ANNALES D'ÉCONOMIE ET DE STATISTIQUE. – N° 55-56 – 1999

Sample Attrition in Panel Data: The Role of Selection on Observables

Robert MOFFIT, John FITZGERALD, Peter GOTTSCHALK *

ABSTRACT. – The traditional formulation of the attrition problem in econometrics treats it as a special case of the partial-population section bias model in which selection (attrition) is based on model unobservables. This paper considers instead the treatment of attrition as a special case of selection on observables. The analysis compares and contrasts the identification assumptions and estimation procedures for this case with those of the usual case of selection on unobservables. Selection on observables case has rarely been considered in the econometric literature on the problem and hence the framing of the problem in these terms, as presented here, is apparently new. The selection on observables problem is made nontrivial by the assumption that selection occurs on endogenous observables; leading examples are lagged dependent variables from earlier periods in the panel. Among other things, it is shown in the paper that (i) weighted least squares using estimated attrition equations to construct the weights is one method of consistent estimation in this case; (ii) simply conditioning on the observables does not, by itself, generate consistent estimates; and (iii) that the model is closely related to the choice-based sampling model.

Attrition dans les données de panel: le rôle de la sélection à partir des observables

RÉSUMÉ. – L'approche traditionnelle du problème d'attrition en économétrie le traite comme un cas particulier des modèles à biais de sélection d'une population partielle dans lequel la sélection (l'attrition) repose sur des variables non observables. Ce papier considère en revanche le traitement de l'attrition comme un cas spécial de sélection à partir des observables. L'analyse compare et oppose les hypothèses d'identification et les méthodes d'estimation dans ce cas avec celles du cas usuel de sélection sur des variables non observables. La sélection à partir des observables a été rarement considérée dans la littérature économétrique, si bien que la présentation du problème dans les termes choisis ici semble nouvelle. Le problème de la sélection à partir des observables est rendu non trivial par l'hypothèse que la sélection repose sur des variables endogènes : les exemples principaux correspondent à des variables dépendantes retardées issues de périodes précédantes dans le panel. Parmi d'autres résultats, il est montré dans cet article que (i) les moindres carrés pondérés utilisant des équations d'attrition estimées pour construire les poids donnent une estimation convergente, (ii) conditionner simplement sur les observables ne produit pas des estimations convergentes et (iii) le modèle est étroitement lié au modèle dont l'échantillonnage repose sur les valeurs de la variable dépendante.

* R. MOFFITT: Johns Hopkins University, J. FITZGERALD: Bowdoin College, P. GOTTSCHALK: Boston College.

This paper was presented at the Seventh International Conference on Panel Data, Paris, June 19-20, 1997. The authors would like to thank J. WOOLDRIDGE G. RIDDER, and an anonymous referee for comments, and the U.S. National Science Foundation for financial support.

1 Introduction

Panel data are now understood to make capable a much wider range of inference than is possible with single cross-sections. However, with this advantage comes the disadvantage of attrition, which is perhaps the most potentially damaging and frequently-mentioned threat to the value of panel data. Despite the potential seriousness of the problem and the frequency with which it is mentioned in applied work using panels, the topic has received relatively little formal econometric attention. Attrition has been thought of in econometrics as part of the selection bias model and attrition was recognized early on as one application of that model (HECKMAN [1978, 1979]). The traditional selection bias model posits attrition bias to arise from selection on unobservables. The earliest and most well-known paper working within this general framework is that of HAUSMAN and WISE [1979], who studied a two-period model of attrition. Other work since that time has worked within the same framework but has relaxed various of the HAUSMAN-WISE assumptions and has expanded the model in different directions and relaxed some of its assumptions (RIDDER [1990, 1992]; NIJMAN and VERBEEK [1992], VAN DEN BERG *et al.* [1994], see VERBEEK and NIJMAN [1996], for a review).

This paper uses the econometric formulation of the partial-population selection bias model – that is, the selection bias model in which only a part of the population is observed – to explore attrition from a different perspective, that of *selection on observables*. Both the identification assumptions and estimation procedures are quite different for this case and the more traditional case of selection on unobservables. Rather surprisingly, the selection on observables case has rarely been considered in the econometric literature on the problem and hence the development presented here is apparently new. [1] What makes the selection on observables problem nontrivial is that selection is assumed to occur on endogenous observables; that is, on variables which are not independent of the error term in the main equation of interest. In the case of panel data, the leading examples of such variables are lagged dependent variables from earlier periods in the panel. Among other things, it is shown in the paper that (i) weighted least squares using estimated attrition equations to construct the weights is one method of consistent estimation in this case; (ii) simply conditioning on the observables does not, by itself, generate consistent estimates; and (iii) that the model is closely related to the choice-based sampling model.

The first half of the paper explicates the selection on observables model and the second half presents illustrative calculations from the Michigan Panel on Income Dynamics, one of the most well-known household survey panels in the U.S.

[1] On the other hand, it has been extensively considered in the survey sampling and statistics literature, using a different framework, as noted below.

2 Models of Selection on Observables and Unobservables

The types of selection we will discuss are applicable to cross sectional models as well as panel data, even though our specific application will be to the latter. For generality, therefore, our setup will initially be formulated as a cross-section model and will be modified for panels subsequently.

We assume that the object of interest is a conditional population density $f(y|x)$ where y is a scalar dependent variable and x is, for illustration, a scalar independent variable. We will work at the population level and ignore sampling considerations. Define A as an attrition dummy equal to 1 if an observation is missing its value of y because of attrition and 0 if not. We assume for the moment that x is observed for all, as would be the case if it were a time-invariant or lagged variable. We therefore observe, and can estimate, only the density $g(y|x, A=0)$. The problem is how to infer f from g. By necessity this will require restrictions of some kind.

Although there are many restrictions possible (in fact, an infinite number), we will focus only on a set of restrictions which can be imposed directly on the attrition function, which we define as the probability function $Pr(A=0 |y,x,z)$. Here z is an auxiliary variable which is assumed to be observable for all units but distinct from x, and whose role and properties will become clear momentarily. The variable y is partially unobserved in this function because it is not observed if $A=1$.

The key distinction we make is between what we term *selection on observables* and *selection on unobservables*. [2] We say that selection on observables occurs when

$$(1) \qquad Pr(A=0|y,x,z) = Pr(A=0|x,z)$$

We say that selection on unobservables occurs simply when (1) fails to hold; that is, when the attrition function cannot be reduced from $Pr(A=0|y,x,z)$.[3]

These definitions may be more familiar when they are restated within the textbook parametric model. Letting $E(y|x) = \beta_0 + \beta_1 x$ and $Pr(A=0|x,z) = F(-\delta_0 - \delta_1 x - \delta_2 z)$, where F is a proper c.d.f., we can state the model

[2] These terms have not, to our knowledge, been utilized in the literature on partial-population sample selection models (*i.e.*, models where a subset of the population is missing information on y). However, the terms have been used in the treatment-effects literature, most extensively and explicitly by HECKMAN and HOTZ [1989] but also by HECKMAN and ROBB ([1985], p. l90). The model and estimation method here is very different from treatment-effects models.

[3] The statistics literature often describes what we are calling selection on observables as "*missing at random*" or "*ignorable*" selection (*e.g.*, LITTLE and RUBIN [1987]). These terms are potentially misleading because there are nonrandom variables affecting selection (namely, x and z) and because selection is not ignorable in the sense that consistent estimates can be obtained if selection is ignored (to the contrary, as will be shown momentarily, biased and inconsistent parameter estimation can result if selection is "*ignored*").

equivalently with error terms ε and ν (suppressing individual i subscripts) as

$$\begin{aligned} &(2) \qquad & Y &= \beta_0 + \beta_1 x + \varepsilon \qquad &&, y \text{ observed if } A = 0 \\ &(3) \qquad & A^* &= \delta_0 + \delta_1 x + \delta_2 z + \nu && \\ &(4) \qquad & A &= 1 \text{ if } A^* \geqslant 0 && \\ & & &= 0 \text{ if } A^* < 0 && \end{aligned}$$

where ν is the random variable whose c.d.f. is F. In the context of this model, selection on unobservables occurs when

$$(5) \qquad z \perp\!\!\!\perp \varepsilon | x \quad \text{but} \quad \varepsilon \, \widetilde{\perp\!\!\!\perp} \, \nu | x$$

and selection on observables occurs when

$$(6) \qquad \nu \perp\!\!\!\perp \varepsilon | x \quad \text{but} \quad \varepsilon \, \widetilde{\perp\!\!\!\perp} \, z | x$$

where the symbols $\perp\!\!\!\perp$ and $\widetilde{\perp\!\!\!\perp}$ denote "*is independent of*" and "*is not independent of*", respectively. The critical assumption, as shown in (6), is that in the selection on observables case, the variable z is endogenous (examples of z will be given below).

We treat the more familiar case of selection on unobservables first, briefly, before turning to the relatively unfamiliar case of selection on observables.

Selection on Unobservables. Working from the parametric form of the model, the conditional mean of y in the nonattriting sample can be written

$$\begin{aligned} (7) \qquad E(y|x,z,A=0) &= \beta_0 + \beta_1 x + E(\varepsilon | x,z,\nu < -\delta_0 - \delta_1 x - \delta_2 z) \\ &= \beta_0 + \beta_1 x + h(-\delta_0 - \delta_1 x - \delta_2 z) \\ &= \beta_0 + \beta_1 x + h'(F(-\delta_0 - \delta_1 x - \delta_2 z)) \end{aligned}$$

where h and h' are functions with unknown parameters. Moving from the first to the second line of the equation requires that the joint distribution of ε and ν be independent of x and z, so that the conditional expectation depends on x and z only through the index. Moving from the second to the third line simply replaces the index by its probability, which is permissible because they have a one-to-one correspondence.

Early implementations of this model assumed a specific bivariate distribution for ε and ν, leading to specific forms of the expectation function (*e.g.*, the inverse Mills ratio for bivariate normality), while more recent implementations have relaxed some of the distributional assumptions in the model by estimating functions h or h' whose arguments are either the attrition index or the attrition probability, respectively (see MADDALA, [1983], for a textbook treatment of the early approach and POWELL, [1994], pp. 2509-2510, for discussions of the more recent approach). Armed with estimates of the parameters of the attrition index or of the predicted attrition probability, equation (7) becomes a function whose parameters can be consistently estimated.

The major difficulty with the selection on unobservables model is the difficulty of identification. Aside from nonlinearities in the h, h', and F functions, identification of β requires an exclusion restriction, namely, that a z exist satisfying the independence property from ε and for which δ_2 is nonzero.

Such a variable is often loosely termed an "*instrument*", although most estimation methods proposed for eqn (7) do not take a textbook instrumental-variables form. Finding a general, suitable instrument for unobservable selection is much more difficult for the case of nonresponse than in many other applications because there are few variables affecting nonresponse that are credibly excluded from the main equation of interest on *a priori* grounds. Although every case is application-specific, characteristics of the individual agent such as those generally included in x are unlikely to be promising sources of instruments because most such characteristics are related to behavior in general and hence to y. More promising are variables external to the individual and not under his control, such as characteristics of the interviewer or the interviewing or data collection process, or even interview payments. But these can be used as appropriate instruments only if the interviewing organization assigns them randomly to the respondents; if, in contrast, they are assigned on the basis of respondent characteristics, they are no longer useful as identifiers. [4] Thus identifying the selection-on-unobservables model can be difficult, and it is useful to consider other cases.

Selection on Observables. Our main goal is instead to explore the implications of the restrictions in eqn(l) and eqn(6). The assumption in equation(6) that z and ε are not independent implies that the estimation methods outlined above for the selection on unobservables model will generate inconsistent parameter estimates; the "*instrument*" z in that model must be strictly exogenous. In the selection on observables model, on the other hand, no exogenous instrument is available but because selection bias arises from an endogenous observable–z–consistent estimation is possible (assuming, of course, that the basic selection on observables assumption in (1) holds).

The critical variable in the selection on observables case is z, a variable which affects attrition propensities but is presumed also to be related to the density of y conditional on x (*i.e.*, z is endogenous to y), and yet is not in the y equation. Such a variable can exist only if the investigator is interested in a "*structural*" y function which we define as a function of a variable x that plays a causal role in a theoretical sense; other variables (*i.e.*, z) are assumed not to "*belong*" in the function. More generally, this situation can arise if the investigator is interested in (say) the expectation of y conditional on x and simply does not wish to condition on z. For a cross-sectional example, consider the standard Becker-Mincer theory of human capital in which earnings is a function of education and experience. In that theory, occupation and industry are jointly chosen with earnings and hence do not "*belong*" in the earnings equation, and one is not interested in conditioning earnings on occupation and industry. Yet estimation of an earnings equation on a sample that is selected on the basis of occupation and industry – often data on workers is available only on selected occupations and industries – will result in biased and inconsistent estimates. The variable z is thus an auxiliary endogenous variable which is jointly determined with y. [5]

[4] In a study of the Michigan Panel Study of Income Dynamics (Fitzgerald *et al.* [1997]), we found that these interview characteristics were not assigned randomly and were therefore not suitable instruments.

[5] As we will discuss below, in the panel data case, a lagged value of y can play the role of z if it is not in the "*structural*" model and if it is related to attrition.

In the presence of selection on such an endogenous variable, it is easy to show that least squares estimation of (2) on the nonattriting sample will generate inconsistent estimates of β and, more generally, that the estimable density $g(y|x, A = 0)$ will not correspond to the complete-population density $f(y|x)$ because the event $A = 0$ is related to y through z. However, $f(y|x)$ can be obtained indirectly, and this furnishes the basis for consistent estimation. Let $f(y,z|x)$ be the complete-population joint density of y and z and let $g(y,z|x,A = 0)$ be the conditional joint density. Then

$$\begin{aligned} \text{(8)} \qquad g(y,z|x,A=0) &= \frac{g(y,z,A=0|x)}{Pr(A=0|x)} \\ &= \frac{Pr(A=0|y,z,x)\,f(y,z|x)}{Pr(A=0|x)} \\ &= \frac{Pr(A=0|z,x)\,f(y,z|x)}{Pr(A=0|x)} \\ &= \frac{f(y,z|x)}{w(z,x)} \end{aligned}$$

where $w(z,x)$ is given by

$$\text{(9)} \qquad w(z,x) = \left[\frac{Pr(A=0|z,x)}{Pr(A=0|x)}\right]^{-1}$$

which is a weight equal to the inverse of the normalized retention (*i.e.*, non-attrition) probability. Hence

$$\text{(10)} \qquad f(y,z|x) = w(z,x)g(y,z|x,A=0)$$

and thus the total-population joint density can be obtained by weighting the conditional density in the nonattriting population. Integrating both sides of (10) over z gives

$$\text{(11)} \qquad f(y|x) = \int_z g(y,z|x,A=0)w(z,x)dz$$

which is the density we seek. From this density we can compute conditional means and other moments of y.

All elements on the right-hand-side are consistently estimable with a sample that is random in all respects except for attrition. The terms inside the brackets in (9) involve the probability of retention in the sample which is, in the parametric model described above, $F(-\delta_0 - \delta_1 x - \delta_2 z)$, which is estimable. The conditional joint density of y and z can also be consistently estimated from a random sample of non-attritors. Thus the complete-population density $f(y|x)$ and its moments (such as its expected value, $\beta_0 + \beta_1 x$ in the parametric model) can also be consistently estimated. That the complete-population density in (11) can be derived by weighting the conditional density by the (normalized) inverse selection probabilities can be used to show weighted least squares can be applied to eqn(2), using the weights in (9), for consistent estimation. [6]

If z is not a determinant of attrition, the weights in (9) equal one and hence all conditional and unconditional densities are equal, and no attrition bias is present. Alternatively, if y and z are independent conditional on x and $A = 0$, the density g in (11) factors and it can again be shown that the unconditional density $f(y|x)$ equals the conditional density, and there is again no attrition bias. [7] These two conditions form the basis for two tests for attrition bias from selection on observables, as noted further below.

While these results are relatively unfamiliar in the econometric literature, they are pervasive in the survey sampling literature, where they form the intellectual justification for the construction and use of attrition-based survey weights (RAO, [1963 1975]; LITTLE and RUBIN, [1987], pp. 55-60 ; see MADOW *et al.*, (1983], for a survey).[8,9]

However, the spirit of the use of WLS here is rather different because it does not lead to a single, "*universal*" weight that can be constructed once and then used for all models and analyses that are conducted with the data set. Instead, the weights are more likely to be model-specific as the z variables entered into the attrition equation may be different for different choices of y.

In the econometrics literature, while weighting formulations are sometimes used as a framework for discussing selection models (*e.g.*, HECKMAN, [1987]), the main point of contact with the models discussed here is the choice-based sampling literature (for discrete y, see MANSKI and LERMAN, [1977] for an early treatment and AMEMIYA, [1985], for a textbook treatment; for continuous y, see HAUSMAN and WISE, [1981], COSSLETT, [1993], and IMBENS and LANCASTER, [1996]). That literature generally considers estimation and identification in samples which are selected directly on the dependent variable, y; weighted maximum likelihood or least squares procedures are often proposed to 'undo' the disproportionate endogenous sampling. The difference in the attrition case is that selection is on an auxiliary variable (z) and not on y itself; and, in addition, the choice-based sampling literature is primarily concerned with stratified sample designs where sampling probabilities are nonzero in all

[6] A consistency proof for WLS is straightforward and is available from the authors upon request. We should emphasize that the application of WLS in this case is unrelated to the heteroskedasticity rationale appearing in most econometrics texts. It is also not in conflict with the conventional view among many applied economists that survey weights can be ignored because they do not affect the consistency of OLS coefficients, for survey weights are often intended only to adjust for sample designs which have stratified the population or differentially sampled it by variables that are exogenous. Here, however, selection is indirectly based on the dependent variable, and not adjusting for attrition results in loss of consistency.

[7] The conditional and unconditional densities are also equal if y and z (conditional on x) are independent in the complete population, but this is not necessary.

[8] The use of weights to adjust for selection on endogenous observables has been considered in the econometric literature by COSSLETT ([1993], pp. 31-32) and WOOLDRIDGE [1987], but these authors are concerned with endogenous sample stratification rather than attrition. HOROWITZ and MANSKI (forthcoming) have considered weighting for attrition and have the result we have reported above which we discovered after our own work.

[9] Much of this literature is concerned with weights intended to adjust for samples which are stratified, often on endogenous variables, rather than to adjust for nonresponse (WLS for this purpose has also been considered in the econometrics literature by COSSLETT [1993] and WOOLDRIDGE [1997]). Although there is a basic similarity to the nonresponse problem, the stratified design case requires a slightly different formulation to justify.

strata, which is different than the attrition case. But otherwise the solutions are closely related. [10]

It should be noted that simply conditioning y on z (as well as x) does not, by itself, solve the problem. This can be seen most simply by observing that the object of interest is $E(y|x)$, not $E(y|x,z)$. Including z in the regressor set will generate biased coefficients on x in a linear-regression model, for example, in the sense that it will not estimate the effect of x on y unconditional on z. Because z is an endogenous variable, it distorts the conditional distribution of y on x. Hence correcting for selection on observables is to be sharply distinguished from the corrections for unobservable selection shown in eqn(7), which involve conditioning on functions of x and z or of the selection probability; those conditioning methods are not appropriate for this case. [11]

Panel Data. With the selection on observables framework thus laid out, the application to panel data is straightforward because the leading case of a z is a lagged value of y. Assuming serial correlation in the y process, such lagged variables will be related to current values of y conditional on x. If attrition is related to lagged y, least squares projection of y on x using the non-attriting sample will yield biased and inconsistent coefficient estimates. Estimation of attrition probabilities and subsequent WLS estimation yields consistent estimation instead. [12]

The most well-known model of attrition in the econometrics literature is the model of HAUSMAN and WISE [1979]; that model has been generalized and extended by RIDDER [1990, 1992], NIJMAN and VERBEEK {1992], VAN DEN BERG *et al.* [1994], and others (see VERBEEK and NIJMAN [1996], for a review). These models generally assume a components structure to the error term, sometimes including individual-specific time-invariant effects and sometimes serially-correlated transitory effects, for example, and impose restrictions on how attrition relates to the components of the structure. A common assumption in some studies in the literature, for example, is that attrition propensities are independent of the transitory effect but not the individual effect; in that case, simple first-differencing (among other methods) can eliminate the bias.

Our approach differs from this past work because of our sharp distinction between identifiability under selection on observables and on unobservables, a distinction not made in these past studies. Many error components models which allow attrition propensities to covary with individual components of the process can be treated within the selection on observables framework because lagged values of y can be mapped into those components. For example, both

[10] WLS for the attrition case has recently been considered by HOROWITZ and MANSKI, who have a result of the type given above in (11) but in different notation, a result of which we became aware after we had completed our analysis.

[11] However, $E(y|x)$ can be obtained by integrating z out of $E(y|x,z)$; hence conditioning on z can be used to derive the object of interest indirectly.

[12] The use of lagged values of y in this role requires the same distinction we noted earlier between structural and auxiliary determinants of contemporaneous y. The use of lagged y as a z makes sense only if the investigator is interested, for theoretical or other purposes, in functions of y not conditioned on those lagged values. An investigator who posits a structural model that includes all lags of y will necessarily have much reduced scope for selection on observables. Thus the relevance of the selection on observables method is model-dependent.

AR(l) and random walk error components processes for ε_t (again, suppressing individual subscripts *i*, which are implicit on all terms) are

$$\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t, \quad -1 < \rho < 1 \tag{12}$$

$$\varepsilon_t = \varepsilon_{t-1} + \eta_t \tag{13}$$

where the η_t are assumed to be i.i.d. Assuming that attrition propensities at time *t* are functions only of those components that appear at $t-1$, we assume

$$A_t^* = \delta_0 + \delta_1 x + \delta_2 \varepsilon_{t-1} + \nu_t \tag{14}$$

and in both cases we assume η_t and ν_t are independent. Letting $y_{t-1} = \beta_0 + \beta_1 x + \varepsilon_{t-1}$ it is clear that (14) can be written equivalently as

$$A_t^* = \delta_0 + \delta_1 x + \delta_2 y_{t-1} + \omega_t \tag{15}$$

(with suitable redefinition of parameters) which fits into the selection on observables framework.

This approach can be generalized in an simple way to models with multiple error components. Suppose that a vector of $t-1$ values of lagged *y* are available when considering attrition at time *t* and that the process for ε over these $t-1$ periods contains *s* unique linear error components. Then the selection on observables framework can be applied if $s \leqslant t-1$ and if the usual rank condition for a unique solution of a set of linear relations is met, for in that case the *s* error components can be exactly mapped into the $t-1$ observables for lagged *y*. This is only a sufficient condition; if, for example, some of the components of *s* are assumed to be independent of the attrition process then a weaker condition is required.

If the number of error components exceeds the number of available lags, or if the component of contemporaneous ε not forecastable from lagged values of *y* covaries with the attrition probability (*e.g.*, shocks to earnings which occur simultaneously with, not prior to, attrition from the sample), or if the lagged *y* captures the biasing error components only imperfectly, then bias remains and must be considered under the selection on unobservables framework. [13] However, a full conditioning on the available data on the history of y in the panel reduces the scope of possible unobservable selection because it isolates the remaining sources of such bias.

Testing. As noted previously, two sufficient conditions for the absence of attrition bias on observables are either that the weights equal one (*i.e.*, *z* does not affect A) or that *z* is independent of *y* conditional on *x*. Specification tests for selection on observables can be based on either of these two conditions. Thus one test is simply to determine whether candidate variables for *z* significantly affect A. A second test is conduct a specification test for whether OLS and WLS estimates of eqn (2) are significantly different, which is an indirect test for whether the identifying variables used in the weights are endogenous.

[13] Excluded from the selection on observables framework is the conventional random effects model with time invariant individual effect *y* and white-noise errors η. In this case, lagged y are only a noisy signal for μ (except as *T* goes to infinity, in which case mean *y* equals μ) and hence attrition propensities will still be dependent on comporaneous *y*, conditional on lagged *y* (and *x*). We thank G. RIDDER for this point.

A HAUSMANN test on the significance of the difference in the weighted and unweighted coefficient estimates is one type of test that could be applied (see DUMOUCHEL and DUNCAN, [1983] for an example of such a test for the stratified sample problem).

Another test for selection on observables which we will perform in the next section is based on an exercise performed by BECKETTI *et al.* [1988] and which we term the "*inversion*" test. In the inversion test, the value of y at the initial wave of the survey, which we denote by y_0, is regressed on x and on future A. The test for attrition selection is based upon the significance of A in that equation. [14] This test must necessarily be closely related to the test we have already described of regressing A on x and y_0, which is z in this case. In fact, the two equations are simply inverses of one another. Despite this, the inversion formulation of the test is also of interest because it shows directly the implications of the model for (initial period) y, which the attrition equations do not. Formally, suppose that the attrition function is taken as the latent index in the parametric model, *i.e.*

$$A^* = \delta_0 + \delta_1 x + \delta_2 z + \nu \tag{16}$$

Inverting this equation, taking expectations, and applying Bayes' Rule, it can be shown that

$$E(y_0|A,x) = \int y_0 f(y_0|x) w(A,y_0,x) dy_0 \tag{17}$$

where

$$w(A,y_0,x) = \frac{Pr(A|y_0,x)}{Pr(A|x)} \tag{18}$$

which is the same as the weight appearing in (9) but including the probability of $A = 1$ as well as $A = 0$. Eqn(17) shows that if the weights all equal one, the conditional mean of y_0 is independent of A and hence A will be insignificant (in the limit) in a regression of y_0 on x and A (the conditional mean of y_0 in the absence of attrition bias is $\beta_0 + \beta_1 x$, so a regression of y_0 on x will yield estimates of this equation). As noted previously, the weights will equal one only if y_0 is not a determinant of A conditional on x. Thus the inversion method is an indirect test of the same restriction as the direct method of estimating the attrition function itself. [15]

However, if the weights do not equal one, it would be difficult to derive equation (17) from the estimates of (16) that we will obtain in our attrition propensity models. To do so would require conducting directly the integration shown in (17). It would be simpler to just estimate a linear approximation to (17) by OLS, as did BECKETTI *et al.* [1988], to determine the magnitude of the effect of A on the intercept and coefficients of the equation for y_0 as a function of x. However, it should be kept in mind that this is not an independent

[14] We assume x to be time-invariant. If it is not, this method requires that only the values of x at the initial wave be included in the equation.

[15] In general, of course, if $v = \alpha + \beta u + \varepsilon$, regressing u on v instead of v on u results in a "*biased*" coefficient on v (*i.e.*, it is not a consistent estimate of the inverse of β). Nothing here contravenes that result. The "*coefficient*" on x in a regression of y on x and A bears no simple relationship to δ_1 or δ_2 in eqn(16), as can be seen from eqn(l7).

test of attrition bias separate from that embodied in estimates of eqn(16); it is only a shorthand means of deriving the implications of our estimates of eqn(l6) for the magnitudes of differences in first-period y conditional on X between attritors and nonattritors.

3 Illustration From the Michigan Panel Study of Income Dynamics

3.1. General Attribution Patterns

The Michigan Panel Study of Income Dynamics (PSID) began in 1968 with a sample of approximately 4800 families drawn from the U.S. noninstitutional population (for a general description of the PSID see HILL, [1992]). Since 1968 families have been interviewed annually and a wide variety of socioeconomic information has been collected; we use data through 1989. [16] Table 1 shows response and nonresponse rates of the original 1968 sample members. The first three columns in the table show the number of individuals remaining in the sample by year – the number in a family unit, the portion in institutions – whom we treat as respondents, to be consistent with practice by PSID staff – and their sum, equal to 18,191 individuals in 1968. As the table indicates in the fourth column, about 88 percent of these individuals remained after the second year, implying an attrition rate of 12 percent. The actual number attriting is shown in the fifth column, with conditional attrition rates shown in parentheses below each count. A smaller proportion left the PSID in each year after the first – generally about 2.5 or 3.0 percent annually. By 1989, only 49 percent of the original number were still being interviewed, corresponding to a cumulative attrition rate of 51 percent.

The table also shows the distribution of the attritors by reason-either because the entire family became nonresponse ("*family unit nonresponse*"), because of death, or because of a residential move which could not be successfully followed. The distribution of attrition by reason has not changed greatly over time, although there is a slight increase in the percent attriting because of death and a slight reduction in the percent attriting because of mobility. Both of these trends are no doubt a result of the increasing age of the 1968 sample. The final column in the table shows the number of individuals who came back into the survey from nonresponse ("*In from nonresponses*") each year. These figures are quite small because, prior to the early l990s, the PSID did not attempt to locate and reinterview attritors.

[16] About three-fifths of the 1968 families were drawn from a representative sampling frame of the U.S. called the "*SRC*" sample, and two-fifths were drawn from a set of individuals in low-income families (mostly in SMSAs) known as the "*SEO*" sample. At the time the survey began, the PSID staff produced weights that were intended to allow users to combine the two samples and to calculate statistics representive of the general population. We use those weights not to adjust for attrition but to adjust for the initial choice-based design.

TABLE 1

Response and Nonresponse Rates in the PSID

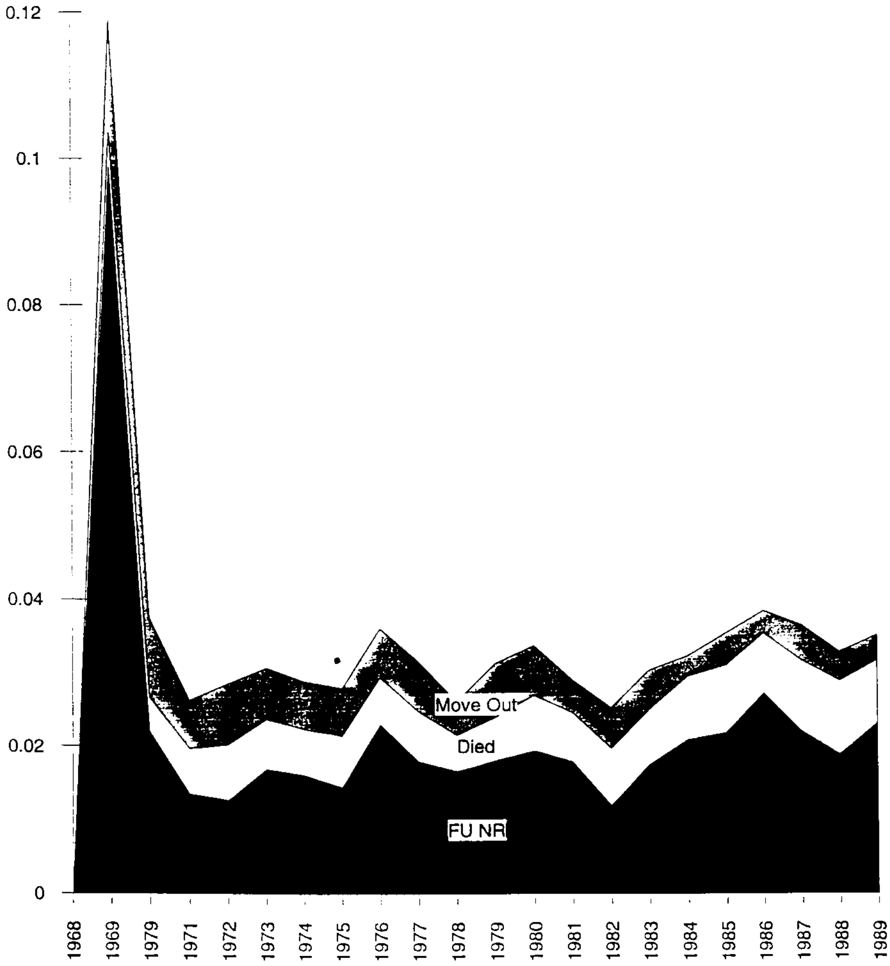
	Remaining in Sample				Attritors*				
Year	In a Family Unit	In an Institution	Total	As a Pct of 1968 Total	Total	Fam. Unit Non-resp.	Died	Moved	In from Non-resp.
1968	17870	384	18191	100.0	–	–	–	–	–
1969	15561	367	16028	88.1	2163 (.119)	1797 (.099)	84 (.005)	282 (.016)	–
1970	15126	333	15459	85.0	600 (.037)	351 (.022)	74 (.005)	175 (.011)	31
1971	14767	322	15089	82.9	404 (.026)	208 (.013)	95 (.006)	101 (.007)	34
1972	14400	293	14693	80.8	429 (.028)	190 (.013)	115 (.008)	124 (.008)	33
1973	13969	307	14276	78.5	449 (.031)	247 (.017)	100 (.007)	102 (.007)	32
1974	13581	307	13888	76.3	410 (.029)	229 (.016)	89 (.006)	92 (.006)	22
1975	13226	302	13528	74.4	386 (0.28)	200 (0.14)	97 (.007)	89 (.006)	26
1976	12785	291	13076	71.9	487 (.036)	310 (0.23)	86 (.006)	91 (.007)	35
1977	12377	310	12687	69.7	411 (.031)	234 (.018)	88 (.007)	89 (.007)	22
1978	12078	320	12398	68.2	330 (.026)	210 (.017)	63 (.005)	57 (.004)	41
1979	11718	316	12034	66.2	387 (.031)	224 (.018)	73 (.006)	90 (.007)	23
1980	11357	305	11662	64.1	405 (.034)	33 (.019)	90 (.007)	82 (.007)	33
1981	11022	340	11362	62.5	337 (.029)	208 (.018)	77 (.007)	52 (.004)	37
1982	10780	326	11106	61.1	285 (.025)	135 (0.12)	88 (.008)	62 (.005)	29
1983	10487	322	10809	59.4	336 (.030)	194 (.017)	83 (.007)	59 (.005)	39
1984	10178	319	10497	57.7	348 (.032)	225 (.021)	93 (.009)	30 (.003)	36
1985	9891	275	10166	55.9	371 (.035)	229 (.022)	96 (.009)	46 (.004)	40
1986	9517	292	9809	53.9	390 (.038)	275 (.027)	84 (.008)	31 (.003)	33
1987	9230	257	9487	52.2	357 (.036)	215 (.022)	94 (.010)	48 (.005)	35
1988	9002	206	9208	50.6	310 (.033)	178 (.019)	95 (.010)	37 (.004)	31
1989	8743	170	8913	49.0	323 (.035)	212 (.023)	79 (.009)	32 (.003)	28

Notes:
Excludes new births and nonsample entrants.
* Figures in parentheses show attrition rates as a percent of the total sample remaining in the prior year (column four).

Figure 1 illustrates the overall attrition hazards graphically. The Figure clearly shows the spike in the hazard in the first year. It is also more noticable in the Figure that there has been a slight upward trend in attrition rates over time, although not large in magnitude.

In other work (FITZGERALD et al., [1997]), we have adjusted these attrition figures for mortality in the population in two ways. First, we excluded individuals who died while in the survey and who could thus be identified as having left for this reason. Second, since some of those who attrited later died, we used national-level mortality rates by age, race, and sex to estimate the number of attritors who were still alive, and to use those as a basis for calculating attrition rates. These calculations lower the attrition rate for the older population to 35 percent and the overall attrition rate to 44 percent.

FIGURE 1
Attrition Hazards: Sample With No New Entrants



3.2. **Attrition Analyses**

We conduct four types of attrition analysis. First, we analyze whether attritors and nonattritors differ in their *x* and *y* characteristics in 1968, the initial year. Second, we estimate probit equations for whether individuals had attrited from the PSID by 1989 using 1968 *x* and *y* as regressors. We do not use *x* and *y* information from the intervening period 1969-1988 because attrition had already begun by 1969 and hence incorporating that information would require a year-by-year hazard model, which we leave for future work. We shall focus on labor income as our example of a *y*, since this is one of the most commonly used outcome variables in analyses using PSID data. Third, we estimate inversion equations by regressing 1968 labor income (*y*) on *x* and on future A and determine whether these yield the same inferences as the attrition probits.

Our fourth analysis addresses the issue of using the information from 1969 to 1988 for a hazard analysis of attrition but only in a limited way. We ignore potential bias from attrition prior to each year and present simple illustrative year-by-year hazards of attrition using the full history of y prior to the attrition point to formulate regressors. [17] We do not formulate formal error components models but instead test whether various functions of lagged *y* (mean, variance, period-specific deviations) covary with attrition. The aim of these attrition equations is to test whether attrition is selective on individual "*dynamics*", as we will discuss.

We conduct all analyses only on male heads of household aged 25-64.

Characteristics of Attritors. We focus on the seemingly simple question of whether 1968 characteristics differ between those who were present in 1989 and those who were not (hence the distributions of *x* and *y* conditional on *A*, in a tabular form). Table 2 shows the mean values of 1968 characteristics by their attrition status as of 1989 – "*always in*" versus "*ever out*" by that year. [18] As the first two columns indicate, attritors and nonattritors have many significant differences in characteristics. Attritors are more likely to be on welfare, less likely to be married, and are older and more likely nonwhite. In addition, attritors have lower levels of education, fewer hours of work, less labor income, and are less likely to own a home and more likely to rent. The clear implication of this pattern is that attritors are concentrated in the lower portion of the socioeconomic distribution. The second moments for labor income in the table indicate that the variance of labor income is greater among attritors than among nonattritors, and, interestingly, that the attritor labor income distribution is more dispersed at the upper tail than the nonattritor distribution. This suggests that, to some degree, some high labor-income families may be more likely to attrite than middle income families.

The last two columns in the table provide an assessment of the effect of mortality. The third and fourth columns disaggregate the "*ever out*" subsample

[17] Unobserved heterogeneity in the attrition equation biases the structural coefficient estimates in the attrition equation but does not by itself cause bias in the main equation after the application of WLS. In eqn(3), there is no requirement that *x* and *v* be independent for WLS to be consistent.

[18] Because only a tiny fraction of attritors ever return – see Table 1 above – those individuals who were "*always in*" between 1968 and 1989 are almost identical to the set of individuals present in 1989, and the set of individuals who were "*ever out*" between 1968 and 1989 is almost identical to those who were nonresponse in 1989.

into those “*not dead*” and those “*dead*” according to whether individuals died while in the PSID (as noted previously, some individuals die after attriting, of which we have no knowledge). Comparing the third column (not dead) with the first two shows that the gap between the Always In and Ever Out is sometimes narrowed by excluding the dead from the attritors, but rarely by very much; indeed, in some circumstances, the gap even increases. The latter occurs when mortality is related to a variable in opposite sign to its relation to attrition conditional on being alive; consequently, ignoring mortality actually makes the selectiveness of attrition seem milder than it actually is.

TABLE 2
1968 Characteristics by Attrition Status

	Always In	**Ever Out**	**Ever Out/ Not Dead**	**Evert Out/ Dead**
Welfare Participation (%)	0.8	1.3	1.4	1.2
Marital Status (%)				
married	95.8	90.1*	87.1	98.1
never married	2.4	3.7*	4.9	0.4
widowed	0.3	1.5*	2.0	0.1
divorced/separated	1.2	4.6*	5.9	1.3
Percent with Annual Hours Worked >0	98.7	94.1*	95.7	89.8
Annual Labor Income	21345	17011	17277	16298
Annual Labor Income for those w/income > 0	21631	18152	18106	18281
Annual Hours Worked fot those w/hours > 0	2378	2246	2268	2182
Variance of log annual labor income for those w/income > 0	.248	.529	.481	.667
Labor income quintile ratios for those w/labor income > 0 :				
Quintile 20/median	.658	.611	.615	.558
Quintile 40/median	.886	.905	.923	.865
Quintile 60/median	1.101	1.139	1.123	1.164
Quintile 80/median	1.392	1.498	1.462	1.493
Education (%) : < 12	31.5	52.5*	50.8	57.2
12	32.8	25.6*	27.3	21.0
12-15	15.8	11.5*	11.5	11.5
16+	19.9	10.4*	10.4	10.4
Race (%): White	92.7	88.3*	87.4	90.7
Black	6.6	10.7*	11.5	8.0
Region (%) : Northeast	24.7	25.8	26.9	22.3
North Central	32.2	27.5*	26.5	30.1
South	26.7	30.1*	29.6	31.2
West	16.4	16.7	17.0	15.7
Age	40.7	45.6*	43.1	52.1
Tenure(%) : Own home	74.9	62.9*	58.0	75.9
Rent	21.5	33.8*	38.9	20.2
Number of children in Familly	2.0	1.5	1.6	1.3
Sample size	1238	1533	1116	417

Notes : Sample weights used.
*: Significantly different from “*Always In*” at 10 % level.

TABLE 3

Attrition Probits

	Model 1		**Model 2**		**Model 3**		**Model 4**	
	Coeff.	$\partial P/\partial x$	Coeff.	$\partial P/\partial x$	Coeff.	$\partial P/\partial x$	Coeff.	$\partial P/\partial x$
Intercept	.334*	.128	.360*	.139	1.770*	.671	1.130*	.417
	(.059)		(.096)		(.454)		(.518)	
Labor Income$^{a)}$	–.0239*	–.0092	–.0272*	–.0105	–.0192*	–.0073	–.0237*	–.0088
	(.0030)		(.0103)		(.0108)		(.0120)	
No Labor Income	.284*	.110	.254	.100	.291	.110	.181	.067
	(.160)		(.177)		(.180)		(.186)	
Labor Income Squared $^{b)}$			.009	.003	.018	.006	.022	.008
			(.025)		(.026)		(.026)	
Black					.074	.028	.037	.014
					(.066)		(.081)	
Other race					.356	.134	.198	.073
					(.248)		(.251)	
Age					–.088*	–.033	–.039	–.014
					(.022)		(.024)	
Age Squared $^{c)}$					.107*	.041	.054*	0.20
					(0.25)		(.028)	
Education <12 Years					.200*	.076	.208*	.077
					(.690)		(.071)	
Some College					–.114	–.043	–.195*	–0.72
					(.096)		(.097)	
College Degree					–.305*	–.116	–.384*	–.142
					(.107)		(.109)	
Northeast							–.051	–.019
							(.939)	
North Central							–.139	–.051
							(.091)	
South							–.120	–.044
							(.088)	
In SEO Sample							–.070	–.025
							(.080)	
Lives in Rural Area (SMSA <1000)							–.271*	–.100
							(0.72)	
Number of Children in Family							–.033*	–0.12
							(.017)	
Presence of Child <6							.095	.035
							(.061)	
Owns House							–.310*	–.114
							(.068)	
Might Move In Future							–.015	–.006
							(.072)	
Income/Needs ratio							.031	.012
							(.033)	
R^2	.028		.028		.044		.068	
Sample Size	2253		2253		2253		2253	
Number Ever out	1074		1074		1074		1074	
Loq like.	–1516.05		–1515.99		–1490.327		–1453.02	

Notes: Excludes known dead. Characteristics leasured in 1968.
*: Significant at 10% level. SEO+SRC weighted. $\partial P/\partial X$ signifies the effect of a unit change in the variable on the probability of attribution evaluated at the mean.
R^2 equals one minus the ratio of the log likehood of the fitted function to the log likelihood of a function with only an intercept.
$^{a)}$ Coefficients multiplied by 10^3. $^{b)}$ Coefficients multiplied by 10^8. $^{c)}$ Coefficients multiplied by 10^2.

Attrition Probits. The first multivariate analysis we present consists of estimates of binary-choice models for the determinants of attrition, using the same data in the tables we have been presenting (*i.e.*, whether having ever been nonresponse by 1989 as a function of 1968 characteristics). We therefore estimate probit equations for the probability of having ever been nonresponse by 1989. [19] As in Table 2, the sample consists of all 1968 male heads 25-64 and all regressors are measured in 1968. The "*y*" variable we focus on is labor income.

Table 3 shows a set of expanding specifications of attrition probits. The first two columns of the table show the effect of labor income on attrition without conditioning on any other regressors ("*No Labor Income*" is a dummy equal to 1 if the individual has no labor income). The results show that the 1968 labor income levels of male heads have a very strong correlation with future nonresponse. Attrition probabilities are quadratic in labor income–lowest at middle income levels and greatest at high and low income levels. Individuals with no labor income at all have higher attrition rates as well.

The third column in the table shows that when "*standard*" earnings-determining variables are added – race, age, and education – labor income remains a significant determinant of attrition. Implicitly, therefore, the residual in a labor income equation containing these regressors is correlated with attrition probabilities. When a large number of other variables – income/needs, home ownership, SEO status, and others – are added, the labor income effects are considerably attenuated, but still remain. We should also note that the R-squareds from these probits are extremely small and never exceed .069. [20] Thus, even where significant correlates of attrition are found, they explain very little of the variation in attrition probabilities in the data. One implication of this result is that weights based on these equations would, in all likelihood, have little effect on estimated outcome equations. [21]

Inversion Tests. As we noted previously, the inversion of our attrition probits – the effect of future attrition on 1968 outcome variables, rather than the other way around – is also of interest. Such regressions were estimated by BECKETT *et al.* [1988] and used as a test for attrition bias. As we noted previously, apart from nonlinearities and some differences in the stochastic assumptions, the results should have the same general tenor as the attrition probits but will show more directly the degree to which regression coefficients in typical outcome equations are affected.

[19] Although we do not estimate a dynamic model of year-by-year attrition, these estimates can be viewed as a model of cumulative attrition that reflects the working-out of a year-by-year model. Since all the regressors are held at their 1968 values, our equation can be viewed as an approximation to the reduced-form model.

[20] The R-squared measure we use is defined in the footnote to the Table and is a common measure of fit in binary-choice models. This measure has recently been shown to have desirable properties relative to other measures (CAMERON and WINDMEIJER, [1997]) and can be interpreted as the proportionate reduction in uncertainty from the fitted model, where uncertainty is defined by an entropy measure.

[21] This statement must be qualified because even weights with very small variance could have a large impact if they are sufficiently highly correlated with the error term and the regressors.

TABLE 4

1968 Log Labor Income Regressions

	SRC and SEO			Combined SRC Only		
	Total	**Always In**	**Difference**	**Total**	**Always In**	**Difference**
Intercept	8.24* (.197)	8.38* (.232)	.14 (.12)	8.28* (.23)	8.35* (0.26)	.08 (.13)
Black	–.249* (.044)	–.272* (.056)	–.022 (.035)	–.173* (.055)	–.195 (0.070)	–.022 (.043)
Other Race	–.221 (.136)	–.246 (.173)	.196* (.106)	–.393* (0.164)	–.193 (.184)	.200* (.0830)
Ed < 12	–.293* (.034)	–.271 (.039)	.023 (.019)	–.291* (.40)	–.244* (.045)	.047* (.020)
Some College	.101* (.037)	.068* (.039)	–.033* (0.14)	.103* (.042)	.098* (.044)	–.005* (.001)
College Degre	.271* (.043)	.283* (.045)	.012 (.011)	.311* (.050)	.334* (.050)	.024* (.008)
Age	.080* (.009)	.074* (.001)	–0.059 (.061)	.080* (.011)	.079* (.013)	–.001 (.007)
Age Squared[a)]	–.948* (.108)	–.856* (.132)	.092 (.075)	–.947* (.125)	–.922* (.149)	.003 (.081)
Northeast	.076* (.039)	.110* (.045)	.034 (.022)	.088* (.047)	.065 (.052)	–.022 (.023)
North central	.045 (0.38)	.006 (.043)	–.039* (.020)	.013 (.043)	–.056 (.048)	–.069* (.021)
South	–.076* (.039)	–.105* (.045)	–.028 (.023)	–.111* (.045)	–.147* (.051)	–.036 (.025)
Sample Size	2182	1159		1406	788	
R^2	.19	.24		.22	.26	
F-Satistic[b)]	50.5	25.7		38.8	27.8	
Variance of Error	.326	.220		.285	.194	

Notes: Standard errors in parentheses.
Sample excludes know dead. SRC+SEO sample are weighted.
*: Significant at 10% Level.
[a)] Coefficient multiplied by 10^3.
[b)] F-Statistic for hypothesis that all coefficients except the intercept are equal to zero.

TABLE 5

1968 Lagged Earning Equations:
Difference in Total And Always-In Samples, Intercept-Only Model

	SRC + SEO	**SRC Only**
Intercept Difference	-.059* (.012)	-.053 (.013)

Notes : Models include all variables shown in Table 4 but allow the intercept to differ for the Total and Always-In Samples. Coefficient equals Total-Sample intercept minus Always-In Sample intercept. Standard errors in parentheses. Samples excludes Known dead
SRC+SEO is weighted
* Significant at the 10 % level.

Table 4 shows 1968 log labor income regressions. [22] Separate regressions are estimated for individuals who were always in the sample through our final year, 1989, and for the total sample in 1968. We compare the total sample and the nonattriting sample – not attritors and nonattritors–because the issue is how different parameter estimates would be from those in the total sample if only the nonattriting sample is used. [23] We show results separately when the SEO sample is included and excluded. None of the coefficients on the variables of most past research interest – Black, Ed<12, College Degree, Age and Age-Squared – are significantly different between the total and nonattriting samples in estimates including the SEO, and the magnitudes of the differences in the coefficients are seldom large from a substantive research point-of-view. Significant differences do appear for the "*Other Race*" and "*Some College*" variables (and one of the region variables), for reasons we have not been able to determine. More significant difference appear for the estimates when the SEO is excluded, but these are again not large in magnitude. In summary, at least for SRC-SEO combined sample, we find very few important effects of attrition on the coefficients. [24]

Wald tests for the joint significance of the differences in all slope coefficients and intercepts generally reject the hypothesis of equality between the vectors. However, when tests are conducted for the equality of the slope coefficients allowing the intercepts to differ, most fail to reject equality. The estimates on the intercept differences (*i.e.*, constraining all coefficients on the other regressors to be the same for the two groups) are shown in Table 5 and are significant in all cases. Thus we conclude that, while the coefficients on "*standard*" variables in labor income and welfare participation equations and, to a lesser extent, marital-status equations, are unaffected by attrition, there are still be differences in the levels of these outcome variables conditional on the regressors.

Attrition Hazards. In the final piece of our analysis, we estimate year-by-year attrition hazards through 1989 using the full history of lagged labor income prior to each time point.

Although we have not developed a formal model of the causes of attrition, it is plausible to hypothesize that not only are low socioeconomic-status individuals likely to attrite (as our results on levels of the relevant variables have demonstrated thus far) but also that individuals with a recent change in earnings are more likely to attrite. Taking this notion one step further, we hypothesize that individuals observed over their full past history to have had above average rates of fluctuations in earnings are more likely to attrite.

[22] Individuals with zero labor income are excluded. While this introduces some noncomparability with our attrition probits as well as raising well-known selection issues, we wish to maintain correspondence with the bulk of the earnings function literature, which also generally conditions on positive income.

[23] The two sets of differences are transforms of one another, but they have different standard errors. Under the null of equality of the true coefficient vectors, the variance of the difference in the estimates is the difference in the separate variances (the variance in the smaller sample must be larger, necessarily, under the null).

[24] We calculated White standard errors for the coefficients but found them to be only 5 percent higher, at most, than those shown . We therefore do not calculate them for the remainder of the analysis.

We conjecture that disruption in general may be related to attrition because it may make individuals either more difficult to locate by the PSID field staff, or less receptive to participation in the panel, or both.

To investigate this issue, we estimate attrition functions with a latent index of the form:

$$A_{it}^{*} = f(y_{i,t-1}, y_{i,t-2}, \ldots, y_{i0}) + x_{i0}^{\theta} + \nu_{it} \tag{19}$$

where the outcome variable, A_{it}, equals 1 if the individual attrites at time t, conditional on still being a respondent at $t-1$. The vector x_{i0} consists of time-invariant "*x*" variables, with coefficient vector θ. Eqn(l9) allows the lagged dependent variables to affect current attrition propensities in a general way (function *f*) but, in our empirical work, we test functions which transform the lagged y into only three different summary variables: (a) the individual-specific mean of the variable over all years since 1968; (b) the individual specific variance of the variable over all years since 1968; and (c) deviations of lagged variables from the individual-specific means.

The first of these measures tests whether attrition is affected by individual-specific mean levels of earnings. This analysis should yield broadly similar findings to those reported in the last section, for they only replace the 1968 values of these variables with their means over a period of years. The second of the statistics measures individual heterogeneity in labor market and earnings turnover. As we noted previously, if attrition covaries with lagged values for these variables, then it follows that models estimated on nonattritors but using the contemporaneous counterparts to these measures as dependent variables (turnover, durations, transition rates, etc.) will be biased provided that the contemporaneous and lagged measures covary as well. The third of the measures tests whether lagged changes ("*shocks*") to these variables affect attrition. This is logically separate from the question of individual heterogeneity in turnover. It relates closely to the issue of whether transitory events affect later attrition, although we cannot be sure of that interpretation because we cannot, by definition, determine whether recent events will persist in the future or not if the individual attrites (and hence whether the events will, in retrospect, be seen to be permanent or transitory shocks). This analysis has implications for bias in the estimation of transition rate models for contemporaneous variables on the nonattriting sample.

For our models we pool all observations on individuals 25-64 in original 1968 sample families for all years 1970-1989 for which they are observed. [25] We estimate logits for whether the individual attrites in the next period as a function of the three summary measures discussed above defined as of the current period. We also include 1968 variables for education, age, and other socioeconomic characteristics. In some runs we include year dummies, which fully capture duration dependence.

[25] We omit 1968 and 1969 so that we can construct at least two lagged variables for individuals last observed in 1970. Note that we do not include individual effects in the estimation and hence the lagged regressors will pick up the influence of those effects. This is acceptable for the purposes of this exercise, which is not to obtain structural coefficients but rather to determine the reduced form relation between attrition and the history of *y*, including its influence through individual effects.

Table 6 shows a series of estimated attrition hazards. Column (1) shows that attrition propensities are significantly negatively affected both by lagged mean earnings as well as earnings in the prior period. The latter implies that negative deviations of current earnings from mean earnings raise the likelihood of attrition. Column (2) shows that the effect of deviations does not extend back beyond the current period. Column (3) tests the effect of the individual specific variance and finds that attrition rates are positively affected by variances, even conditioning on current period and lagged mean earnings. Column (4) shows that this result is robust to the inclusion of age and year dummies, for it might be the case that if attrition rates vary with calendar year or age, this might create spurious estimates since earnings vary with year and age.[26] However, column (5) shows that the inclusion of several standard socioeconomic variables (education, race, etc.) is sufficient to render insignificant the effect of lagged mean earnings on attrition rates, a result not surprising inasmuch as permanent earnings are likely to be more predictable by such regressors than are earnings deviations or earnings variances. The latter two remain significant even after inclusion of the additional regressors.

TABLE 6
Dynamic Attribution Models with Focus on Lagged Earnings (Logit Coefficients)

	(1)	(2)	(3)	(4)	(5)
$\overline{y}$	–.20*	–.24*	–.28*	–.26*	–.07
	(.07)	(.08)	(.08)	(.08)	(.09)
y_{t-1}	–.22*	–.17*	–.18*	–.20*	–.15*
	(.06)	(.08)	(06)	(06)	(.07)
y_{t-2}	–	–.09	–	–	–
		(. 09)	–	–	–
Var(y)	–	–	.32*	.33*	.38*
			(. 09)	(. 09)	(. 09)
Time Dummies and age	n	n	n	y	y
Other Characts.[a)]	n	n	n	n	y
R^2	.018	.017	.020	.025	.043

Notes : Dep. var. is 1 if individual attrites in next period, 0 if not. $\overline{y}$ is the mean earnings from 1968 to current period; y_{t-1} and y_{t-2} are earnings in the current period and one period back; and var(y) is the variance of earnings from 1968 to the current period. The coefficients on the first three variables are multiplied by 10^4 and the coefficient on the fourth is multiplied by 10^8.

Standard errors in parentheses. For R-squared definitions, see Table 3.

*: significant at the 10 percent level.

a) Education, race, region, age of youngest child, rural residence, homeowner.

[26] The year dummies show no significant duration dependence in the hazard after 1970.

These results, therefore, are consistent with attrition being selective on stability. Therefore it should be expected that measures of second moments, of turnover and hazard rates, and of related variables should be smaller in the nonattriting PSID sample than in the population as a whole.

Although these results clearly demonstrate a tendency for men with more unstable histories to attrite, the seriousness of the problem for the PSID is difficult to judge. The R-squared values in these attrition equations are uniformly very small, as shown in the tables, which implies that attrition along these dimensions may not have a large effect on the comparable contemporaneous measures on the nonattriting sample from selection on these observables. This cannot be known for certain because the size of the bias depends not only on the R-squared values, but also on the size of relation of these lagged instability measures with both the regressors in the main outcome equation of interest and with the error term in that equation (recall the model discussed previously). However, weights based on these equations could be developed which would capture dynamic effects and which could be used in specification tests to test the importance of their effect on estimates of outcome equations.

6 Summary

We have demonstrated in this paper that a selection bias model relevant to the problem of attrition in panel data is a model in which selection is on endogenous observables, such as lagged dependent variables. Consistent estimates can be obtained with weighted least squares by using weights constructed from estimated selection probability functions. This model ignores the selection on unobservables that has been the focus of the traditional attrition model and therefore future research would profitably devise a general model to incorporate both.

In our application to the Michigan Panel Study of Income. Dynamics, we find some evidence of attrition bias in the earnings of male heads of household but that these primarily affect the intercepts of earnings equations and not the coefficients. We also find that attrition has been selective on stability of earnings profiles, and suggest that weights based on lagged variables measuring stability be used to eliminate this source of bias.

We have not considered any of the detailed estimation issues involved in using weighted procedures with the estimated attrition functions we describe nor have we addressed efficiency issues. These are appropriate topics for future research.

• References

AMENIYA T., (1985). – *Advanced Econometrics*; Cambridge: Harvard University Press.

BECKETTI S., GOULD W., LILLARD L., WELCH F., (1988). – “The Panel Study of Income Dynamics after Fourteen Years: An Evaluation”, *Journal of Labor Economics* 6, p. 472-492.

CAMERON A.C., WINDMEIJER F.A.G., (1997). – “ An R-squared Measure of Goodness of Fit for Some Common Nonlinear Regression models”, *Journal of Econometrics* 77, pp. 329-342.

COSSLETT S., (1993). – “Estimation from Endogenously Stratified Samples”, In *Handbook of Statistics*, Vol 11, eds. G.S. Maddala, C.R. Rao, and H.D. Vinod, eds. Elsevier.

DUMOUCHEL W., DUNCAN G., (1983). – “Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples”, *Journal of the American Statistical Association* 78, pp. 535-543.

FITZGERALD J., GOTTSCHALK P., MOFFITT R., (1997). – A Study of Sample Attrition in the Michigan Panel Study of Income Dynamics, *Mimeographed*, Johns Hopkins University.

HAUSMAN J., WISE D., (1979). – “Attrition Bias in Experimental and Panel Data : The Gary Income Maintenance Experiment”, *Econometrica* 47, pp. 455-474.

HAUSMAN J., WISE D., (1981). – “Stratification on Endogenous Variables and Estimation: The Gary Income Maintenance Experiment”, In *Structural Analysis of Discrete Data with Econometric Applications*, eds C. Manski and D. McFadden, Cambridge: MIT Press.

HECKMAN J., (1978). – “Dummy Endogenous Variables in a Simultaneous Equations System”, *Econometrica* 46, pp. 931-960.

HECKMAN J., (1979). – “Sample Selection Bias as a Specification Error”, *Econometrica* 47, pp. 153-162.

HECKMAN J., (1987). – “Selection Bias and Self-Selection”. In *The New Palgrave : A Dictionary of Economics*, eds. J. Eatwell, M. Milgate, and P. Newman, Vol.IV London: Macmillan.

HECKMAN J., HOTZ V.J., (1989). – “Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social programs” *Journal of the American Statistical Association* 84, pp. 862-874.

HECKMAN J., ROBB R., (1985). – “Alternative Methods for Evaluating the Effects of Interventions”, In *Longitudinal Analysis of Labor Market data*, eds. J. Heckman and B. Singer. Cambridge: Cambridge University Press.

HILL M., (1992). – *The Panel Study of Income Dynamics*: *A User’s Guide*, Newbury Park, Ca. : Sage Publications.

HOROWITZ J., MANSKI C., (1998). – “Censoring of Outcomes and Regressors Due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations”, *Journal of Econometrics* 84, pp. 37-58.

IMBENS G., LANCASTER T., (1996). – “Efficient Estimation and Stratified Sampling”. *Journal of Econometrics* 74 pp. 289-318.

LITTLE R., RUBIN D., (1987). – *Statistical Analysis with Missing Data*, New-York: Wiley.

MADDALA G.S., (1983). – *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge : Cambridge University Press.

MADOW W., OLKIN I., RUBIN D., (1983). – eds. *Incomplete Data in Sample Surveys*, 3 volumes, New York: Academic Press.

MANSKI C., LERMAN S., (1977). – “The Estimation of Choice Probabilities from Choice-Based Samples”, *Econometrica* 45, 1977-1988.

NIJMAN T., VERBEEK M., (1992). – “Nonresponse in Panel Data: The Impact on Estimates of a Life Cycle Consumption Function”, *Journal of Applied Econometrics* 7, pp. 243-257.

POWELL J., (1994). – “Estimation of Semi-Parametric Models”, In *Handbook of Econometrics*, Vol. IV, eds. R. Engle and D. McFadden, Amsterdam and New York: North-Holland.

RAO C.R., (1965). – “On Discrete Distribution Arising Out of Methods of Ascertainment”, In *Classical and Contagious Discrete Distributions*, ed. G.P. Patil, Calcutta: Statistical Publishing Society.

RAO C.R., (1985). – “Weighted Distributions Arising Out of Methods of Ascertainment: What Populations Does a Sample Represent?”. In *A Celebration of Statistics*, eds. A. Atkinson and S. Fienberg, New York: Springer-verlag.

RIDDER G., (1990). – “Attrition in Multi-Wave Panel Data” In *Panel Data and Labor Market Studies*, eds. J. Hartog, G. Ridder, and J. Theeuwes, Amsterdam: North-Holland.

RIDDER G., (1992). – “An Empirical Evaluation of Some Models for Non-Random Attribution in Panel Data”, *Mimeographed*, University of Groningen.

VAN DER BERG G., LINDEBOOM M., RIDDER G., (1994). – “ Attrition in Longitudinal Panel Data and the Empirical Analysis of Dynamic Labour Market Behavior”. *Journal of Applied Econometrics* 99, pp. 421-435.

VERBEEK M., NIJMAN T., (1996). – “Incomplete Panels and Selection Bias”. In *The Econometrics of Panel Data*, eds. L. Matyas and P. Sevestre, Dordrecht: Kluwer.

WOOLDRIDGE J., (1997). – “Asymptotic Properties of Weighted M-Estimators for Variable Probability Samples”, *Mimeographed*, Michigan State University.