ELSEVIER Journal of Econometrics 68 (1995) 115–132

JOURNAL OF Econometrics

Selection corrections for panel data models under conditional mean independence assumptions

Jeffrey M. Wooldridge

Department of Economics, Michigan State University, East Lansing, MI 48824-1038, USA

Abstract

Some new methods for testing and correcting for sample selection bias in panel data models are proposed. The assumptions allow the unobserved effects in both the regression and selection equations to be correlated with the observed variables; the error distribution in the regression equation is unspecified; arbitrary serial dependence in the idiosyncratic errors of both equations is allowed; and all idiosyncratic errors can be heterogeneously distributed. Compared with maximum likelihood and other estimators derived under fully parametric assumptions, the new estimators are much more robust and have significant computational advantages.

Key words: Panel data; Sample selection; Fixed effects; Conditional mean independence; Two-step estimation
JEL classification: C23

1. Introduction

This paper offers some new methods for testing and correcting for selection bias in linear unobserved components panel data models. The model studied is of the fixed effects (FE) variety in the sense that the unobserved component is allowed to be correlated with the observable explanatory variables, and we impose no distributional assumptions on the unobserved effect. In addition, we do not require the idiosyncratic errors in the regression equation to have a known distribution, and they can have serial dependence of unspecified form.

Tom Mroz, Peter Schmidt, Jeff Zabel, and two anonymous referees provided very helpful comments. Financial support from a Sloan Foundation Research Fellowship is gratefully acknowledged.

SSDI 0304407694 01645 G

These features distinguish the paper from other recent work on correcting for selection bias in linear panel data models. Verbeek and Nijman (1992) (hereafter VN) consider a random effects (RE) model under the assumptions of normality and serial independence of the idiosyncratic errors in both the selection and regression equations, and the time-constant unobserved effects in the selection and regression equations are assumed to be normally distributed. Nijman and Verbeek (1992) (hereafter NV) and Zabel (1992) study almost the same model except they allow both unobserved effects to be correlated with the observables. Vella and Verbeek (1992) extend NV to allow for functions of the endogenous censoring variable to appear among the explanatory variables.

In order to make the selection equation simple to estimate we make a normality assumption on the errors in the selection equation; but we allow these errors to display arbitrary serial correlation and unconditional heteroskedasticity. The key assumptions that lead to estimable equations that identify the structural parameters – the conditional mean independence assumptions – are much weaker than full joint distributional assumptions on the time-constant unobservables and the idiosyncratic errors.

In addition to their wide array of applicability, the new procedures are relatively computationally simple. Current corrections for the fully parametric model require bivariate numerical integration due to the time dimension and the integrating out of the unobserved effect. Allowing for even simple forms of serial correlation would make the parametric approach almost prohibitively difficult. The methods offered here require a standard probit or Tobit regression for each time period followed by a multivariate linear regression, regardless of the time series properties of the errors.

In Section 3 we offer some variable addition tests for selection bias based on fixed effects estimation using the unbalanced panel. These tests have some advantages over the variable addition tests suggested by VN (1992). In addition to the absence of selection bias, the latter tests maintain a null hypothesis of no correlation between the unobserved effect and the regressors. Such tests could reject even in the absence of selection bias if the unobserved effect and the regressors are correlated, or if selection is based on the unobserved effect. Here, the unobserved effect and regressors are allowed to be arbitrarily correlated and selection may depend on the unobserved effect, so in this sense the tests are complementary to the Hausman test proposed by VN that compares the fixed effects estimators from the balanced and unbalanced panels. Our variable addition tests are easily computable after probit or Tobit estimation for each time period followed by fixed effects estimation, and they are explicitly robust to arbitrary dependence and heterogeneity in the errors of the regression equation.

In what follows we cover two cases that arise in practice. The first is when the selection variable is partially observed. The leading application of this case is to labor studies where the regression equation is, say, a wage equation, and selection depends on whether or not individuals are working. If a person is working, then her/his hours are recorded, and selection is determined by nonzero hours worked. The other case is the more widely applicable one, where only a binary selection indicator is observed. One might expect that more flexibility is available in the first case because more information is known, and we show here that this is the case.

A final note on the sampling assumption. Throughout we assume that the cross-section observations are independent, identically distributed. The independence assumption is crucial but the identical distribution assumption is mostly for notational ease. Unless otherwise stated there is no assumption on the dependence across time. This is possible because the asymptotic analysis is for fixed T and N going to infinity.

2. Consistency of fixed effects under exogenous selection

We first briefly investigate assumptions under which the usual fixed effects estimator on the unbalanced panel is consistent. The model is the usual linear, unobserved effects model for i.i.d. cross-section observations: for any i,

$$y_{it} = \alpha_i + \boldsymbol{x}_{it}\beta + u_{it}, \quad t = 1, \ldots, T, \tag{2.1}$$

where $\boldsymbol{x}_{it}$ is $1 \times K$ and β is the $K \times 1$ vector of interest. We assume that N cross-section observations are available and the asymptotic analysis is as $N \to \infty$. We explicitly cover the case where α_i is allowed to be correlated with $\boldsymbol{x}_{it}$, so that all elements of $\boldsymbol{x}_{it}$ are time-varying; this allows for time dummies and interactions of time dummies with time-constant variables.

If all T periods are available for any cross-section drawn from the population, then a sufficient condition for fixed effects (and a variety of other procedures) to be consistent as $N \to \infty$ is:

Asuumption 1. $\mathrm{E}(u_{it} | \alpha_i, \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}) = 0, \quad t = 1, 2, \ldots, T.$ ■

Under Assumption 1, fixed effects is consistent and $\sqrt{N}$-asymptotically normal. Since the regressors for *all* time periods are in the conditioning set at any time t, the assumption implies that the $\boldsymbol{x}_{it}$ are *strictly exogenous conditional* on the latent effect α_i (see, for example, Chamberlain, 1982). Technically speaking, Assumption 1 is more restrictive than needed for fixed effects to be consistent, but it gives a natural interpretation for β and it extends in a straightforward manner to the sample selection case.

For the usual FE standard errors and test statistics to be valid we should add to Assumption 1 the assumption

$$\mathrm{E}(\boldsymbol{u}_i \boldsymbol{u}_i' | \alpha_i, \boldsymbol{x}_i) = \sigma^2 \boldsymbol{I}_T, \tag{2.2}$$

where $\boldsymbol{u}_i = (\boldsymbol{u}_{i1}, \ldots, \boldsymbol{u}_{iT})'$ and $\boldsymbol{x}_i \equiv (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT})$. Alternatively, we could relax (2.2) by replacing $\sigma^2 \boldsymbol{I}_T$ with Ω, a positive definite $T \times T$ matrix. Then, as in Kiefer (1980), fixed effects GLS can be applied. We will not use GLS-type procedures because we make *no* assumptions about $\mathrm{E}(\boldsymbol{u}_i \boldsymbol{u}_i' | \alpha_i, \boldsymbol{x}_i)$.

We will not cover the random effects case in detail, except to occasionally note how the assumptions would have to be strengthened. In addition to Assumption 1, random effects add the assumption

$$\mathrm{E}(\alpha_i | \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}) = \mathrm{E}(\alpha_i). \tag{2.3}$$

This is too restrictive for many applications, so we will not impose it here.

Now consider the fixed effects estimator on the unbalanced panel. The vector of selection indicators for each i is denoted $\boldsymbol{s}_i = (s_{i1}, \ldots, s_{iT})'$, and for now we assume only that $(\boldsymbol{x}_{it}, y_{it})$ is observed if $s_{it} = 1$. The FE estimator can be written

$$\hat{\beta} = \left(N^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \ddot{\boldsymbol{x}}_{it}' \ddot{\boldsymbol{x}}_{it} \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \ddot{\boldsymbol{x}}_{it}' \ddot{y}_{it} \right), \tag{2.4}$$

where, for $s_{it} = 1$,

$$\ddot{\boldsymbol{x}}_{it} \equiv \boldsymbol{x}_{it} - T_i^{-1} \sum_{r=1}^{T} s_{ir} \boldsymbol{x}_{ir}, \tag{2.5}$$

$$\ddot{y}_{it} \equiv y_{it} - T_i^{-1} \sum_{r=1}^{T} s_{ir} y_{ir}, \tag{2.6}$$

$$T_i = \sum_{t=1}^{T} s_{it}. \tag{2.7}$$

Expanding (2.4) shows that the following is sufficient for FE on the selected subsample to be consistent and asymptotically normal (as $N \to \infty$):

Assumption 1′. $\mathrm{E}(u_{it} | \alpha_i, \boldsymbol{x}_i, \boldsymbol{s}_i) = 0, \quad t = 1, 2, \ldots, T.$ ■

Note that – just as it is not sufficient to put only $\boldsymbol{x}_{it}$ in the conditioning set at time t – it is not sufficient to put only s_{it} in the conditioning set at time t. In other words, under Assumption 1′ the selection process is *strictly* exogenous conditional on α_i and x_i. As with Assumption 1, this assumption allows for arbitrary serial correlation and heteroskedasticity in the (u_{it}).

Technically, Assumption 1′ is more restrictive than the exogeneity assumptions used by VN in their analysis of the fixed effects estimator. (VN state conditions directly in terms of $\ddot{u}_{it} \equiv u_{it} - T_i^{-1} \sum_{r=1}^{T} s_{ir} u_{ir}$.) But Assumption 1′ is simple to state and is less cumbersome to apply to cases where selection is based on the exogenous variables. In fact, because Assumption 1′ puts no restrictions on how $\boldsymbol{s}_i$ relates to $(\alpha_i, \boldsymbol{x}_i)$, it follows that selection can depend on $(\alpha_i, \boldsymbol{x}_i)$ in an arbitrary fashion (contrast VN, 1992, fn. 5).

For the usual fixed effects variance–covariance matrix and inference to be valid an additional assumption is needed. Sufficient is

$$\mathrm{E}(\boldsymbol{u}_i\boldsymbol{u}_i' \mid \alpha_i, \boldsymbol{x}_i, \boldsymbol{s}_i) = \sigma^2 \boldsymbol{I}_T. \tag{2.8}$$

If (2.8) does not hold we can obtain a robust variance–covariance matrix estimator by applying the results from the appendix: let $\boldsymbol{w}_{it} = \ddot{\boldsymbol{x}}_{it}$, $\hat{e}_{it} = \ddot{y}_{it} - \ddot{\boldsymbol{x}}_{it}\hat{\beta}$, and set $\hat{\boldsymbol{D}} = \boldsymbol{0}$ in Appendix equations (A.3)–(A.5).

Consistency of random effects estimation using the unbalanced or balanced panel requires an additional assumption. The most natural is

$$\mathrm{E}(\alpha_i \mid \boldsymbol{x}_i, \boldsymbol{s}_i) = \mathrm{E}(\alpha_i). \tag{2.9}$$

An important limitation of (2.9) is that it rules out selection that depends on the unobserved effect α_i. Tests that maintain Assumption 1′ and (2.9) under H_0 cannot distinguish between selection that depends on α_i and selection that does not. Because fixed effects is consistent under Assumption 1′ only, the tests we derive in the next section maintain only that assumption – or its natural variant when the selection variable is partially observed – under H_0.

3. Some variable addition tests for selection bias

We now derive some simple variable addition tests of selection bias. There are unlimited tests one can derive that have the proper size under the null, but one must be guided by power and computational considerations. Here we use variable addition tests similar in spirit to VN (1992), but our auxiliary regressors are either Tobit residuals or inverse Mills ratios, and we explicitly allow for serial correlation and heteroskedasticity of unknown form in $\{u_{it}\}$.

3.1. Testing when the selection variable is partially observed

We first obtain a test under the assumption that the (latent) variable determining selection can be observed whenever it is nonnegative. To motivate the test we write down a particular selection mechanism, but it should be remembered that the selection mechanism need not be correctly specified in any sense; it is simply a vehicle for obtaining a sensible test. In Section 4 we formalize the assumptions on the selection mechanism.

Unless otherwise stated we assume that the explanatory variables $\boldsymbol{x}_{it}$ are observed for all $t = 1, 2, \ldots, T$. The variable y_{it} is observed if $s_{it} = 1$ and not otherwise. For each $t = 1, 2, \ldots, T$, define a latent variable

$$h_{it}^* = \delta_{t0} + \boldsymbol{x}_{i1}\delta_{t1} + \cdots + \boldsymbol{x}_{iT}\delta_{tT} + v_{it}, \tag{3.1}$$

where v_{it} is independent of $(\alpha_i, \boldsymbol{x}_i)$, $v_{it} \sim \text{Normal}(0, \sigma_t^2)$, and δ_{tr} is a $K \times 1$ vector of unknown parameters, $r = 1, 2, \ldots, T$. At this point it should be mentioned that independence between v_{it} and α_i is often too strong an assumption for deriving selection *corrections*, so we will not impose such an assumption in Section 4. Here we are only deriving a test of selection bias, and because it is convenient to do so, we proceed as if v_{it} and α_i are independent. As will be seen in Section 4, the test we derive here also falls out under much less restrictive assumptions.

The binary selection indicator is defined as $s_{it} \equiv 1[h_{it}^* > 0]$. In this subsection we also assume that, for all t, the censored variable $h_{it} \equiv \max(0, h_{it}^*)$ is observed; this is what we mean by partial observability of the selection variable. Technically, the null hypothesis is a strengthened version of Assumption 1′, where $\boldsymbol{h}_i \equiv (h_{i1}, h_{i2}, \ldots, h_{iT})'$ replaces $\boldsymbol{s}_i$ in the conditioning set; but the interpretation of the assumption is essentially the same.

Note that we use $\boldsymbol{x}_{it}$ as the explanatory variables in both (2.1) and (3.1). Technically this does not affect our ability to carry out testing and correcting for selection bias, but it helps if one has exclusion restrictions in (2.1). This is accomplished in what follows by simply omitting the appropriate elements from $\boldsymbol{x}_{it}$ in (2.1). We do not show this explicitly to keep the notation as simple as possible.

The mechanism described by (3.1) is the reduced form for many popular selection equations. For example, consider the structural selection equation

$$h_{it}^* = \xi_i + \boldsymbol{x}_{it}\delta + a_{it}, \tag{3.2}$$

where (ξ_i, a_{it}) is jointly normally distributed with $\mathrm{E}(a_{it}) = 0$ and a_{it} independent of $\boldsymbol{x}_i$; ξ_i is a time-constant unobserved effect. If ξ_i is independent of $\boldsymbol{x}_i$, as in a random effects-type specification, then (3.1) holds with $\delta_{t0} = \delta_0 \equiv \mathrm{E}(\xi_i)$, $\delta_{tr} \equiv 0$, $r \neq t$, and $\delta_{tt} \equiv \delta$. Note that $v_{it} = (\xi_i - \delta_0) + a_{it}$ in this case, so the v_{it} are probably serially dependent.

To allow ξ_i to be correlated with $\boldsymbol{x}_i$ we can specify, as in Chamberlain (1980) and VN (1992),

$$\xi_i = \eta_0 + \boldsymbol{x}_{i1}\eta_1 + \cdots + \boldsymbol{x}_{iT}\eta_T + c_i, \tag{3.3}$$

where c_i is independent of $\boldsymbol{x}_i$ with a zero mean normal distribution. Then (3.1) holds with $\delta_{t0} \equiv \eta_0$, $\delta_{tr} \equiv \eta_r$, $r \neq t$, $\delta_{tt} \equiv \eta_t + \delta$, and

$$v_{it} \equiv c_i + a_{it}. \tag{3.4}$$

Again, unless a_{it} has a very specific type of autocorrelation, v_{it} is serially correlated. Also, $\mathrm{var}(v_{it})$ can change over time if $\mathrm{var}(a_{it})$ does.

Another model that can be expressed as in (3.1) is the dynamic model

$$h_{it}^* = \delta_0 + \rho h_{i,t-1}^* + \boldsymbol{x}_{it}\delta + a_{it}, \tag{3.5}$$

where a_{it} is a mean zero normal random variable independent of $\boldsymbol{x}_i$. Then, assuming that h_{i0}^* given $\boldsymbol{x}_i$ is normally distributed with linear conditional expectation, (3.5) can be written as in (3.1). The same conclusion holds if a time-constant unobserved effect of the form (3.3) is added to (3.5).

We now derive a conditional expectation that leads to a test for selection bias. Because $\boldsymbol{s}_i$ is a function of $(\boldsymbol{x}_i, \boldsymbol{v}_i)$, where $\boldsymbol{v}_i \equiv (v_{i1}, \dots, v_{iT})'$, a sufficient condition for Assumption 1′ – with $\boldsymbol{h}_i$ or $\boldsymbol{s}_i$ in the conditioning set – is

$$\mathrm{E}(u_{it} \mid \alpha_i, \boldsymbol{x}_i, \boldsymbol{v}_i) = 0, \qquad t = 1, 2, \dots, T. \tag{3.6}$$

This equation also suggests useful alternatives that imply selectivity bias. The simplest such alternative is

$$\mathrm{E}(u_{it} \mid \alpha_i, \boldsymbol{x}_i, \boldsymbol{v}_i) = \mathrm{E}(u_{it} \mid v_{it}) = \rho v_{it}, \qquad t = 1, 2, \dots, T, \tag{3.7}$$

for some unknown scalar ρ. Later we will use assumptions similar to, but not nearly as restrictive as, (3.7) when correcting for selection bias. For now, note that it states that u_{it} is mean independent of $(\alpha_i, \boldsymbol{x}_i, v_{i1}, \dots, v_{i,t-1}, v_{i,t+1}, \dots, v_{iT})$, conditional on v_{it}. In addition, the regression of u_{it} on v_{it} is linear and constant across t.

Under the alternative (3.7) we have

$$\mathrm{E}(y_{it} \mid \alpha_i, \boldsymbol{x}_i, \boldsymbol{v}_i, \boldsymbol{s}_i) = \mathrm{E}(y_{it} \mid \alpha_i, \boldsymbol{x}_i, \boldsymbol{v}_i) = \alpha_i + \boldsymbol{x}_{it}\beta + \rho v_{it}. \tag{3.8}$$

From (3.8) it follows that, if we could observe v_{it} when $s_{it} = 1$, then we could test H_0 by including the v_{it} as an additional regressor in fixed effects estimation and testing H_0: $\rho = 0$ using standard methods. While v_{it} cannot be observed, it can be estimated whenever $s_{it} = 1$ because v_{it} is simply the error in a Tobit model. Thus, we propose the following test for selection bias when h_{it} is observed.

Procedure 3.1 (h_{it} is observed; valid under Assumption 1′ with $\boldsymbol{h}_i$ replacing $\boldsymbol{s}_i$):

(i) For each t, estimate the equation

$$h_{it} = \max(0, \boldsymbol{x}_i \delta_t + v_{it})$$

by standard Tobit, where now $\boldsymbol{x}_i = (1, \boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \dots, \boldsymbol{x}_{iT})$ and $\delta_t \equiv (\delta_{t0}, \delta'_{t1}, \dots, \delta'_{tT})'$. (Henceforth, for notational simplicity, $\boldsymbol{x}_i$ is defined to have unity as its first element.) For $s_{it} = 1$, define $\hat{v}_{it} = h_{it} - \boldsymbol{x}_i \hat{\delta}_t$.

(ii) Estimate the equation

$$\ddot{y}_{it} = \ddot{\boldsymbol{x}}_{it}\beta + \rho\ddot{v}_{it} + \mathrm{error}_{it} \tag{3.9}$$

by pooled OLS using those observations for which $s_{it} = 1$, where $\ddot{\boldsymbol{x}}_{it}$ and $\ddot{y}_{it}$ are defined in (2.5) and (2.6), and

$$\ddot{v}_{it} \equiv \hat{v}_{it} - T_i^{-1} \sum_{r=1}^{T} s_{ir}\hat{v}_{ir}.$$

(iii) Test H_0: $\rho = 0$ using the t-statistic for $\hat{\rho}$. Unless assumption (2.8) is maintained under H_0, a serial correlation and heteroskedasticity-robust standard error should be computed for $\hat{\rho}$. The robust variance matrix for $\hat{\theta} \equiv (\hat{\beta}', \rho)'$ is obtained by setting $\boldsymbol{w}_{it} \equiv (\ddot{\boldsymbol{x}}_{it}, \ddot{v}_{it})$ and $\hat{e}_{it} \equiv \ddot{y}_{it} - \ddot{\boldsymbol{x}}_{it}\hat{\beta} - \hat{\rho}\ddot{v}_{it}$ and using these in Appendix equations (A.3)–(A.5); under H_0 one can take $\boldsymbol{D} \equiv \boldsymbol{0}$ in (A.4). ■

Rather than unconstrained Tobit estimation for each t one might prefer to impose restrictions on the δ_{tr} derived from one of the structural selection models mentioned earlier. For example, under (3.2), (3.3), and a homoskedasticity assumption on the a_{it},

$$h_{it}^* = \eta_0 + \boldsymbol{x}_{i1}\eta_1 + \cdots + \boldsymbol{x}_{iT}\eta_T + \boldsymbol{x}_{it}\delta + v_{it}, \tag{3.10}$$

and $v_{it}|\boldsymbol{x}_i \sim$ Normal $(0, \sigma_v^2)$, $t = 1, 2, \ldots, T$. Even though the v_{it} are *not* serially independent, one can use pooled Tobit to estimate $(\eta_0, \eta_1, \ldots, \eta_T)$, δ, and σ_v^2. Simply define $\boldsymbol{w}_{it} \equiv (1, \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}, \boldsymbol{x}_{it})$ and $\theta \equiv (\eta_0, \eta_1', \ldots, \eta_T', \delta')$, and estimate the equations $h_{it} = \max(0, \boldsymbol{w}_{it}\ \theta + v_{it})$ as if on one long cross-section. Because the h_{it} are not independent across t given $\boldsymbol{x}_i$, the usual Tobit standard errors and test statistics are inappropriate, but the estimators are still consistent and $\sqrt{N}$-asymptotically normal. The consistency proof is quite simple and is based on the Kullback–Leibler information criterion, as in White (1994); a proof is available upon request.

Another popular simplification – see for example Mundlak (1978), NV (1992), and Zabel (1992) – is to assume that ξ_i depends only on the time average of $\boldsymbol{x}_{it}$, in which case (3.10) can be replaced by

$$h_{it}^* = \eta_0 + \bar{\boldsymbol{x}}_i\eta + \boldsymbol{x}_{it}\delta + v_{it}, \tag{3.11}$$

where η is a $K \times 1$ vector of parameters. Again, (η_0, η, δ) can be consistently estimated by pooled Tobit, ignoring the serial dependence in v_{it}.

Simplest of all is to carry out the test assuming that ξ_i is independent of $\boldsymbol{x}_i$ and obtaining the $\hat{v}_{it}$ from pooled Tobit estimation on the equations $h_{it} = \max(0, \delta_0 + \boldsymbol{x}_{it}\delta + v_{it})$. In thinking about these alternative procedures one should remember that the method for obtaining $\hat{v}_{it}$ does not affect the asymptotic size of the test, but it should affect the power.

A final comment before turning to the case where only s_{it} is observed. As noted earlier, many variable addition tests are valid tests of Assumption 1′. The test in Procedure 3.1 seems reasonable because the alternative (3.7) can be derived from a standard model of sample selection. Also, as we show in Section 4.1, under certain assumptions this procedure produces consistent estimates of β under sample selection. Nevertheless, for testing it is just as legitimate to replace $\hat{v}_{it}$ with h_{it} in Procedure 3.1. This approach does have two advantages: it circumvents having to do the T cross-section Tobits to obtain $\hat{v}_{it}$, and it can be used when $\boldsymbol{x}_{it}$ is observed only when $s_{it} = 1$.

3.2. Testing when only the selection indicator is observed

An alternative to Procedure 3.1 is needed when h_{it} is not observed. The starting point is still Eq. (3.8), but we must now condition only on $\mathbf{s}_i$ rather than $\mathbf{v}_i$. Using iterated expectations, this gives

$$\begin{aligned} \mathrm{E}(y_{it}|\alpha_i, \mathbf{x}_i, \mathbf{s}_i) &= \alpha_i + \mathbf{x}_{it}\beta + \rho \mathrm{E}(v_{it}|\alpha_i, \mathbf{x}_i, \mathbf{s}_i) \\ &= \alpha_i + \mathbf{x}_{it}\beta + \rho \mathrm{E}(v_{it}|\mathbf{x}_i, \mathbf{s}_i), \end{aligned} \tag{3.12}$$

under the assumption – used for deriving the test – that $\mathbf{v}_i$ is independent of $(\alpha_i, \mathbf{x}_i)$. If the v_{it} were independent across t then $\mathrm{E}(v_{it}|\mathbf{x}_i, \mathbf{s}_i) = \mathrm{E}(v_{it}|\mathbf{x}_i, s_{it})$. As mentioned earlier, this is unrealistic when a structural selection equation contains an unobserved effect. Still, it is computationally much easier to replace $\mathrm{E}(v_{it}|\mathbf{x}_i, \mathbf{s}_i)$ with $\mathrm{E}(v_{it}|\mathbf{x}_i, s_{it})$, so we do that here for the purposes of obtaining a simple test. Then the conditional expectation we need to estimate is

$$\mathrm{E}(v_{it}|\mathbf{x}_i, s_{it} = 1) = \mathrm{E}(v_{it}|\mathbf{x}_i, v_{it} > -\mathbf{x}_i\delta_t). \tag{3.13}$$

Assuming now that the variance of v_{it} is unity, we have

$$\mathrm{E}(v_{it}|\mathbf{x}_i, v_{it} > -\mathbf{x}_i\delta_t) = \lambda(\mathbf{x}_i\delta_t), \tag{3.14}$$

where $\lambda(\cdot)$ denotes the inverse Mills ratio. This leads to the following procedure:

Procedure 3.2 (s_{it} is observed; valid under Assumption 1′):

(i) For each t, estimate the equation

$$\mathrm{P}(s_{it} = 1|\mathbf{x}_i) = \Phi(\mathbf{x}_i\delta_t), \tag{3.15}$$

using standard probit. For $s_{it} = 1$, compute $\hat{\lambda}_{it} \equiv \lambda(\mathbf{x}_i\hat{\delta}_t)$.

(ii) Estimate Eq. (3.9) with $\ddot{\lambda}_{it} = \hat{\lambda}_{it} - T_i^{-1}\sum_{r=1}^{T} s_{ir}\hat{\lambda}_{ir}$ in place of $\ddot{v}_{it}$, again using only those observations for which $s_{it} = 1$.

(iii) Test H_0: $\rho = 0$ using the t-statistic for ρ. Again, a serial correlation and heteroskedasticity-robust standard error is warranted unless (2.8) is maintained under H_0. ■

The same remarks concerning restricting the δ_{tr} – in this case using pooled probit – hold here as in the case where h_{it} is observed.

This procedure is different from the ones proposed by VN (1992). VN suggest variable addition tests in a random effects framework using the variables T_i, $\prod_{r=1}^{T} s_{ir}$, and $s_{i,t-1}$. The first two of these could not be implemented in a fixed effects framework because the added variables have no time variation. The third of these could be after dropping the first time period. There are other possibilities that can be used in place of $\hat{\lambda}_{it}$ in a variable addition test during fixed effects estimation; for example, $\sum_{r \neq t}^{T} s_{ir}$ and $\prod_{r \neq t}^{T} s_{ir}$. Such tests have the advantage of computational simplicity and the need to only observe $\boldsymbol{x}_{it}$ when $s = 1$. Procedure 3.2 can be viewed as an extension of Heckman's (1976) procedure to an unobserved effects framework.

4. Correcting for sample selection bias

We now consider correcting for selection bias, and again we consider the two cases that h_{it}^* is partially observed and that only s_{it} is observed. In the former case two different corrections are offered. The first is obtained directly from the testing Procedure 3.1 in Section 3, and it has the advantage of not specifying how α_i depends on $(\boldsymbol{x}_i, \boldsymbol{v}_i)$. The second procedure relaxes (implicit) assumptions on the errors in the selection equation by imposing a linearity assumption on the regression of α_i on $(\boldsymbol{x}_i, v_{it})$.

The regression equation is still given by (2.1). For interpretive reasons it is useful to think of Assumption 1 as also holding, but it is not directly assumed in what follows. We first formalize the selection mechanism.

Assumption 2. Define h_{it}^* as in (3.1), where v_{it} is independent of $\boldsymbol{x}_i$ and $v_{it} \sim \text{Normal}(0, \sigma_t^2)$. Let $h \equiv \max(0, h_{it}^*)$, $s_{it} \equiv 1[h_{it}^* > 0]$, $t = 1, \dots, T$. ■

Assumption 2 was one of the assumptions used to motivate the tests for selection bias in Section 3; unlike there, we must now take this assumption seriously. Further, we now need to be very careful with our assumptions about relationships among α_i, u_{it}, and v_{it} so as not to rule out structural selection equations of the type covered in Section 3.

Although we assume normality of v_{it} in Assumption 2, the temporal dependence in $\{v_{it}\colon t = 1, 2, \dots, T\}$ is entirely unrestricted. This is important for examples like (3.2) and (3.5) because – whether or not the a_{it} are independent across t – the v_{it} can never be counted on to be serially independent. Note also that the v_{it} are allowed to have different variances in different time periods.

4.1. Selection corrections when the selection variable is partially observed

We begin with an assumption that allows us to correct for selection bias in a fixed effects estimation framework. It is a combination of a conditional mean independence assumption and a linearity assumption.

Assumption 3. For some zero-mean random variable ζ_i and all $t = 1, 2, \dots, T$,

$$\mathrm{E}(u_{it} \mid \alpha_i, \zeta_i, \boldsymbol{x}_i, \boldsymbol{v}_i) = \mathrm{E}(u_{it} \mid \zeta_i, v_{it}) = \zeta_i + \rho v_{it}. \quad ■ \tag{4.1}$$

The first equality in (4.1) is a particular conditional mean independence assumption. Namely, Assumption 3 implies that there is a latent variable ζ_i such that, conditional on ζ_i, u_{it} is mean independent of $\boldsymbol{x}_i$ and v_{ir}, $r \neq t$. This assumption is aimed at the leading case of the structural selection model (3.2), where the errors a_{it} in the structural selection equation are zero-mean normally distributed, $(\boldsymbol{u}_i, \boldsymbol{a}_i)$ is independent of $(\alpha_i, \xi_i, \boldsymbol{x}_i)$, ξ_i has the representation in (3.3), and

$$\mathrm{E}(u_{it} | \boldsymbol{a}_i) = \rho a_{it}. \tag{4.2}$$

Then $\rho a_{it} = \mathrm{E}(u_{it} | \alpha_i, \xi_i, \boldsymbol{x}_i, \boldsymbol{a}_i) = \mathrm{E}(u_{it} | \alpha_i, c_i, \boldsymbol{x}_i, \boldsymbol{v}_i)$, or $\mathrm{E}(u_{it} | \alpha_i, c_i, \boldsymbol{x}_i, \boldsymbol{v}_i) = -\rho c_i + \rho v_{it}$. Thus, Assumption 3 holds by taking $\zeta_i = -\rho c_i$.

The linearity in v_{it} assumed in (4.1) can be relaxed by adding, for example, polynomials in v_{it}. This gives even more flexibility – and explicitly allows the joint distribution of (u_{it}, v_{it}) to be nonnormal – but we focus on the linear case for simplicity.

The limitations of Assumption 3 can be seen by studying (4.2) a little more closely. A sufficient condition for (4.2) with linear conditional expectations is that (u_{it}, a_{it}) is uncorrelated with a_{ir}, $r \neq t$ (although serial correlation in $\{u_{it}\}$ is unrestricted). Unfortunately, (4.2) can fail if there is serial correlation in $\{a_{it}\}$; below we offer a different assumption that explicitly allows such serial correlation.

Under Assumptions 2 and 3 we have

$$\mathrm{E}(y_{it} | \alpha_i, \zeta_i, \boldsymbol{x}_i, \boldsymbol{v}_i) = \omega_i + \boldsymbol{x}_{it}\beta + \rho v_{it}, \tag{4.3}$$

where $\omega_i \equiv \alpha_i + \zeta_i$. Arguments similar to this have appeared in Smith and Blundell (1986) and Vella (1992) in the context of limited dependent variable models in pure cross-section contexts under distributional assumptions. Here we derive it in a selection context under weaker conditions and, more importantly, without making restrictive assumptions on the unobserved components or the time series properties of $\{u_{it}: t = 1, 2, \ldots, T\}$.

The strategy for the selection correction follows from (4.3). Because $\boldsymbol{s}_i$ is a function of $(\boldsymbol{x}_i, \boldsymbol{v}_i)$, (4.3) implies that

$$\mathrm{E}(y_{it} | \omega_i, \boldsymbol{x}_i, \boldsymbol{v}_i, \boldsymbol{s}_i) = \omega_i + \boldsymbol{x}_{it}\beta + \rho v_{it}, \tag{4.4}$$

which means that $\boldsymbol{x}_{it}$ and v_{it} are strictly exogenous conditional on ω_i. Estimation of β and ρ proceeds *exactly* as in Procedure 3.1. But, because ρ is generally different from zero, the asymptotic variance of $\hat{\theta} \equiv (\hat{\beta}', \hat{\rho})'$ needs to be adjusted due to the preliminary estimation of δ.

Procedure 4.1.1 (h_{it} observed; valid under Assumptions 2 and 3):

Steps (i) and (ii) are carried out exactly as in Procedure 3.1.

(iii) To estimate the asymptotic variance of $\hat{\theta}$ using the results in the Appendix, define $\hat{\boldsymbol{w}}_{it} = (\ddot{\boldsymbol{x}}_{it}, \ddot{v}_{it})$ and $\hat{e}_{it} \equiv \ddot{y}_{it} - \ddot{\boldsymbol{x}}_{it}\hat{\beta} - \hat{\rho}\ddot{v}_{it}$. Define a $(K+1)\times$

$T(1 + TK)$ matrix by $G_{it} \equiv (0' \; z_{it}')'$, where 0 is a $K \times T(1 + TK)$ matrix of zeros and the $1 \times T(1 + TK)$ vector z_{it} is defined as

$$z_{it} \equiv \frac{1}{T_i}(s_{i1}x_i, \ldots, s_{i,t-1}x_i, \; -(T_i - 1)x_i, s_{i,t+1}x_i, \ldots, s_{iT}x_i).$$

Then

$$\hat{D} \equiv N^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \hat{w}_{it}' \hat{\theta}' G_{it}. \tag{4.5}$$

Next, for each t, let $\hat{r}_{it}$ be minus the inverse of the average estimated Hessian (over the entire cross-section) times the estimated score of the Tobit log-likelihood function for observation i; delete the last element of this vector because $\hat{\sigma}_t^2$ does not appear in $\hat{v}_{it}$. Thus, $\hat{r}_{it}$ is a $(1 + TK) \times 1$ vector for each i and t. The expressions for the average Hessian and score for each i, evaluated at $\hat{\delta}_t$, are given in Maddala (1983, Sect. 6.3). Form the $T(1 + TK) \times 1$ vector $\hat{r}_i$ by stacking $\{\hat{r}_{i1}, \hat{r}_{i2}, \ldots, \hat{r}_{iT}\}$. Obtain $\hat{A}$ as in the Appendix equation (A.3), define $\hat{q}_i$ and $\hat{p}_i$ as in (A.4), and use these to construct $\hat{B}$ in (A.5). Then $\widehat{\text{Avar}}(\hat{\theta}) = \hat{A}^{-1}\hat{B}\hat{A}^{-1}/N$. ■

Procedure 4.1.1 is useful when h_{it} is observed and heterogeneity and serial correlation in the errors of a structural selection model such as (3.2) can be ruled out. If the expected value of u_{it} given v_i depends on v_{ir}, $r \neq t$, even after conditioning on a time-constant effect, then Procedure 4.1.1 does not generally produce a consistent estimator of β. If we are willing to make an assumption about α_i, then we can replace (4.1) with an assumption that allows for heterogeneity and unrestricted serial correlation in the errors of both the selection and regression equations.

Assumption 3'. For $t = 1, 2, \ldots, T$,

(i) $$\mathrm{E}(u_{it} | x_i, v_{it}) = \mathrm{E}(u_{it} | v_{it}) = \mathrm{L}(u_{it} | v_{it}), \tag{4.6}$$

(ii) $$\mathrm{E}(\alpha_i | x_i, v_{it}) = \mathrm{L}(\alpha_i | 1, x_{i1}, \ldots, x_{iT}, v_{it}), \tag{4.7}$$

where $\mathrm{L}(\cdot|\cdot)$ denotes the linear projection operator. ■

Consider first Assumption 3'(i). The first equality in (4.6) states that u_{it} is mean independent of x_i conditional on v_{it}. The key observation is that, unlike in (4.1), the entire history $(v_{i1}, \ldots, v_{iT})$ does *not* appear in the first conditioning set in (4.6); this is done on purpose so that the nature of any serial dependence in the v_{it} is entirely unrestricted. The conditional mean independence assumption (4.6) always holds if (u_{it}, v_{it}) is independent of x_i, something usually maintained in selection contexts. (Recall that we have already assumed that v_{it} is independent of x_i in Assumption 2, but we have not required u_{it} to be independent of x_i.) Also, (4.6) imposes no restriction on the temporal dependence of $\{u_{it}\}$, or on how u_{it} relates to v_{ir}, $r \neq t$.

The second equality in (4.6) is much less important, and could be relaxed. For simplicity, and because it is the leading case, we have assumed that $\mathrm{E}(u_{it} \mid v_{it})$ is linear, which allows us to write

$$\mathrm{E}(u_{it} \mid v_{it}) = \rho_t v_{it}, \tag{4.8}$$

for some scalar ρ_t. We can allow for more flexibility by, say, adding to (4.8) the quadratic term $v_{it}^2 - \sigma_t^2$. For simplicity we focus on (4.8).

Practically speaking, Assumption 3′(i) is much weaker than Assumption 3. Apparently, this comes at the cost of needing an assumption on the unobserved effect α_i, given here by Assumption 3′(ii). Without the term v_{it}, that assumption is similar to an assumption used by Chamberlain (1980) in the context of unobserved component probit models. Here, except for linearity of $\mathrm{E}(\alpha_i \mid \boldsymbol{x}_i, v_{it})$, the distribution of α_i given $(\boldsymbol{x}_i, v_{it})$ is otherwise unrestricted for all t; for example, conditional heteroskedasticity of unknown form is allowed. Assumption 3′(ii) is hardly for free, but it is notably less restrictive than – and implied by – the assumptions made in previous work on sample selection with panel data. In particular, Assumption 3′(ii) always holds when (α_i, v_{it}) conditional on $\boldsymbol{x}_i$ is bivariate normal with constant variance matrix and expectation linear in $\boldsymbol{x}_i$ (recall that $\mathrm{E}(v_{it} \mid \boldsymbol{x}_i) = 0$ and $\mathrm{E}(v_{it}^2 \mid \boldsymbol{x}_i) = \sigma_t^2$ already hold under Assumption 3). Unlike a random effects specification, Assumption 3′(ii) allows α_i to be correlated with $\boldsymbol{x}_i$. Also, that assumption places no restrictions on the serial dependence of $\{v_{it}\}$, and so it applies to all structural selection models discussed in Section 3 with arbitrary serial correlation in $\{a_{it}\}$.

To see how Assumption 3′ allows us to correct for selection bias, first note that for each t the linear predictor in (4.7) can always be written as

$$\mathrm{L}(\alpha_i \mid \boldsymbol{x}_i, v_{it}) = \psi_{t0} + \boldsymbol{x}_{i1}\psi_{t1} + \boldsymbol{x}_{i2}\psi_{t2} + \cdots + \boldsymbol{x}_{iT}\psi_{tT} + \phi_t v_{it}, \tag{4.9}$$

where ψ_{t0} is a scalar and ψ_{tr}, $r = 1, \ldots, T$, are $K \times 1$ vectors. This representation turns out to be too general for identifying the vector β. The key point is that, under Assumptions 2 and 3′, the ψ_{tr} are necessarily *constant* across t; that is, the coefficients on the $\boldsymbol{x}_{ir}$ are the same regardless of which v_{it} is also being conditioned on. This is crucial to the approach and it is a simple application of the law of iterated expecations: for any t,

$$\begin{aligned}\mathrm{E}(\alpha_i \mid \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}) &= \psi_{t0} + \boldsymbol{x}_{i1}\psi_{t1} + \boldsymbol{x}_{i2}\psi_{t2} + \cdots + \boldsymbol{x}_{iT}\psi_{tT} + \phi_t \mathrm{E}(v_{it} \mid \boldsymbol{x}_i) \\ &= \psi_{t0} + \boldsymbol{x}_{i1}\psi_{t1} + \boldsymbol{x}_{i2}\psi_{t2} + \cdots + \boldsymbol{x}_{iT}\psi_{tT} \qquad (4.10) \\ &= \psi_0 + \boldsymbol{x}_{i1}\psi_1 + \boldsymbol{x}_{i2}\psi_2 + \cdots + \boldsymbol{x}_{iT}\psi_T. \qquad (4.11)\end{aligned}$$

Eq. (4.10) follows because $\mathrm{E}(v_{it} \mid \boldsymbol{x}_i) = 0$ under Assumption 2, and Eq. (4.11) is just the fact that the coefficients in the linear projection of α_i onto $\boldsymbol{x}_i$ are necessarily time-invariant. Thus, we have established the important conclusion that the vector of coefficients appearing on $\boldsymbol{x}_i$ in $\mathrm{E}(\alpha_i|\boldsymbol{x}_i, v_{it})$ is the same for all t.

Given Assumptions 2 and 3′ we can easily identify β. Write

$$\mathrm{E}(y_{it}|\boldsymbol{x}_i, v_{it}) = \boldsymbol{x}_i\psi + \boldsymbol{x}_{it}\beta + \gamma_t v_{it}, \tag{4.12}$$

where $\gamma_t = \rho_t + \phi_t$. Thus, for each t we are left with a population regression that contains a linear combination of the explanatory variables from all time periods, the term $\boldsymbol{x}_{it}\beta$, and the additional term $\gamma_t v_{it}$. Without the term $\gamma_t v_{it}$, (4.12) is similar to Chamberlain (1982). The important difference is that Chamberlain can work with the linear projection $\mathrm{L}(y_{it}|\boldsymbol{x}_i)$, while here, because of the sample selection problem, we need (4.12) to be a conditional expectation. Still, we do not need to assume a particular conditional distribution or a constant conditional second moment for α_i.

We could extend (4.7) by adding a quadratic term, $v_{it}^2 - \sigma_t^2$, and even interaction terms such as $v_{it}\boldsymbol{x}_i$. It is easily shown using iterated expectations that these modifications for $\mathrm{E}(\alpha_i|\boldsymbol{x}_i, v_{it})$ are entirely consistent with Assumptions 2 and 3′(i).

We can also assume that α_i depends on $\boldsymbol{x}_i$ only through the time average. In this case, $\boldsymbol{x}_i$ would be replaced with $(1, \bar{\boldsymbol{x}}_i)$ in (4.12). This conserves on parameters but also imposes restrictions on the relationship between α_i and $\boldsymbol{x}_i$ that could be violated, especially if the $\boldsymbol{x}_{it}$ are trending.

In using (4.12) to motivate estimators for β it is important to remember that v_{ir} for $r \neq t$ is not included in the conditioning set (remember we did this on purpose to place as few restrictions as possible on time dependence in the idiosyncratic errors). This means that v_{it} is not strictly exogenous in (4.12). Consequently, estimators such as feasible GLS applied to (4.12) are generally inconsistent for β. Although other procedures could be used – such as minimum distance estimation – we will focus on the simplest consistent estimator, pooled OLS.

Because s_{it} is a function of $(\boldsymbol{x}_i, v_{it})$, (4.12) implies that

$$\mathrm{E}(y_{it}|\boldsymbol{x}_i, v_{it}, s_{it} = 1) = \boldsymbol{x}_i\psi + \boldsymbol{x}_{it}\beta + \gamma_t v_{it}. \tag{4.13}$$

This makes it clear that a pooled OLS procedure consistently estimates β.

Procedure 4.1.2 (h_{it} observed; valid under Assumptions 2 and 3′):

(i) Obtain $\hat{v}_{it}$ when $s_{it} = 1$ from T separate Tobit equations, just as in Procedure 3.1. For $s_{it} = 1$ define the $1 \times (1 + TK + K + T)$ vector

$$\hat{\boldsymbol{w}}_{it} = (1, \boldsymbol{x}_{i1}, \ldots, , \boldsymbol{x}_{iT}, \boldsymbol{x}_{it}, 0, \ldots, 0, \hat{v}_{it}, 0, \ldots, 0).$$

(ii) Obtain $\hat{\theta} = (\hat{\psi}', \hat{\beta}', \hat{\gamma}')'$ as the pooled OLS estimator in the equation

$$y_{it} = \hat{\boldsymbol{w}}_{it}\theta + \mathrm{error}_{it}, \qquad s_{it} = 1. \tag{4.14}$$

This yields

$$\hat{\theta} \equiv \left(\sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \hat{\boldsymbol{w}}'_{it} \hat{\boldsymbol{w}}_{it} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \hat{\boldsymbol{w}}'_{it} y_{it} \right), \tag{4.15}$$

which will be consistent and $\sqrt{N}$-asymptotically normal under Assumptions 2 and 3′, and standard regularity conditions.

(iii) Estimate $\text{Avar}(\hat{\theta})$ as follows: define the OLS residuals $\hat{e}_{it} = y_{it} - \hat{\boldsymbol{w}}_{it}\hat{\theta}$, $s_{it} = 1$. Define the matrix $\hat{\boldsymbol{D}}$ as in (4.5) except now $\boldsymbol{G}_{it}$ is the $(1 + TK + K + T) \times T(1 + TK)$ matrix

$$\boldsymbol{G}_{it} = \begin{pmatrix} 0 & 0 \dots 0 & 0 & 0 \dots 0 \\ 0 & 0 \dots 0 & \boldsymbol{Z}_{it} & 0 \dots 0 \end{pmatrix}; \tag{4.16}$$

each zero in the first row block of $\boldsymbol{G}_{it}$ is a $(1 + TK + K) \times (1 + TK)$ matrix and each zero in the second row block is a $T \times (1 + TK)$ matrix. (The matrix $\boldsymbol{Z}_{it}$ is in the tth column block of this matrix.) The $T \times (1 + TK)$ matrix $\boldsymbol{Z}_{it}$ is defined as

$$\boldsymbol{Z}_{it} = (0' \; 0' \; \dots \; 0' \;\; -\boldsymbol{x}'_i \; 0' \; \dots \; 0')', \tag{4.17}$$

where each zero in $\boldsymbol{Z}_{it}$ is a $1 \times (1 + TK)$ vector and the $1 \times (1 + TK)$ vector $-\boldsymbol{x}_i$ is in the t row. Now finish exactly as in Procedure 4.1.1 (that is, first form the $T(1 + TK) \times 1$ vector $\hat{\boldsymbol{r}}_i$, then construct $\hat{\boldsymbol{q}}_i$, $\hat{\boldsymbol{p}}_i$, $\hat{\boldsymbol{A}}$, and $\hat{\boldsymbol{B}}$). ■

Note that a test of the T restrictions H_0: $\gamma = 0$ is easily carried out by constructing a Wald statistic. This can be used as a test for selection bias, but it has a drawback compared with the test in Section 3. Namely, the test would maintain Assumption 3′(ii) under H_0 even though we know this assumption is not needed for fixed effects to consistently estimate β.

4.2. A selection correction when only the selection indicator is observed

If h_{it} is not observed then the fixed effects approach to the selection correction is not available, even under Assumptions 2 and 3. Nevertheless, Procedure 4.1.2 is easily modified to handle the case where only s_{it} is observed. Thus, we impose Assumptions 2 and 3′ in this subsection. Without loss of generality we now assume that $E(v_{it}^2) = 1$.

In place of (4.13) we must find the expectation of y_{it} given $(\boldsymbol{x}_i, s_{it} = 1)$. This yields

$$E(y_{it} | \boldsymbol{x}_i, s_{it} = 1) = \boldsymbol{x}_i \psi + \boldsymbol{x}_{it} \beta + \gamma_t \lambda(\boldsymbol{x}_i \delta_t), \tag{4.18}$$

leading to the following procedure.

Procedure 4.2 (s_{it} observed; valid under Assumptions 2 and 3′):

(i) For each $t = 1, 2, \ldots, T$ estimate Eq. (3.15) by standard probit. For $s_{it} = 1$ obtain the inverse Mills ratio $\lambda(\boldsymbol{x}_i\hat{\delta}_t)$. For $s_{it} = 1$ define $\hat{\boldsymbol{w}}_{it} \equiv (1, \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}, \boldsymbol{x}_{it}, 0, \ldots, 0, \hat{\lambda}_{it}, 0, \ldots, 0)$.

(ii) Obtain $\hat{\theta} \equiv (\hat{\psi}', \hat{\beta}', \hat{\gamma}')'$ as the pooled OLS estimator in Eq. (4.14).

(iii) Estimate $\mathrm{Avar}(\hat{\theta})$ as follows: first define the OLS residuals $\hat{e}_{it} \equiv y_{it} - \hat{\boldsymbol{w}}_{it}\hat{\theta}$, $s_{it} = 1$. Define the matrix $\hat{\boldsymbol{D}}$ as in (4.5), where $\boldsymbol{G}_{it}$ is defined as in (4.16) but $\boldsymbol{Z}_{it}$ is now defined by replacing $-\boldsymbol{x}_i$ with $\hat{v}_{it}\boldsymbol{x}_i$ in (4.17), where $\hat{v}_{it}$ is the derivative of $\lambda(\cdot)$ evaluated at $\boldsymbol{x}_i\hat{\delta}_t$. For each t let $\hat{\boldsymbol{r}}_{it}$ be the $(1 + TK) \times 1$ vector equal to minus the inverse of the average estimated Hessian times the estimated score of the probit log-likelihood function for observation i; these are given in Maddala (1983, Sect. 2.5). Now proceed exactly as in Procedure 4.1.1. ■

5. Concluding remarks

The selection corrections proposed in this paper have been derived for an unobserved effects linear model under the assumption of strict exogeneity of the regressors conditional on the unobserved effect. We used the strict exogeneity assumption to allow α_i to be correlated with $(\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{iT})$. If a random effects type specification is preferred, then one can simply replace $\boldsymbol{x}_i\psi$ in (4.12) and (4.18) with a constant; Procedure 4.1.2 or 4.2 can then be applied to the simplified equation. Even in a random effects framework the assumptions used here are much weaker than usual: the u_{it} can be arbitrarily heterogeneous and serially dependent.

A more difficult case to handle is when (2.1) contains a lagged dependent variable, which is necessarily correlated with α_i and is only observed for a subset of time periods. Such an extension would allow estimation of dynamic wage equations with unobserved effects while accounting for the sample selection bias due to current and lagged wage being observed for only those working in two consecutive periods. Preliminary work shows that, under reasonable extensions of the current assumptions, it is possible to allow for variables in (2.1) that are not strictly exogenous and that are observed for only a selected subset of time periods.

Appendix

In Sections 3 and 4 we rely on some general results concerning pooled OLS estimation using a selected sample. For brevity we present only the formulas for the asymptotic variance; derivations are available on request.

For each i let $\{(\boldsymbol{w}_{it}, y_{it}, s_{it})\colon t = 1, \ldots, T\}$ be a random draw from some population, where $\boldsymbol{w}_{it}$ is $1 \times \boldsymbol{G}$ and y_{it} and s_{it} are scalars. The pair $(\boldsymbol{w}_{it}, y_{it})$ is observed only if the selection indicator is unity: $s_{it} = 1$. Suppose that

$$\mathrm{E}(y_{it} \mid \boldsymbol{w}_{it}, s_{it} = 1) = \boldsymbol{w}_{it}\theta, \tag{A.1}$$

where θ is a $\boldsymbol{G} \times 1$ vector. Then, we can write (A.1) in error form as

$$y_{it} = \boldsymbol{w}_{it}\theta + e_{it}, \quad \mathrm{E}(e_{it} \mid \boldsymbol{w}_{it}, s_{it} = 1) = 0, \qquad t = 1, 2, \ldots, T.$$

The pooled OLS estimator on the selected sample can be written as

$$\hat{\theta} = \left(N^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \boldsymbol{w}_{it}' \boldsymbol{w}_{it}\right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \boldsymbol{w}_{it}' y_{it}\right),$$

and this is easily shown to be consistent. In the applications here we need to obtain the asymptotic variance of $\sqrt{N}(\hat{\theta} - \theta)$ when some elements of $\boldsymbol{w}_{it}$ are estimated in a preliminary stage, as in Newey (1984) and Pagan (1984). Suppose that $\boldsymbol{w}_{it} = \boldsymbol{w}_{it}(\delta)$, where δ is a $Q \times 1$ vector of unknown parameters (typically only a subset of $\boldsymbol{w}_{it}$ will actually depend on δ). Let $\hat{\delta}$ be a $\sqrt{N}$-asymptotically normal estimator of δ with representation

$$\sqrt{N}(\hat{\delta} - \delta) = N^{-1/2} \sum_{i=1}^{N} \boldsymbol{r}_i + \mathrm{o}_p(1), \tag{A.2}$$

where $\{\boldsymbol{r}_i\}$ is a $Q \times 1$ i.i.d. sequence with $\mathrm{E}(\boldsymbol{r}_i) = \boldsymbol{0}$. Typically, $\boldsymbol{r} = \boldsymbol{r}_i(\delta)$, and we estimate this as $\hat{\boldsymbol{r}}_i = \hat{\boldsymbol{r}}_i(\hat{\delta})$. A simplifying assumption that holds in our applications is that $\mathrm{E}[s_{it} \nabla_\delta \boldsymbol{w}_{it}(\delta)' e_{it}] = \boldsymbol{0}$, where $\nabla_\delta \boldsymbol{w}_{it}(\delta)'$ is the $G \times Q$ gradient of $\boldsymbol{w}_{it}(\delta)'$; this simplifies estimation of the asymptotic variance.

Let $\hat{\theta}$ now denote the OLS estimator from the regression on the selected sample, but where $\hat{\boldsymbol{w}}_{it} \equiv \boldsymbol{w}_{it}(\hat{\delta})$ is used in place of $\boldsymbol{w}_{it}$. It can be shown that

$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} \mathrm{Normal}(\boldsymbol{0}, \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}),$$

where

$$\boldsymbol{A} \equiv \mathrm{E}\left(\sum_{t=1}^{T} s_{it} \boldsymbol{w}_{it}' \boldsymbol{w}_{it}\right) \quad \text{and} \quad \boldsymbol{B} = \mathrm{var}(\boldsymbol{p}_i) = \mathrm{E}(\boldsymbol{p}_i \boldsymbol{p}_i');$$

the $G \times 1$ vector $\boldsymbol{p}_i$ is defined as $\boldsymbol{p}_i = \boldsymbol{q}_i - \boldsymbol{D}\boldsymbol{r}_i$, where $\boldsymbol{r}_i$ is from (A.2), $\boldsymbol{q}_i \equiv \sum_{t=1}^{T} s_{it} \boldsymbol{w}_{it}' e_{it}$ is a $G \times 1$ vector, and $\boldsymbol{D} \equiv \mathrm{E}(\sum_{t=1}^{T} s_{it} \boldsymbol{w}_{it}' \theta' \nabla_\delta \boldsymbol{w}_{it}(\delta)')$ is a $G \times Q$ matrix.

To estimate $\mathrm{Avar}(\hat{\theta}) \equiv \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}/N$, first define

$$\hat{\boldsymbol{A}} \equiv N^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \hat{\boldsymbol{w}}_{it}' \hat{\boldsymbol{w}}_{it} \quad \text{and} \quad \hat{\boldsymbol{D}} \equiv N^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \hat{\boldsymbol{w}}_{it}' \hat{\theta}' \nabla_\delta \boldsymbol{w}_{it}(\hat{\delta})', \tag{A.3}$$

where $\nabla_\delta \boldsymbol{w}_{it}(\hat{\delta})'$ is the $G \times Q$ gradient of $\hat{\boldsymbol{w}}_{it}(\delta)'$ evaluated at $\hat{\delta}$. Let $\hat{\boldsymbol{r}}_i \equiv \boldsymbol{r}_i(\hat{\delta})$ be estimates of $\hat{\boldsymbol{r}}_i$, and let $\hat{e}_{it} = y_{it} - \hat{\boldsymbol{w}}_{it}\hat{\theta}$ be the OLS residuals for all i and t such that $s_{it} = 1$. Then for each $i = 1, 2, \ldots, N$, define

$$\hat{\boldsymbol{q}}_i = \sum_{t=1}^{T} s_{it}\hat{\boldsymbol{w}}_{it}'\hat{e}_{it} \quad \text{and} \quad \hat{\boldsymbol{p}}_i = \hat{\boldsymbol{q}}_i - \hat{\boldsymbol{D}}\hat{\boldsymbol{r}}_i, \qquad i = 1, \ldots, N. \tag{A.4}$$

A consistent estimator of $\boldsymbol{B}$ is

$$\hat{\boldsymbol{B}} \equiv N^{-1} \sum_{i=1}^{N} \hat{\boldsymbol{p}}_i \hat{\boldsymbol{p}}_i'. \tag{A.5}$$

The asymptotic variance of $\hat{\theta}$ is estimated as $\widehat{\mathrm{Avar}}(\hat{\theta}) = \hat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{B}}\hat{\boldsymbol{A}}^{-1}/N$, and the asymptotic standard errors are obtained as the square roots of the diagonal elements of this matrix. If the elements of $\hat{\boldsymbol{w}}_{it}$ depending on $\hat{\delta}$ have coefficients in θ equal to zero – as when testing for exclusion of the generated regressors – then $\boldsymbol{D} = \boldsymbol{0}$ and one can take $\hat{\boldsymbol{D}} \equiv \boldsymbol{0}$.

References

Chamberlain, G., 1980, Analysis of covariance with qualitative data, Review of Economic Studies 47, 225–238.

Chamberlain, G., 1982, Multivariate regression models for panel data, Journal of Econometrics 18, 5–46.

Heckman, J., 1976, The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models, Annals of Economic and Social Measurement 5, 475–492.

Kiefer, N.M., 1980, Estimation of fixed effect models for time series of cross-sections with arbitrary intertemporal covariance, Journal of Econometrics 14, 195–202.

Maddala, G.S., 1983, Limited dependent and qualitative variables in econometrics (Cambridge University Press, Cambridge).

Mundlak, Y., 1978, On the pooling of time series and cross section data, Econometrica 46, 69–85.

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

Nijman, T. and M. Verbeek, 1992, Nonresponse in panel data: the impact on estimates of a life cycle consumption function, Journal of Applied Econometrics 7, 243–257.

Pagan, A., 1984, Econometric issues in the analysis of regressions with generated regressors, International Economic Review 25, 221–247.

Smith, R. and R. Blundell, 1986, An exogeneity test for a simultaneous Tobit model with an application to labor supply, Econometrica 54, 679–685.

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

Vella, F., 1992, Simple tests for sample selection bias in censored and discrete choice models, Journal of Applied Econometrics 7, 413–421.

Vella, F. and M. Verbeek, 1992, Estimating and testing simultaneous equation panel data models with censored endogenous variables, Unpublished manuscript (Department of Economics, Rice University, Houston, TX).

White, H., 1994, Estimation, inference and specification analysis (Cambridge University Press, Cambridge).

Zabel, J.E., 1992, Estimating fixed and random effects with selectivity, Economics Letters 40, 269–272.