Chapter 4

Combining Panel Data Sets with Attrition and Refreshment Samples

Keisuke Hirano

Guido W. Imbens

Geert Ridder

Donald B. Rubin[1]

[1]The authors thank Joshua Angrist, Gary Chamberlain, Jerry Hausman, and participants in presentations at Harvard–MIT, NYU, UCLA, Aarhus, the International Statistical Institute meetings in Instanbul, and especially students at the University of Wisconsin–Madison for comments.

4.1 Introduction

In economics and other fields, researchers often wish to consider statistical models that allow for more complex relationships than can be inferred using only cross-sectional data. Panel, or longitudinal, data, in which the same units are observed repeatedly at different points in time, can often provide the richer data needed for such models (Chamberlain, 1984; Hsiao, 1986; Baltagi, 1995). Although panel data allow researchers to address more complex models than cross-sectional data, missing data problems can be more severe in panels. In particular, even units who respond in initial waves of the panel may drop out of the sample in subsequent waves, so that the subsample with complete data for all waves of the panel can be less representative of the population than the original sample (e.g., Hausman and Wise, 1979; Ridder, 1990; Verbeek and Nijman, 1992; Abowd, Crepon, Kramarz, and Trognon, 1995; Vella, 1998).

Sometimes, in the hope of mitigating the effects of attrition without losing the fundamental advantages of panel data over cross-sections, panel data sets are augmented by replacing units who have dropped out with new units randomly sampled from the original population. Following Ridder (1992), who used these replacement units to test conventional models for attrition, we call such additional samples *refreshment samples*. Data sets that include a refreshment component include the Dutch Transportation Panel analyzed in this paper, and the Health and Retirement Survey. In addition *rotating panels* such as the Current Population Survey can be interpreted as including a refreshment sample every period.

In this paper we explore the benefits of refreshment samples for inference in the pres-

ence of attrition. The two themes of the paper are, first, that refreshment samples can improve inference under conventional models by providing additional sample information, and second, that refreshment samples allow for estimation of more general models of attrition without requiring auxiliary assumptions on distributions of the response variables. Thus, refreshment samples are potentially a relatively inexpensive way to improve the quality of longitudinal surveys.

In the following section we lay out the structure of the data and the inferential problem. In Section 4.3 we describe two conventional models for attrition in panel data. The first model, essentially based on the *missing at random* assumption (MAR, Rubin, 1976; Little and Rubin, 1987), allows the probability of attrition to depend on lagged but not on contemporaneous variables. This case is sometimes also referred to as *selection on observables* (Moffitt, Fitzgerald, and Gottschalk 1997). The second model (denoted by HW in the remainder of the paper, given its origin in a model developed by Hausman and Wise (1979)), allows the probability of attrition to depend on contemporaneous, but not on lagged, variables. Related models have also been referred to as *selection on unobservables* (Moffitt, Fitzgerald, and Gottschalk 1997) because attrition partly depends on variables that are not observed when the unit drops out. In Sections 4.4 and 4.5 we present the main theoretical contributions. We develop a model for attrition that includes the MAR and HW models as special cases. This Additive Nonignorable (AN) model is identified with no testable implications from a panel data set with a refreshment sample. We first discuss in Section 4.4 the identification issues in a simple two-period context with a single binary variable in both periods, and generalize the model to allow for multi-valued variables as

well as time-invariant covariates in Section 4.5.

In Section 4.6 we apply the ideas presented in this paper to a panel data set on travel behavior in the Netherlands, the Dutch Transportation Panel (DTP). This data set is based on a survey of Dutch households concerning their use of various modes of transportation. For a number of years households were asked to keep a detailed travel diary for an entire week each year. For every trip taken by a household member detailed information was gathered including destination, time, and mode of transportation. Attrition was severe, and because the considerable effort required to respond to the survey was directly related to the value of one of the variables (total number of trips taken by household members), it is plausible that among those who responded in the first wave the willingness to cooperate in the second wave depended on the value of these variables in either first or second period. Because MAR essentially rules out dependence on second period variables, and the HW model rules out dependence on first period variables, this makes for a potentially interesting comparison of the performance of the MAR and HW models. In this application we implement the models by multiply imputing the missing data according to the various missing data models, and compare the results both in terms of estimates of the relation of the change in the number of trips to the change in income as well as in terms of estimates of the attrition process itself. Imputation has the advantage over joint estimation of the model for attrition and the substantive model of interest that it does not the researcher from to adapt the estimation procedures to allow for missing data. Given the complex nature of modern methods for estimating panel data models including non-differentiable objective functions (e.g., Honore, 1992) and kernel methods

(e.g., Kyriazidou, 1997), this can be a substantial benefit.

4.2 The Sampling Framework

In the first time period we draw a random sample of size N_p from a fixed population. For each unit i, for $i = 1, \ldots, N_P$, we observe an outcome variable Z_{i1}. For a subset of size N_{CP} of this sample, we observe in the second period a second variable Z_{i2}; the remaining N_{IP} units have dropped out of the panel and all of their Z_{i2} are missing. We refer to the first subsample, as the "complete panel" subsample, the second subsample, as the "incomplete panel" subsample, and the two combined as the "panel", with $N_P = N_{CP} + N_{IP}$. In addition to the panel data set we draw in the second period a new random sample from the original population, the "refreshment" subsample, of size N_R. For these units we observe Z_{i2} but not Z_{i1}. Finally, for all $N = N_{CP} + N_{IP} + N_R$ units we observe a vector of time-invariant covariates denoted by X_i.

We formalize the data-generating process as follows. Each unit in the population is randomly assigned a three-valued sampling indicator A_i. If assigned $A_i = 2$, unit i is designated to be part of the panel and will be approached in both periods. If assigned $A_i = 1$, unit i is designated to be part of the refreshment sample and will be approached only in the second period. Finally, if assigned $A_i = 0$, the unit will not be approached at all. We assume that all units respond the first time they are approached: if assigned $A_i = 2$, unit i will respond in the first period, and thus Z_{i1} and X_i will be recorded, whereas if assigned $A_i = 1$, unit i will respond in the second period and Z_{i2} and X_i will be recorded. Not all units, however, respond the second time they are approached. In

particular, not all units assigned $A_i = 2$ will respond in the second period. Let W_i be an indicator denoting the willingness to respond repeatedly; $W_i = 1$ implies that unit i, if approached in the second period after already having responded in the first period, will respond, and Z_{i2} will be recorded, and $W_i = 0$ implies that unit i, if approached in the second period after already having responded in the first period, will not respond, and Z_{i2} will not be recorded. The willingness to respond W_i is observed if and only if the researcher attempts to get a second response from unit i; that is we observe W_i only when $A_i = 2$. We can use the survey design variable, A_i, and the willingness to respond indicator, W_i, to define two missing data indicators, D_{i1} and D_{i2}. When $D_{i1} = 1$, we observe Z_{i1} and when $D_{i1} = 0$ we do not observe Z_{i1}. Similarly, when $D_{i2} = 1$ we observe Z_{i2} and when $D_{i2} = 0$ we do not observe Z_{i2}:

$$D_{i1} = 1\{A_i = 2\}, \qquad \text{and} \qquad D_{i2} = 1\{A_i = 1\} + 1\{A_i = 2\} \cdot W_i.$$

The two missing data indicators are always observed. Table 4.2 illustrates the missing data pattern. Note that the missing data pattern is not *monotone* (Little and Rubin 1987); for some units Z_{i1} is missing but Z_{i2} is observed whereas for others Z_{i2} is missing but Z_{i1} is observed. The missing data pattern resembles that studied in the literature on estimation of cell frequencies in contingency tables with known marginals (e.g., Little and Wu, 1991; Little and Wang, 1996). The main differences are that we do not assume exact knowledge of the marginal distributions, allow for continuous as well as discrete variables, and focus on more general estimands than cell probabilities.

The design variable A_i is under control of the surveyor, and by assumption it satisfies

Table 4.1: MISSING DATA PATTERN: OBSERVED (×) OR MISSING (–), OR ACTUAL VALUES (0, 1, OR 2)

Missing Data Indicators		Design Variable	Individual Characteristics			
D_{i1}	D_{i2}	A_i	W_i	Z_{i1}	Z_{i2}	X_i
0	0	0	–	–	–	–
0	1	1	–	–	×	×
1	1	2	1	×	×	×
1	0	2	0	×	–	×

the following independence condition:[2]

$$A_i \perp W_i, Z_{i1}, Z_{i2}, X_i. \tag{4.1}$$

In the remainder of this section we make some comments on this sampling framework and provide some motivation for the special focus of the paper.

First, the substantive interest is in the joint distribution of (Z_{i1}, Z_{i2}, X_i), or possibly the conditional distribution of (Z_{i1}, Z_{i2}) given X_i. The distribution of the willingness to respond is of concern solely because its properties can affect our ability to recover the distributions of interest. The willingness to respond repeatedly W_i, or its complement, the attrition indicator $1 - W_i$ is interpreted as an individual characteristic. It has the unusual feature that it can only be revealed by actions of the surveyor: by approaching a person in the first period, that is by assigning unit i the value $A_i = 2$, this willingness to

[2] In fact this independence condition is stronger than necessary. In practice we are primarily interested in the conditional distribution of (Z_{i1}, Z_{i2}) given X_i rather than the joint distribution of (Z_{i1}, Z_{i2}, X_i), and for features of this distribution it suffices that

$$A_i \perp W_i, Z_{i1}, Z_{i2} \mid X_i,$$

combined with positive probabilities for all values of A_i conditional on X_i.

respond gets revealed in the second period. It is also important to stress again that there is no intrinsic interest in the distribution of the willingness to respond.

Second, we assume throughout the analysis that we always oberve Z_{i1} when we assign $A_i = 2$ to unit i, and similarly we always observe Z_{i2} when we assign $A_i = 1$ to unit i. In general, however, there might be non-response of the standard cross-section type present, where we know nothing about units other than that they did not respond, or where we know some variables but not others for some units. Ridder (1992) discusses these issues for the particular data set we use in this paper. We ignore such issues here to focus on the specific panel data problem of the use of refreshment samples when there is attrition of units who are initially prepared to respond but choose not to do so in subsequent waves of the panel.

A third issue is the focus on the specification of the attrition probability

$$Pr(W_i = 1|Z_{i1}, Z_{i2}, X_i).$$

To justify this focus, consider the joint distribution of (Z_{i1}, Z_{i2}, X_i). Assuming that the attrition probability is strictly less than one,

$$f(Z_{i1}, Z_{i2}, X_i) = \frac{f(Z_{i1}, Z_{i2}, X_i|W_i = 1) \cdot Pr(W_i = 1)}{Pr(W_i = 1|Z_{i1}, Z_{i2}, X_i)}. \tag{4.2}$$

Because $A_i \perp W_i, Z_{i1}, Z_{i2}, X_i$, the two factors in the numerator of the righthand side can be rewritten as

$$f(Z_{i1}, Z_{i2}, X_i|W_i = 1) = f(Z_{i1}, Z_{i2}, X_i|W_i = 1, A_i = 2) = f(Z_{i1}, Z_{i2}, X_i|D_{i1} = 1, D_{i2} = 1),$$

and

$$Pr(W_i = 1) = Pr(W_i = 1|A_i = 2) = Pr(D_{i2} = 1|D_{i1} = 1),$$

both of which can be estimated directly from the panel data set. Specification of the attrition probability $Pr(W_i = 1|Z_{i1}, Z_{i2}, X_i)$ therefore specifies the joint distribution of (Z_{i1}, Z_{i2}, X_i).

The final comment concerns inference under a specific model for attrition. Given knowledge of the conditional probability $Pr(W_i = 1|Z_{i1}, Z_{i2}, X_i)$, inference can proceed in different ways. Assuming the probability of $W_i = 1$ conditional on Z_{i1}, Z_{i2}, and Z_{i2} is greater than zero, one can use the inverse of this conditional probability to weight the complete panel, that is, the observations with $A_i = 2$ and $W_i = 1$, and use weighted versions of the complete data estimation techniques (e.g., Hansen, Hurwitz, and Madow, 1996; Imbens and Hellerstein, 1994; Nevo, 1995). Alternatively, one can use this conditional probability to impute the missing values and use complete data estimation techniques on the imputed data sets (Rubin, 1987, 1996). See Brownstone and Valetta (1996) and Stasny and Reagan (1997) for recent economic applications. In the application in Section 4.6 we use the second of these approaches. One reason for our choice is that the theoretical models we develop in Sections 4.4 and 4.5 are conveniently implemented using the Data Augmentation (DA) algorithm proposed by Tanner and Wong (1987), which as a by-product generates imputed data sets that allow for standard complete-data estimation.[3]

[3]The imputation approach is particularly convenient for our application because it simultaneously allows us to deal with the fact that one of the variables, income, is only observed to lie in one of four intervals. This creates complications even in the absence of attrition if, for example, we wish to regress the level of income on another variable. Here we impute the value of income as part of the general imputation procedure.

4.3 Models for Panel Data with Attrition

In this section we review two models that have been proposed in the literature to address the problem of attrition in panel data. In terms of the notation defined in the previous section, we only have the subsample with $A_i = 2$. In Section 4.4.1 we shall see that these models employ assumptions that, although to some extent unavoidable in the context for which they were designed, can be relaxed substantially in the presence of refreshment samples. In the application in Section 4.6 we shall evaluate the appropriateness of these assumptions for the particular data set analyzed.

4.3.1 Missing at Random (MAR)

The first model makes the assumption that Z_{i2} is missing at random (MAR) in the panel,

$$W_i \perp Z_{i2} \mid Z_{i1}, X_i \quad \text{(MAR)}, \tag{4.3}$$

implying that if the parameters of the missing data process are distinct from those of the data distribution, then the missing data process is *ignorable* (Rubin, 1976; Little and Rubin, 1987).

The special case arising when

$$W_i \perp Z_{i1}, Z_{i2}, X_i \quad \text{(MCAR)},$$

is referred to as *missing completely at random* (MCAR). In that case no bias results from limiting the analysis to the complete panel with $D_{i1} = D_{i2} = 1$.

4.3.2 The Hausman–Wise (HW) model for Attrition

The second model for panel data with attrition we consider is closely related to a model developed by Hausman and Wise (1979), and more generally is related to models developed for cross-sectional surveys by Heckman (1979). Hausman and Wise allow the probability of attrition in the second period to depend on the contemporaneous, that is second period, variables Z_{i2} but assume that the first period variables do not affect this probability:

$$W_i \perp Z_{i1} \mid Z_{i2}, X_i \quad \text{(HW)}. \tag{4.4}$$

The original formulation of the Hausman–Wise model (Hausman and Wise 1979) also restricts the joint distribution of Z_{i1} and Z_{i2} and assumes normality of some of the variables, but these restrictions can partly be relaxed and do not concern us here. The appeal of these models is that they can reflect optimal behavior of the respondent whose effort in responding is related to the anticipated value of Z_{i2}. An implication is that the distribution of Z_{i2} given (Z_{i1}, X_i) for those with $W_i = 0$ differs in a systematic way from the distribution for those with $W_i = 1$. Again the MCAR case is a special case of this attrition model. In a conventional panel survey with no refreshment samples, neither MAR or HW are testable without auxiliary assumptions.

4.4 A Simple Example with Binary Variables

In this section we assume that Z_{i1} and Z_{i2} are binary variables and suppress the conditioning on time-invariant covariates X_i. Denote the conditional probability $Pr(Z_{i2} = 1|Z_{i1} = z, W_i = w)$ by q_{zw}, and the probability $Pr(Z_{i1} = z, W_i = w)$ by r_{zw}. In large

samples we can learn the value of r_{zw} for $z, w \in \{0, 1\}$ because the subsample with $A_i = 2$ is a random sample from the population, and for this subsample we always observe Z_{i1} and W_i. Similarly we can learn in large samples the values of q_{z1} for $z = 0, 1$, because $A_i \perp (W_i, Z_{i1}, Z_{i2})$ implies that the subsample with $A_i = 2$, $W_i = 1$ and $Z_{i1} = z$ is a random sample from the subpopulation with $W_i = 1$ and $Z_{i1} = z$, and for this subsample we always observe Z_{i2}. The data with $A_i = 2$, however, contain no information on q_{z0} because we never observe Z_{i2} if $W_i = 0$ and $A_i = 2$.

The subsample with $A_i = 1$ allows us to deduce in large samples the marginal distribution of Z_{i2}. Since

$$Pr(Z_{i2} = 1) = \sum_{z,w} q_{zw} \cdot r_{zw},$$

knowledge of marginal distribution of Z_{i2} implies a single linear restriction on the two remaining parameters q_{10} and q_{00} in terms of the directly estimable parameters q_{01}, q_{11} and r_{00}, r_{01}, r_{10}, and r_{11}. The panel and refreshment sample combined therefore do *not* enable us to estimate the values of q_{00} and q_{10} uniquely from the population distribution of the observed data, although they do imply a linear restriction on these two parameters.

4.4.1 Testable Implications of the MAR and HW Models in the Presence of Refreshment Samples

The MAR and HW models do not require the refreshment sample for estimation of q_{00} and q_{10}. The independence assumptions (4.3) and (4.4) each imply two restrictions on the eight parameters r_{zw} and q_{zw} that are sufficient for identication of q_{00} and q_{10}. Specifically,

MAR implies

$$q_{00} = q_{01},$$

and

$$q_{10} = q_{11}.$$

Under the HW assumption the relation between q_{00} and q_{10} and the directly estimable parameters is more complex:

$$q_{00} = \frac{r_{10} \cdot r_{01} \cdot (1 - q_{01}) - r_{11} \cdot r_{00} \cdot (1 - q_{11})}{r_{00} \cdot r_{11} \cdot q_{11} \cdot (1 - q_{01})/q_{01} - r_{11} \cdot r_{00} \cdot (1 - q_{11})},$$

and

$$q_{10} = \frac{q_{00} \cdot r_{00} \cdot q_{11} \cdot r_{11}}{q_{01} \cdot r_{01} \cdot r_{10}}.$$

Under either of these two models, therefore, we can estimate all eight parameters from the panel alone. In each case these estimates imply a marginal distribution for Z_{i2}. This distribution can be compared to the distribution of Z_{i2} in the refreshment sample to test the attrition model that generated it.

To illustrate these issues we use in this section a subset of the data set that will be analyzed in more detail in Section 4.5. We define a binary variable indicating whether the total number of trips for a household during the survey week was less than or equal to twenty-five. Table 4.2 summarizes the sample information for this variable and Table 4.3 presents estimates of the six parameters that are directly estimable from the panel data alone as well as estimates for q_{00} and q_{10} under the MAR and HW assumptions.

Assuming MAR, the panel subsample with $A_i = 2$ leads to the estimates $\hat{q}_{00} = \hat{q}_{01} = 0.074$, and $\hat{q}_{10} = \hat{q}_{11} = 0.602$, which in turn implies that the marginal probability of

Table 4.2: SUMMARY STATISTICS DUTCH TRANSPORTATION PANEL: Z_{it} IS INDICATOR FOR NUMBER OF TRIPS IN THE PERIOD t LESS THAN OR EQUAL TO 25, AND W_i IS INDICATOR FOR WILLINGNESS TO RESPOND IN THE SECOND PERIOD.

Subsample	Z_{i1}	Z_{i2}	W_i	No of obs
Complete Panel	0	0	1	832
	0	1	1	66
	1	0	1	53
	1	1	1	80
Incomplete Panel	0	–	0	518
	1	–	0	215
Refreshment Sample	–	0	–	520
	–	1	–	136

Table 4.3: ESTIMATES BASED ON THE PANEL SUBSAMPLE

	Directly Estimable Parameters			Model–based Estimates	
				MAR	HW
	$W_i = 0$	$W_i = 1$	q_{z1}	q_{z0}	q_{z0}
$Z_{i1} = 0$	$\hat{r}_{00} = 0.294$	$\hat{r}_{01} = 0.509$	$\hat{q}_{01} = 0.074$	$\hat{q}_{00} = 0.074$	$\hat{q}_{00} = 0.306$
$Z_{i1} = 1$	$\hat{r}_{10} = 0.122$	$\hat{r}_{11} = 0.075$	$\hat{q}_{11} = 0.602$	$\hat{q}_{10} = 0.602$	$\hat{q}_{10} = 0.894$

the number of trips in the second period exceeding twenty-five is $\hat{r}_{00} \cdot \hat{q}_{00} + \hat{r}_{01} \cdot \hat{q}_{01} + \hat{r}_{10} \cdot \hat{q}_{10} + \hat{r}_{11} \cdot \hat{q}_{11} = 0.178$. This differs from the marginal probability of the number of trips in the second period exceeding twenty-five implied by the refreshment sample, which is $136/(136 + 520) = 0.207$. A likelihood ratio test, however, with a nominal $\mathcal{X}^2(1)$ distribution, gives a test statistic of 2.2, so the difference is not statistically significant at conventional levels.

Assuming HW and again ignoring sampling error, the two proportions that cannot

directly be estimated from the data are

$$\hat{q}_{00} = \frac{\hat{r}_{10} \cdot \hat{r}_{01} \cdot (1 - \hat{q}_{01}) - \hat{r}_{11} \cdot \hat{r}_{00} \cdot (1 - \hat{q}_{11})}{\hat{r}_{00} \cdot \hat{r}_{11} \cdot \hat{q}_{11} \cdot (1 - \hat{q}_{01})/\hat{q}_{01} - \hat{r}_{11} \cdot \hat{r}_{00} \cdot (1 - \hat{q}_{11})} = 0.306,$$

$$\hat{q}_{10} = \frac{\hat{q}_{00} \cdot \hat{r}_{00} \cdot \hat{q}_{11} \cdot \hat{r}_{11}}{\hat{q}_{01} \cdot \hat{r}_{01} \cdot \hat{r}_{10}} = 0.894,$$

leading to an estimate for the marginal probability of the number of trips in the second period being less than or equal to twenty-five of 0.282, substantially different from the refreshment sample estimate of 0.207. A likelihood ratio test gives 7.8, with a nominal $\mathcal{X}^2(1)$, statistically significant at the 0.05 level.

4.4.2 The Additive Nonignorable (AN) Model

The above discussion demonstrates that MAR and HW models have testable implications if refreshment samples are available, suggesting that more general models may be identified. We therefore proceed to develop a model that generalizes MAR and HW in a way that has no testable implications. Suppose we model, with no essential loss of generality[4] given the binary nature of Z_{i1} and Z_{i2}, the probability of response as

$$Pr(W_i = 1 | Z_{i1} = z_1, Z_{i2} = z_2) = g(\alpha_0 + \alpha_1 \cdot z_1 + \alpha_2 \cdot z_2 + \alpha_3 \cdot z_1 \cdot z_2), \qquad (4.5)$$

for some known, increasing $g(\cdot)$ satisfying $\lim_{a \to -\infty} g(a) = 0$, $\lim_{a \to \infty} g(a) = 1$. With Z_{i1} and Z_{i2} binary this saturates the model, implying that the choice of $g(\cdot)$ is irrelevant, and the model places essentially no restrictions on the data-generating process. Assuming MAR (HW) in this context amounts to assuming $\alpha_2 = \alpha_3 = 0$ ($\alpha_1 = \alpha_3 = 0$), and in each case the choice of $g(\cdot)$ is irrelevant.

[4]Other than that we continue to require the attrition probability to be strictly between zero and one.

Without restrictions on α the model in (4.5) is saturated. The discussion in the introduction to Section 4.4 therefore implies that this model is not identified, and we cannot estimate all four parameters α_1, α_2, α_3, and α_4 consistently from a random sample of $(D_{i1} \cdot Z_{i1}, D_{i2} \cdot Z_{i2}, D_{i1}, D_{i2})$ from the population. Figure 1 illustrates this issue in (q_{00}, q_{10}) space for the data from Table 4.2. All values of (q_{00}, q_{10}) between zero and one are consistent with the panel data. For a given $g(\cdot)$, every point (q_{00}, q_{10}), combined with the data in Table 4.2 corresponds to a unique set of values for $(\alpha_0, \alpha_1, \alpha_2, \alpha_3)$ in (4.5). The "$\times$" indicates the MAR point $(q_{00}, q_{10}) = (0.074, 0.602)$, where $\alpha_2 = \alpha_3 = 0$. The "+" indicates the HW point $(q_{00}, q_{10}) = (0.306, 0.894)$ where $\alpha_1 = \alpha_3 = 0$. Finally, the solid line indicates the set of (q_{00}, q_{10}) that are consistent with $Pr(Z_{i2} = 1) = 0.207$ given the values of the directly estimable parameters.

Figure 1 illustrates the main issues of the paper. The most general model, not restricting the α's, corresponds to the entire (q_{00}, q_{10}) space. Given point estimates of the directly estimable parameters, a model, specified in terms of restrictions on the conditional probability $Pr(W_i = 1|Z_{i1} = z_1, Z_{i2} = z_2)$, corresponds to a point or set of points in (q_{00}, q_{10}) space. For example the MAR model that specifies $(Pr(W_i = 1|, Z_{i1} = z, Z_{i2} = 1) = Pr(W_i = 1|Z_{i1} = z, Z_{i2} = 0)$, or, equivalently, $\alpha_2 = \alpha_3 = 0$ in (4.5), corresponds to the point marked by "$\times$". Our approach can now be described as follows. We wish to develop models that satisfy two conditions. First, the model should always be consistent with the data, which in terms of this figure means that the intersection of the set of (q_{00}, q_{10}) consistent with the model and the set of points on the solid line (i.e., the set of points that corresponds to the marginal distribution for Z_{i2} estimated from the refreshment sample)

is nonempty for any set of observed values of $\hat{q}_{zw}$ and $\hat{r}_{zw}$. Second, the model should be identified. That is, this intersection should contain only a single point. The models imposing MAR or HW *a priori* fail the first condition because they are not always consistent with the data, and the general unrestricted model fails the second because it is not identified.

The alternative family of models we suggest has the form

$$Pr(W_i = 1 | Z_{i1} = z_1, Z_{i2} = z_2) = g(\alpha_0 + \alpha_1 \cdot z_1 + \alpha_2 \cdot z_2), \tag{4.6}$$

for unrestricted values of the unknown parameters α_0, α_1, and α_2. This model rules out an interaction term between Z_{i1} and Z_{i2}, but allows for non-ignorable models by allowing α_2 to differ from zero. To reflect the additivity of the index in the $g(\cdot)$ function in first and second period variables we refer to this as the additive non-ignorable (AN) model. Note that both the MAR and HW models are special cases of this model.

In Figure 2 we illustrate, for the case of a logistic $g(a) = \exp(a)/(1 + \exp(a))$, the set of points consistent with both the model and the directly estimable q_{zw} and r_{z1} (the solid curve). Note that there is a single point of intersection between this set and the set of points consistent with the marginal distribution of Z_{i2} (solid line). Also note that the set of points consistent with the model includes both the MAR and HW points. The latter is trivial because the MAR (HW) point corresponds to $\alpha_2 = 0$ ($\alpha_1 = 0$) in the model. Both points are true in general, irrespective of the choice of $g(\cdot)$ and the population distributions, as the following result shows.

Theorem 1 *For any quadruple* $q_{01}, q_{11} \in (0,1)$, *any quadruple* $r_{zw} \in (0,1)$ *with* $\sum_{zw} r_{zw} =$

1, *and any continuous and increasing* $g(\cdot)$ *with* $\lim_{a\to-\infty} g(a) = 0$ *and* $\lim_{a\to\infty} g(a) = 1$, *there is a unique quintuple* $(\alpha_0, \alpha_1, \alpha_2, \hat{q}_{00}, \hat{q}_{10})$ *with* $\hat{q}_{00}, \hat{q}_{10} \in (0,1)$ *such that the following five conditions are satisfied:*

$$g(\alpha_0) = (1 - q_{01})r_{01}/((1 - q_{01})r_{01} + (1 - \hat{q}_{00})r_{00}), \tag{4.7}$$

$$g(\alpha_0 + \alpha_1) = (1 - q_{11})r_{11}/((1 - q_{11})r_{11} + (1 - \hat{q}_{10})r_{10}), \tag{4.8}$$

$$g(\alpha_0 + \alpha_2) = q_{01}r_{01}/(q_{01}r_{01} + \hat{q}_{00}r_{00}), \tag{4.9}$$

$$g(\alpha_0 + \alpha_1 + \alpha_2) = q_{11}r_{11}/(q_{11}r_{11} + \hat{q}_{10}r_{10}), \tag{4.10}$$

and

$$\hat{q}_{00}r_{00} + \hat{q}_{10}r_{10} + q_{01}r_{01} + q_{11}r_{11} = q_{00}r_{00} + q_{10}r_{10} + q_{01}r_{01} + q_{11}r_{11}. \tag{4.11}$$

PROOF: See Appendix A.

An important consequence of our approach is that the solutions $\hat{q}_{10}$ and $\hat{q}_{00}$ depend on the choice of $g(\cdot)$ function. Every $g(\cdot)$ function corresponds in Figure 2 to a curve approaching $(0,0)$, going through both the MAR and HW points, and approaching $(1,1)$. Nevertheless the exact point of intersection with the set of points corresponding to the restriction from the marginal distribution of Z_{i2} will in general depend on the choice of $g(\cdot)$. This differs qualitatively from both the MAR and HW models where the functional form of the selection probability is immaterial. For example, if the probability of attrition does not depend on Z_{i2}, $g(\cdot)$ cancels from the restrictions in equations 4.7–4.10. The question arises how sensitive the results are to alternative choices of $g(\cdot)$. We therefore estimate the same model using a normal distribution function, or

$g(a) = \Phi(a) = \int_{-\infty}^{a}(1/\sqrt{(2\pi)})\exp(-z^2/2)dz$. The dashed curve in Figure 2 illustrates the resulting set of points consistent with the panel data and the probit version of the AN model. It is clear that, as in conventional binary response models, there is very little difference between the logit and the probit model. This is not surprising given that both approach the points $(0,0)$ and $(1,1)$, as well as go through the MAR and HW points. The difference between the two models around the intersection with the set of points consistent with the refreshment sample is so small that in Figure 3 we enlarge the area in the rectangle around this intersection in Figure 2.

4.4.3 Estimates in the Binary Case

Now let us return to the binary data example and consider estimation of the joint distribution of (Z_{i1}, Z_{i2}) using the four different models and different combinations of data. For ease of exposition we focus on a single feature of this distribution

$$Pr(Z_{i1} = 0, Z_{i2} = 1) = r_{00}q_{00} + r_{01}q_{01}. \tag{4.12}$$

The rows in Table 4.4 correspond to the different models for attrition. The first row is based on the MCAR assumption. The second row is based on the MAR assumption. The third row presents estimates based on the HW model, and the next two rows are based on the AN model, using the logistic and normal distribution function for $g(\cdot)$. The different columns correspond to different data sets. The first column presents estimates based on the complete panel data set alone (observations with $A_i = 2$ and $W_i = 1$). Only the MCAR model is identified in that case, so only estimates for this model are reported. The next column presents estimates based on the panel data alone (all observations with $A_i = 2$).

Now the HW and MAR models are identified as well, so estimates are reported for the first three models. Finally, the last column reports estimates based on all observations. In this case all models are identified (MCAR, MAR, and HW are in fact overidentified), and estimates are reported for all five models, MCAR, MAR, HW, and the logit and probit versions of AN.

In the last row we also report nonparametric bounds on the probability of $Pr(Z_{i1} = 0, Z_{i2} = 1)$, in the spirit of work by Manski (1995) and Horowitz and Manski (1998). These bounds demonstrate the identifying power of the various parts of the data set. In the first set of bounds on $Pr(Z_{i1} = 0, Z_{i2} = 1) = r_{00}q_{00} + r_{01}q_{01}$ we assume that we have knowledge of the complete panel subsample and the marginal probability of attrition. This allows us to estimate r_{01} and q_{01}. Nothing is known about q_{00} (other than that it is between zero and one) and r_{00} is only known to lie in the interval $(0, \hat{r}_{00} + \hat{r}_{10}) = (0, 0.416)$ since we only know the frequency of attrition. In the second set of bounds we add the incomplete panel subsample. This allows us to estimate r_{00} as 0.294 but still nothing is known about q_{00}. In the last set of bounds we add the refreshment sample. This narrows down the interval for q_{00} from $(0, 1)$ to $(0.007, 0.422)$. In each case the bounds narrow with the additional information.

An important feature of Table 4.4 is that the differences in estimates between models generally decrease in magnitude the more data are used. Consider for example the differences between the HW and MAR estimates of $Pr(Z_{i1} = 0, Z_{i2} = 1)$ based on panel data alone: 0.043 and 0.079 respectively. The estimates are considerably closer when we also use the refreshment sample: 0.048 and 0.075, and both are closer to the AN estimates

Table 4.4: MAXIMUM LIKELIHOOD ESTIMATES OF $Pr(Z_{i1} = 0, Z_{i2} = 1)$

Model	$g(\cdot)$	Complete Panel ($A_i = 2,\ W_i = 1$)	Panel ($A_i = 2$)	Panel and Refreshment Sample ($A_i = 1, 2$)
MCAR		0.051	0.079	0.070
MAR			0.079	0.075
HW			0.043	0.048
AN	logit			0.073
AN	probit			0.073
Bounds		(0.038, 0.454)	(0.038, 0.332)	(0.040, 0.162)

of 0.073. The exact difference between the probit and logit version of the AN model are extremely small: 0.07341 for the logit version, and 0.07339 for the probit version. In the binary case the estimates based on the MAR and HW models do not depend on the choice of $g(\cdot)$.

4.5 Identification with Multi-valued and Time-invariant Variables

In this section we generalize the identification result in Section 4.4.3 to allow for multi-valued and time-invariant variables. All variables, Z_{i1}, Z_{i2}, and X_i are assumed to have finite (but possibly very large) support.

Theorem 2 *Let $f(z_1, z_2, x)$ be the joint probability function of (Z_{i1}, Z_{i2}, X_i), and let $p(z_1, z_2, x)$ be the conditional probability that $W_i = 1$ given (Z_{i1}, Z_{i2}, X_i), and let $0 <$*

$p(z_1, z_2, x) < 1$ for all (z_1, z_2, x) in the support of (Z_{i1}, Z_{i2}, X_i). Let $\{(z_{1k}, z_{2k})\}_{k=1}^{K}$ be the support of (Z_{i1}, Z_{i2}). Finally, let $g(\cdot)$ be a continuous, increasing function with $\lim_{a \to -\infty} g(a) = 0$, and $\lim_{a \to \infty} g(a) = 1$.

Then there is a unique set of functions $\hat{f}(z_1, z_2, x)$, $k_0(x)$, $k_1(z, x)$ and $k_2(z, x)$ such that for some $(\overline{z}_1, \overline{z}_2)$ in the support of (Z_{i1}, Z_{i2}):

(i) $k_1(\overline{z}_1, x) = 0$, $k_2(\overline{z}_2, x) = 0$,

(ii) $\sum_{z_2} \hat{f}(z_1, z_2, x) = \sum_{z_2} f(z_1, z_2, x)$,

(iii) $\sum_{z_1} \hat{f}(z_1, z_2, x) = \sum_{z_1} f(z_1, z_2, x)$,

(iv)

$$\hat{f}(z_1, z_2, x) = f(z_1, z_2, x) \cdot \frac{g(k_0(x) + k_1(z_1, x) + k_2(z_2, x))}{p(z_1, z_2, x)}.$$

Proof: see Appendix A

The theorem implies that given any joint distribution of (Z_{i1}, Z_{i2}, X_i), and given any attrition probability $p(z_1, z_2, x)$, we can find a joint distribution of (Z_{i1}, Z_{i2}, X_i) with the additive nonignorable attrition model that is "observationally equivalent" under our sampling frame. That is, take a distribution $f(z_1, z_2, x)$, and an attrition probability $p(z_1, z_2, x)$. Then we can find another distribution $\hat{f}(z_1, z_2, x)$ and another attrition probability function $\hat{p}(z_1, z_2, x)$ that leads to the same directly estimable distributions. This means the implied joint distribution of (Z_{i1}, X_i) is the same:

$$\sum_{z_2} \hat{f}(z_1, z_2, x) = \sum_{z_2} f(z_1, z_2, x).$$

In addition the joint distribution of (Z_{i2}, X_i) is the same:

$$\sum_{z_1} \hat{f}(z_1, z_2, x) = \sum_{z_1} f(z_1, z_2, x).$$

Finally, the conditional distribution given $W_i = 1$ is the same:

$$\frac{\hat{f}(z_1, z_2, x) \cdot \hat{p}(z_1, z_2, x)}{\sum_{z_1, z_2, x} \hat{f}(z_1, z_2, x) \cdot \hat{p}(z_1, z_2, x)} = \frac{f(z_1, z_2, x) \cdot p(z_1, z_2, x)}{\sum_{z_1, z_2, x} f(z_1, z_2, x) \cdot p(z_1, z_2, x)}.$$

At this point there are clearly many such distributions $\hat{f}(\cdot)$ and attrition probabilities $\hat{p}(\cdot)$, including the distribution that generated the data. The theorem implies, however, that we can find a solution that imposes a particular structure on $\hat{p}(\cdot)$, namely that it can be written as additive in z_1 and z_2: $\hat{p}(z_1, z_2, x) = g(k_0(x) + k_1(z_1, x) + k_2(z_2, x))$ for the given choice of $g(\cdot)$. Because we can do this irrespective of the original distribution $f(z_1, z_2, x)$ and the attrition probability $p(z_1, z_2, x)$, the theorem implies that the model has no testable implications, unlike the MAR and HW models. In addition the theorem implies that this solution is unique, or that the model is identified.

Note that condition (i) is a normalization that is required because of the inclusion of a constant in the additive nonignorable model. The fact that the functions k_1 and k_2 are unrestricted makes it impossible to identify the distribution function $g(\cdot)$. It is well-known that any probability model for a binary dependent variable can be expressed as a logit by choosing an appropriate functional form for the dependence on the explanatory variables. The same is true in the AN model. Restricting these functions imposes restrictions on $g(\cdot)$.

4.6 Travel Behavior in The Netherlands

We apply the models discussed in Sections 4.3 through 4.5 to a data set on travel behavior in the Netherlands. First we describe the data. Then we specify a fully parametric

model which we use only to impute the missing data. Finally we repeatedly estimate the substantive model (a linear regression model in differences) on the complete data sets and average the estimates to get the reported estimates.

Although the identification results in the previous section are nonparametric in the sense that no distributional assumptions are made concerning the conditional distribution of (Z_{i1}, Z_{i2}) given X_i, and no assumptions are made regarding the functional form of $k_0(x)$, $k_1(x,z)$ and $k_2(x,z)$, we shall make make such assumptions here. The reason is that with a modest sample size, estimates without such smoothing assumptions would be likely to have poor sampling properties. Since we are only using the parametric model to impute missing values, and since the nonparametric identification result of the previous section ensures that the imputation is not being solely driven by the parametric form, we do not believe that much is lost in practice. In a sufficiently large sample the parametric restrictions could be tested by classical methods, or nonparametric imputation could be carried out.

4.6.1 The Data

The DTP collected information regarding travel behavior for six years. The households in the first wave were interviewed in March 1984, just before an increase in the price of public transportation, and the last wave was interviewed in March 1989. Households were approached twice a year, in the spring and in the fall. The spring waves involved face-to-face interviews, and the fall waves were postal surveys. We use data from the spring waves of 84 and 85, and, except for this discussion of the data, we refer to these waves

as the first and second wave.[5] Of the sample of 6128 households that were selected for the first wave, 2886 (47%) agreed to participate in the panel. In the following analysis we ignore potential biases induced by initial nonresponse. Of these households, 2185 were approached for an interview, and 1764 (81%) of these provided the required information. The main purpose of the survey was to collect detailed information on household travel demand. Every household member over 11 years of age was asked to keep a travel diary, in which he/she reported all trips during a particular week. A trip starts when a household member leaves the home, and it ends upon return. In the sequel we concentrate on the total number of trips made by all household members in the survey week.

The DTP has been plagued by heavy attrition. Only 38% of the original sample participated in all seven waves of the panel. The attrition after the first wave was 41%. It is not unusual that the attrition rate is the highest in the second wave of a panel study. However, in the DTP the design of the second (fall) wave increased the attrition after the first wave. In the fall wave of 1984, which was a postal survey, household members were asked to keep a travel diary. A substantial fraction (21%) of the households did not respond to this request and were dropped from the panel. In later fall waves a travel diary was not asked for. If we correct for this additional attrition, the attrition rate in wave 3 is about 20%, which is about the same as the attrition observed after the first wave in the refreshment samples. In the third (spring 85) wave a refreshment sample of 656

[5]The households in the first wave were obtained by a stratified sample in 20 municipalities of varying sizes. The municipalities were selected on the basis of the number of inhabitants and the availability of different types of public transportation. The sample was stratified by household type (combination of age of head and household composition), and net household income.

households was added to the panel survey[6]

In this paper we only use the first two spring waves. Table 4.5 gives the some summary statistics for the subsamples defined by the value of the design variable A_i, the willingness to participate in the second period, W_i and by period. Included are the mean and standard deviation of the number of trips, the fraction in each of the four income categories (less than 17,000 guilders, between 17,000 and 24,000 guilders, between 24,000 and 38,000 guilders, and more than 38,000 guilders), and the fraction living in a central city. The first row gives the summary statistics for the first wave for those individuals who stayed in the sample for both waves. The second row gives the summary statistics for the first wave for individuals who dropped out of the sample after the first wave. The third row gives t-statistics for the difference in means between the two subsamples, corresponding to the MCAR null hypothesis that the willingness to respond W_i is independent of the number of trips in both periods, the indicators for the income categories, and the city indicator.

The average number of trips in the periods 1 and 2, computed from the unbalanced panel only, is 55.0 and 54.9, respectively, an indication that travel demand has not changed. This is confirmed by the refreshment sample, as Ridder (1992) shows that the decrease in the refreshment sample average is due to differences in the sampling fractions in the strata. However, in the balanced panel there is a (statistically significant) decrease in the average number of trips from 61.8 to 54.9. Moreover, the households that stay in the panel make on average (significantly) more trips than households that drop out after the first wave. Hence, the probability of attrition is negatively correlated with the time average of

[6]Again we ignore the initial nonresponse in the refreshment sample.

the number of trips and positively correlated by an increase after the first wave.

The fourth row gives the statistics for the individuals who stayed in the panel in both periods, and the fifth row gives the results for the refreshment sample. For the last group we do not know the value of W_i. The last row reports t-statistics for the difference in means in the two subsamples. Here the implicit MCAR null hypothesis is that both the design variable A_i and the willingness to respond W_i are independent of all the other variables in the model. Both sets of t-statistics clearly demonstrate that the data are not missing completely at random, and therefore that using only the complete panel with $D_{i1} = D_{i2} = 1$ may be very misleading. These tests do not reflect on the adequacy of the MAR and HW assumptions.

The two key variables, number of trips and income, are both characterized by a high degree of persistence. The correlation between first and second period values of the number of trips for the subsample who stays in the panel in both periods is 0.79. The fraction of individuals in this subsample who stays in the same income category is 0.72.

Table 4.5: Summary Statistics by Period

Period	A_i	W_i	Sample Size	Number of Trips mean	(s.d.)	Earn1 $< 17,000$	Earn2 $\geq 17,000$ $< 24,000$	Earn3 $\geq 24,000$ $< 38,000$	Earn4 $\geq 38,000$	City
1	0	1	1031	61.8	(34.9)	0.14	0.21	0.35	0.29	0.07
1	0	0	733	45.5	(33.2)	0.25	0.23	0.28	0.24	0.14
t-stat for diff. of means				9.9		14.0	2.4	7.7	5.4	14.3
2	0	1	1031	54.9	(30.8)	0.11	0.20	0.38	0.30	0.07
2	1	–	656	51.8	(32.9)	0.15	0.25	0.35	0.25	0.27
t-stat for diff. of means				1.9		6.9	5.6	2.6	5.1	25.1

4.6.2 The Model

We make the following modeling assumptions, using T_{1i} and T_{2i} to denote the number of trips per household for the first and second period, Y_{1i} and Y_{2i} to denote household income for the first and second period, and C_i to denote whether the household lives in a city. Conditional on living in a city, the joint distribution of the logarithm of household income in both periods and the logarithm of the number of trips in both periods (adding one for each household to avoid problems with the less than one percent of the total number of observations with zero trips in the survey week), is assumed to be multivariate normal:

$$\left(\begin{array}{c} \ln(T_{1i}+1) \\ \ln Y_{1i} \\ \ln(T_{2i}+1) \\ \ln Y_{2i} \end{array} \right) \Bigg| \alpha, \mu, \Sigma, C_i \sim \mathcal{N}\left(\mu_0 + \mu_1 \cdot C_i, \Sigma\right).$$

We also assume that, conditional on first and second period income and number of trips, the probability of attrition has a logistic form:

$$Pr(W_i = 1 | T_{1i}, Y_{1i}, T_{2i}, Y_{2i}, C_i, \alpha, \mu, \Sigma)$$

$$= \frac{\exp(\alpha_0 + \alpha_1 \cdot \ln(T_{1i}+1) + \alpha_2 \cdot \ln Y_{1i} + \alpha_3 \cdot \ln(T_{2i}+1) + \alpha_4 \cdot \ln Y_{2i} + \alpha_5 \cdot C_i)}{1 + \exp(\alpha_0 + \alpha_1 \cdot \ln(T_{1i}+1) + \alpha_2 \cdot \ln Y_{1i} + \alpha_3 \cdot \ln(T_{2i}+1) + \alpha_4 \cdot \ln Y_{2i} + \alpha_5 \cdot C_i)}.$$

To create imputed values for the missing data, as well as to obtain draws from the posterior distribution of the parameters of the model, we use Markov Chain Monte Carlo (MCMC) methods (Geman and Geman, 1984; Gelfand, Hills, Racine–Poon, and Smith, 1990; Gelman and Rubin, 1992; Tanner, 1993; Geweke, 1997) and in particular the DA algorithm developed by Tanner and Wong (1987). Recent economic applications include

Albert and Chib (1993), Lancaster (1995), Geweke and Keane (1997), and Chamberlain and Hirano (1997). The algorithm is discussed in more detail in Appendix B.

Given the imputed data sets we estimate the quantities of interest, e.g., regression coefficients defined in terms of the complete data sets. The approximate variances of the estimates of the regression coefficients are obtained by adding the average complete data variance and the variance of the estimates over the imputed data sets. See Rubin (1987) for details.

4.6.3 The Results

We estimate eight versions of the model. The versions differ by the missing data model, MCAR, MAR, HW or AN, and by the data set used, complete panel, panel or panel and refreshment sample. Given a specific model for the missing data process we create a number of imputed data sets. The primary interest here is in estimates of the regression coefficients in the regression of the change in log of the number of trips on the change in the log of earnings:

$$\ln(T_{2i}+1) - \ln(T_{1i}+1) = \beta_0 + \beta_1 \cdot (\ln Y_{2i} - \ln Y_{1i}) + \varepsilon_i.$$

We focus on the estimate of β_1, the income elasticity of the number of trips.

For comparison the regression coefficient on the first period logarithm of earnings in the regression of the number of trips on the first period logarithm of number of trips using the complete data subsample with $A_i = 2$, $W_i = 1$, using the MCAR model to impute the actual level of earnings, is 0.63 with a standard error of 0.13. The estimates for the MCAR, MAR, and HW model are close together around 0.13, irrespective of the

Table 4.6: Estimates of Income Elasticity of Travel

Model	Complete Panel		Panel		Panel and Refreshment	
	est.	s.e.	est.	s.e.	est.	s.e.
MCAR	0.130	(0.064)	0.139	(0.073)	0.136	(0.073)
MAR			0.135	(0.073)	0.139	(0.073)
HW			0.120	(0.070)	0.134	(0.066)
AN					0.217	(0.145)

data set used. The estimate for the AN model is somewhat larger, with a considerably higher standard error. Overall it appears that the estimates of the elasticity are relatively insensitive to the model for the attrition process. Another important result is that, as in the binary variable example in Section 4.4.3, adding the refreshment sample reduces the difference between the MAR and HW models considerably.

To further interpret the difference between in particular the AN results and the results based on the more restrictive models Table 4.7 presents the posterior means and standard deviations for the parameters of the probability of the willingness to respond. The poste-

Table 4.7: Posterior Means and Standard Deviations of Parameters of Willingness to Respond

model	data	$\ln(T_{i1}+1)$		$\ln(Y_{i1})$		$\ln(T_{i2}+1)$		$\ln(Y_{i2})$		C_i	
		mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
MAR	panel	0.70	(0.08)	0.03	(0.11)	0	-	0	-	-0.39	(0.14)
MAR	all	0.70	(0.08)	0.04	(0.11)	0	-	0	-	-0.39	(0.14)
HW	panel	0	-	0	-	1.30	(0.16)	-0.15	(0.21)	-0.31	(0.16)
HW	all	0	-	0	-	0.85	(0.14)	0.06	(0.18)	-0.38	(0.15)
AN	all	3.21	(0.35)	0.19	(0.62)	-3.49	(0.46)	0.19	(0.83)	-0.50	(0.19)

rior standard deviations for the parameters under the AN model are approximately five times higher than under the restricted models. This is partly due to the high correlation between first and second period variables. When we free up the coefficients on both first and second period variables this leads to considerable uncertainty in their estimates. At the same time, however, the posterior distributions suggest that both first and second period number of trips are important in determining the probability of attrition. This implies that both the MAR and HW models are inadequate in describing the attrition process, although they appear to do reasonably well in recovering the elasticity of interest. Note also that the estimates for the MAR model are the same whether based on the panel alone or on the panel and refreshment sample. Teh reason is that under the MAR restrictions the refreshment sample is not informative about the attrition process because it only contributes about righthand side variables in the logistic regression. In contrast, because the HW model relies on the relation between first and second period variables to transform the directly estimable correlation between first period variables and attrition into an estimate of the correlation between second period variables and the attrition indicator, the presence of the refreshment sample affects estimates of the parameters of the attrition process.

4.7 Conclusion

Panel data sets can provide a much richer set of variables than cross-sections, but they often are subject to more severe missing data problems. Adding a sample consisting of new units randomly drawn from the original sample as replacements for units who have

dropped out of the panel, a so-called refreshment sample, can be helpful in mitigating the effects of attrition in two ways. First, refreshment samples allow for estimation of richer models, potentially resolving differences between selection models common in the statistical literature and those popular in the econometric literature. Second they can make estimation of conventional models more robust and precise. In this paper we have developed a family of models to incorporate the presence of refreshment samples. We demonstrate in an application to a Dutch data set on travel behavior that such models are feasible and can lead to somewhat different results than models assuming the missing data process is ignorable, or conventional econometric models for panel data with attrition.

Appendix A: Proofs

Proof of Theorem 1

We can rewrite equations (4.7)–(4.10) as

$$\alpha_0 = g^{-1}\Big((1-q_{01})r_{01}/((1-q_{01})r_{01}+(1-\hat{q}_{00})r_{00})\Big), \tag{4.13}$$

$$\alpha_0+\alpha_1 = g^{-1}\Big((1-q_{11})r_{11}/((1-q_{11})r_{11}+(1-\hat{q}_{10})r_{10})\Big), \tag{4.14}$$

$$\alpha_0+\alpha_2 = g^{-1}\Big(q_{01}r_{01}/(q_{01}r_{01}+\hat{q}_{00}r_{00})\Big), \tag{4.15}$$

$$\alpha_0+\alpha_1+\alpha_2 = g^{-1}\Big(q_{11}r_{11}/(q_{11}r_{11}+\hat{q}_{10}r_{10})\Big). \tag{4.16}$$

Eliminating α_0, α_1 and α_2 leaves the restriction $h(\hat{q}_{00},\hat{q}_{10})=0$ where

$$\begin{aligned} h(q_{10},q_{00}) = {} & g^{-1}\left(\frac{q_{11}r_{11}}{q_{11}r_{11}+q_{10}r_{10}}\right)+g^{-1}\left(\frac{(1-q_{01})r_{01}}{(1-q_{01})r_{01}+(1-q_{00})r_{00}}\right) \\ & -g^{-1}\left(\frac{q_{01}r_{01}}{q_{01}r_{01}+q_{00}r_{00}}\right)-g^{-1}\left(\frac{(1-q_{11})r_{11}}{(1-q_{11})r_{11}+(1-q_{10})r_{10}}\right). \end{aligned} \tag{4.17}$$

Because of continuity of $h(\cdot,\cdot)$, and because $h(\cdot,\cdot)$ is increasing in q_{00} and decreasing in q_{10}, this restriction defines an implicit function $\bar{q}_{10}(q_{00})$ with the following properties:

$$\frac{\partial \bar{q}_{10}}{\partial a}(a)>0, \quad \lim_{a\downarrow 0}\bar{q}_{10}(a)=0, \quad \text{and} \quad \lim_{a\uparrow 0}\bar{q}_{10}(a)=1.$$

Now consider the restriction (4.11). It defines a function

$$\tilde{q}_{10}(a) = \frac{q_{00}r_{00}+q_{10}r_{10}-ar_{00}}{r_{10}},$$

with the properties

$$\frac{\partial \tilde{q}_{10}}{\partial a}(a)<0, \quad \tilde{q}_{10}(0)>0, \quad \text{and} \quad \tilde{q}_{10}(1)<1.$$

Hence there is a unique value $\hat{q}_{00}$) solving

$$\bar{q}_{10}(a) = \tilde{q}_{10}(a),$$

and $\hat{q}_{10} = \bar{q}_{10}(\hat{q}_{00})$. ■

The proof for Theorem 2 consists of three parts. First we prove two lemmas. The first lemma states that the solution $\hat{f}(z_1, z_2, x)$ can be characterized as the solution to a different problem. The second lemma states that the solution to the second problem is unique. Then we put these two results together. The discrete points of the support of the joint distribution of (Z_{i1}, Z_{i2}) are organized in a sequence $\{(z_{1k}, z_{2k})\}_{k=1}^{K}$.

Lemma 1 *Let $f(z_1, z_2)$ be the joint probability function of (Z_{i1}, Z_{i2}), let $p(z_1, z_2)$ be the conditional probability that $W_i = 1$ given (Z_{i1}, Z_{i2}), and let $0 < p(z_1, z_2) < 1$ for all (z_1, z_2) in the support of (Z_{i1}, Z_{i2}). Let $g(\cdot)$ be a continuous, increasing function with $\lim_{a\to-\infty} g(a) = 0$, and $\lim_{a\to\infty} g(a) = 1$. Let $\{(z_{1k}, z_{2k})\}_{k=1}^{K}$ be the support of (Z_{i1}, Z_{i2}), and let*

$$q = Pr(W_i = 1) = \sum_{k=1}^{K} f(z_{1k}, z_{2k}) \cdot p(z_{1k}, z_{2k}),$$

$$\pi_{z_1\cdot} = Pr(Z_{i1} = z_1) = \sum_{z_2} f(z_1, z_2),$$

$$\pi_{\cdot z_2} = Pr(Z_{i2} = z_2) = \sum_{z_1} f(z_1, z_2),$$

and

$$\pi_{k|1} = Pr(Z_{i1} = z_{1k}, Z_{i2} = z_{2k}|W_i = 1) = f(z_{1k}, z_{2k}) \cdot p(z_{1k}, z_{2k})/q.$$

Then if there is a set of functions $\hat{f}(z_1, z_2)$, $k_1(z)$ and $k_2(z)$ and a constant k_0 such that for some $(\overline{z}_1, \overline{z}_2)$ in the support of (Z_{i1}, Z_{i2}):

(i), $k_1(\overline{z}_1) = 0,\ k_2(\overline{z}_2) = 0,$

(ii), $\sum_{z_2} \hat{f}(z_1, z_2) = \sum_{z_2} f(z_1, z_2),$

(iii), $\sum_{z_1} \hat{f}(z_1, z_2) = \sum_{z_1} f(z_1, z_2),$

(iv),

$$\hat{f}(z_1, z_2) = f(z_1, z_2) \cdot \frac{p(z_1, z_2)}{g(k_0 + k_1(z_1) + k_2(z_2))},$$

then

$$p_k = \hat{f}(z_{1k}, z_{2k})$$

satisfies the first order conditions for a solution to the constrained maximization problem

$$\max_{p_1,\ldots,p_K | \pi_{k|1} \cdot q < p_k \leq 1} \sum_{k=1}^{K} \pi_{k|1} \cdot h(p_k/\pi_{k|1}), \qquad \text{subject to } \sum_{k=1}^{K} p_k \cdot 1\{z_{1k} = z_1\} - \pi_{z_1 \cdot} = 0, \tag{4.18}$$

$$\sum_{k=1}^{K} p_k \cdot 1\{z_{2k} = z_2\} - \pi_{\cdot z_2} = 0, \text{ for all } z_1, z_2, \quad \text{and } \sum_{k=1}^{K} p_k = 1,$$

where,

$$h(a) = \begin{cases} -\int_a^{2q} g^{-1}(q/s)ds & q < a < 2q \\ \int_{2q}^{a} g^{-1}(q/s)ds & 2q \leq a, \end{cases}$$

and $h(a)$ *not defined for* $a \leq q$.

Proof of Lemma 1:

The argument consists of showing that $p_k = \hat{f}(z_{1k}, z_{2k})$ solves the first order conditions for a solution to the maximization program, with the following Lagrange multipliers. For the restriction

$$\sum_k p_k \cdot 1\{z_{1k} = z_1\} - \pi_{z_1 \cdot},$$

the Lagrange multiplier is λ_{z_1}. The number of restrictions is equal to the number of points of support of Z_{i1}, K_1, minus 1. The omitted value is denoted by $\overline{z}_1$. For the restriction

$$\sum_k p_k \cdot 1\{z_{2k} = z_2\} - \pi_{\cdot z_2},$$

the Lagrange multiplier is γ_{z_2}. The number of restrictions equals the number of points of support of Z_{i2}, K_2, minus 1, with the omitted value denoted by $\overline{z}_2$. For the adding up restriction the Lagrange multiplier is δ.

The first order condition for p_k is:

$$h'(p_k/\pi_{k|1}) - \delta - \sum_{z_1} \lambda_{z_1} \cdot 1\{z_{1k} = z_1\} - \sum_{z_2} \gamma_{z_2} \cdot 1\{z_{2k} = z_1\} = 0.$$

By assumption the derivative of $h'(\cdot)$ is invertible, with the inverse equal to $q/g(\cdot)$, so that the solution for p_k is

$$\begin{aligned} p_k &= \pi_{k|1} \cdot (h')^{-1}\Big(\delta + \sum_{z_1} \lambda_{z_1} \cdot 1\{z_{1k} = z_1\} + \sum_{z_2} \gamma_{z_2} \cdot 1\{z_{2k} = z_2\}\Big) \\ &= \pi_{k|1} \cdot \frac{q}{g\Big(\delta + \sum_{z_1} \lambda_{z_1} \cdot 1\{z_{1k} = z_1\} + \sum_{z_2} \gamma_{z_2} \cdot 1\{z_{2k} = z_2\}\Big)} \\ &= \pi_{k|1} \cdot \frac{q}{g(k_0 + k_1(z_{1k}) + k_2(z_{2k}))}. \end{aligned}$$

with k_1, k_2, and k_0 defined as $k_1(z_{1k}) = \sum_{z_1} \lambda_{z_1} \cdot 1\{z_{1k} = z_1\}$, $k_2(z_{2k}) = \sum_{z_2} \gamma_{z_2} \cdot 1\{z_{2k} = z_2\}$, and $k_0 = \delta$.

If we substitute this solution in the restrictions, we obtain a system of $K_1 + K_2 - 1$ equations in the same number of unknowns: k_0 and the values of the functions k_1, k_2 on the support of Z_{i1}, Z_{i2}, except $\overline{z}_1$, $\overline{z}_2$

$$\pi_{z_1 \cdot} - \sum_{k=1}^{K} \frac{q\pi_{k|1}}{g(k_0 + k_1(z_1) + k_2(z_{2k}))} 1\{z_{1k} = z_1\} = 0$$

$$\pi_{\cdot z_2} - \sum_{k=1}^{K} \frac{q\pi_{k|1}}{g(k_0 + k_1(z_{1k}) + k_2(z_2))} 1\{z_{2k} = z_2\} = 0$$

$$1 - \sum_{k=1}^{K} \frac{q\pi_{k|1}}{g(k_0 + k_1(z_{1k}) + k_2(z_{2k}))} = 0$$

Because $0 < p(z_1, z_2) < 1$, the variables in this system are bounded. If we let $k_1(z_1)$, $k_2(z_2)$, k_0 increase, then the left-hand sides of the equations become nonnegative for a finite value of these variables, and irrespective of the values taken by the other (bounded) variables. If we let these variables decrease, then the left-hand sides become nonpositive, again for a finite value and irrespective of the values taken by the other (bounded) variables. Because the left-hand sides are continuous functions and there exists a bounded set such that these functions are nonnegative and nonpositive on the boundaries, we can invoke a fixed point theorem to show that the system of equations indeed has a (bounded) solution. Substitution of this solution in the equation for p_k gives the desired result. ■

Lemma 2 *Let $\{z_{1k}, z_{2k}\}_{k=1}^{K}$ be the support of a pair of discrete random variables with probability $0 < \pi_k < 1$ for $k = 1, \ldots, K$, and let $p(z_1, z_2)$ be a function such that $0 < p(z_{1k}, z_{2k}) < 1$ for all $k = 1, \ldots, K$ with $q = \sum_k p(z_{1k}, z_{2k}) \cdot \pi_k$. Let $\pi_{z_1 \cdot} = \sum_k \pi_k \cdot 1\{z_{1k} = z_1\}$, and $\pi_{\cdot z_2} = \sum_k \pi_k \cdot 1\{z_{2k} = z_2\}$. Furthermore, let $\pi_{k|1} = \pi_k \cdot p(z_{1k}, z_{2k})/q$. Finally let $h(\cdot)$ be any function defined on (q, ∞) such that the inverse of the derivative of $h(\cdot)$ is equal to $q/g(a)$, for some increasing and continuous function $g(a)$ with $\lim_{a \to -\infty} g(a) = 0$, and $\lim_{a \to \infty} g(a) = 1$.*

Then the optimization program

$$\max_{p_1, \ldots, p_K | q\pi_{k|1}} < p_k \leq 1 \sum_{i=1}^{K} \pi_{k|1} \cdot h(p_k/\pi_{k|1}), \qquad \text{subject to } \sum_j p_j \cdot 1\{\kappa_j = x\} - \pi_{z_1 \cdot}, \quad (4.19)$$

$$\sum_j p_j \cdot 1\{\gamma_j = y\} - \pi_{\cdot z_2}, \quad \text{and } \sum p_k = 1,$$

for all z_1 in the support of Z_{i1}, and for all z_2 in the support of Z_{i2}, has a unique solution.

Proof of Lemma 2:

First consider the function $h(\cdot)$. Its second derivative is negative. In addition, $\lim_{a \downarrow q} h'(a) = \infty$, and $\lim_{a \to \infty} h'(a) = -\infty$. Hence $h(\cdot)$ is bounded from above by some number $\bar{h}$.

Hence we are maximizing a concave function over a convex set. If the set over which the function is maximized were compact, this would guarantee the existence of a unique solution. However, the restriction $p_k > q\pi_{k|1}$ implies the set is not compact. There are two possibilities. If the limit $a \downarrow q h(a) = \underline{h \to \infty}$, we can extend the definition of $h(\cdot)$ and maximize the function over a compact set. If the limit $a \downarrow q$ diverges, we can restrict the set of p_k to those such that the objective function is greater than $c - \varepsilon$, where c is the value at $p_k = \pi_{k|1}$. This set will then be compact and the corresponding solution will be unique and in the interior. ■

Proof of Theorem 2:

For a given value for x, we can apply Lemma's 1 and 2 to prove the existence and uniqueness of $\hat{f}(z_1, z_2, x)$, $k_0(x)$, $k_1(z_1, x)$ and $k_2(z_2, x)$. ■

Appendix B: MCMC Algorithms

Although the specific details of the MCMC simulations discussed below depend on the particular models used for the conditional distribution of $(T_{1i}, Y_{1i}, T_{2i}, Y_{2i})$ given C_i, and

the conditional probability $Pr(W_i = 1|T_{1i}, Y_{1i}, T_{2i}, Y_{2i}, C_i)$, for most conventional models MCMC methods will be easy to implement. In our implementation, the chains consist of six steps, the first four dealing with imputing the missing data given current parameter values, and the last two drawing from the posterior distributions of the parameters given imputed and observed data.

First, given initial values of the parameters, we impute the missing values for T_{2i} and Y_{2i} for units with $A_i = 2$ and $W_i = 0$, conditioning on T_{1i}, Y_{1i}, C_i and $W_i = 0$. We implement this by drawing from the conditional distribution of $\ln(T_{2i}+1)$ and $\ln Y_{2i}$ given T_{1i}, Y_{1i} and C_i, which is bivariate normal, and rejecting draws with probability equal to $Pr(W_i = 1|T_{1i}, Y_{1i}, T_{2i}^{\text{imputed}}, Y_{2i}^{\text{imputed}}, C_i)$.

Second, we imputed values for T_{1i} and Y_{1i} for units with $A_i = 1$. The conditional distribution of $\ln(T_{1i} + 1)$ and $\ln Y_{1i}$ given T_{2i}, Y_{2i}, C_i and the parameters is bivariate normal and straightforward to draw from.

In third step we draw, for units with $A_i = 1$, given parameters, T_{2i}, Y_{2i}, C_i and given the previously imputed values T_{1i}^{imputed} and Y_{1i}^{imputed}, the attrition indicator W_i using a binomial distribution.

In the fourth step we impute earnings Y_{1i} for units with $D_{i1} = 1$ and Y_{2i} for units with $D_{i2} = 1$ given the observed indicators for the four ranges, the observed or imputed willingness to respond W_i, the other variables and the parameters. Two methods were used to impute the continuous earnings variable. In one method unrestricted normally distributed random variables are drawn without conditioning on the observed range or on W_i. These draws are then rejected if they are outside the appropriate range, and also

rejected with a probability depending on the value of W_i. This simple method can be computationally very burdensome and lead to many rejected draws. A second method was therefore used if the first one did not lead to an acceptable draw with 30 attempts. In the second method a piecewise linear approximation to the normal distribution inside the appropriate range was used with the draws rejected at an appropriate rate to generate draws from the appropriate truncated normal distribution, whose draws were then rejected with some probability depending on the value of W_i. See Gelman, Carlin, Stern, and Rubin (1995), who call this method "trapezoidal approximation followed by rejection sampling", Ripley (1987), and Hammersley and Handscomb (1964).

In the fifth step we draw from the posterior distribution of μ and Σ given observed and imputed data. Given standard prior distributions these posterior distributions are straightforward to draw from. We use an improper, flat, prior distribution on all elements of $\mu = (\mu_0', \mu_1')'$ and an improper prior distribution on Σ proportional to $|\Sigma|^{-2}$.

Finally, in the sixth step, we draw from the posterior distribution of α given observed and imputed data, using the Metropolis–Hastings algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller and Teller, 1953; Hastings, 1970). We assume prior independence of the components of α, using normal prior distributions centered around zero with standard deviations equal to the square root of the average square of the corresponding variables. This leads to a prior standard deviation for α_1 of 1, for α_2 of 4, for α_3 of 3, for α_4 of 4, for α_5 of 3, and for α_6 of 0.3.

We first ran one chain of 100,000 iterations, and used this to draw overdispersed starting values for a number of independent chains, also of length 100,000. More precisely,

given the first sequence of values for α, $\{\alpha_{(k)}\}_{k=1}^{100,000}$, we choose a number of draws, $\alpha_{(k)}$, and used for each $\alpha_{(k)}$ the starting value $\alpha_{(k)} + (\alpha_{(k)} - \overline{\alpha})$. The values $\alpha_{(k)}$ chosen corresponded to the ten draws from the original chain with the highest or lowest values for any of the elements of $\alpha_{(k)}$ in the first sequence of 100,000 iterations. We then used the Gelman–Rubin (1992) criteria to monitor convergence of the chains. The first long chain used zero starting values for the slope coefficients because maximum likelihood estimates are difficult to obtain.

Fig 1: MAR (*), HW (+), and Restrictions Implied by Second Period Marginal Distribution of Trips2

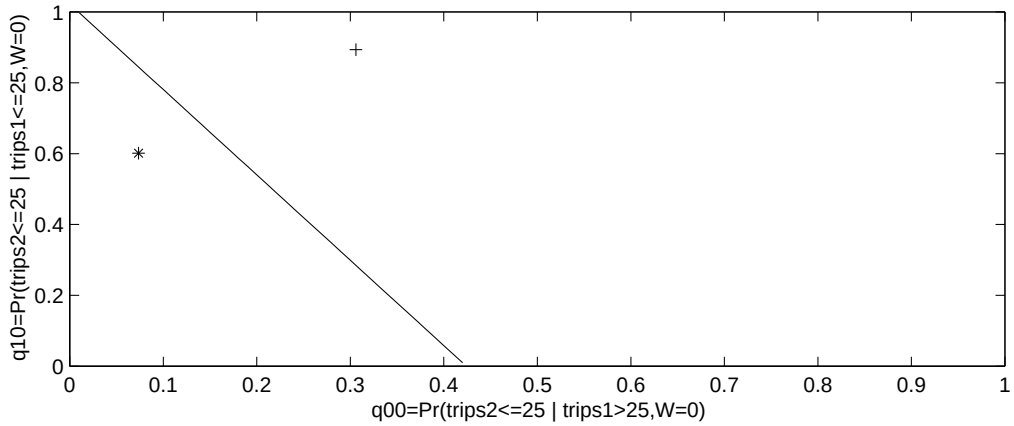

Fig 2: Logistic and Probit Model with Additive Nonignorably Missing Data Process

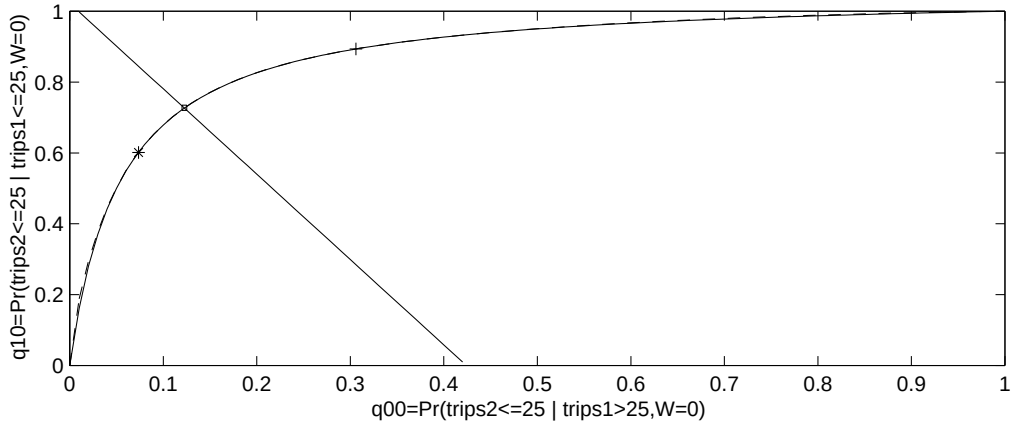

Fig 3: Logistic and Probit Model with Additive Nonignorably Missing Data Process