



Problems with the Control Variable Approach in Achieving Unbiased Estimates in Nonlinear Models in the Presence of Many Instruments

Jinyong Hahn¹ · Jerry Hausman²

Accepted: 27 October 2021 / Published online: 15 December 2021

© The Author(s), under exclusive licence to The Indian Econometric Society 2021

Abstract

We present a pseudo-panel model and argue that the control variable approach is subject to the many instrument problem, since it uses the predicted value of the endogenous variable. We show how the bias can be analytically characterized. Finally, we demonstrate the problems of split sample cross fitting.

Keywords Control variable · Many instruments · Nonlinear models

JEL Classification C33 · C36

Introduction

The “control variable” approach has been used in various nonlinear models to address the endogeneity problems.¹ The purpose of this paper is to examine (i) whether the control variable approach is also subject to the bias problem due to the many instruments problem as pointed out by Bekker (1994) for the linear models, and if so, (ii) whether the cross-fitting advocated by the modern machine learning type estimators² eliminates such bias. The many instrument problem is essentially a problem of bias due to many nuisance parameters, which can be understood by

¹ The control variable approach was developed as early as in Hausman (1978). See Blundell and Powell (2004), e.g., for recent application of the control variable approach.

² See, e.g., Chernozhukov et al. (2017).

✉ Jerry Hausman
jhausman@mit.edu

Jinyong Hahn
hahn@econ.ucla.edu

¹ UCLA Economics, BOX 951477, 8283 Bunche Hall, Los Angeles, CA 90095-1477, USA

² MIT Economics, 77 Massachusetts Avenue, Bldg E52-300, Cambridge, MA 02139, USA

using the incidental parameters problem in the panel data. Using a pseudo-panel analysis,³ we demonstrate that (i) the control variable approach is indeed subject to the many instrument problem; and (ii) the cross-fitting does not remove the bias. The negative result arises primarily because the control variable approach essentially used the fitted value of the endogenous variable which creates a finite sample bias problem. The bias and its correction in the control variable approach can in principle be understood from the perspective of the quite large general literature⁴ on bias correction, but because our focus is on the situation where there is a large number of instruments, we use a pseudo-panel analysis⁵ and answer these questions.

The bias of the control variable approach is relatively straightforward to understand. In linear simultaneous equations models, it is well-known that the 2SLS is equivalent to a control variable estimator. See, e.g., Hausman (1978). Therefore, it is quite natural to expect that the control variable approach is subject to the bias problem even in nonlinear models. The problem of cross fitting is not as immediately obvious, which we try to answer using the pseudo-panel analysis in the current paper. In the current section, we just explain what the cross fitting means in control variable estimation, and why it may be intuitively appealing.

Bias in the 2SLS estimator in finite samples has been long recognized. Nagar (1959) proposed the first estimator to remove this finite sample bias.⁶ As has been recognized more recently, the bias can be especially important in the many instrument problem which occurs often with increased size of data sets as Bekker (1994) and Hahn and Hausman (2002, 2003) demonstrate and Hansen et al. (2008) explore empirically. The bias problem in 2SLS is quite important with a number of subsequent papers proposing methods to remove the finite sample bias. The Nagar estimator removes bias by analytically adjusting the estimating equation, and it only holds for linear models. For a linear model

$$\begin{aligned}y &= X\theta + \varepsilon \\X &= Z\pi + \eta\end{aligned}$$

with the instrument matrix Z , the usual 2SLS estimator solves

$$0 = \hat{X}'(y - X\hat{b}), \quad (1)$$

while Nagar (1959) estimator solves

³ It is essentially the same bias problem analyzed by Cattaneo et al. (2019). Their analysis is predicated on the assumption that the first stage takes the form of an OLS estimation, which may be restrictive for certain applications. In order to facilitate the analysis, we present a pseudo-panel model to make the same point, and demonstrate the problems of split sample cross fitting. The panel analogy makes it easy to understand the reason why the split sample cross fitting does not remove the bias.

⁴ The long list includes Cox and Snell (1968), Firth (1993), and Rilstone et al. (1996), to list a few.

⁵ See Arellano and Hahn (2007), e.g., for a survey of the panel literature.

⁶ Biases in the finite samples may be removed either by correcting the estimator or by fixing the moments (estimating equation). The interpretation of Nagar's estimator is that it is a result of fixing the moment.

$$0 = \tilde{X}'(y - X\tilde{b}), \tag{2}$$

where $\hat{X} = PX$, $\tilde{X} = \left(P - \frac{k}{n-k}Q\right)X$, and $P = Z(Z'Z)^{-1}Z'$ and $Q = I - P$ denote the usual projection matrices. Here, n and k denote the number of observations and the number of instruments. Note that Nagar’s bias correction $\frac{k}{n-k}Q$ is roughly proportional to k/n , which can be understood to be the ratio between the “number of nuisance parameters” and the sample size, where the nuisance parameters here are the first stage OLS coefficients.

This approach can be motivated by observing that the moment underlying 2SLS is biased

$$E[(PX)'(y - X\theta)] = E[(Z\pi + P\eta)'\varepsilon] = k\sigma_{\varepsilon\eta},$$

where $\sigma_{\varepsilon\eta}$ denotes the covariance between the i th elements of ε and η , while the moment underlying Nagar’s estimator is unbiased

$$E\left[\left(\left(P - \frac{k}{n-k}Q\right)X\right)'(y - X\theta)\right] = E\left[\left(Z\pi + \left(P - \frac{k}{n-k}Q\right)\eta\right)'\varepsilon\right] = 0.$$

The lack of bias in Nagar (1959) moment can be understood from the perspective that the noise of estimating the instrument \hat{X} used in the moment (1) for 2SLS is correlated with the error ε in the second stage, which is eliminated by using the instrument \tilde{X} . As such, we can understand the cross-fit estimator using sample splitting as sharing a similar spirit as Nagar (1959) estimator. Specifically, we can see that the moment underlying the cross fit estimator

$$0 = \check{X}'(y - X\check{b})$$

has an unbiased moment, i.e., $E[\check{X}'(y - X\theta)] = 0$. Here,

$$\check{X} = \begin{bmatrix} \check{x}_{(1),1} \\ \vdots \\ \check{x}_{(1),m} \\ \check{x}_{(2),1} \\ \vdots \\ \check{x}_{(2),m} \end{bmatrix} = \begin{bmatrix} z'_1 \hat{\pi}_{(2)} \\ \vdots \\ z'_m \hat{\pi}_{(2)} \\ z'_{m+1} \hat{\pi}_{(1)} \\ \vdots \\ z'_{2m} \hat{\pi}_{(1)} \end{bmatrix},$$

where $n = 2m$, we split the sample into two equal sized subsamples, and $\hat{\pi}_{(1)}$ and $\hat{\pi}_{(2)}$ are first stage estimators based on the first and second subsamples.

We ask whether such an interpretation would lead to a reasonable inference for *nonlinear* models. Our conclusion is that it does not. We show that it is impossible in general to remove the bias of the moment equation by manipulating the first stage estimator alone. We do this analysis by considering nonlinear models of endogeneity with many instruments, and showing that the moment equation with the cross fit estimator does not eliminate a bias due to nonlinearity, and as a consequence, does not have the desired unbiasedness property.

Pseudo-Panel Model

Our model of interest is nonlinear models with endogeneity such as the probit model with endogenous regressors, where

$$y_i = 1(x_i \bar{\delta} + \varepsilon_i \geq 0),$$

$$x_i = z_i' \pi + \eta_i,$$

and (ε_i, η_i) have a bivariate normal distribution. The model has a built-in nonlinearity, and therefore, the endogeneity is probably best handled by the control variable approach. In the particular case of probit models, Rivers and Vuong (1988) solved the problem by writing

$$y_i = 1(x_i \delta + \rho(x_i - z_i' \pi) + \zeta_i),$$

$$x_i = z_i' \pi + \eta_i,$$

which generates a consistent estimator as long as (ε_i, η_i) have a bivariate normal distribution. (We assume that ζ_i has a standard normal distribution, i.e., the parameters δ and ρ reflect such normalization.)

In order to examine the consequence of many instruments, we adopt the strategy of interpreting the nuisance parameters (due to many instruments) as incidental parameters similar to the fixed effects in panel data. Therefore, we consider a special case that has a panel representation:

$$y_{it} = 1(x_{it} \delta + \rho(x_{it} - \alpha_i) + \zeta_{it})$$

$$x_{it} = \alpha_i + \eta_{it}. \tag{3}$$

This is a model where the first stage is characterized by n dummy instruments, and α_i denotes the first stage coefficient for the i th dummy instrument, i.e., $\pi = (\alpha_1, \alpha_2, \dots, \alpha_n)$.

The usual two step estimator can be understood to be the method of moments estimator

$$0 = E[x_{it} - \alpha_i]$$

$$0 = E \left[\begin{matrix} m(z_{it}, \theta, \alpha_i) x_{it} \\ m(z_{it}, \theta, \alpha_i) (x_{it} - \alpha_i) \end{matrix} \right]$$

where

$$m(z_{it}, \theta, \alpha_i) = \frac{y_{it} - \Phi(x_{it} \delta + \rho(x_{it} - \alpha_i))}{\Phi(x_{it} \delta + \rho(x_{it} - \alpha_i)) [1 - \Phi(x_{it} \delta + \rho(x_{it} - \alpha_i))]}$$

and Φ denotes the cumulative distribution function of a standard normal distribution.

Bias of Panel Two Step Estimator

In this section, we review the panel literature, and discuss how the bias of panel data estimator can be interpreted. The framework in this section provides a basis of understanding the problem of cross-fit estimator presented in Sect. 5.

The model in the previous section is a special case of the nonlinear panel data estimator

$$\begin{aligned}
 0 &= \sum_{i=1}^M \underline{v}(z_{it}, \hat{\theta}, \hat{\gamma}_i) \\
 0 &= \sum_{i=1}^n \sum_{t=1}^M \underline{u}(z_{it}, \hat{\theta}, \hat{\gamma}_i).
 \end{aligned}$$

For reasons that will become clearer later, we use the symbol M to denote the time series dimension of the panel data. Hahn and Newey (2004) and Arellano and Hahn (2007, 2016) are among the few who analyzed the finite sample bias from the large n , large T asymptotic approximation point of view. For our purpose, it is useful to make an explicit assumption that the fixed effects γ are multi-dimensional, and that v is of the same dimension as γ . We let J denote $\dim(\theta)$.

We provide a brief summary of the finite sample bias from the literature. It is convenient to analyze the general panel estimator in terms of the efficient score

$$\begin{aligned}
 0 &= \sum_{i=1}^M \underline{v}(z_{it}, \hat{\theta}, \hat{\gamma}_i) \\
 0 &= \sum_{i=1}^n \sum_{t=1}^M \underline{U}(z_{it}, \hat{\theta}, \hat{\gamma}_i)
 \end{aligned}$$

where

$$\begin{aligned}
 \underline{U}(z_{it}, \theta, \gamma_i) &= \underline{u}(z_{it}, \theta, \gamma_i) - \underline{\Delta}_i \underline{v}(z_{it}, \theta, \gamma_i), \\
 \underline{\Delta}_i &= E[u_{it}^{\gamma_i}] E[v_{it}^{\gamma_i}]^{-1}.
 \end{aligned}$$

Here, the $E[\underline{v}_{it}^{\gamma_i}] = E[\partial \underline{v}(z_{it}, \theta_0, \gamma_{i0}) / \partial \gamma_i']$ and $E[\underline{u}_{it}^{\gamma_i}] = E[\partial u(z_{it}, \theta_0, \gamma_{i0}) / \partial \gamma_i']$ are evaluated at the ‘truth’. The asymptotic distribution of $\sqrt{nM}(\hat{\theta} - \theta)$ is asymptotically normal with variance equal to

$$\left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \underline{\mathcal{I}}_i \right)^{-1} \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n E[\underline{U}_{it}^2] \right) \left(\left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \underline{\mathcal{I}}_i \right)^{-1} \right)' \tag{4}$$

and mean equal to

$$\left(\lim_{n \rightarrow \infty} \frac{\sqrt{n}}{\sqrt{M}} \right) \underline{B} \tag{5}$$

where

$$\underline{B} = \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \underline{\mathcal{I}}_i \right)^{-1} \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \underline{b}_i \right) \tag{6}$$

for

$$\underline{\mathcal{I}}_i \equiv -E \left[\frac{\partial U_{it}(\theta_0, \gamma_{i0})}{\partial \theta'} \right], \tag{7}$$

$$\begin{aligned} \underline{b}_i &= (\underline{b}_{i,1}, \dots, \underline{b}_{i,J})', \\ \underline{b}_{i,j} &= -\text{trace} \left((E[\underline{v}_{it}^{\gamma_i}])^{-1} E[\underline{v}_{it} \underline{U}_{it,j}^{\gamma_i}] \right) \\ &\quad + \frac{1}{2} \text{trace} \left(E[\underline{U}_{it,j}^{\gamma_i}] (E[\underline{v}_{it}^{\gamma_i}])^{-1} E[\underline{v}_{it} \underline{v}_{it}'] \left((E[\underline{v}_{it}^{\gamma_i}])^{-1} \right)' \right), \end{aligned} \tag{8}$$

and $\underline{b}_{i,j}$ and $\underline{U}_{it,j}$ denote the j -th components of \underline{b}_i and \underline{U}_{it} . In other words, the $1/M$ bias is given by the formula \underline{B}/M . Here, the $1/M$ bias denotes the approximate bias of $\hat{\theta}$ based on the asymptotic bias (5) of $\sqrt{nM}(\hat{\theta} - \theta)$. Because the number of fixed effects is equal to n , and the sample size is equal to nM , we can see that the ratio between the “number of nuisance parameters” and the sample size is $1/M$, and that it is of the same order of magnitude of the bias of 2SLS as discussed by Nagar (1959).

Applying this result to the two-step estimation case where $M = T$ and the fixed effects are scalars,

$$\begin{aligned} 0 &= \sum_{i=1}^T v(z_{it}, \hat{\alpha}_i) \\ 0 &= \sum_{i=1}^n \sum_{t=1}^T u(z_{it}, \hat{\theta}, \hat{\alpha}_i) \end{aligned}$$

we have the asymptotic variance of $\sqrt{nT}(\hat{\theta} - \theta)$ equal to

$$\left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathcal{I}_i \right)^{-1} \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n E[U_{it}^2] \right) \left(\left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathcal{I}_i \right)^{-1} \right)'$$

and the approximate bias equal to

$$\frac{1}{T} \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathcal{I}_i \right)^{-1} \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n b_i \right), \tag{9}$$

where

$$U(z_{it}, \theta, \alpha_i) = u(z_{it}, \theta, \alpha_i) - \Delta_i v(z_{it}, \alpha_i), \tag{10}$$

$$\Delta_i \equiv \frac{E[u_{it}^{\alpha_i}]}{E[v_{it}^{\alpha_i}]}, \tag{11}$$

$$\mathcal{I}_i \equiv -E\left[\frac{\partial U_{it}}{\partial \theta'}\right], \tag{12}$$

$$b_i = -\frac{E[v_{it}U_{it}^{\alpha_i}]}{E[v_{it}^{\alpha_i}]} + \frac{1}{2} \frac{E[U_{it}^{\alpha_i \alpha_i}]E[v_{it}^2]}{(E[v_{it}^{\alpha_i}])^2}. \tag{13}$$

Further Analysis of the Bias Formula

In this section, we analyze the formula (13) in two important models, and show that the bias formula simplifies for linear models, but not in the probit models. In Sect. 5, we will use this difference to illustrate why the bias in the probit models cannot be removed by cross fitting.

Because $E[v_{it}U_{it}^{\alpha_i}] = E[v_{it}u_{it}^{\alpha_i}] - \Delta_i E[v_{it}v_{it}^{\alpha_i}]$, and $E[U_{it}^{\alpha_i \alpha_i}] = E[u_{it}^{\alpha_i \alpha_i}] - \Delta_i E[v_{it}^{\alpha_i \alpha_i}]$, we can see that the bias formula simplifies a little bit if $\Delta_i = 0$ or $v_{it}^{\alpha_i}$ is constant. Under this condition, we can see

$$b_i = -\frac{E[v_{it}u_{it}^{\alpha_i}]}{E[v_{it}^{\alpha_i}]} + \frac{1}{2} \frac{E[u_{it}^{\alpha_i \alpha_i}]E[v_{it}^2]}{(E[v_{it}^{\alpha_i}])^2}.$$

The condition $\Delta_i = 0$ is satisfied if $E[u_{it}^{\alpha_i}] = 0$, i.e., under Neyman orthogonality. The condition that $v_{it}^{\alpha_i}$ is constant is satisfied if v_{it} is an affine function of α_i .

In order to understand these conditions, consider the panel model with n dummy IV's

$$\begin{aligned} y_{it} &= x_{it}\theta + \varepsilon_{it}, \\ x_{it} &= \alpha_i + \eta_{it}. \end{aligned} \tag{14}$$

If our 2SLS estimator solves

$$\begin{aligned} 0 &= \sum_{i=1}^T (x_{it} - \hat{\alpha}_i), \\ 0 &= \sum_{i=1}^n \sum_{t=1}^T \hat{\alpha}_i (y_{it} - x_{it}\hat{\theta}), \end{aligned} \tag{15}$$

we see that $v_{it}^{\alpha_i} = -1$. We also see that $E[u_{it}^{\alpha_i}] = E[y_{it} - x_{it}\theta] = E[\varepsilon_{it}] = 0$ so the condition $A_i = 0$ is also satisfied. The 2SLS for the pseudo panel model is special because $u_{it}^{\alpha_i} = 0$. This implies that the bias formula is very simple and satisfies $b_i = -E[v_{it}u_{it}^{\alpha_i}]/E[v_{it}^{\alpha_i}]$. This plays an important role in understanding the properties of split sample cross fitting for 2SLS.

We should recognize that these conditions are not satisfied for the probit model with endogenous regressors. In fact, the special nature of 2SLS, i.e., $u_{it}^{\alpha_i} = 0$, can be argued to be an implication of the IV type interpretation of 2SLS. If the 2SLS is interpreted to be a regression using the fitted value from the first stage as a regressor in the second stage, we see that the 2SLS solves

$$\begin{aligned} 0 &= \sum_{t=1}^T (x_{it} - \hat{\alpha}_i), \\ 0 &= \sum_{i=1}^n \sum_{t=1}^T \hat{\alpha}_i (y_{it} - \hat{\alpha}_i \hat{\theta}). \end{aligned} \tag{16}$$

Here, we can easily see that $u_{it}^{\alpha_i} \neq 0$ in general. In general, control variable approach requires that it be used as a regressor, so we should expect $u_{it}^{\alpha_i} \neq 0$ in general.

Split Sample Cross Fit Estimator

In this section, we will use the framework of Sect. 3, and analyze the bias of the cross fit estimator after sample splitting. We assume that the sample is split into two, and we use the estimate α_i from one subsample to be used as part of u in the second half of the sample cross fit. In other words, in order to understand the issue, we will assume now that the data consists of

$$z_i = (z_{i1}, \dots, z_{iT}) = (q_{i1}, \dots, q_{iM}, r_{iM}, \dots, r_{iM}),$$

i.e., we will assume that $T = 2M$, and write q and r for the first and second half of the observations. We will write the split sample cross fit estimator as

$$\begin{aligned} 0 &= \sum_{t=1}^M v(q_{it}, \hat{\alpha}_{1,i}) \\ 0 &= \sum_{t=1}^M v(r_{it}, \hat{\alpha}_{2,i}) \\ 0 &= \sum_{i=1}^n \sum_{t=1}^M (u(q_{it}, \hat{\theta}, \hat{\alpha}_{2,i}) + u(r_{it}, \hat{\theta}, \hat{\alpha}_{1,i})) \end{aligned}$$

with the recognition that $\hat{\alpha}_{1,i}$ and $\hat{\alpha}_{2,i}$ are estimators of $\alpha_{1,i} = \alpha_{2,i} = \alpha_i$. In order to see the resemblance to the panel model, we will write it

$$0 = \sum_{i=1}^M v_{(S)}(q_{it}, r_{it}, \hat{\theta}, \hat{\gamma}_i) = \sum_{i=1}^M \begin{bmatrix} v(q_{it}, \hat{\theta}, \hat{\alpha}_{1,i}) \\ v(r_{it}, \hat{\theta}, \hat{\alpha}_{2,i}) \end{bmatrix},$$

$$0 = \sum_{i=1}^M u_{(S)}(q_{it}, r_{it}, \hat{\theta}, \hat{\gamma}_i) = \sum_{i=1}^M \left(u(q_{it}, \hat{\theta}, \hat{\alpha}_{2,i}) + u(r_{it}, \hat{\theta}, \hat{\alpha}_{1,i}) \right).$$

In other words, the split sample cross fit estimator can be analyzed by adopting a perspective that the fixed effects are multidimensional. The result for the multi-dimensional fixed effects is already available from Arellano and Hahn (2016), which we will utilize here.

It can be shown⁷ that the efficient score is

$$\begin{aligned} \underline{U}_{(S)}(q_{it}, r_{it}, \theta, \alpha_{1,i}, \alpha_{2,i}) &= (u(q_{it}, \theta, \alpha_{2,i}) - \Delta_i v(q_{it}, \theta, \alpha_{1,i})) \\ &\quad + (u(r_{it}, \theta, \alpha_{1,i}) - \Delta_i v(r_{it}, \theta, \alpha_{2,i})), \end{aligned}$$

where the Δ_i is identical to the one in (11). Note that at $\alpha_{1,i} = \alpha_{2,i} = \alpha_{1,i}$, we see that the counterparts of \underline{U} and $\underline{\mathcal{I}}_i$ are

$$\begin{aligned} \underline{U}_{(S)}(q_{it}, r_{it}, \theta, \alpha_{1,i}, \alpha_{2,i}) &= U(q_{it}, \theta, \alpha_i) + U(r_{it}, \theta, \alpha_i), \\ \underline{\mathcal{I}}_{(S),i} &\equiv -E \left[\frac{\partial (U(q_{it}, \theta, \alpha_i) + U(r_{it}, \theta, \alpha_i))}{\partial \theta} \right] = 2\mathcal{I}_i, \end{aligned}$$

where the U and \mathcal{I}_i on the RHS are identical to the ones in (10) and (12). We therefore see that the asymptotic distribution of $\sqrt{nM}(\hat{\theta} - \theta)$ is normal with variance equal to

$$\frac{1}{2} \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathcal{I}_i \right)^{-1} \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \text{Var}(U_{it}) \right) \left(\left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathcal{I}_i \right)^{-1} \right)'$$

It follows that the asymptotic distribution of $\sqrt{nT}(\hat{\theta} - \theta) = \sqrt{n(2M)}(\hat{\theta} - \theta) = \sqrt{2}\sqrt{nM}(\hat{\theta} - \theta)$ is normal with variance equal to

$$\left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathcal{I}_i \right)^{-1} \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \text{Var}(U_{it}) \right) \left(\left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathcal{I}_i \right)^{-1} \right)'$$

In other words, the asymptotic variance of $\sqrt{nT}(\hat{\theta} - \theta)$ does not change.

As for the bias, we see that the counter part of b_i is given by

⁷ Algebraic details are collected in the appendix.

$$\underline{b}_{(S),i} = 2\Delta_i \frac{E[v_{it}v_{it}^{\alpha_i}]}{E[v_{it}^{\alpha_i}]} + E[U_{it}^{\alpha\alpha}] \frac{E[v_{it}^2]}{(E[v_{it}^{\alpha_i}])^2}$$

so

$$\underline{b}_{(S),i} = 2 \left(b_i + \frac{E[v_{it}u_{it}^{\alpha_i}]}{E[v_{it}^{\alpha_i}]} \right)$$

and the implied bias is

$$\begin{aligned} & \frac{1}{M} \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathcal{I}_{(S),i} \right)^{-1} \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \underline{b}_{(S),i} \right) \\ &= \frac{2}{T} \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathcal{I}_{(S),i} \right)^{-1} \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \underline{b}_{(S),i} \right) \\ &= \frac{2}{T} \left(2 \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathcal{I}_i \right)^{-1} \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \left(2\Delta_i \frac{E[v_{it}v_{it}^{\alpha_i}]}{E[v_{it}^{\alpha_i}]} + \frac{E[U_{it}^{\alpha\alpha}]E[v_{it}^2]}{(E[v_{it}^{\alpha_i}])^2} \right) \right) \tag{17} \\ &= \frac{1}{T} \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathcal{I}_i \right)^{-1} \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \left(2\Delta_i \frac{E[v_{it}v_{it}^{\alpha_i}]}{E[v_{it}^{\alpha_i}]} + \frac{E[U_{it}^{\alpha\alpha}]E[v_{it}^2]}{(E[v_{it}^{\alpha_i}])^2} \right) \right). \end{aligned}$$

We now compare the bias of the split sample cross fit estimator with the full sample estimator. We first rewrite the bias (9) with the full sample plug in estimator as

$$\frac{1}{T} \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathcal{I}_i \right)^{-1} \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \left(-\frac{E[v_{it}u_{it}^{\alpha_i}]}{E[v_{it}^{\alpha_i}]} + \Delta_i \frac{E[v_{it}v_{it}^{\alpha_i}]}{E[v_{it}^{\alpha_i}]} + \frac{1}{2} \frac{E[U_{it}^{\alpha\alpha}]E[v_{it}^2]}{(E[v_{it}^{\alpha_i}])^2} \right) \right) \tag{18}$$

using $U_{it}^{\alpha_i} = u_{it}^{\alpha_i} - \Delta_i v_{it}^{\alpha_i}$. Comparing (17) with (18), we can see that the split sample cross fit affects the bias in three ways:

1. It eliminates the bias $-E[v_{it}u_{it}^{\alpha_i}]/E[v_{it}^{\alpha_i}]$ due to the correlation between v_{it} and $u_{it}^{\alpha_i}$.
2. It magnifies the bias $\Delta_i E[v_{it}v_{it}^{\alpha_i}]/E[v_{it}^{\alpha_i}]$ due to the correlation between v_{it} and $v_{it}^{\alpha_i}$ by a factor of two.
3. It magnifies the bias $\frac{1}{2} E[U_{it}^{\alpha\alpha}]E[v_{it}^2]/(E[v_{it}^{\alpha_i}])^2$ due to the variance of v_{it} by a factor of two.

This is all intuitive. The finite sample bias is due to the noise of estimating α_i , which may be correlated with the second stage moment u . The split sample cross fit estimator severs this correlation, which explains the first effect. On the other hand, the split sample estimator effectively uses half the sample size for estimation of each α_i , which leads to the second and third effects.

We saw that in the pseudo panel 2SLS (15) with the IV interpretation, $\Delta_i = 0$ and $u_{it}^{\alpha_i}$. This implies that the bias of the full sample estimator takes a simple form $-E[v_{it}u_{it}^{\alpha_i}]/E[v_{it}^{\alpha_i}]$, and it is completely eliminated by the split sample cross fit.

Note that $E[v_{it}v_{it}^{\alpha_i}] = 0$ if $v_{it}^{\alpha_i}$ is constant as in (3). Even then, we should expect that (i) the bias is not removed by the cross fitting in general, although (ii) it is removed in the special case where $E[u_{it}^{\alpha_i}] = 0$.

Getting back to our panel rendition of the probit model with endogenous regressor, we see that

$$U_{it} = \begin{bmatrix} x_{it}m(z_{it}, \theta, \alpha_i) \\ (x_{it} - \alpha_i)m(z_{it}, \theta, \alpha_i) \end{bmatrix} + \Delta_i(x_{it} - \alpha_i)$$

where

$$\begin{aligned} \Delta_i &= E \left[\begin{array}{c} \frac{\partial(x_{it}m(z_{it}, \theta, \alpha_i))}{\partial\alpha_i} \\ \frac{\partial((x_{it} - \alpha_i)m(z_{it}, \theta, \alpha_i))}{\partial\alpha_i} \end{array} \right] \\ &= E \left[\begin{array}{c} x_{it}(\frac{\partial m(z_{it}, \theta, \alpha_i)}{\partial\alpha_i}) \\ (x_{it} - \alpha_i)(\frac{\partial m(z_{it}, \theta, \alpha_i)}{\partial\alpha_i}) - m(z_{it}, \theta, \alpha_i) \end{array} \right] \\ &= E \left[\begin{array}{c} x_{it}\phi(x_{it}\theta + \rho\eta_{it}) \\ (x_{it} - \alpha_i)\phi(x_{it}\theta + \rho\eta_{it}) \end{array} \right] \rho, \end{aligned}$$

where we use

$$E \left[\frac{\partial m(z_{it}, \theta, \alpha_i)}{\partial\alpha_i} \middle| x_{it}, \eta_{it} \right] = \phi(x_{it}\theta + \rho\eta_{it})\rho.$$

It can be seen that $U_{it}^{\alpha_i} \neq 0$, so we cannot expect the cross fitting estimator to remove the bias.

Note that the probit model is just one of the examples where the control variable is used as part of a nonlinear regression. We should therefore expect that (i) the control variable based estimator to have the many IV problem, and (ii) the problem is not solved by cross fitting. (In fact, the pseudo panel 2SLS (15) with the regression interpretation would be such that the bias due to $u_{it}^{\alpha_i}$ will not be eliminated by the split sample cross fit.)

Modified Objective Function for the Second Step

In this section, we review the panel literature, discuss a method of bias removal in the context of control variable estimation, which suggests how the bias can be corrected in principle. One can conjecture with high confidence that the bias can be removed from traditional methods of bias correction such as jackknife⁸, but it may be useful to find a

⁸ See Cattaneo et al. (2019). Hahn et al. (2019) recently reviewed higher order bias correction in a general framework. Cattaneo et al. (2019) results are predicated on first stage estimation based on many covariates, and therefore, more relevant for the many IV type situation.

simple alternative to these computationally intensive procedures. Panel literature discussed various methods of bias correction in the recent past, so one can imagine that it would work even in non-panel settings with some modifications. It is not clear how to frame an asymptotic sequence of models such that the biases in *non*-panel models with a large number of nuisance parameters can be easily understood and corrected, which we leave for future research.

Consider the moment (16) of the linear model, with a twist that the fitted value from the first stage is used as a regressor in the second stage. In particular, assume that $x_{it} = \alpha_i + \eta_{it}$, which implies that

$$v(z_{it}, \alpha_i) = x_{it} - \alpha_i$$

and that the bias formula (13) takes the form

$$b_i = E[\eta_{it}u_{it}^{\alpha_i}] + \frac{1}{2}E[u_{it}^{\alpha_i}]E[\eta_{it}^2].$$

If we further assume that η_{it} is i.i.d. over i and t , the formula further simplifies to

$$b_i = E[\eta_{it}u_{it}^{\alpha_i}] + \frac{\sigma_\eta^2}{2}E[u_{it}^{\alpha_i}].$$

In Sect. 5, we saw that the term $E[\eta_{it}u_{it}^{\alpha_i}]$ can be eliminated by sample split cross fit, but the second term actually gets magnified.

We consider changing the moment equation altogether, adopting Arellano and Hahn (2007) proposal to correct the bias of the moment equation. For this purpose, we assume that the moment u is obtained in the maximization of some objective function

$$\sum_i \sum_t \psi(z_{it}, \theta, \hat{\alpha}_i)$$

with respect to θ , i.e., assume that

$$u(z_{it}, \theta, \alpha_i) = \frac{\partial \psi(z_{it}, \theta, \alpha_i)}{\partial \theta}.$$

We then have

$$\begin{aligned} \frac{\partial u(z_{it}, \theta, \alpha_i)}{\partial \alpha_i} &= \frac{\partial}{\partial \theta} \left(\frac{\partial \psi(z_{it}, \theta, \alpha_i)}{\partial \alpha_i} \right) \\ \frac{\partial^2 u(z_{it}, \theta, \alpha_i)}{\partial \alpha_i^2} &= \frac{\partial}{\partial \theta} \left(\frac{\partial^2 \psi(z_{it}, \theta, \alpha_i)}{\partial \alpha_i^2} \right). \end{aligned}$$

This suggests that we can adopt the proposal in Arellano and Hahn (2007), and consider maximizing

$$\sum_i \left(\sum_i \psi(z_{it}, \theta, \hat{\alpha}_i) - \frac{1}{T} \sum_i \left(v(z_{it}, \hat{\alpha}_i) \frac{\partial \psi(z_{it}, \theta, \hat{\alpha}_i)}{\partial \alpha_i} + \frac{\hat{\sigma}_\eta^2}{2} \frac{\partial^2 \psi(z_{it}, \theta, \hat{\alpha}_i)}{\partial \alpha_i^2} \right) \right),$$

where $\hat{\sigma}_\eta^2 = \frac{1}{nT} \sum_{i=1}^T \sum_{t=1}^T (x_{it} - \hat{\alpha}_i)^2$ with the corresponding moment equation⁹

$$0 = \sum_i \left(\sum_i u(z_{it}, \hat{\theta}, \hat{\alpha}_i) - \frac{1}{T} \sum_i \left(v(z_{it}, \hat{\alpha}_i) u^{\alpha_i}(z_{it}, \hat{\theta}, \hat{\alpha}_i) + \frac{\hat{\sigma}_\eta^2}{2} u^{\alpha_i \alpha_i}(z_{it}, \hat{\theta}, \hat{\alpha}_i) \right) \right). \tag{19}$$

Summary

Using a pseudo-panel model, we have demonstrated that the control variable approach is subject to the many instrument problem, since it uses the predicted value of the endogenous variable. It is essentially the same bias problem analyzed by Cattaneo et al. (2019), who advocated the use of the jackknife to remove the higher order bias. It would be interesting to develop a method of analytic bias correction in the non-panel setting, which we leave for future research.

Detailed Derivations for Section 5

Noting that

$$\begin{aligned} \underline{\Delta}_i &= E \left[u_{(S),it}^{\gamma_i} \right] \left(E \left[v_{(S),it}^{\gamma_i} \right] \right)^{-1} \\ &= \left(E \left[u^{\alpha_i}(r_{it}, \theta, \alpha_{1,i}) \right], E \left[u^{\alpha_i}(q_{it}, \theta, \alpha_{2,i}) \right] \right) \left(E \left[\begin{matrix} v^{\alpha_i}(q_{it}, \theta, \alpha_{1,i}) & 0 \\ 0 & v^{\alpha_i}(r_{it}, \theta, \alpha_{2,i}) \end{matrix} \right] \right)^{-1} \\ &= \left(\frac{E \left[u^{\alpha_i}(z_{it}, \theta, \alpha_i) \right]}{E \left[v^{\alpha_i}(z_{it}, \theta, \alpha_1) \right]}, \frac{E \left[u^{\alpha_i}(z_{it}, \theta, \alpha_i) \right]}{E \left[v^{\alpha_i}(z_{it}, \theta, \alpha_1) \right]} \right) \equiv (\Delta_i, \Delta_i), \end{aligned}$$

where the Δ_i is identical to the one in (11), we see that the efficient score is

⁹ In the case where

$$\psi(z_{it}, \theta, \hat{\alpha}_i) = -\frac{1}{2} (y_{it} - \hat{\alpha}_i \theta)^2,$$

which corresponds to the 2SLS for the linear model

$$\begin{aligned} x_{it} &= \alpha_i + \eta_{it} \\ y_{it} &= x_{it} \theta + \varepsilon_{it}, \end{aligned}$$

we can show that the moment equation (19) produces Nagar’s estimator. Algebraic details are collected in the appendix, which also contains a discussion of the relationship to Cattaneo et al. (2019).

$$\begin{aligned} \underline{U}_{(S)}(q_{it}, r_{it}, \theta, \alpha_{1,i}, \alpha_{2,i}) &= (u(q_{it}, \theta, \alpha_{2,i}) + u(r_{it}, \theta, \alpha_{1,i})) - (\Delta_i v(q_{it}, \theta, \alpha_{1,i}) + \Delta_i v(r_{it}, \theta, \alpha_{2,i})) \\ &= (u(q_{it}, \theta, \alpha_{2,i}) - \Delta_i v(r_{it}, \theta, \alpha_{2,i})) + (u(r_{it}, \theta, \alpha_{1,i}) - \Delta_i v(q_{it}, \theta, \alpha_{1,i})) \\ &\equiv U_{(S,1)}(q_{it}, r_{it}, \theta, \alpha_{1,i}) + U_{(S,2)}(q_{it}, r_{it}, \theta, \alpha_{2,i}) \\ &= (u(q_{it}, \theta, \alpha_{2,i}) - \Delta_i v(q_{it}, \theta, \alpha_{1,i})) + (u(r_{it}, \theta, \alpha_{1,i}) - \Delta_i v(r_{it}, \theta, \alpha_{2,i})) \end{aligned}$$

Note that at $\alpha_{1,i} = \alpha_{2,i} = \alpha_{1,i}$, we see that the counterparts of \underline{U} and $\underline{\mathcal{I}}$ are

$$\begin{aligned} \underline{U}_{(S)}(q_{it}, r_{it}, \theta, \alpha_{1,i}, \alpha_{2,i}) &= U(q_{it}, \theta, \alpha_i) + U(r_{it}, \theta, \alpha_i), \\ \underline{\mathcal{I}}_{(S),i} &\equiv -E \left[\frac{\partial (U(q_{it}, \theta, \alpha_i) + U(r_{it}, \theta, \alpha_i))}{\partial \theta'} \right] = 2\mathcal{I}_i, \end{aligned}$$

where the U and \mathcal{I}_i on the RHS are identical to the ones in (10) and (10). We therefore see that the asymptotic distribution of $\sqrt{nM}(\hat{\theta} - \theta)$ is normal with variance equal to

$$\begin{aligned} &\left(\lim_{n \rightarrow \infty} 2 \frac{1}{n} \sum_{i=1}^n \mathcal{I}_i \right)^{-1} \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \text{Var}(U(q_{it}, \theta, \alpha_i) + U(r_{it}, \theta, \alpha_i)) \right) \left(\lim_{n \rightarrow \infty} 2 \frac{1}{n} \sum_{i=1}^n \mathcal{I}_i \right)^{-1'} \\ &= \frac{1}{2} \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathcal{I}_i \right)^{-1} \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \text{Var}(U_{it}) \right) \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathcal{I}_i \right)^{-1'}, \end{aligned}$$

where we used independence between q_{it} and r_{it} , which implies

$$\text{Var}(U(q_{it}, \theta, \alpha_i) + U(r_{it}, \theta, \alpha_i)) = \text{Var}(U(q_{it}, \theta, \alpha_i)) + \text{Var}(U(r_{it}, \theta, \alpha_i)).$$

As for the bias, we simplify notations a little bit by assuming that $J = 1$, and recognize that the counter parts of the components of \underline{b}_i are given by

$$\begin{aligned} \left(E \left[\underline{v}_{(S),it}^{\gamma_i} \right] \right)^{-1} &= \begin{bmatrix} (E[v^{\alpha_i}(q_{it}, \alpha_i)])^{-1} & 0 \\ 0 & (E[v^{\alpha_i}(r_{it}, \alpha_i)])^{-1} \end{bmatrix}, \\ E \left[\underline{v}_{(S),it} \underline{U}_{(S),it}^{\gamma_i} \right] &= E \left[\begin{bmatrix} v(q_{it}, \alpha_i) \\ v(r_{it}, \alpha_i) \end{bmatrix} \begin{bmatrix} u^{\alpha_i}(r_{it}, \theta, \alpha_{1,i}) & u^{\alpha_i}(q_{it}, \theta, \alpha_{2,i}) \\ -\Delta_i v^{\alpha_i}(q_{it}, \theta, \alpha_{1,i}) & -\Delta_i v^{\alpha_i}(r_{it}, \theta, \alpha_{2,i}) \end{bmatrix} \right] \\ &= \begin{bmatrix} E \left[v(q_{it}, \alpha_i) \begin{pmatrix} u^{\alpha_i}(r_{it}, \theta, \alpha_{1,i}) \\ -\Delta_i v^{\alpha_i}(q_{it}, \alpha_{1,i}) \end{pmatrix} \right] & E \left[v(q_{it}, \alpha_i) \begin{pmatrix} u^{\alpha_i}(q_{it}, \theta, \alpha_{2,i}) \\ -\Delta_i v^{\alpha_i}(r_{it}, \alpha_{2,i}) \end{pmatrix} \right] \\ E \left[v(r_{it}, \alpha_i) \begin{pmatrix} u^{\alpha_i}(r_{it}, \theta, \alpha_{1,i}) \\ -\Delta_i v^{\alpha_i}(q_{it}, \alpha_{1,i}) \end{pmatrix} \right] & E \left[v(r_{it}, \alpha_i) \begin{pmatrix} u^{\alpha_i}(q_{it}, \theta, \alpha_{2,i}) \\ -\Delta_i v^{\alpha_i}(r_{it}, \alpha_{2,i}) \end{pmatrix} \right] \end{bmatrix} \\ &= \begin{bmatrix} -\Delta_i E[v(q_{it}, \alpha_i) v^{\alpha_i}(q_{it}, \alpha_i)] & E[v(q_{it}, \alpha_i) u^{\alpha_i}(q_{it}, \theta, \alpha_i)] \\ E[v(r_{it}, \alpha_i) u^{\alpha_i}(r_{it}, \theta, \alpha_i)] & -\Delta_i E[v(r_{it}, \alpha_i) v^{\alpha_i}(r_{it}, \alpha_i)] \end{bmatrix}, \end{aligned}$$

so

$$\begin{aligned} & \left(\left(E \left[\underline{v}_{(S),it}^{\gamma_i} \right] \right)^{-1} E \left[\underline{v}_{(S),it} \underline{U}_{(S),it}^{\gamma_i} \right] \right) \\ &= \begin{bmatrix} \left(E \left[v^{\alpha_i} (q_{it}, \alpha_i) \right] \right)^{-1} & 0 \\ 0 & \left(E \left[v^{\alpha_i} (r_{it}, \alpha_i) \right] \right)^{-1} \end{bmatrix} \\ & \times \begin{bmatrix} -\Delta_i E \left[v_{it} v_{it}^{\alpha_i} \right] & E \left[v_{it} U_{it}^{\alpha_i} \right] + \Delta_i E \left[v_{it} v_{it}^{\alpha_i} \right] \\ E \left[v_{it} U_{it}^{\alpha_i} \right] + \Delta_i E \left[v_{it} v_{it}^{\alpha_i} \right] & -\Delta_i E \left[v_{it} v_{it}^{\alpha_i} \right] \end{bmatrix} \\ &= \begin{bmatrix} -\left(E \left[v^{\alpha_i} (q_{it}, \alpha_i) \right] \right)^{-1} \Delta_i E \left[v_{it} v_{it}^{\alpha_i} \right] & \left(E \left[v^{\alpha_i} (q_{it}, \alpha_i) \right] \right)^{-1} E \left[v_{it} U_{it}^{\alpha_i} \right] \\ \left(E \left[v^{\alpha_i} (r_{it}, \alpha_i) \right] \right)^{-1} E \left[v_{it} U_{it}^{\alpha_i} \right] & -\left(E \left[v^{\alpha_i} (r_{it}, \alpha_i) \right] \right)^{-1} \Delta_i E \left[v_{it} v_{it}^{\alpha_i} \right] \end{bmatrix}, \end{aligned}$$

and

$$-\text{trace} \left(\left(E \left[\underline{v}_{(S),it}^{\gamma_i} \right] \right)^{-1} E \left[\underline{v}_{(S),it} \underline{U}_{(S),it}^{\gamma_i} \right] \right) = 2\Delta_i \frac{E \left[v_{it} v_{it}^{\alpha_i} \right]}{E \left[v_{it}^{\alpha_i} \right]}. \tag{20}$$

We also have

$$\begin{aligned} E \left[\underline{U}_{(S),it}^{\gamma_i} \right] &= E \begin{bmatrix} u^{\alpha_i, \alpha_i} (r_{it}, \theta, \alpha_{1,i}) - \Delta_i v^{\alpha_i, \alpha_i} (q_{it}, \theta, \alpha_{1,i}) & 0 \\ 0 & u^{\alpha_i, \alpha_i} (q_{it}, \theta, \alpha_{2,i}) - \Delta_i v^{\alpha_i, \alpha_i} (r_{it}, \theta, \alpha_{2,i}) \end{bmatrix} \\ &= \begin{bmatrix} E \left[U^{\alpha\alpha} (z_{it}, \theta, \alpha_i) \right] & 0 \\ 0 & E \left[U^{\alpha\alpha} (z_{it}, \theta, \alpha_i) \right] \end{bmatrix}, \end{aligned}$$

and

$$E \left[\underline{v}_{(S),it} \underline{v}'_{(S),it} \right] = E \left[\begin{bmatrix} v(q_{it}, \alpha_i) \\ v(r_{it}, \alpha_i) \end{bmatrix} \begin{bmatrix} v(q_{it}, \theta, \alpha_{1,i}) & v(r_{it}, \theta, \alpha_{2,i}) \end{bmatrix} \right] = \begin{bmatrix} E \left[v_{it}^2 \right] & 0 \\ 0 & E \left[v_{it}^2 \right] \end{bmatrix}$$

so

$$\begin{aligned} & \frac{1}{2} \text{trace} \left(E \left[\underline{U}_{(S),it}^{\gamma_i} \right] \left(E \left[\underline{v}_{(S),it}^{\gamma_i} \right] \right)^{-1} E \left[\underline{v}_{(S),it} \underline{v}'_{(S),it} \right] \left(\left(E \left[\underline{v}_{(S),it}^{\gamma_i} \right] \right)^{-1} \right)' \right) \\ &= \frac{1}{2} \text{trace} \begin{bmatrix} E \left[U^{\alpha\alpha} \right] \frac{E \left[v_{it}^2 \right]}{\left(E \left[v_{it}^{\alpha_i} \right] \right)^2} & 0 \\ 0 & E \left[U^{\alpha\alpha} \right] \frac{E \left[v_{it}^2 \right]}{\left(E \left[v_{it}^{\alpha_i} \right] \right)^2} \end{bmatrix} = E \left[U^{\alpha\alpha} \right] \frac{E \left[v_{it}^2 \right]}{\left(E \left[v_{it}^{\alpha_i} \right] \right)^2} \end{aligned} \tag{21}$$

Combining (20) and (21), we see that the counterpart of (8) is given by

$$\underline{b}_{(S),i} = 2\Delta_i \frac{E \left[v_{it} v_{it}^{\alpha_i} \right]}{E \left[v_{it}^{\alpha_i} \right]} + E \left[U^{\alpha\alpha} \right] \frac{E \left[v_{it}^2 \right]}{\left(E \left[v_{it}^{\alpha_i} \right] \right)^2}$$

Detailed Derivations for Section 6

In this case, we see that

$$\begin{aligned}
 u(z_{it}, \theta, \alpha_i) &= \frac{\partial \psi(z_{it}, \theta, \alpha_i)}{\partial \theta} = \hat{\alpha}_i(y_{it} - \hat{\alpha}_i \theta) \\
 \frac{\partial u(z_{it}, \theta, \alpha_i)}{\partial \alpha_i} &= \frac{\partial}{\partial \theta} \left(\frac{\partial \psi(z_{it}, \theta, \alpha_i)}{\alpha_i} \right) = y_{it} - 2\hat{\alpha}_i \theta \\
 \frac{\partial^2 u(z_{it}, \theta, \alpha_i)}{\partial \alpha_i^2} &= \frac{\partial}{\partial \theta} \left(\frac{\partial^2 \psi(z_{it}, \theta, \alpha_i)}{\alpha_i^2} \right) = -2\theta
 \end{aligned}$$

so the moment equation (19) is now

$$0 = \sum_i \left(\sum_t \hat{\alpha}_i (y_{it} - \hat{\alpha}_i \hat{\theta}) - \frac{1}{T} \sum_t \left((x_{it} - \hat{\alpha}_i) (y_{it} - 2\hat{\alpha}_i \hat{\theta}) + \hat{\sigma}_\eta^2 (-\hat{\theta}) \right) \right)$$

Because $\hat{\alpha}_i = \bar{x}_i$ in this case, we can now write the moment as

$$0 = \sum_i \left(\sum_t \bar{x}_i (y_{it} - \bar{x}_i \hat{\theta}) - \frac{1}{T} \sum_t \left((x_{it} - \bar{x}_i) (y_{it} - 2\bar{x}_i \hat{\theta}) + \hat{\sigma}_\eta^2 (-\hat{\theta}) \right) \right)$$

Noting that

$$\begin{aligned}
 \sum_t \bar{x}_i (y_{it} - \bar{x}_i \hat{\theta}) &= T\bar{x}_i \bar{y}_i - T\bar{x}_i^2 \hat{\theta}, \\
 \frac{1}{T} \sum_t (x_{it} - \bar{x}_i) (y_{it} - 2\bar{x}_i \hat{\theta}) &= \frac{1}{T} \sum_t (x_{it} - \bar{x}_i) (y_{it} - \bar{y}_i), \\
 \hat{\sigma}_\eta^2 &= \frac{1}{nT} \sum_{i=1}^T \sum_{t=1}^T (x_{it} - \bar{x}_i)^2,
 \end{aligned}$$

we can rewrite the moment as

$$\begin{aligned}
 0 &= \sum_i \left(\left(T\bar{x}_i \bar{y}_i - T\bar{x}_i^2 \hat{\theta} \right) - \frac{1}{T} \sum_t (x_{it} - \bar{x}_i) (y_{it} - \bar{y}_i) + \left(\frac{1}{T} \sum_{i=1}^T \sum_{t=1}^T (x_{it} - \bar{x}_i)^2 \right) \hat{\theta} \right) \\
 &= \sum_i \left(T\bar{x}_i \bar{y}_i - \frac{1}{T} \sum_t (x_{it} - \bar{x}_i) (y_{it} - \bar{y}_i) \right) \\
 &\quad - \left(\sum_i T\bar{x}_i^2 - \frac{1}{T} \sum_{i=1}^T \sum_{t=1}^T (x_{it} - \bar{x}_i)^2 \right) \hat{\theta}
 \end{aligned}$$

or

$$\hat{\theta} = \frac{\sum_i \sum_t \left(\bar{x}_i \bar{y}_i - \frac{1}{T} (x_{it} - \bar{x}_i) (y_{it} - \bar{y}_i) \right)}{\sum_i \sum_t \left(\bar{x}_i^2 - \frac{1}{T} (x_{it} - \bar{x}_i)^2 \right)} = \frac{\sum_i \sum_t \left(\bar{x}_i \bar{y}_i - \frac{1}{T} \tilde{x}_{it} \tilde{y}_{it} \right)}{\sum_i \sum_t \left(\bar{x}_i^2 - \frac{1}{T} \tilde{x}_{it}^2 \right)}$$

We now show that Nagar’s formula $\left(X' \left(P - \frac{k}{N} M \right) X \right)^{-1} X' \left(P - \frac{k}{N} M \right) y$ reduces to the expression above for our particular case with n dummy instruments. Because

$$Z = \begin{bmatrix} \ell & 0 \\ & \ddots \\ 0 & \ell \end{bmatrix}$$

where ℓ is the T -dimensional column vector of ones, we have

$$P = Z(Z'Z)^{-1}Z = \begin{bmatrix} \frac{1}{T} \ell \ell' & 0 \\ & \ddots \\ 0 & \frac{1}{T} \ell \ell' \end{bmatrix},$$

$$M = \begin{bmatrix} I - \frac{1}{T} \ell \ell' & 0 \\ & \ddots \\ 0 & I - \frac{1}{T} \ell \ell' \end{bmatrix}.$$

Noting that $k = n$ and $N = nT$ in this case, we obtain

$$\left(X' \left(P - \frac{k}{N} M \right) X \right)^{-1} X' \left(P - \frac{k}{N} M \right) y = \frac{\sum_i \sum_t \left(\bar{x}_i \bar{y}_i - \frac{1}{T} \tilde{x}_{it} \tilde{y}_{it} \right)}{\sum_i \sum_t \left(\bar{x}_i^2 - \frac{1}{T} \tilde{x}_{it}^2 \right)}.$$

To conclude, Nagar’s estimator can be understood to be the variation/generalization of (19).

We also discuss the higher order bias in non-panel setting. Although the higher order properties of 2SLS are well known, it would be interesting to compare it with the recent analysis of general two-step estimators developed by Cattaneo et al. (2019). For

$$\psi(z_i, \theta, \hat{\mu}_i) = -\frac{1}{2} (y_i - \hat{\mu}_i \theta)^2$$

with

$$m(z_i, \theta, \hat{\mu}_i) = \frac{\partial \psi(z_i, \theta, \hat{\mu}_i)}{\partial \theta} = \hat{\mu}_i (y_i - \hat{\mu}_i \theta)$$

where $\hat{\mu}_i = z_i' \hat{\pi}$ is the fitted value in the regression of the first stage $x_i = w_i' \pi + \eta_i$ is regressed on w_i , we see that $\dot{m}(z_i, \theta, \mu_i) = y_i - 2\mu_i \theta$ and $\ddot{m}(z_i, \theta, \mu_i) = -2\theta$ in Cattaneo et al. (2019) notations, so their bias is equal to

$$\frac{1}{N} \sum_{i=1}^N E[(y_i - 2\mu_i\theta)\eta_i]p_{ii} + (-\theta)\sigma_\eta^2 \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N p_{ij}^2 \tag{22}$$

divided by $E[\mu_i^2]$.

We will simplify their expression a bit, using the fact that $\sum_{j=1}^N p_{ij}^2$ is the (i, i) -element of PP' , and P is symmetric and idempotent, i.e.,

$$\sum_{j=1}^N p_{ij}^2 = p_{ii}, \tag{23}$$

which allows (22) to be rewritten as

$$\frac{1}{N} \sum_{i=1}^N E[(y_i - 2\mu_i\theta)\eta_i]p_{ii} + (-\theta)\sigma_\eta^2 \left(\frac{1}{N} \sum_{i=1}^N p_{ii} \right) = \frac{1}{N} \sum_{i=1}^N \left(E[(y_i - 2\mu_i\theta)\eta_i] + (-\theta)\sigma_\eta^2 \right) p_{ii}.$$

Because

$$y_i - \mu_i\theta = x_i\theta + \varepsilon_i - \mu_i\theta = (\mu_i + \eta_i)\theta + \varepsilon_i - \mu_i\theta = \eta_i\theta + \varepsilon_i,$$

we have

$$\begin{aligned} E[(y_i - 2\mu_i\theta)\eta_i] &= E[(\eta_i\theta + \varepsilon_i - \mu_i\theta)\eta_i] = \theta\sigma_\eta^2 + \sigma_{\varepsilon\eta}, \\ \sum_{i=1}^N p_{ii} &= \text{trace}(P) = k, \\ \sum_{i=1}^N \sum_{j=1}^N p_{ij} &= \text{trace}(PP') = \text{trace}(P) = k, \end{aligned}$$

we obtain further simplification of (22)

$$\frac{1}{N} \sum_{i=1}^N \left(E[(y_i - 2\mu_i\theta)\eta_i] + (-\theta)\sigma_\eta^2 \right) p_{ii} = \frac{k}{N} \sigma_{\varepsilon\eta}$$

Therefore, their bias reduces to

$$\frac{\frac{k}{N}(\theta\sigma_\eta^2 + \sigma_{\varepsilon\eta}) - \frac{k}{N}(\theta\sigma_\eta^2)}{E[\mu_i^2]} = \frac{k}{N} \frac{\sigma_{\varepsilon\eta}}{E[\mu_i^2]} \tag{24}$$

i.e., the usual bias of 2SLS.

The idea behind (19) and their bias (24), if applied to the current framework, suggest that we may consider a modified moment equation

$$\begin{aligned}
0 &= \sum_i \hat{\mu}_i (y_i - \hat{\mu}_i \theta) - \frac{K}{N} \sum_{i=1}^N (y_i - x_i \theta) (x_i - \hat{\mu}_i) \\
&= x' P (y - (Px) \theta) - \frac{K}{N} (y - x \theta)' M x \\
&= x' P (y - x \theta) - \frac{K}{N} x' M (y - x \theta) \\
&= x' \left(P - \frac{K}{N} \right) (y - x \theta),
\end{aligned}$$

which yields Nagar's estimator.

Funding None.

Data Availability Not applicable.

Code Availability (Software Application or Custom Code) Not applicable.

Declarations

Conflicts of interests None.

References

- Arellano, M., and J. Hahn. 2007. Understanding Bias in Nonlinear Panel Models: Some Recent Developments. In *Advances in Economics and Econometrics*, ed. R. Blundell, W.K. Newey, and T. Persson. Cambridge: Cambridge University Press.
- Arellano, M., and J. Hahn. 2016. A Likelihood-based Approximate Solution to the Incidental Parameter Problem in Dynamic Nonlinear Models with Multiple Effects. *Global Economic Review* 45: 251–274.
- Bekker, P.A. 1994. Alternative Approximations to the Distributions of Instrumental Variable Estimators. *Econometrica* 92: 657–681.
- Blundell, R.W., and J.L. Powell. 2004. Endogeneity in Semiparametric Binary Response Models. *Review of Economic Studies* 71: 655–679.
- Cattaneo, M., M. Jansson, and X. Ma. 2019. Two-Step Estimation and Inference with Possibly Many Included Covariates. *Review of Economic Studies* 86: 1095–1122.
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W.K. Newey, and J.M. Robins. 2017. Double/Debiased Machine Learning for Treatment and Structural Parameters. *Econometrics Journal* 21: C1–C68.
- Cox, D.R., and E.J. Snell. 1968. A General Definition of Residuals. *Journal of the Royal Statistical Society Series B (Methodological)* 30: 248–275.
- Firth, D. 1993. Bias Reduction of Maximum Likelihood Estimates. *Biometrika* 80: 27–38.
- Hahn, J., and J. Hausman. 2002. A New Specification Test for the Validity of Instrumental Variables. *Econometrica* 70: 163–189.
- Hahn, J., and J. Hausman. 2003. Weak Instruments: Diagnosis and Cures in Empirical Econometrics. *American Economic Review: Papers and Proceedings* 93: 118–125.
- Hahn, J., and W. Newey. 2004. Jackknife and Analytical Bias Reduction for Nonlinear Panel Models. *Econometrica* 72: 1295–1319.
- Hahn, J., G. Kuersteiner, and W. Newey. 2019. "Bias Correction and Higher Order Efficiency". Unpublished Working Paper.

- Hansen, C., J. Hausman, and W. Newey. 2008. Estimation with Many Instrumental Variables. *Journal of Business & Economic Statistics* 26: 398–422.
- Hausman, J.A. 1978. Specification Tests in Econometrics. *Econometrica* 46: 1251–1271.
- Nagar, A.L. 1959. The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations. *Econometrica* 27: 575–595.
- Rilstone, P., V.K. Srivastava, and A. Ullah. 1996. The Second-Order Bias and Mean Squared Error of Nonlinear Estimators. *Journal of Econometrics* 75: 369–395.
- Rivers, D., and Q.H. Vuong. 1988. Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models. *Journal of Econometrics* 39: 347–366.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.