Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Estimating panel data models in the presence of endogeneity and selection

Anastasia Semykina[a,*], Jeffrey M. Wooldridge[b]

[a] Department of Economics, Florida State University, Tallahassee, FL 32306-2180, United States

[b] Department of Economics, Michigan State University, East Lansing, MI 48824-1038, United States

ARTICLE INFO

Article history:
Received 2 May 2008
Received in revised form
15 February 2010
Accepted 1 March 2010
Available online 15 March 2010

Keywords:
Fixed effects
Instrumental variables
Sample selection
Mills ratio
Semiparametric

ABSTRACT

We consider estimation of panel data models with sample selection when the equation of interest contains endogenous explanatory variables as well as unobserved heterogeneity. Assuming that appropriate instruments are available, we propose several tests for selection bias and two estimation procedures that correct for selection in the presence of endogenous regressors. The tests are based on the fixed effects two-stage least squares estimator, thereby permitting arbitrary correlation between unobserved heterogeneity and explanatory variables. The first correction procedure is parametric and is valid under the assumption that the errors in the selection equation are normally distributed. The second procedure estimates the model parameters semiparametrically using series estimators. In the proposed testing and correction procedures, the error terms may be heterogeneously distributed and serially dependent in both selection and primary equations. Because these methods allow for a rather flexible structure of the error variance and do not impose any nonstandard assumptions on the conditional distributions of explanatory variables, they provide a useful alternative to the existing approaches presented in the literature.

1. Introduction

Due to the increased availability of longitudinal data and recent theoretical advances, panel data models have become widely used in applied work in economics. In many applications, particularly when the cross-sectional unit is a person, family, or firm, the panel data set is unbalanced. Standard methods such as fixed effects and random effects are easily modified to allow for unbalanced panels, but simply implementing the algebraic modifications ignores an important question: Why is the panel unbalanced? If the missing time periods result from self-selection, applying standard methods may result in inconsistent estimation.

A number of studies have addressed the problems of heterogeneity and selectivity under the assumption of strictly exogenous explanatory variables. Various tests of selection bias were proposed by Verbeek and Nijman (1992), Wooldridge (1995) and Hsiao et al. (2008). Wooldridge (1995) proposes a correction procedure, where the errors in the selection equation are assumed to be normally distributed; however, heterogeneous error distributions and arbitrary serial correlation are permitted. Kyriazidou (1997) considers a semiparametric approach, which relies on the conditional exchangeability assumption and removes both the unobserved effects and selection terms by differencing.[1] Rochina-Barrachina (1999) uses differencing to eliminate the time-constant unobserved effect, and assumes a trivariate normal error distribution to model the selection correction term.

The estimators of Wooldridge (1995), Kyriazidou (1997) and Rochina-Barrachina (1999) help us to resolve the endogeneity issues that arise because of the correlated unobserved effects. However, endogeneity biases may also arise due to a nonzero correlation between explanatory variables and idiosyncratic errors. This type of endogeneity can become an issue due to omission of relevant time-varying factors, simultaneous responses to idiosyncratic shocks, or measurement error. The resulting biases cannot be removed via differencing or fixed effects estimation, and hence, require special consideration.

An estimator that allows for endogenous variables was proposed by Vella and Verbeek (1999), who provide a method for estimating panel data models with censored endogenous regressors and selection. Kyriazidou (2001) considers estimation of dynamic panel data models with selection. Lewbel (unpublished manuscript) proposes an estimator based on the assumption that one of the explanatory variables is conditionally independent of unobserved heterogeneity and idiosyncratic errors and is conditionally continuously distributed on a large support. The approach

* Corresponding author. Tel.: +1 850 644 4557; fax: +1 850 644 4535.
E-mail addresses: asemykina@fsu.edu (A. Semykina), wooldri1@msu.edu (J.M. Wooldridge).

[1] The conditional exchangeability assumption does not always hold in practice — for example, if error variances change over time. Additionally, identification problems may arise when using Kyriazidou's estimator. For a detailed discussion of these issues, see Dustmann and Rochina-Barrachina (2007).

0304-4076/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2010.03.039

employs weighting to address selection and removes unobserved effects via differencing. Extensions of Kyriazidou's (1997) estimator were considered by Askildsen et al. (2003), Charlier et al. (2001) and Dustmann and Rochina-Barrachina (2007). Dustmann and Rochina-Barrachina (2007) also suggest combining generalized method of moments (GMM) with Rochina-Barrachina's (1999) estimator to account for endogeneity. Most related to the present paper are studies by Dustmann and Rochina-Barrachina (2007), Gonzalez-Chapela (unpublished manuscript) and Winder (unpublished manuscript), who consider extensions of Wooldridge's (1995) estimator. The latter two papers focus on applications and contain only a brief discussion of the underlying theory. Dustmann and Rochina-Barrachina (2007) provide a more elaborate theoretical discussion, but their approach involves using fitted values – a method that becomes problematic when nonlinear functions of explanatory variables are used.

In this study, we contribute to the existing literature in several ways. We extend the approach of Wooldridge (1995) to consider testing for contemporaneous selection bias when some variables may be correlated with idiosyncratic errors. The test is based on the within transformation and has an important advantage over alternative testing procedures because it is valid in the presence of arbitrary correlation between unobserved heterogeneity and explanatory variables. Moreover, we propose simple variable addition tests that can be used to detect selection biases that are due to violations of strict exogeneity.

We also propose two estimators that can be used to correct for selection bias. The first approach is parametric. Similar to Wooldridge (1995), it assumes normality of the errors in the selection equation to derive the correction term. As an alternative approach, we propose an estimator, where the selection problem is addressed semiparametrically using series estimators, as in Newey (2009). Both estimators permit heterogeneously distributed and serially dependent errors in the selection and primary equations. Thus, our methods are complementary to Kyriazidou's approach (Kyriazidou, 1997) in that they allow for arbitrary dynamics in the errors of both equations. Moreover, our estimators serve as a useful alternative to Lewbel's (unpublished manuscript) estimator because we do not require the availability of a conditionally independent covariate.

2. The model

Consider a correlated unobserved effects panel data model

$$y_{it} = x_{it}\beta + c_{i1} + u_{it1}, \quad t = 1, \ldots, T, \tag{1}$$

where x_{it} is a $1 \times K$ vector of explanatory variables, c_{i1} is the unobserved effect (which may be correlated with x_{it}), and u_{it1} is an idiosyncratic error. Assume that $\mathrm{E}(u_{it1}|x_{i1}, \ldots, x_{iT}, c_{i1}) \equiv \mathrm{E}(u_{it1}|x_i, c_{i1}) \neq 0$ for some element(s) of x_{it}, as occurs in simultaneous equations, models, models with measurement error and models with time-varying omitted variables. Additionally, define a $1 \times L$ $(L > K)$ vector of instruments, z_{it}, which is strictly exogenous conditional on c_{i1}: $\mathrm{E}(u_{it1}|z_{i1}, \ldots, z_{iT}, c_{i1}) \equiv \mathrm{E}(u_{it1}|z_i, c_{i1}) = 0$.[2] This permits for unspecified correlation between z_{it} and c_{i1}, but requires z_{it} to be uncorrelated with $\{u_{ir1} : r = 1, \ldots, T\}$. Unless stated otherwise, vectors x_{it} and z_{it} always contain a (possibly time-specific) intercept. In fact, z_{it} includes all the variables in x_{it} that are exogenous in (1). All variables in x_{it} and z_{it} are assumed to be time-varying.

We introduce selection (or incidental truncation) by defining a latent variable, s_{it}^*,

$$s_{it}^* = z_{it}\delta_t + c_{i2} + u_{it2}, \quad t = 1, \ldots, T. \tag{2}$$

Here c_{i2} is an unobserved effect and u_{it2} is an idiosyncratic error. Also, define

$$s_{it} = 1[s_{it}^* > 0] = 1[z_{it}\delta_t + c_{i2} + u_{it2} > 0], \tag{3}$$

where $1[\cdot]$ is the indicator function. By definition, s_{it} is a selection indicator that equals one if y_{it} is observed, and zero otherwise. In what follows, we assume that z_{it} is always observed, while endogenous variables in x_{it} may have missing values for $s_{it} = 0$.

3. Testing for selection bias

One way to test for contemporaneous selection bias is to follow the approach of Verbeek and Nijman (1992) and model $\mathrm{E}(c_{i1} + u_{it1}|z_{it}, s_{it})$ in Eq. (1). Alternatively, Eq. (1) can be augmented by functions of exogenous variables that belong to z_{it}, but not x_{it}, as suggested by Hsiao et al. (2008). In both cases, the t-test or the Wald test for joint significance of additional terms can be used as a test for selection bias.[3] However, here we build on the test proposed by Wooldridge (1995), which uses fixed effects to remove unobserved heterogeneity, and hence, permits arbitrary correlation between c_{i1} and z_{it}. In applications of IV methods to panel data, it is often important to allow the heterogeneity to be correlated with the instruments. In other words, the instruments are not exogenous with respect to the composite error, $c_{i1} + u_{it1}$, but only with respect to the idiosyncratic error, u_{it1}. Generally, time-variation in policies might be exogenous in cases where the level of a policy variable is systematically related to past history, as captured by c_{i1}.

The test is derived under the assumption

$$u_{it2}|z_i, c_{i2} \sim \mathrm{Normal}(0, 1), \quad t = 1, \ldots, T, \tag{4}$$

which allows arbitrary serial dependence in $\{u_{it2}\}$.

Define $\bar{z}_i \equiv T^{-1}\sum_{t=1}^{T} z_{it}$ and use Mundlak's (1978) modeling device to model c_{i2} as

$$c_{i2} = \bar{z}_i\xi + a_{i2}, \tag{5}$$

$$a_{i2}|z_i \sim \mathrm{Normal}(0, \sigma_a^2), \quad t = 1, \ldots, T, \tag{6}$$

which assumes that c_{i2} is related to z_i only through the time averages of the exogenous variables, while the remainder, a_{i2}, is independent of z_i. A popular less restrictive specification proposed by Chamberlain (1980) is: $c_{i2} = z_{i1}\xi_1 + \cdots + z_{iT}\xi_T + a_{i2}$. Mundlak's specification is a special case that imposes $\xi_1 = \cdots = \xi_T$. The advantage of Mundlak's model is that it conserves on degrees of freedom, which is important especially when T is large. In what follows, we use (5).

Combining (2) through (6) gives

$$s_{it} = 1[z_{it}\delta_t + \bar{z}_i\xi_t + v_{it2} > 0] \tag{7}$$

$$v_{it2}|z_i \sim \mathrm{Normal}(0, 1 + \sigma_a^2), \quad t = 1, \ldots, T, \tag{8}$$

where $v_{it2} = a_{i2} + u_{it2}$, and the time-varying coefficients on $\bar{z}_i$ can arise if $\mathrm{Var}(u_{it2})$ changes over time or if we use c_{it2} instead of c_{i2} in Eq. (3).

Furthermore, similar to Wooldridge (1995), suppose (u_{it1}, v_{i2}) is independent of (z_i, c_{i1}), where $v_{i2} = (v_{i12}, \ldots, v_{iT2})'$, and

[2] Strict inequality $(L > K)$ is needed for identification, as discussed in more detail in Section 4.

[3] Verbeek and Nijman (1992) and Hsiao et al. (2008) consider panel data models with exogenous explanatory variables.

(u_{it1}, v_{it2}) is independent of $(v_{i12}, \ldots, v_{i,t-1,2}, v_{i,t+1,2}, \ldots, v_{iT2})$. Then, if $\mathrm{E}(u_{it1}|v_{it2})$ is linear,

$$\mathrm{E}(u_{it1}|z_i, c_{i1}, v_{i2}) = \mathrm{E}(u_{it1}|v_{i2}) = \mathrm{E}(u_{it1}|v_{it2}) = \rho v_{it2}, \quad t = 1, \ldots, T, \tag{9}$$

where, for now, we assume ρ to be constant across time. Independence of v_{i2} and c_{i1} would not be a good assumption if v_{i2} contains an unobserved effect, as we expect, but, at this point, we are using these assumptions to motivate a test for selection bias. In Section 4 we will be more formal about stating assumptions used for a consistent correction procedure.

The proposed test is based on the fixed effects two-stage least squares (FE-2SLS) estimator, which is consistent on an unbalanced panel only if $\mathrm{E}(u_{it1}|z_i, c_{i1}, s_i) = 0$, with $s_i \equiv (s_{i1}, \ldots, s_{iT})$. If selection is not random, this expectation generally depends on s_{it} and z_{it}, $t = 1, \ldots, T$. Under the previous assumptions, however, we can write

$$\mathrm{E}(u_{it1}|z_i, c_{i1}, s_i) = \rho \mathrm{E}(v_{it2}|z_i, c_{i1}, s_i) = \rho \mathrm{E}(v_{it2}|z_i, s_{it}), \quad t = 1, \ldots, T, \tag{10}$$

so that the augmented primary equation is

$$y_{it} = x_{it}\beta + c_{i1} + \rho \mathrm{E}(v_{it2}|z_i, s_{it}) + e_{it1}, \quad t = 1, \ldots, T, \tag{11}$$

where, by construction, $\mathrm{E}(e_{it1}|z_i, c_{i1}, s_i) = 0$, $t = 1, \ldots, T$. It follows that, if we knew $\mathrm{E}(v_{it2}|z_i, s_{it})$, then a test for selection bias is obtained by testing $H_0 : \rho = 0$ in (11), which can be estimated by FE-2SLS. In fact, we need only $\mathrm{E}(v_{it2}|z_i, s_{it} = 1)$, which from the usual probit calculation is equal to $\lambda(z_{it}\delta_t^a + \bar{z}_i\xi_t^a)$, where $\delta_t^a \equiv \frac{\delta_t}{\sqrt{1+\sigma_a^2}}$, $\xi_t^a \equiv \frac{\xi_t}{\sqrt{1+\sigma_a^2}}$, and $\lambda(\cdot)$ denotes the inverse Mills ratio. To sum up, the following procedure can be used to test for sample selection bias:

Procedure 3.1. (i) For each t, use probit to estimate the equation: $\mathrm{P}(s_{it} = 1|z_i) = \Phi(z_{it}\delta_t^a + \bar{z}_i\xi_t^a)$. Use the resulting estimates to obtain $\hat{\lambda}_{it} \equiv \lambda(z_{it}\hat{\delta}_t^a + \bar{z}_i\hat{\xi}_t^a)$.

(ii) For $s_{it} = 1$, use FE-2SLS to estimate the equation: $y_{it} = x_{it}\beta + c_{i1} + \rho\hat{\lambda}_{it} + e_{it1}$; use $z_{it1}, \hat{\lambda}_{it}$ as instruments, where $z_{it1} \subset z_{it}$ but z_{it1} has at least dimension K. In addition to $\hat{\lambda}_{it}$, one can also add the interactions of $\hat{\lambda}_{it}$ with time dummies to allow ρ to be different across t.

(iii) Use the t-statistic to test $H_0 : \rho = 0$, or the Wald test to test $H_0 : \rho_1 = \cdots = \rho_T = 0$.

For the purpose of testing, the variance–covariance matrix does not need to be adjusted for the first-step estimation. However, the test should be carried out using statistics that are robust to heteroskedasticity and serial correlation in $\{u_{it1}\}$. See Wooldridge (1995) for the case of strictly exogenous regressors.

The procedure allows for choosing z_{it1} – the instruments for x_{it} – as a strict subset of z_{it}. Implicitly, 2SLS estimates a reduced form for endogenous elements of x_{it}. While only fitted values are used, the finite-sample performance of the test may be enhanced by allowing z_{it} to contain at least one element that affects selection but is not used directly as an IV for x_{it}. Plus, the procedure is convincing only if we have at least one IV for each endogenous element of x_{it} and then another exogenous variable that affects selection. But choosing z_{it1} as a strict subset of z_{it}, a researcher is forced to distinguish between which exogenous variables would be good instruments in the absence of selection, and which exogenous variables help predict selection.

The test summarized by Procedure 3.1 only checks for contemporaneous selection. Several simple variable addition tests can be used to test whether u_{it1} is correlated with s_{ir}, $r \neq t$. Specifically, we can add time-varying functions of selection indicators as explanatory variables and obtain simple t or joint Wald tests. For example, we can add $s_{i,t-1}$ or $s_{i,t+1}$ to (1) and test their significance. Two other possibilities are $\sum_{r=1}^{t-1} s_{ir}$ and $\sum_{r=t+1}^{T} s_{ir}$. For cases of attrition, where attrition is an absorbing state, neither $s_{i,t-1}$ nor $\sum_{r=1}^{t-1} s_{ir}$ varies across i for the selected sample, so they cannot be used to test for attrition bias. However, $s_{i,t+1}$ and $\sum_{r=t+1}^{T} s_{ir}$ can be used. These tests can be applied even when there is missing data on elements of x_{it} or z_{it}.

4. Correcting for selection bias

The FE-2SLS estimator used in Section 3 permits arbitrary correlation between s_{it} and c_{i1}, which is an attractive property of the fixed effects methods. However, such flexibility does come at a price. The FE-2SLS estimator is consistent only if selection is strictly exogenous conditional on the unobserved effect, which is unlikely to be true, particularly if the errors in the selection equation contain an unobserved effect, and hence, are inevitably serially correlated. In such a case, the conditional expectation $\mathrm{E}(u_{it1}|z_i, s_i)$ has a very complicated form. In contrast, the pooled two stage least squares (pooled 2SLS) estimator is consistent if $\mathrm{E}(c_{i1} + u_{it1}|z_{it}, s_{it}) = 0$, which allows for an arbitrary relation between u_{it1} and s_{ir}, $r \neq t$. We use this property of the pooled 2SLS estimator to derive the correction procedures.

Assume that the model summarized by Eqs. (1) through (3) holds. Additionally, model the unobserved effect as $c_{i1} = f(z_i) + a_{i1}$, where $f(\cdot)$ is a known function, and $\mathrm{E}(a_{i1}|z_i) = 0$. Although $f(\cdot)$ can be any known function, linearity of the conditional mean is likely to hold in linear models. Therefore, assume

$$c_{i1} = f(z_i) + a_{i1} = \bar{z}_i\eta + a_{i1}, \tag{12}$$

where again, a more flexible Chamberlain's (1980) specification is a possibility. Condition (12) is similar in spirit to the within transformation and produces the fixed-effects slope estimators when using a balanced panel with $x_{it} = z_{it}$, $t = 1, \ldots, T$ (Mundlak, 1978). In an unbalanced panel, (12) is different from the usual within transformation because it models c_{i1} as a function of $\bar{z}_i = T^{-1}\sum_{t=1}^{T} z_{it}$ (that are not distorted by selection) rather than $\tilde{z}_i = (\sum_{i=1}^{T} s_{it})^{-1}\sum_{t=1}^{T} s_{it}z_{it}$. Thus, the model in (12) is ideologically similar to fixed effects, but is free of selection biases, which makes it an attractive modeling device.

Given condition (12), we can plug into (1) and obtain

$$y_{it} = x_{it}\beta + \bar{z}_i\eta + a_{i1} + u_{it1} = x_{it}\beta + \bar{z}_i\eta + v_{it1}, \quad t = 1, \ldots, T, \tag{13}$$

where $v_{it1} \equiv a_{i1} + u_{it1}$, $\mathrm{E}(v_{it1}|z_i) = 0$. In an unbalanced panel, write this equation as

$$y_{it} = x_{it}\beta + \bar{z}_i\eta + \mathrm{E}(v_{it1}|z_i, s_{it}) + e_{it1}, \tag{14}$$

where $\mathrm{E}(e_{it1}|z_i, s_{it}) = 0$, $t = 1, \ldots, T$, by construction. If we know $\mathrm{E}(v_{it1}|z_i, s_{it})$, Eq. (14) can be consistently estimated by pooled 2SLS; the proof is straightforward.

Note that it is not necessary that $\mathrm{E}(e_{it1}|z_i, s_i) = 0$; in fact, generally e_{it1} will be correlated with selection indicators, s_{ir}, for $r \neq t$. This is a key benefit of the current approach: we can ignore selection in other time periods that might be correlated with v_{it1}. It is particularly important when $\mathrm{E}(v_{it1}|z_i, s_{i,t-1}) \neq 0$ even though $\mathrm{E}(v_{it1}|z_i, s_{it}) = 0$ might hold, or if $\mathrm{E}(v_{it1}|z_i, s_{it}) \neq 0$ and $\{v_{it2}\}$ are serially correlated.

4.1. Parametric correction

This and the following subsection focus on estimating Eq. (14). A formal set of assumptions that allow us to derive $\mathrm{E}(v_{it1}|z_i, s_{it})$ in a parametric setting is as follows:

Assumption 4.1.1. (i) z_{it} is always observed while (x_{it}, y_{it}) is observed when $s_{it} = 1$; (ii) selection occurs according to Eqs. (7) and (8); (iii) c_{i1} satisfies (12); (iv) $\mathrm{E}(v_{it1}|z_i, v_{it2}) \equiv \mathrm{E}(u_{it1}+a_{i1}|z_i, v_{it2}) = \mathrm{E}(u_{it1} + a_{i1}|v_{it2}) = \gamma v_{it2}, t = 1, \ldots, T$.

From parts (iii) and (iv) of Assumption 4.1.1 it follows that

$$y_{it} = x_{it}\beta + \bar{z}_i\eta + \gamma \mathrm{E}(v_{it2}|z_i, s_{it}) + e_{it1}, \tag{15}$$

where $\mathrm{E}(e_{it1}|z_i, s_{it}) = 0, t = 1, \ldots, T$. In Section 3, we have already derived $\mathrm{E}(v_{it2}|z_i, s_{it} = 1)$, which is all we need. With a slight abuse of notation, write the equation for $s_{it} = 1$ as

$$y_{it} = x_{it}\beta + \bar{z}_i\eta + \gamma\lambda_{it} + e_{it1}, \quad t = 1, \ldots, T. \tag{16}$$

The final estimating equation is obtained by substituting $\hat{\lambda}_{it}$ for λ_{it} in Eq. (16). We summarize the method for estimating β with the following procedure:

Procedure 4.1.1. (i) For each t, use probit to estimate the equation: $\mathrm{P}(s_{it} = 1|z_i) = \Phi(z_{it}\delta_t^a + \bar{z}_i\xi_t^a)$. Use the resulting estimates to obtain $\hat{\lambda}_{it}$.

(ii) For $s_{it} = 1$, use pooled 2SLS to estimate the equation: $y_{it} = x_{it}\beta + \bar{z}_i\eta + \gamma\hat{\lambda}_{it} + e_{it1}$; use $z_{it1}, \bar{z}_i, \hat{\lambda}_{it}$ as instruments, where $z_{it1} \subset z_{it}$ but z_{it1} has at least dimension K. In addition to $\hat{\lambda}_{it}$, one can also add the interactions of $\hat{\lambda}_{it}$ with time dummies to allow γ to be different across t.

(iii) Estimate the asymptotic variance as described in the Appendix.

The reasons for choosing z_{it1} to be a strict subset of z_{it} in obtaining a test for selection bias also hold here, and may even be more important. Suppose, for example, w_{it} is the only endogenous element of x_{it}. Then we should focus on what the reduced form of w_{it} would be apart from the sample selection issue. Viewed in this way, z_{it1} are the variables in the (population) reduced form for w_{it}, and then z_{it} should contain at least one additional variable that affects selection.

Instead of using analytical formulae for the asymptotic variance, one can apply "panel bootstrap". This involves resampling cross-sectional units (and all time periods for each unit sampled) and using the bootstrap sample to approximate the distribution of the parameter vector. Such a bootstrap estimator will be consistent for $N \to \infty$ and T fixed.

To perform Procedure 4.1.1, one should have a sufficient number of instruments, i.e. $L > K$. If this condition is not met, λ_{it} will be well approximated by a linear function of most of its range, which will lead to multicollinearity. Moreover, instead of employing a two-step estimator, one can stack the moment conditions from the two steps and estimate the parameters jointly using a more efficient GMM estimator, similar to Meijer and Wansbeek (2007).

4.2. Semiparametric correction

Here, we relax the assumption of normally distributed errors in the selection equation and propose a semiparametric estimator that is robust to a wide variety of actual error distributions. As demonstrated below, semiparametric correction permits identification of parameters in β only in the presence of an exclusion restriction. To emphasize this condition, let $z_{it} = (z_{it1}, z_{it2})$, where z_{it1} is a $1 \times L_1$ vector of exogenous variables (with $K \leq L_1 < L$) that are used to instrument for x_{it} in Eq. (16). It is assumed that all exogenous elements of x_{it} are included in z_{it1}. Because the intercept is not identified when estimating the model semiparametrically, the constant is excluded from x_{it}, z_{it1}, and z_{it}.

To derive the estimating equation, we formulate the following set of assumptions:

Assumption 4.2.1. (i) z_{it} is always observed while (x_{it}, y_{it}) is observed when $s_{it} = 1$; (ii) selection occurs according to Eq. (7); (iii) c_{i1} satisfies (12), so that Eq. (14) holds; (iv) the distribution of (v_{it1}, v_{it2}) is either independent of z_i or is a function of the selection index $(z_{it}\delta_t + \bar{z}_i\xi_t)$.

Assumption 4.2.1 does not specify a particular form of the error distribution, which makes the resulting estimator robust to variations in the distribution of (v_{it1}, v_{it2}). Moreover, it permits serial correlation in $\{v_{it2}\}$, as well as arbitrary relationships between v_{it1} and v_{is2} for $s \neq t$. Part (iv) of Assumption 4.2.1, albeit somewhat restrictive, is routinely used in the literature on semiparametric estimation (Powell, 1994).

From parts (ii) and (iv) of Assumption 4.2.1 it follows that

$$\begin{aligned}\mathrm{E}(v_{it1}|z_i, s_{it} = 1) &= \mathrm{E}(v_{it1}|v_{it2} > z_{it}\delta_t + \bar{z}_i\xi_t, z_i) \\ &= \varphi_t(z_{it}\delta_t + \bar{z}_i\xi_t) \equiv \varphi_{it},\end{aligned} \tag{17}$$

where $\varphi_t(\cdot)$ is an unknown function that may be different in each time period. Thus, combining Eqs. (14) and (17), we can write for $s_{it} = 1$:

$$y_{it} = x_{it}\beta + \bar{z}_i\eta + \varphi_{it} + e_{it1}, \quad t = 1, \ldots, T. \tag{18}$$

To estimate Eq. (18), we use an approach similar to the one proposed by Newey (2009) and employ series estimators to approximate the unknown function $\varphi_t(\cdot)$. Specifically, the focus is on power series and splines — estimators that are commonly used in economic applications. These are the polynomial and piecewise polynomial functions of the selection index, respectively, and can be easily implemented in practice. In case of splines, the attention is limited to splines with fixed evenly spaced knots.

For estimation purposes it may be preferred to limit the size of the selection index, which in the case of the power series estimator can be done by applying a strictly monotonic transformation $\tau_{it} \equiv \tau(z_{it}\delta_t + \bar{z}_i\xi_t)$. Several simple possibilities proposed by Newey (1994, 2009) are logit transformation $[\tau_{it} = \{1+\exp(z_{it}\delta_t + \bar{z}_i\xi_t)\}^{-1}]$, standard normal transformation $[\tau_{it} = \Phi(z_{it}\delta_t + \bar{z}_i\xi_t)]$, and the inverse Mills ratio. Such a transformation will not alter consistency of the estimator, but will reduce both the effect of outliers and multicollinearity in the approximating terms (Newey, 1994). Similarly, B-splines can be used in place of usual splines to avoid the multicollinearity problem.

Define the vector of M approximating functions $p(\tau_{it}) = (p_1(\tau_{it}), p_2(\tau_{it}), \ldots, p_M(\tau_{it}))$, and let $p_{it} \equiv p(\tau_{it})$. As discussed below, a consistent estimator of β can be obtained by applying pooled 2SLS to Eq. (18), where φ_{it} is replaced with a linear combination of approximating functions $p(\hat{\tau}_{it})$, $\hat{\tau}_{it} \equiv \tau(z_{it}\hat{\delta}_t + \bar{z}_i\hat{\xi}_t)$, assuming that consistent estimators of δ_t and ξ_t (and hence, τ_{it}) are available.

Before formulating consistency assumptions, it is convenient to write the estimator explicitly. Define vectors $w_{it} = (x_{it}, \bar{z}_i)$, $h_{it} = (z_{it1}, \bar{z}_i)$, $q_{it} = (z_{it}, \bar{z}_i)$, $\theta = (\beta', \eta')'$, and $\pi_t = (\delta_t', \xi_t')'$. Also, define linear projections of w_{it} and h_{it} on the approximating functions, $\hat{p}_{it} \equiv p(\hat{\tau}_{it})$:

$$\hat{m}_{it}^w = \hat{p}_{it}\left(\sum_{i=1}^N s_{it}\hat{p}_{it}'\hat{p}_{it}\right)^{-1}\left(\sum_{i=1}^N s_{it}\hat{p}_{it}'w_{it}\right),$$

$$\hat{m}_{it}^h = \hat{p}_{it}\left(\sum_{i=1}^N s_{it}\hat{p}_{it}'\hat{p}_{it}\right)^{-1}\left(\sum_{i=1}^N s_{it}\hat{p}_{it}'h_{it}\right), \quad t = 1, \ldots, T. \tag{19}$$

Using the results for partial regression, the estimator of θ can be written as

$$\hat{\theta} = \left\{ \sum_{t=1}^{T} \sum_{i=1}^{N} s_{it}(w_{it} - \hat{m}_{it}^{w})' h_{it} \left(\sum_{t=1}^{T} \sum_{i=1}^{N} s_{it}(h_{it} - \hat{m}_{it}^{h})' h_{it} \right)^{-1} \right.$$
$$\left. \times \sum_{t=1}^{T} \sum_{i=1}^{N} s_{it}(h_{it} - \hat{m}_{it}^{h})' w_{it} \right\}^{-1} \sum_{t=1}^{T} \sum_{i=1}^{N} s_{it}(w_{it} - \hat{m}_{it}^{w})' h_{it}$$
$$\times \left(\sum_{t=1}^{T} \sum_{i=1}^{N} s_{it}(h_{it} - \hat{m}_{it}^{h})' h_{it} \right)^{-1} \sum_{t=1}^{T} \sum_{i=1}^{N} s_{it}(h_{it} - \hat{m}_{it}^{h})' y_{it}. \quad (20)$$

Notice that linear projections $\hat{m}_{it}^{w}$ and $\hat{m}_{it}^{h}$ are semiparametric estimators of conditional means, $m_t^w \equiv \mathrm{E}(w_{it} | q_{it}\pi_t, s_{it} = 1)$ and $m_t^h \equiv \mathrm{E}(h_{it} | q_{it}\pi_t, s_{it} = 1)$, respectively. In other words, the estimator can be obtained by removing the selection effect via "demeaning", and then applying pooled 2SLS estimator to the transformed data. In this sense, the estimator in (20) is similar to Robinson's estimator (Robinson, 1988).

Given the expression in (20), we can specify the identification assumption:

Assumption 4.2.2. (i) For $A \equiv \sum_{t=1}^{T} \mathrm{E}[s_{it}(w_{it} - m_t^w)'(h_{it} - m_t^h)]$, $\mathrm{rank}(A) = K + L$; (ii) for $B \equiv \sum_{t=1}^{T} \mathrm{E}[s_{it}(h_{it} - m_t^h)'(h_{it} - m_t^h)]$, $\mathrm{rank}(B) = L_1 + L$; (iii) for $\Omega \equiv \mathrm{E}[(\sum_{t=1}^{T} s_{it}(h_{it} - m_t^h)' e_{it1})(\sum_{t=1}^{T} s_{it} e_{it1}(h_{it} - m_t^h))]$, $\mathrm{rank}(\Omega) = L_1 + L$.

Assumption 4.2.2 imposes certain restrictions on the instruments and explanatory variables. In particular, it implies that the number of variables in z_{it} should be strictly greater than the number of elements in z_{it1}. If this is not the case, "demeaned" instruments may be perfectly linearly related, so that matrices A, B and Ω will not have full rank. The usual requirement that demeaned instruments are sufficiently correlated with demeaned endogenous variables applies.

The following regularity conditions are the same as or similar to those stated in Newey (2009).

Assumption 4.2.3. (i) $\mathrm{E}(s_{it}\|w_{it}\|^{2+\nu}) < \infty$ for some $\nu > 0$, $t = 1, \ldots, T$, where the Euclidean norm is defined as $\|C\| = [\mathrm{tr}(C'C)]^{1/2}$; (ii) $\mathrm{E}(s_{it}\|h_{it}\|^2) < \infty$ for $t = 1, \ldots, T$; (iii) $\mathrm{Var}(w_{it} | q_{it}\pi_t, s_{it} = 1)$ is bounded for $t = 1, \ldots, T$; (iv) $\mathrm{Var}(h_{it} | q_{it}\pi_t, s_{it} = 1)$ is bounded for $t = 1, \ldots, T$; (v) $\mathrm{E}(e_{it1}^2 | q_{it}\pi_t, s_{it} = 1)$ is bounded for $t = 1, \ldots, T$.

Assumption 4.2.3 imposes restrictions on conditional and unconditional moments of the variables. These conditions permit the use of the law of large numbers and central limit theorem, as well as secure that series approximations lead to the consistent estimation of the approximated functions.

We further assume that a semiparametric estimator of π_t is available and satisfies the following assumption:

Assumption 4.2.4. For some ψ_{it}, $\sqrt{N}(\hat{\pi}_t - \pi_t) = N^{-1/2}\sum_{i=1}^{N} \psi_{it} + o_p(1) \xrightarrow{d} \mathrm{Normal}(0, V_t)$, and there exists an estimator $\hat{V}_t$, such that $\hat{V}_t \xrightarrow{p} V_t = \mathrm{E}(\psi_{it}\psi_{it}')$ for $t = 1, \ldots, T$.

Assumption 4.2.4 states that the first-step semiparametric estimator can be approximated as a sample average and is $\sqrt{N}$-consistent and asymptotically normal. Such estimators exist and are described in the literature, the estimators of Ichimura (1993) and Klein and Spady (1993) being the well-known examples.

The last assumption defines properties of φ_t, conditional variable means, and approximating functions.

Assumption 4.2.5. (i) Functions φ_t, m_t^w, and m_t^h are continuously differentiable in their argument of orders d, d_w and d_h, respectively, for $t = 1, \ldots, T$; (ii) The distribution of $\tau(q_{it}\pi_t)$ has an absolutely continuous component with p.d.f. bounded away from zero on its support, which is compact. The first and second derivatives of $\tau(q_{it}\hat{\pi}_t)$ with respect to the selection index are bounded for $\hat{\pi}_t$ in a neighborhood of π_t. All variables in q_{it} are bounded; (iii) $M \to \infty$, $N \to \infty$ so that $\sqrt{N}M^{-d-d_h+1} \to 0$ and (a) $p(\tau)$ is a power series, $d \geq 5$, and $M^7/N \to 0$; or (b) $p(\tau)$ is a spline of degree l, with $l \geq d_h - 1$, $d \geq 3$, and $M^4/N \to 0$.

Smoothness conditions in part (i) of Assumption 4.2.5 control for the bias when functions φ_t, m_t^w, and m_t^h are approximated by power series or splines. These conditions, combined with parts (iii)–(v) of Assumption 4.2.3, guarantee that $\hat{\varphi}_t$, $\hat{m}_t^w$, and $\hat{m}_t^h$ converge in probability to their true values as the number of approximating terms grows. Similarly, additional smoothness requirements in part (iii) of Assumption 4.2.5 are necessary to ensure consistent estimation of θ and the first derivative of φ_t. These smoothness assumptions are not restrictive and are commonly used in the literature. Part (ii) of Assumption 4.2.5 imposes restrictions on the transformation function and the variables in the selection equation. Boundedness of τ_{it} and h_{it} is not restrictive in practice, while the requirement for τ_{it} to have p.d.f. which is bounded away from zero is somewhat restrictive. Both conditions are needed, however, for series approximations to work.

Proposition 4.2.1. *Under Assumptions 4.2.1–4.2.5, $\hat{\theta}$ is consistent and $\sqrt{N}$-asymptotically normal for θ.*

The proof of Proposition 4.2.1 follows, with some modifications, from the proof provided in Newey (2009). Thus, under Assumptions 4.2.1–4.2.5, a consistent estimator of θ can be obtained by implementing the following procedure:

Procedure 4.2.1. (i) For each t, use a semiparametric estimator that satisfies Assumption 4.2.4 to obtain $\hat{\pi}_t$, $t = 1, \ldots, T$; compute $p(\hat{\tau}_{it})$.

(ii) For $s_{it} = 1$, estimate Eq. (18) (with φ_{it} replaced by the set of approximating functions) by pooled 2SLS using z_{it1}, $\bar{z}_i$, $p(\hat{\tau}_{it})$ as instruments. One can allow the selection correction to be different in each time period by adding the appropriate interaction terms in the regression.

(iii) Estimate the asymptotic variance as described in the Appendix.

5. Simulations

Monte Carlo simulations were used to study the finite-sample properties of the test and estimators summarized by Procedures 3.1, 4.1.1 and 4.2.1. Simulations were performed for $N = 200, 500$ and $T = 5, 10$.[4] In all experiments, the computed size of the test was close to the nominal size. The power of the test increases with N and T, as well as when $\mathrm{Corr}(u_1, u_2)$, increases. However, when the proportion of the variance due to unobserved heterogeneity rises, the selection bias becomes smaller, and hence, the power of the test is reduced.

When examining the properties of the proposed estimators, we compare their performance to OLS, 2SLS, FE, and FE-2SLS. In the absence of unobserved heterogeneity, endogeneity and selection, all estimators have very small biases; however, the correction procedures produce larger standard errors and hence, larger root mean square errors (RMSEs). Once we introduce unobserved heterogeneity, the correction procedures are preferred to OLS and 2SLS, but are inferior to FE and FE-2SLS because of the relatively large RMSE. After adding endogeneity and selection, the estimators discussed in Section 4 perform better than all other estimators. Even though their standard errors are relatively large, the computed biases remain small, so that the RMSEs are the smallest.

[4] The detailed results are available from the authors upon request.

6. Conclusion

In this paper, we considered estimation of panel data models with endogeneity and selection, where endogeneity is conditional on the unobserved effect. The methods discussed in this paper should provide a useful tool for applied economic research.[5] The proposed tests offer robust ways of testing for selection bias in the presence of endogenous regressors. The correction procedures provide an important alternative to some existing methods, as they allow for time-specific variances and arbitrary serial dependence in idiosyncratic errors. The considered semiparametric estimator shares the properties of all semiparametric estimators in the sense that it is robust to a wide variety of error distributions.

An avenue for further research is in relaxing the single-index assumption for the selection equation. Semiparametric and nonparametric procedures that relax the separability of the unobserved effect from the effects of other variables in the selection equation (see, for example Altonji and Matzkin, 2005), can add to the flexibility of the approach.

Acknowledgements

We thank the editor Cheng Hsiao, the associate editor and three anonymous referees for their useful comments.

Appendix

In this section, we present the formulae for the asymptotic variance of the estimators discussed in Procedures 4.1.1 and 4.2.1. Assume that either Eq. (16) or Eq. (18) holds. Define $\hat{w}_{it} = (x_{it}, \bar{z}_i, 0, \ldots, 0, \hat{d}_{it}, 0, \ldots, 0)$ and $\hat{h}_{it} = (z_{it1}, \bar{z}_i, 0, \ldots, 0, \hat{d}_{it}, 0, \ldots, 0)$, respectively, where $\hat{d}_{it} = \hat{\lambda}_{it}$ if using Procedure 4.1.1, and $\hat{d}_{it} = \hat{p}_{it}$ if using Procedure 4.2.1. In the primary equation, define $\hat{\theta} = (\hat{\beta}', \hat{\eta}', \hat{\gamma}_1', \ldots, \hat{\gamma}_T')'$, where $\hat{\gamma}_t$ is a scalar when using parametric correction, and it is an $M \times 1$ vector when using series approximations. In the selection equation, let $\hat{\pi}_t = (\hat{\delta}_t', \hat{\xi}_t')'$, and $\hat{\pi} = (\hat{\pi}_1', \hat{\pi}_2', \ldots, \hat{\pi}_T')'$.

Then,

$$\mathrm{A}\hat{\mathrm{v}}\mathrm{ar}(\hat{\theta}) = (\hat{C}'\hat{D}^{-1}\hat{C})^{-1}\hat{C}'\hat{D}^{-1}\hat{G}\hat{D}^{-1}\hat{C}(\hat{C}'\hat{D}^{-1}\hat{C})^{-1}/N,$$

where

$$\hat{C} = N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\hat{h}_{it}'\hat{w}_{it}, \qquad \hat{D} = N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\hat{h}_{it}'\hat{h}_{it},$$

$$\hat{G} = N^{-1}\sum_{i=1}^{N}\hat{g}_i\hat{g}_i',$$

with $\hat{g}_i = \sum_{t=1}^{T} s_i\hat{h}_{it}'\hat{e}_{it1} - \hat{F}\hat{\psi}_i$ and $\hat{e}_{it1} = y_{it} - \hat{w}_{it}\hat{\theta}$. If we use Procedure 4.1.1,

$$\hat{F} = -N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big[0, \ldots, 0, s_{it}\hat{h}_{it}'\hat{\gamma}_t q_{it}\hat{\lambda}_{it} \cdot (q_{it}\hat{\pi}_t + \hat{\lambda}_{it}), 0, \ldots, 0\Big]. \tag{21}$$

And, if we use Procedure 4.2.1,

$$\hat{F} = -N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[0, \ldots, 0, s_{it}\hat{h}_{it}'q_{it}\frac{\mathrm{d}\hat{p}_{it}\hat{\gamma}_t}{\mathrm{d}(q_{it}\hat{\pi}_t)}, 0, \ldots, 0\right]. \tag{22}$$

The expressions for $\hat{\psi}_i$ are obtained by stacking $\hat{\psi}_{it}$ for each i and depend on the first-step estimator. In the semiparametric case, $\hat{\psi}_{it}$ refers to the influence function mentioned in Assumption 4.2.4. In the parametric case, the formula for $\hat{\psi}_{it}$ follows from standard results for probit.

References

Altonji, J.G., Matzkin, R.L., 2005. Cross section and panel data estimators for nonseparable models with endogenous regressors. Econometrica 73, 1053–1102.

Askildsen, J.E., Baltagi, B.H., Holmas, T.H., 2003. Wage policy in the health care sector: a panel data analysis of nurses' labour supply. Health Economics 12, 705–719.

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

Charlier, E., Melenberg, B., van Soest, A., 2001. An analysis of housing expenditure using semiparametric models and panel data. Journal of Econometrics 101, 71–107.

Dustmann, C., Rochina-Barrachina, M.E., 2007. Selection correction in panel data models: an application to the estimation of females' wage equations. Econometrics Journal 10, 263–293.

Gonzalez-Chapela, J., 2004, On the price of recreation goods as a determinant of female labor supply (unpublished manuscript).

Hsiao, C., Shen, Y., Wang, B., Weeks, G., 2008. Evaluating the effectiveness of Washington state repeated job search services on the employment rate of prime-age female welfare recipients. Journal of Econometrics 145, 98–108.

Ichimura, H., 1993. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. Journal of Econometrics 58, 71–120.

Klein, R.L., Spady, R.H., 1993. An efficient semiparametric estimator for binary response models. Econometrica 61, 387–421.

Kyriazidou, E., 1997. Estimation of a panel data sample selection model. Econometrica 65, 1335–1364.

Kyriazidou, E., 2001. Estimation of dynamic panel data sample selection models. Review of Economic Studies 68, 543–572.

Lewbel, A., 2005, Simple endogenous binary choice and selection panel model estimators, Boston College (unpublished manuscript).

Meijer, E., Wansbeek, T., 2007. The sample selection model from a method of moments perspective. Econometric Reviews 26, 25–51.

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

Newey, W.K., 1994. The asymptotic variance of semiparametric estimators. Econometrica 62, 1349–1382.

Newey, W.K., 2009. Two-step series estimation of sample selection models. Econometrics Journal 12, S217–S229.

Powell, J.L., 1994. Estimation of semiparametric models. In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics, Vol. 4. North Holland, Amsterdam, pp. 2444–2521.

Robinson, P.M., 1988. Root-N-consistent semiparametric regression. Econometrica 56, 931–954.

Rochina-Barrachina, M.E., 1999. A new estimator for panel data sample selection models. Annales d'Economie et de Statistique 55/56, 153–181.

Vella, F., Verbeek, M., 1999. Two-step estimation of panel data models with censored endogenous variables and selection bias. Journal of Econometrics 90, 239–263.

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

Winder, K.L, 2004, Reconsidering the motherhood wage penalty (unpublished manuscript).

Wooldridge, J.M., 1995. Selection corrections for panel data models under conditional mean independence assumptions. Journal of Econometrics 68, 115–132.

[5] An application of the methods by the authors to estimating earnings equations for females is available upon request.