Econometrica, Vol. 69, No. 6 (November, 2001), 1645–1659

COMBINING PANEL DATA SETS WITH ATTRITION AND REFRESHMENT SAMPLES

BY KEISUKE HIRANO, GUIDO W. IMBENS, GEERT RIDDER, AND DONALD B. RUBIN[1]

1. INTRODUCTION

IN ECONOMICS AND OTHER social sciences, researchers often wish to consider statistical models that allow for more complex relationships than can be inferred using only cross-sectional data. Panel, i.e., 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 (e.g., Chamberlain (1984), Hsiao (1986), Baltagi (1995), Arellano and Honoré (forthcoming)). Missing data problems, however, can be more severe in panels, because even those units that respond in initial waves of the panel may drop out of the sample in subsequent waves (e.g., Hausman and Wise (1979), Robins and West (1986), Ridder (1990), Verbeek and Nijman (1992), Abowd, Crépon, Kramarz, and Trognon (1995), Fitzgerald, Gottschalk, and Moffitt (1998), and Vella (1998)). Sometimes, in the hope of mitigating the effects of such attrition, panel data sets are augmented by replacing the units that have dropped out with new units randomly sampled from the original population. Following Ridder (1992), who used such replacement units to test alternative models for attrition, we call such additional samples *refreshment samples*.

Here we explore the benefits of refreshment samples for inference in the presence of attrition. Two general approaches are often used to deal with attrition in panel data sets when refreshment samples are not available. One model, 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 that have missing values. The other model (denoted by HW in the remainder of the paper, given the similarity to a model developed by Hausman and Wise (1979)), allows the probability of attrition to depend on such contemporaneous, but not on lagged, variables. Both sets of models have some theoretical plausibility, but they rely on fundamentally different restrictions on the dependence of the attrition process on the time path of the variables. They can therefore lead to very different inferences, especially regarding the dynamic aspects of the underlying process. However, in many cases panel data alone cannot be used to distinguish between them. For example, in the two-period case without a refreshment sample both models are essentially just-identified (except for some inequality restrictions under the HW model). If the panel data set is augmented with a refreshment sample, however, the data can distinguish between the MAR and HW models. The two models have in that case testable restrictions and it is possible to estimate a more general class of models that nests both MAR and HW as special cases. This class of models, which we label Additive

[1] We thank Joshua Angrist, Gary Chamberlain, Jerry Hausman, and participants in presentations at Harvard-MIT, NYU, UCLA, Aarhus, the International Statistical Institute meetings in Instanbul, students at the University of Wisconsin, three referees and a coeditor for comments. Imbens acknowledges financial support through a research fellowship from the Alfred P. Sloan Foundation and Grant SBR 9511718 from the National Science Foundation.

Non-ignorable (AN), is characterized by an additivity restriction on the attrition process. We show that the AN class of models is nonparametrically just-identified in the presence of refreshment samples.

In the next section we set up the general problem of attrition in a two-period panel data set, discuss traditional restrictions on the attrition process that have been used in economics and statistics, and discuss how refreshment samples can be used to relax some of these restrictions. To provide intuition for the identification results, we study in Section 3 a simplified version of the problem, involving only binary variables. Section 4 contains our main result, a theorem that makes precise the extent to which refreshment samples aid in identification under more general models of attrition. Section 5 concludes by discussing extensions to multi-wave panels.

2. GENERAL MODEL

We consider a two-period panel data set with attrition. Let Z_{it} be a vector containing all time-varying variables for unit i in time t (including both exogenous and endogenous variables), and let X_i be a vector containing all time-invariant variables for unit i.[2] In the first time period we draw a random sample of size N_P from a fixed population; we refer to this as the *panel*. For each unit i in this sample, for $i = 1, \ldots, N_P$, we observe X_i and Z_{i1}. For a subset of size N_{BP} of this sample, we observe in the second period a second variable Z_{i2}; we refer to this subset as the *balanced panel*. The remaining $N_{IP} = N_P - N_{BP}$ units have dropped out of the panel and all of their Z_{i2} are missing; this will be called the *incomplete panel*.

In addition to the panel data set, in the second period we draw a new random sample from the original population, the *refreshment* subsample, of size N_R.[3] For these units we observe Z_{i2} and X_i, but not Z_{i1}.

We assume that all units respond the first time they are approached: if approached in period 1 (as part of the panel), unit i will respond in the first period, and thus Z_{i1} and X_i will be recorded, whereas if first approached in the second period (as part of the refreshment sample), unit i will respond in the second period and Z_{i2} and X_i will be recorded.[4] Not all units, however, respond the second time they are approached. 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 again, so that Z_{i2} will be recorded, whereas $W_i = 0$ implies that unit i, if approached in the second period after already having responded in the first period, will choose not to respond, so that Z_{i2} will not be recorded. The willingness to respond indicator, 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 if unit i is part of the panel, but W_i is missing if unit i is part of the refreshment sample.

We want to recover the joint distribution of (Z_1, Z_2, X), or possibly the conditional distribution of (Z_1, Z_2) given X. This immediately identifies any parameter that can be

[2] Whenever the meaning is clear from the context we drop the subscript i in the remainder.

[3] We could allow for stratified sampling based on X_i in both random samples without any modification of our main results.

[4] In many panel data sets there is also initial nonresponse. For clarity of exposition we focus on attrition issues ignoring such initial nonresponse. In practice one might wish to account for this as well, but our results do not add to the methods already available for dealing with this issue (e.g., Little and Rubin (1987)).

defined as a functional of one of these joint distributions. The conditional distribution of W given (Z_1, Z_2, X) is of concern solely because its properties can affect our ability to recover the distribution of interest. Throughout the paper we assume that the conditional probability of responding in the second period is strictly positive, so that we can write

$$f(Z_1, Z_2, X) = \frac{f(Z_1, Z_2, X|W=1) \cdot \Pr(W=1)}{\Pr(W=1|Z_1, Z_2, X)}.$$

Because we can directly estimate $f(Z_1, Z_2, X|W=1)$ from the balanced panel, and $\Pr(W=1)$ from the full panel, identification of the willingness-to-respond probability $\Pr(W=1|Z_1, Z_2, X)$ implies that the joint distribution of (Z_1, Z_2, X) is identified. Given identification, inference can proceed in a number of distinct ways. One possibility is to weight the complete-panel observations by the inverse of the attrition probabilities (e.g., Hansen, Hurwitz, and Madow (1953), Hellerstein and Imbens (1999)). An alternative is to use the attrition probabilities to (multiply) impute the missing values (e.g., Rubin (1987), Brownstone and Valletta (1996)). A third approach is to jointly estimate the model of interest with the attrition model, either in a parametric or semiparametric setting. Here we focus solely on identification. The working paper version of this paper (Hirano, Imbens, Ridder, and Rubin (1998)) discusses estimation using multiple imputation in this context.

Next let us consider two models that have been used for inference with panel data in the presence of attrition. The first model makes the assumption that Z_2 is missing at random (MAR) in the panel, to yield

$$W \perp Z_2 | Z_1, X \quad \text{(MAR)}, \tag{1}$$

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)). This case is sometimes also referred to as *selection on observables* (e.g., Moffitt, Fitzgerald, and Gottschalk (1999)) because we can write the attrition probability as

$$\Pr(W=1|Z_1, Z_2, X) = \Pr(W=1|Z_1, X),$$

with the probability of attrition only depending on Z_1 and X, which are always observed. The MAR model is just-identified in the absence of further restrictions on the joint distribution of the variables. An application to panel data is Marini, Olsen, and Rubin (1980).

The second model for panel data with attrition we consider is closely related to a model used by Hausman and Wise (1979), and more generally is related to models developed for sample selection in cross-sectional surveys by Heckman (1976, 1979). A generalized version of Hausman and Wise's model allows the probability of attrition in the second period to depend in an arbitrary fashion on the contemporaneous variables Z_2, as well as on X, but assumes that the first period variables do not affect this probability:

$$W \perp Z_1 | Z_2, X \quad \text{(HW)}. \tag{2}$$

Related models have also been referred to as *selection on unobservables* (e.g., Moffitt, Fitzgerald, and Gottschalk (1999)) because attrition partly depends on variables that are not observed when the unit drops out:

$$\Pr(W=1|Z_1, Z_2 X) = \Pr(W=1|Z_2, X),$$

with Z_2 not observed if $W = 0$. The appeal of these models is that they can reflect optimal behavior of the respondent if incentives for responding depend on current or future, unobserved, rather than past, observed, values.

To illustrate the main issues, consider the following simple model for a two-period panel:

$$Y_{it} = \alpha_i + \gamma \cdot t + \beta \cdot X_i \cdot t + \varepsilon_{it},$$

where $X_i \in \{0, 1\}$ is a group indicator. We are interested in β, the difference in time trends for the two groups, allowing for unit-specific intercepts. Assuming ε_{it} is independent of X_i, we can estimate β by regressing the change $Y_{i2} - Y_{i1}$ on X_i:

$$Y_{i2} - Y_{i1} = \gamma + \beta \cdot X_i + \varepsilon_{i2} - \varepsilon_{i1},$$

leading to a standard difference-in-differences estimator (e.g., Blundell and MaCurdy (2000)):

$$\hat{\beta} = \left(\overline{Y_{i2|X_i=1}} - \overline{Y_{i1|X_i=1}}\right) - \left(\overline{Y_{i2|X_i=0}} - \overline{Y_{i1|X_i=0}}\right).$$

Now suppose there is attrition in the second period. A parametric version of the MAR model could specify the attrition indicator as

$$W_i = 1\{\pi_0 + \pi_1 \cdot X_i + \pi_2 \cdot Y_{i1} + \eta_i > 0\}, \tag{3}$$

with a standard normal distribution for η_i, independent of (X_i, Y_{i1}, Y_{i2}). A corresponding parametric version of the HW model could specify

$$W_i = 1\{\pi_0 + \pi_1 \cdot X_i + \pi_3 \cdot Y_{i2} + \eta_i > 0\}, \tag{4}$$

with η_i again standard normal and independent of (X_i, Y_{i1}, Y_{i2}). Although these models can be generalized somewhat (for example by allowing more general functions of Y_{i1} and X_i, or of Y_{i2} and X_i, and non-normal disturbances), with only panel data available one cannot introduce dependence on Y_{i2} in the MAR model, or dependence on Y_{i1} in the HW model, without relying heavily on functional form and distributional assumptions. In other words, some exclusion restriction is needed for identification, and the MAR and HW models differ in the exclusion restrictions they impose.

The choice between the two different models without a refreshment sample therefore relies on theoretical considerations. Such considerations can often support either model. Suppose that the disutility of responding again in the second period depends on the value of the second period variables, Z_{i2}. For example, the effort in responding may be related to some of the current responses. If each individual realizes the burden to respond, this would be supportive of the HW model where the contemporaneous values affect the probability of responding. On the other hand, the decision whether to respond may be related to past experiences—if in the first period the effort in responding was high, an individual may be less inclined to respond in the second period. This may be particularly relevant if the decision to respond is made before the respondent knows the value of the future variables. If expectations about future variables are formed using past experience, then some form of the MAR model, in which attrition depends on past values, could be appropriate. In practice one may therefore not wish to rule out either the HW or MAR models a priori.

With a refreshment sample one can use the data to help distinguish between the two models, as well as estimate more general models. A parametric example of the class of models we introduce has the form

$$W_i = 1\{\pi_0 + \pi_1 \cdot X_i + \pi_2 \cdot Y_{i1} + \pi_3 \cdot Y_{i2} + \eta_i > 0\},$$

allowing dependence of the attrition decision on both Y_{i1} and Y_{i2}, and thus nesting both the MAR and HW models, as well as leading to testable restrictions on those models. As in the MAR and HW models, we can generalize this model to allow the index to depend on general functions of Y_{i1} and Y_{i2}, as well as interactions with X_i, but an important limitation is that we cannot allow for interactions between Y_{i1} and Y_{i2}, demonstrating that a refreshment sample does not lead to full nonparametric identification of the attrition process. In the next section we make the previous claim precise for a simple case with all variables binary. In Section 4 we present the general results and formalize the class of models identified.

3. BINARY CASE

To provide intuition for the main identification result, we begin by examining the simpler case where both Z_1 and Z_2 are binary scalars, without time-invariant covariates X. In this case the joint distribution of (W, Z_1, Z_2) is fully described by the eight probabilities $\Pr(W = w, Z_1 = z_1, Z_2 = z_2)$ for $w, z_1, z_2, \in \{0, 1\}$. We parameterize these eight probabilities in terms of the joint distribution of Z_1 and W and the conditional distribution of Z_2 given Z_1 and W:

$$q_{zw} = \Pr(Z_2 = 1 | Z_1 = z, W = w),$$

and

$$r_{zw} = \Pr(Z_1 = z, W = w).$$

This parameterization is particularly convenient for separating the identifying information from the panel and the refreshment sample. In large samples we can learn the value of r_{zw} for all $z, w \in \{0, 1\}$ from the panel alone, because the panel is a random sample from the population, and for this subsample we always observe Z_1 and W. Similarly we can learn the values of q_{z1}, for $z \in \{0, 1\}$, from the panel, because the balanced panel with $W = 1$ and $Z_1 = z$ is a random sample from the subpopulation with $W = 1$ and $Z_1 = z$, and for this subsample we always observe Z_2. The panel alone therefore identifies six of the eight probabilities that describe the joint distribution of (W, Z_1, Z_2). The panel data, however, contain no direct information concerning the remaining two probabilities q_{00} and q_{10}, because in the panel we never observe Z_2 if $W = 0$.

The refreshment sample allows us to deduce in large samples the marginal distribution of Z_2, captured by a scalar parameter $\Pr(Z_2 = 1)$. Since

$$\Pr(Z_2 = 1) = \sum_{x,\, w} q_{zw} \cdot r_{zw},$$

knowledge of the marginal distribution of Z_2 corresponds to a single linear restriction on the two remaining parameters q_{10} and q_{00}, given the parameters $q_{01}, q_{11}, r_{00}, r_{01}, r_{10}$, and r_{11} that are identified from the panel alone. 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 without an additional assumption—the refreshment

sample implies only a single restriction on the two remaining parameters and does not achieve full nonparametric identification.

First we discuss the testable restrictions in the MAR and HW models given the presence of refreshment samples. Neither the MAR nor the HW model requires the refreshment sample for estimation of q_{00} and q_{10}. The independence assumptions (1) and (2) each imply two restrictions on the eight parameters r_{zw} and q_{zw} that are sufficient for identification of q_{00} and q_{10} from the panel alone. Specifically, the MAR restrictions imply

$$q_{00} = q_{01} \quad \text{and} \quad q_{10} = q_{11}. \tag{5}$$

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

$$q_{00} = \frac{r_{10} \cdot r_{01} \cdot (1 - q_{01}) - r_{11} \cdot r_{00} \cdot (1 - q_{11})}{r_{00} \cdot r_{11} \cdot q_{11} \cdot (1 - q_{01})/q_{01} - r_{11} \cdot r_{00} \cdot (1 - q_{11})} \tag{6}$$

and

$$q_{10} = \frac{q_{00} \cdot r_{00} \cdot q_{11} \cdot r_{11}}{q_{01} \cdot r_{01} \cdot r_{10}}. \tag{7}$$

Under either of these two models, we can therefore estimate all eight parameters solely from the panel, thereby leading to an estimate for the marginal distribution of Z_2. This indirect estimate of the marginal distribution can be compared to the direct estimate based on the refreshment sample to test the specific attrition model.

To illustrate these issues we use a subset of the Dutch Transportation Panel, a survey on transportation usage by Dutch households that incorporated refreshment samples in its design (see Meurs and Ridder (1992) and Ridder (1992) for more details on this data set). 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 I summarizes the sample information for this variable and Table II presents estimates of the directly estimable parameters r_{zw} and q_{z1}. Table III presents estimates of the remaining parameters q_{z0} and the implied marginal probability $\Pr(Z_2 = 1)$ under the MAR and HW models.

TABLE I

SUMMARY STATISTICS FOR DUTCH TRANSPORTATION PANEL: Z_{it} INDICATES WHETHER THE NUMBER OF TRIPS IN THE PERIOD t IS LESS THAN OR EQUAL TO 25, AND W_i INDICATES WILLINGNESS TO RESPOND IN THE SECOND PERIOD ($N = 2420$)

Subsample	Z_{i1}	Z_{i2}	W_i	No of obs.
Balanced Panel	0	0	1	832
($N_{BP} = 1031$)	0	1	1	66
	1	0	1	53
	1	1	1	80
Incomplete Panel	0	—	0	518
($N_{IP} = 733$)	1	—	0	215
Refreshment Sample	—	0	—	520
($N_R = 656$)	—	1	—	136

TABLE II

ESTIMATES AND STANDARD ERRORS FOR DUTCH TRANSPORTATION PANEL EXAMPLE: DIRECTLY ESTIMABLE PARAMETERS

	$W=0$	$W=1$	q_{z1}
$Z_1=0$	$\hat{r}_{00}=0.294$ (0.011)	$\hat{r}_{01}=0.509$ (0.012)	$\hat{q}_{01}=0.074$ (0.009)
$Z_1=1$	$\hat{r}_{10}=0.122$ (0.008)	$\hat{r}_{11}=0.075$ (0.006)	$\hat{q}_{11}=0.602$ (0.043)

First we can test whether the hypothesis that the willingness to respond indicator is independent of (Z_1, Z_2), that is, whether the claim that the missing data are *Missing Completely At Random* (MCAR, Little and Rubin (1987)), is consistent with the data by testing independence of Z_1 and W in the panel. A likelihood ratio test gives 72.2, with a Chi-squared distribution with one degree of freedom under the null hypothesis, suggesting the data clearly reject the hypothesis that the missing data process is MCAR.

Next, consider estimation of the MAR and HW models using only the panel data. For the six directly estimable parameters, r_{zw} and q_{z1}, the MAR and HW models agree exactly. Assuming MAR, the panel subsample leads to the estimates

$$\hat{q}_{00}=\hat{q}_{01}=0.074$$

and

$$\hat{q}_{10}=\hat{q}_{11}=0.602,$$

implying that the marginal probability of the number of trips in the second period being less than or equal to twenty-five is $\sum_{z,w}\hat{r}_{zw}\cdot\hat{q}_{zw}=0.178$, with a standard error of 0.020 (see Table III). This result can be compared to the marginal probability of the number of trips in the second period being less than or equal to twenty-five implied by the refreshment sample, which is $136/(136+520)=0.207$, with a standard error of 0.016. The likelihood ratio test statistic for the MAR null hypothesis is 2.2, with a $\chi^2(1)$ distribution under the null hypothesis, so that this difference is not statistically significantly different from zero at conventional levels.

TABLE III

ESTIMATES AND STANDARD ERRORS FOR DUTCH TRANSPORTATION PANEL EXAMPLE: MODEL-BASED ESTIMATES

MAR	HW	AN
$\hat{q}_{00}=0.074$ (0.009)	$\hat{q}_{00}=0.306$ (0.072)	$\hat{q}_{00}=0.123$ (0.032)
$\hat{q}_{10}=0.602$ (0.043)	$\hat{q}_{10}=0.894$ (0.097)	$\hat{q}_{10}=0.727$ (0.068)
$\Pr(\widehat{Z}_2=1)=0.178$ (0.020)	$\Pr(\widehat{Z}_2=1)=0.282$ (0.021)	$\Pr(\widehat{Z}_2=1)=0.207$ (0.016)

Under the HW assumptions, the two probabilities q_{00} and q_{10} that cannot be estimated directly from the data are

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

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

The implied estimate of the marginal probability that $Z_2 = 1$ is 0.282 (with a standard error of 0.021), substantially different from the refreshment sample estimate of 0.207. The likelihood ratio test statistic for the HW null hypothesis is 7.8, with a nominal $\chi^2(1)$ distribution, implying a statistically significant difference at the 0.05 level. Thus, in this particular example, the MAR assumption is better supported by the extra information in the refreshment sample than the HW assumption.

The above discussion demonstrates that the MAR and HW models have testable implications if refreshment samples are available, suggesting that more general models may be identified. We now propose a model that generalizes MAR and HW in a way that fully exhausts the additional information provided by the refreshment sample and has no testable implications.[5] To do so it is convenient to characterize the MAR and HW models in a different way than the restrictions (5) for the MAR model and (6)–(7) for the HW model. Note that, with no essential loss of generality[6] given the binary nature of Z_1 and Z_2, the probability of response can be written as

$$\Pr\big(W = 1 \big| Z_1 = z_1, Z_2 = z_2\big) = g\big(\alpha_0 + \alpha_1 \cdot z_1 + \alpha_2 \cdot z_2 + \alpha_3 \cdot z_1 \cdot z_2\big), \tag{8}$$

for some known, strictly increasing $g(a)$ satisfying $\lim_{a\to-\infty} g(a) = 0$, $\lim_{a\to\infty} g(a) = 1$. With Z_1 and Z_2 binary, this specification saturates the model, implying that the choice of $g(\cdot)$ is irrelevant, and the model places no restrictions on the data-generating process. Assuming MAR (HW) in this context amounts to assuming $\alpha_2 = \alpha_3 = 0$ ($\alpha_1 = \alpha_3 = 0$ respectively), and in each case the choice of $g(\cdot)$ is irrelevant. This representation shows that the two restrictions that each model imposes have one restriction in common: $\alpha_3 = 0$.

The class of models we propose has the form

$$\Pr\big(W = 1 \big| Z_1 = z_1, Z_2 = z_2\big) = g\big(\alpha_0 + \alpha_1 \cdot z_1 + \alpha_2 \cdot z_2\big), \tag{9}$$

for unrestricted values of the unknown parameters α_0, α_1, and α_2. This model rules out an interaction term between Z_1 and Z_2, but allows for non-ignorable models by allowing α_2 to differ from zero.[7] To reflect the additivity of the index in the $g(\cdot)$ function in the first and second period variables, we refer to this as the Additive Non-ignorable (AN)

[5] An alternative is to characterize the set of values of the parameters of interest consistent with the panel and refreshment sample, following the bounds approach developed by Manski (1995). This is straightforward for the current example with all variables binary, and is carried out in the working paper version of the current paper (Hirano, Imbens, Ridder, and Rubin (1998)). Obtaining bounds in this way is more complicated in cases with continuous time-varying variables.

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

[7] Rosenbaum and Rubin (1983) and Scharfstein, Rotnitzky, and Robins (1999) use similar models to investigate sensitivity to MAR assumptions without the additional information that would identify these models.

model. Note that both the MAR and HW models are special cases of this model, as the AN model imposes only the restriction $\alpha_3 = 0$ that is common to MAR (which imposes both $\alpha_2 = 0$ and $\alpha_3 = 0$) and HW (which imposes both $\alpha_1 = 0$ and $\alpha_3 = 0$).

An important feature of our approach is that the solutions $\hat{q}_{10}$ and $\hat{q}_{00}$ in principle depend on the choice of the $g(\cdot)$ function. This is not the case for the MAR and HW models where the choice of $g(\cdot)$ is immaterial. Because the restriction in the AN model rules out an interaction, one needs to specify the metric in which the variables enter additively. One interpretation of the choice of the $g(\cdot)$ function is that it represents a choice of distance function. In this view one can interpret the estimate of the joint distribution of (Z_1, Z_2) as that closest to the one estimated by the MAR model, subject to the restriction on the second period marginal distribution implied by the refreshment sample. This interpretation has a close connection to the empirical likelihood literature (Qin and Lawless (1994), Imbens, Spady, and Johnson (1998)).

Now let us return to the binary data example and consider estimation of the joint distribution of (Z_{i1}, Z_{i2}) using data from both panel and refreshment samples. The estimates for r_{zq} and q_{z1} are the same as for the MAR and HW models presented in Table II. The last column in Table III presents the AN estimates for q_{00} and q_{10}. These estimates are consistent with both the panel and the refreshment sample, and reconcile some of the differences between the MAR and HW models. The AN estimates in Table III are based on the logistic model, with $g(a) = \exp(a)/(1+\exp(a))$. Estimates based on the probit model differ only slightly. For q_{00} and q_{10} the logistic-based estimates are 0.12258 and 0.72657, the probit model based estimates are 0.12251 and 0.72673, and the linear probability model based estimates are 0.12217 and 0.72755. There appears to be little sensitivity to the choice of $g(\cdot)$.

4. MAIN RESULTS

In this section we generalize the identification result in Section 3 to allow for multivalued time-dependent and time-invariant variables. The basic intuition from the previous section is that in the presence of refreshment samples we can allow for dependence on both first and second period variables, but that we cannot allow them to interact in the index of the attrition model.

First consider the case without time-invariant variables. The main result is that without restrictions on the joint distribution of (Z_1, Z_2), one can identify attrition models of the form

$$\Pr\big(W_i = 1 \big| Z_1 = z_1, Z_2 = z_2\big) = g\big(k_0 + k_1(z_1) + k_2(z_2)\big),$$

for known $g(\cdot)$, and subject to normalizations on $k_1(\cdot)$ and $k_2(\cdot)$, e.g., $k_1(0) = k_2(0) = 0$. As in the binary case, the MAR model is generalized to allow the attrition probability to depend on the value of the second period variable, but interactions between first and second period variables are still ruled out. Generalizing the binary case, the functions of z_1 and z_2 are allowed to be completely general.

The extension to the case with time-invariant variables allows $k_0, k_1(\cdot)$, and $k_2(\cdot)$ to be arbitrary functions of x:

$$\Pr\big(W_i = 1 \big| Z_1 = z_1, Z_2 = z_2, X = x\big) = g\big(k_0(x) + k_1(z_1, x) + k_2(z_2, x)\big),$$

with the normalization now at every value of x, e.g., $k_1(0, x) = k_2(0, x) = 0$.

THEOREM 1: *Let the conditional distributions of Z_1, Z_2 given $X = x$ and given $X = x$, $W = 1$ have density functions $f(z_1, z_2|x)$ and $f(z_1, z_2|x, W = 1)$, respectively, with respect to the (product) Lebesgue measure. Let $S(x) = \{(z_1, z_2)|f(z_1, z_2|x) > 0\}$ be the support of the population conditional probability density function. Suppose that*:

(A1) *the support of $f(z_1, z_2|x, W = 1)$ coincides with $S(x)$ almost surely with respect to x*;

(A2) *the conditional density functions $f(z_1, z_2|x)$ and $f(z_1, z_2, x|W = 1)$ are square integrable almost surely with respect to x*;

(A3) *g is a differentiable, strictly 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 $k_0(x), k_1(z_1, x)$, and $k_2(z_2, x)$, defined for almost all values of x, such that:

(i) *for some $\bar{z}_1(x), \bar{z}_2(x)$ in $S(x)$, $k_1(\bar{z}_1(x), x) = k_2(\bar{z}_1(x), x) = 0$, for almost all x*;

(ii) *for almost all (z_1, z_2) and almost all x*

$$\int \frac{\Pr(W = 1|x)}{g(k_0(x) + k_1(z_1, x) + k_2(z_2, x))} f(z_1, z_2|x, W = 1)\, dz_2 = f_1(z_1|x),$$

$$\int \frac{\Pr(W = 1|x)}{g(k_0(x) + k_1(z_1, x) + k_2(z_2, x))} f(z_1, z_2|x, W = 1)\, dz_1 = f_2(z_2|x).$$

PROOF: See Appendix.

Given the result in the theorem, the probability of responding is

$$\Pr(W = 1|z_1, z_2, x) = g(k_0(x) + k_1(z_1, x) + k_2(z_2, x)).$$

The implied estimate of the joint distribution of (Z_1, Z_2) given X is then

$$f(z_1, z_2|x) = \frac{f(z_1, z_2|x, W = 1) \cdot \Pr(W = 1|x)}{g(k_0(x) + k_1(z_1, x) + k_2(z_2, x))}. \tag{10}$$

The first equation in (ii) then ensures that the implied first period marginal distribution is consistent with the information in the panel. The second ensures that the second period marginal distribution is consistent with the information in the refreshment sample. For any choice of g there is therefore exactly one AN model that is compatible with the restrictions. Note that (i) is an obvious normalization of the functions due to the inclusion of a constant $k_0(x)$ in the AN model.

In the theorem we assumed that the support of the joint distribution conditional on $W = 1$ coincides with the support $S(x)$ of the unconditional joint distribution.[8] If the support of the observed distribution is strictly smaller than that of the population distribution, then $k_1(z_1, x) = -\infty$ if $(z_1, z_2) \in S(x)$ for some z_2 (and hence z_1 is in the support of $f_1(z_1|x)$, but $f(z_1, z_2|x, W = 1) = 0$ for almost all z_2). The values of z_2 with $k_2(z_2, x) = -\infty$ can be found in the same way. Of course, the joint population distribution cannot be recovered for these values of z_1, z_2 because they are never observed in the panel.

In practice one may not have sufficient data given the dimension of the variables to estimate the attrition process completely nonparametrically (within the class of AN models). In that case one may wish to impose additional smoothness on the process using relatively

[8] Up to a set of Lebesque measure 0.

flexible parameterizations. As in the example in Section 2, one may wish to specify the attrition process as

$$W = 1\{\beta_0' X + \beta_1' Z_1 + \beta_2' Z_2 + \varepsilon > 0\},$$

with ε standard normal, or include additional polynomial terms involving Z_1 and X, or Z_2 and X. The identification result in this paper implies that the identification does not rely on any of these functional form or distributional assumptions. In practice the estimates would be more robust asymptotically than those based on the same model using only data from the panel.

5. CONCLUSION

Panel data sets can provide a much richer amount of information than cross-sections, but they often are subject to more severe missing data problems through attrition. Adding a sample consisting of new units randomly drawn from the original sample to replace 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, it can make estimation of conventional models more robust and precise, and allow for testing of these models. Second, the presence of a refreshment sample allows for estimation of richer models, potentially resolving differences between selection models common in the statistical literature and those popular in the econometric literature. In this paper we propose a class of models to incorporate the information in refreshment samples that naturally extends the most commonly used specifications for the attrition process.

The discussion in this paper has focused on two-period panels. With more than two periods one can have multiple refreshment samples, one for each period from the second onwards. These refreshment samples can themselves be followed over time as panels, possibly with attrition, or they can be pure cross-sections. In both cases the refreshment samples provide additional information regarding the attrition process. The form of the models identified in those cases can be found using the same type of empirical likelihood approach discussed in Section 3 to motivate the AN model, where we estimate the joint distribution as the one closest to the joint distribution for the complete panel subject to the restrictions implied by the refreshment samples.

Dept. of Economics, University of California at Los Angeles, 405 Hilgard Ave., Los Angeles, CA 90095-1477, U.S.A., and Dept. of Economics, University of Miami, P.O. Box 248126, Coral Gables, FL 33124-6550, U.S.A.; khirano@miami.edu

Dept. of Economics, University of California at Los Angeles, 405 Hilgard Ave., Los Angeles, CA 90095-1477, U.S.A., University of California at Berkeley, and NBER; imbens@econ.ucla.edu

Dept. of Economics, Johns Hopkins University, Baltimore, MD, 21218, U.S.A., and Dept. of Economics, University of Southern California, University Park Campus, Los Angeles, CA 90089, U.S.A.; gridder@jhu.edu

and

Dept. of Statistics, Science Center, Harvard University, Cambridge, MA 02138, U.S.A.; rubin@hustat.harvard.edu

Manuscript received September, 1998; final revision received September, 2000.

APPENDIX: PROOF OF THEOREM 1

The proof consists of three parts. First, we prove that the equations in (ii) are equivalent to the first-order conditions of a constrained maximization problem. Next, we prove that the solution to this problem is unique. Finally, the theorem follows directly by the combination of these two results. Let x be an arbitrary value of X such that $S(x)$ is not empty. In the sequel we fix x.

In the constrained maximization problem, we maximize a functional,[9] defined on the vector space $\mathcal{V}$ of square (Lebesgue) integrable functions f: $\Re^2 \to \Re$. A sufficient condition for a p.d.f. to be square integrable is that f be bounded on $\Re^2$. By definition, $\mathcal{V}$ is an L_2 space and hence a Hilbert space.

The constraints are a set of inequality and equality constraints. The equality constraints are defined by (linear) mappings from $\mathcal{V}$ to $\mathcal{V}_1$, the Hilbert space of square integrable density functions (with respect to the Lebesgue measure) on $\Re$, of which $f(z_1|x)$ and $f(z_2|x)$ are elements. Inequalities involving f are defined using the convex cone of non-negative functions defined on $\Re^2$. With the L_2 norm this cone is regular, a fact that will be used in Lemma 2. In the sequel, the dots in e.g. $f(\cdot,\cdot|x, W=1)$ or $f(\cdot, z_2|x)$ indicate that we consider the function and not the value taken by the function at a particular point in its domain.

LEMMA 1: *Let x be such that $S(x)$ is not empty. Consider the constrained maximization problem*

$$\max_{f\in\mathcal{V}} \iint f(z_1, z_2|x, W=1) h\left(\frac{f(z_1, z_2|x)}{f(z_1, z_2|x, W=1)}\right) dz_1\, dz_2, \tag{11}$$

subject to the inequality constraints

$$f(\cdot,\cdot|x) \geq f(\cdot,\cdot|x, W=1)\Pr(W=1|x), \tag{12}$$

and the linear restrictions

$$\int f(\cdot, z_2|x)\, dz_2 = f_1(\cdot|x), \tag{13}$$

$$\int f(z_1, \cdot|x)\, dz_1 = f_2(\cdot|x). \tag{14}$$

The function $h : (q(x), \infty) \to \Re$ is defined by

$$h(a) = \begin{cases} -\int_a^{2q(x)} g^{-1}(q(x)/s)ds, & q(x) < a < 2q(x), \\ \int_{2q(x)}^{a} g^{-1}(q(x)/s)ds, & 2q(x) \leq a, \end{cases} \tag{15}$$

with $q(x) = \Pr(W=1|x)$ and $g(\cdot)$ a differentiable, strictly increasing function with $\lim_{a\to-\infty} g(a) = 0$ and $\lim_{a\to\infty} g(a) = 1$.

Then the first-order conditions for the maximum are equivalent to the system of integral equations

$$\int \frac{\Pr(W=1|x)}{g(k_0(x)+k_1(\cdot, x)+k_2(z_2 x))} f(\cdot, z_2|x, W=1)\, dz_2 = f_1(\cdot|x),$$

$$\int \frac{\Pr(W=1|x)}{g(k_0(x)+k_1(z_1, x)+k_2(\cdot, x))} f(z_1, .|x, W=1)\, dz_1 = f_2(\cdot|x),$$

with $k_0(x) \in \Re, k_1(\cdot, x), k_2(\cdot, x) \in \mathcal{V}_1$, and the normalization $k_1(\bar{z}_1(x), x) = k_2(\bar{z}_1(x), x) = 0$ for some $\bar{z}_1(x), \bar{z}_2(x)$ in $S(x)$.

[9] Luenberger (1969) discusses functional optimization problems.

PROOF: The two equality restrictions can be written as $T_1(f)=0$, $T_2(f)=0$ with 0 the null element in $\mathcal{V}_1$, and

$$T_1(f)=\int f(\cdot,z_2|x)\,dz_2-f_1(\cdot|x), \tag{16}$$

$$T_2(f)=\int f(z_1,\cdot|x)\,dz_1-f_2(\cdot|x). \tag{17}$$

The combined restrictions can be expressed as $T(f)=0$ with T a mapping from $\mathcal{V}$ to $\mathcal{V}_1\times\mathcal{V}_1$. The mappings T_1, T_2 are linear and hence Fréchet differentiable. The Fréchet derivatives of T_1, T_2 are linearly independent for all $f\in\mathcal{V}$, so that all f that satisfy the restrictions are regular points. This linear independence follows from the fact that $T_1'(f)$ is a function of z_1, $T_2'(f)$ of z_2. Linear independence follows if the joint conditional distribution of Z_1, Z_2 given X is nonsingular. Because h is differentiable the maximand is Fréchet differentiable. The necessary condition for a maximum at f is that f is a stationary point of the Lagrangian functional with $(k_1(\cdot,x),k_2(\cdot,x))\in\mathcal{V}_1\times\mathcal{V}_1$, which is the dual of the image space of T:

$$\begin{aligned}L(f)=&\iint f(z_1,z_2|x,W=1)\Pr(W=1|x)\\&\cdot h\left(\frac{f(z_1,z_2|x)}{f(z_1,z_2|x,W=1)\Pr(W=1|x)}\right)dz_1\,dz_2\\&+\int k_1(z_1,x)\left(f_1(z_1|x)-\int f(z_1,z_2|x)\,dz_2\right)dz_1\\&+\int k_2(z_2,x)\left(f_2(z_2|x)-\int f(z_1,z_2|x)\,dz_1\right)dz_2.\end{aligned}$$

The stationary point is found by setting the Fréchet derivative of the Lagrangian with respect to f equal to 0, i.e. the zero element in $\mathcal{V}$. The Fréchet derivatives follow directly from the observations that, e.g.,

$$\begin{aligned}&\int k_1(z_1,x)\left(f_1(z_1|x)-\int f(z_1,z_2|x)\,dz_2\right)dz_1\\&\quad=\int k_1(z_1,x)f_1(z_1|x)\,dz_1-\iint f(z_1,z_2|x)k_1(z_1,x)\,dz_1\,dz_2\end{aligned}$$

is linear in f and hence its Fréchet derivative is equal to $k_1(\cdot,x)$. Hence the stationary point satisfies both

$$h'\left(\frac{f(\cdot,\cdot|x)}{f(\cdot,\cdot|x,W=1)}\right)=k_1(\cdot,x)+k_2(\cdot,x), \tag{18}$$

and the equality constraints. This determines the functions $k_1(\cdot,x)$ and $k_2(\cdot,x)$ up to a normalization (we can add and subtract a constant). To obtain a unique solution we set $k_1(\bar z_1(x))=k_2(\bar z_1(x))=0$ for some $\bar z_1(x), \bar z_2(x)$ in $S(x)$. This determines $k_0(x)$.

Substitution of h yields

$$f(\cdot,\cdot|x)=\frac{f(\cdot,\cdot,x|W=1)\Pr(W=1|x)}{g(k_1(\cdot,x)+k_2(\cdot,x)+k_0(x))}, \tag{19}$$

and substitution of this solution in the equality restrictions gives the desired result.

Next we prove the following Lemma.

LEMMA 2: *The maximization problem in Lemma* 1 *has a unique solution* $f\in\mathcal{V}$.

PROOF: The function h is strictly concave on $(\Pr(W=1|x),\infty)$. Hence, in Lemma 1 we maximize a strictly concave functional on a convex set defined by inequality and equality constraints. In such a maximization problem, the necessary first order conditions are also sufficient for a unique global maximum. Hence, it suffices to show that there is an f that satisfies the first order conditions, and this is equivalent to the existence of a solution to the system of simultaneous integral equations:

$$\int \frac{\Pr(W=1|x)}{g(k_0(x)+k_1(\cdot,x)+k_2(z_2,x))} f(\cdot,z_2|x,W=1)\,dz_2 = f_1(\cdot|x),$$

$$\int \frac{\Pr(W=1|x)}{g(k_0(x)+k_1(z_1,x)+k_2(\cdot,x))} f(z_1,\cdot|x,W=1)\,dz_1 = f_2(\cdot|x).$$

The left-hand side of these equations is monotonically decreasing in $k_0(x)+k_1(\cdot,x)+k_2(\cdot,x)$ and hence in $k_1(\cdot,x)$, and $k_2(\cdot,x)$. Moreover, if $k_0(x)+k_1(\cdot,x)+k_2(\cdot,x)\to\infty$, then the left-hand side is equal to $f_1(\cdot|x,W=1)$ and $f_2(\cdot|x,W=1)$, respectively, and hence not larger than the right-hand side. If $k_0(x)+k_1(\cdot,x)+k_2(\cdot,x)\to-\infty$, the left-hand side will become larger than the right-hand side. Because the cone of nonnegative functions in the Hilbert space of square integral functions is regular, we can use Theorem 8.3.17 of Hutson and Pym (1980) to show that these equations have a solution. This solution is uniformly bounded from below. This implies that f defined by

$$f(z_1,z_2|x) = \frac{f(z_1,z_2|x,W=1)}{g(k_0(x)+k_1(z_1,x)+k_2(z_2,x))} \tag{20}$$

satisfies both the first-order conditions and the inequality constraints.

REFERENCES

ABOWD, J. M., B. CRÉPON, F. KRAMARZ, AND A. TROGNON (1995): " A la Recherche des Moments Perdus: Covariance Models for Unbalanced Panels with Endogenous Death," NBER Technical Working Paper No. 180, Cambridge, Mass.

ARELLANO, M., AND B. HONORÉ (2000): "Panel Data," forthcoming in *Handbook of Econometrics*, Vol. 5, ed. by J. Heckman and E. Leamer. Amsterdam: North Holland.

BALTAGI, B. (1995): *Econometric Analysis of Panel Data*. New York: Wiley.

BLUNDELL, R., AND T. MACURDY (2000): "Labour Supply: A Review of Alternative Approaches," in *Handbook of Labour Economics*, Vol. 3, ed. by O. Ashenfelter and D. Card. Amsterdam: North Holland, pp. 1559–1695.

BROWNSTONE, D., AND R. VALLETTA (1996): "Modeling Earnings Measurement Error: A Multiple Imputation Approach," *Review of Economics and Statistics*, 78, 705–717.

CHAMBERLAIN, G. (1984): "Panel Data," in *Handbook of Econometrics*, Vol I, ed. by Z. Griliches and M. Intriligator. Amsterdam: North Holland.

FITZGERALD, J., P. GOTTSCHALK, AND R. MOFFITT (1998): "An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics," *Journal of Human Resources*, 33, 251–299.

HANSEN, M., W. HURWITZ, AND W. MADOW (1953): *Sample Survey Methods and Theory*, Vol. I and II. New York: Wiley.

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

HECKMAN, J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables, and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 5, 475–492.

——— (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–361.

HIRANO, K., G. IMBENS, G. RIDDER, AND D. RUBIN (1998): "Combining Panel Data Sets with Attrition and Refreshment Samples," National Bureau of Economic Research, Technical Working Paper 230.

HSIAO, C. (1986): *Analysis of Panel Data*. Cambridge: Cambridge University Press.

HUTSON, V., AND J. S. 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, 333–357.

LITTLE, R. J. A., AND D. B. RUBIN (1987): *Statistical Analysis with Missing Data*. New York: Wiley.

LUENBERGER, D. G. (1969): *Optimization by Vector Space Methods*. New York: Wiley.

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

MARINI, M., A. OLSEN, AND D. RUBIN (1980): "Maximum Likelihood Estimation in Panel Studies with Missing Data," *Sociological Methodology*, Ch. 11, 314–357.

MEURS, H., AND G. RIDDER (1992): "Attrition and Response Effects in the Dutch Mobility Panel," Working Paper, University of Amsterdam.

MOFFITT, R., J. FITZGERALD, AND P. GOTTSCHALK (1999): "Sample Selection in Panel Data: The Role of Selection on Observables," *Annales d'Economie et de Statistique*, 55–56, 129–152.

QIN, J., AND J. LAWLESS (1994): "Generalized Estimating Equations," *Annals of Statistics*, 22, 300–325.

RIDDER, G. (1990): "Attrition in Multi-wave Panel Data," in *Panel Data and Labor Market Studies*, ed. by J. Hartog, G. Ridder, and J. Theeuwes. Amsterdam: North Holland, pp. 45–68.

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

ROBINS, P., AND R. WEST (1986): "Sample Attrition and Labor Supply Response in Experimental Panel Data," *Journal of Business and Economic Statsistics*, 4, 329–338.

ROSENBAUM, P., AND D. RUBIN (1983): "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome," *Journal of the Royal Statistical Society, Series B*, 45, 212–218.

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

——— (1987): *Multiple Imputation*. New York: Wiley.

SCHARFSTEIN, D., A. ROTNITZKY, AND J. ROBINS (1999): "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models," *Journal of the American Statistical Association*, 94, 1096–1120.

VERBEEK, M., AND T. NIJMAN (1992): "Testing for Selectivity Bias in Panel Data Models," *International Economic Review*, 33, 681–703.

VELLA, F. (1998): "Estimating Models with Sample Selection Bias," *Journal of Human Resources*, 33, 127–169.