INTERNATIONAL ECONOMIC REVIEW
Vol. 33, No. 3, August 1992

TESTING FOR SELECTIVITY BIAS IN PANEL DATA MODELS*

BY MARNO VERBEEK AND THEO NIJMAN[1]

We discuss several tests to check for the presence of selectivity bias in estimators based on panel data. One approach to test for selectivity bias is to specify the selection mechanism explicitly and estimate it jointly with the model of interest. Alternatively, one can derive the asymptotically efficient LM test. Both approaches are computationally demanding. In this paper, we propose the use of simple variable addition and (quasi-) Hausman tests for selectivity bias that do not require any knowledge of the response process. We compare the power of these tests with the asymptotically efficient test using Monte Carlo methods.

1. INTRODUCTION

Missing observations are a rule rather than an exception in panel data sets. It is common practice in applied economic analysis of panel data to analyze only the observations on units for which a complete time series is available. Since the seminal contributions of Heckman (1976, 1979) and Hausman and Wise (1979) it is well known that inferences based on either the balanced sub-panel (with the complete observations only) or the unbalanced panel without correcting for selectivity bias, may be subject to bias if the nonresponse is endogenously determined. Even if the response process is known, estimation of the full model including a response equation explaining the missing observations, is, in general, rather cumbersome (compare Ridder 1990, Verbeek 1990). Therefore, it is worthwhile to have some simple tests to check for the presence of selectivity bias which can be performed first. An obvious choice for such a test is the Lagrange Multiplier test, which requires estimation of the model under the null hypothesis only. As will be shown in this paper, the computation of the LM test statistic is still rather cumbersome and, in addition, its value is highly dependent on the specification of the response mechanism and the distributional assumptions. In this paper we will therefore consider several simpler tests to check for the presence of selectivity bias without the necessity of having to estimate the full model or to specify a response equation. A consequential advantage of these tests is that they can be performed in a simple way in cases with wave nonresponse, where all observations on the variables of the model are missing for some individuals in some periods, as well as item nonresponse, where only information on the endogenous variable is missing.

For ease of presentation we will in this paper restrict attention to the linear

* Manuscript received March 1990; final revision received September 1991.

[1] Helpful comments of Bertrand Melenberg, Arie Kapteyn, Peter Kooreman, Arthur van Soest and two anonymous referees are gratefully acknowledged. The authors have benefitted from financial support of the Netherlands Organization for Scientific Research (N.W.O.) and the Royal Netherlands Academy of Arts and Sciences (K.N.A.W.), respectively.

regression model, although several of the tests can straightforwardly be generalized to nonlinear models. Consider

$$y_{it} = x_{it}\beta + \alpha_i + \varepsilon_{it}, \quad t = 1, \ldots, T; \quad i = 1, \ldots, N, \tag{1}$$

where x_{it} is a k dimensional row vector of exogenous variables relating to the ith cross sectional unit at period t, β is a column vector of unknown parameters of interest, α_i and ε_{it} are unobserved i.i.d. random variables with expectation zero and variance σ_α^2 and σ_ε^2, respectively, which are mutually independent. The variables in x_{it} are assumed to be strictly exogenous, i.e., $E\{\varepsilon_{is}|x_{it}\} = 0$ for all i, s, t and $E\{\alpha_i|x_{it}\} = 0$ for all i, t. For simplicity we assume that the model does not contain an intercept term and that means have been removed from all data. T and N denote the number of periods and the number of cross sectional units (individuals, households, firms) in the panel, respectively.

Whether or not observations for y_{it} are available is denoted by the dummy variable r_{it}, such that $r_{it} = 1$ if y_{it} is observed and $r_{it} = 0$ otherwise. In addition, we define $c_i = \Pi_{t=1}^{T} r_{it}$, so that $c_i = 1$ if and only if y_{it} is observed for all t. Observations on x_{it} are assumed to be available when $r_{it} = 1$. A commonly used assumption to describe the process generating r_{it} is based on a latent variable specification. In that case, r_{it} is determined by the sign of r_{it}^*, given by, for example,

$$r_{it}^* = z_{it}\gamma + \xi_i^* + \eta_{it}, \quad t = 1, \ldots, T; \quad i = 1, \ldots, N, \tag{2}$$

with z_{it} a row vector of exogenous variables, possibly containing (partly) the same variables as x_{it}, and η_{it} an unobserved random variable. The term ξ_i^* accounts for unobserved time-invariant individual-specific effects. Now, $r_{it} = 1$ if $r_{it}^* > 0$ and zero otherwise. For the moment however, we shall not use additional assumptions on the process that determines r_{it}. Only in Section 4, where the LM test is discussed, we shall assume that specification (2) holds.

When estimating β in (1) using the available observations one is implicitly conditioning upon the outcome of the selection process, i.e., upon $r_{it} = 1$. The problem of selectivity bias arises from the fact that this conditioning may affect the unobserved determinants of y_{it}, in particular, this may occur if the indicator variable r_{it} is not independent of the individual effect α_i or the error term ε_{it}. Similar problems arise if one concentrates attention to the complete observations only, i.e., to those cross-sectional units for which a complete time series is available (forming a balanced sub-panel). In this case one is implicitly conditioning upon $c_i = 1$ ($r_{i1} = \ldots r_{iT} = 1$).

In this paper attention will be paid to several simple testing procedures that can be used to check whether selectivity bias is seriously present. First, in Section 2, we analyze two well known estimators, the fixed effects (FE) and the random effects (RE) estimator, and discuss the conditions for no selectivity bias in these estimators. It appears that the condition that r_{it} is independent of both α_i and ε_{it} in (1) is not necessary (though sufficient) for consistency. Moreover, it is shown that the fixed effects estimator is more robust for selectivity bias than the random effects estimator. Section 3 shows how differences between the FE and RE estimators

from a balanced and unbalanced design can be used to construct simple (quasi-) Hausman tests of selectivity bias. Moreover, some simple variable addition tests are suggested. Neither of these tests does require knowledge of the process that determines r_{it}.

In Section 4 we introduce and specify a latent variable specification to describe the selection process r_{it}. If this description is correct and data are available to identify its unknown parameters, the Lagrange Multiplier test for independence of r_{it} and $\alpha_i + \varepsilon_{it}$ can be computed and is asymptotically efficient. Moreover, it is possible to use a two step estimation and testing procedure based on the results of Heckman (1976, 1979). Both of these tests are computationally not very attractive. To illustrate the findings of Section 2 and to obtain some idea about the power of the tests proposed in Section 3, we perform a Monte Carlo study, the results of which are reported in Sections 5 and 6. Finally, Section 7 contains some concluding remarks.

2. SELECTIVITY BIAS IN THE FIXED AND RANDOM EFFECTS ESTIMATORS

In this section we derive conditions for consistency of the fixed effects (or "within") estimator for the regression coefficients β in (1). Subsequently, we consider the random effects estimator. Since most panel data sets are characterized by a large number of cross sectional observations covering a fairly short time period, we shall concentrate on consistency for $N \to \infty$ and keep T fixed. T is assumed to be strictly larger than one.

If we define $\tilde{x}_{it}$ as the value of x_{it} in deviation from its (observed) individual mean, i.e.

$$\tilde{x}_{it} = x_{it} - \sum_{s=1}^{T} x_{is} r_{is} \Big/ \sum_{s=1}^{T} r_{is}, \quad \text{if } \sum_{s=1}^{T} r_{is} > 0 \tag{3}$$

$$= 0 \qquad \text{otherwise,}$$

and analogously for $\tilde{y}_{it}$, the FE estimator based on the unbalanced panel is given by (compare Hsiao 1986, p. 31)[2]

$$\hat{\beta}_{FE}(U) = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{x}_{it}' \tilde{x}_{it} r_{it} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{x}_{it}' \tilde{y}_{it} r_{it} \right) \tag{4}$$

and the one based on the balanced sub-panel by

$$\hat{\beta}_{FE}(B) = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{x}_{it}' \tilde{x}_{it} c_i \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{x}_{it}' \tilde{y}_{it} c_i \right). \tag{5}$$

[2] This estimator is only defined if at least one individual is observed more than once; for finite samples there will generally be a small but nonzero probability that this is not the case, but for practical purposes this can be ignored. Similar remarks hold for all other estimators presented below.

Obviously, $\hat{\beta}_{FE}(.)$ is unbiased and consistent[3] for β if selection is determined independently of α_i and ε_{it}. Using $\tilde{y}_{it} = \tilde{x}_{it}\beta + \tilde{\varepsilon}_{it}$, one immediately sees that this condition is too strong, since independence of $r_i = (r_{i1}, \ldots, r_{iT})'$ and the transformed error term $\tilde{\varepsilon}_{it}$ also guarantees unbiasedness and consistency. It is straightforward to show that an even weaker condition for consistency of $\hat{\beta}_{FE}(U)$ and $\hat{\beta}_{FE}(B)$ is that[4]

$$E\{\tilde{\varepsilon}_{it}|r_i\}r_{it} = 0, \quad t = 1, \ldots, T; \quad i = 1, \ldots, N \tag{6}$$

or

$$E\{\tilde{\varepsilon}_{it}|c_i\}c_i = 0, \quad t = 1, \ldots, T; \quad i = 1, \ldots, N, \tag{7}$$

respectively. Consequently, a sufficient condition[5] for both conditions (6) and (7) to hold is that

$$E\{\tilde{\varepsilon}_{it}|r_i\} = 0 \quad t = 1, \ldots, T; \quad i = 1, \ldots, N. \tag{8}$$

First of all, it should be noted that (8) does not involve α_i. Thus, the fact that selection (indicated by r_{it}) depends upon the individual effects α_i in the model of interest does not introduce a selectivity bias in the fixed effects estimators. In addition, if selection affects the conditional expectation of each of the error terms $\varepsilon_{i1}, \ldots, \varepsilon_{iT}$ in the same way, selectivity bias will also not occur. In all these cases selectivity may have an effect on the structural equation (1), but since this effect is fixed for a given individual over all periods in which its dependent variable is observed, it is absorbed in the fixed effect and no consistency problems arise for the FE estimator. In Section 4 some more attention to condition (8) will be paid in the context of the latent variable equation (2) explaining r_{it}.

Next we consider the random effects estimator (compare Hsiao 1986, p. 34 ff.). First, we stack the observations for each cross sectional unit into vectors and matrices, i.e.

$$y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{pmatrix}, \quad X_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{iT} \end{pmatrix}, \quad \varepsilon_i = \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iT} \end{pmatrix}.$$

Let T_i denote the number of periods unit i is observed, i.e. $T_i = \sum_{t=1}^{T} r_{it}$. For each cross sectional unit we define a $T_i \times T$ matrix R_i transforming y_i into the T_i-dimensional vector of observed values y_i^{obs}, say. This matrix R_i is obtained by deleting the rows of the T-dimensional identity matrix corresponding to the unobserved elements. Now we can write $y_i^{obs} = R_i y_i$. Defining $\iota = (1, 1, \ldots, 1)'$

[3] Throughout the paper, we assume that the usual regularity conditions are met.

[4] The conditional expectations in the sequel are also conditional on the exogenous variables, but for the sake of notation these are omitted.

[5] A case in which this sufficient condition is not necessarily met, but condition (6) holds, is the situation where observations are missing deterministically (given x_{it}) ($E\{r_{it}|x_{it}\} = r_{it} = 0$), for example, if being on vacation implies nonresponse.

of dimension T, the variance covariance matrix of the error term in (1) can be written as

$$\Omega = V\{\iota\alpha_i + \varepsilon_i\} = \sigma_\alpha^2 \iota\iota' + \sigma_\varepsilon^2 I.$$

Writing $\Omega_i = R_i \Omega R_i'$ and $X_i^{obs} = R_i X_i$, the random effects estimator based on the unbalanced panel is given by[6]

$$\hat{\beta}_{RE}(U) = \left(\sum_{i=1}^{N} X_i^{obs'} (\Omega_i)^{-1} X_i^{obs} \right)^{-1} \left(\sum_{i=1}^{N} X_i^{obs'} (\Omega_i)^{-1} y_i^{obs} \right). \tag{9}$$

If only the complete observations in the panel are used the random effects estimator is given by

$$\hat{\beta}_{RE}(B) = \left(\sum_{i=1}^{N} X_i' \Omega^{-1} X_i c_i \right)^{-1} \left(\sum_{i=1}^{N} X_i' \Omega^{-1} y_i c_i \right), \tag{10}$$

Note that these estimators can easily be computed using OLS on transformed data even if the unbalanced panel is used (see, e.g., Baltagi 1985 or Wansbeek and Kapteyn 1989).

The estimators $\hat{\beta}_{RE}(.)$ are consistent if

$$E\{\alpha_i + \varepsilon_{it} | r_i\} = 0, \quad t = 1, \ldots, T; \quad i = 1, \ldots, N. \tag{11}$$

Clearly, this condition is stronger than condition (8) needed for consistency of the fixed effects estimator and consequently, we can conclude that the fixed effects estimator is more robust with respect to nonrandom selectivity than the random effects estimator. This may be a reason to prefer the fixed effects estimator although of course some efficiency is lost by this choice if in fact condition (11) holds. Assuming normality of the error terms in (1) and a probit model to describe the selection process r_{it}, this point is further elaborated in Section 4.

Before we propose several simple tests to check for the presence of selectivity bias, it is important to note two things. First, the conditions for consistency of the fixed effects and random effects estimators are different and, second, there is no reason why the inconsistencies in estimators based on the balanced sub-panel and those on the unbalanced panel would coincide. These two points enable us to construct tests for the presence of selectivity bias (or, in fact, for consistency of the FE of RE estimators) using only the four simple estimators presented above. This will be the main theme of the next section.

3. SIMPLE TESTS FOR SELECTIVITY BIAS

In Section 2 four estimators of β have been presented which are all consistent in the absence of nonrandom selection (i.e. if r_{it} is independent of α_i and ε_{it}), namely

[6] For expository purposes we ignore the fact that in practice unknown variances have to be replaced by consistent estimates.

the fixed effects estimators based on the balanced sub-panel and the unbalanced panel and the random effects estimators based on the balanced and unbalanced panel. In general, it is quite unlikely that the pseudo true values, i.e. the probability limits under the true data generating process, of either two estimators are identical, unless both estimators are consistent. Therefore, it is possible to construct a test for selectivity bias based on the differences between either two, three or four estimators.

Let us stack all four estimators into a $4k$ dimensional vector $\hat{\beta}$ as follows.

$$\hat{\beta} = (\hat{\beta}_{FE}(B)', \hat{\beta}_{FE}(U)', \hat{\beta}_{RE}(B)', \hat{\beta}_{RE}(U)')'. \tag{12}$$

Under weak regularity assumptions $\hat{\beta}$ is asymptotically normally distributed according to

$$\sqrt{N}(\hat{\beta} - \bar{\beta}) \xrightarrow{L} N(0, V), \tag{13}$$

where $\bar{\beta}$ denotes the vector of pseudo true values. From (13) it immediately follows that the hypothesis $D\bar{\beta} = 0$ can be tested using

$$\xi_D = N\hat{\beta}'D'(D\hat{V}D')^{-}D\hat{\beta}, \tag{14}$$

which is asymptotically distributed as a central Chi-square with d degrees of freedom under the null hypothesis $D\bar{\beta} = 0$, where A^{-} denotes a generalized inverse of A and d is the rank of DVD'.

In order to be able to compute the test statistics in (14) for appropriate choices of D, an estimator for the full matrix V is needed. Using the definitions of the four estimators given in (4), (5), (9) and (10), it is a straightforward exercise to determine their variances and their covariances. Denoting $V_{11} = V\{\hat{\beta}_{FE}(B)\}$, $V_{22} = V\{\hat{\beta}_{FE}(U)\}$, $V_{33} = V\{\hat{\beta}_{RE}(B)\}$ and $V_{44} = V\{\hat{\beta}_{RE}(U)\}$, it follows that all blocks in the matrix V are a function of the variance covariance matrices of the four estimators in $\hat{\beta}$ only. In particular, it holds that

$$V = \begin{pmatrix} V_{11} & V_{22} & V_{33} & V_{44} \\ & V_{22} & V_{22}V_{11}^{-1}V_{33} & V_{44} \\ & & V_{33} & V_{44} \\ & & & V_{44} \end{pmatrix}. \tag{15}$$

Using (15) any test statistic given in (14) can easily be computed. Two obvious candidates from the tests that compare two out of four possible estimators, are those comparing the fixed or random effects estimators from the balanced sub-panel and the unbalanced panel, where $D = D_1 = [I \; -I \; 0 \; 0]$ or $D = D_2 = [0 \; 0 \; I \; -I]$, respectively. Two other choices, $D_3 = [I \; 0 \; -I \; 0]$ and $D_4 = [0 \; I \; 0 \; -I]$, result in the standard Hausman specification test for uncorrelated individual effects (see, e.g., Hsiao 1986, p. 48) and its generalization to an unbalanced panel, respectively. A fifth test compares the FE estimator in the balanced sub-panel and the RE estimator in the unbalanced panel ($D_5 = [I \; 0 \; 0 \; -I]$), while for the last possible test $D_6 = [0 \; I \; -I \; 0]$. Obviously, alternative tests which compare three or more estimators of β are possible.

Since the tests proposed above are based on the comparison of two estimators for the same parameter vector and since some special cases correspond to well known Hausman tests in the literature we shall refer to them as (quasi-) Hausman tests. Unlike in the standard case our tests are based on estimators which are all inconsistent under the alternative. In the very unlikely case where all estimators would have identical asymptotic biases these tests will have no power at all. Keeping this in mind the null hypotheses ($H_0^i: D_i\bar{\beta} = 0$) of the tests above can be translated into hypotheses in terms of estimator consistency.

Let us define

$$H_0^{FE}: E\{\tilde{\varepsilon}_{it}|r_i\} = 0 \text{ (the fixed effects estimators are consistent),}$$

and

$$H_0^{RE}: E\{\alpha_i + \varepsilon_{it}|r_i\} = 0 \text{ (the RE and FE estimators are consistent).}$$

The null hypothesis (denoted by H_0) of nonrandom selection, i.e. the hypothesis that r_{it} and α_i and ε_{it} are independent, is the strongest hypothesis (since it implies all the others). However, for conducting inferences it is not relevant whether H_0 is true or not, but whether H_0^{RE} or H_0^{FE} are correct, since inferences will be based on either the random effects or the fixed effects estimator. Notice that the latter hypothesis is implied by the former, i.e. whenever the random effects estimator is consistent, the fixed effects estimator is consistent as well. The (quasi-) Hausman tests may be appropriate instruments for checking the consistency of these estimators, although they are only able to test for the weaker hypotheses H_0^i. Consequently, a rejection of H_0^i (for some $i = 1, \ldots, 6$) by the corresponding Hausman test, implies that H_0^{RE} should be rejected. If H_0^1 is rejected, H_0^{FE} should be rejected as well. However, the converse is not true.

Note that if both H_0^{RE} and H_0^{FE} are false, all estimators are inconsistent. In that case knowledge of the selection process can be used to model selection simultaneously with model (1) to obtain consistent random effects or fixed effects estimators correcting for selectivity. However, the joint estimation of a selection process and model (1) may be computationally demanding, unless some simplifying distributional assumptions are made. See, for example, Ridder (1990) or Verbeek (1990). In addition, the restrictions needed to identify β may be stronger than one would like, while the resulting estimates will depend heavily on the available prior information (compare Manski 1989, 1990).

Note that only the first test statistic (based on D_1) is appropriate for checking H_0^{FE}, while any other test statistic can be used for H_0^{RE}. The optimal testing procedure seems to be to test for the stronger hypothesis first (H_0^{RE}), and, if this test rejects, test subsequently for the weaker one (H_0^{FE}). Of course, it is preferable to use the most powerful test out of all possible tests for the hypothesis H_0^{RE}. However, the analysis of statistical power is extremely difficult if not impossible, not only because the test statistics are not mutually independent, but also because we are working with Hausman specification tests for which the null hypotheses H_0^i cannot be written down in a simple parametric form. Therefore, standard results on

the power of Hausman tests (compare Holly 1982) and on sequential testing (see, e.g., Mizon 1977, Holly 1987) are not directly applicable in this situation.

Of course alternative tests for selectivity can be constructed. Remember that selectivity bias in model (1) occurs because the conditional expectation of the error term $\alpha_i + \varepsilon_{it}$ does not equal zero. If this conditional expectation $E\{\alpha_i + \varepsilon_{it}|r_i\}$ were known (possibly apart from one or more proportionality factors) one could add it as an extra regressor (or combination of regressors) in (1) such that the new error term has expectation zero (given x_{it} and r_i). Subsequently, the parameters in the extended model can be estimated consistently using standard methods. This is the essence of the well known two step estimation procedure in the cross sectional sample selection model proposed by Heckman (1976, 1979) and the simple two step estimators for models with censored endogenous regressors and sample selection suggested by Vella (1990). An application for the case of nonresponse in panel data is presented by Nijman and Verbeek (1990).

Of course, the conditional expectation $E\{\alpha_i + \varepsilon_{it}|r_i\}$ is not known (or identifiable) unless the selection process is known (or identifiable), and therefore this procedure will have the same drawbacks as joint estimation of the model and the selection process, although the computational burden may be somewhat less. As a testing procedure it may be worthwhile to try to approximate the conditional expectation in a simple way and to check whether it enters model (1) significantly. Since $E\{\alpha_i + \varepsilon_{it}|r_i\}$ will be a function of r_i, the functional form of which depends upon the joint distribution of $\alpha_i + \varepsilon_{it}$ and r_i, one can think of two more or less distinct ways of approximating it. Firstly, one can have one or more variables, z_{it}, say, that are likely to determine the probability of selection (i.e. affect the distribution of r_i), and enter these variables in a convenient form, for example as a low order polynomial. The resulting test would then be a joint test of the hypothesis that, conditional on x_{it}, y_{it} does not depend on (this function of) z_{it} and the hypothesis of no selectivity bias. Alternatively, one can choose some function of r_i itself, from which it is known that it should not enter the model significantly under the hypothesis of no selectivity bias. The resulting test is a test of the selectivity bias hypothesis only. In the sequel we shall concentrate on this second approach and consider three possible variables that can be included in the regression equation. First, $T_i = \sum_{s=1}^{T} r_{is}$, the number of waves individual i participates, second $c_i = \prod_{s=1}^{T} r_{is}$, a 0-1 variable equal to 1 if and only if individual i is observed in all periods and third, $r_{i,t-1}$, indicating whether individual i is observed in the previous period or not. Note that $r_{i,0} = 0$ by assumption. To test the significance of these variables in (1) we are forced to use the unbalanced panel since in the balanced panel the added variables are identical for all individuals and thus incorporated in the intercept term. Since the additional variables are constant over time for each individual in the first two cases, the corresponding parameters are not identified in the case where the individual effects α_i are treated as fixed. We shall therefore concentrate attention to random effects estimators.

Although one could expect that the added variables have an influence on the relationship between y_{it} and x_{it} if there is selective nonresponse, there is no reason why this effect would be linear and thus the power of the tests may be doubtful. If we denote the coefficient for the added variable w, say, by γ_w then the null

hypothesis of the variable addition test is $H_0^w: \gamma_w = 0$. Note that H_0 implies H_0^w but that the converse is not true.

4. SPECIFICATION OF THE RESPONSE MECHANISM AND THE LM TEST FOR SELECTIVITY BIAS

In this section we assume that response r_{it} is determined by a random effects probit model, an assumption which is often made in empirical applications (compare Hausman and Wise 1979, Nijman and Verbeek 1990, Ridder 1990). Under this assumption and assuming normality of the error terms in (1) it is possible to derive the LM test statistic for the null hypothesis that r_{it} is independent of the unobserved determinants of y_{it} (α_i and ε_{it}). Furthermore, we pay some more attention to the conditions for consistency of the FE and RE estimators in the context of this example.

Suppose r_{it} is determined by the sign of a latent variable r_{it}^*, which is generated by

$$r_{it}^* = z_{it}\gamma + \xi_i^* + \eta_{it}, \quad t = 1, \ldots, T; \quad i = 1, \ldots, N, \tag{16}$$

where z_{it} is a row vector of exogenous variables, usually containing partly the same variables as x_{it}, η_{it} denotes an unobserved random variable and ξ_i^* is an individual-specific effect. In order to account for possible correlation between ξ_i^* and the explanatory variables z_{it}, we follow Chamberlain (1984) in assuming that,

$$\xi_i^* = z_{i1}\pi_1 + z_{i2}\pi_2 + \cdots + z_{iT}\pi_T + \xi_i, \tag{17}$$

where ξ_i is independent of all z_{it}'s. Substitution in (16) yields

$$r_{it}^* = z_{it}\gamma + z_{i1}\pi_1 + z_{i2}\pi_2 + \cdots + z_{iT}\pi_T + \xi_i + \eta_{it}. \tag{18}$$

To be able to identify the parameters in (18) it is essential to assume that observations on z_{it} are available for both $r_{it} = 1$ and $r_{it} = 0$. Note that this assumption is not required when performing the (quasi-) Hausman tests or variable addition tests proposed in Section 3. The unobserved random variables in (1) and (18) are assumed to be normally distributed according to

$$\begin{pmatrix} \varepsilon_i \\ \eta_i \\ \xi_i \\ \alpha_i \end{pmatrix} \sim N\left(0, \begin{pmatrix} \sigma_\varepsilon^2 I & & & \\ \sigma_{\varepsilon\eta} I & \sigma_\eta^2 I & & \\ 0 & 0 & \sigma_\xi^2 & \\ 0 & 0 & \sigma_{\alpha\xi} & \sigma_\alpha^2 \end{pmatrix}\right), \tag{19}$$

where $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iT})'$ and $\eta_i = (\eta_{i1}, \ldots, \eta_{iT})'$. For identification of the probit model we will impose (as usual) $\sigma_\eta^2 + \sigma_\xi^2 = 1$. Of course, one can test the model assumptions implied by (18) and (19) along the lines discussed in, for example, Lee (1984) and Lee and Maddala (1985).

Under these assumptions the expectation of ε_{it} given selection is given by (see the Appendix)

$$(20)\quad E\{\varepsilon_{it}|r_i\} = \frac{\sigma_{\varepsilon\eta}}{\sigma_\eta^2}\left(E\{\xi_i + \eta_{it}|r_i\} - \frac{\sigma_\xi^2}{\sigma_\eta^2 + T\sigma_\xi^2}\sum_{s=1}^{T} E\{\xi_i + \eta_{is}|r_i\}\right),$$

while the conditional expectation of α_i given selection is given by (see the Appendix)

$$(21)\quad E\{\alpha_i|r_i\} = \frac{\sigma_{\alpha\xi}}{\sigma_\eta^2 + T\sigma_\xi^2}\sum_{s=1}^{T} E\{\xi_i + \eta_{is}|r_i\}.$$

The conditional expectation $E\{\xi_i + \eta_{it}|r_i\}$ is a complicated function (see the Appendix) of the variables in z_{it} and reduces to "Heckman's (1979) lambda" if there is no individual effect in the probit model ($\sigma_\xi^2 = 0$).

Under the normality assumption the independence of r_i and $(\alpha_i, \varepsilon_{it})$ is equivalent to $\sigma_{\alpha\xi} = \sigma_{\varepsilon\eta} = 0$. Clearly, this condition implies that (11) holds, implying consistency of both the random effects and the fixed effects estimators. For the transformed error term $\tilde{\varepsilon}_{it}$ (20) implies that

$$(22)\quad E\{\tilde{\varepsilon}_{it}|r_i\} = \frac{\sigma_{\varepsilon\eta}}{\sigma_\eta^2}\left(E\{\xi_i + \eta_{it}|r_i\} - \sum_{s=1}^{T} r_{is}E\{\xi_i + \eta_{is}|r_i\}\Big/\sum_{s=1}^{T} r_{is}\right).$$

From this it immediately follows that condition (8) is fulfilled and the fixed effects estimator is consistent if either $\sigma_{\varepsilon\eta} = 0$ or $E\{\xi_i + \eta_{it}|r_i\}$ does not vary over time. The latter condition implies (see the Appendix) that there is no selectivity bias if the probability of an individual of being observed is constant over time, even if $\sigma_{\varepsilon\eta} \neq 0$. This will occur when $z_{it}\gamma$ is constant over time. Since (22) does not contain $\sigma_{\alpha\xi}$, a correlation between the individual effects in the structural equation (1) and the probit equation (18) does not result in a bias in the fixed effects estimator.

The condition that $E\{\xi_i + \eta_{it}|r_i\}$ does not vary with t is clearly not sufficient for consistency of $\hat{\beta}_{RE}$. For the latter we either need that $E\{\xi_i + \eta_{it}|r_i\}$ is constant and $T \to \infty$ (since the FE-estimator and the RE-estimator are equivalent when T tends to infinity)[7] or that $E\{\xi_i + \eta_{it}|r_i\}$ is constant and $\sigma_{\alpha\xi} + \sigma_{\varepsilon\eta} = 0$, which does not seem to be very likely in practice.

The actual magnitude of the inconsistencies of the estimators is determined by the projection of the conditional expectations derived above on the (transformed) x_{it}'s. Although it is possible to analyze the effects of changes in model parameters on the conditional expectation of the (transformed) error term analytically (compare Ridder 1990), it is, in general, virtually impossible to give analytical expressions in terms of the model parameters for projections of these expectations on the explanatory variables, i.e. of the biases in the estimators. To obtain some insight in the numerical importance of the bias in the four estimators discussed above, we will present some numerical results in the next section.

Given the model in (1) and (18) and the assumed normality of the error terms in (19) is it possible to write down the likelihood function (compare Ridder 1990) and

[7] This equivalence also holds when the model is not correctly specified, as in our case.

to derive the Lagrange Multiplier test statistic for the null hypothesis $H_0: \sigma_{\varepsilon\eta} = \sigma_{\alpha\xi} = 0$. The loglikelihood function involves the joint distribution of the observed y-values in y_i^{obs} and the response indicator r_i. In particular, the loglikelihood contribution of individual i is given by

$$L_i = \log f(y_i^{obs}, r_i) = \log f(r_i | y_i^{obs}) + \log f(y_i^{obs}), \tag{23}$$

where we are using $f(.)$ as generic notation for any density/mass function. The second term in the right-hand side of (23) is the log of a T_i-variate normal density function, while the first term is the loglikelihood function of a (conditional) T-variate probit model (see the Appendix for details).

Denoting the full vector of parameters involved in (23) (including $\sigma_{\alpha\xi}$ and $\sigma_{\varepsilon\eta}$) by θ, the Lagrange Multiplier test statistic is given by

$$\xi_{LM} = \sum_{i=1}^{N} \frac{\partial L_i'}{\partial \theta} \left(\sum_{i=1}^{N} \frac{\partial L_i}{\partial \theta} \frac{\partial L_i'}{\partial \theta} \right)^{-1} \sum_{i=1}^{N} \frac{\partial L_i'}{\partial \theta} \Bigg|_{\theta = \hat{\theta}_0} \tag{24}$$

where $\hat{\theta}_0$ is the ML estimate for θ under $H_0: \sigma_{\alpha\xi} = \sigma_{\varepsilon\eta} = 0$. Since there does not appear to be any form of block diagonality of the Fisher information matrix under the null, the scores with respect to all parameters in the model are required to compute this test statistic from the first derivatives of the loglikelihood. For the cross sectional case the LM test for selectivity is discussed in Lee and Maddala (1985).

Because under H_0 the two terms in the right-hand side of (23) depend on nonoverlapping subsets of the vector of parameters, the score contributions with respect to the parameters in (1) can be found in Hsiao (1986, p. 39),[8] while those for the parameters in (18) can be derived from a standard random effects probit likelihood (see the Appendix). The most difficult score contributions are those with respect to the two covariances $\sigma_{\alpha\xi}$ and $\sigma_{\varepsilon\eta}$; the latter even requires double numerical integration (see the Appendix). Because estimation under H_0 requires numerical integration (for each individual) for the probit part of the model and computation of each score contribution also requires numerical integration over one or two dimensions, the LM test is rather unattractive in applied work.

For the cross sectional sample selection model Heckman (1976, 1979) proposed a simple way to test for selectivity bias and to obtain consistent estimators. As discussed in Ridder (1990) this method can be generalized to the case of panel data, where two correction terms to equation (1) are added instead of just the one variable known as Heckman's lambda (or the inverted Mill's ratio). These two correction terms are the conditional expectations of the two error terms (α_i and ε_{it}) given the sampling scheme, as given in (20) and (21) evaluated at the (consistent) parameter estimates of the probit model under the null hypothesis (see Nijman and Verbeek 1990 for an application). The two unknown covariances $\sigma_{\alpha\xi}$ and $\sigma_{\varepsilon\eta}$ are not included in these correction terms but are the corresponding true coefficients in

[8] Note that (3.3.20) in Hsiao (1986) contains a printing error: the first $-$ sign on the second line should read a $+$ sign.

equation (1). Obviously, consistent estimation of these coefficients $\sigma_{\alpha\xi}$ and $\sigma_{\varepsilon\eta}$ allows one to check whether nonresponse is selective or not. Since estimation of the parameters in the response equation as well as computation of the conditional expectation of $\xi_i + \eta_{it}$ in (20) and (21) requires numerical integration, these generalized Heckman (1979) method is still computationally unattractive. Therefore, it may be worthwhile to have some simple variables that can be used instead to approximate the true correction terms to check for selective nonresponse, for example those suggested in the previous section.

If the specification of the response process in (18) is correct, the Lagrange Multiplier test is known to be asymptotically efficient for testing the null hypothesis H_0. To obtain some idea about the power of the alternative simple tests we performed a Monte Carlo study under this assumption, the results of which are presented in the next two sections. In Section 5 we introduce the Monte Carlo model and present estimates for the pseudo true values of the four estimators in (12), giving insight in the importance of the selectivity bias in these estimators. In Section 6 some numerical results on the power of the simple tests in comparison with the Lagrange Multiplier test are presented.

5. NUMERICAL RESULTS ON THE PSEUDO TRUE VALUES OF THE RE AND FE ESTIMATORS

In this section we will present some numerical results on the pseudo true values of $\hat{\beta}_{FE}$ and $\hat{\beta}_{RE}$, defined as the probability limits of these estimators under the true data generating process. For expository purposes we consider a simple model consisting of equations (1) and (18) with only one exogenous variable included besides the constant term.

This exogenous variable ($z_{it} = x_{it}$) is assumed to be generated by a Gaussian AR(1) process with mean zero, autocorrelation coefficient ρ_x and variance σ_x^2. For simplicity we have imposed equality of all π_t's in (17). The model used for simulation is thus given by

$$y_{it} = \beta_1 x_{it} + \alpha_i + \varepsilon_{it} \tag{25}$$

$$r_{it}^* = \gamma_0 + \gamma_1 x_{it} + \pi \bar{x}_i + \xi_i + \eta_{it} \tag{26}$$

where $\bar{x}_i$ is the average value of the x_{it}'s over time. We concentrate on a model with only one explanatory variable, since it elucidates the discussion most clearly. Including an additional variable in (25) that is uncorrelated with x_{it} essentially would not change the results, while inclusion of a variable that is correlated with x_{it} would result in biases that depend heavily on the sign and magnitude of this correlation. Similar remarks hold for the inclusion of additional variables in (26).

We consider two possible specifications for the selection equation, one in which π is a priori set to zero (in which case the probability of selection in period t is determined by x_{it}), and one in which γ_1 is a priori set to zero such that the average value of x_{it} over time determines the selection probability. Given this choice of specification, the relative biases of the estimators for β_1 in this model, defined as $(\bar{\beta}_1 - \beta_1)/\beta_1$, where $\bar{\beta}_1$ is the pseudo true value of the respective estimators for β_1,

TABLE 1
RELATIVE INCONSISTENCIES (IN PERCENT) IN THE FE AND RE ESTIMATORS FROM A BALANCED AND UNBALANCED PANEL

Reference situation (REF): $T = 3$, $R_y^2 = 0.1$, $R_r^2 = 0.9$, $\rho_\alpha = 0.1$, $\rho_x = 0.7$, $p_0 = 0.5$, $\rho_\xi = 0.1$ and $\rho_{\alpha\xi} = 0.5$

estimator	REF	$R_y^2 = 0.9$	$R_r^2 = 0.1$	$\rho_\alpha = 0.9$	$\rho_x = 0.3$	$p_0 = \Phi(1)$	$\rho_\xi = 0.9$	$\rho_{\alpha\xi} = 0.9$
A:				$\pi = 0$, $\rho_{\varepsilon\eta} = 0.9$				
FE(B)	−78	−8	−49	−25	−90	−61	−28	−77
RE(B)	−79	−9	−49	−27	−93	−61	−39	−81
FE(U)	−98	−10	−50	−33	−101	−77	−37	−98
RE(U)	−116	−13	−53	−39	−115	−88	−56	−121
B:				$\pi = 0$, $\rho_{\varepsilon\eta} = 0$				
RE(B)	−6	−1	−5	−2	−6	−6	−17	−11
RE(U)	−6	−1	−4	−6	−7	−5	−19	−12
C:				$\gamma_1 = 0$, $\rho_{\varepsilon\eta} = 0.9$				
RE(B)	−34	−3	−38	−1	−17	−27	−17	−35
RE(U)	−74	−7	−44	−4	−41	−61	−32	−75

1. Relative inconsistency of an estimator is defined as its pseudo true value minus the true value divided by the true value (multiplied by 100 percent).
2. The number of replications in each situation is chosen such that all (Monte Carlo) standard errors are smaller than 0.5 percent.
3. All simulation results are obtained using the NAG-library subroutines G05CCF and G05DDF.
4. From the analytical results we know that the fixed effects estimators are consistent in panels B and C, which was confirmed by the Monte Carlo results.

depend on T, the number of time periods, and the following eight hyperparameters.

$\rho_\alpha = \sigma_\alpha^2(\sigma_\alpha^2 + \sigma_\varepsilon^2)^{-1}$, the importance of the individual effect in equation (25);
$\rho_\xi = \sigma_\xi^2$, the importance of the individual effect in the selection equation;
ρ_x, the autocorrelation coefficient of x_{it};
$p_0 = \Phi(\gamma_0)$, the (unconditional) probability of observation when $x_{it} = 0$ for all t;
$R_y^2 = \beta_1^2\sigma_x^2(\beta_1^2\sigma_x^2 + \sigma_\alpha^2 + \sigma_\varepsilon^2)^{-1}$, the (theoretical) R^2 of equation (1);
R_r^2, the (theoretical) R^2 of the selection equation;
$R_r^2 = \gamma_1^2\sigma_x^2(\gamma_1^2\sigma_x^2 + 1)^{-1}$ if $\pi = 0$, or
$R_r^2 = \pi^2\sigma_{\bar{x}}^2(\pi^2\sigma_{\bar{x}}^2 + 1)^{-1}$ if $\gamma_1 = 0$, with $\sigma_{\bar{x}}^2 = \sigma_x^2(3 + 4\rho_x + 2\rho_x^2)/9$ (the variance of $\bar{x}_i$);
$\rho_{\varepsilon\eta} = \sigma_{\varepsilon\eta}/\sigma_\varepsilon\sigma_\eta$, the correlation between the error shocks in (25) and (26);
$\rho_{\alpha\xi}$, the correlation between the individual effects in (25) and (26).

If we assume that all correlations are nonnegative, all of the hyperparameters are restricted to the interval [0, 1], so that one has some more feeling what "small" and "large" values for these parameters mean. Without loss of generality, it is assumed that $\gamma_1 \geq 0$ or $\pi \geq 0$. In Table 1 estimated relative biases (relative differences between the estimated pseudo true values and the true values) of the four estimators discussed above are given for several combinations of parameter values

and $T = 3$. The number of replications is chosen in such a way that all (Monte Carlo) standard errors are smaller than 0.005. In the table the parameter values are chosen as follows. There is one "reference situation" characterized by $T = 3$, $R_y^2 = 0.1$, $R_r^2 = 0.9$, $\rho_\alpha = 0.1$, $\rho_x = 0.7$, $\rho_0 = 0.5$, $\rho_\xi = 0.1$ and $\rho_{\alpha\xi} = 0.5$. Three alternative combinations of π and $\rho_{\varepsilon\eta}$ are considered given in panels A, B and C. The columns in the table correspond to the reference situation (REF) or this situation with only one of the parameter values changed. For example, the column with heading $\rho_x = 0.3$ refers to the reference situation given above with $\rho_x = 0.3$ instead of 0.7. If $\pi = 0$ and $\rho_{\varepsilon\eta} = 0.9$ (panel A) we see in this column that the fixed effects estimator based on the balanced panel suffers from an inconsistency of −90 percent, while the same figure for the random effects estimator from the unbalanced panel is −115 percent. The standard errors implied by the Monte Carlo experiment are such that the true relative inconsistencies are with a 95 percent probability within a 1 percent point range of the reported values.

Although, as always, it is difficult to draw definitive conclusions from results for specific parameter values the results in Table 1 suggest the following points.

The biases in the estimators can be substantial. In some cases it is even possible that the sign of the pseudo true value is opposite to the sign of the true value of β_1. Moreover, like other simulation results (not reported in this paper) suggest, if the true β_1 parameter is equal to zero (which implies that $R_y^2 = 0$), a significant effect of the explanatory variable on y_{it} can be found. This phenomenon is also known from the standard (cross section) sample selection model of Heckman (1979).

Although the fact that the conditions for the fixed effects estimator to be consistent are weaker than those for the random effects estimator does not necessarily imply that the bias in the latter is always larger than that in the first, our simulations show that this is in fact the case. If there is a difference between the RE and FE pseudo true values, it is in favor of the latter estimator. This result is caused by the fact that we have assumed that $\rho_{\alpha\xi} > 0$. In the not very likely situation where $\rho_{\alpha\xi} < 0$ and $\rho_{\varepsilon\eta} > 0$, the bias in the random effects estimator may in fact be smaller. If the amount of bias is used as criterion for choosing an estimator, it is obvious from our analytical and numerical results that the fixed effects estimator is likely to be preferable to the random effects estimator.

For almost all situations we consider, the bias in the estimator based on the unbalanced panel is larger (in absolute value) than that in the same estimator based on the balanced panel; if it is smaller the difference between the two estimates is negligible given the size of the Monte Carlo experiment. This somewhat surprising result suggests that a balanced panel may be preferred to an unbalanced panel. A possible explanation for this result might be that the individuals that are not observed in all periods have on average a lower probability of being observed, thus also a lower probability in those periods they are observed, implying a larger correction term in the regression equation. In the standard sample selection model of Heckman this would mean that for those individuals Heckman's lambda deviates more from zero.

Keeping all parameters fixed at some level except one, it may be possible to say something about the change of the bias if that one parameter is changed. It is evident from the analytical results and also from the numerical results above that a

rise in R_y^2 will cause a decrease in the absolute value of the bias, simply because a rising R_y^2 diminishes the role of the error terms α_i and ε_{it}. On the other hand, a rise in R_r^2 increases the absolute value of the bias, since it increases the correlation between the probabilities of being observed and the explanatory variable(s) x_{it}. For $p_0 \geq \frac{1}{2}(\gamma_0 \geq 0)$, an increase in p_0 diminishes this correlation and therefore decreases the absolute value of the bias. Obviously, increasing the (absolute values of the) correlation coefficients $\rho_{\varepsilon\eta}$ or $\rho_{\alpha\xi}$ (already being nonnegative) causes a rise in the absolute value of the bias of all estimators. A more important individual effect in equation (25), ρ_α, seems to reduce the absolute value of the bias; the effect of ρ_x and ρ_ξ is ambiguous.

6. NUMERICAL RESULTS ON THE POWER OF THE TESTS

In Section 3 a number of tests were proposed which can be used to check whether selectivity bias is present or not. In this section we present numerical results on the power properties of the quasi-Hausman tests, the variable addition tests and the LM test of Section 4 for the Monte Carlo model introduced in Section 5. We shall not consider the generalized Heckman test because it is as hard to compute but asymptotically less powerful than the asymptotically optimal Lagrange Multiplier test.

For simplicity we restrict ourselves to an analysis of the asymptotic local power. That is, we consider the power of our tests under a sequence of local alternatives, in general $\bar{\theta} = \theta_0 + \delta/\sqrt{N}$ for some vector δ, where θ_0 denotes the parameter value under the null hypothesis (compare Engle 1984 or Holly 1987). Under such a sequence of local alternatives our tests (or their χ^2 equivalents) are asymptotically noncentrally χ^2 distributed, with a decentrality parameter Δ determined by δ. For the (quasi-) Hausman tests, for example, and a sequence of local alternatives given by $\bar{\beta} = \beta + \bar{\delta}/\sqrt{N}$ it holds that

$$(27)\quad \xi_R = N\hat{\beta}'R'(R\hat{V}R')^{-}R\hat{\beta} \xrightarrow{L} \chi_d^2(\bar{\delta}'R'(RVR')^{-}R\bar{\delta}) = \chi_d^2(\Delta_R),\ N \to \infty.$$

Since the power of a test is a direct function of its decentrality parameter, we report decentrality parameters only.

We interpret the particular alternative implied by the Monte Carlo model as being one in a sequence of local alternatives. For all cases in the Monte Carlo set-up we choose a sample size[9] of $N = 25{,}000$ to estimate the pseudo true values $\bar{\theta}$ by $\hat{\bar{\theta}}$. We estimate δ by $\hat{\delta} = \sqrt{n}(\hat{\bar{\theta}} - \theta_0)$, which gives us (an estimate for) the decentrality parameter for sample size n. In Table 2 decentrality parameters for $n = 500$ are reported. From these decentrality parameters one can compute the probability of rejection of the null hypothesis in a sample of 500 observations based on an approximation by the asymptotic distribution. Considering, for example, the reference situation in panel A ($\pi = 0$, $\rho_{\varepsilon\eta} = 0.9$), we see that the Hausman test comparing the RE estimators from the unbalanced panel and the balanced sub-

[9] Sample size refers to the number of individuals in the panel, including those that are observed only once or twice.

TABLE 2

DECENTRALITY PARAMETERS OF THE CHI-SQUARE DISTRIBUTIONS OF SEVERAL TESTS FOR SELECTIVITY BIAS AT $n = 500$ AND $T = 3$

Reference situation (REF): $T = 3$, $R_y^2 = 0.1$, $R_r^2 = 0.9$, $\rho_\alpha = 0.1$, $\rho_x = 0.7$, $p_0 = 0.5$, $\rho_\xi = 0.1$ and $\rho_{\alpha\xi} = 0.5$

test	DF	REF	$R_y^2 =$ 0.9	$R_r^2 =$ 0.1	$\rho_\alpha =$ 0.9	$\rho_x =$ 0.3	$p_0 =$ $\Phi(1)$	$\rho_\xi =$ 0.9	$\rho_{\alpha\xi} =$ 0.9
A:					$\pi = 0$, $\rho_{\varepsilon\eta_i} = 0.9$				
Quasi-Hausman tests:									
1	1	1.41	1.27	0.07	2.00	0.31	1.52	0.26	1.05
2	1	7.23	6.00	0.06	3.55	1.53	7.48	1.84	7.33
3	1	0.85	0.72	0.03	1.76	0.60	0.43	1.13	0.72
4	1	2.07	1.81	0.01	3.55	0.85	1.43	1.37	1.66
5	2	2.04	1.64	0.04	2.02	0.89	1.83	1.39	2.49
6	2	7.27	6.04	0.10	4.25	1.69	7.48	2.44	7.34
Variable addition tests:									
7	1	0.01	0.01	0.04	0.14	0.03	0.11	0.10	0.04
8	1	0.03	0.03	0.00	0.24	0.04	0.04	0.17	0.14
9	1	0.02	0.01	0.01	0.02	0.00	0.14	0.03	0.02
Lagrange Multiplier test:									
LM	2	55.1	49.2	5.46	31.3	58.5	57.3	14.1	66.3
B:					$\pi = 0$, $\rho_{\varepsilon\eta} = 0$				
Quasi-Hausman tests:									
2	1	0.07	0.06	0.00	0.02	0.00	0.02	0.00	0.00
3	1	0.12	0.35	0.06	0.72	0.09	0.01	0.81	0.41
4	1	0.06	0.45	0.04	0.18	0.04	0.00	0.81	0.38
5	2	0.17	0.44	0.00	0.79	0.12	0.02	0.98	0.57
6	2	0.15	0.36	0.07	0.73	0.11	0.05	0.84	0.46
Variable addition tests:									
7	1	0.09	0.07	1.88	0.61	0.32	0.22	1.23	0.59
8	1	0.06	0.09	1.31	0.39	0.17	0.21	0.98	0.64
9	1	0.00	0.12	0.16	0.00	0.14	0.04	0.27	0.15
Lagrange Multiplier test:									
LM	2†	1.33	0.13	4.12	4.95	1.06	1.15	5.92	3.74
C:					$\gamma_1 = 0$, $\rho_{\varepsilon\eta} = 0.9$				
Quasi-Hausman tests:									
2	1	19.6	19.4	0.10	1.47	11.4	20.7	8.96	17.7
3	1	19.9	18.3	3.73	6.68	22.4	15.2	4.35	19.3
4	1	16.0	14.7	1.56	2.50	15.1	12.7	3.93	15.3
5	2	30.6	29.3	3.73	7.60	27.1	28.3	11.9	29.2
6	2	29.4	28.4	3.74	6.86	24.5	27.5	11.4	27.9
Variable addition tests:									
7	1	29.9	27.6	0.09	3.92	36.7	27.0	16.0	24.9
8	1	22.7	21.6	0.08	3.16	30.6	21.6	13.9	18.5
9	1	2.80	2.29	0.05	0.04	5.85	2.10	0.59	2.19
Lagrange Multiplier test:									
LM	2	75.9	73.6	13.7	20.1	83.8	66.8	12.1	83.0

1. Fixed Effects (Balanced vs. Unbalanced)
2. Random Effects (Balanced vs. Unbalanced)
3. Unbalanced (Random Effects vs. Fixed Effects)
4. Fixed Effects, Balanced vs. Random Effects, Unbalanced
5. Balanced (FE vs. RE) and Unbalanced (FE vs. RE)
6. RE (Balanced vs. Unbalanced) and FE, Balanced vs. RE, Unbalanced
7. $\Sigma_t\ r_{it}$
8. $\Pi_t\ r_{it}$
9. $r_{i,\,t-1}$

†If the restriction $\rho_{\varepsilon\eta} = 0$ is imposed a priori, this test has one degree of freedom.

Estimated decentrality parameters are based on 25,000 individual observations. Estimates for decentrality parameters for sample size n can be obtained by multiplying the numbers by $n/500$.

TABLE 3
PROBABILITIES OF REJECTION (AT 5 PERCENT) FOR SEVERAL DECENTRALITY PARAMETERS

	Decentrality parameter							
DF	0	1	2	3	4	5	10	20
1	0.05	0.17	0.29	0.41	0.52	0.61	0.89	0.99
2	0.05	0.13	0.23	0.32	0.42	0.50	0.82	0.99

panel has a decentrality parameter of 7.23, implying a 77 percent probability of rejection at a nominal size of 5 percent (if $n = 500$). If the available sample contains 1000 individuals, the decentrality parameter is twice as large (14.46) corresponding to a 97 percent probability of rejection. Similarly, the implied probabilities of rejection (at a nominal size of 5 percent) for six (quasi-) Hausman tests, three variable addition tests and the LM test for any number of observations can be computed using Table 3.

Note that the estimated decentrality parameters in Table 2 are not normally but (noncentrally) Chi-square distributed, which makes computation of confidence intervals difficult. Based on the asymptotic normality of the parameter estimators the variance of $\hat{\Delta}$ approximately satisfies

$$V\{\hat{\Delta}\} = \frac{n^2}{N^2}\left(d + \frac{N}{n}\Delta\right) \tag{28}$$

where d is the number of degrees of freedom, and where we use the fact that $N/n\hat{\Delta}$ is Chi-square distributed. It is important to note that this variance increases with the true value Δ. For large enough Δ the corresponding standard error for $N = 25{,}000$ and $n = 500$ is (approximately) given by $0.283\sqrt{\Delta}$.

Looking at panel A of Table 2 first, where both H_0^{FE} and H_0^{RE} are false, we see that in this case none of the variable addition tests has any power. Obviously, these variables are under these data generating processes not capable of approximating the Heckman (1979) like correction terms. This is probably due to the fact that our simple variables are not capable of capturing the time variation in these correction terms (due to $z_{it}\gamma$). With regard to the Hausman tests, the results in Table 2 suggest that the test based on comparison of the random effects estimators in the balanced and the unbalanced panel (the second test) is more powerful than all other tests based on comparison of two estimators. Looking at the tests that compare two pairs of estimators (the fifth and the sixth test in Table 2), the latter seems to perform relatively well, although it is not performing uniformly better than the best one degree of freedom test. The test statistic based on comparing all four estimators (which is not reported in the Table) does not result in a very powerful test compared to those tests based on two pairs of estimators, since the additional degree of freedom has a much more dominant effect on the power than a (fairly small) rise in the decentrality parameter. For panel A of Table 2, the LM test is obviously far more powerful then any Hausman test. Note that the power of all tests reduces substantially if the R^2 of the selection equation is reduced from 0.9 to 0.1; the bias

in the estimators is however still substantial (53 percent for the random effects estimator from the unbalanced panel).

If $\sigma_{\varepsilon\eta} = 0$, i.e. if the error shocks in the structural equation and the selection equation are uncorrelated, but $\sigma_{\alpha\xi} \neq 0$ (so H_0^{FE} is true and H_0^{RE} is not; panel B) all tests seem to have limited power only. Even the power of the LM test is very limited in this case, in which, of course, the null hypothesis H_0 is only violated in one direction ($\sigma_{\alpha\xi} \neq 0$). Since the bias in the fixed effects estimators is zero in this case, while that in the random effects estimators is small (compare Table 1), this does not seem to be a situation to worry about.

As shown in panel C of Table 2, the power of all tests appears to be larger in the case where the response is determined by an individual effect which is correlated with the regressor ($\pi \neq 0$ and $\gamma_1 = 0$) than in the case where the regressor itself determines the response ($\pi = 0$ and $\gamma_1 \neq 0$). Note that for the Hausman tests comparing FE and RE estimators we have a standard situation in which one of the estimators in the test statistic is consistent even if the null hypothesis does not hold. Remarkably, the variable addition tests have fairly good power properties as well, especially the one based on adding the number of waves an individual is participating ($\Sigma_t r_{it}$). The one based on including $r_{i,t-1}$ has only very limited power. Concerning the Hausman tests, the one comparing the RE and FE estimator in the unbalanced panel, which is the standard Hausman test for uncorrelated individual effects, has the largest power of the one degree of freedom tests. In some cases it is worthwhile to combine two restrictions and perform a two degrees of freedom test. It should be clear from the simulation results in the table that it is well possible that the standard Hausman (1978) specification test for testing the hypothesis that the individual effects are uncorrelated with the explanatory variables rejects due to the presence of selectivity bias.

Unfortunately, none of the simple tests seems to have uniformly better power properties than the others, so we cannot recommend one particular test. The power of all tests seems to depend crucially on the fact whether H_0^{FE} is false or, if it is true, why H_0^{FE} is true ($\sigma_{\varepsilon\eta} = 0$ or $\gamma_1 = 0$?). In the latter case ($\gamma_1 = 0$) the power of most simple tests is quite reasonable, while it is not if $\sigma_{\varepsilon\eta} = 0$. In line with the Monte Carlo results above, we are tempted to say that both the second and the third Hausman test (RE, balanced versus unbalanced, and unbalanced, FE versus RE, respectively) perform relatively well and may be a good choice in applied work. The best choice for a variable addition test seems to be to include $\Sigma_t r_{it}$ in the structural equation.

So far, we have only considered numerical analyses for a three wave panel ($T = 3$). If T increases, the number of individuals in the balanced subpanel (keeping all parameters fixed) will decrease, which may increase the differences found between the estimators from the balanced and the unbalanced panel. Moreover, the difference between the fixed effects estimator and the random effects estimator for a given sample will get smaller, since the weight of the between estimator in the random effects estimator is inversely related with T (compare Hsiao 1986, p. 36). This suggests that the power of the Hausman tests comparing estimators from the balanced and unbalanced panel will increase with T and that of the standard Hausman specification tests will decrease with T. For larger T the second Hausman

test (comparing the random effects estimators from the balanced and unbalanced panel) is probably the most attractive way to test hypothesis H_0^{RE}.

7. CONCLUDING REMARKS

In this paper we suggested several simple tests to check the presence of selective nonresponse in a panel data model. We considered the selectivity bias of the fixed and random effects estimators and showed that the FE estimator is more robust to nonresponse biases than the RE estimator. Several simple Hausman tests have been suggested which are based on the differences in the pseudo true values of these estimators. Furthermore, some variable addition tests are proposed which can be used to test for selectivity bias. Neither of these tests requires estimation of the model under selectivity nor a specification of the response mechanism.

Our theoretical results show that the conditions for consistency of a fixed effects estimator are weaker than that for a consistent random effects estimator. In addition, a Monte Carlo study shows that the bias of the FE estimator is likely to be smaller than that of the RE estimator in cases where both estimators are inconsistent. The numerical results also indicate that the bias resulting from a balanced sub-panel is likely to be smaller than that from the unbalanced panel.

Although the proposed Hausman and variable addition tests have poor power properties in some cases, they may be a good instrument for checking the importance of the selectivity problem. In particular when response is partly determined by an individual effect which is correlated with the regressor the power of several Hausman tests and variable addition tests is quite reasonable in comparison with the Lagrange Multiplier test. For practical purposes at least two Hausman tests can be recommended: the one comparing the random effects estimators from the balanced and unbalanced panel, and the one comparing the RE and FE estimators in the unbalanced panel (the standard Hausman test for correlated individual effects). A test that is even simpler is the variable addition test including $T_i = \Sigma_t r_{it}$ in the specification of equation (1). This test also seems to perform quite reasonable in practice.

For ease of presentation attention in this paper was restricted to the linear regression model, although several of the tests can straightforwardly be generalized to nonlinear models. For example, for any model that is identified from both the unbalanced panel and the balanced sub-panel, it is possible to compute a simple Hausman test comparing the corresponding two estimators. Moreover, adding T_i or c_i as an additional explanatory variable is possible in virtually any kind of model and consequently, its significance can be tested straightforwardly, yielding very simple checks for the presence of selectivity bias.

Tilburg University, The Netherlands

APPENDIX
SOME TECHNICAL DETAILS

The Derivation of (20) and (21). From (19) it is readily verified that

$$\begin{pmatrix} \alpha_i \\ \varepsilon_i \\ \xi_i \iota + \eta_i \end{pmatrix} \sim N\left(0, \begin{pmatrix} \sigma_\alpha^2 & 0 & \sigma_{\alpha\xi}\iota' \\ & \sigma_\varepsilon^2 I & \sigma_{\varepsilon\eta} I \\ & & \sigma_\eta^2 + \sigma_\xi^2 \iota\iota' \end{pmatrix}\right) \tag{29}$$

which yields

$$E\{\varepsilon_i | \xi_i \iota + \eta_i\} = \frac{\sigma_{\varepsilon\eta}}{\sigma_\eta^2}\left(I - \frac{\sigma_\xi^2}{T\sigma_\xi^2 + \sigma_\eta^2}\iota\iota'\right)(\xi_i \iota + \eta_i), \tag{30}$$

and proves (20) and (22) if we use the definition of $\tilde{\varepsilon}_{it}$ and take expectations conditional upon $r_{i1}, \ldots, r_{iT}$. It also follows that

$$E\{\alpha_i | \xi_i \iota + \eta_i\} = \frac{\sigma_{\alpha\xi}}{\sigma_\eta^2}\iota'\left(I - \frac{\sigma_\xi^2}{T\sigma_\xi^2 + \sigma_\eta^2}\iota\iota'\right)(\xi_i \iota + \eta_i), \tag{31}$$

which proves (21) after taking conditional expectations upon $r_{i1}, \ldots, r_{iT}$.

Moreover, since $E\{\xi_i | r_i\}$ is fixed over time and since (dropping the $z_{is}\pi_s$ terms for notational convenience)

$$E\{\eta_{it} | r_i\} = \int \frac{\phi\left(\dfrac{z_{it}\gamma + \xi_i}{\sigma_\eta}\right)}{\Phi\left(\dfrac{z_{it}\gamma + \xi_i}{\sigma_\eta}\right)} f(\xi_i | r_i)\, d\xi_i \quad \text{if } r_{it} = 1 \tag{32}$$

where ϕ and Φ are the standard normal density and distribution function, respectively, and $f(\xi_i | r_i)$ is the conditional density of ξ_i given selection (see Ridder 1990), it is evident that there is no selectivity bias if $z_{it}\gamma$ is constant over time, i.e. if the probability of an individual of being observed is constant for all t.

The Lagrange Multiplier Test Statistic for Selectivity Bias. The loglikelihood contribution of an individual i in the full model is given by

$$L_i = \log f(r_i | R_i y_i) f(R_i y_i) \tag{33}$$

where $f(r_i | R_i y_i)$ is the likelihood function of a (conditional) T-variate probit model and $f(R_i y_i)$ is the likelihood function of a T_i-dimensional linear error components model (compare Hsiao 1986, p. 38). The second term is simple and can be written as

$$\log f(R_i y_i) = -\frac{T_i}{2}\log 2\pi - \frac{T_i - 1}{2}\log \sigma_\varepsilon^2 - \frac{1}{2}\log(\sigma_\varepsilon^2 + T_i\sigma_\alpha^2) \tag{34}$$

$$-\frac{1}{2\sigma_\varepsilon^2}\sum_{t=1}^{T} r_{it}(\tilde{y}_{it}-\tilde{x}_{it}\beta)^2-\frac{T_i}{2(\sigma_\varepsilon^2+T_i\sigma_\alpha^2)}(\bar{y}_i-\bar{x}_i\beta)^2.$$

The first term in (33) is somewhat more complicated because we have to derive the conditional distribution of the error term in the probit model. From (19) and defining $v_{it}=r_{it}(\alpha_i+\varepsilon_{it})$ (where r_{it} is treated as nonstochastic), the conditional expectation of the error term $\xi_i+\eta_{it}$ is given by

$$(35)\quad E\{\xi_i+\eta_{it}|v_{i1},\ldots,v_{iT}\}=r_{it}\frac{\sigma_{\varepsilon\eta}}{\sigma_\varepsilon^2}\left(v_{it}-\frac{\sigma_\alpha^2}{\sigma_\varepsilon^2+T_i\sigma_\alpha^2}\sum_{s=1}^{T}v_{is}\right)+\frac{\sigma_{\alpha\xi}}{\sigma_\varepsilon^2+T_i\sigma_\alpha^2}\sum_{s=1}^{T}v_{is}=c_{it},\ \text{say}.$$

Using (19) the conditional variance of $\xi_i+\eta_{it}$ can also be derived. It is straightforward to show that the conditional distribution of $\xi_i+\eta_{it}$ given $v_{i1},\ldots,v_{iT}$ corresponds with the (unconditional) distribution of the sum of three normal variables $u_{it}+\nu_{1i}+r_{it}\nu_{2i}$ whose distribution is characterized by

$$E\{\nu_{1i}\}=E\{\nu_{2i}\}=0,\quad E\{u_{it}\}=c_{it},$$

$$V\{u_{it}\}=\sigma_\eta^2-r_{it}\sigma_{\varepsilon\eta}^2/\sigma_\varepsilon^2=s_t^2,\quad \text{say}$$

$$V\{\nu_{1i}\}=\sigma_\xi^2-T_i\sigma_{\alpha\xi}^2(\sigma_\varepsilon^2+T_i\sigma_\alpha^2)^{-1}=\omega_1,\quad \text{say}$$

$$V\{\nu_{2i}\}=\sigma_{\varepsilon\eta}^2\sigma_\alpha^2\sigma_\varepsilon^{-2}(\sigma_\varepsilon^2+T_i\sigma_\alpha^2)^{-1}=\omega_2,\quad \text{say}$$

$$\operatorname{cov}\{\nu_{1i},\nu_{2i}\}=-\sigma_{\alpha\xi}\sigma_{\varepsilon\eta}(\sigma_\varepsilon^2+T_i\sigma_\alpha^2)^{-1}=\omega_{12},\quad \text{say}$$

and all other covariances equal to zero. For notational convenience we do not explicitly add an index i to the (co)variances s_t and ω. Note that $c_{it}=0$, $s_t^2=\sigma_\eta^2$, $\omega_1=\sigma_\xi^2$ and $\omega_2=0$ under H_0. Like in the unconditional error components probit model (compare Heckman 1981), the likelihood function can be written as (dropping the $z_{is}\pi_s$ terms for notational convenience)

$$(36)\qquad f(r_i|R_iy_i)=E_\theta\left\{\prod_{t=1}^{T}\Phi\left(d_{it}\frac{z_{it}\gamma+c_{it}+\nu_{1i}+r_{it}\nu_{2i}}{s_t}\right)\right\}$$

where the expectation is taken over ν_{1i} and ν_{2i}, and $d_{it}=2r_{it}-1$. It is this likelihood function that has to be differentiated w.r.t. the unknown parameters γ, σ_ξ^2, $\sigma_{\alpha\xi}$ and $\sigma_{\varepsilon\eta}$. However, the expectation operator depends on the unknown parameter vector θ (because the density of ν_{1i} and ν_{2i} is not defined with respect to the same measure under H_0 and the alternative), implying that the order of taking expectations and differentiating is not interchangeable. This problem can easily be solved by defining two new integration variables that are both standard

normally distributed (under the null and the alternative), τ_1 and τ_2, say. Then we obtain

$$(37)\quad f(r_i|R_i y_i) = \iint \prod_{t=1}^{T} \Phi\left(d_{it}\,\frac{z_{it}\gamma + c_{it} + a_{it}\tau_1 + b_{it}\tau_2}{s_t}\right)\phi(\tau_1)\phi(\tau_2)\,d\tau_1\,d\tau_2$$

where

$$a_{it} = \omega_1^{1/2} + r_{it}\omega_{12}\omega_2^{-1/2} \quad \text{and} \quad b_{it} = r_{it}(\omega_2 - \omega_{12}^2\omega_1^{-1})^{1/2}.$$

Since $f(R_i y_i)$ does not depend on $\sigma_{\varepsilon\eta}$ and $\sigma_{\alpha\xi}$, differentiating the log of the expression above and evaluating the result under H_0 yields the scores w.r.t. the two covariances. Using the fact that for any element ψ of the parameter vector (γ, σ_η^2, $\sigma_{\varepsilon\eta}$, $\sigma_{\alpha\xi}$),

$$(38)\qquad \frac{\partial L_i}{\partial \psi} = \frac{\partial f(r_i|R_i y_i)}{\partial \psi}\,f(r_i|R_i y_i)$$

with

$$(39)\quad \frac{\partial f(r_i|R_i y_i)}{\partial \psi} = \iint \sum_{s=1}^{t} \prod_{t=1,\,t\neq s}^{T} \Phi_t(.)\,\frac{\partial \Phi_s(.)}{\partial \psi}\,\phi(\tau_1)\phi(\tau_2)\,d\tau_1\,d\tau_2,$$

the score w.r.t. $\sigma_{\alpha\xi}$ can easily be derived using the following equality (under H_0)

$$(40)\qquad \frac{\partial \Phi_t(.)}{\partial \sigma_{\alpha\xi}} = \phi\left(d_{it}\,\frac{z_{it}\gamma + \sigma_\xi\tau_1}{\sigma_\eta}\right)\frac{d_{it}}{\sigma_\eta}\left(\frac{\partial c_{it}}{\partial \sigma_{\alpha\xi}} + \frac{\partial \omega_1^{1/2}}{\partial \sigma_{\alpha\xi}}\,\tau_1\right).$$

Similarly, for $\sigma_{\varepsilon\eta}$, we use

$$(41)\quad \frac{\partial \Phi_t(.)}{\partial \sigma_{\varepsilon\eta}} = \phi\left(d_{it}\,\frac{z_{it}\gamma + \sigma_\xi\tau_1}{\sigma_\eta}\right)\frac{d_{it}}{\sigma_\eta}\left(\frac{\partial c_{it}}{\partial \sigma_{\varepsilon\eta}} + r_{it}\tau_2\sigma_\alpha^2\sigma_\varepsilon^{-2}(\sigma_\varepsilon^2 + T_i\sigma_\alpha^2)^{-1}\right),$$

from which the score w.r.t. $\sigma_{\varepsilon\eta}$ under H_0 can easily be derived. Note that both τ_1 and τ_2 occur in the integrand such that numerical integration over two dimensions will be required.

For the scores w.r.t. γ and $\sigma_\xi^2 = 1 - \sigma_\eta^2$ it suffices under H_0 to look at $\partial f(r_i)/\partial\gamma$ and $\partial f(r_i)/\partial\sigma_\xi$, where (compare Heckman 1981)

$$(42)\qquad f(r_i) = \int \prod_{t=1}^{T} \Phi\left(d_{it}\,\frac{z_{it}\gamma + \sigma_\xi\tau_1}{\sigma_\eta}\right)\phi(\tau_1)\,d\tau_1.$$

Both scores will require numerical integration over one dimension.

REFERENCES

BALTAGI, B. H., "Pooling Cross-Sections with Unequal Time-Series Lengths," *Economics Letters* 18 (1985), 133–136.

CHAMBERLAIN, G., "Panel Data," in Z. Griliches and M. D. Intriligator, eds., *Handbook of Econometrics*, Vol. 2 (Amsterdam: North Holland, 1984), 1247–1318.

ENGLE, R. F., "Wald, Likelihood Ratio and Lagrange Multiplier Tests in Econometrics," in Z. Griliches and M. D. Intrilligator, eds., *Handbook of Econometrics*, Vol. 2 (Amsterdam: North Holland, 1984), 775–826.

HAUSMAN, J. A., "Specification Tests in Econometrics," *Econometrica* 46 (1978), 1251–1271.

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

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

———, "Sample Selection Bias as a Specification Error," *Econometrica* 47 (1979), 153–161.

———, "Statistical Models for Discrete Panel Data," in C. F. Manski and D. McFadden, eds., *Structural Analysis of Discrete Data with Econometric Applications* (Cambridge: MIT Press, 1981), 114–178.

HOLLY, A., "A Remark on Hausman's Specification Test," *Econometrica* 50 (1982), 749–759.

———, "Specification Tests: An Overview," in T. F. Bewley, ed., *Advances in Econometrics, Fifth World Congress*, Vol. 1 (Cambridge: Cambridge University Press, 1987), 59–97.

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

LEE, L. F., "Tests for the Bivariate Normal Distribution in Econometric Models with Selectivity," *Econometrica* 52 (1984), 843–863.

——— AND G. S. MADDALA, "The Common Structure of Tests for Selectivity Bias, Serial Correlation, Heteroskedasticity and Non-normality in the Tobit Model," *International Economic Review* 26 (1985), 1–20.

MANSKI, C. F., "Anatomy of the Selection Problem," *The Journal of Human Resources* 24 (1989), 343–360.

———, "The Selection Problem," Working Paper No. 9012, Social Systems Research Institute, University of Wisconsin, 1990.

MIZON, G. E., "Inferential Procedures in Nonlinear Models: An Application to a UK Industrial Cross Section Study of Factor Substitution and Returns to Scale," *Econometrica* 45 (1977), 1221–1242.

NIJMAN, T. E. AND M. VERBEEK, "Nonresponse in Panel Data: The Impact on Estimates of a Life Cycle Consumption Function," mimeo, Tilburg University, 1990.

RIDDER, G., "Attrition in Multi-Wave Panel Data," in J. Hartog, G. Ridder and J. Theeuwes, eds., *Panel Data and Labor Market Studies* (Elsevier: North-Holland, 1990), 45–67.

VELLA, F., "A Simple Estimator for Simultaneous Models with Censored Endogenous Regressors," mimeo, Rice University, 1990.

VERBEEK, M., "On the Estimation of a Fixed Effects Model with Selectivity Bias," *Economics Letters* 34 (1990), 267–270.

WANSBEEK, T. J. AND A. KAPTEYN, "Estimation of the Error Components Model with Incomplete Panels," *Journal of Econometrics* 41 (1989), 341–361.