}(Z_1)$, $\overline{f}(Z_2)$ and $\overline{f}(Z_3)$ denote squared summable (Lebesgue integrable in the continuous case) marginal distributions observable from the panel and the two refreshment samples. The solution $f_{MAN}(Z^T)$ solves the following optimization problem[7]:

$$\max_{f(Z^3)} \sum_{Z_1,Z_2,Z_3} f(Z^3|D^3=1) h\left(\frac{f(Z^3)}{f(Z^3|D^3=1)}\right) \text{ subject to} \tag{2.12}$$

$$f(Z^3|D^3=1)P(D^3=1) < f(Z^3) \quad \forall Z^3 \tag{2.13}$$

$$\sum_{Z_1,Z_2,Z_3} f(Z^3) = 1 \tag{2.14}$$

$$\sum_{Z_2,Z_3} f(Z^3) = \overline{f}(Z_1) \quad \forall Z_1 \tag{2.15}$$

$$\sum_{Z_1,Z_3} f(Z^3) = \overline{f}(Z_2) \quad \forall Z_2 \tag{2.16}$$

$$\sum_{Z_1,Z_2} f(Z^3) = \overline{f}(Z_3) \quad \forall Z_3 \tag{2.17}$$

The function $h(t)$ is continuously differentiable and strictly concave. The function $h(a)$ is continuously differentiable and strictly concave. It must be chosen such that the function $G(a) \equiv (h')^{-1}(a)$ is differentiable and strictly increasing with $\lim_{a\to-\infty} G(a) = 0$ and $\lim_{a\to\infty} G(a) = 1$.

The choice of h corresponds to choosing a measure of discrepancy to be minimized. The inequality restrictions ensure that the observation probabilities take values between zero and one. To see this, note that

$$f(Z^3) = \frac{1}{P(D^3=1|Z^3)} f(Z^3|D^3=1)P(D^3=1).$$

[7]For convenience we use distributions with discrete support in this exposition. The argument is essentially the same in the continuous case.

Theorem 1. *Let $f(Z^3)$ and $f(Z^3|D^3 = 1)$ be squared-summable discrete probability functions (squared Lebesgue integrable in the continuous case) with coinciding support. Then, the Marginal AN optimization problem has a unique solution. Moreover, the first order conditions imply the following restrictions on the observation probabilities:*

$$P(D^3 = 1|Z^3) = G(k_0 + k_1(Z_1) + k(Z_2) + k(Z_3)). \tag{2.18}$$

The functions $k_1(Z_1)$, $k_2(Z_2)$ and $k_3(Z_3)$ are arbitrary real-valued squared-summable sequences (squared Lebesgue integrable in the continuous case) normalized to equal zero for some value in the support of Z_1, Z_2 and Z_3, respectively, to allow for the inclusion of the intercept k_0 in the index. The function $G(t)$ is 1-to-1 related to h.

Proof. See Appendix. □

The index functions $k_1(Z_1)$, $k_2(Z_2)$ and $k_3(Z_3)$ mentioned in the theorem are the (sequences of) Lagrange multipliers (functional Lagrange multipliers in the continuous case) associated with the restrictions (2.15), (2.16) and (2.17), respectively. The normalization is necessary as restriction (2.14) renders one of the restrictions in (2.15), (2.16) and (2.17) redundant. The function h is one-to-one related to G. A popular choice for h is the likelihood metric $h(a) = \ln(a)$ corresponding to the linear probability model for the attrition function G. The Generalized Exponential Tilting norm proposed by Nevo (2002) corresponds to the logit model. For other choices see Imbens et al. (1998), Baggerly (1998) and Read and Cressie (1988).

Nevo (2003) applies this attrition model to study mobility patterns in the Dutch Transportation Panel, a panel data set with attrition for which refreshment samples were collected. He did not develop the information theoretic interpretation of this solution. It is the only generalization of the two-period AN model to panels with more than two waves that has been suggested in the literature. The uniqueness of the solution implies the identification of $f_{MAN}(Z^T)$. The restrictions (2.15), (2.16) and (2.17) imply that this solution is consistent with the information available in the refreshment samples and $f(Z_1)$ from the panel. However, the model has two shortcomings. First, it does not encompass SMAR, as the model imposes restrictions on the observation probability, not the observation hazards. Second, it is over-identified. To see this, the optimization from which it is derived does not force $f_{MAN}(Z_2|Z_1, D_2 = 1)$ to equal the observable distribution $f(Z_2|Z_1, D_2 = 1, D_3 = 0)$, as the latter distribution plays no role in the optimization. More formally, while the first order conditions imply restrictions on the observation probabilities, they fail to imply condition (2.8) from section 2.4.2. The MAN attrition model therefore has testable implications. Immediate generalization of the AN model proposed in Hirano et al. (2001) to multi-wave panels leads to an attrition model that has undesirable properties.

2.5.2 Time-consistent AN Attrition

The MAN model has testable implications because it fails to use all the data available in the panel. This can be avoided by using a sequential approach, as will be shown below. To motivate this approach, we first define time consistency.

If there is no attrition, we can view a panel study with T waves as revealing a sequence of population distributions $f(Z_1), f(Z^2), \ldots, f(Z^T)$. Each additional wave that becomes available adds one more population distribution to this sequence. The last population distribution is the population distribution of interest. Note that $f(Z^T)$ has all $f(Z^t)$ with $t < T$ as its (multivariate) marginal distributions. In other words, features of $f(Z^T)$ that pertain only to its marginal $f(Z^t)$ can be obtained by using the first t waves only. We refer to this property as *time-consistency*.

If the panel suffers from attrition, additional assumptions are needed to obtain a sequence of identified population distribution *solutions*. The MAN attrition model above, for instance, is not time-consistent as the balanced panel changes with each additional wave.

Definition. (Time consistent attrition)

An attrition model A is called *time-consistent* if it implies a sequence of population distribution solutions $f_A(Z_1), f_A(Z^2), \ldots, f_A(Z^T)$ such that

$\int f_A(Z^t) dZ_t = f_A(Z^{t-1})$ for all $t = 2, \ldots, T$.

Time-consistency is especially desirable for the analysis of ongoing panel studies, or for analyses that use only a first set of waves from a completed panel study. It becomes a natural assumption in a sequential approach. The sequential approach outlined below replaces the single optimization of the MAN attrition model by a sequence of optimizations, each of which uses the solution from its predecessor. These solutions can only be obtained if the observation probabilities are restricted to not depend on future values of Z. The recursion ensures that there are no testable implications; viewing the panel as a sequence of balanced panels ensures that all available information is used. It is implicitly defined below by taking the third wave optimization as representative.

Definition. (Time consistent AN Attrition)

Let $f(Z_t)$ be squared summable (integrable) for all t and let $\overline{f}(Z^2)$ be the distribution identified by the MAN model using only the first two waves of the panel. This solution corresponds to the AN solution proposed in Hirano et al. (2001). $\overline{f}(Z_3)$ denotes the marginal distribution obtained from the third period refreshment sample. The solution $f_{TCAN}(Z^3)$ solves the following optimization problem:

$$\max_{f(Z^3)} \sum_{Z_1,Z_2,Z_3} f(Z^3|D^3=1) h_3\left(\frac{f(Z^3)}{f(Z^3|D^3=1)}\right) \quad \textit{subject to}$$

$$f(Z^3|D^3=1)P(D^3=1) < f(Z^3) \quad \forall Z^3 \tag{2.19}$$

$$\sum_{Z_1,Z_2,Z_3} f(Z^3) = 1 \tag{2.20}$$

$$\sum_{Z_3} f(Z^3) = \overline{f}(Z^2) \quad \forall Z^2 \tag{2.21}$$

$$\sum_{Z_1,Z_2} f(Z^3) = \overline{f}(Z_3) \quad \forall Z_3 \tag{2.22}$$

The function $h_3(a)$ is continuously differentiable and strictly concave. It must be chosen such that the function $G_3(a) \equiv (h_3')^{-1}(a)$ is differentiable and strictly increasing with $\lim_{a\to-\infty} G_3(a) = 0$ and $\lim_{a\to\infty} G_3(a) = 1$.

Any sequence of population distribution solutions thus obtained is by construction time-consistent. Note that the notation h_t is used instead of h, as it is in principle possible to use a different discrepancy measure in each optimization. As the discrepancy measure h is one-to-one related with the attrition model G, the notation G_t is used below.

Theorem 2. *Let $f(Z^3)$ and $f(Z^3|D^3 = 1)$ be squared-summable discrete probability functions (squared Lebesgue integrable in the continuous case) with coinciding support. Moreover, let the observation probabilities be restricted to not depend on future values of Z, i.e. $P(D_2 = 1|Z^3) = P(D_2 = 1|Z^2)$. Then, the Time-Consistent AN optimization has a unique solution that has no testable implications. Moreover, the first order conditions imply the following restrictions on the observation probabilities:*

$$P(D^3 = 1|Z^3) = G_3(k_0 + k_1(Z^2) + k_2(Z_3)). \tag{2.23}$$

The functions $k_1(Z^2)$ and $k_2(Z_3)$ are arbitrary functions (squared summable / integrable) that are normalized to equal zero for some value in the support of Z^2 and Z_3, respectively, in order to allow for the inclusion of the intercept k_0 in the index.

Proof. See Appendix. □

The TCAN model imposes restrictions on a sequence of observation probabilities as opposed to a sequence of observation hazards. Its solution is by construction consistent with all the information in the refreshment samples and, as the first order conditions imply (2.7), (2.8) and (2.9), no testable implications arise. However, from (2.23), it fails to encompass SMAR.

2.5.3 The Sequential Additively Non-ignorable attrition model

The MAN and TCAN attrition models fail because they restrict the observation probability representation of attrition. To encompass SMAR the attrition model implied by the solutions of the sequence of optimizations need to restrict the observation hazards. The number of parameters can be reduced by imposing Sequential Attrition. When the optimization is over $f(Z_3|Z^2)$, Sequential Attrition can be imposed by optimizing over $f(Z_3|Z^2, D_2 = 1)$ after substitution. Again, the recursion is defined by taken the third wave optimization as representative.

Definition. (Sequential AN attrition)

Let $f(Z_t)$ be squared summable (integrable) for all t and let $\overline{f}(Z^2)$ be the distribution identified by the SAN model using only the first two waves of the panel. $\overline{f}(Z_3)$ denotes the marginal distribution obtained from the third period refreshment sample. The solution $f_{SAN}(Z^3)$ solves the following optimization problem:

$$\max_{f(Z^3|Z^2)} \sum_{Z_1,Z_2,Z_3} f(Z_3|Z^2, D^3 = 1)\bar{f}(Z^2) h_3 \left(\frac{f(Z_3|Z^2)\bar{f}(Z^2)}{f(Z_3|Z^2, D^3 = 1)\bar{f}(Z^2)} \right) \text{ subject to} \tag{2.24}$$

$$f(Z_3|Z^2, D^3 = 1)P(D_3 = 1|D^2 = 1, Z^2) < f(Z_3|Z^2) \; \forall Z^3 \tag{2.25}$$

$$\sum_{Z_1,Z_2,Z_3} f(Z_3|Z^2)\overline{f}(Z^2) = 1 \tag{2.26}$$

$$\sum_{Z_3} f(Z_3|Z^2)\overline{f}(Z^2) = \overline{f}(Z^2) \; \forall Z^2 \tag{2.27}$$

$$\sum_{Z_1,Z_2} f(Z_3|Z^2)\overline{f}(Z^2) = \overline{f}(Z_3) \; \forall Z_3 \tag{2.28}$$

$$f(Z_3|Z^2, D_2 = 1) = f(Z_3|Z^2) \; \forall Z^3 \tag{2.29}$$

The function $h_3(a)$ is continuously differentiable and strictly concave. It must be chosen such that the function $G_3(a) \equiv (h_3')^{-1}(a)$ is differentiable and strictly increasing with $\lim_{a\to-\infty} G_3(a) = 0$ and $\lim_{a\to\infty} G_3(a) = 1$.

The distribution $f(Z_3|Z^2, D^3 = 1)\bar{f}(Z^2)$ is the recursive analog of the sampled population distribution. They coincide if the attrition in the second wave satisfies MAR. Restriction (2.27) reduces to an adding-up restriction. It implies time-consistency of the solution, if this solution exists. The solution implied by the SAN attrition model can be interpreted as the population distribution that is as close as possible, in an information theoretic sense, to the sampled population distribution, while being consistent with the information contained in the refreshment samples. This contrasts with the MAN model, which finds the distribution that is as close as possible to the balanced panel.

Theorem 3. *Let $f(Z_3|Z^2)$ and $f(Z_3|Z^2)$ be squared-summable discrete probability functions (squared Lebesgue integrable in the continuous case) with coinciding support. The solution to the Sequential AN attrition optimization satisfies the following conditions:*

(i) Any population distribution $f_{SAN}(Z^3)$ implied by the SAN model is consistent with all the information in the panel and the refreshment samples.

(ii) The SAN attrition model implies a unique solution $f_{SAN}(Z^3)$.

(iii) Any solution $f_{SAN}(Z^3)$ is time-consistent.

(iv) The first order conditions imply the following restrictions on the observation hazards:

$$P(D_3 = 1|D_2 = 1, Z^3) = G_3(k_0 + k_1(Z^2) + k_2(Z_3)).$$

The functions $k_1(Z^2)$ and $k_2(Z_3)$ are arbitrary real-valued functions (squared summable/integrable) normalized to equal zero for some value in the support of Z^2 and Z_3, respectively, in order to allow for the inclusion of the intercept k_0 in the index.

Proof. See Appendix. □

The SAN attrition model has SMAR as a special case. Although this is clear from the theorem, it is instructive to see this from the optimization that characterizes the SAN solution. Consider the situation where the refreshment sample restriction (2.28) is not binding. Its (functional) Lagrange multiplier $k_2(Z_3)$ will then equal zero for all values of Z_3 . If the second period refreshment sample restrictions are also not binding in the second wave optimization, $f_s(Z^3)$ is consistent with the information available in both refreshment samples. As the solution to the SAN optimization problem is unique, it must equal $f_s(Z^3)$, with attrition hazards equal to those of SMAR. Note also that in this case the discrepancy equals zero, implying that the SMAR solution is independent of the choice of h.

If restriction (2.28) is binding, the attrition is non-ignorable and the choice of discrepancy will matter. In that case $f_{SAN}(Z^3)$ will differ from $f_s(Z^3)$ but (i) implies that the two distributions will be observationally equivalent in the panel. This is achieved by minimizing the discrepancy between them, as is shown in the proof. Compared to $f_s(Z^3)$, the advantage of the SAN solution is that it is also consistent with the information in the refreshment samples. It therefore satisfies all three properties mentioned at the start of this section, as well as time-consistency.

Just identification can be verified in the binary example: under Sequential Attrition 6 additional restrictions are required. Imposing AN in both periods corresponds to imposing the restrictions $\gamma_{12} = 0$ and $\delta_{13} = \delta_{23} = \delta_{123} = 0$. The last two restrictions come from the requirement that the solution be consistent with $P(Z_2 = 1)$ and $P(Z_3 = 1)$.

2.5.4 The Relationship between h_t and G_t

As mentioned above, the measure of discrepancy corresponding to a particular choice of h_t has a one-to-one relationship with G_t. The choice

$$h_t(a) = -(a - p_t)\ln(\frac{a}{p_t} - 1) + (a - p_t)$$

can be shown to correspond to $G_t(a) = G(a) = \frac{\exp(a)}{1+\exp(a)}$, a logit model for the attrition hazard in all waves. The constant p_t denotes $P(D_t = 1|D^{t-1} = 1, Z^{t-1})$. The logit model will be used in the application in section 2.8. The discrepancy in Hirano et al. (2001) requires $p_2 = P(D^2 = 1)$ to correspond with a logit model. This occurs because the SAN model maximizes over conditional distributions. This implies that $f_{SAN}(Z^2)$ and $f_{AN}(Z^2)$ will differ when derived from the same G_2. In terms of the restrictions imposed on the index function of the observation probability, the SAN model coincides with AN in panels with two waves.

2.6 Estimation of the SAN model by Weighted GMM

This section discusses the weighted GMM estimator. The discussion will focus on the just identified Method of Moment estimator. The over-identified GMM can be more challenging numerically but is conceptually the same. We will assume the following standard conditions to hold:

Assumption. *Let Z^T have support S_Z, a compact subset of $\Re^p$. Consider a k-vector of parameters of interest $\theta^* \epsilon \Theta$, where Θ is a compact subset of $\Re^k$, that uniquely solves a set of k moment equations $E[\phi(Z^T, \theta^*)] = 0$. The moment function $\phi : S_Z \times \Theta \to \Re^k$ is twice continuously differentiable with respect to θ and measurable in Z^T, and $E[\phi(Z^T, \theta^*)\phi(Z^T, \theta^*)']$ and $E[\frac{\delta\phi(Z^T,\theta^*)}{\delta\theta'}]$ are of full rank.*

Under MCAR attrition θ_0 can be estimated consistently by the analog estimator

$$\hat{\theta} \text{ such that } \frac{1}{N_{BP}} \sum_{i \epsilon BP} \phi(Z_i^T, \hat{\theta}) = 0.$$

If attrition depends on the values taken by one or more of the Z_t, the empirical distribution of the balanced panel will not be a consistent estimate of $f(Z^T)$, and the foregoing estimate is biased and inconsistent. However, using the identity $f(Z^T) = \frac{P(D^T=1)}{P(D^T=1|Z^T)} f(Z^T|D^T = 1)$,we have

$$E[\phi(Z^T, \theta_0)] = \int \phi(Z^T, \theta_0) \frac{P(D^T = 1)}{P(D^T = 1|Z^T)} f(Z^T|D^T = 1) dZ^T = 0,$$

so that the correct analog estimate becomes the weighted method of moments estimator

$$\hat{\theta} \text{ such that } \frac{1}{N_{BP}} \sum_{i \epsilon BP} w_i(Z_i^T, \alpha)\phi(Z_i^T, \hat{\theta}) = 0,$$

where the weights depend on the unknown parameter α and are equal to

$$w_i(\alpha) \equiv w_i(Z_i^T, \alpha) = \frac{P(D_i^T = 1)}{P(D_i^T = 1|Z_i^T, \alpha)}. \quad \textbf{(2.30)}$$

The dimension of α is discussed below. The key idea of the weighted GMM estimator is to apply GMM on the weighted balanced panel. These weights are constructed to ensure that the weighted balanced panel has a distribution that coincides with the population distribution $f(Z^T)$. This population distribution is identified only under restrictions on the observation probability $P(D^T = 1|Z^T)$. The attrition models discussed in section 2.4 and 2.5 nonparametrically just-identify this population distribution. Each of these models imply a potentially different set of weights and hence a different weighted GMM estimate.

Once the weights are estimated, it is straightforward to estimate θ. The next sub-section considers the estimation of the weights.

2.6.1 Estimation of the Weights

The weights are estimated in a procedure that involves T steps. These steps will be described below. Before that, note that the nonparametric just identification of the SAN model suggests that, in principle, weights can be constructed by estimating the functional Lagrange multipliers $k_1(Z^{t-1})$ and $k_2(Z_t)$. This corresponds to an infinite dimensional α in (2.30). For panels with two waves, Bhattacharya (2008) proposes a nonparametric sieve estimator. For larger T this approach becomes computationally infeasible and a flexible parametric approach is more attractive.

The key idea of the estimator proposed here is the following. A finite dimensional α can be obtained by replacing knowledge of the complete marginal distribution $f(Z_t)$, obtainable from the refreshment samples, by a finite set of moments of these distributions. Matching on only a few moments is unlikely to lead to the same solution, but, as more and more moments are matched, the SAN solution obtained in this way will get arbitrary close.[8] Moreover, if the parameters of interest are functionals of only first and second moments of the population distribution, as in a linear panel data model, matching on first and second moments suffices to find the exact solution. To see this, note that , although the SAN solution obtained in this way will differ from the SAN solution obtained by using all information available in the refreshment sample, they will agree up to first and second moments. In general, however, the order of moments required will depend on the population model of interest.

To estimate the weights, consider the denominator in (2.30). Sequential Attrition implies

$$w_i(\alpha) = \frac{\prod_{t=2}^{T} P(D_t = 1|D^{t-1} = 1)}{\prod_{t=2}^{T} P(D_t = 1|D^{t-1} = 1, Z^t, \alpha_t)} = \prod_{t=2}^{T} \frac{P(D_t = 1|D^{t-1} = 1)}{P(D_t = 1|D^{t-1} = 1, Z^t, \alpha_t)} \tag{2.31}$$

$$= \prod_{t=2}^{T} w_{ti}(Z^t, \alpha_t)$$

with $\alpha' = (\alpha_2', \ldots, \alpha_T')$. The weights depend on the parameters of the observation hazard, α_t, in each wave where attrition occurs.

In the first step, we estimate the moments of the distributions $f(Z_1), \ldots, f(Z_T)$. This can be done using the refreshment samples. For each t we can construct the vector $\bar{h}_{ti} = \bar{h}_t(Z_{ti})$ that has expectation h_t^* obtainable from $f(Z_t)$. By defining $h_{ti} = h_t(Z_{ti}) = (\bar{h}_{ti} - h_t^*)$, we have $E[h_{ti}] = 0$. Note that, as the re-

[8]Formally, $f(Z_t)$ needs to be uniformly integrable to be characterized by its moments (see van der Vaart (2000)).

freshment samples consist of random draws from the population distribution, these expectations are zero in the *target* population, not necessarily in the *sampled* population, i.e. the (wave t) balanced panel. A leading choice for h is to include raw moments up to some order but other choices are possible. For example, if Z_t only contains one variable, y_t, and first moments are matched, we have $\bar{h}_{ti} = y_{ti}$ and $h_{ti} = (y_{ti} - \mu_t^*)$. In the setup of panel data with attrition and refreshment samples, these moments will usually not be known with certainty. Consistent estimates $\hat{h}_t$ can easily be obtained by solving $\frac{1}{N_{rs}} \sum_{i=1}^{N_{rs}} (\bar{h}_{ti} - h_{t1}^*) = 0$. We can collect these estimates in $\hat{h} = (\hat{h}_1, \ldots, \hat{h}_T)$, an R by 1 vector of moments, and define $h_i = ((\bar{h}_{1i} - \hat{h}_1), \ldots, (\bar{h}_{Ti} - \hat{h}_T)) = (\bar{h}_i - h_0)$.

The second step starts by estimating α_2. We use the set of observations and variables that correspond to the balanced panel in wave 2, i.e. observations with $d_{i2} = 1$ (see Table 2.1). We look for weights $w_{2i}(\alpha_2)$ such that the weighted (second wave) balanced panel has sample moments of $\bar{h}(Z_1)$ and $\bar{h}(Z_2)$ equal to $\hat{h}_1$ and $\hat{h}_2$, respectively. From the literature on information theoretic alternatives to GMM (Qin and Lawless (1994), Imbens (1997), Imbens et al. (1998)), we have that $\alpha_2' = (\alpha_{21}', \alpha_{22}')$ are the Lagrange multipliers corresponding to these moment restrictions. Estimates of $\alpha_2' = (\alpha_{21}', \alpha_{22}')$ solve

$$
\begin{aligned}
\tfrac{1}{N_{BP2}} \sum_{i=1}^{N_{BP2}} w_{2i}(\alpha_2) h_{1i} &= \tfrac{1}{N_{BP2}} \sum_{i=1}^{N_{BP2}} \frac{\hat{P}(D_2 = 1)}{G(\alpha_{21}' h(Z_{1i}) + \alpha_{22}' h(Z_{2i}))} h(Z_{1i}) = 0 \\
\tfrac{1}{N_{BP2}} \sum w_{2i}(\alpha_2) h_{2i} &= \tfrac{1}{N_{BP2}} \sum_{i=1}^{N_{BP2}} \frac{\hat{P}(D_2 = 1)}{G(\alpha_{21}' h(Z_{1i}) + \alpha_{22}' h(Z_{2i}))} h(Z_{2i}) = 0 \\
\tfrac{1}{N_{BP2}} \sum (w_{2i}(\alpha_2) - 1) &= 0
\end{aligned}
\tag{2.32}
$$

where G denotes the logistic cdf and , $\hat{P}(D_2 = 1) = \hat{P}(D_2 = 1|D_1 = 1) = \frac{1}{N_p}\sum_{i=1}^{N_p} d_{2i}$, the estimated response fraction in wave 2. The weights $w_{2i}(\alpha_2)$ can now be constructed as

$$w_{2i}(\hat{\alpha}_2) = \frac{\hat{P}(D^2 = 1)}{G(\hat{\alpha}_{21}' h(Z_{1i}) + \hat{\alpha}_{22}' h(Z_{2i}))} = \frac{\hat{P}(D^2 = 1)}{\hat{P}(D_2 = 1|Z^2, \hat{\alpha}_2)}$$

They force the moments of $\bar{h}(Z_1)$ and $\bar{h}(Z_2)$ to equal $\hat{h}_1$ and $\hat{h}_2$ in the *weighted* sampled population. With $\check{h}_{2i}' = (h_{1i}', h_{2i}')$, they force $w_{2i}(\alpha_2)\check{h}_{2i}$ to have average zero in the (second-period) balanced panel. The weights themselves are normalized to average to one. This implies that the weights $\tilde{w}_{2i} = {}^{w_{2i}(\alpha)}/_{N_{BP2}}$ sum to 1, with $\sum_{i=1}^{N_{BP2}} \tilde{w}_{2i}(\hat{\alpha}_2)\bar{h}_{2i} = \hat{h}_2$. The weights in later waves are normalized in a similar way.

By moving from refreshment sample marginal distributions to their corresponding marginal moments, we have essentially replaced the functions $k_1(Z_1)$ and $k_2(Z_2)$ by polynomials of chosen order in moments of Z_1 and Z_2. The SAN attrition model in the third wave involves $k(Z_1, Z_2)$. The moments-equivalent therefore involves cross-moments, such as $E[y_1 y_2]$. Although these moments are not available from the refreshment samples, they are identified by the SAN model.[9] In the second part of the second step the weights $w_{2i}(\hat{\alpha}_2)$ that are estimated in the first part, are used to estimate these moments. If we collect the necessary cross-moments in the vector h_{c2}^*, an estimate $\hat{h}_{c2}$ is obtained by solving $\frac{1}{N_{BP2}}\sum_{i=1}^{N_{BP2}} w_2(\hat{\alpha}_2)(\bar{h}_{c2}(Z_1, Z_2) - h_{c2}^*) = 0$.

Like the second step, the third step estimates α_3 as well as the cross moments

[9]The distribution $f(Z_1, Z_2)$ that is identified by the SAN model in its current form is the distribution that is as close as possible to the balanced panel while being consistent with $\hat{h}_1$ and $\hat{h}_2$.

needed for $\hat{\alpha}_4$. Only estimation of α_3 needs clarification. The third period SAN attrition model involves the moments corresponding to $k(Z_1, Z_2)$ and $k(Z_3)$. The marginal moments of Z_1 and Z_2 are estimated by $\hat{h}_1$ and $\hat{h}_2$ above, the cross moments by $\hat{h}_{c2}$. If we collect all these moments in $\check{h}_{31}$, we have that $\check{h}_{31i} = (\bar{h}_{31}(Z_1, Z_2) - \check{h}_{31})$ has zero mean in the target population. The marginal moments of Z_3 are estimated by $\check{h}_{32} = \hat{h}_3$. With $\check{h}_3' = (\check{h}_{31}', \check{h}_{32})$, we have

$$\frac{1}{N_{BP3}} \sum_{i=1}^{N_{BP3}} w_{2i}(\hat{\alpha}_2) w_{3i}(\alpha_3) \check{h}_{31i} =$$

$$\frac{1}{N_{BP3}} \sum_{i=1}^{N_{BP3}} w_{2i}(\hat{\alpha}_2) \frac{\hat{P}(D_3 = 1 | D_2 = 1)}{G(\alpha_{31}' \check{h}_{31i}(Z_{1i}, Z_{2i}) + \alpha_{22}' \check{h}_{32}(Z_{3i}))} \check{h}_{31}(Z_{1i}, Z_{2i}) = 0$$

$$\frac{1}{N_{BP3}} \sum_{i=1}^{N_{BP3}} w_{2i}(\hat{\alpha}_2) w_{3i}(\alpha_3) \check{h}_{32i} =$$

$$\frac{1}{N_{BP3}} \sum_{i=1}^{N_{BP3}} w_{2i}(\hat{\alpha}_2) \frac{\hat{P}(D_3 = 1 | D_2 = 1)}{G(\alpha_{31}' \check{h}_{31i}(Z_{1i}, Z_{2i}) + \alpha_{32}' \check{h}_{32}(Z_{3i}))} \check{h}_{32}(Z_{3i}) = 0$$

$$\frac{1}{N_{BP3}} \sum (w_{2i}(\hat{\alpha}_2) w_{3i}(\alpha_3) - 1) = 0$$

Estimates of $\alpha_3' = (\alpha_{31}', \alpha_{32}')$ solve these equations. With $\check{h}_{3i}' = (h_{31i}', h_{32i}')$ they ensure that $w_{2i}(\hat{\alpha}_2) w_{3i}(\hat{\alpha}_3) \check{h}_{3i}$ averages to zero in the (third wave) balanced panel. Again, the weights must average to one.

This procedure continues until the last parameters, α_T are estimated, allowing us to calculate $w_T(\hat{\alpha}_T)$. This gives us the estimated weights $w_i(\hat{\alpha}) = \Pi_{t=2}^{T} w_{ti}(\hat{\alpha}_t)$.

2.6.2 Estimation of the Parameters of Interest θ

Once the weights are estimated, it is easy to find point estimates of the parameters of interest θ. The weights ensure that the weighted balanced panel can be considered to be random draws from the population distribution of interest. The weighted GMM estimator $\hat{\theta}_{WGMM}$ solves the set of equations

$$\frac{1}{N_{BP}} \sum_{i=1}^{N_{BP}} w_i(\hat{\alpha}) \phi(Z_i^T, \theta) = 0.$$

Note that all method of moments estimates of the α's are just-identified by construction. The number of moment conditions in ϕ may well be larger than the dimension of θ. In that case, point estimates and standard errors can only be obtained by solving the estimating equations for all parameters simultaneously. Obviously, the choice of weighting matrix will also influence the estimates in that case. The next section summarizes the estimating equations described in this section and derives the approximate repeated sampling behaviour of the estimator.

2.7 The Sampling Distribution of the Weighted GMM Estimator

From the refreshment samples we obtain auxiliary information in the form of the estimate $\hat{h}' = (\hat{h}'_1, \hat{h}'_2, \ldots, \hat{h}'_T)$, together with an approximation of its variance matrix $\triangle/N_{rs}$. As the refreshment samples consist of random draws from the population of interest, $\hat{h} - h_0$ is independent of the observations in the panel, $\left\{Z_i^T\right\}_{i=1}^{N_p}$. As the auxiliary moments h_0 are not known but estimated, we require the vector $h(Z^T) = \bar{h}(Z^T) - \hat{h}$ to have expectation zero. The repeated sampling behaviour of our estimator is approximated by letting N_p and N_{rs} go to infinity with their ratio $^{N_{rs}}/_{N_p}$ converging to a constant k. This is the only case of practical interest because if the sample size of the refreshment samples were to increase at a faster rate than the sample size in the panel, then in large samples the sampling variation in $\hat{h}$ could be ignored. In the opposite case the auxiliary information would not be informative. For ease of exposition, but without loss of generality, we take k to equal some integer value. We can then think of our observations in the panel as consisting of Z_i^T and $(\bar{h}_{i1}, \ldots, \bar{h}_{ik})$ for $i = 1, \ldots, N_p$. The row-vector $\bar{h}_{1i}$, for instance, contains observations on all variables observed in all the refreshment samples. In particular, it contains the i^{th} observation from the first N_p such observations. The i^{th} observation of the last N_p such observations are in $\bar{h}_{ik}$. With $\psi' = (\psi_1', \ldots, \psi_{T+1}') = (h_0', \tilde{\alpha}_2', \ldots, \tilde{\alpha}_T', \theta')$, $\tilde{\alpha}_t' = (\hat{\alpha}_t', \hat{h}_{ct}')$, and $\tilde{\alpha}_T = \hat{\alpha}_T$, the estimating equations for $\hat{\psi}$ are

$$
g(\hat{\psi}) = \begin{pmatrix} g_1(\hat{h}) \\ g_2(\tilde{\alpha}_2) \\ g_3(\tilde{\alpha}_3) \\ \vdots \\ g_T(\tilde{\alpha}_T) \\ g_{T+1}(\hat{\theta}) \end{pmatrix} = \frac{1}{N_p}\sum_{i=1}^{N_p} \begin{pmatrix} \frac{1}{k}\sum_{j=1}^{k}(\bar{h}_{ij} - \hat{h}) \\ \begin{bmatrix} d_{2i}w_{2i}(\hat{\alpha}_2)(\bar{h}_{2i} - \hat{h}_2) \\ d_{2i}(w_{2i}(\hat{\alpha}_2) - 1) \\ d_{2i}w_{2i}(\hat{\alpha}_2)(\bar{h}_{c2i} - \hat{h}_{c2}) \end{bmatrix} \\ \begin{bmatrix} d_{3i}w_{2i}(\hat{\alpha}_2)w_{3i}(\hat{\alpha}_3)(\check{h}_{3i} - \check{h}_3) \\ d_{3i}(w_{2i}(\hat{\alpha}_2)w_{3i}(\hat{\alpha}_3) - 1) \\ d_{3i}w_{2i}(\hat{\alpha}_2)w_{3i}(\hat{\alpha}_3)(\bar{h}_{c3i} - \hat{h}_{c3}) \end{bmatrix} \\ \vdots \\ \begin{bmatrix} d_{Ti}w_i(\hat{\alpha})(\check{h}_{Ti} - \check{h}_T) \\ d_{Ti}(w_i(\hat{\alpha}) - 1) \end{bmatrix} \\ \begin{bmatrix} d_{Ti}w_i(\hat{\alpha})\phi(\hat{\theta}) \end{bmatrix} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}
$$

where $\tilde{\alpha}_t' = (\hat{\alpha}_t', \hat{h}_{ct}')$ for $t = 1, \ldots, T-1$, $\tilde{\alpha}_T = \hat{\alpha}_T$ and $\tilde{\alpha}' = (\tilde{\alpha}_2', \ldots, \tilde{\alpha}_T')$. The vector $\check{h}_t$ collects all the moments and cross-period moments estimated up and until wave $t-1$ as well as the moments estimated from the refreshment sample of wave t, i.e. $\check{h}_t' = (\check{h}_{t-1}', \hat{h}_{c,t-1}', \hat{h}_t')$. The SAN model does not require cross-period moments to correct for attrition in the second wave. Cross-period moments estimated in wave $t-1$ are used to estimate $\hat{\alpha}_t$ and are for that reason included in $\check{h}_t$. Moreover, the weights $w_{ti}(\hat{\alpha}_t)$ are required to average to 1 for all t. Solving the first equation leads to $\hat{h} = \frac{1}{kN_p}\sum_{i=1}^{N_p}\sum_{j=1}^{k}\bar{h}_{ij}$. Solving the last set of equations amounts to estimating θ by the method of moments using the weighted Balanced Panel.

The following two theorems describe the approximate repeated sampling behaviour of the estimator in large samples.

Theorem 4. *Let the observations in the panel and the refreshment samples be* *iid with the empirical distribution from the panel converging to the sampled population distribution. When N_p and N_{rs} go to infinity with $N_{rs}/N_p = k$, and standard regularity conditions hold, we have*

$$\begin{pmatrix} \hat{h} \\ \tilde{\alpha} \\ \hat{\theta} \end{pmatrix} \xrightarrow{p} \begin{pmatrix} h_0 \\ \tilde{\alpha}_0 \\ \theta_0 \end{pmatrix}$$

Proof. See Appendix. □

Theorem 5. *When the moment conditions satisfy standard regularity conditions, the conditions stated in theorem 4 hold, and the matrices below are of full rank, the asymptotic distribution of the weighted GMM estimator is given by*

$$\sqrt{N}(\begin{pmatrix} \hat{h} \\ \tilde{\alpha} \\ \hat{\theta} \end{pmatrix} - \begin{pmatrix} h_0 \\ \tilde{\alpha}_0 \\ \theta_0 \end{pmatrix}) \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Gamma^{-1}\Omega(\Gamma')^{-1}\right)$$

where $\Gamma = \begin{bmatrix} -I_R & 0 & 0 & \cdots & 0 \\ G_{21} & G_{22} & 0 & \cdots & 0 \\ G_{31} & G_{32} & G_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ G_{T+1,1} & G_{T+1,2} & G_{T+1,3} & \cdots & G_{T+1,T+1} \end{bmatrix}$ *and*

$$\Omega = \begin{bmatrix} \triangle/k & 0 & 0 & 0 \\ 0 & \Omega_{22} & \ldots & \Omega_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \Omega_{T+1,2} & \cdots & \Omega_{T+1,T+1} \end{bmatrix}.$$

These matrices are partitioned conformably with the stacked moment conditions. Then, for all $s, t \epsilon \{2, \ldots T+1\}$ *the matrices* $G_{t,t-r}$ *in* Γ *are defined as* $E_s\left[\frac{\delta g_t}{\delta \psi_{t-r}}\right]$ *for* $r = 0, \ldots, T$ *and* $\Omega_{st} \equiv E_s\left[g_s g_t'\right]$, *where* $E_s[\cdot]$ *denotes expectations taken with respect to the sampled population distribution. The matrices* $\tilde{G}_{t,t-r}$ *are the corresponding block matrices in* Γ^{-1}. *They can be found by the following recursion formulae:*

$$\begin{aligned} \tilde{G}_{tt} &= G_{tt}^{-1} & r &= 0 \\ \tilde{G}_{t,t-r} &= -\tilde{G}_{tt} \sum_{l=t-r}^{t-1} G_{tl} \tilde{G}_{l,t-r} & r &> 0. \end{aligned}$$

The inverse of Γ *can therefore be obtained by inverting its nonzero block-matrices.*

Proof. See Appendix. □

2.8 Application to the DTP

A detailed description of The Dutch transportation Panel (DTP) can be found in Meurs and Ridder (1992) and Ridder (1992). The original purpose of the DTP was to evaluate the effect of price increases on the use of public transportation. Every member of the households that cooperated was asked to report all trips during a particular week. A trip starts when the home is left and ends on returning home. It is counted irrespective of the means of transportation chosen. Table 2.2 shows the missing data pattern. Only 1037 of the 1770 households that responded in the first wave continued to respond in the

Obs. Indicators			Count	Percentage
D_1	D_2	D_3		
1	1	1	859	49%
1	1	0	178	10%
1	0	0	733	41%
0	1	1	479	
0	1	0	176	
0	0	1	515	

Table 2.2: Missing data pattern in the Dutch Transportation Panel. The Percentage column only refers to observations in the panel.

second wave. To offset this attrition a refreshment sample was drawn consisting of 655 households. These households were also approached in the third wave, leading to a refreshment panel. Only the second wave cross-section is used here. In the third wave another 178 households dropped out of the panel. A second refreshment sample was obtained consisting of 515 households.

Following Hirano et al. (2001), we define Z_t to be a binary indicator variable that equals 1 if the total number of trips during the survey week was less than or equal to 25. The parameters of interest are the probabilities $\pi_{z_1,z_2,z_3} \equiv P(Z_1 = z_1, Z_2 = z_2, Z_3 = z_3)$ that define the joint probability distribution. Below, I present estimates of the feature π_{000}, obtained under different sets of identifying assumptions. In this binary example, the SAN model maintains $\gamma_3 = \gamma_{12} = \gamma_{13} = \gamma_{23} = \gamma_{123} = 0$ and $\delta_{13} = \delta_{23} = \delta_{123} = 0$ in (2.4). Estimates of the remaining parameters are given in Table 2.3. The binary nature of the variables implies that only first moments need to be matched. To illustrate the potential benefits of collecting refreshment samples, the SMAR and MCAR estimates are also given. The SMAR estimates use all the information in the panel, but not the refreshment samples, while MCAR only uses the balanced panel.

Population	MCAR coeff.	MCAR s.e.	SMAR coeff.	SMAR s.e.		SAN coeff.	SAN s.e.	
π_{000}	0.75	0.035	0.70	0.012		0.64	0.037	
Attrition (t=2)								
$const$			0.55	0.06	*	0.06	0.07	*
Z_1			-1.03	0.12	*	-0.89	0.25	*
Z_2						-0.33	0.38	
Attrition (t=3)								
$const$			1.69	0.09	*	1.66	0.09	*
Z_1			-0.33	0.34		-0.83	0.28	*
Z_2			-0.09	0.34		-0.38	0.26	
Z_1Z_2			-0.17	0.52		-0.07	0.42	
Z_3						-0.61	0.25	*

Table 2.3: Estimates of π_{000} obtained using the MCAR, SMAR and SAN attrition models. The * indicates that the coefficient is significantly different from zero at the 5% level.

The parameter π_{000} represents the fraction of households in the population that reported more than 25 trips in each of the three waves. The estimation results in Table 2.3 suggest that these households are over-represented in the balanced panel. Ignoring the attrition problem gives an estimate of 75%. The SMAR estimates show that a lower estimate is obtained in the sampled population distribution. The SAN model, that also requires consistency with the refreshment samples, provides a further downward correction to 64%; over-representation persists to some extent even in the sampled population distribution. A selection on observables approach ignores this. Inspection of the attrition parameter estimates reveals that the selection on observables hypothesis is not rejected in the second wave. It is the drop-out in the third wave that seems to depend on unobservable characteristics of the households.

2.9 Conclusion

Selection bias due to attrition can be mitigated or even avoided by using attrition models that are sufficiently unrestrictive to allow for a wide range of potential selection. Hirano et al. (2001) propose the Additively Nonignorable (AN) attrition model to correct for the potential selectivity of the attrition in panels with two waves. Generalizing their identification strategy to panels with more than two waves is shown to lead to an attrition model that is over-identified, does not encompass SMAR and is not time-consistent.

In section 2.5 of this chapter, the Sequential Additively Nonignorable attrition model was developed. The SAN model has three key properties. First, the model identifies the population distribution in panels with any number of waves and has no testable implications in the panel. Second, the implied population distribution is consistent with the refreshment samples. Third, it encompasses SMAR. It is also shown to be time-consistent.

A weighted GMM estimator is proposed and its consistency and asymptotic normality were derived. Application to the Dutch Transportation Panel suggested that attrition in the DTP is nonignorable. Ignoring the attrition would involve a balanced panel in which households that make relatively many trips in each of the three waves are over-represented.

Appendix

Proof of Theorem 1:

The proof is analogous to the proof in Hirano et al. (2001). The continuous case follows directly from the discrete case under the qualifications mentioned in the theorem. The MAN optimization maximizes a strictly concave functional defined on the vector space of squared summable sequences l_2 (the space L_2 of squared Lebesgue integrable functions from R^3 to R in the continuous case). The inequality constraints are defined using the convex cone of nonnegative functions defined on R^3. With the l_2 (L_2) norm this cone is regular. The functional is maximized over the convex set defined by the inequality restrictions and the linear equality restrictions. The solution, if it exists, is unique. As l_2 (L_2) is a Hilbert space, and hence complete, the existence is established by the monotone fixed point theorem (Hutson and Pym (1980)). Because h is differentiable, the maximand is Fréchet differentiable. The stationary point of the Lagrangian satisfies

$$h'\left(\frac{f(Z^3)}{f(Z^3|D^3=1)}\right) = k_0 + k_1(Z_1) + k_2(Z_2) + k_3(Z_3). \tag{2.33}$$

The functions $k_1(Z_1)$, $k_2(Z_1)$, $k_3(Z_1)$ and k_0 are the (functional) Lagrange multipliers corresponding to the restrictions. They are determined by (2.33) up to a normalization. The normalization mentioned in the theorem is necessary as restriction (2.14) renders one of the restrictions in (2.15), (2.16) and (2.17) redundant. The functional Lagrange multipliers take values in the dual space of l_2 (Luenberger (1969)). As the dual space of l_2 (and L_2) is isometrically isomorphic to itself, it follows that the k functions are squared summable (integrable). Equation (2.33) implies

$$f(Z^3) = (h')^{-1}(k_0 + k_1(Z_1) + k_2(Z_2) + k_3(Z_3))f(Z^3|D^3 = 1)$$

and, using (2.1) we have

$$P(D^3 = 1|Z^3) = \frac{P(D^3 = 1)}{(h')^{-1}(k_0 + k_1(Z_1) + k_2(Z_2) + k_3(Z_3))} \equiv G(k_0+k_1(Z_1)+k_2(Z_2)+k_3(Z_3)).$$

Proof of Theorem 2:

Uniqueness follows from arguments identical to those in the proof of Theorem 1. The stationary point of the Lagrangian satisfies

$$h_3'\left(\frac{f(Z^3)}{f(Z^3|D^3 = 1)}\right) = k_0 + k_1(Z_1, Z_2) + k_3(Z_3) \tag{2.34}$$

which implies

$$f(Z^3) = (h_3')^{-1}(k_0 + k_1(Z_1) + k_2(Z_2) + k_3(Z_3))f(Z^3|D^3 = 1).$$

Using (2.1) the restrictions on the observation probabilities are

$$
\begin{aligned}
P(D^3 &= 1|Z^3) = \frac{P(D^3=1)}{(h_3')^{-1}(k_0+k_1(Z_1,Z_2)+k_3(Z_3))} \\
&\equiv G_3(k_0+k_1(Z_1)+k_2(Z_2)+k_3(Z_3)).
\end{aligned}
$$

To verify that the TCAN attrition model has no testable implications, consider the equations (2.7), (2.8) and (2.9). The left-hand side of equation (2.7) is obtained by multiplying the TCAN solutions mentioned above. As $(h_3')^{-1}(k_0 + k_1(Z_1, Z_2)+k_3(Z_3))$ cancels, the resulting expression equals the right-hand side of the equation. In equation (2.8) we have

$$
\begin{aligned}
&\sum_{Z_3}\left(1-\frac{P(D^3=1|Z^3)}{P(D_2=1|Z^2)}\right)P(D_2=1|Z^2)f(Z^3) \\
&= f(Z^2, D_2=1) - \sum_{Z_3} f(Z^3, D^3=1) \\
&= f(Z^2, D_2=1) - f(Z^2, D^3=1) \\
&= P(D_2=1, D_3=0)f(Z^2|D_2=1, D_3=0).
\end{aligned}
$$

Finally, equation (2.9) gives

$$
\begin{aligned}
&\sum_{Z_2}\left(1-P(D_2=1|Z^2)\right)f(Z^2) \\
&= \sum_{Z_2}\left(1-\frac{P(D_2=1)}{(h_2')^{-1}(\cdot)}\right)(h_2')^{-1}(\cdot)f(Z^2|D_2=1) \\
&= f(Z_1) - f(Z_1, D_2=1) \\
&= P(D_2=0)f(Z_1|D_2=0),
\end{aligned}
$$

ignoring the index function in the notation. This completes the proof.

Proof of theorem 3:

For condition (ii), uniqueness of the solution, note that squared summability of the product of two squared summable sequences follows from the Cauchy-Schwartz inequality for l_2. The same is true for squared Lebesgue integrability in L_2. With this in mind, uniqueness follows from squared summability of $f(Z_1)$ and arguments identical to those in the proof of Theorem 1. Time consistency follows directly from the restrictions (2.27) in the optimization. Substituting the Sequential Attrition restriction leads to an optimization over $f(Z_3|Z^2, D_2 = 1)$. The stationary point of the Lagrangian then satisfies

$$\bar{f}(Z^2)\left[h'\left(\frac{f(Z_3|Z^2, D_2 = 1)}{f(Z_3|Z^2, D^3 = 1)}\right) - (k_0 + k_1(Z_1, Z_2) + k_3(Z_3))\right] = 0. \tag{2.35}$$

On the support of $f(\bar{Z}^2)$ this implies

$$f(Z_3|Z^2, D_2 = 1) = (h')^{-1}(k_0 + k_1(Z_1) + k_2(Z_2) + k_3(Z_3))f(Z_3|Z^2, D^3 = 1).$$

Note that

$$f(Z_3|Z^2, D_2 = 1) = \frac{P(D_3 = 1|D_2 = 1, Z^2)}{P(D_3 = 1|D^2 = 1, Z^3)}f(Z_3|Z^2, D^3 = 1).$$

The restrictions on the observation hazards follow from combining the last two equations:

$$\begin{aligned} P(D_3 &= 1|D_2 = 1, Z^3) = \frac{P(D_3 = 1|D_2 = 1, Z^2)}{(h')^{-1}(k_0 + k_1(Z_1, Z_2) + k_3(Z_3))} \\ &\equiv G_3(k_0 + k_1(Z_1, Z_2) + k_3(Z_3)). \end{aligned}$$

This shows condition (iv) in the Theorem.

Condition (i) consists of two parts. Consistency with the refreshment samples is directly imposed in the optimization. Consistency with the information in the panel can be verified using equations (2.7), (2.8) and (2.9). Minimization of the discrepancy ensures that these equations are satisfied. Indeed, the first order conditions imply that the SAN solution satisfies $f(Z_3|Z^2) = (h_3')^{-1}(k_0 + k_1(Z_1, Z_2) + k_3(Z_3))f(Z_3|Z^2, D^3 = 1)$ and $P(D_3 = 1|D_2 = 1, Z^3) = \frac{P(D_3=1|D_2=1,Z^2)}{(h_3')^{-1}(k_0+k_1(Z_1,Z_2)+k_3(Z_3))}$. The second wave solutions are defined in the same way. Substitution of these solutions in the left-hand side of equation (2.7) gives

$$\begin{aligned} \frac{P(D_3 = 1|D_2 = 1, Z^2)}{(h_3')^{-1}(\cdot)} \frac{P(D_2 = 1|Z_1)}{(h_2')^{-1}(\cdot)} (h_3')^{-1}(\cdot) f(Z_3|Z^2, D^3 &= 1) \\ (h_2')^{-1}(\cdot) f(Z_2|Z_1, D_2 = 1) f(Z_1), \end{aligned}$$

where index functions are ignored in the notation. The functions $(h_2')^{-1}(\cdot)$ and $(h_3')^{-1}(\cdot)$ cancel, and substitution of

$$f(Z_2|Z^2, D_2 = 1) = \frac{P(D_3 = 1|D_2 = 1, Z_1)}{P(D_3 = 1|D_2 = 1, Z^2)} f(Z_2|Z_1, D^3 = 1)$$

and

$$f(Z_1) = \frac{P(D_3 = 1|D_2 = 1)P(D_2 = 1)}{P(D_3 = 1|D_2 = 1, Z_1)P(D_2 = 1|Z_1)} f(Z_1|D^3 = 1)$$

gives the result.

In equation (2.8), Sequential Attrition implies

$$\sum_{Z_3} \left(1 - \frac{P(D_3 = 1|D_2 = 1, Z^2)}{(h_3')^{-1}(\cdot)}\right) P(D_2 = 1|Z^3) f(Z^3) = \frac{P(D_2 = 1|Z_1)}{(h_2')^{-1}(\cdot)} f(Z^2) -$$
$$\sum_{Z_3} \frac{P(D_3 = 1|D_2 = 1, Z^2)}{(h_3')^{-1}(\cdot)} \frac{P(D_2 = 1|Z_1)}{(h_2')^{-1}(\cdot)} (h_3')^{-1}(\cdot) f(Z_3|Z^2, D^3 = 1)$$
$$(h_2')^{-1}(\cdot) f(Z_2|Z_1, D_2 = 1) f(Z_1).$$

The functions $(h_2')^{-1}(\cdot)$ and $(h_3')^{-1}(\cdot)$ cancel and $f(Z_3|Z^2, D^3 = 1)$ sums to 1. We then obtain

$$f(Z^2, D_2 = 1) - f(Z^2, D^3 = 1) = P(D_2 = 1, D_3 = 0) f(Z^2|D_2 = 1, D_3 = 0).$$

Finally, using Sequential Attrition and the absence of return, the left-hand side of equation (2.9) obeys

$$\begin{aligned}
\sum_{Z_2}\sum_{Z_3} P(D_2 &= 0|Z^3)f(Z^3) = \sum_{Z_2} P(D_2 = 0|Z^2)f(Z^2) \\
&= \sum_{Z_2}\left(1 - \frac{P(D_2 = 1|Z_1)}{(h_2')^{-1}(\cdot)}\right)(h_2')^{-1}(\cdot)f(Z_2|Z_1, D_2 = 1)f(Z_1) = \\
&= f(Z_1) - \sum_{Z_2} P(D_2 = 1|Z_1)f(Z_2|Z_1, D_2 = 1)f(Z_1) \\
&= f(Z_1)(1 - P(D_2 = 1|Z_1)) = P(D_2 = 0)f(Z_1|D_2 = 0).
\end{aligned}$$

This concludes the proof.

Proof of Theorem 4:

The vector of stacked moment conditions has expectation zero at the true parameter values. The regularity assumptions include compactness of the parameter space and continuity of the moment functions in their parameters (almost everywhere). Standard GMM theory then provides the result (see Newey (1984) and Newey and McFadden (1994)).

Proof of theorem 5:

The regularity conditions now include twice continuous differentiability of the moment conditions in a neighbourhood of the true value, with probability approaching one. Theorem 6.1 of Newey and McFadden (1994) implies the asymptotic distribution. The recursion formulae follow from partitioned inversion of Γ.

Chapter 3

Correcting for Attrition Bias with Refreshment Samples in the ELSA Panel

3.1 Introduction

Concerns about the population ageing and early retirement of older workers from the labor market abound in Britain and in most of the Western world. As a consequence, there is a considerable literature that considers such things as changes to the retirement age, incentives to encourage pension saving and the relationship between health and retirement (Banks et al. (1998), French (2005), Rice et al. (2010)). Studies that investigate these issues using panel data need to take attrition into account. In this chapter I study the transition out of the labor market into inactivity of elderly people using the English

Longitudinal Study of Ageing (ELSA). Evidence of selection in this example would suggest that attrition in the ELSA panel is nonrandom.

The outline of this chapter is as follows. Section 2 gives a short description of the ELSA panel together with a description of the way I handle item nonresponse and return. In section 3 some preliminary evidence of the effect of attrition is discussed. Section 4 defines the attrition models considered in this chapter and later sections describe my estimation results.

3.2 The ELSA Panel

The ELSA sample was designed to represent people aged 50 and over, living in private households in England. The sampling frame consists of the three waves of the Health Survey for England: 1998, 1999 and 2001. These samples were nationally representative. The first wave of the ELSA panel took place in 2002-2003. Eligible sample members who responded at this stage are called the Core Members of the ELSA panel. Later waves were obtained in 2004-2005, 2007-2008 and 2009-2010, giving a current total of four waves.

3.2.1 The Data

ELSA collects a large variety of information from its respondents. The interviews contain questions relating to health, social participation, work and pension, income and assets, housing cognitive function, expectations, psychosocial health and demographics. The appendix gives an overview of the selected set of variables that I use in this chapter and how they were derived.

I use only observations relating to core members for which a full interview was conducted in person.[1] In the third and fourth wave a refreshment sample was collected. These are new random samples from the population of interest that are collected in order to refresh the sample members that were lost due to attrition. The third wave refreshment sample consists only of small subset of age-cohorts, and is for that reason not used here.

The ELSA data can be obtained via the Economic and Social Data Service (ESDS). Extensive documentation is available in the Technical reports (Taylor et al. (2007), Sholes et al. (2008), Sholes et al. (2009) and Banks et al. (2010)) and the corresponding User Guides.

3.2.2 Attrition and Persistence Rates

For the purpose of studying attrition, each individual can be classified as belonging to one of 9 sub-populations. One can can distinguish between the balanced panel (BP), consisting of individuals that responded in all four waves, and three incomplete panels (IP2, IP3, IP4). Individuals in IP4 attrit from the panel in the fourth wave. The sub-populations IP3 and IP2 are defined similarly. Some members of IP3 and IP2 returned to the panel in later waves. Table 1 depicts all resulting sub-populations, together with their frequency of occurrence. The indicator variable D_1 takes the value 1 in sub-populations for

[1]Some interviews were conducted by proxy. This means that a person other than the respondent, responded on behalf of the core sample member. This occurred when the respondent was physically or cognitively impaired or in hospital or temporary care. As only a subset of the questions were asked, including such observations would increase item non-response considerably and potentially inflate measurement error. For these reasons proxy interviews are not considered in this study.

which an interview was successfully issued in the first wave.[2] The variables D_2, D_3 and D_4 are defined in the same way for later waves. In the first wave, a total of 11029 full personal interviews were successfully issued.

Sub-population	D_1	D_2	D_3	D_4	Freq	%
Balanced panel (BP)	1	1	1	1	5695	52%
Incomplete Panel 4 (IP4)	1	1	1	0	1204	11%
Incomplete Panel 3 (IP3)	1	1	0	0	1369	12%
return in 4 (IP3R4)	1	1	0	1	238	2%
Incomplete Panel 2 (IP2)	1	0	0	0	2072	19%
return in 3 (IP2R3)	1	0	1	0	123	1%
return in 4 (IP2R4)	1	0	0	1	144	1%
return in 3 and 4 (IP2R34)	1	0	1	1	184	2%
Total panel					11029	100%
Refreshment Sample 4 (RS4)	0	0	0	1	2230	100%

Table 3.1: Response patterns for core members of the ELSA panel.

From Table 3.1 we can construct the rates of persistence in staying in the panel: the fraction of individuals that continue to respond (i.e. full interview successfully issued) in the next wave out of those that responded in the current wave. These rates of persistence are given in Table 3.2.

Persistence	Freq	%
wave 1	11029	
wave 1 - wave 2	8506	77%
wave 2 - wave 3	6899	81%
wave 3 - wave 4	5695	83%

Table 3.2: Persistence rates for core members of the ELSA panel.

[2]The ELSA data contain the variables indoutw1, indoutw2, indoutw3 and indoutw4 that indicate whether or not a full interview was successfully issued for this individual in the four respective waves. These variables were used for the calculations in Table 1. If a full interview was successfully issued, there may still be item-nonresponse.

3.2.3 Item Nonresponse

Even respondents for which a full interview was successfully issued may not have answered all questions in the interview, leading to item-nonresponse. This different type of missing data can severely hamper the analysis of potential bias due to attrition if not properly taken into account. To illustrate the problem, consider a univariate analysis on some variable, X say. Attrition analysis considers an individual that responds in all waves to belong to the balanced panel sub-population. This is because the respondent has shown willingness to respond to the survey in all four waves. However, if this individual has *item*-nonresponse in the fourth wave on X, it cannot be distinguished from an IP4 individual in analyses that involve only X. This implies that correction for attrition becomes dependent on the item-nonresponse pattern, a different type of missing data altogether. It is therefore desirable to make a clear separation between item non-response and attrition.

First, I will assess to what extent item nonresponse influences attrition patterns in the ELSA panel. Table 3.3 shows the non-response pattern for a number of variables separately. Again, precise descriptions of these variables and how they were derived can be found in the appendix.

From Table 3.3 it is clear that item-nonresponse has a strong impact on the resulting response patterns. A balanced panel of 5695 observations can be reduced to 1917 just by including the variable longill into the analysis. Note that the percentages in Table 3.3 no longer add up to 100% because many new response patterns that do not appear in Table 3.1 can now occur due to item-nonresponse.

	phealth		longill		mhealth		work		age		holiday		clubs	
Sub-population	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%
BP	2857	26%	1917	17%	5709	52%	5780	52%	5927	54%	4327	39%	3874	35%
IP4	629	6%	714	6%	1189	11%	1125	10%	1153	10%	1180	11%	1124	10%
IP3	653	6%	882	8%	1343	12%	1287	12%	1289	12%	1292	12%	1210	11%
IP3R4	116	1%	286	3%	230	2%	255	2%	233	2%	464	4%	486	4%
IP2	1007	9%	1734	16%	2047	19%	1933	18%	1950	18%	2145	19%	2165	20%
IP2R3	66	1%	234	2%	131	1%	137	1%	131	1%	248	2%	300	3%
IP2R4	64	1%	210	2%	139	1%	163	1%	160	1%	187	2%	214	2%
IP2R34	86	1%	264	2%	185	2%	229	2%	186	2%	318	3%	387	4%
Total Panel	5478	50%	6241	57%	10937	99%	10909	99%	11029	100%	10161	92%	9760	88%

Table 3.3: Response pattern frequencies and percentages by variable. In the first wave, a total of 11029 full personal interviews were successfully issued.

Most analyses use more than one variable. For such multivariate analyses, the instance of item nonresponse on only one variable can affect the whole simultaneous response pattern. Table 3.5 reports the simultaneous response rates for a few *sets* of variables. The variables included in each set are indicated in Table 3.4. The variables income and wealth are currently not available in wave 4.

As expected, the number of available observations within each sub-population drops even more dramatically.

Table 3.3 and Table 3.5 show two important consequences of item non-response. Firstly, nonresponse leads to a substantially lower amount of usable responses, especially when a multivariate analysis is conducted. Secondly, it can lead to re-assignment of observations over the sub-populations in Table 3.5 that can invalidate attrition analysis. For attrition analysis, it is crucial that members of one sub-sample (e.g. IP4) do not get re-assigned to another sub-sample (e.g. IP3, because the fourth wave suffered from item-nonresponse). Indeed, what matters for attrition is the willingness to respond on the survey-level, not the willingness to respond on the item-level. To illustrate that re-assignment must have occurred, consider Table 3.5. The variable sets in the columns of this table are defined in such a way that set 1,2,3 and 4 contain an increasing number of variables (see Table 3.4). Hence, the number of respondents in each row cannot increase from one column to the next, except when re-assignment occurred. Inspection of Table 3.5 shows that it did. For instance, the response frequencies for the sub-population IP2R3 even increase monotonically. The only way to make sure that each individual retains its original sub-population membership is to discard observations that have item-nonresponse *within the*

Variable set	phealth (3)	work (2)	age (3)	mhealth (2)	income	wealth	longill (2)	holiday (2)	clubs (2)
set1	1	1	1						
set2	1	1	1	1					
set3	1	1	1	1	1				
set4	1	1	1	1	1	1	1	1	1

Table 3.4: Several sets of variables. Each row indicates which variables are included in the set: included variables take the value 1 in that row. The variable names of categorical variables are followed by the number of their categories within parentheses.

	set1		set2		set3		set4	
Sub-population	Freq	%	Freq	%	Freq	%	Freq	%
BP	2747	25%	2730	24%	2654	24%	552	13%
IP4	612	6%	604	5%	590	5%	274	3%
IP3	670	6%	670	6%	668	6%	395	3%
IP3R4	137	1%	138	1%	158	1%	169	1%
IP2	1016	9%	1007	9%	1005	9%	915	8%
IP2R3	64	1%	67	1%	70	1%	140	1%
IP2R4	67	1%	66	1%	81	1%	112	1%
IP2R34	114	1%	120	1%	136	1%	121	1%
Total Panel	5427	49%	5402	49%	5362	49%	2678	24%

Table 3.5: Simultaneous response pattern by sets of variables. In the first wave, a total of 11029 full personal interviews were successfully issued.

sub-populations defined by response on the survey-level. This ensures that the attrition problem addressed in this chapter remains clearly separated from the problem of item non-response. This procedure has no impact on the estimation results if non-response at the item level is random. The response pattern shown in Table 3.6 results from this adjustment for re-assignment.

3.2.4 Return

The estimation method that will be used in later sections of this chapter requires that the response pattern in the panel is monotone. This rules out that an individual that dropped out in a certain wave of the panel returns later. In the ELSA panel, as we have seen, the percentage of individuals that return after dropping out is very small. To obtain a monotone pattern we include observations from IP3R4 (120) in IP3 and observations from IP2R4 (63), IP2R3 (61) and IP2R34 (90) in IP2. The responses given in the wave of return are not used. Table 3.7 and Table 3.8 depict the simultaneous reponse rates and per-

Sub-population	set1 Freq	set1 %	set2 Freq	set2 %	set3 Freq	set3 %	set4 Freq	set4 %
BP	2747	51%	2730	51%	2654	51%	552	39%
IP4	612	11%	599	11%	585	11%	129	9%
IP3	666	12%	658	12%	651	13%	204	14%
IP3R4	120	2%	119	2%	116	2%	34	2%
IP2	1011	19%	994	19%	987	19%	475	33%
IP2R3	61	1%	61	1%	61	1%	11	1%
IP2R4	63	1%	63	1%	63	1%	14	1%
IP2R34	90	2%	89	2%	83	2%	12	1%
Total Panel	5370	100%	5313	100%	5200	100%	1431	100%
RS4	2229		2220		2220		1036	
Discarded	57		89		162		1247	

Table 3.6: Simultaneous response pattern by sets of variables, adjusted for re-assignment.

sistence rates after adjusting for item-nonresponse and return. All of these rates will differ depending on the set of included variables. The last column of both tables repeats the results from Table 3.1 and Table 3.2 obtained for all successfully issued full personal interviews.

The results in the tables can be summarized as follows: after adjusting for item non-response in such a way that re-assignment to other sub-panels does not occur, about half the data is preserved when the first, second or third sets of variables are used. Moreover, in that case the response fractions are similar to the response fractions at the survey level. Using the fourth set of variables preserves less data and changes the response pattern. In what follows, I will use the third set of variables.

	set1		set2		set3		set4		Table 3.1	
Sub-population	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%
BP	2747	51%	2730	51%	2654	51%	552	39%	5695	52%
IP4	612	11%	599	11%	585	11%	129	9%	1204	11%
IP3	786	15%	777	15%	767	15%	238	17%	1607	15%
IP2	1225	23%	1207	23%	1194	23%	512	36%	2523	23%
Total Panel	5370	100%	5313	100%	5200	100%	1431	100%	11029	100%
RS4	2229		2220		2220		1036		2230	

Table 3.7: Simultaneous response pattern by sets of variables, adjusted for re-assignment and return.

Persistence	set1		set2		set3		set4		Table 3.2	
	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%
wave 1	5370		5313		5200		1431		11029	
wave 1 - wave 2	4145	77%	4106	77%	4006	77%	919	64%	8506	77%
wave 2 - wave 3	3359	81%	3329	81%	3239	81%	681	74%	6899	81%
wave 3 - wave 4	2747	82%	2730	82%	2654	82%	552	81%	5695	83%

Table 3.8: Persistence rates for core members of the ELSA panel, adjusted for re-assignment and return.

3.3 Preliminary Evidence of Selection

Many panel data analyses ignore the attrition problem by using only the balanced panel. If the people that continue to respond are different from the people that leave the panel with respect to the variables of interest, this leads to selection bias. As a preliminary analysis, sample means of sub-populations with different response patterns can be compared. Table 3.9 reports differences in average income. As this variable is continuously measured, the regression results in the table are easy to interpret. For labor market status, the variable of interest, a nonlinear model would be required. Previous analysis has shown that income and wealth are key financial determinants of when people retire (Blundell et al. (2002)).

Table 3.9 shows cross-sectional regressions of income on waves1 and waves2, binary variables that indicate the duration of panel-membership. The variable wave1 equals 1 if the individual dropped out of the panel after one wave, and zero otherwise. Wave2 equals one in case of two consecutive waves of participation. Only the first three waves of the ELSA panel are used here, as the variable income is not yet available for the fourth wave. The results in the first regression column show that, compared to the balanced panel, average net weekly income is significantly lower in IP2. With 91.97, the difference is substantial. The sample average in IP3 is 73.13 lower than in the balanced panel. By setting these mean-comparisons up as OLS regressions, other regressors can be included. This enables us to investigate whether the differences in income persist when more similar individuals are compared within the sub-populations defined by the response patterns. After inclusion of the

	income		income		income		income	
	coeff	s.e.	coeff	s.e.	coeff	s.e.	coeff	s.e
cons	396.71	6.50	612.83	15.41	294.71	10.61	498.67	18.09
waves1	-91.97	18.32	-49.66	16.56	-55.52	18.56	-42.63	16.81
waves2	-73.13	17.30	-40.81	15.01	-33.10	17.47	-24.06	15.22
phealth (poor)					-56.24	24.18	-4.53	21.41
phealth (fair)					-51.70	15.70	-9.12	13.95
work					178.29	16.13	181.39	14.60
age (50-59)					120.18	17.46	40.74	16.16
age (60-69)					91.72	13.92	42.15	12.44
mhealth					-74.68	15.70	-30.60	14.04
male			4.73	10.43			-2.17	10.44
single			-184.30	11.31			-161.00	11.44
educ (gce,cse)			-117.52	15.00			-120.67	14.90
educ (no)			-247.13	13.10			-201.10	13.47
educ (other/foreign)			-207.66	20.21			-180.79	20.20
mt1child			-25.52	11.74			-3.06	11.79

Table 3.9: Regressions of income on different sets of explanatory variables and attrition indicators.

time-constant variables, the differences are still significant, as the second column of Table 3.9 shows. The same result is obtained when the time-constant variables are replaced by time-varying ones. The last column shows that, if both sets of variables are included, average income in IP3 is no longer significantly different from average income in the balanced panel. This suggests that an attrition model that corrects for selection on observables is required. The differences in average income between IP2 and the balanced panel, indicated by the significance of waves1, hints at potential selection on unobservables.

3.4 Attrition Models

In each wave, a panel study obtains responses on a vector of variables Z. For each individual, the vector of responses in wave t is denoted by Z_t. The binary indicator D_{it} takes the value 1 if responses are obtained for individual i in wave t. For simplicity, I assume that for all individuals approached in the first wave responses were obtained. In the ELSA panel this assumption is reasonable (details can be found in the documentation). The presence of attrition in a panel with two waves implies that the distribution $f(Z_2|Z_1, D_2 = 0)$ is not observed. Attrition can be modelled by restricting the conditional probability of observation in both waves. This observation probability would be unrestricted if $P(D_2 = 1|Z_1, Z_2) = G(k(Z_1, Z_2))$, where G denotes some cdf function and k denotes the index function. For example, in a logit model, G would be the cdf of the logistic distribution and k would be a linear function of the variables in Z_1 or Z_2. Any particular choice of the index function corresponds with a particular unobserved distribution $f(Z_2|Z_1, D_2 = 0)$. Attrition models can hence be described by the restrictions they place on the index function k. What follows is a discussion of the four most commonly used attrition models for panels with two waves.

Ignoring the attrition problem amounts to assuming that the index function is a constant, $P(D_2 = 1|Z_1, Z_2) = G(k_0)$. In other words, the probability of observation does not vary with Z_1 or Z_2. The population distribution *solution* implied by the assumption of MCAR attrition, $f_{MCAR}(Z_1, Z_2)$, then equals the balanced panel distribution $f(Z_1, Z_2|D_2 = 1)$. Any complete-cases analysis of panel data implicitly assumes MCAR.

A less restrictive alternative is Missing At Random (MAR). This attrition model assumes that $P(D_2 = 1|Z_1, Z_2) = G(k_0 + k_1(Z_1))$. The observation probability is allowed to vary with Z_1 in arbitrary ways, via the unrestricted function k_1, but cannot depend on Z_2. As Z_1 is observed for all individuals in the panel, MAR is called ignorable. In the econometrics literature it is sometimes referred to as *selection on observables* (Fitzgerald and Moffitt (1998)).

Non-ignorable models allow for selection on unobservables in that they have the observation probability depend on the partially observed Z_2. An example of this is provided by the attrition model suggested by Hausman and Wise (1979): $P(D_2 = 1|Z_1, Z_2) = G(k_0 + k_2(Z_2))$. Note that HW admits selection on unobservables but at the same time rules out selection on observables. Both the MAR and HW attrition models are nonparametrically just-identified: they are identified and have no testable implications and are hence observationally equivalent.

If a refreshment sample is available in the second wave, Hirano et al. (2001) show that the Additively Non-ignorable (AN) attrition model identifies the population distribution, with observation probability $P(D_2 = 1|Z_1, Z_2) = G(k_0 + k_1(Z_1) + k_2(Z_2))$. The AN model admits Non-ignorable attrition and does not rule out selection on observables. This model is generalized to panels with any number of waves in chapter 2 of this thesis. The resulting Sequential Additively Non-ignorable (SAN) attrition model imposes restrictions on the attrition hazards to achieve just-identification:

$$P(D_t = 1|D_{t-1} = 1, \ldots, D_1 = 1) = G(k_0 + k(Z_1, \ldots, Z_{t-1}) + k(Z_t)).$$

where the functions G denotes a cdf, e.g a logit. In a four wave panel where Z_t consists of a single binary indicator, these attrition hazards reduce to

$$P(D_2 = 1|Z^2) = G(\alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2) \tag{3.1}$$

$$P(D_3 = 1|D_2 = 1, Z^3) = G(\beta_0 + \beta_1 Z_1 + +\beta_2 Z_2 + \beta_{12} Z_1 Z_2 + \beta_3 Z_3) \tag{3.2}$$

$$P(D_4 = 1|D_3 = 1, Z^4) = G(\gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_2 + \gamma_3 Z_3 \tag{3.3}$$

$$+ \gamma_{12} Z_1 Z_2 + \gamma_{13} Z_1 Z_3 + \gamma_{23} Z_2 Z_3 + \gamma_4 Z_4). \tag{3.4}$$

3.5 Estimation of the SAN Attrition Model

For the ELSA panel, currently only a fourth wave refreshment sample is available.[3] Consequently, there is no good reason to include a non-ignorable

[3] As mentioned earlier, no *complete* refreshment samples were obtained in previous waves. In the third wave ELSA did collect a refreshment sample, but this sample included 50-53 year

attrition model for attrition in the second wave. Indeed, any HW attrition model has an observationally equivalent solution derived from MAR (Hirano et al. (2001), Bhattacharya (2008)). For the same reason, attrition in the third wave can take the form of MAR.[4] For the analysis of attrition in the ELSA panel this is unfortunate, given the results discussed in section 3.3. As in the fourth wave a refreshment sample was collected, an AN model for attrition in the fourth wave identifies the parameters of interest.

Estimation of the SAN model can proceed by the weighted GMM estimator proposed in chapter 2. This estimator chooses weights that are such that the weighted balanced panel is consistent with all the information in the panel and the refreshment samples. Estimation of the weights is simplified by matching on a set of moments instead of on the full marginal distributions obtained from the refreshment samples. The SAN solution can be approximated by matching on a large set of such moments.

3.5.1 Moment Conditions

The parameters of interest are the probabilities of transition into inactivity in the second, third and fourth waves, respectively. Let Z_t denote the value taken by the binary variable work in wave t. This variable takes the value one for individuals that are active in the labor market and zero otherwise. The

olds only. This was done to "refresh" the cohort of youngest core members in the ELSA panel. In the first wave, the youngest eligible core member was approximately aged 50. In the third wave these individuals were hence aged 54. In wave 4 refreshment sample members of all cohorts were sampled and as such auxiliary information was gathered about all age cohorts contained in the original panel. It is for this reason that we only consider *this* sample as a refreshment sample.

[4]Section 7 proposes the GHW model for attrition, which allows for non-ignorability of attrition in the third wave under some additional restrictions.

transition probabilities can be characterized as the solutions of the following moment equations:

$$
\begin{aligned}
E[I(Z_2 &= 0|Z_1 = 1) - \theta_1] = 0 \\
E[I(Z_3 &= 0|Z_2 = 1) - \theta_2] = 0 \\
E[I(Z_4 &= 0|Z_3 = 1) - \theta_3] = 0,
\end{aligned}
$$

where $I(A)$ equals 1 if A holds and is zero otherwise. In the absence of attrition, method of moments estimates of the transition probabilities θ_1, θ_2 and θ_3 can be obtained by calculating the relevant sample fractions. To correct for the attrition in the ELSA panel the weighted balance panel must be used. Weights that allow for SAN attrition use (3.4), adding 16 parameters to estimate. Imposing MAR in the second and third wave corresponds the restrictions $\alpha_2 = \beta_3 = 0$. The binary nature of Z implies that only first moments need to be matched.

3.5.2 Estimation Results

The estimation results are reported in Table 3.10. For brevity, the attrition parameter estimates are categorized in three categories: selection on observables in wave t indicates that one or more coefficients of Z in earlier waves are significantly different from zero. This implies that the attrition problem in that wave cannot be ignored, and MCAR estimates are potentially misleading. Selection on unobservables in wave t implies that the coefficient of Z_t was significant. A selection on observables (MAR) model is not sufficient in that case, and AN estimates are more credible.

The transition probability estimates show that the MCAR underestimates the transition probabilities in the second and third wave. This suggests that individuals that exit the labor market are under-represented in the balanced panel. In other words, individuals that exit the labor market are more likely to attrit. A possible explanation for this result is that individuals that retired relatively recently are more likely to have moved. It could also be that both attrition and retirement are related to health.

The results can also be linked to the analysis in section (3.3). There it was shown that the average income in the balanced panel is relatively high, when compared to IP2 and IP3. Individuals that earn more tend to prefer a phasing out approach to retirement (Banks et al. (2010)). Phasing out means that transition to part-time work precedes transition into retirement. As Z_t equals one for both part-time and full-time workers, working less hours does not correspond to a transition in Z. Less transitions will hence be observed in the balanced panel, where average income is relatively high. A conditional analysis could be more revealing. Such an analysis would, however, involve many more parameters and is beyond the scope of this chapter.

Not surprisingly, the attrition results show evidence of selection on observables in all waves where attrition occurs. The AN model that uses the information contained in the fourth wave refreshment sample does not indicate non-ignorability of the attrition.

Transition	MCAR coeff	MCAR s.e.	MAR-MAR-MAR coeff	MAR-MAR-MAR s.e.	MAR-MAR-AN coeff	MAR-MAR-AN s.e.
wave 1- wave 2 (θ_1)	0.178	0.026	0.213	0.031	0.212	0.028
wave 2 - wave 3 (θ_2)	0.163	0.029	0.221	0.034	0.200	0.031
wave 3 - wave 4 (θ_3)	0.237	0.032	0.218	0.042	0.224	0.040

Attrition wave number	type of selection	MCAR	MAR-MAR-MAR	MAR-MAR-AN
wave 2	observables	-	yes	yes
	unobservables	-	-	-
wave 3	observables	-	yes	yes
	unobservables	-	-	-
wave 4	observables	-	yes	yes
	unobservables	-	-	no

Table 3.10: Estimates of transition probabilities into inactivity under several attrition models.

3.6 Non-ignorable Attrition in Waves Without Refreshment Samples

If attrition is counteracted by collecting refreshment samples, the SAN attrition model identifies the population distribution and selection on unobservables can be admitted. A potential weakness of the analysis in section 3.5 is that in the ELSA panel no such refreshment samples have been collected in waves two and three. By exploiting the moment conditions in a slightly different way, selection on unobservables can however be allowed for. In this section I propose the Generalized Hausman and Wise (GHW) attrition model that can be used for that purpose. It will be shown how it can accommodate potential selection on unobservables due to attrition in the third wave of the panel.

3.6.1 The Generalized Hausman and Wise Attrition Model

The SAN attrition model maintains two sets of restrictions on the third wave attrition hazard, sequential attrition and additive non-ignorability. Below, I clarify the nature of these assumptions in our binary application. The population distribution $f(Z_1, Z_2, Z_3)$ has 8 parameters in this case. Attrition can be taken into account by adding a model for $P(D_1 = 1, D_2 = 1, D_3 = 1|Z_1, Z_2, Z_3)$. Sequential attrition then implies that this probability equals

$$P(D_1 = 1, D_2 = 1, D_3 = 1|Z_1, Z_2, Z_3) = \Pi_{t=2}^{3} P(D_t = 1|D_{t-1} = 1, \ldots, D_1, Z_1, \ldots, Z_t).$$

The two attrition hazards can be parameterized as

$$P(D_2 = 1|D_1 = 1, Z_1, Z_2, Z_3) = G(\gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_2 + \gamma_{12} Z_1 Z_2) \quad (3.5)$$

$$P(D_3 = 1|D_2 = 1, Z_1, Z_2, Z_3) = G(\delta_0 + \delta_1 Z_1 + \delta_2 Z_2 + \delta_3 Z_3 + \delta_{12} Z_1 Z_2 \quad (3.6)$$

$$+\delta_{13} Z_1 Z_3 + \delta_{23} Z_2 Z_3 + \delta_{123} Z_1 Z_2 Z_3), \quad (3.7)$$

resulting in an additional 12 parameters. Models that have no testable implications can be obtained by requiring that the pair $(f(Z_1, Z_2, Z_3), P(D_1 = 1, D_2 = 1, D_3 = 1|Z_1, Z_2, Z_3))$ be consistent with the observable distributions $f(Z_1, Z_2, Z_3|D_1 = 1, D_2 = 1, D_3 = 1)$ from the balanced panel, $f(Z_1, Z_2|D_2 = 1, D_3 = 0)$ from IP3, $f(Z_1|D_2 = 0)$ from IP2, as well as the response fractions $P(D_2 = 1)$ and $P(D_3 = 1|D_2 = 1)$. These provide 14 restrictions, implying that another 6 restrictions are needed to identify the model. The SAN model imposes

$$\gamma_{12} = \delta_{13} = \delta_{23} = \delta_{123} = 0.$$

The remaining two restrictions are obtained by requiring consistency with $P(Z_2 = 1)$ and $P(Z_3 = 1)$, obtainable from second and third wave refreshment samples. As no refreshment samples are available for the ELSA panel in the second and third wave, the latter two restrictions cannot be imposed here.

Identification can be achieved by fixing γ_2 and δ_3 at a particular value, as proposed by Rotnitzky et al. (1998). Alternatively, we may impose restrictions on the joint population distribution to reduce the number of parameters. There is a considerable literature (Baker and Laird (1988), Baker (1995), Chambers and Welsh (1993), Conaway (1993), Fitzmaurice et al. (1995)) that imposes a mixture of distributional assumptions on the population distribution and functional form assumptions on the attrition probabilities. However, leaving the population distribution unrestricted has the advantage that the identification becomes independent of the population model used. I therefore impose the restrictions $\gamma_2 = \gamma_{12} = 0$, corresponding to MAR attrition in the second wave. The remaining 4 restrictions can be obtained by requiring that the third wave attrition does not depend on Z_1, implying that $\delta_1 = \delta_{12} = \delta_{13} = \delta_{123} = 0$. For the general case, this corresponds to a first order Markov assumption on the third period attrition hazard:

$$P(D_3 = 1 | D_2 = 1, D_1 = 1, Z_1, Z_2, Z_3) = G(k_0 + k_1(Z_2, Z_3)).$$

I will refer to this model as the Generalized Hausman and Wise (GHW) model. Its advantage is that it allows for non-ignorable attrition in third wave without requiring a refreshment sample. The disadvantage is that it does so by ruling out some, but not all, forms of selection on observables. This contrasts with the Hausman and Wise model in the second wave, that rules out *all* forms of selection on observables to achieve non-ignorability. It must be stressed, though, that, even in the third wave, an observationally equivalent solution can be derived from MAR.

3.6.2 Estimation Results for the GHW Model

The estimates using the GHW model are shown in Table 3.11. To facilitate comparison, the MAR estimates from Table 3.10 are repeated in the first column of this table.

		MAR-MAR-MAR		MAR-GHW-GHW		MAR-GHW-AN	
Transition		coeff	s.e.	coeff	s.e.	coeff	s.e.
wave 1- wave 2 (θ_1)		0.213	0.031	0.198	0.034	0.201	0.029
wave 2 - wave 3 (θ_2)		0.221	0.034	0.189	0.032	0.196	0.032
wave 3 - wave 4 (θ_3)		0.218	0.042	0.221	0.041	0.225	0.044
Attrition							
wave number	type of selection						
wave 2	observables	yes		yes		-	
	unobservables	-		-		-	
wave 3	observables	yes		yes		yes	
	unobservables	-		no		no	
wave 4	observables	yes		yes		yes	
	unobservables	-		no		no	

Table 3.11: Estimates of transition probabilities into inactivity using the MARGHW attrition model.

The estimates of the transition probabilities are similar to those in Table 3.10. The MCAR attrition model is rejected in favor of selection on observables. No evidence is found for selection on unobservables. The results in Table 3.10 and Table 3.11 can be summarized as follows: although the attrition in the ELSA panel cannot be ignored, it can be dealt with by maintaining selection on observables.

3.7 Summary and Conclusion

In this chapter, estimates were obtained for the probability of transition into inactivity for elderly people using the English Longitudinal Study of Ageing (ELSA). The aim of the study was to investigate if the estimates are affected by the potentially nonrandom attrition in the ELSA panel. By exploiting the information available in the fourth wave refreshment sample, attrition in the fourth wave could be permitted to be non-ignorable through the use of the SAN attrition model. Attrition in earlier waves was restricted to MAR.

The estimates show that ignoring the attrition by using only the balanced panel, leads to underestimation of the transition probabilities. This suggests that individuals that exit the labor market are more likely to attrit. There is insufficient evidence of non-ignorable attrition in the fourth wave. Non-ignorability was ruled out by the attrition model in the second and third wave.

The Generalized Hausman and Wise attrition model was proposed as an alternative to MAR in the third wave. The advantage of this model is that it allows for non-ignorable attrition in the absence of refreshment samples. The disadvantage is that it does so by ruling out some, but not all, forms of selection on observables. The estimates that were obtained were not suggestive of non-ignorable attrition in the third wave.

Appendix

To ensure that the study is reproducible, Table 3.12 lists the variables that were used. Table 3.13 shows the derived variables.

Variable name	Description	Coding	Time-constant
Health			
hehelf	self-reported health	poor, fair, good, very good, excellent	no
heill	longstanding illnesses?	yes, no	no
helim	Do these limit your activities?	yes, no	no
psceda	depressed last week?	yes, no	no
Work			
wpdes	work	retired, empl., self-empl, unempl, disabl, look after home, semi-retired	no
Income			
totinc_bu	total net income	weekly	no
nettotw_bu	total net (non-pension) wealth	pounds	no
Other			
indager	age	years	no
dhsex	respondent's sex	male, female	yes
dimar	marital status	never married, married, re-married, legally separated, divorced, widowed	yes
scptpa5	holiday abroad in past year	yes, no	no
scorg9	Not a member of clubs	yes, no	no
edqual	Highest educ. qual.	degree, higher educ, gce, cse, other, no semi-sk man, unsk man, armed forces	yes
dhnch	# living non-res children	0-13	yes

Table 3.12: Variable descriptions of the variables used in the analysis.

Variable name	Description	Derived from	Coding	Time-Constant
phealth	Self-reported physical health	hehelf, hegenh	poor, fair, good	no
longill	Longstanding limiting illness	heill, helim	yes, no	no
mhealth	depressed during last week?	psceda	yes, no	no
work	Active in the labor market	wpdes	yes, no	no
income	total net weekly income	totinc_bu	in pounds	no
wealth	total net (non-pension) wealth	nettotw_bu	in pounds	no
age	Age	indager	in years	no
agec	Age category	indager	50-59, 60-69, 70+	no
male	Respondent is male	dhsex	yes. no	yes
single	Not married or not re-married	dimar	yes, no	yes
holiday	Holiday abroad last year	scptpa5	yes, no	no
clubs	Member of any clubs?	scorg09	yes, no	no
educ	Highest attained qualification	edqual	higher/gce,cse/other/no	yes
mt1child	>1 living (non-res) child	dhnch	yes, no	yes

Table 3.13: Variable descriptions of derived variables.

Chapter 4

EM estimation of Panel data Models with Nonignorable Attrition and Refreshment Samples

4.1 Introduction

Panel-data are obtained by repeatedly observing units (e.g. firms, countries or households). Compared to that of a single cross-section or a single time series, the use of panel data permits the identification of more elaborate models. At the same time, missing data problems often become more severe, especially unit non-response. Indeed, in many panel surveys the fraction of units that leave the panel at a certain time period is non-negligible. This type of nonresponse is called attrition. When not taken into account, attrition can be particularly harmful if it is related to the variables of interest. In that case the attrition is called selective. If the attrition in a particular time-period is related to the value contemporaneously taken by some variable of interest it is called nonignorable.

Until recently, tests for the selectivity of attrition as well as estimation methods meant to correct for the resulting bias were based on stochastic censoring models of response behaviour. A joint (conditional) model for the outcome variable of interest and the presence of the unit in the current wave of the panel is specified. Often, the outcome variable is assumed to relate to the conditioning variables by means of some standard panel data regression model, e.g. a random effect model. Possible non-ignorability of the attrition is admitted by allowing for correlation between the unobserved components of this joint model, see e.g. Hausman and Wise (1979), Ridder (1990), Ridder (1992) and Verbeek and Nijman (1992). Identification of these models relies on either the availability of an instrument or on functional form restrictions. Moreover, different panel data regression models give rise to different correlation structures.

Hirano et al. (2001) have proposed a way out of this deadlock by using of refreshment samples. These are random samples from the population of interest. This additional data source allows for the nonparametric just-identification of the population distribution of the outcome variables in a panel of two waves under weak assumptions on the response probabilities. Their results are generalized to panels with any number of waves in chapter 2 of this thesis. The chapter also proposes a weighted GMM estimator to estimate parameters that solve a set of moment equations in the population. The SAN attrition model implies a population distribution solution that is consistent with all the information in the panel and the refreshment samples. Any particular model, e.g. a linear regression model relating Z to X, imposes a restriction on this distribution and therefore inherits this correction.

In this chapter, I propose an Expectation-Maximization (EM) algorithm for maximum likelihood to estimate the parameters of a generic parametric panel data model under SAN attrition. The algorithm is non-standard because the SAN model allows for non-ignorable attrition. Almost all EM algorithms proposed in the missing data literature require Missing At Random. In addition, the values taken by regressor variables are usually assumed to be constant over time. I do not make this assumption in this chapter.

The parametric nature of maximum likelihood implies that more restrictions are made than strictly necessary for nonparametric identification. A GMM approach would be less restrictive. Moreover, when the likelihood is specified, estimation could proceed by using the expected score in the GMM approach. However, the EM algorithm proposed in this chapter has the advantage that it can be used in cases where the score function is hard to obtain.

The outline of the chapter is as follows. Section 4.2 considers a two-period panel and discusses the Additively Non-ignorable attrition model of Hirano et al. (2001) and its specializations. Section 4.3 discusses the SAN generalization. Section 4.4 describes the advantages of EM over a direct likelihood approach. The EM algorithm for discrete time-varying regressors is presented in section 4.5. The discreteness assumption is then relaxed in section 4.6.

Section 4.7 provides an interpretation of the algorithm as a weighted MAR procedure. The final section comments and concludes.

4.2 Identification of Population Models with Attrition and Refreshment Samples

This section outlines the attrition problem, the sampling process under consideration and the identifying assumption imposed on the attrition process. In the sequel, Z is the variable (possibly a vector) of interest, i.e. the dependent or endogenous variable[1], and X is a (vector of) independent or exogenous variable(s). We consider a panel of two waves that we label as 1 and 2. The joint distribution of Z_1, Z_2 given X in the population has density

$$f(Z_1, Z_2|X) \tag{4.1}$$

A model, e.g. a linear regression model that relates Z_1 and Z_2 to X (or Z_1 to X and Z_2 to Z_1 and X if one considers a model with lagged dependent variables), is a restriction on this joint density that may involve a vector of parameters θ.

We assume that we always observe X and Z_1, but that as a consequence of attrition we fail to observe Z_2 for some fraction of the population. Let D be

[1]More precisely, a vector Z may also contain time-varying covariates that are not completely observed, as will be discussed in section 4. The characterizing feature of variables that are contained in Z is that they are only partially observed for some subjects due to attrition. The set of variables that are observed in every time-period for every subject are contained in X. For ease of exposition, for the moment it is assumed that only endogenous variables have missing values.

the indicator of the observation of Z_2, with $D = 1$ if Z_2 is observed and with $D = 0$ if it is not observed. Then our observation process allows us to recover

$$f(Z_1|X) \tag{4.2}$$

$$f(Z_1, Z_2|X, D = 1) \tag{4.3}$$

To isolate the attrition problem, note that the observation process is uninformative with respect to $f(Z_2|Z_1, X, D = 0)$. Therefore, without additional information the population distribution is not identified[2]. To model attrition we specify a model for the observation probability

$$Pr(D = 1|Z_1, Z_2, X) = G(k(Z_1, Z_2, X)) \tag{4.4}$$

with G being a (prespecified) c.d.f. of some continuous distribution, e.g. the logistic or standard normal, and k being some real-valued index-function.

In many panel studies attrition does not come as a surprise. Indeed, the designers anticipate attrition and, because they want to keep the number of units approximately constant, they 'refresh' the panel by 'replacing' the units that are lost by new (randomly selected) units. In our notation this amounts to observing the distribution of Z_2 (given X)

$$f(Z_2|X) \tag{4.5}$$

[2]This does not mean that the sampling process is completely uninformative with respect to (4.1). One may be able to put bounds on the set of distributions that are consistent with the available information. See Manski (1995) for details.

The key result in Hirano et al. (2001) is that if we restrict k to be additive in Z_1 and Z_2,

$$k(Z_1, Z_2, X) = k_0(X) + k_1(Z_1, X) + k_2(Z_2, X) \tag{4.6}$$

then we can uniquely identify the population distribution in (4.1) (and hence any model that is a restriction of this distribution). Hirano et al. (2001) refer to (4.4) with restriction (4.6) as the Additively Non-ignorable (AN) model. The motivation for identification by additive non-ignorability is that AN estimates can be interpreted as minimizing the discrepancy with the balanced panel estimates among the set of estimates that satisfy the restrictions imposed by the refreshment sample and the distribution $f(Z_1)$, observable from the panel.

Hirano et al. (2001) show that this is a case of just identification; the AN restriction does not impose any testable restrictions on distributions of observables.

Before we relate the AN restriction to other restrictions proposed in the literature, it is convenient to first consider a more restrictive version of the attrition process. Suppose, therefore, that that the functions k_0, k_1 and k_2 are known to be linear in X, Z_1 and Z_2, respectively. Moreover, assume that G is the standard normal distribution function Φ. This reduces (4.4) to

$$Pr(D = 1|Z_1, Z_2, X) = \Phi(\alpha_1' Z_1 + \alpha_2' Z_2 + \alpha_X' X)) \tag{4.7}$$

with the α's denoting parameter vectors. In what follows, this model is referred to as the linear AN assumption. Now we relate this assumption to the ones most commonly used in practice. To facilitate this discussion, first note that

$$Pr(Z_1, Z_2|X) = Pr(Z_1, Z_2|X, D = 1)\frac{Pr(D = 1|X)}{Pr(D = 1|Z_1, Z_2, X)}. \qquad (4.8)$$

In the terminology of Little and Rubin (1987), the most stringent assumption is Missing Completely At Random (MCAR). This assumes that the observed distribution equals the population distribution. It is apparent from (4.8) that this can be denoted as $D \perp Z_1, Z_2|X$ or $k_1(Z_1, X) = k_2(Z_2, X) = 0$ in (4.6), which implies $\alpha_1 = \alpha_2 = 0$ in (4.7). Essentially, this states that, although there are data missing, there is no missing data problem. Inference based on completely observed units does not lead to biased results in this case. Indeed, any empirical study that disregards incomplete observations implicitly assumes MCAR.

The second, and most popular, assumption is called Missing At Random (MAR). Although tolerating a non-random sample from the population distribution, this requires random draws from the conditional distribution of the (not completely observed) Z_2 given the (completely observed) Z_1 and X. This identifies the joint population distribution by assuming $D \perp Z_2|X$ or $k_2(Z_2, X) = 0$ in (4.6), which implies $\alpha_2 = 0$ in (4.7). This assumption is attractive for two reasons. First, it identifies the population distribution non-parametrically from the unbalanced panel without the need of a refreshment sample. Second, under certain conditions the attrition process is ignorable when inference is likelihood-based. Specifications of response models that obey the MAR assumption end up cancelling from the likelihood. Therefore, no explicit specification of the response model is necessary, other then restricting it to belong to the MAR-class of models. Hence the term ignorability.

In observational studies it is often not appropriate to assume ignorability of the attrition process. Hausman and Wise (1979) developed an attrition model that corresponds with $\alpha_1 = 0$ in (4.7). An obvious generalization is $D \perp Z_1 | X$ or $k_1(Z_1, X) = 0$ in (4.6). The attrition is therefore allowed to depend on values taken by the variables of interest that are not always observed. As a consequence, even the observed *conditional* distribution is a distorted version of its population counterpart. This type of attrition is called non-ignorable.

All assumptions discussed above are nested within the AN assumption. The refreshment sample provides just enough identifying power to permit the attrition process to be non-ignorable (unlike MAR) while allowing MAR as a nested sub-model (unlike HW). The next section discusses the generalization of the AN model to panels with more than two waves.

4.3 Identification of Population Models With More Than Two Periods

Longitudinal studies often continue for more than two periods. In this section the Sequential Additively Non-ignorable class of attrition models is discussed. This model extends the AN model to multi-wave panels. First, we have to be precise about the sampling scheme. Consider a panel with N_p subjects, indexed by i, that are approached for T periods. Subject i responds for T_i periods, and, when $T_i < T$, will not be approached in the remaining periods. The set of observations that belong to subjects that responded in all T time periods is referred to as the Balanced Panel (BP). The set of observations with unit non-response from a certain period onwards is referred to as the Incomplete Panel (IP). More specifically, IP_t is the subset of IP with attrition occurring in period t, with $2 \leq t \leq T$. The panel, consisting of the balanced panel and the incomplete panels, has N_p observations. In obvious notation, we have $N_p = N_{CP} + N_{IP}$ with $N_{IP} = \Sigma_{t=2}^{T} N_{IP_t}$. The absence of return implies a missing data pattern in the panel that is monotone (Little and Rubin (1987)). Monotonicity disappears after inclusion of the refreshment samples. In multi-wave panels attrition occurs several times. A random refreshment sample is drawn in each wave, starting from the second wave. Each refreshment sample yields N_{RS_t} additional observations. In total, the refreshment samples contain $N_{RS} = \Sigma_{t=2}^{T} N_{RS_t}$ additional subjects.

Our objective is to combine the balanced panel, the incomplete panel and the refreshment samples with a structure imposed on the attrition process so as to identify the joint population distribution of the variables of interest, $f(Z_1, \ldots, Z_T|X)$. The attrition process is described by the distribution of the T dimensional vector of binary variables $(D_{i1}, \ldots, D_{iT})$ conditional on the values taken by the variables of interest $Z_1, \ldots, Z_T$ and a set of covariates X. A key assumption for identification is Sequential Attrition (SA). This assumption implies

$$\Pr\{D_1 = 1, \ldots, D_T = 1|Z_1, \ldots, Z_T, X\} = \Pi_{t=2}^{T} \Pr\{D_t = 1|\mathbf{D}_{t-1} = 1, \mathbf{Z}_t, X\} \quad (4.9)$$

omitting the subject index i. Boldfaced symbols like $\mathbf{Z}_t$, denote the history of the variable in question, i.e. the t-vector $(Z_t, \ldots, Z_1)$. Decomposition (4.9) states that the attrition hazard in period t does not depend on the values taken by the variables of interest at any period later than t.

It suffices to illustrate the identification result in chapter 2 for panels with three waves. Given the identification of $f(Z_1, Z_2|X)$, from the AN model of Hirano et al. (2001), identification of $f(Z_1, Z_2, Z_3|X)$ requires identification of $f(Z_3|Z_2, Z_1, X)$. Since there have now been two realizations of dropout, the observation process is uninformative about *two* distributions, namely $f(Z_3|Z_1, Z_2, D_2 = 1, D_3 = 0, X)$ and $f(Z_3|Z_1, Z_2, D_2 = 0, D_3 = 0, X)$. Sequential attrition implies

$$f(Z_t|\mathbf{Z}_{t-1}, \mathbf{D}_{t-1} = \mathbf{1}, X) = f(Z_t|\mathbf{Z}_{t-1}, X). \tag{4.10}$$

The identifying power of the sequential attrition assumption resides in the implication that the unobserved distribution $f(Z_3|Z_1, Z_2, W_2 = 0, W_3 = 0, X)$ is not informative for the identification of $f(Z_3|Z_1, Z_2, X)$. This suggests that the identification of $f(Z_1, \ldots, Z_T|X)$ can be pursued by sequentially identifying $f(Z_t|\mathbf{Z}_{t-1}, X)$ for $2 \leq t \leq T$. For each value of t there is one observed distribution $f(Z_t|\mathbf{Z}_{t-1}, \mathbf{D}_t = \mathbf{1}, X)$, effectively one unobserved distribution $f(Z_t|\mathbf{Z}_{t-1}, \mathbf{D}_{t-1} = \mathbf{1}, D_t = 0, X)$ and a refreshment sample identifying $f(Z_t|X)$. This resembles the situation studied in Hirano et al. (2001), which suggests that the SAN attrition model is identified. This is formally shown in chapter 2. The definition of the SAN model is:

Definition 6. The sequence of attrition hazards $p_t = \Pr\{D_t = 1|\mathbf{D}_{t-1} = \mathbf{1}, \mathbf{Z}_t, X\}$ for $2 \leq t \leq T$ obey the Sequential Additively Non-ignorable attrition model if

$$p_t = G(k_0(X) + k_1(\mathbf{Z}_{t-1}, X) + k_2(Z_t, X)),$$

where $k_0(X)$, $k_1(Z_1, Z_2)$ and $k_2(Z_3)$ are arbitrary squared Lebesgue integrable functions. They are normalized to equal zero in some point in the support of

(Z_t, X) to allow for the inclusion of the constant k_0.

4.4 EM-algorithm For General Panel Data Models

This section describes maximum likelihood estimation of SAN models with an EM algorithm. I argue that the algorithm has clear advantages over a direct likelihood approach. The algorithm is described in terms of generic densities and is hence generally applicable.

4.4.1 Direct Likelihood

Maximum likelihood estimation usually involves numerical optimization over the parameter space. When a model incorporates some missing data mechanism to reflect that the data are only partially observed, the simultaneous log-likelihood is often more difficult to optimize; the missing data process may affect the global concavity that is present in the original log-likelihood. Consider the log-likelihood of an observation in the balanced panel when the panel has two waves:

$$\ln L_{BP} = \ln\{f(Z_1, Z_2|X)\} + \ln\{Pr(D = 1|Z_1, Z_2, X)\} \qquad (4.11)$$

In the incomplete panel, Z_2 is not observed and therefore needs to be integrated out. Additional notation is needed to describe the implications. The characterizing feature of a variable that is contained in Z is that it is only *partially* observed for some subjects due to attrition. The set of variables that are observed in *every* time-period for *every* subject are contained in X. This is a natural division of variables when discussing attrition in panel data in general terms. In applications, however, it is customary to distinguish between *variables of interest* and *explanatory variables* or *covariates*. The extension of panel data models with an attrition component therefore requires a notation that combines the two notations. We use y to denote the variable of interest, or endogenous variable. The vector of time-varying explanatory variables and the vector of time-constant variables are denoted by x and z, respectively. The notations are related by $Z = (y, x)$ and $X = (z)$. Note that the SAN identification results imply that the vector of time-varying variables x is, like y, allowed to be missing in an additively non-ignorable way.

We are now able to distinguish between endogenous and explanatory variables in (4.11). Empirical studies often examine how the location of the distribution of y varies with x and z. In parametric models the permitted class of such relations is indexed by some parameter β. The panel data model of interest can be denoted by $f(y_1, y_2|x_1, x_2, z; \beta)$. To reflect our interest in β, rewrite (4.11) as

$$\ln L_{BP} = \ln\{f(y_1, y_2|x_1, x_2, z; \beta)\} + \ln\{Pr(D = 1|y_1, y_2, x_1, x_2, z; \alpha)\} \qquad (4.12)$$

with α denoting the attrition parameters. To obtain the log-likelihood for an observation in the incomplete panel, y_2 and x_2 are removed in (4.12). Consider the first term of that equation. Integrating out y_2 causes no special problems, other than perhaps unfavorably affecting the shape of the original likelihood. With y_2 removed, the removal of x_2 requires the specification of the distribution $f(x_2|x_1, z; \pi)$. This implies that the parameters π will need to be estimated. The log-likelihood requires a third term $\ln\{f(x_2|x_1, z; \pi)\}$ in (4.12). We would rather avoid making assumptions on this distribution since there is no interest in its value, but the above discussion shows that its specification is required. In the second term, y_2 can be averaged out using $f(y_2|y_1, x_1, x_2, z, \beta)$. Removal of x_2 requires the specified $f(x_2|x_1, z; \pi)$, together with the assumption that y does not Granger-cause x. The latter notion can be defined as

Assumption. *Consider* $f(\mathbf{y}_T|\mathbf{x}_T, z) = \Pi_{t=2} f(y_t|\mathbf{y}_{t-1}, \mathbf{x}_T, z)$. *The variable* y *does not Granger-cause* x *if*

$$f(y_t|\mathbf{y}_{t-1}, \mathbf{x}_T, z) \;=\; f(y_t|\mathbf{y}_{t-1}, \mathbf{x}_t, z) \;\; \textit{for all} \;\; 2 \leq t \leq T.$$

The result of integrating out y_2 in the second summand depends on the value of β. Integrating out x_2 from the result generally depends on π for each such value of β. This induces cross-restrictions between the parameters β, α and π. It follows that log-likelihood will in general need to be maximized over the complete vector of parameters $\theta \equiv (\beta, \alpha, \pi)$.

To summarize, the direct likelihood approach involves three problems. First, it requires the specification of $f(x_2|x_1, z; \pi)$ and estimation of its parameters π. Second, desirable properties of the population model likelihood are not necessarily retained in the incomplete panel likelihood. Third, optimization over the complete vector of parameters (β, α, π) is required. The latter is particularly inconvenient if the vector of nuisance parameters π is of high dimension.

4.4.2 EM algorithm

The EM algorithm, originally proposed by Dempster et al. (1977), has been shown to be applicable in a wide range of models with intractable likelihoods (see Ruud (1991) and MacLachlan and Krishnan (1997) for a review). The algorithm is iterative and each iteration consists of two successive steps, called the E-step and the M-step, respectively. The main idea is to exchange a difficult optimization problem with a sequence of less difficult optimization problems.

Consider again a two-wave panel with attrition. The algorithm starts by choosing an initial guess $\theta^{(0)}$ of the parameter vector θ, equal to (β, α, π). Each iteration yields an update of this value. The resulting sequence of estimates $\theta^{(i)}$ converges to the maximum likelihood estimate of θ, if the model is sufficiently regular. In iteration $(i+1)$, the E-step for an observation in the incomplete panel computes the *expected* log-likelihood. The expectation is taken over the distribution of the missing data y_2^* and x_2^* given the observed data y_1, x_1 and z and the current guess of the parameter value $\theta^{(i)}$. It thus computes the function $Q(\theta|\theta^{(i)})$, which for the incomplete panel is defined as

$$Q(\theta|\theta^{(i)}) = \tag{4.13}$$

$$\int\int \ln\left\{f(y_1, y_2^*|x_1, x_2^*, z; \beta)\right\} f(y_2^*, x_2^*|y_1, x_1, z, D=0; \theta^{(i)}) dy_2^* dx_2^* + \tag{4.14}$$

$$\int\int \ln\left\{\Pr\left\{D=0|y_1, y_2^*, x_1, x_2^*, z; \alpha\right\}\right\} f(y_2^*, x_2^*|y_1, x_1, z, D=0; \theta^{(i)}) dy_2^* dx_2^* \tag{4.15}$$

Note that, again, specification of $f(x_2^*|x_1, z)$ cannot be avoided. For the balanced panel the Q function equals the population model log-likelihood. The treatment of observations in the refreshment sample is postponed until the next section. In the M-step the entire Q-function is then maximized over θ to obtain $\theta^{(i+1)}$.

There are two things to note about this algorithm. The first is that in most applications a specialized form is used. If the log-likelihood of the behavioral model of interest is linear in the sufficient statistics – this occurs for instance when the population model belongs to the exponential family – the E-step reduces to calculating the expected sufficient statistics. For models including an attrition component, said sufficient statistics are not readily available. The second thing to note is that many applications assume MAR. This implies that there is no need to condition on $D = 0$ in (4.13). In many applications, this allows for analytic derivation of $Q(\theta|\theta_i)$, particularly when MAR is used in conjunction with a population model from the exponential family. The main attraction of the AN family of attrition models is that it allows for non-ignorable attrition. Therefore, the expectation is taken over the distribution

$$f(y_2^*, x_2^*|y_1, x_1, z, D = 0; \theta^{(i)}) = \tag{4.16}$$

$$\frac{\Pr\{D=0|y_1,y_2^*,x_1,x_2^*,z;\alpha^{(i)}\}\ f(y_2^*,x_2^*|y_1,x_1,z;\theta^{(i)})}{\iint \Pr\{D=0|y_1,y_2^*,x_1,x_2^*,z;\alpha^{(i)}\}\ f(y_2^*,x_2^*|y_1,x_1,z;\theta^{(i)})\ dy_2^* dx_2^*} \tag{4.17}$$

Taking the expectation over the distribution in (4.16) may be difficult analytically due to the conditioning on the event $\{D = 0\}$. Yet it is straightforward to perform the E-step by simulation. Indeed, consider the following scheme to draw from (4.16) in iteration $(i + 1)$, given the value of $\theta^{(i)}$:

(i) Draw x_2^* from $f(x_2|x_1, z; \pi^{(i)})$

(ii) Draw y_2^* from $f(y_2|y_1, x_1, z; \beta^{(i)})$

(iii) Retain (y_2^*, x_2^*) with probability $\Pr\{D = 0|y_1, y_2^*, x_1, x_2^*, z; \alpha^{(i)}\}$

The realization (y_2^*, x_2^*) thus obtained can be used to evaluate the complete data log-likelihood for an observation (y_1, x_1, z) in the incomplete panel. Repeating this procedure a large number of times, M say, and averaging over the M resulting values of the log-likelihood, gives the expected log-likelihood for this observation.

Although EM cannot avoid specification of $f(x_2|x_1, z; \pi^{(i)})$, it resolves two of the three problems associated with direct likelihood maximization. To see this, note that in (4.13) the expectation is taken over a *known* distribution, since it is parameterized by $\theta^{(i)}$, the known result of the previous iteration. This avoids the cross-restrictions introduced in the direct likelihood approach. This implies that separate maximization over β, α, and π is valid. This is a major advantage when the parameter space is relatively large. Moreover, note that the E-step yields an average of potentially well-behaved complete data log-likelihoods. This may be helpful in preserving smoothness properties.

The next section discusses an example that illustrates how to incorporate refreshment samples.

4.5 An Example with Discrete Regressors

Consider a panel of three waves with attrition in the second and third wave. If in both periods refreshment samples are drawn, we obtain the data structure represented by the thick-lined part of figure 4.5.The figure shows subjects organized in rows and the variables Z_t organised in columns. The observations

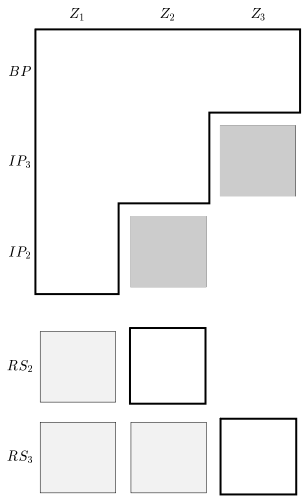

Figure 4.1: Data structure including imputations for the EM algorithm. The panel consists of the balanced panel (BP) and the two incomplete panel subpopulations (IP2 and IP3). Imputations for (y_2, x_2) are indicated in dark grey. Imputations for (x_2) light grey.

that are part of the balanced panel (BP) are situated at the top of the figure. The incomplete panels and the refreshment samples are also contained in the same *thick-lined* part of the figure. Together they describe the observations in the panel.

The shaded blocks in the figure refer to the imputations that are needed for the EM-algorithm proposed in this chapter.

To illustrate the EM algorithm, it suffices to consider the static random effects panel data model. Extension to more general models follows directly. Let y be continuous and let x denote a time-varying discrete regressor, taking values in the set $\{1, \ldots, A\}$, with x_t denoting the value taken by x at period t.[3] In a regression model, x is represented by a vector of time-varying binary variables

[3]When $T = 3$ there are A^3 possible realizations (x_1, x_2, x_3). Each such realization can be considered a cell in a three dimensional table. The extension to more than one discrete regressor simply corresponds to taking more cells.

$\tilde{d}'_{it} = (d_{1it}, \ldots, d_{Ait})$. The binary variable d_{jit} equals 1 if x takes the value j for individual i in period t. The event $\{x_3 = j_3 | x_2 = j_2, x_1 = j_1)$ can be defined similarly, and is denoted by $d_{3|2,1}$. When a time-specific effect is included, the model is described by

$$\begin{aligned} y_{i1} &= \tilde{\beta}_0 + \tilde{\beta}'_1 \tilde{d}_{i1} + \gamma_i + \delta_1 + \varepsilon_{i1} \\ y_{i2} &= \tilde{\beta}_0 + \tilde{\beta}'_1 \tilde{d}_{i2} + \gamma_i + \delta_2 + \varepsilon_{i2} \\ y_{i3} &= \tilde{\beta}_0 + \tilde{\beta}'_1 \tilde{d}_{i3} + \gamma_i + \delta_3 + \varepsilon_{i3} \end{aligned} \tag{4.18}$$

$$D^*_{i2} = \alpha^{(2)}_0 + \alpha^{(2)}_{y1} y_{i1} + \alpha^{(2)}_{y2} y_{i2} + \alpha^{(2)\prime}_{x1} \tilde{d}_{i1} + \alpha^{(2)\prime}_{x2} \tilde{d}_{i2} + \xi^{(2)}_i \tag{4.19}$$

$$D^*_{i3} = \alpha^{(3)}_0 + \alpha^{(3)}_{y1} y_{i1} + \alpha^{(3)}_{y2} y_{i2} + \alpha^{(3)}_{y3} y_{i3} + \alpha^{(3)\prime}_{x1} \tilde{d}_{i1} + \alpha^{(3)\prime}_{x2} \tilde{d}_{i2} + \alpha^{(3)\prime}_{x3} \tilde{d}_{i3} + \xi^{(3)}_i \tag{4.20}$$

where D^* in (4.19) and (4.20) denotes the latent propensity to stay in the the panel. Specifically, D^*_2 defines $\Pr\{D_2 = 1 | \mathbf{y}_2, \mathbf{x}_2\}$ while D^*_3 specifies $\Pr\{D_3 = 1 | D_2 = 1, \mathbf{y}_3, \mathbf{x}_3\}$ with $D_t = I\{D^*_t \geq 0\}$. The unobservables in (4.18) obey the standard random effects model assumptions whereas the ξ's are iid with distribution function Φ.

Given an initial guess $\theta^{(0)}$ of the parameter-vector θ, iteration $(i+1)$ of the EM algorithm reads

(1) Given $\theta^{(i)}$:

(a) impute y^*_3, x^*_3 in IP_3 and y^*_2, x^*_2 in IP_2.

(b) impute x^*_1 in RS_2 and x^*_2, x^*_1 in RS_3.

(2) Given M imputations:

(a) maximize $Q(\beta|\theta^{(i)})$ to obtain $\beta^{(i+1)}$

(b) maximize $Q(\alpha|\theta^{(i)})$ to obtain $\alpha^{(i+1)}$

(c) maximize $Q(\pi|\theta^{(i)})$ to obtain $\pi^{(i+1)}$

The first step creates the imputations, the greyed parts in the figure. Step $(1a)$, $(2a)$ and $(2b)$ follow the procedure discussed above. The notation $Q(\beta|\theta^{(i)})$ indicates that maximization over the three sets of parameters can be separated. Consider, for instance, the expected log-likelihood of an observation in IP_3. This is the average of the M log-likelihoods, indexed by m,

$$\ln L^m_{IP_3} = \ln \phi((y^*_3, \mathbf{y}_2) - \mu(\tilde{\beta}, \gamma, \delta); \Sigma_\varepsilon) + \tag{4.21}$$

$$\ln \{\Pr \{D_2 = 1|\mathbf{y}_2, \mathbf{x}_2; \alpha^{(2)}\}\} + \ln \{\Pr \{D_3 = 1|D_2 = 1, y^m_3, \mathbf{y}_2, x^m_3, \mathbf{x}_2; \alpha^{(3)}\}\} \tag{4.22}$$

$$+ \ln \{f(x^m_3|\mathbf{x}_2; \pi_{x_3|x_2,x_1})\} + \ln \{f(x_2|x_1; \pi_{x_2|x_1})\} + \ln \{f(x_1; \pi_{x_1})\}, \tag{4.23}$$

with $\mu(\tilde{\beta}, \gamma, \delta) = \mu(\beta)$ denoting the conditional mean and $\phi(\cdot, \Sigma)$ the normal density with covariance Σ. The optimization needed to obtain $\beta^{(i+1)}$ ignores terms in which β does not occur. The contribution of an observation in IP_3 therefore is a weighted average of complete data contributions for a random effects model. Likewise, a weighed average of complete data probit contributions is obtained for α. For β and α, the balanced panel and refreshment samples do not require the E-step. Computationally, estimation of $\beta^{(i+1)}$ and $\alpha^{(i+1)}$ is roughly comparable to the complete data case.

The above discussion assumes that we are able to draw the necessary x's in step $(1b)$ and maximize $Q(\pi|\theta^{(i)})$ in step $(2c)$. The remainder of this section will pursue these objectives. Again, the argument is given for a general discrete regressor and a panel with T waves, with three period panel taken as a normative example.

As for the distribution of the regressor, we assume that the conditional distribution $f(x_t|\mathbf{x}_{t-1})$ is multinomial for each t between 2 and T, and a realization $(x_t|\mathbf{x}_{t-1})$ has probability $\pi_{x_t|\mathbf{x}_{t-1}}$ of occurring. In addition, the distribution $f(x_1)$ is multinomial with probabilities π_{x_1}. The main disadvantage of this choice is the number of parameters. As the multinomial distribution without any restrictions on the parameter space is saturated – it does not restrict the data in any way – the dimension of the parameter space is correspondingly high. Often, this parameter space (or some reparameterized version of it) is restricted to enforce a lower dimensional problem, see e.g. Agresti (1990)). However, no restrictions on the parameter space of π will be imposed here. As inference on β is our objective, it is not of primary interest to estimate π efficiently. The dimensionality of π has the virtue that it effectively avoids the specification of $f(x_t|\mathbf{x}_{t-1}, z)$. If we are able to deal with this dimensionality in step $(2b)$, this solves part of the first problem mentioned earlier.

It turns out that estimating π *is* easy. To see this, consider (4.23). The multinomial log-likelihood is linear in the binary data, as an observations like $(x_3|x_2, x_1)$ has log-likelihood contribution $d_{3|2,1} \ln\{\pi_{x_3|x_2,x_1}\}$. This simplifies the estimation of π considerably. Indeed, for an observations with x_3 missing, in the E-step this contribution becomes

$$\frac{1}{M}\sum_m \left(d^m_{3|2,1} \ln\{\pi_{x_3|x_2,x_1}\}\right) = \left(\frac{1}{M}\sum_m d^m_{3|2,1}\right) \ln\{\pi_{x_3|x_2,x_1}\} = \bar{d}_{3|2,1} \ln\{\pi_{x_3|x_2,x_1}\}. \tag{4.24}$$

This equals the contribution obtained when taking $\bar{d}_{3|2,1}$ as pseudo-data replacing the unobserved x_3 in IP_3. The sum of all such partial contribution factors $\bar{d}_{3|2,1}$ in IP_3 is denoted by $\bar{n}^{IP_3}_{d_{3|2,1}}$.

Since our model is static, the refreshment samples only require imputed regressors. It suffices to discuss RS_3. The required imputations are easy to obtain. The multinomial probabilities $\pi^{(i)}_{x_1,x_2|x_3}$ follow directly from $\pi^{(i)}$. From (x_1^m, x_2^m) we obtain $n^m_{d_1}$, $n^m_{d_{2|1}}$ and $n^m_{d_{3|2,1}}$ and averaging is over the M imputations yields $\bar{n}^{RS_3}_{d_1}$, $\bar{n}^{RS_3}_{d_{2|1}}$ and $\bar{n}^{RS_3}_{d_{3|2,1}}$. As in the incomplete panel, all of these appear linearly in the log-likelihood. The above implies that the maximum likelihood estimate $\hat{\pi}_{x_3|x_2,x_1}$ can be obtained in much the same way as in the multinomial complete data case, by calculating sample fractions. Indeed, $\hat{\pi}_{x_3|x_2,x_1}$ can be obtained by calculating the fraction of (pseudo) observations falling into the cell $x_3|x_2,x_1$. The calculation of $\hat{\pi}_{x_2|x_1}$ and $\hat{\pi}_{x_1}$ is equally simple:

$$\hat{\pi}_{x_3|x_2,x_1} = \frac{n^{BP}_{d_{3|2,1}} + \bar{n}^{IP_3}_{d_{3|2,1}} + \bar{n}^{RS_3}_{d_{3|2,1}}}{n^{BP}_{d_{\cdot|2,1}} + \bar{n}^{IP_3}_{d_{\cdot|2,1}} + \bar{n}^{RS_3}_{d_{\cdot|2,1}}} \tag{4.25}$$

$$\hat{\pi}_{x_2|x_1} = \frac{n^{BP}_{d_{2|1}} + \bar{n}^{IP_3}_{d_{2|1}} + \bar{n}^{IP_2}_{d_{2|1}} + \bar{n}^{RS_2}_{d_{2|1}} + \bar{n}^{RS_3}_{d_{2|1}}}{n^{BP}_{d_{\cdot|1}} + n^{IP_3}_{d_{\cdot|1}} + \bar{n}^{IP_2}_{d_{\cdot|1}} + \bar{n}^{RS_2}_{d_{\cdot|1}} + \bar{n}^{RS_3}_{d_{\cdot|1}}} \tag{4.26}$$

$$\hat{\pi}_{x_1} = \frac{n^{BP}_{d_1} + n^{IP_3}_{d_1} + n^{IP_2}_{d_1} + \bar{n}^{RS_2}_{d_1} + \bar{n}^{RS_3}_{d_1}}{n^{BP}_{\cdot} + n^{IP_3}_{\cdot} + n^{IP_2}_{\cdot} + \bar{n}^{RS_2}_{\cdot} + \bar{n}^{RS_3}_{\cdot}} \tag{4.27}$$

where the symbol $\cdot$ denotes summation over the argument. By doing this analysis conditional on each z, the case of time-constant variables is also covered.

In summary, the EM algorithm solves all three problems associated with direct likelihood approach if the time-varying regressors are discrete variables. No restrictions on their distribution need be imposed and estimation of the nuisance parameters π amount to calculating sample fractions. The next section will discuss the continuous case.

4.6 Continuous and Discrete Regressors

The simplicity of obtaining maximum likelihood estimates of the parameters π of the distribution of the time-varying regressors generalizes from the multinomial distribution to the exponential family of distributions. An important property of the exponential family of distributions is that the log-likelihood is linear in the sufficient statistics. This property can be exploited in much the same way as the more restrictive property that the log-likelihood of the multinomial distribution is linear in the data. The vector of pseudo sufficient statistics S_m of iteration m is obtained by taking into account all relevant observed and imputed observations of imputation m in the balanced panel, incomplete panels and the refreshment samples. The vectors $\{S_m\}$ thus obtained are then averaged over the M imputations to obtain the pseudo sufficient statistic $\bar{S}$, from which the expected log-likelihood is obtained. As in the complete data case, the maximand of this log-likelihood is obtained by solving

$$E_\pi S = \bar{S} \tag{4.28}$$

for π (see Lehmann (1983)). Explicit solutions of these equations are often available.

As an example of a continuous regressor, consider the normal distribution for (x_1, x_2, x_3). This distribution has as sufficient statistics the set of sample sums s_t and sample sum of cross products s_{ts} with s and t indexing time. The set of parameters π of this simultaneous distribution can be reparameterized to parameters of conditional distributions $\pi_{3|2,1}$, $\pi_{2|1}$ and π_1. For instance, $\pi_{3|2,1}$ contains the parameters of the normal linear regression of x_3 on x_2 and x_1, including the error variance. After each imputation the estimates of $\pi_{3|2,1}$, $\pi_{2|1}$ and π_1 imply unique values for the sufficient statistics[4]. Averaging over M imputations gives the desired expected sufficient statistics. The M-step then calculates

$$\hat{\mu_t} = \bar{s}_t/n \tag{4.29}$$

$$\hat{\sigma_{st}} = \bar{s}_{st}/n - \hat{\mu}_s\hat{\mu}_t \tag{4.30}$$

which imply the values of $\hat{\pi}_{3|2,1}$, $\hat{\pi}_{2|1}$ and $\hat{\pi}_1$.

The procedures for continuous and discrete time-varying regressors can be combined to allow for mixed types. Consider discrete time-varying regressors D having a multinomial distribution. Estimation of its parameters π_D follows the discussion in the former section. Conditional on the cell taken by D, the set of continuous time-varying regressors C is normally distributed with cell-dependent mean and cell-constant covariance. This model is called the general location model. It belongs to the exponential family of distributions. When D does not Granger cause C, the estimation of π_C follows the exposition above.

4.7 Imputations and Weights

The imputations employed in the algorithm outlined in the former two sections have a useful interpretation as a weighting procedure. This conceptual

[4]With imputations the missing data pattern becomes monotone. The sufficient statistics can therefore be calculated conveniently by means of the sweep operator (Little and Rubin (1987)).

observation can sometimes be made operational in order to clear away the need for imputations in the refreshment samples. For the exponential family this is shown below.

It is well-known that the sufficient statistics for the exponential family take the form $\sum T(x_i)$, where the sum is taken over the observations and the function T differs for different members of the exponential family. For the multinomial distribution $T(x_c) = I\{x_c\}$ for all cells c and for the bivariate normal distribution $T(x_1, x_2) = (x_1, x_2, x_1^2, x_2^2, x_1x_2)$. Consider an observation x_3 in RS_3 with imputation (x_1^m, x_2^m) and log-likelihood contribution $f(x_3|x_2, x_1; \pi_{3|2,1})$. Observations x_3 appear with relative frequency $f(x_3; \pi_3)$ in RS_3, whereas complete data contributions appear with relative frequency $f(x_3|x_2, x_1; \pi_{3|2,1})$. The appropriate relative frequencies can be obtained by re-weighting the contributions as they appear in RS_3. This mimics the analysis of Manski and Lerman (1977) in the sense that the data generating mechanism does not match the likelihood employed. In their case, the likelihood corresponded to exogenous stratification whereas the data obeyed a choice based sampling scheme. However, the weighting procedure described above is not a construction to achieve consistency, but an interpretation of the calculations in the E-step. It interprets *how* maximum likelihood achieves efficiency.

For the multinomial example, the distribution over which the expectation is taken in this E-step has a known parameter vector $\pi^{(i)}$, the current guess of π. The imputation procedure results in the partial log-likelihood contributions in (4.24). The weighting interpretation follows by the observation that

$$\bar{d}_3 = \frac{\pi^{(i)}_{x_1,x_2|x_3}}{\pi^{(i)}_{x_1,x_2}},$$

a weight that can be easily obtained from $\pi^{(i)}$ without the need for imputations. Note that, as x_3 occurs at a rate π_{x_3}, this procedure results in the correct relative frequencies. For the normal case, we need to calculate the expectation of x_1, x_2, x_1^2, x_2^2 and $x_1 x_2$ given the value x_3 and $\theta^{(i)}$. Together with similar calculations for RS_2, these values can be combined with observed data to obtain the relevant expected sufficient statistics. By their additive nature and the fact that they appear linearly in the log-likelihood, the expected sufficient statistics re-weight the total log-likelihood contribution of RS. A similar weighting mechanism is operative in the incomplete panels, but there the computation of weights is more complicated because of the attrition at work in the panel. It is conceptually interesting however, because it identifies the imputation procedure employed here as an iteratively re-weighted MAR procedure. This will be made precise below.

The imputations make sure that (y_3^m, x_3^m) appear with relative frequency $f(y_3, x_3|\mathbf{y}_2, \mathbf{x}_2, \mathbf{D}_2 = \mathbf{1}, D_3 = 0; \theta^{(i)})$. Then, BP and IP_3 together yield the appropriate relative frequencies $f(y_3, x_3|\mathbf{y}_2, \mathbf{x}_2, \mathbf{D}_2 = \mathbf{1}; \theta^{(i)})$. This is equivalent to using only the balanced panel, with the log-likelihood contribution of an observation $(y_3, x_3|\mathbf{y}_2, \mathbf{x}_2)$ premultiplied by

$$\frac{\Pr\left\{D_3 = 1|D_2 = 1, \mathbf{y}_3, \mathbf{x}_3; \theta^{(i)}\right\}}{\Pr\left\{D_3 = 1|D_2 = 1, \mathbf{y}_2, \mathbf{x}_2; \theta^{(i)}\right\}} \tag{4.31}$$

A conditional, three period equivalent of (4.8) shows that these are the weights required. Note that the calculation of the weights involves the numerical integration over $f(y_3, x_3|\mathbf{y}_2, \mathbf{x}_2; \theta^{(i)})$. As a result, imputing the panel is usually more attractive.

This weighting interpretation allows us to clarify the workings of the algorithm. We do this by decomposing the construction of the algorithm in three steps.

After a single imputation, a monotone missing data pattern results. Under MAR, such a pattern can be conveniently analyzed by factoring the likelihood in conditional likelihoods, if the parameters of the conditional parametrization are distinct. Maximum likelihood estimates can then be obtained by separate optimization of these conditional log-likelihoods. The Hessian of the log-likelihood corresponding to this parametrization is block-diagonal, which makes it easy to obtain standard errors. This illustrates the convenience of assuming MAR. Under sequential attrition, the remaining missing data *can* be interpreted as MAR[5]. By averaging over M imputations, a weighting procedure results. Finally, the embedding of this weighting procedure in an iterative EM algorithm, makes these weights available. Therefore, the EM algorithm is an *iteratively re-weighted MAR* algorithm. In the same way conventional EM algorithms can be interpreted as iteratively re-weighted MCAR algorithms.

4.8 Conclusion

In this chapter, an EM algorithm was proposed to estimate Sequential Additively Non-ignorable attrition models. This class of panel data models allows for attrition that is potentially non-ignorable in every wave. Almost all EM

[5] Notice that in our analysis the *complete* data structure, i.e. the data *after* imputation, has a monotone pattern that permits the above MAR approach.

algorithms proposed in the missing data literature require Missing At Random. In addition, the values taken by regressor variables are, more often than not, assumed to be constant over time. The algorithm proposed here does not require this.

It was shown that estimation by direct maximization of the likelihood has three disadvantages. First, it requires the specification of $f(x_2|x_1, z; \pi)$ and estimation of its parameters π. Second, desirable properties of the population model likelihood are not necessarily retained in the incomplete panel likelihood. Third, maximization over the complete vector of parameters (β, α, π) is required. The latter is particularly inconvenient if the vector of nuisance parameters π is of high dimension.

The proposed EM algorithm solves these problems. Moreover, when the time-varying variables are discrete, it was shown that the nuisance parameters π can be estimated by simply calculating sample fractions. When, conditional on the discrete regressors, the continuous regressors are distributed according to a distribution from the exponential family, its parameters can be estimated with similar ease.

Chapter 5

Summary and Conclusions

Attrition in Panel Data can lead to estimation results that suffer from selection bias. This potential selectivity can be taken into account by including an attrition model into the analysis. The restrictions imposed by these attrition models rule out certain forms of selection and permit others. The least restrictive attrition models allow for the widest range of potential selection. The aim of this thesis was to show how the set of restrictions can be reduced by exploiting the information contained in refreshment samples. Moreover, to apply these models, new estimation methods were needed.

In chapter two it was shown that, in the absence of refreshment samples, the sampled population distribution is observationally equivalent to the population distribution. This population distribution solution is obtained by the Sequential Missing At Random (SMAR) attrition model. If refreshment samples are available, this attrition model has testable implications, implying that less restrictive attrition models exist. The Sequential Additively Non-ignorable (SAN) attrition model, proposed in chapter 2, does not have testable implications. The population distribution implied by the SAN model is by construction consistent with the information contained in the refreshment samples. Moreover, this distribution is shown to be observationally equivalent to the sampled population distribution, implying that it is consistent with all the information in the panel as well. As the SAN solution is unique, it therefore nonparametrically just-identifies the population distribution. Because no restrictions are imposed on the population model, this identification result is applicable to any population model of interest.

Efficient estimators of SAN models have been obtained. Most parameters of interest can be characterized as the solution of a set of population moment equations. In the absence of attrition, these parameters can then be estimated by the Generalized Method of Moments (GMM). For panels with attrition, a weighted GMM estimator was proposed that efficiently estimates the parameters of interest. The estimator is applicable under SAN attrition and all its specializations. The asymptotic properties of this estimator were derived in chapter 2. In Chapter 4 a second way to estimate the parameters was proposed. The EM algorithm proposed there delivers maximum likelihood estimates of the parameters of interest, while avoiding many of the numerical difficulties associated with a direct likelihood approach. Chapter 2 and 3 contain empirical applications.

Bibliography

Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.

Baggerly, K. A. (1998). Empirical likelihood as goodness-of-fit measure. *Biometrika 85*, 535–547.

Baker, S. G. (1995). Marginal regression for repeated binary data with outcome subject to non-ignorable non-response. *Biometrics 51*, 1042–1052.

Baker, S. G. and N. M. Laird (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association 83*(401), 62–69.

Banks, J., R. Blundell, and Tanner (1998). Is there a retirement-savings puzzle? *The American Economic Review 88*, 769–788.

Banks, J., E. Breeze, R. Crawford, P. Demakakos, C. Oliveira, E. Gjonça, R. Green, d. Hussey, M. Kumari, C. Lessof, M. Marmot, A. McMunn, A. Muriel, J. Nazroo, S. Nunn, Z. Oldfield, A. Shankar, M. Stafford, A. Steptoe, G. Tetlow, K. Ward, N. Wood, and P. Zaninotto (2010). Financial circumstances, health and well-being of the older population in england: The 2008 english longitudinal study of ageing. Technical report, UK data Archive Study Number 5050.

Banks, J., C. Lessof, J. Nazroo, N. Rogers, M. Stafford, and A. Steptoe (2010). Financial circumstances, health and well-being of the older population in england. Technical report, Institute for Fiscal Studies.

Bhattacharya, D. (2008). Inference in panel data models under attrition caused by unobservables. *Journal of Econometrics 144*(2), 430–446.

Blundell, R., C. Meghir, and S. Smith (2002). Pension incentives and the pattern of early retirement. *The Economic Journal 112*, C153–C170.

Chambers, R. L. and A. H. Welsh (1993). Log-linear models for survey data with non-ignorable non-response. *Journal of the Royal Statistical Society B 55*(1), 157–170.

Conaway, M. R. (1993). Non-ignorable non-response models for time-ordered categorical variables. *Applied Statistics 42*(1), 105–115.

Dempster, A., N. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B 39*, 1–38.

Fitzgerald, John, F. A. P. G. and R. M. Moffitt (1998). An analysis of sample attrition in panel data: the panel study of income dynamics. *Journal of Human Resources 33*(2), 251–299.

Fitzmaurice, G. M., G. Molenberghs, and S. R. Lipsitz (1995). Regression models for longitudinal binary responses with informative drop-outs. *Journal of the Royal Statistical Society B 57*(4), 691–704.

French, E. (2005). The effects of health, wealth, and wages on labour supply and retirement behaviour. *Review of Economic Studies 72*, 395–427.

Haberman, S. (1984). Adjustment by minimum discriminant information. *The Annals of Statistics*, 971–988.

Hausman, J. A. and D. A. Wise (1979). Attrition bias in experimental and panel data: The gary income maintenance experiment. *Econometrica 47*(2), 455–473.

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.

Hutson, V. and J. Pym (1980). *Applications of Functional Analysis and Operator Theory*. New York: Academic Press.

Imbens, G., R. Spady, and P. Johnson (1998). Information theoretic approaches to inference in moment condition models. *Econometrica 66*(2), 333–357.

Imbens, G. W. (1997). One-step estimators for over-identified generalized method of moments models. *Review of Ec 64*, 359–383.

Ireland, C. and S. Kullback (1968). Contingency tables with given marginals. *Biometrika 55*, 179–188.

Lehmann, E. (1983). *Theory of Point Estimation*. New York: Wiley.

Little, R. and 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.

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

Luenberger, D. G. (1969). *Optimization by Vector Space Methods*. John Wiley & Sons.

MacLachlan, G. J. and T. Krishnan (1997). *The EM Algorithm and Extensions*. New York: Wiley.

Manski, C. and S. Lerman (1977). The estimation of choice probabilities from choice-based samples. *Econometrica 45*, 1977–1988.

Manski, C. F. (1995). *Identification Problems in the Social Sciences*. Cambridge: Harvard University Press.

Manski, C. F. (2003). *Partial Identification of Probability Distributions (Springer Series in Statistics)*. Springer.

Meurs, H. and G. Ridder (1992). Attrition and response effects in the dutch mobility panel. Technical report, University of Amsterdam.

Nevo, A. (2002). Sample selection and information-theoretic alternatives to GMM. *Journal of Econometrics 107*(1-2), 149–157.

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

Newey, W. and D. McFadden (1994). Large Sample Estimation and Hypothesis Testing. *Handbook of Econometrics 4*, 2111–2245.

Newey, W. K. (1984). A method of moments interpretation of sequential estimators. *Economics Letters 14*, 201–206.

Qin, J. and J. Lawless (1994). Empirical likelihood and general estimating equations. *ANNALS OF STATISTICS 22*, 300–300.

Read, T. and N. Cressie (1988). *Goodness-of-fit Statistics for Discrete Multivariate Data*. Springer-Verlag, New York.

Rice, N., I. Lang, W. Henley, and D. Melzer (2010). Common health predictors of early retirement: findings from the english longitudinal study of ageing. *Age and ageing 40*, 54–61.

Ridder, G. (1990). Attrition in multi-wave panel data. In R. Hartog and Theeuwes (Eds.), *Panel Data and Labor Market Studies*, pp. 45–68. North Holland.

Ridder, G. (1992). An empirical evaluation of some models for non-random attrition in panel data. *Structural Change and Economic Dynamics 3*(2), 337–355.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association 90*(429), 106–121.

Rotnitzky, A., J. M. Robins, and D. O. Scharfstein (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association 93*(444), 1321–1339.

Rubin, D. (1976). Inference and missing data. *Biometrika 63*, 581.

Ruud, P. A. (1991). Extensions of estimation methods using the EM algorithm. *Journal of Econometrics 49*, 305–341.

Sholes, S., J. Medina, H. Cheshire, K. Cox, E. Hacker, and C. Lessof (2008). Health, wealth and the lifestyle of the older population in england: the 2004 english longitudinal study of ageing. Technical report, UK data Archive Study Number 5050.

Sholes, S., J. Medina, H. Cheshire, K. Cox, E. Hacker, and C. Lessof (2009). Living in the 21^{st} century: Older people in england, the 2006 english longitudinal study of ageing. Technical report, UK data Archive Study Number 5050.

Taylor, R., L. Conway, L. Calderwood, c. Lessot, H. Cheshire, K. Cox, and S. Sholes (2007). Health, wealth and the lifestyle of the older population in england: The 2002 english longitudinal study of ageing. Technical report, UK data Archive Study Number 5050.

van der Vaart, A. W. (2000). *Asymptotic Statistics (Cambridge Series in Statistical and Probabilist Mathematics)*. Cambridge University Press.

Verbeek, M. and T. Nijman (1992). Testing for selectivity bias in panel data models. *International Economic Review 33*(3), 681–703.