GENES

ADRES

Correlated Random Effects Models with Endogenous Explanatory Variables and Unbalanced Panels

Author(s): Riju Joshi and Jeffrey M. Wooldridge

Source: *Annals of Economics and Statistics*, June 2019, No. 134 (June 2019), pp. 243-268

Published by: GENES on behalf of ADRES

Stable URL: https://www.jstor.org/stable/10.15609/annaeconstat2009.134.0243

CORRELATED RANDOM EFFECTS MODELS WITH ENDOGENOUS EXPLANATORY VARIABLES AND UNBALANCED PANELS

RIJU JOSHI[a] AND JEFFREY M. WOOLDRIDGE[b]

This paper shows how the correlated random effects approach can be extended to linear panel data models when instrumental variables are needed and the panel is unbalanced. We obtain the algebraic equivalence between the fixed effects two stage least squares (FE2SLS) estimator and a pooled 2SLS (P2SLS) estimator on a transformed equation. This equivalence allows us to obtain fully robust Hausman tests comparing random effects 2SLS (RE2SLS) and FE2SLS. In addition, we obtain an equivalence result for control function estimates and FE2SLS estimates in an unbalanced panel. We use this result to obtain a robust variable addition Hausman test that effectively compares the FE and FE2SLS estimates. We illustrate the tests using an unbalanced panel on student performance and spending at the school level.

JEL Codes: C12, C18, C33, C36.
Keywords: Unbalanced Panel Data, Variable Addition Hausman Test, Fixed Effects, Correlated Random Effects, Control Function, Specification Tests.

1. INTRODUCTION

Panel data sets have become indispensable in contemporary empirical work in the social and behavioral sciences. As discussed in

Hsiao (1985, 2014) and Baltagi (2008), one reason for the popularity of panel data is that it allows one to model dynamic relationships at disaggregated levels. Probably more important is that following the same units over time allows one to account for systematic differences across units – individuals, firms, schools, cities, and so on – that can lead to more convincing estimates of economic parameters and causal effects.

Perhaps the most common estimator applied to panel data, the fixed effects (FE) estimator, allows for an additive, time-constant unobservable – typically called "unobserved heterogeneity" – to be arbitrarily correlated with time-varying explanatory variables. This is one form of endogeneity of explanatory variables, but the usual FE estimator assumes the covariates are strictly exogenous with respect to the time-varying (idiosyncratic) errors. The same is true of differencing methods, the most common being first differencing (FD). Like the FE estimator, the FD approach eliminates endogeneity due to heterogeneity but not endogeneity caused by idiosyncratic errors. The random effects (RE) estimator assumes, for consistency, that the explanatory variables are uncorrelated with the unobserved heterogeneity as well as the idiosyncratic errors. The correlated random effects (CRE) approach can be used to unify the FE and RE approaches. As is well known from Mundlak (1978), adding the time averages of the time-varying covariates reproduces the FE estimates on the time-varying covariates. This means that the CRE approach allows for the heterogeneity to be arbitrarily correlated with the time-varying explanatory variables. However, it rules out the possibility that explanatory variables are correlated with

[a]Department of Economics, Portland State University, 1721 SW Broadway, Suite 241R Portland, Oregon, 97201, United States. riju@pdx.edu

[b]Corresponding author. Department of Economics, Michigan State University, 110 Marshall- Adams Hall, East Lansing, Michigan, 48824-1038, United States. wooldri1@msu.edu

time-varying innovations across any time periods.

There are many situations where one would like to allow for both sources of endogeneity: that due to omitted heterogeneity and that due to time-varying unobservables. If time-varying instrumental variables can be found that are suitably exogenous, combining instrumental variables (IV) approaches with the FE and FD transformations can be quite powerful. For example, in the context of studying the effects of criminal justice variables on crime rates, Levitt (1995, 1996) uses IV approaches after eliminating heterogeneity at either the state or city level via FE or FD. Papke (2005) Papke (2005) and Papke and Wooldridge (2008) use a correlated random effects approach along with instrumental variables, in linear and nonlinear models, to allow per student spending to be correlated with unobserved school district effects and shocks to student performance at the district level. There are also instrumental variable versions of difference-in-differences methods, where

As discussed recently in Wooldridge (2018), the CRE approach has benefits even when applied to unbalanced panels. In the case of linear models, the CRE approach unifies the FE and RE estimators in the sense that each is obtained as a special case. Further, it is straightforward to include time-constant covariates in the CRE estimating equation; unlike with FE, one can estimate coefficients on time-constant variables. Finally, and most importantly, the CRE approach allows for straightforward specification testing that is completely robust to heteroskedasticity and serial correlation. In particular, Wooldridge (2018) shows that a fully robust variable addition Hausman test is available from the Mundlak (1978) device in the general unbalanced case.

In this paper we extend the framework for the linear model in Wooldridge (2018) to the case of endogenous explanatory variables. While it is known that fixed effects IV (FEIV) and random effects IV (REIV) methods can be applied to unbalanced panels, to the best of our knowledge these methods have not been unified in a CRE framework that allows unbalanced panels. There are several reasons this is important for empirical researchers. First, by showing that a certain CRE estimator reduces to the FEIV estimator, we can understand how the Mundlak (1978) CRE assumption turn out to be completely harmless, even in the case of unbalanced panels with instrumental variables. An immediate consequence of this equivalance is that we can obtain a fully robust Hausman test that compares the FEIV and REIV estimators, something that is practically important because obtaining inference robust to heteroskedasticity and serial correlation is important in modern applied work. In some cases, the chosen instrumental variables may be exogenous with respect to both the heterogeneity and idiosyncratic errors, in which case one would typically prefer REIV over FEIV for efficiency reasons.

We also derive robust specification tests that allow us to test whether endogeneity is caused by correlation of the explanatory variables with the idiosyncratic errors, allowing for correlation of all explanatory variables (and outside instruments) with unobserved heterogeneity. We do this by showing that a certain control function estimator in the context of FE is equivalent to the FEIV estimator, and then using an exclusion restriction test on the control functions. In effect, we obtain a fully robust Hausman test for unbalanced panels that allows us to choose between FE and FEIV. Such tests can be difficult to compute using standard software as they often are not robust to heteroskedasticity and serial correlation, and obtaining the necessary generalized inverses to obtain a standard Hausman statistic can be tricky.

We must emphasize that in this paper we study what is known as the *complete cases estimator*, which is the estimator that uses a pair (i, t), where i indexes the cross-sectional unit and t the time series, only when a full set of variables are observed – dependent variable, explanatory variables, and instruments. This is the estimator that is computed by default in all econometrics packages: it is an "all-or-nothing" estimator. [Wooldridge (2018) studies the same estimator but without instrumental variables.] We do not study imputation schemes; nor do we model the selection process. There are many papers that propose selection methods in the case where exogenous variables are available in all time periods, and then variations on Heckman corrections are available. Just a few include Verbeek and Nijman (1992), Wooldridge (1995), Kyriazidou (1997), Rochina-Barrachina (1999), and Semykina and Wooldridge (2010). The last paper allows data to be missing on explanatory variables but then one must have enough instruments, observed in every time period, to treat all such variables as endogenous. None of these papers considers the standard specification tests – effectively comparing FEIV and REIV, comparing FE and FEIV – in the context of unbalanced panels.

As is well known, FE approaches to estimation – unlike RE approaches – allow selection to be correlated with unobserved heterogeneity. The same is true when these methods are used in conjunction with instrumental variables. Because we show that a CRE approach on unbalanced panels leads to FEIV, the CREIV approach also allows arbitrary correlation between selection and heterogeneity. We propose some simple tests that can be used to detect nonrandom selection in the context of FEIV.

The rest of the paper is organized as follows. Section 2 introduces the general model and introduces the convention for allowing unbalanced panels. In Section 3 we review the fixed effects 2SLS (FE2SLS) and RE 2SLS estimators on unbalanced panels, and we provide a set of sufficient conditions for consistency and asymptotic normality. Our perspective here is a microeconometric one, and so we treat T, the number of time periods, fixed, and let the cross-sectional dimension, N, get large.

Section 4 establishes the algebraic equivalence between the FE2SLS estimates and a general class of pooled 2SLS (P2SLS) estimates on a transformed equation. As a special case, CRE versions of RE2SLS and P2SLS are shown to equal the FE2SLS estimator, where all estimators are based on the complete cases. In Section 5, we use the algebraic result in Section 4 to develop a simple, fully-robust variable addition Hausman specification test to compare Random Effects 2SLS and Fixed Effects 2SLS estimators.

In Section 6, we consider the control function approach to detect the endogeneity of explanatory variables with respect to idiosyncratic errors, allowing for unbalanced panels. The basis for the test is showing that using usual FE on the unbalanced panel and including first-stage residuals produces the FEIV estimates. In Section 7, we briefly talk about an empirical strategy that could be followed as protocol for approaching endogeneity issues in a linear model with unbalanced panels. We illustrate this strategy and our theoretical findings with an empirical application in Section 8, studying the effects of spending on student performance in an unbalanced panel of Michigan schools. Section 9 concludes the paper and suggests some future directions.

2. MODEL AND SAMPLING SCHEME

We start with an assumption that serves to introduce a model which includes the special cases considered later.

Assumption 2.1: For a unit i drawn from the population,

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{w}_i\boldsymbol{\delta} + c_i + u_{it},\ t = 1, 2, ..., T. \tag{1}$$

The vector of time-varying covariates is denoted by $\mathbf{x}_{it}$, which has dimension $1 \times K$ and can contain nonlinear functions such as squares, logarithm, interactions, and so on. Typically, one would include time-period dummies, or maybe other functions of time, and they would be treated as exogenous. We are interested in estimating $\boldsymbol{\beta}$ (or at least a subset of $\boldsymbol{\beta}$). Generally, we want to allow elements of $\mathbf{x}_{it}$ to be correlated with unobserved heterogeneity, c_i. We are also interested in cases where some elements of $\mathbf{x}_{it}$ can be correlated with the idiosyncratic errors u_{it}, or even u_{ir} for $r \neq t$. Such variables are said to be endogenous with respect to the idiosyncratic errors, $\{u_{it}\}$. To allow such variables, we assume the existence of a vector of instrument variables (or instruments), $\mathbf{z}_{it}$, which has dimension $1 \times L$. Below we will be clear about the exogeneity assumptions assumed for $\mathbf{z}_{it}$ for the different specification tests. Generally, some elements of $\mathbf{x}_{it}$ (including time-period dummies) will be included in $\mathbf{z}_{it}$. The order condition $L \geq K$ will be necessary for estimating $\boldsymbol{\beta}$ using instrumental variables methods.

The vector $\mathbf{w}_i$ denotes the set of time-invariant variables (including an overall intercept). We view the $\mathbf{w}_i$ as control variables that are included in random effects estimation approaches. When $\mathbf{w}_i$ does appear, it is assumed to be uncorrelated with both c_i and $\{u_{it} : t = 1, 2, ..., T\}$. One can think of the assumption that $\mathbf{w}_i$ is uncorrelated with c_i as holding essentially by construction, by projecting the original heterogeneity onto the time-constant variables $\mathbf{w}_i$. As we will discuss later, it can be important to include such variables when trying to choose between random effects and fixed effects methods for estimating $\boldsymbol{\beta}$. We are not interested in estimating $\boldsymbol{\delta}$ in this paper, and so we will not test assumptions about $\mathbf{w}_i$. When we are comparing usual fixed effects and fixed effects IV methods, $\mathbf{w}_i$ will not appear.

To allow for an unbalanced panel, we introduce a binary selection indicator s_{it}:

$$(2) \qquad s_{it} \equiv \begin{cases} 1 & \text{if and only if } (y_{it}, \mathbf{x}_{it}, \mathbf{w}_i, \mathbf{z}_{it}) \text{ is fully observed} \\ 0 & \text{otherwise} \end{cases}$$

The indicator s_{it} is a *complete cases indicator*, and the time series of selection indicators , $\mathbf{s}_i \equiv \{s_{i1}, s_{i2}, ..., s_{iT}\}$, tells us the pattern of complete observations for each i. We only use (i, t) pair in estimation if $s_{it} = 1$. The number of time periods for which unit i has a complete set of data is denoted by $T_i = \sum_{r=1}^{T} s_{ir}$.

When referring to asymptotic properties of estimators or tests in the remainder of the paper, we assume that a random sample is drawn from an underlying population that consists of a large number of units for whom data on T time periods are potentially observable. This is formally stated as:

Assumption 2.2: The draws $\{(y_{it}, \mathbf{x}_{it}, \mathbf{w}_i, \mathbf{z}_{it}, s_{it}) : t = 1, ..., T\}$ are independent and identically distributed across i.

Random sampling in the cross section means that it is simple to carry out the asymptotics with T fixed and the cross-sectional dimension $N \to \infty$ fixed. We do not formally state moment conditions that imply the weak law of large numbers and central limit theorem hold. By including the s_{it} as part of the random draws, we can be very precise about when the unbalanced nature of the panel does not affect consistency of the different estimators.

In the next section, we briefly review three estimators that play a role in developing the specification tests. There we will discuss the exogeneity assumptions and rank conditions that ensure consistency of the estimators. For now, an informal discussion is useful. Consider (1), which contains two sources of unobservables. If we do not have instrumental variables available, then the main issue in practice is choosing between random effects and fixed effects, where we assume that it is $\boldsymbol{\beta}$ that is of interest. Typically, the FE estimator is less precise – sometimes much less – and so the empirical researcher, for efficiency reasons, may want to use the RE estimator if it is not rejected by the data.

When instrumental variables are available, to allow elements of $\mathbf{x}_{it}$ to be correlated with c_i, u_{it}, or both – the situation is more complicated. For example, we may think we have instruments $\mathbf{z}_{it}$ that are exogenous with respect to both c_i and $\{u_{ir} : r = 1, ..., T\}$, in which case using a random effects version of 2SLS is attractive. However, if the instruments are correlated with c_i but not $\{u_{ir} : r = 1, ..., T\}$, then fixed effects 2SLS is consistent whereas RE2SLS is not. Thus, we are interested in obtaining simple, robust tests that allow us to choose between RE2SLS and FE2SLS.

Alternatively, we may resign ourselves that unobserved heterogeneity is generally correlated with explanatory variables and any instrumental variables we might find. But maybe the explanatory variables $\mathbf{x}_{it}$ are exogenous with respect to $\{u_{ir} : r = 1, ..., T\}$, in which case standard FE is consistent and should be used. (Note that the variables $\mathbf{w}_i$ would not be in the model.) But if we find evidence, using instruments $\mathbf{z}_{it}$, that $\mathbf{x}_{it}$ is correlated with $\{u_{ir} : r = 1, ..., T\}$, then we should use FE2SLS. In other words, we want simple, robust tests that allow us to choose between FE and FE2SLS.

In the next section we provide assumptions under which the standard estimators are consistent and asymptotically normal. We then turn to obtaining specification tests that are fully robust to heteroskedasticity and serial correlation and also apply to unbalanced panels.

3. ESTIMATION METHODS

In this section, we describe popular approaches available to estimate the parameters in (1). We characterize the different estimators and describe their statistical properties under various assumptions.

3.1. *Pooled Two Stage Least Squares (P2SLS) Estimation*

A generic approach to estimation is to apply two stage least squares (2SLS) directly to a panel data equation. In fact, the estimators that underly all specification tests in this paper can be computed as pooled 2SLS applied to a suitably transformed equation.

To make the notation simple, put all explanatory variables – time-varying and time-constant, exogenous or endogenous – into the vector $\mathbf{x}_{it}$. Similarly, let $\mathbf{z}_{it}$ denote the row vector of instrumental variables. Write the equation to be estimated as

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + v_{it}, t = 1, 2, ..., T$$

where we need not be specific about the nature of v_{it}. Typically, we think of it as a composite error, $v_{it} = c_i + u_{it}$, but this is not necessary for describing pooled 2SLS (P2SLS). Because we want to allow for unbalanced panels, we use the selection indicators, s_{it}, to define the complete cases estimator.

Assuming sufficient rank conditions hold for the unbalanced panel in a sample of size

N, we can write the P2SLS as

$$(3)\qquad \hat{\boldsymbol{\beta}}_{P2SLS} = \left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\mathbf{x}_{it}'\mathbf{z}_{it}\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\mathbf{z}_{it}'\mathbf{z}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\mathbf{z}_{it}'\mathbf{x}_{it}\right)\right]^{-1} \cdot \left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\mathbf{x}_{it}'\mathbf{z}_{it}\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\mathbf{z}_{it}'\mathbf{z}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\mathbf{z}_{it}'y_{it}\right)\right]$$

By plugging in for y_{it} and using simple algebra, it is easy to obtain sufficient conditions for the consistency and $\sqrt{N}$-asymptotic normality of $\hat{\beta}_{P2SLS}$ (with T fixed).

Assumption P2SLS.1: $\mathbb{E}(s_{it}\mathbf{z}_{it}'v_{it}) = 0, t = 1, 2, ..., T$.

Because $\mathbf{z}_{it}$ almost always includes a constant, or at least time-period dummies, at a minimum Assumption P2SLS.1 requires v_{it} to be uncorrelated with selection. If $v_{it} = c_i + u_{it}$, then selection should be uncorrelated with heterogeneity and the idiosyncratic error. These assumptions can be applied to any equation where the estimator is obtained as pooled 2SLS, and so it is a very general result. Sufficient for P2SLS is the mean independence assumption $\mathbb{E}(v_{it}|\mathbf{z}_{it}, s_{it}) = 0, t = 1, 2, ..., T$.

Assumption P2SLS.2: (i) rank $\left[\sum_{t=1}^{T}\mathbb{E}\left(s_{it}\mathbf{z}_{it}'\mathbf{x}_{it}\right)\right] = K$. (ii) $\sum_{t=1}^{T}\mathbb{E}\left(s_{it}\mathbf{z}_{it}'\mathbf{z}_{it}\right)$ is nonsingular.

Assumption P2SLS.2 is the standard rank condition on the selected subpopulation, with part (i) being the most important. As usual, it requires a sufficient number of instruments that are partially correlated with the endogenous elements of $\mathbf{x}_{it}$. It could fail in cases where the rank condition holds in the population and the selection mechanism selects too small a subpopulation. Part (ii) requires that, in the selected subpopulation, there is no perfect collinearity among the instruments, a fairly weak requirement unless the instruments include exact linear dependencies in the orginal population.

Under Assumptions P2SLS.1 and P2SLS.2, and standard regularity conditions, $\hat{\boldsymbol{\beta}}_{P2SLS}$ is consistent and $\sqrt{N}\left(\hat{\boldsymbol{\beta}}_{P2SLS} - \boldsymbol{\beta}\right)$ is asymptotically normal, where the asymptotic variance generally has a sandwich form; see, for example, Wooldridge (2010) (Section 11.2) in the balanced case. There are versions of homoskedasticity and no serial correlation assumptions on $\{v_{it} : t = 1, 2, ..., T\}$ that ensure that the usual, nonrobust inference is valid. These assumptions are too restrictive for almost all applications. With large N, one should obtain fully robust variance-covariance matrix that is available from generalized method of moments. See, for example, Wooldridge (2010), Section 11.2.

3.2. *Fixed Effects Two Stage Least Squares (FE2SLS) Estimation*

The fixed effects 2SLS estimator can be obtained as a pooled 2SLS applied to a time-demeaned equation. Now write the equation explicitly with unobserved heterogeneity:

$$(4)\qquad y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}, t = 1, 2, ..., T,$$

where now $\mathbf{x}_{it}$ includes only time-varying variables, and same for the instruments, $\mathbf{z}_{it}$. Define the time averages, but only using the complete cases, as

$$\bar{y}_i = T_i^{-1}\sum_{r=1}^{T} s_{ir}y_{ir},\ \bar{\mathbf{x}}_i = T_i^{-1}\sum_{r=1}^{T} s_{ir}\mathbf{x}_{ir},\ \bar{u}_i = T_i^{-1}\sum_{r=1}^{T} s_{ir}u_{ir},\ \bar{\mathbf{z}}_i = T_i^{-1}\sum_{r=1}^{T} s_{ir}\mathbf{z}_{ir}.$$

Notice that even if we have, say, no missing data on $\mathbf{z}_{it}$, we still compute the time average $\bar{\mathbf{z}}_i$ using only the complete cases. Clearly if $T_i = 0$ then we cannot use observation i, and such units cannot be part of the implied population.

The FE2SLS estimator is pooled 2SLS applied to the within equation (or time-demeaned equation)

$$(5) \qquad (y_{it} - \bar{y}_i) = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\boldsymbol{\beta} + (u_{it} - \bar{u}_i), t = 1, ..., T$$

using instruments $(\mathbf{z}_{it} - \bar{\mathbf{z}}_i)$, where we only use the selected sample. Defining the deviations from means, as $\ddot{y}_{it} = (y_{it} - \bar{y}_i)$, $\ddot{\mathbf{x}}_{it} = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)$, and $\ddot{\mathbf{z}}_{it} \equiv (\mathbf{z}_{it} - \bar{\mathbf{z}}_i)$, the equation can be written as

$$\ddot{y}_{it} = \ddot{\mathbf{x}}_{it}\boldsymbol{\beta} + \ddot{u}_{it}, t = 1, 2, ..., T$$

and then the pooled 2SLS estimator applied to this equation is

$$(6) \qquad \hat{\boldsymbol{\beta}}_{FE2SLS} = \left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{z}}_{it}\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{z}}_{it}'\ddot{\mathbf{z}}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{z}}_{it}'\ddot{\mathbf{x}}_{it}\right)\right]^{-1} \cdot \left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{z}}_{it}\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{z}}_{it}'\ddot{\mathbf{z}}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{z}}_{it}'\ddot{y}_{it}\right)\right]$$

To state sufficient conditions for the consistency and $\sqrt{N}$-asymptotic normality of FE2SLS estimator on the unbalaced panel, define the entire history of the instruments and the selection indicators as

$$\mathbf{z}_i = (\mathbf{z}_{i1}, ..., \mathbf{z}_{iT}), \mathbf{s}_i = (s_{i1}, ..., s_{iT}).$$

Assumption FE2SLS.1: $\mathbb{E}(u_{it}|\mathbf{z}_i, \mathbf{s}_i, c_i) = 0, t = 1, 2..., T.$

Notice that Assumption FE2SLS.1 imposes strict exogeneity on the instruments and the selection indicator with respect to the idiosyncratic error. in other words, $\mathbf{z}_{ir}$ and s_{ir} must be uncorrelated with u_{it} for all r and t. Thus, it rules out situations where the instruments and selection react to changes in u_{ir} – in any time period, r. However, selection may be arbitrarily correlated with $\mathbf{z}_i$ and the unobserved heterogeneity, c_i. One of the benefits of an FE approach is that selection due to unobserved, time constant differences do not cause inconsistency.

Assumption FE2SLS.2: (i) rank $\left[\sum_{t=1}^{T}\mathbb{E}\left(s_{it}\ddot{\mathbf{z}}_{it}'\ddot{\mathbf{x}}_{it}\right)\right] = K$. (ii) $\sum_{t=1}^{T}\mathbb{E}\left(s_{it}\ddot{\mathbf{z}}_{it}'\ddot{\mathbf{z}}_{it}\right)$ is nonsingular.

The rank conditions in FE2SLS.2 mean, at a minimum, that each element of $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$ must have sufficient time variation. Part (i) requires that, after the within i means are removed, we need strong enough instruments for the rank condition to hold. Incidentally, note that any unit i with $T_i = 1$ drops out of the FE estimation. This is as it should be, as such units are uninformative for estimating $\boldsymbol{\beta}$. This does not cause any kind of sample selection problem provided Assumption FE2SLS.1 holds.

There is a set of homoskedasticity and no serial correlation assumptions on $\{u_{it} : t = 1, 2, ..., T\}$ that imply the "usual" asymptotic variance estimator can be used, but these are usually too strong. Wooldridge (2010) (Section 11.2) shows a fully robust asymptotic variance in the balanced cased. The unbalanced case requires no special consideration, and many software packages routinely compute a "sandwich" form that is robust to arbitrary serial correlation and heteroskedasticity.

3.3. *Random Effects Two Stage Least Squares (RE2SLS) Estimation*

The most difficult estimator to define is traditionally called the random effects 2SLS estimator. It is important to understand that, while this estimator is typically obtained under a strong set of assumptions on the composite error term, it is consistent without such assumptions. Specifically, write the model again as in (4), where c_i is the unobserved heterogeneity and $\{u_{it} : t = 1, 2, ..., T\}$ are the idiosyncratic errors. The RE variance-covariance structure is obtained by assuming

$$(7) \qquad \mathbb{C}(c_i, u_{it}) = 0, t = 1, ..., T$$

$$(8) \qquad \mathbb{V}(u_{it}) = \sigma_u^2, t = 1, ..., T$$

$$(9) \qquad \mathbb{C}(u_{ir}, u_{it}) = 0, \text{all } r \neq t$$

In fact, in the traditional analysis, all of these assumptions should hold conditional on the history of the instruments, $\mathbf{z}_i$. Under these assumptions, the composite error $v_{it} = c_i + u_{it}$ has the following properties:

$$(10) \qquad \mathbb{V}(v_{it}) = \sigma_c^2 + \sigma_u^2, t = 1, ..., T$$

$$(11) \qquad \mathbb{C}(v_{ir}, v_{it}) = \sigma_c^2, \text{all } r \neq t$$

This leads to the so-called *random effects variance-covariance structure* for the $T \times T$ matrix $\mathbb{V}(\mathbf{v}_i)$, where $\mathbf{v}_i$ is the $T \times 1$ vector of composite errors. Now suppose that the instruments satisfy the following exogeneity requirements:

Assumption RE2SLS.1: (i) $\mathbb{E}(u_{it}|\mathbf{z}_i, c_i, \mathbf{s}_i) = 0$, $t = 1, 2, ..., T$. (ii) $\mathbb{E}(c_i|\mathbf{z}_i, \mathbf{s}_i) = \mathbb{E}(c_i) = 0$.

Note that Assumption RE2SLS.1(i) is identical to the strict exogeneity requirement in FE2SLS.1. But part (ii) is new, and it requires both the instruments and selection are exogenous with respect to c_i, too. This can be a very strong requirement.

As discussed in Wooldridge (2010) (Section 11.2) in the balanced case, Assumption RE2SLS.1 and the particular variance-covariance structure leads to a particular generalized IV analysis, which is often called the RE2SLS estimator. Fortunately, the assumptions in (7), (8), and (9) turn out to be irrelevant for the consistency of the RE2SLS estimator. What is needed is the RE2SLS.1 and and appropriate rank condition.

To state the rank condition, it is helpful to characterize the RE2SLS estimator as a particular pooled 2SLS estimator, Hausman and Taylor (1981) in the balanced case. To this end, we act as if the RE variance-covariance structure holds. Define a quasi-time-demeaning value, for each i, as

$$(12) \qquad \theta_i = 1 - \left[\frac{\sigma_u^2}{(\sigma_u^2 + T_i\sigma_c^2)}\right]^{\frac{1}{2}},$$

which depends on the number of time periods observed, T_i. If we new σ_c^2 and σ_u^2 – actually, knowing their ratio would suffice – then we know θ_i. Then, the RE2SLS estimator is obtained by applying pooled 2SLS to the equation

$$(13) \qquad (y_{it} - \theta_i \bar{y}_i) = (\mathbf{x}_{it} - \theta_i \bar{\mathbf{x}}_i)\boldsymbol{\beta} + (u_{it} - \theta_i \bar{u}_i)$$

using instruments $\mathbf{z}_{it} - \theta_i \bar{\mathbf{z}}_i$ and only the complete cases. Clearly, if $\theta_i = 0$ for all i, we obtain the usual P2SLS estimator, and if $\theta_i = 1$ we obtain the FE2SLS estimator. When $\theta_i \neq 1$, both $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$ can include time-constant variables, a point we return to in the next section. We can use this characterization to state the rank condition for the RE2SLS estimator. Let

$$\begin{aligned} \check{\mathbf{x}}_{it} &= \mathbf{x}_{it} - \theta_i \bar{\mathbf{x}}_i \\ \check{\mathbf{z}}_{it} &= \mathbf{z}_{it} - \theta_i \bar{\mathbf{z}}_i \\ \check{u}_{it} &= u_{it} - \theta_i \bar{u}_i \end{aligned}$$

Assumption RE2SLS.2: (i) rank $\left[\sum_{t=1}^{T} \mathbb{E}\left(s_{it}\check{\mathbf{z}}_{it}'\check{\mathbf{x}}_{it}\right)\right] = K$. (ii) $\sum_{t=1}^{T} \mathbb{E}\left(s_{it}\check{\mathbf{z}}_{it}'\check{\mathbf{z}}_{it}\right)$ is nonsingular.

It is straightforward to show that applying pooled 2SLS to (13) is consistent and asymptotically normal, as Assumption RE2SLS.1 implies, by iterated expectations, that

$$\mathbb{E}\left(s_{it}\check{\mathbf{z}}_{it}'\check{u}_{it}\right) = \mathbf{0}, t = 1, ..., T$$

and the rank condition is immediate. Importantly, we need not maintain any of the second moment assumptions, unconditionally or conditionally, stated in (7), (8), and (9). Wooldridge (2010) makes the same point in the balanced case. Without such assumptions, inference needs to be fully robust, but that is easily done using sandwich formulations of the asymptotic variances; see Wooldridge (2010) for the balanced case. The RE2SLS will not be asymptotically efficient, but it is still popular to use it, as well as FE2SLS, when the ideal second moment assumptions do not hold.

Any RE estimation inserts estimates for σ_c^2 and σ_u^2, and standard estimators are available in the literature. We refer to reader to Baltagi and Chang (1994), Hsiao (2014), and Wooldridge (2010). However, it is important to remember that these estimators need not be consistently estimating, say, a common σ_u^2 because the actual variances might change over time. Or, if there is serial correlation in $\{u_{it} : t = 1, ..., T\}$, the usual estimators, which exploit the specific serial correlation in $\{v_{it} : t = 1, ..., T\}$ implied by (11), do not consistently estimate σ_c^2. For consistency of the estimator, this does not matter. We only need to assume that the estimators converge to *something*, and this is guaranteed under weak regularity conditions with random sampling across i. In any, let $\hat{\sigma}_c^2$ and $\hat{\sigma}_u^2$ be two such estimators. Then θ_i gets replaced with

$$\hat{\theta}_i = 1 - \left[\frac{\hat{\sigma}_u^2}{(\hat{\sigma}_u^2 + T_i\hat{\sigma}_c^2)}\right]^{\frac{1}{2}},$$

and then pooled 2SLS is applied to (13) using IVs $\mathbf{z}_{it} - \hat{\theta}_i \bar{\mathbf{z}}_i$. The feasible estimator is

$$(14)\qquad \hat{\boldsymbol{\beta}}_{RE2SLS} = \left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\check{\mathbf{x}}_{it}'\check{\mathbf{z}}_{it}\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\check{\mathbf{z}}_{it}'\check{\mathbf{z}}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\check{\mathbf{z}}_{it}'\check{\mathbf{x}}_{it}\right)\right]^{-1} \cdot \left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\check{\mathbf{x}}_{it}'\check{\mathbf{z}}_{it}'\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\check{\mathbf{z}}_{it}'\check{\mathbf{z}}_{it}'\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\check{\mathbf{z}}_{it}'\check{y}_{it}\right)\right],$$

where we reuse the notation $\check{\mathbf{x}}_{it}$, $\check{\mathbf{z}}_{it}$, and $\check{y}_{it}$ when replacing θ_i with $\hat{\theta}_i$.

Wooldridge (2010) discusses how inserting $\hat{\sigma}_c^2$ and $\hat{\sigma}_u^2$ do not change the $\sqrt{N}$-asymptotic distribution of the RE2SLS estimator under the balanced versions of Assumptions RE2SLS.1 and RESLS.2. One must use robust inference, of course, because we are not imposing (7), (8), or (9).

4. A USEFUL ALGEBRAIC EQUIVALENCE RESULT

Wooldridge (2010) obtains a general equivalence result for estimating a transformed equation by Pooled OLS on the unbalanced panel. As a special case, both the RE and POLS estimators on the time-varying variables are shown to equal the FE estimator. We extend the equivalence result for the case of 2SLS estimation. It is important to emphasize that the result in this section is purely algebraic; we only have to assume that certain data matrices have full rank in the selected sample. Of course, we are motivated by the equation that we began with, now written explicitly so that the time-contant variables, $\mathbf{w}_i$, are separated from the time-varying variables, $\mathbf{x}_{it}$:

$$(15)\qquad y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{w}_i\boldsymbol{\delta} + c_i + u_{it}$$

Now, we can apply any of the three estimation methods we have covered to estimate $\boldsymbol{\beta}$, and both pooled 2SLS and RE2SLS can be used to estimate $\boldsymbol{\delta}$. By transforming this equation and augmented it with the time averages of the instruments, we can obtain a unified estimating framework. Specifically, consider the equation

$$(16)\qquad (y_{it} - \theta_i\bar{y}_i) = (\mathbf{x}_{it} - \theta_i\bar{\mathbf{x}}_i)\boldsymbol{\beta} + (1-\theta_i)\mathbf{w}_i\boldsymbol{\delta} + (1-\theta_i)\bar{\mathbf{z}}_i\boldsymbol{\xi} + v_{it} - \theta_i\bar{v}_i,$$

where we follow the notation of Sections 2 and 3 and θ_i can be any unit-specific scalar. Using the selected sample ($s_{it} = 1$), apply pooled 2SLS using instruments

$$(\mathbf{z}_{it} - \theta_i\bar{\mathbf{z}}_i), (1-\theta_i)\mathbf{w}_i, (1-\theta_i)\bar{\mathbf{z}}_i.$$

Note that $(1-\theta_i)\mathbf{w}_i$ and $(1-\theta_i)\bar{\mathbf{z}}_i$ act as their own instruments. Call the estimators from the augmented equation $\hat{\boldsymbol{\beta}}_{A2SLS}$, $\hat{\boldsymbol{\delta}}_{A2SLS}$, and $\hat{\boldsymbol{\xi}}_{A2SLS}$.

Proposition 4.1: Define regressors and instruments as $\check{\mathbf{r}}_{it} = [(\mathbf{x}_{it} - \theta_i\bar{\mathbf{x}}_i), (1-\theta_i)\mathbf{w}_i, (1-\theta_i)\bar{\mathbf{z}}_i]$ and $\check{\mathbf{q}}_{it} = [(\mathbf{z}_{it} - \theta_i\bar{\mathbf{z}}_i), (1-\theta_i)\mathbf{w}_i, (1-\theta_i)\bar{\mathbf{z}}_i]$, respectively. If $\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\check{\mathbf{q}}_{it}'\check{\mathbf{r}}_{it}$ and $\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\check{\mathbf{q}}_{it}'\check{\mathbf{q}}_{it}$ have full rank then

$$\hat{\boldsymbol{\beta}}_{A2SLS} = \hat{\boldsymbol{\beta}}_{FE2SLS}.$$

The proof is given in the appendix, and uses an extension of the Frisch-Waugh-Lovell partialling out theorem to the case of 2SLS.

The proposition, which is purely algebraic in nature, has some useful implications. First, if $\theta_i = 1$ for all i, $(1-\theta_i)\mathbf{w}_i$ and $(1-\theta_i)\bar{\mathbf{z}}_i$ drop out, and then we have simply defined the FE2SLS estimator. The interesting cases are when $\theta_i \neq 1$. One important leading case is $\theta_i = 0$ – so the variables are not transformed but $\bar{\mathbf{z}}_i$, along with $\mathbf{w}_i$, in included in the equation. The other is when θ_i is chosen to deliver the RE2SLS estimator, in which case $\theta_i < 1$ for all i unless $\hat{\sigma}_u^2 = 0$, which very rarely happens. In other words, if we start with the untransformed equation

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{w}_i\boldsymbol{\delta} + \bar{\mathbf{z}}_i\boldsymbol{\xi} + v_{it},$$

and estimate it by pooled 2SLS using IVs $(\mathbf{z}_{it}, \mathbf{w}_i, \bar{\mathbf{z}}_i)$ or by RE2SLS using the same instruments, we obtain the FE2SLS estimates for $\boldsymbol{\beta}$. Again, this is an exact result, not an approximate or asymptotic results.

We will use this result in the next section to test for correlation between the unobserved heterogeneity, c_i, in the original model, and $\bar{\mathbf{z}}_i$. Before doing so, it is important to highlight a couple of points. First, the equivalence does not hold unless only the complete cases are used to obtain the time averages. Second, if $\mathbf{z}_{it}$ includes aggregate time effects – typically a set of dummy variables – their time averages must also be included. That is because with an unbalanced panel the time average of a dummy variable for the second time period, say $d2_t$, has a time average that generally varies across i: $\overline{d2}_i = T_i^{-1}\sum_{r=1}^{T} s_{it} d2_t$. In the balanced case, $\overline{d2}_i = 1/T$ for all i, and the same with all other time dummies.

5. ROBUST HAUSMAN VARIABLE ADDITION TEST COMPARING RE2SLS AND FE2SLS

In this section we will obtain a variable addition version of the Hausman test using the algebraic equivalance from the previous section. Consider again the equation for a random draw i:

$$(17) \qquad y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{w}_i\boldsymbol{\delta} + c_i + u_{it}, t = 1, ..., T,$$

where we have $\mathbf{z}_{it}$ as instruments for $\mathbf{x}_{it}$. If we apply pooled 2SLS or, more likely, RE2SLS directly to this equation, consistency requires that $\mathbf{z}_{it}$ is uncorrelated with c_i. FE2SLS relaxes this assumption, but then the FE2SLS estimator of β may be imprecise. The Hausman (1978) can be used to construct a test that compares $\hat{\beta}_{RE2SLS}$ and $\hat{\beta}_{FE2SLS}$, but the traditional form of the test is not robust to serial correlation in $\{u_{it} : t = 1, ..., T\}$ or heteroskedasticity in $\{u_{it} : t = 1, ..., T\}$ or c_i [conditional on $(\mathbf{w}_i, \mathbf{z}_i)$]. Here we derive a fully robust variable addition test (VAT) instead.

Wooldridge (2010) motivates the test in the balanced case as follows. In addition to the strict exogeneity assumption

$$(18) \qquad \mathbb{E}\left(u_{it}|\mathbf{w}_i, \mathbf{z}_i, c_i\right) = 0, t = 1, 2, ..., T,$$

which is maintained by FE2SLS as well as RE2SLS, we would like to detect violations of the extra RE assumption:

$$(19) \qquad \mathrm{H}_0 : \mathbb{E}\left(c_i|\mathbf{w}_i, \mathbf{z}_i\right) = \mathbb{E}\left(c_i\right) = 0,$$

where the assumption of zero mean is simply a normalization. In order to obtain a variable addition test, we need an alternative. FollowingArellano (1993), we define the *Hausman alternative hypothesis* as

$$\text{(20)} \qquad \mathrm{H}_1 : \mathbb{E}\left(c_i|\mathbf{w}_i, \mathbf{z}_i\right) = \mathbb{E}\left(c_i|\mathbf{z}_i\right) = \xi_0 + \bar{\mathbf{z}}_i\boldsymbol{\xi}$$

We can then write

$$\begin{aligned} c_i &= \xi_0 + \bar{\mathbf{z}}_i\boldsymbol{\xi} + a_i \\ \mathbb{E}\left(a_i|\mathbf{w}_i, \mathbf{z}_i\right) &= 0 \end{aligned}$$

and write an expanded equation as

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{w}_i\boldsymbol{\delta} + \xi_0 + \bar{\mathbf{z}}_i\boldsymbol{\xi} + a_i + u_{it}, t = 1, ..., T.$$

The null hypothesis is

$$\mathrm{H}_0 : \boldsymbol{\xi} = \mathbf{0}.$$

Given (18), the exogeneity assumptions for consistency of RE2SLS (and P2SLS) hold with respect to the error components a_i and u_{it}. We can drop ξ_0 becase $\mathbf{w}_i$ should include a consant. In any case, the augmented equation should have a constant.

One could ask why $\mathbf{w}_i$ is not included in (20). The reason is that we are not trying to identify the effects of time-constant variables because we treat the $\mathbf{w}_i$ as control variables. If instead the alternative is

$$\mathbb{E}\left(c_i|\mathbf{w}_i, \mathbf{z}_i\right) = \xi_0 + \mathbf{w}_i\boldsymbol{\lambda} + \bar{\mathbf{z}}_i\boldsymbol{\xi}$$

then, under the null, we are allowing $\mathbf{w}_i$ to be correlated with c_i. But this would simply change the coefficient on $\mathbf{w}_i$ to $\boldsymbol{\delta} + \boldsymbol{\lambda}$, and nothing of substance would change because we are using $\mathbf{w}_i$ to proxy for unobserved heterogeneity, anyway. In fact, we could have started the model with $\mathbf{w}_i$ and then projected the heterogeneity onto $\mathbf{w}_i$ to arrive at (17). The point is that the only null we can test is that the time-varying instruments in $\mathbf{z}_{it}$ are uncorrelated with c_i, as that determines whether we will use RE2SLS or FE2SLS.

When the panel is unbalanced, we use the conclusion of Proposition 4.1. Namely, even in the unbalanced case, estimation of the equation

$$\text{(21)} \qquad y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{w}_i\boldsymbol{\delta} + \bar{\mathbf{z}}_i\boldsymbol{\xi} + a_i + u_{it}, t = 1, ..., T$$

by RE2SLS, using the complete cases, delivers $\hat{\boldsymbol{\beta}}_{FE2SLS}$. In other words, the difference between FE2SLS and RE2SLS is still the inclusion of $\bar{\mathbf{z}}_i$: if $\bar{\mathbf{z}}_i$ is omitted, we obtain the RE2SLS estimator of $\boldsymbol{\beta}$, and if it is included, we obtain the FE2SLS estimator. It does not matter what else we include in $\mathbf{w}_i$, including time constant functions of the selection indicators. For example, one might think of including dummy variables for the different values of T_i. But including these in $\mathbf{w}_i$ does not change that we obtain $\hat{\boldsymbol{\beta}}_{FE2SLS}$ when applying RE to (21). Therefore, the only time-constant variables that are relevant are in $\bar{\mathbf{z}}_i$, and that is why we restrict out test to these elements.

The test statistic is the familiar Wald test statistic:

$$\text{(22)} \qquad W = \hat{\boldsymbol{\xi}}'\hat{\mathbf{V}}_{\boldsymbol{\xi}}^{-1}\hat{\boldsymbol{\xi}}$$

where $\hat{\boldsymbol{\xi}}$ is the RE estimator from (21) and $\hat{\mathbf{V}}_{\boldsymbol{\xi}}$ is a suitable estimator of the asymptotic variance of $\hat{\boldsymbol{\xi}}$. We prefer that a fully robust variance estimator, that allows for any pattern of serial correlation and heteroskedasticity in $\{a_i + u_{it} : t = 1, 2, ..., T\}$ be used. Under the null hypothesis

$$\begin{aligned} \mathbb{E}\left(u_{it}|\mathbf{w}_i, \mathbf{z}_i, c_i, \mathbf{s}_i\right) &= 0,\ t = 1, 2, ..., T. \\ \mathbb{E}\left(c_i|\mathbf{w}_i, \mathbf{z}_i, \mathbf{s}_i\right) &= \mathbb{E}\left(c_i\right) \end{aligned}$$

and the RE2SLS rank condition, $W \overset{d}{\rightarrow} \chi^2_L$, where L is the dimensionof $\bar{\mathbf{z}}_i$. Of course, one could impose the RE second moment assumptions and then use a nonrobust test, but that would be against the spirit of current empirical work with panel data where one should usual use cluster-robust standard errors and test statistics. Fortunately, popular econometrics packages have built-in routines for obtaining the the RE2SLS estimators along with cluster-robust inference, and so implementing our test is rather easy – much easier than a robust form of the test based directly on the difference $\hat{\boldsymbol{\beta}}_{FE2SLS} - \hat{\boldsymbol{\beta}}_{RE2SLS}$. Of course, one should determine whether the differences in the FE and RE estimates are practically different in addition to just looking at the outcome of a statistical test.

6. ROBUST HAUSMAN VARIABLE ADDITION TEST COMPARING FE AND FE2SLS

The test in the previous section takes as given that we need instruments for at least some elements of $\mathbf{x}_{it}$, and then the goal is to determine whether to use the instruments in an RE or FE approach. In other words, we are testing whether the instruments are correlated with the unobserved heterogeneity, essentially conceding that some elements of $\mathbf{x}_{it}$ are endogenous. If we take as a starting point the perspective that instrumental variables (and explanatory variables) are likely correlated with heterogeneity, we might instead want to choose between using the usual FE estimator – that is, without instruments – and the FE2SLS estimator. The question becomes whether it is enough to remove the heterogeneity in order to render $\mathbf{x}_{it}$ endogenous. Or, do we need to still allow instruments because some elements of $\mathbf{x}_{it}$ are correlated with u_{it} (and therefore violate strict exogeneity).

In this section we show how to test the null hypothesis that $\mathbf{x}_{it}$ is exogenous with respect to $\{u_{ir} : r = 1, 2, ..., T\}$. We again require time-varying instrumental variables, but now we assume only that the are strictly exogenous with respect to $\{u_{it} : t = 1, ..., T\}$. They can be arbitrarily correlated with the heterogeneity.

One can, of course, use a traditional Hausman test by comparing the FE and FE2SLS coefficients, but it has some shortcomings as discussed in Wooldridge (2010). For one, it almost always involves calculating a generalized inverse, and software may not be particularly good about computing the appropriate degrees of freedom. More importantly, the traditional Hausman test assumes that the idiosyncratic errors are serially uncorrelated and homoskedastic.

In this section, we use control function approach in the context of fixed effects estimation to obtain a test of the null hypothesis that the variables we suspect are endogenous are actually exogenous (with respect to the idiosyncratic errors). It is well known that the usual 2SLS estimator can be obtained using a control function approach, and this leads to a simple test of the null of exogeneity. See, for example, Chapter 6 of Wooldridge (2010). We extend that result to FE estimation on an unbalanced panel data set, and this leads immediately to a simple, fully robust test.

It is useful now to express the model, which includes only time-varying variables, by separating the exogenous and endogenous variables. To this end, write

(23) $$y_{it1} = \mathbf{z}_{it1}\boldsymbol{\gamma}_1 + \mathbf{y}_{it2}\boldsymbol{\alpha}_1 + c_{i1} + u_{it1}, t = 1, ..., T,$$

where y_{it1} denotes the response variable. The explanatory variables $\mathbf{z}_{it1}$ are assumed to be strictly exogenous with respect to $\{u_{it1} : t = 1, ..., T\}$, whereas we suspect the elements of and $\mathbf{y}_{it2}$, a $1 \times G_1$ vector, are correlated with $\{u_{it1} : t = 1, ..., T\}$. Both variables are allowed to be correlated with c_{i1}. Provided we have suitable time-varying instruments, instrumental variables we can consistently estimate $\boldsymbol{\gamma}_1$ and $\boldsymbol{\alpha}_1$ by FE2SLS. Let $\mathbf{z}_{it} = (\mathbf{z}_{it1}, \mathbf{z}_{it2})$ where the $1 \times L_2$vector $\mathbf{z}_{it2}$ serves as the external instruments for $\mathbf{y}_{it2}$, and so we require, at a minimum, $L_2 \geq G_1$.

Underlying FE2SLS is a first stage equation that is estimated by fixed effects. In what follows, we do not put any restrictions on the reduced form for $\mathbf{y}_{it2}$, but it is helpful to think of it as

(24) $$\mathbf{y}_{it2} = \mathbf{z}_{it}\boldsymbol{\Gamma}_2 + \mathbf{c}_{i2} + \mathbf{v}_{it2},$$

where $\boldsymbol{\Gamma}_2$ is $L \times G_1$. The idea is to think of endogeneity of $\mathbf{y}_{it2}$ with respect to u_{it1} as correlation between u_{it1} and $\mathbf{v}_{it2}$. To this end, write a linear equation

$$u_{it1} = \mathbf{v}_{it2}\boldsymbol{\rho}_1 + e_{it1}, t = 1, ..., T$$

where

(25) $$\mathbb{E}\left(e_{it1}|\mathbf{z}_i, \mathbf{y}_{i2}, c_{i1}, \mathbf{s}_i\right) = 0, t = 1, 2, ..., T.$$

Now plug in for u_{it1} and rearrange:

(26) $$y_{it1} = \mathbf{z}_{it1}\boldsymbol{\gamma}_1 + \mathbf{y}_{it2}\boldsymbol{\alpha}_1 + \mathbf{v}_{it2}\boldsymbol{\rho}_1 + c_{i1} + e_{it1}, t = 1, ..., T$$

Condition (25) is sufficient for FE estimation on (26), using the selected sample, to be consistent, provided the rank condition holds. This will be the case provided we have sufficient time-varying instruments $\mathbf{z}_{it2}$ for $\mathbf{y}_{it2}$. To this end, write the equation in deviations from the complete cases time averages as

(27) $$\ddot{y}_{it1} = \ddot{\mathbf{z}}_{it1}\boldsymbol{\gamma}_1 + \ddot{\mathbf{y}}_{it2}\boldsymbol{\alpha}_1 + \ddot{\mathbf{v}}_{it2}\boldsymbol{\rho}_1 + \ddot{e}_{it1}, t = 1, ..., T,$$

where the heterogeneity, c_{i1}, gets swept away by the within transformation, as always. If we knew $\ddot{\mathbf{v}}_{it2}$ we could apply pooled OLS, which would give the FE estimator. Instead, we replace $\ddot{\mathbf{v}}_{it2}$ with the FE residuals from the first stage in (24). To be precise, the steps are summarized:

Procedure 6.1:

1. Estimate (24) using the complete cases ($s_{it} = 1$) and obtain the fixed effects residuals,

$$\widehat{\ddot{\mathbf{v}}}_{it2} = \ddot{\mathbf{y}}_{it2} - \ddot{\mathbf{z}}_{it}\hat{\boldsymbol{\Gamma}}_2$$

If $\mathbf{y}_{it2}$ has more than one element, the FE estimations can be done on each separately.

2. Estimate the equation

(28) $$\ddot{y}_{it1} = \ddot{\mathbf{z}}_{it1}\boldsymbol{\gamma}_1 + \ddot{\mathbf{y}}_{it2}\boldsymbol{\alpha}_1 + \widehat{\ddot{\mathbf{v}}}_{it2}\boldsymbol{\rho}_1 + error_{it_1}$$

by pooled OLS.

3. Compute a (robust) Wald statistic for the null hypothesis

$$\mathrm{H}_0 : \boldsymbol{\rho}_1 = \mathbf{0}$$

This is the same as estimating the equation

$$y_{it1} = \mathbf{z}_{it1}\boldsymbol{\gamma}_1 + \mathbf{y}_{it2}\boldsymbol{\alpha}_1 + \widehat{\ddot{\mathbf{v}}}_{it2}\boldsymbol{\rho}_1 + error_{it1} \tag{29}$$

by fixed effects and testing the same null. Under H_0, the statistic is asymptotically distributed as $\chi^2_{G_1}$, where G_1 is the dimension of $\mathbf{y}_{it2}$.

Proposition 6.1: Consider estimating (29) using the usual fixed effects estimator on the complete cases. Call these $\hat{\gamma}_{CF,1}$, $\hat{\boldsymbol{\alpha}}_{CF,1}$, and $\hat{\boldsymbol{\rho}}_{CF,1}$, respectively. Then

$$\hat{\gamma}_{CF,1} = \hat{\gamma}_{FE2SLS,1} \text{ and } \hat{\boldsymbol{\alpha}}_{CF,1} = \hat{\boldsymbol{\alpha}}_{FE2SLS,1}$$

where the FE2SLS estimators are obtained from (23) using instruments $\mathbf{z}_{it} = (\mathbf{z}_{it1}, \mathbf{z}_{it2})$ on the selected subsample.

The proof is shown in the appendix.

The point of this result is that whether we include $\widehat{\ddot{\mathbf{v}}}_{it2}$ or not in (29) determines whether we are using the usual FE or the FE2SLS estimator. In this way, the motivation for the test is very similar to that in Section 5. Note that the proper degrees of freedom are obtained by just testing exclusion restrictions on the variables $\widehat{\ddot{\mathbf{v}}}_{it2}$, are there is not need for generalized inverses. Most important, it is essentially trivial these days to obtain a fully robust test statistic.

Section 11.2 of Wooldridge (2010) discusses why, under H_0, there is no need to adjust the Wald statistic for the estimation of $\widehat{\ddot{\mathbf{v}}}_{it2}$ – that is, for the fact that $\hat{\boldsymbol{\Gamma}}_2$ replaces $\boldsymbol{\Gamma}_2$. But this is true only under the null. If one rejects the null and opts for the FE2SLS estimator, one should apply FE2SLS directly to (23) and obtain (robust) standard errors and confidence intervals, as well as joint test statistics, from there.

7. A STRATEGY FOR THE APPLIED ECONOMETRICIAN

We have so far suggested two kinds of specification tests. One is to detect endogeneity of the explanatory variables with respect to the time-varying idiosyncratic errors (Section 6) and the other is to detect endogeneity of the instruments with the time-constant unobserved effect (Section 5). How should one use these tests? While there are always issues about pretesting, and which tests should be done first, we think a sensible strategy is to start with the test in Section 6 to determine whether instrumental variables estimation is needed in a fixed effects setting. A failure to reject the null means that we might conclude the explanatory variables are strictly exogenous with respect to the time-varying idiosyncratic errors, and we can opt for the usual FE estimator. We might further test the usual RE estimator against the FE estimator using the robust, regression-based tests discussed in Wooldridge (2018). If we reject the null, then we are concluding that we need to use the instrumental variables. It could be that the instruments are exogenous with respect to the heterogeneity (as well as the idioysyncratic errors). We can test this using the VAT in Section 5. If we fail to reject, we can opt for the RE2SLS estimator.

Pretesting Issues

Guggenberger (2010) highlights the pretesting issue in using the traditional (nonrobust) Hausman test to choose between random effects and fixed effects in the case without instrumental variables. The key issue is that when the Hausman test incorrectly fails to rejects the null – that is, we commit a Type II error – we will use the incorrect estimation method. In addition, pretesting generally affects the size of any subsequent tests and the coverage rates of confidence intervals. We have nothing much to add to the issue of pretesting except to note that, because of the regression-based nature of our tests, the pretesting issue for Hausman-type tests are essentially similar to the issue of deciding whether to include a set of explanatory variables (say, $\mathbf{x}_2$) when estimating the effects of another set of variables ($\mathbf{x}_1$). In Section 5, our pretest is on the time averages, $\bar{\mathbf{z}}_i$. In Section 6, the pretest is on the fixed effects residuals, $\widehat{\ddot{\mathbf{v}}}_{it2}$. [The fact that these are estimates likely makes the pretesting problem even worse.] Unlike Guggenberger (2010), our tests can be made completely robust, but that does not change the basic nature of the pre-testing problem.

If one does not want the analysis to be subjected to pretesting biases, one can always use the FE2SLS estimator – assuming there is no pretesting bias in, say, choosing instruments – and accept the consequences of perhaps getting noisy estimates.

Testing Strict Exogeneity of Selection and the Instruments

The estimator that imposes the fewest assumptions on the instruments and explanatory variables is the FE2SLS estimator. Even if using FE2SLS is what emerges from applying the specification tests, there are some threats to validity of FE2SLS. First, the instruments are assumed to be strictly exogenous with respect to the idiosyncratic errors (but allowed to be arbitrarily correlated with the heterogeneity). Second, because of the unbalanced panel, we might be concerned that selection, s_{it}, is not strictly exogenous with respect to $\{u_{it} : t = 1, ..., T\}$. In the context of FE estimator, Wooldridge (2010), borrowing an idea fromVerbeek and Nijman (1992) in the context of random effects, suggests adding lags or leads of the selection indicator to the equation estimated by FE. This idea is easily extended to the current context. Estimate the augmented equation

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \psi s_{i,t+1} + c_i + u_{it}, t = 1, ..., T$$

be FE2SLS using the complete cases (and losing period T). The null hypothesis is

$$H_0 : \psi = 0,$$

which means that, after controlling for endogeneity of elements of $\mathbf{x}_{it}$ using instruments $\mathbf{z}_{it}$, and allowing for heterogeneity to be correlated with the explanatory variables, instruments, and selection, test whether having a complete case in $t + 1$ helps predict the outcome in period t. If $\psi \neq 0$ then the FE2SLS estimator on the selected sample is generally inconsistent. Note that one needs at least $T = 3$ time periods to carry out this test. Alternative, replace $s_{i,t+1}$ with $s_{i,t-1}$ and use periods $t = 2, ..., T$. Other functions of the selection indicators can be included instead, such as, at time t, how many future time periods are missing. We cannot include s_{it} at time t because, by definition, we use period t in the estimation if and only if $s_{it} = 1$. Thus, putting in leads and lags of the selection indicator might not have good power if s_{it} and u_{it} are uncorrelated. If complete cases are available for the instruments, Semykina and Wooldridge (2010) show how to obtain Heckman-type tests to determine if s_{it} and u_{it} are uncorrelated.

We can also test, at the same time, strict exogeneity of the proposed instruments by including $s_{i,t+1}\mathbf{z}_{i,t+1}$, or a subset, along with $s_{i,t+1}$ and using a joint Wald test. Again, this does not directly test whether $\mathbf{z}_{it}$ is correlated with u_{it}, but it does test an implication of strict exogeneity of the instruments.

If we can rule out selection bias then we can rely on the FE2SLS estimator.

8. EMPIRICAL ILLUSTRATION

To illustrate the methods above, we consider the problem of estimating the effects of spending on student performance. We use the data on standardized test scores from 1992 to 1998 at Michigan schools to determine the effects of spending on math test outcomes for fourth graders.Papke (2005) studies this using school-level data and a linear functional form. Papke and Wooldridge (2008) extend the analysis by recognizing the fractional nature of the pass rates. Specifically, they use fractional response models for the district level panel data. Both find non-trivial effects of spending on test pass rates. Since we deal with linear models in our paper, our analysis is closer to Papke and Wooldridge (2008).

8.1. *Background*

Funding for K-12 schools in Michigan dramatically changed in 1994 from local, property-tax based system to a statewide system supported primarily through a higher sales tax. The primary goal of this policy change was to equalize spending and this was reflected in the rise of per-pupil spending. Papke (2005) studies the effect of this policy change on student performance. The data used comes from annual Michigan School Reports (MSRs). The outcome variable of study is the percentage of students passing the Michigan Educational Assessment Program (MEAP) math test for 4th graders: *math4*. The key explanatory variable is log of average per pupil expenditure: $\log(avgrexpp)$ which serves as a measure of per-pupil spending. The data used in Papke (2005) is an unbalanced panel data. The second column of Table 3 gives a snapshot of the missing nature of the data: about 66 percent of the total of 1,643 schools have data on all the four years (1995-1998). A simple Verbeek and Nijman (1992) test verifies that the selection is not correlated with the idiosyncratic shocks. Specifically, we add the lagged value of the selection indicator to our model and found it insignificant. This allows us to apply our tests to this empirical problem. In this empirical illustration, we revisit this problem taking the note of the incomplete nature of the panel. In addition, we use Stata 15 and have fully robust standard errors as "clustering"'is available for all estimation methods.

8.2. *Results*

Since our purposes are primarily illustrative, we focus on the simple specification:

$$math4_{it} = \eta_t + \beta_1 \log(avgrexpp_{it}) + \beta_2 lunch_{it} + \beta_3 \log(enroll_{it}) + c_{i1} + u_{it1} \tag{30}$$

where i indexes school and t indexes year. The η_t are separate time period intercepts, captured by adding time dummies. The covariate vector is $\mathbf{x}_{it} = [\log(avgrexpp_{it}), lunch_{it}, \log(enroll_{it})]$. Papke (2005) argues that $\log(avgrexpp_{it})$ could be endogenous as spending could be correlated with the idiosyncratic shocks u_{it}. She uses the log of the district foundation grant,

$\log(found_{it})$, as an instrument. The variables $lunch_{it}$ and $\log(enroll_{it})$ act as their own instruments, and so $\mathbf{z}_{it} = [\log(found_{it}), lunch_{it}, \log(enroll_{it})]$.

We begin by conducting a test to check the endogeneity of the explanatory variables. As mentioned before, Papke (2005) argues that the primary variable of interest $\log(avgrexpp_{it})$, may be endogenous in the sense that it is correlated with the time varying idiosyncratic errors. She verifies this claim using a fully robust Hausman test that compares the Pooled OLS and Pooled 2SLS estimators. In this paper, we further verify the endogeneity of $\log(avgrexpp_{it})$ using the control function approach as described in Section 6. More specifically, we begin by estimating the reduced form equation

$$\log(avgrexpp_{it}) = \phi_t + \pi_1 lunch_{it} + \pi_2 \log(enroll_{it}) + \pi_3 \log(found_{it}) + c_{i2} + \nu_{it2} \tag{31}$$

by fixed effects. Denote the FE residuals by $\widehat{v}_{it2}$. The control function equation is

$$math4_{it} = \psi_t + \beta_1 \log(avgrexpp_{it}) + \beta_2 lunch_{it} + \beta_3 \log(enroll_{it}) + \rho\widehat{\widehat{v}}_{it2} + error_{it} \tag{32}$$

We estimate this equation using fixed effects. The results are given in Column (1) of Table 1. We see verification of the general equivalence: the CF estimates are identical to the FE2SLS estimates given in Column (4). To check for the endogeneity of $\log(avrgexpp)$, we check the significance of the estimate of the coefficient on $\hat{\nu}_{it2}$ using a cluster-robust standard error. We emphasize that the standard errors are valid under the null of no endogeneity of the explanatory variables. The conclusion is marginal: we would reject at the at 10 percent level but not 5 percent. The magnitudes of FE and FE2SLS estimates differ considerably, but, as in much empirical work, it is difficult to settle on one over the other.

Next, we compare RE2SLS and FE2SLS, and verify that the CRE approach produces the FE2SLS estimates. The results are given in Columns (3), (4), and (5) of Table 1. Both RE2SLS and FE2SLS estimation methods give a statistically significant estimate of the coefficient on $\log(avgrexpp)$. The results verify that the effects of spending on student performance are non-trivial. This is consistent with the results obtained in Papke (2005) and Papke and Wooldridge (2008).

We find that the RE2SLS estimates are quite different from the FE2SLS estimates and this motivates testing for correlation between the instrument with the school level heterogeneity. We use the robust VAT version of the Hausman test developed in Section 5. This also allows us to verify our equivalence results. Recall that we model the individual heterogeneity as $c_{i1} = \xi_0 + \bar{\mathbf{z}}_i\boldsymbol{\xi} + a_{i1}$ and substitute it in. The estimating equation becomes

(33)
$$\begin{aligned} math4_{it} = {} & \eta_t + \beta_1 \log(avgrexpp)_{it} + \beta_2 lunch_{it} + \beta_3 \log(enroll)_{it} \\ & + \xi_1 \overline{lunch_i} + \xi_2 \overline{\log(enroll)_i} + \xi_3 \overline{\log(found)_i} + \xi_4 \overline{y96}_i + \xi_4 \overline{y97}_i + \xi_4 \overline{y98}_i + a_{i1} + e_{it1} \end{aligned}$$

Note that the averages of the time dummies are included in the equation in order to obain the equivalance between REIV applied to this equation and FEIV applied to the original equation. In other words, $\bar{\mathbf{z}}_i$ include the averages of the year dummies. This is an important aspect in which our analysis differs due to the unbalanced nature of the panel. Since different individuals have different T_i, we are averaging over different time periods for different i. Thus the time averages of the aggregate time variables changes across i. Also, only the complete cases are used to compute the averages of all variables.

We estimate the (33) by RE2SLS and the results are given in Column (5) of Table 1. The estimates verify our equivalence result for the unbalanced panels with endogeneity. To check for the correlation of our instrument $\log(found_i)$ with the unobserved heterogeneity c_i, we check the joint significance of the coefficients on $\bar{\boldsymbol{z}}_i$. This translate into checking the joint significance of the coefficients on $\overline{lunch_i}$, $\overline{\log(enroll)_i}$, $\overline{\log(found)_i}$, $\overline{y96_i}$, $\overline{y97_i}$, and $\overline{y98_i}$. The $\boldsymbol{\chi}^2$ statistic and the p-value of the test, which is essentially zero, are given in Table 2. We find that the variables are jointly significant, illustrating a non-zero correlation between the instruments and the time-invariant unobserved individual-specific heterogeneity. The coefficient on $\overline{lunch_i}$ is especially strong, and perhaps including that only would be sufficient to account for correlation between the heterogeneity and the instruments. In any case, the RE2SLS approach is rejected.

9. CONCLUDING REMARKS

In this paper, we provide a unifying framework for FEIV and REIV estimation methods applied to linear unbalanced panel data models. We use the CRE approach for unbalanced panels when explanatory variables are allowed to be correlated with time-varying idiosyncratic errors. We derive an important algebraic result showing the equivalence of a CRE estimator on a transformed equation and FE2SLS estimator. A byproduct of this result is fully robust Hausman tests for unbalanced panels that facilitate comparison of the FE2SLS and RE2SLS estimators. In addition, we also derive a specification test to detect the endogeneity of the explanatory variables with respect to the idiosyncratic errors in a framework that allows correlation between explanatory variables and time-constant unobserved heterogeneity. Specifically, we obtain an algebraic equivalence result between FE estimator of an equation augmented with a control function and the usual FE2SLS estimator. This result then gives us a fully robust Hausman test for unbalanced panels that allows us to choose between FE and FE2SLS estimator.

An extension of the CRE approach developed in this paper is to models with unit-specific slopes, where we model the slopes as functions of the time averages of the exogenous variables, as well as functions of T_i. Wooldridge (2018) shows some ways of doing this when instrumental variables are not needed. Among other things, we would have a simple way to test the null hypothesis that the slopes are not heterogeneous.

A critical assumption throughout this paper is that the selection mechanism is exogenous with respect to the idiosyncratic shocks. This assumption seldom holds in most unbalanced panels. Thus analysis for specifications for unbalanced panels where the exogenous sampling does not hold is another area in which we would like to further extend our analysis.

TABLE I
RESULTS

	(1)	(2)	(3)	(4)	(5)
math4	CF	FE	RE2SLS	FE2SLS	CRE2SLS
$\log(avgrexpp)$	47.00**	5.084	19.47***	47.00*	47.00*
	(23.32)	(3.457)	(2.687)	(25.03)	(25.04)
$lunch$	-0.00466	0.0202	-0.378***	-0.00466	-0.00466
	(0.0458)	(0.0433)	(0.0143)	(0.0467)	(0.0467)
$\log(enrol)$	6.483	-3.174	-0.523	6.483	6.483
	(5.629)	(2.217)	(0.891)	(6.161)	(6.164)
$y96$	-2.558	1.598***	0.0966	-2.558	-2.558
	(2.367)	(0.549)	(0.519)	(2.534)	(2.535)
$y97$	-6.629**	-1.427**	-3.025***	-6.629**	-6.629**
	(2.984)	(0.610)	(0.568)	(3.186)	(3.188)
$y98$	6.111*	11.72***	10.05***	6.111*	6.110*
	(3.235)	(0.664)	(0.614)	(3.446)	(3.447)
$\overline{lunch}$					-0.455***
					(0.0487)
$\overline{\log(enrol)}$					-4.516
					(4.185)
$\overline{\log(found)}$					-20.40
					(17.10)
$\overline{y96}$					-2.202
					(4.575)
$\overline{y97}$					2.436
					(4.578)
$\overline{y98}$					-16.24***
					(4.732)
$\hat{\ddot{v}}_2$	-42.48*				
	(23.52)				
Constant	-361.8	38.61	-80.65***	-361.8	-136.9*
	(222.2)	(35.03)	(24.73)	(238.9)	(73.77)
Observations	5,913	5,913	5,913	5,913	5,913
R-squared	0.249	0.249			
Number of schid	1,643	1,643	1,643	1,643	1,643

Standard errors of Column 1 are valid under the null of no endogeneity of the covariates
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

TABLE II

SPECIFICATION TEST

$H_O : \boldsymbol{\xi} = \mathbf{0}$	χ_2	p value	degrees of freedom
FE2SLS vs RE2SLS	105.6668	1.64e-20	6

TABLE III

SUMMARY OF T_i

	Overall		**Between**		Within
T_i	Freq.	Percent	Freq.	Percent	Percent
1	52	0.79	13	0.79	100
2	316	4.81	79	4.81	100
3	1848	28.12	462	28.12	100
4	4356	66.28	1089	66.28	100
Total	6572	100	1643	100	100

(Number of Schools = 1643)

10. APPENDIX

10.1. *Proof of Proposition 4.1*

We apply the Frisch-Waugh-Lovell (FWL) theorem for instrument variables, hereafter FWL-IV, to the equation

$$(y_{it} - \theta_i \bar{y}_i) = (\mathbf{x}_{it} - \theta_i \bar{\mathbf{x}}_i)\boldsymbol{\beta} + (1 - \theta_i)\mathbf{w}_i \boldsymbol{\delta} + (1 - \theta_i)\bar{\mathbf{z}}_i \boldsymbol{\xi} + error_{it}$$

where the IVs for $(\mathbf{x}_{it} - \theta_i \bar{\mathbf{x}}_i)$ are $(\mathbf{z}_{it} - \theta_i \bar{\mathbf{z}}_i)$. The other explanatory variables act as their own IVs. To obtain a useful expression for $\hat{\beta}_{A2SLS}$, we first orthogonalize the instruments with respect to the regressors that are treated as exogenous. This results in the multivariate regression

$$(\mathbf{z}_{it} - \theta_i \bar{\mathbf{z}}_i) = (1 - \theta_i)\bar{\mathbf{z}}_i \boldsymbol{\Phi}_\mathbf{1} + (1 - \theta_i)\mathbf{w}_i \boldsymbol{\Phi}_\mathbf{2} + \mathbf{error}_{it}$$

We now show that applying pooled OLS on this equation, using $s_{it} = 1$, yields matrix coefficients $\hat{\boldsymbol{\Phi}}_1 = \mathbf{I}_L$ (identity matrix) and $\hat{\boldsymbol{\Phi}}_2 =$ (zero matrix). To show this, apply FWL to this regression. So first run a pooled regression of $(\mathbf{z}_{it} - \theta_i \bar{\mathbf{z}}_i)$ on $(1 - \theta_i)\bar{\mathbf{z}}_i$ for $s_{it} = 1$.

The coefficient will be

$$\begin{aligned}\hat{\mathbf{\Pi}}_1 &\equiv \left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(1-\theta_i)^2\bar{\mathbf{z}}_i'\bar{\mathbf{z}}_i\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(1-\theta_i)^2\bar{\mathbf{z}}_i'(\mathbf{z}_{it}-\theta_i\bar{\mathbf{z}}_i)\right]\\ &= \left[\sum_{i=1}^{N} T_i(1-\theta_i)^2\bar{\mathbf{z}}_i'\bar{\mathbf{z}}_i\right]^{-1}\left[\sum_{i=1}^{N}(1-\theta_i)\bar{\mathbf{z}}_i'\sum_{t=1}^{T} s_{it}(\mathbf{z}_{it}-\theta_i\bar{\mathbf{z}}_i)\right]\\ &= \left[\sum_{i=1}^{N} T_i(1-\theta_i)^2\bar{\mathbf{z}}_i'\bar{\mathbf{z}}_i\right]^{-1}\left[\sum_{i=1}^{N}(1-\theta_i)\bar{\mathbf{z}}_i'(T_i\bar{\mathbf{z}}_i-T_i\theta_i\bar{\mathbf{z}}_i)\right]\\ &= \left[\sum_{i=1}^{N} T_i(1-\theta_i)^2\bar{\mathbf{z}}_i'\bar{\mathbf{z}}_i\right]^{-1}\left[\sum_{i=1}^{N} T_i(1-\theta_i)^2\bar{\mathbf{z}}_i'\bar{\mathbf{z}}_i\right] = \mathbf{I}_L\end{aligned}$$

Therefore, residuals from this regression are

$$(\mathbf{z}_{it}-\theta_i\bar{\mathbf{z}}_{it}) - (\bar{\mathbf{z}}_i-\theta_i\bar{\mathbf{z}}_i) = \mathbf{z}_{it}-\bar{\mathbf{z}}_i \equiv \ddot{\mathbf{z}}_{it}$$

To continue with FWL, run a POLS on the regression of

$$(1-\theta_i)\mathbf{w}_i \text{ on } (1-\theta_i)\bar{\mathbf{z}}_i \text{ for } s_{it}=1$$

As it is clear, the residuals from this regression would only depend on i; call these $\tilde{\mathbf{e}}_i$; the precise formula is unecessary because they are necessarily orthogonal to the $\ddot{\mathbf{z}}_{it}$ in the selected sample:

$$\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\tilde{\mathbf{e}}_i'\ddot{\mathbf{z}}_{it} = \sum_{i=1}^{N}\tilde{\mathbf{e}}_i'\left(\sum_{t=1}^{T} s_{it}\ddot{\mathbf{z}}_{it}\right) = \mathbf{0}$$

because, by construction, $\sum_{t=1}^{T} s_{it}\ddot{\mathbf{z}}_{it}$. Therefore, we have shown that $\hat{\mathbf{\Phi}}_2 = \mathbf{0}$. It follows immediately that $\hat{\mathbf{\Phi}}_1 = \hat{\mathbf{\Pi}}_1 = \mathbf{I}_L$. Therefore, the residuals for the first step of the FWL-IV theorem are simply

$$\ddot{\mathbf{z}}_{it} = \mathbf{z}_{it}-\bar{\mathbf{z}}_i \text{ for } s_{it}=1.$$

In step 2 of the FWL-IV theorem, we use the complete cases to estimate the equation

$$(y_{it}-\theta_i\bar{y}_i) = (\mathbf{x}_{it}-\theta_i\bar{\mathbf{x}}_i)\boldsymbol{\beta} + error_{it}$$

by pooled 2SLS using instruments $\ddot{\mathbf{z}}_{it}$ – because we just showed these are the residuals from step one. By the FWL-IV theorem, the estimate is $\hat{\boldsymbol{\beta}}_{A2SLS}$. We can write it as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{A2SLS} = &\left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(\mathbf{x}_{it}-\theta_i\bar{\mathbf{x}}_i)'\ddot{\mathbf{z}}_{it}\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{z}}_{it}'\ddot{\mathbf{z}}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{z}}_{it}'(\mathbf{x}_{it}-\theta_i\bar{\mathbf{x}}_i)\right)\right]^{-1}\\ &\cdot\left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(\mathbf{x}_{it}-\theta_i\bar{\mathbf{x}}_i)'\ddot{\mathbf{z}}_{it}\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{z}}_{it}'\ddot{\mathbf{z}}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{z}}_{it}'(y_{it}-\theta_i\bar{y}_i)\right)\right].\end{aligned}$$

Now

$$\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(\mathbf{x}_{it}-\theta_i\bar{\mathbf{x}}_i)'\ddot{\mathbf{z}}_{it} = \sum_{i=1}^{N}\left(\sum_{t=1}^{T} s_{it}\mathbf{x}_{it}'\ddot{\mathbf{z}}_{it} - \sum_{t=1}^{T} s_{it}\theta_i\bar{\mathbf{x}}_i'\ddot{\mathbf{z}}_{it}\right)$$

and

$$\begin{aligned}\sum_{t=1}^{T} s_{it}\mathbf{x}_{it}'\ddot{\mathbf{z}}_{it} &= \sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{z}}_{it}\\ \sum_{t=1}^{T} s_{it}\theta_i\bar{\mathbf{x}}_i'\ddot{\mathbf{z}}_{it} &= \theta_i\bar{\mathbf{x}}_i'\sum_{t=1}^{T} s_{it}\ddot{\mathbf{z}}_{it} = \mathbf{0}\end{aligned}$$

because, for each i, deviations from means sum to zero over the complete cases. It follows that

$$\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}(\mathbf{x}_{it}-\theta_i\bar{\mathbf{x}}_i)'\ddot{\mathbf{z}}_{it} = \sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{z}}_{it}.$$

Similarly,

$$\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{z}}_{it}'(y_{it}-\theta_i\bar{y}_i) = \sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{z}}_{it}'\ddot{y}_{it}$$

We have shown that $\hat{\boldsymbol{\beta}}_{A2SLS}$ can be written as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{A2SLS} = &\left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{z}}_{it}\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{z}}_{it}'\ddot{\mathbf{z}}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{z}}_{it}'\ddot{\mathbf{x}}_{it}\right)\right]^{-1}\\ &\cdot\left[\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{z}}_{it}\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{z}}_{it}'\ddot{\mathbf{z}}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{z}}_{it}'\ddot{y}_{it}\right)\right]\end{aligned}$$

and this is precisely the FE2SLS estimator using the complete cases, $\hat{\boldsymbol{\beta}}_{FE2SLS}$. This completes the proof.

10.2. *Proof of Proposition 6.1*

The estimators are motivated by the equation of interest,

(34) $$y_{it1} = \mathbf{z}_{it1}\boldsymbol{\gamma}_1 + \mathbf{y}_{it2}\boldsymbol{\alpha}_1 + c_{i1} + u_{it1},\ t = 1, ..., T,$$

and the reduced form equation for $\mathbf{y}_{it2}$,

(35) $$\mathbf{y}_{it2} = \mathbf{z}_{it}\boldsymbol{\Gamma}_2 + \mathbf{c}_{i2} + \mathbf{v}_{it2}.$$

But the result here is purely algebraic.

The CF approach is to first step estimate the reduced form equation using fixed effects.

This is done by time demeaning (35):

$$(36) \qquad \ddot{\mathbf{y}}_{it2} = \ddot{\mathbf{z}}_{it}\boldsymbol{\Gamma}_2 + \ddot{\mathbf{v}}_{it2}$$

and applying POLS using the selected sample ($s_{it} = 1$). Denote the FE residuals as $\widehat{\ddot{\mathbf{v}}}_{it2}$. The control function equation is obtained by augmenting (34) with $\widehat{\ddot{\mathbf{v}}}_{it2}$:

$$(37) \qquad y_{it1} = \mathbf{z}_{it1}\boldsymbol{\gamma}_1 + \mathbf{y}_{it2}\boldsymbol{\alpha}_1 + \widehat{\ddot{\mathbf{v}}}_{it2}\boldsymbol{\rho}_1 + error_{it1}$$

We now show that the FE estimators of $(\boldsymbol{\gamma}_1, \boldsymbol{\alpha}_1)$ in (37) are identical to the FE2SLS estimators from (34).

To obtain the expressions for $\hat{\boldsymbol{\gamma}}_{CF,1}$, $\hat{\boldsymbol{\alpha}}_{CF,1}$, define

$$\mathbf{x}_{it1} \equiv (\mathbf{z}_{it1}, \mathbf{y}_{it2}) \text{ and } \theta_1 \equiv \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\alpha}_1 \end{pmatrix}$$

The estimating equation becomes:

$$y_{it1} = \mathbf{x}_{it1}\theta_1 + \widehat{\ddot{\mathbf{v}}}_{it2}\boldsymbol{\rho}_1 + error_{it1}$$

The FE transformation of this equation is

$$(38) \qquad \ddot{y}_{it1} = \ddot{\mathbf{x}}_{it1}\theta_1 + \widehat{\ddot{\mathbf{v}}}_{it2}\boldsymbol{\rho}_1 + error_{it1}$$

FE estimator for θ_1 is obtained by doing a POLS on (38) using $s_{it} = 1$.

To obtain an expression for $\hat{\theta}_{CF,1}$, we use the Frisch-Waugh-Lovell partialling out theorem, which involves two steps. The first is to do POLS of $\ddot{\mathbf{x}}_{it1}$ on $\widehat{\ddot{\mathbf{v}}}_{it2}$ for $s_{it} = 1$. Let the matrix of coefficients be $\hat{\boldsymbol{\Theta}}_2$. Now

$$\hat{\boldsymbol{\Theta}}_2 = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \widehat{\ddot{\mathbf{v}}}'_{it2} \widehat{\ddot{\mathbf{v}}}_{it2} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \widehat{\ddot{\mathbf{v}}}'_{it2} \ddot{\mathbf{x}}_{it1} \right)$$

Further,

$$\sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \widehat{\ddot{\mathbf{v}}}'_{it2} \ddot{\mathbf{x}}_{it1} = \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \widehat{\ddot{\mathbf{v}}}'_{it2} \left[\ddot{\mathbf{z}}_{it1}, \ddot{\mathbf{y}}_{it2} \right]$$

and, because $\widehat{\ddot{\mathbf{v}}}_{it2}$ are the FE residuals on the selected sample,

$$\sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \widehat{\ddot{\mathbf{v}}}'_{it2} \ddot{\mathbf{z}}_{it} = \mathbf{0}$$

by the first order condition of POLS on the time-demeaned variables. This, of course, implies

$$\sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \widehat{\ddot{\mathbf{v}}}'_{it2} \ddot{\mathbf{z}}_{it1} = \mathbf{0}$$

Next, we can write (for the selected sample)

$$\ddot{\mathbf{y}}_{it2} = \ddot{\mathbf{z}}_{it}\hat{\boldsymbol{\Gamma}}_2 + \widehat{\ddot{\mathbf{v}}}_{it2}$$

and so

$$\begin{aligned} \sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\widehat{\ddot{\mathbf{v}}}_{it2}'\ddot{\mathbf{y}}_{it2} &= \left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\widehat{\ddot{\mathbf{v}}}_{it2}'\ddot{\mathbf{z}}_{it}\right)\hat{\boldsymbol{\Gamma}}_2 + \sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\widehat{\ddot{\mathbf{v}}}_{it2}'\widehat{\ddot{\mathbf{v}}}_{it2} \\ &= \mathbf{0}\cdot\hat{\boldsymbol{\Gamma}}_2 + \sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\widehat{\ddot{\mathbf{v}}}_{it2}'\widehat{\ddot{\mathbf{v}}}_{it2}. \end{aligned}$$

Combining with the above, we have shown

$$\hat{\boldsymbol{\Theta}}_2 = [\mathbf{0}|\mathbf{I}_{G_1}]$$

and the matrix of zeros is $G_1 \times L_1$. We have shown the FWL residuals are

$$\begin{aligned} \widehat{\ddot{\mathbf{r}}}_{it1} &= \ddot{\mathbf{x}}_{it1} - \widehat{\ddot{\mathbf{v}}}_{it2}\,[\mathbf{0}|\mathbf{I}_{G_1}] \\ &= \left[\ddot{\mathbf{z}}_{it1}, \ddot{\mathbf{y}}_{it2} - \widehat{\ddot{\mathbf{v}}}_{it2}\right] \end{aligned}$$

But

$$\ddot{\mathbf{y}}_{it2} - \widehat{\ddot{\mathbf{v}}}_{it2} = \ddot{\mathbf{z}}_{it}\hat{\boldsymbol{\Gamma}}_2 = \widehat{\ddot{\mathbf{y}}}_{it2},$$

the fitted values from the first-stage regression the time-demeaned variables (using the selected sample). Therefore, by the FWL theorem, $\hat{\theta}_{CF,1}$ is the POLS estimator

$\ddot{y}_{it1}$ on $\ddot{\mathbf{z}}_{it1}$, $\widehat{\ddot{\mathbf{y}}}_{it2}$ using $s_{it} = 1$

and this is precisely the FE2SLS estimator on the selected sample. This completes the proof.

REFERENCES

ARELLANO, M. (1993): "On the testing of correlated effects with panel data," *Journal of econometrics*, 59(1-2), 87–97.

BALTAGI, B. (2008): *Econometric analysis of panel data*. John Wiley & Sons.

BALTAGI, B. H., AND Y.-J. CHANG (1994): "Incomplete panels: A comparative study of alternative estimators for the unbalanced one-way error component regression model," *Journal of Econometrics*, 62(2), 67–89.

GUGGENBERGER, P. (2010): "The impact of a Hausman pretest on the size of a hypothesis test: The panel data case," *Journal of Econometrics*, 156(2), 337–343.

HAUSMAN, J. A. (1978): "Specification tests in econometrics," *Econometrica: Journal of the econometric society*, pp. 1251–1271.

HAUSMAN, J. A., AND W. E. TAYLOR (1981): "Panel data and unobservable individual effects," *Econometrica: Journal of the Econometric Society*, pp. 1377–1398.

HSIAO, C. (1985): "Benefits and limitations of panel data," *Econometric Reviews*, 4(1), 121–174.

——— (2014): *Analysis of panel data*, no. 54. Cambridge university press.

KYRIAZIDOU, E. (1997): "Estimation of a panel data sample selection model," *Econometrica: Journal of the Econometric Society*, pp. 1335–1364.

LEVITT, S. D. (1995): "Using Electoral Cycles in Police Hiring to Estimate the Effect of Policeon Crime," Discussion paper, National Bureau of Economic Research.

——— (1996): "The effect of prison population size on crime rates: Evidence from prison overcrowding litigation," *The quarterly journal of economics*, 111(2), 319–351.

MUNDLAK, Y. (1978): "On the pooling of time series and cross section data," *Econometrica: journal of the Econometric Society*, pp. 69–85.

PAPKE, L. E. (2005): "The effects of spending on test pass rates: evidence from Michigan," *Journal of Public Economics*, 89(5-6), 821–839.

PAPKE, L. E., AND J. M. WOOLDRIDGE (2008): "Panel data methods for fractional response variables with an application to test pass rates," *Journal of Econometrics*, 145(1-2), 121–133.

ROCHINA-BARRACHINA, M. E. (1999): "A new estimator for panel data sample selection models," *Annales d'Economie et de Statistique*, pp. 153–181.

SEMYKINA, A., AND J. M. WOOLDRIDGE (2010): "Estimating panel data models in the presence of endogeneity and selection," *Journal of Econometrics*, 157(2), 375–380.

VERBEEK, M., AND T. NIJMAN (1992): "Testing for selectivity bias in panel data models," *International Economic Review*, pp. 681–703.

WOOLDRIDGE, J. M. (1995): "Selection corrections for panel data models under conditional mean independence assumptions," *Journal of econometrics*, 68(1), 115–132.

——— (2010): *Econometric analysis of cross section and panel data*. MIT press.

——— (2018): "Correlated random effects models with unbalanced panels," *Journal of Econometrics*.