Specification testing in panel data models estimated by fixed effects with instrumental variables

Carrie Falls
Department of Economics
Michigan State University

Abstract

I show that a handful of the regressions based tests traditional to cross-sectional or time series models can be extended to panel data models with correlated fixed effects. Specifically I extend the tests for endogeneity, overindentification, and nonlinearities developed by Wooldridge (1994). This results in regression based tests that can be easily made robust to arbitrary heteroskedasticity and/or cluster serial correlation.

I would like to thank Jeff Wooldridge for his comments, assistance, and overwhelming patience. Any contribution of my work is truly a result of his teachings.

1. Introduction

Estimation methods have become increasing complex as the availability of longitudinal data sets have grown. While there are clearly many benefits to these more sophisticated models, specification testing within such models can be arduous.

Traditional estimation methods for static panel data models are the within-group and the generalized least squares (GLS) estimators. The key distinction between the two methods is in the treatment of the unobserved component. The GLS estimator assumes the unobserved component and the observed explanatory variables are uncorrelated where as the within-group estimator allows for correlation between the unobserved component and the time averages of the explanatory variables. The focus of this paper is specification testing in the context of panel data models estimated using the within-group estimator. Throughout the paper we allow for arbitrary correlation between the time averages of the regressors and the unobserved component. We extend robust regression based methods, for testing hypotheses about the conditional mean, that have been traditionally applied to either purely cross-sectional or time-series models. Specifically, we derive test for endogeneity, over-identification, and non-linearities for the within-group estimator.

In deriving the above test we assume that at least some of the observed explanatory variables are correlated with the idiosyncratic errors. This necessitates an estimation method involving instrumental variables along with the within transformation.

The test statistics are derived under minimal assumptions pertaining to the distribution of the unobserved component. We assume the unobserved component is a random variable that is invariant through time for each unit of the cross-section. The time invariance allows us to apply standard fixed effects estimation techniques using instruments. One could, of course, allow the unobserved component to trend, assuming we had enough time periods to difference the data before applying the within transformation.

The remainder of the paper is organized as follows: section 2 reviews the existing testing literature pertaining to panel data models. Section 3 develops the model and assumptions necessary for consistent estimation of the population parameters. Section 4 derives the test statistic and its asymptotic distribution. Section 5 lays out each of the three tests specifically and section 6 presents a brief conclusion.

2. Literature Review

Panel data is unique in that it has both a cross-section and time-series dimension. This allows us to construct specification tests by applying either T or N asymptotics. The decision of which dimension you assume fixed verses growing infinitely is imperative. Assuming the time-series dimension tending off to infinity changes the whole setting of the model; while the unobserved component is no longer a problem the time-series nature of the observed explanatory variables is. Traditionally the approach in panel data models has been to assume the time dimension is fixed while the cross-sectional dimension tends to infinity. However, Baltagi discusses panel data models containing correlated unobserved components using T asymotiotics.

In deriving the test statistic we assume the cross-section to grow while the time-series dimension remains fixed. We discuss the implications of allowing the time-dimension to grow in the conclusion. Therefore, the focus of the literature review is on panel data models where the cross-sectional dimension is thought to be significantly large relative to the time-dimension.

Much of the testing literature pertaining to panel data models builds on the work of Hausman (1978). In general, the Hausman test can be applied anytime an econometric model can be consistently estimated under the alternative hypothesis as well as under the null. The test is based on comparing the two estimates. Since, under the null hypothesis both estimation procedures are consistent, therefore, observing a statistical difference between the two provides evidence against the null. The standard, and most widely applied, version of the test assumes that the estimation procedure under the null hypothesis is more efficient. This assumption simplifies the calculation of the variance covariance matrix. While it is possible to compute the variance covariance matrix without maintaining such efficiency assumptions, most statistical packages do not allow for computation of the robust version directly.

Hausman and Taylor (1981) derive estimation and testing methods in the context of linear panel data models containing correlated fixed effects where interest lies in the parameters associated with observed time invariant explanatory variables. More specifically they develop an estimation procedure using instruments to estimate the parameters of the observed time-invariant variables. Where the instruments are the within transformations of the time-varying explanatory variables that are assumed to have no relationship with the unobserved component. Therefore, Hausman and Taylor do not rely on instrument from outside the model. They also extend the work of Hausman in the context of testing for correlation between the unobserved component and the included explanatory variables.

Metcalf (1996) extends above procedures of Hausman and Taylor to models containing endogenous variables in addition to the correlated fixed effects. That is, Metcalf requires instruments outside of the model. The test statistic developed is then pertaining to possible correlation between the instruments and the unobserved component.

Ahn and Low (1996) further extend the testing literature regarding panel data models through reformulating the Hausman test in the context of GMM estimation. They note that the Hausman test statistic for testing correlation between the unobserved component and the regressors implies that the individual means or time averages of the regressors are exogenous. Their alternative GMM statistic incorporates a much broader set of moment conditions signifying that each of the time-varying explanatory variables is exogenous. The key feature of the alternative GMM test statistic is that unlike the Hausman test it has power in detecting nonstationary coefficients of the regressors.

The tests derived in this paper differ in two distinct ways: first most of the existing tests focus on testing for correlation between the unobserved effect and observed explanatory variables where as here, we assume the unobserved effects are correlated with the included explanatory variables. The focus here is on testing for correlation between the observed explanatory variables and the idiosyncratic error. Second, the tests derived in this paper focus on computational ease for both the non-robust and robust test statistic. Where, the robust version is robust to arbitrary forms of heteroskedasticity as well as within-group serial correlation. While many statistical packages will compute the non-robust version of the Hausman statistic following a regression; it in general is much more tedious obtaining the robust version. In addition, it's not clear how to apply the Hausman test to panel data models containing both correlated unobserved components and endogenous explanatory variables.

The test statistics derived in this paper extend the regression based tests derived by Wooldridge (1994). Wooldridge develops a score type test statistic in the context of time series models estimated by two-stage least squares. The appeal of the tests is both computational ease and flexibility. The flexibility of the test results from a partialling out method. This allows for misspecification indicators that are non-linear functions of the instruments, for instruments that are not strictly exogenous, and for errors that that are not assumed to be independent or even uncorrelated. The resulting test statistic can be applied to a variety of hypotheses including endogeneity, serial correlation, overindentification, and nonlinearities, all of which can be easily made robust to arbitrary forms of heteroskedasticity and serial correlation.

3. Model and Assumptions

Let $\{(\mathbf{z}_{it}, \mathbf{x}_{it}, y_{it}): i = 1,\ldots\ldots N;\ t = 1,\ldots\ldots\ldots T\}$ be a sequence of observations, where $\mathbf{z}_{it}$ is a $1\times L$ vector of instruments, $\mathbf{x}_{it}$ is a $1\times K$ vector of time-varying explanatory variables, and y_{it} is the dependent variable. We assume random sampling in the cross-section and fixed T. Generally, the set of instruments and the set of explanatory variables will overlap. For example, it is often the case that time dummies are included in the set of explanatory variables. We are interested in the standard unobserved effects model

$$y_{it} = \mathbf{x}_{it}\beta + c_i + u_{it} \qquad (i = 1,\ldots\ldots..N\ ;\ t = 1,\ldots\ldots\mathrm{T}). \tag{2.1}$$

Where, c_i is the time-invariant unobserved heterogeneity and u_{it} is the idiosyncratic error. We are interested in consistently estimating the K $\times 1$ vector β. Throughout the paper we allow for arbitrary correlation between c_i and $\mathbf{x}_{it}$, thus we eliminate c_i using the fixed effects or within transformation:

(2.2) $$\ddot{y}_{it} = \ddot{x}_{it}\beta + \ddot{u}_{it} \qquad (i = 1,........N\ ;\ t = 1,\ldots\ldots T).$$

The time demeaning necessitates strict exogeneity for the consistency of the estimation procedures. That is

(2.3) $$E(\mathbf{x}'_{is}u_{it}) = 0, \qquad (s,t = 1,.......T).$$

Maintaining we know the fixed effects estimation of β is consistent and asymptotically normal. However, if we believe that at least some of the elements of $\boldsymbol{x}_{it}$ are correlated with the idiosyncratic error then the standard fixed effects estimation procedure is inconsistent. In such cases the existence of instruments allows for estimation of β by pooled two-stage least squares (2SLS) using time-demeaned instruments.

Under the appropriate assumptions, this estimation procedure is consistent and asymptotically normal. We then analyze the model (2.1) under the following assumptions.

Assumption A.1 $$E(\mathbf{z}'_{is}u_{it}) = 0 \qquad (s,t = 1,.......T).$$

A sufficient condition for assumption A.1 is the zero conditional mean assumption

(2.4) $$E(u_{it} \mid \mathbf{z}_i) = 0.$$

Where, $\mathbf{z}_i$ is a T $\times$L matrix of instrumental variables. The general analysis assumes A.1 since it is a weaker condition. However, it is necessary to assume 2.4 in the test for nonlinearities as shown in the examples. The next assumption imposed is the identification condition.

Assumption A.2 Let $\ddot{\mathbf{x}}_{it}^{*} \equiv L(\ddot{\mathbf{x}}_{it} \mid \ddot{\mathbf{z}}_{it}) = \ddot{\mathbf{z}}_{it}\Pi$, where, $\Pi \equiv \sum_{t=1}^{T}\left[E(\ddot{\mathbf{z}}_{it}'\ddot{\mathbf{z}}_{it})\right]^{-1} E(\ddot{\mathbf{z}}_{it}'\ddot{\mathbf{x}}_{it})$ and

$L(\ddot{\mathbf{x}}_{it}| \ddot{\mathbf{z}}_{it})$ denotes the linear projection of $\ddot{\mathbf{x}}_{it}$ onto $\ddot{\mathbf{z}}_{it}$. Then let

$$E(\ddot{\mathbf{X}}_{i}^{*\prime}\mathbf{X}_{i}) \equiv \mathbf{A} \tag{2.5}$$

Where A is positive definite, that is, the rank of $\mathbf{A}$ = K. Thus a necessary condition is the order condition, meaning $L \geq K$.

The pooled 2SLS estimator of β can be expressed as

$$\hat{\beta} = (\sum_{i=1}^{N}\hat{\ddot{\mathbf{X}}}_{i}'\hat{\ddot{\mathbf{X}}}_{i})^{-1}(\sum_{i=1}^{N}\hat{\ddot{\mathbf{X}}}_{i}'\mathbf{y}_{i}) = \beta + (\sum_{i=1}^{N}\hat{\ddot{\mathbf{X}}}_{i}'\hat{\ddot{\mathbf{X}}}_{i})^{-1}(\sum_{i=1}^{N}\hat{\ddot{\mathbf{X}}}_{i}'\mathbf{u}_{i}). \tag{2.6}$$

Where, $\hat{\ddot{\mathrm{x}}}_{it} = \ddot{\mathbf{z}}_{it}\hat{\Pi}$ are the fitted values from the first-stage regression of

$$\ddot{\mathbf{x}}_{it} \text{ on } \ddot{\mathbf{z}}_{it}; \quad i = 1,........N \; ; t = 1,.......T. \tag{2.6}$$

Under assumptions A.1 and A.2 and finite second moment assumptions

$$\hat{\beta} \xrightarrow{p} \beta.$$

This follows from the law of large numbers (LLN); since $E(\ddot{\mathbf{X}}_{i}^{*\prime}\mathbf{u}_{i})=0$ under A.1,

$E(\ddot{\mathbf{X}}_{i}^{*\prime}\ddot{\mathbf{X}}_{i}^{*})$ is nonsingular underA.2, and $\hat{\Pi} \xrightarrow{p} \Pi$.

By the central limit theorem (CLT)

$$N^{-1/2}\sum_{i=1}^{N}\ddot{\mathbf{X}}_{i}^{*}\mathbf{u}_{i} \xrightarrow{d} \text{normal } (\mathbf{0},\mathbf{B}). \tag{2.7}$$

Where

$$\mathbf{B} \equiv E(\ddot{\mathbf{X}}_{i}^{*\prime}\mathbf{u}_{i}\mathbf{u}_{i}\ddot{\mathbf{X}}_{i}^{*}). \tag{2.8}$$

Then the first order asymptotic representation is

(2.9) $$\sqrt{N}(\hat{\beta}-\beta) = \mathbf{A}^{-1} N^{-1/2} \sum_{i=1}^{N} \ddot{\mathbf{X}}_{\mathbf{i}}^{*\prime}\mathbf{u}_{\mathbf{i}} + o_p(1).$$

Equation (2.9) shows that the first-stage regression does not affect the limiting distribution of the pooled 2SLS estimator. That is, the same limiting distribution is obtained when $\hat{\ddot{x}}_{it}$ replaces $\ddot{\mathbf{x}}_{\mathbf{it}}^{*}$, or when $\hat{\Pi}$ replaces Π. Without imposing further assumptions

(2.10) $$\sqrt{N}(\hat{\beta}-\beta) \xrightarrow{d} \text{normal}(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}).$$

Thus the asymptotic variance of $\hat{\beta}$ is

(2.11) $$\text{avar}(\hat{\beta}) = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}/N.$$

We could of course simplify (2.11) by imposing homoskedasticity and no serial correlation assumptions. However, we are interested in the fully robust statistic. Later as a special case we consider the case of no serial correlation and/or homoskedasticity.

3. Specification Tests

The specification tests developed here focus on when and if the estimation procedure outline in section two is consistent and necessary. That is we want to test if a particular set of variables are related to the idiosyncratic error. For example, if interest lies in testing when the two stage method is necessary we would want to test whether or not $\mathbf{x}_{it}$ is endogenous. In deriving the test statistic and its properties we generalize by defining a set of misspecification indicators that are to be tested for correlation with the idiosyncratic error. This allows us to model the three tests regarding the conditional mean of equation 2.2 in one setup.

Let $\mathbf{v}_{it}(\gamma)$ be a $1 \times Q$ vector of misspecification indicators that may possibly depend on some $G \times 1$ vector of parameters (γ) and elements of $\{\mathbf{z}_{it}, \mathbf{x}_{it}, y_{it}\}$. Where we assume a consistent estimator of γ, that is, $\sqrt{N}(\hat{\gamma} - \gamma) = O_p(1)$.

The tests take the null hypothesis to be

(3.1) $$H_o: \ E(\mathbf{v}'_{it} u_{is}) = 0, \qquad s,t = 1,\ldots\ldots T.$$

That is, we are interested in testing whether $\mathbf{v}_{it}$ is strictly exogenous in (2.1). This is equivalent to

(3.2) $$H_o: \ E(\ddot{\mathbf{v}}'_{it} \ddot{u}_{it}) = 0.$$

In deriving a test of (3.1) we note that

$$E(\ddot{\mathbf{v}}'_{it} \ddot{u}_{it}) = E(\ddot{\mathbf{v}}'_{it} u_{it}).$$

Thus the test statistics derived take the null hypothesis to be

(3.3) $$H_o: \ E(\ddot{\mathbf{v}}'_{it} u_{it}).$$

While the equivalence between (3.1) and (3.3) simplifies the derivation of the test statistics it is not without cost, we have to restrict the analysis to tests about the conditional mean. For example we cannot take $\mathbf{v}_{it} = (u_{it-1}, \ldots u_{it-q})$, since under the null hypothesis of no serial correlation, (3.3) is naturally violated.

Tests of (3.1) are derived under assumptions that are extensions of those made in section 2; when each is used it is assumed to hold under the null. However, the robust forms of the tests are most easily estimated under the alternative; we therefore analyze the behavior of the tests under local alternatives.

The first assumption is

Assumption B.1 Assumption A.1 holds and, in addition

(i) $E[\ddot{\mathbf{v}}'_{it}(\gamma)u_{it}] = 0 \quad s,t = 1,\ldots.T$

(ii) $E[\nabla \ddot{\mathbf{v}}'_{it}(\gamma)u_{it}] = 0$.

Where $\nabla \ddot{\mathbf{v}}_{it}(\gamma)$ denotes the gradient of $\ddot{\mathbf{v}}_{it}$ with respect to γ. Part (ii) of B.1 ensures that the estimation of γ does not effect the limiting distribution of the test statistic so long as $\sqrt{N}(\hat{\gamma} - \gamma) = O_p(1)$. Requiring B.1 (ii) is innocuous since it is trivially satisfied in the test for endogeneneity and overidentification under A.1 and for the test of nonlinearities it is satisfied under the zero conditional mean assumption (2.4).

The sample analog of (3.3) is

(3.4) $$N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\ddot{\mathbf{v}}}'_{it}\hat{u}_{it}$$

Where "^" denotes that each function is evaluated at $(\hat{\beta}, \hat{\gamma})$. To use (3.4) as a test of (3.3), entails deriving the limiting distribution of

(3.5) $$N^{-1/2}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\ddot{\mathbf{v}}}'_{it}\hat{u}_{it}$$

under the null hypothesis. However, it is often the case that the limiting distribution of (3.5) is different from the limiting distribution of

$$N^{-1/2}\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{\mathbf{v}}'_{it}u_{it}.$$

That is, we need to know when the estimation of β does not effect the limiting distribution. Since B.1 (ii) ensures that the estimation of $\ddot{\mathbf{v}}_{it}$ does not affect the limiting distribution so long as $\sqrt{N}(\hat{\gamma} - \gamma) = O_p(1)$; deriving a quadratic form of (3.5) that has an asymptotic chi-squared distribution is straightforward, although nothing ensures that the resulting statistic will be easy to compute via regressions. A simpler approach is to transform (3.5) so that the first stage estimation does not affect the limiting distribution. This results in robust test statistics that are easy to compute using a series of regressions, and are asymptotically equivalent so long as a consistent estimation procedure is used in the first-stage estimation as shown by Wooldridge (1990). This brings us to our second assumption.

Assumption B.2 Assumption A.2 holds and, in addition, let

$$\ddot{\mathbf{x}}_{it}^{*} \equiv L(\ddot{\mathbf{x}}_{it} | \ddot{\mathbf{z}}_{it}, \ddot{\mathbf{v}}_{it}).$$

Then

$$\text{rank } \mathbf{T}^{-1}\sum_{\mathbf{t}=1}^{\mathbf{T}} \mathbf{E}(\mathbf{r}_{it}'\mathbf{r}_{it}) = Q.$$

Where $\mathbf{r}_{it}$ is a $1 \times Q$ vector of time-demeaned population residuals from the linear projection of $\ddot{\mathbf{v}}_{it}$ onto $\ddot{\mathbf{x}}_{it}^{*}$

$$\mathbf{r}_{it} \equiv \ddot{\mathbf{v}}_{it} - L(\ddot{\mathbf{v}}_{it} | \ddot{\mathbf{x}}_{it}^{*}).$$

We adjust (3.5) by replacing $\hat{\ddot{\mathbf{v}}}_{it}$ with $\hat{\mathbf{r}}_{it}$. This partialling out of $\ddot{\mathbf{v}}_{it}$ essentially allows us to ignore the first-stage estimation when deriving the limiting distribution of

(3.6) $$N^{-1/2}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\mathbf{r}}_{it}'\hat{u}_{it} .$$

Where $\hat{\mathbf{r}}_{it}$ is obtained from the regression

(3.7) $$\hat{\ddot{\mathbf{v}}}_{it} \text{ on } \hat{\ddot{\mathbf{x}}}_{it}^{*}; \qquad i = 1,\ldots\ldots N \; ; t = 1,\ldots\ldots T.$$

Where $\hat{\ddot{\mathbf{x}}}_{it}^{*}$ are the fitted values obtained from the regression

(3.8) $$\ddot{\mathbf{x}}_{it} \text{ on } \hat{\ddot{\mathbf{v}}}_{it}, \ddot{\mathbf{z}}_{it}; \qquad i = 1,\ldots\ldots N \; ; t = 1,\ldots\ldots T.$$

However, in the test for endogentiety and overidentification $\hat{\ddot{\mathbf{v}}}_{it}$ can be omitted from (3.8) since in both tests $\ddot{\mathbf{v}}_{it}$ is a strict subset of $\ddot{\mathbf{z}}_{it}$ under the null hypothesis. We discus when $\hat{\ddot{\mathbf{v}}}_{it}$ can be omitted from (3.8) in the test for nonlinearities in the examples.

In the derivation of the first-order asymptotic representation of (3.6) we assume B.1 and B.2 hold under the null hypothesis. The first step shows that estimation of $\mathbf{r}_{it}$ does not affect the first-order asymptotic distribution. That is,

$$N^{-1/2}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\mathbf{r}}_{it}'\hat{u}_{it} = N^{-1/2}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{r}_{it}'\hat{u}_{it} + o_p(1)$$

under H_o. This can be verified by a mean-value expansion, since $\hat{\mathbf{r}}_{it}$ is a linear function of $(\hat{\ddot{\mathbf{v}}}_{it}(\hat{\gamma}), \ddot{\mathbf{z}}_{it})$ B.1(ii) ensures that $N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\ddot{\mathbf{x}}}_{it}^{*\prime}\hat{u}_{it} \xrightarrow{p} 0$. Next we show that estimation of u_{it} does not affect the first-order asymptotic distribution by a mean-value expansion of

$$N^{-1/2}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{r}_{it}'\hat{u}_{it} = N^{-1/2}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{r}_{it}'u_{it} + N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{r}_{it}'\ddot{\mathbf{x}}_{it}\sqrt{N}(\hat{\beta} - \beta) = N^{-1/2}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{r}_{it}'u_{it} + o_p(1)$$

By the CLT, $\sqrt{N}(\hat{\beta}-\beta) = O_p(1)$ and by definition of $\ddot{\mathbf{x}}_{it}^{*}$, $\ddot{\mathbf{x}}_{it} = \ddot{\mathbf{x}}_{it}^{*} + g_{it}$ where $E(\ddot{\mathbf{z}}_{it}'g_{it}) = 0$, $E(\ddot{\mathbf{v}}_{it}'g_{it}) = 0$. Therefore, $E(\mathbf{r}_{it}'g_{it}) = 0$, thus $E(\mathbf{r}_{it}'\ddot{\mathbf{x}}_{it}) = E(\mathbf{r}_{it}'\ddot{\mathbf{x}}_{it}^{*}) = 0$; since $\mathbf{r}_{it}$ is the population residual from the linear projection of $\ddot{\mathbf{v}}_{it}$ onto $\ddot{\mathbf{x}}_{it}^{*}$, then by the LLN $N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{r}_{it}'\ddot{\mathbf{x}}_{it} \xrightarrow{p} 0$. This asymptotic equivalence simplifies finding the limiting distribution of (3.6) since we can ignore the fact that $\mathbf{r}_{it}$ and u_{it} have been estimated.

Define

(3.9) $$C \equiv \text{var}\left(\sum_{t=1}^{T} \mathbf{r}_{it}' u_{it}\right).$$

The test statistic proposed is

(3.10) $$\hat{\delta} \equiv \left(N^{-1/2}\sum_{i=1}^{N}\sum_{t=1}^{T} \hat{\mathbf{r}}_{it}'\hat{u}_{it}\right)'\hat{\mathbf{C}}^{-1}\left(N^{-1/2}\sum_{i=1}^{N}\sum_{t=1}^{T} \hat{\mathbf{r}}_{it}'\hat{u}_{it}\right)$$

$$= \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \hat{\mathbf{r}}_{it}'\hat{u}_{it}\right)'(N\hat{\mathbf{C}})^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} \hat{\mathbf{r}}_{it}'\hat{u}_{it}\right).$$

Provided a consistent estimator of **C** is used $\hat{\delta} \xrightarrow{d} \chi^2$ with Q degrees of freedom under B.1 and B.2. Of course the actual form of the test statistic depends on the estimator used for **C**. We first derive the fully robust version of the test. That is, a test statistic that is robust to heteroskedasticity of unknown form and to cluster serial correlation. A consistent estimator of (3.9) is then

(3.11) $$\hat{\mathbf{C}} = N^{-1}\sum_{i=1}^{N}\sum_{s=1}^{T}\sum_{t=1}^{T} \hat{\mathbf{r}}_{it}'\hat{\mathbf{r}}_{is}\hat{u}_{is}\hat{u}_{it}.$$

Therefore the proposed test statistic in (3.10) takes the form

(3.12) $$\hat{\delta} = \left(\sum_{\hat{\delta}\, i=1}^{N}\sum_{t=1}^{T} \hat{\mathbf{r}}_{it}'\hat{u}_{it}\right)'\left(\sum_{i=1}^{N}\sum_{s=1}^{T}\sum_{t=1}^{T} \hat{\mathbf{r}}_{it}'\hat{\mathbf{r}}_{is}\hat{u}_{is}\hat{u}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} \hat{\mathbf{r}}_{it}'\hat{u}_{it}\right).$$

The above test statistic, in general, cannot be computed via regressions. However, there exist a test statistic that is asymptotically equivalent to (3.12) under H_o and local alternatives that can be easily computed using simple regression. The alternative statistic is the fully robust Wald statistic of $\hat{\lambda}$ from the regression

(3.13) $$\hat{u}_{it} \text{ on } \hat{\mathbf{r}}_{it}; \quad i = 1,\ldots\ldots N;\ t = 1,\ldots\ldots T.$$

Where $\hat{\lambda}$ is the Q×1 vector of coefficients on $\hat{\mathbf{r}}_{it}$. This results in a test statistic of the form

$$(3.14) \qquad \hat{\Gamma} \equiv (\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\mathbf{r}}_{it}'\hat{u}_{it})'(\sum_{i=1}^{N}\sum_{s=1}^{T}\sum_{t=1}^{T}\hat{\mathbf{r}}_{it}'\hat{\mathbf{r}}_{is}\hat{e}_{is}\hat{e}_{it})^{-1}(\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\mathbf{r}}_{it}'\hat{u}_{it}).$$

Where, $\hat{e}_{it}$ is the residual from the regression in (3.13). Thus the regression form of the test replaces ($\hat{u}_{is}\hat{u}_{it}$) with ($\hat{e}_{is}\hat{e}_{it}$) in the estimation of **C**, that is (3.14) is just (3.12) estimated under the alternative that $\lambda \neq 0$. It can be shown that for sequence of local alternatives $\lambda_N = \lambda N^{-1/2}$, where $u_{it} = \mathbf{r}_{it}\lambda_N + e_{it}$; the test statistic in (3.14) has a limiting non-central chi-squared distribution with non-centrality parameter equal to λ. This facilitates interpreting a rejection since $\lambda_N = \lambda N^{-1/2} \longrightarrow 0$ as $N \longrightarrow \infty$ under the sequence of local alternatives, as well as under H_o where $\hat{\lambda} \xrightarrow{p} 0$.

The more restrictive form of (3.10) adds the following assumptions.

Assumption B.3 The homoskedasticity assumption

$$E(u_{it}^2 \mid \mathbf{z}_i,\mathbf{v}_i) = \sigma^2_u.$$

Assumption B.4 The no serial correlation assumption

$$E(u_{it}u_{it+j} \mid \mathbf{z}_i,\mathbf{v}_i) = 0 \quad j \geq 1.$$

Under B.3 and B.4 a consistent estimator of **C** is

$$(3.15) \qquad \hat{\mathbf{C}} = \hat{\sigma}^2 N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\mathbf{r}}_{it}'\hat{\mathbf{r}}_{it}.$$

With

$$\hat{\sigma}^2 = (N(T-1))^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{u}_{it}^2$$

This gives the following form for the test statistic in (3.10)

$$(3.16) \quad \hat{\delta} = (\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\mathbf{r}}_{it}'\hat{u}_{it})'((N(T-1))^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{u}_{it}^2\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\mathbf{r}}_{it}'\hat{\mathbf{r}}_{it})^{-1}(\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\mathbf{r}}_{it}'\hat{u}_{it}) = N(T-1)R_u^2 \sim \chi^2_Q$$

where R_u^2 is the R-squared from the regression in (3.13).

4. Examples

Example 4.1 (testing for endogeneity)). Partition the model as

(4.1) $$y_{it} = \mathbf{x}_{it1}\beta_1 + \mathbf{x}_{it2}\beta_2 + c_i + u_{it} .$$

Where, $\mathbf{x}_{it1}$ is maintained to be exogenous. The null hypothesis of interest is then H_o: $E(\ddot{\mathbf{x}}'_{it2}u_{it}) = 0$. The model is estimated by Fixed Effects under the null, that is, $\mathbf{z}_{it} = \mathbf{x}_{it.}$ The test statistic is based on the sample covariance between u_{it} and the fitted values from the regression

(4.2) $$\ddot{\mathbf{x}}_{it2} \text{ on } \ddot{\mathbf{w}}_{it} ;\ \ i = 1,........N\ ;\ t = 1,.......T$$

where $\ddot{\mathbf{w}}_{it}$ is a set of instruments that contains $\ddot{\mathbf{x}}_{it1}$ but not $\ddot{\mathbf{x}}_{it2}$. A necessary condition is that $\ddot{\mathbf{w}}_{it}$ contain at least as many variables as $\ddot{\mathbf{x}}_{it}$. We then define $\mathbf{r}_{it} \equiv \ddot{\mathbf{x}}_{it2}^* - \ddot{\mathbf{x}}_{it}\theta$, where $\ddot{\mathbf{x}}_{it2}^*$ is the population fitted value from the regression in (4.2). Thus $\ddot{\mathbf{v}}_{it}(\gamma)' \equiv \ddot{\mathbf{x}}_{it2}^* = \ddot{\mathbf{w}}_{it}\gamma$. Therefore, the test is really testing if a particular linear combination of the instruments $\ddot{\mathbf{w}}_{it}$ are correlated with u_{it}.

Procedure (4.1)

(1) Obtain $\hat{u}_{it}$ form the estimation of (2.1) by fixed effects.

(2) Obtain the predicted values of $\ddot{\mathbf{x}}_{it2}^*$ as the fitted values from the regression of $\ddot{\mathbf{x}}_{it2}$ on $\ddot{\mathbf{w}}_{it}$.

(3) Regress the fitted values from step two on $\ddot{\mathbf{x}}_{it}$ and obtain the residuals $\hat{\mathbf{r}}_{it}$.

(4) Regress $\hat{u}_{it}$ on $\hat{\mathbf{r}}_{it}$.

Then the robust version of the test is attained from computing the fully robust version of the Wald test and the non-robust test is just N(T-1) times the R-squared.

Example 4.2 (Testing for overidentifying restrictions)

There $\hat{u}_{it}$ denotes the pooled 2SLS residuals using instruments $\ddot{\mathbf{z}}_{it}$. To test for overidentifying restrictions we must have L > K where Q ≡ L-K. Let $\ddot{\mathbf{v}}_{it}$ be the 1 ×Q vector of possible overidentifying restrictions, thus the null hypothesis is

H_o: E($\ddot{\mathbf{v}}'_{it}u_{it}$) =0. We define $\mathbf{r}_{it} \equiv \ddot{\mathbf{v}}_{it} - L(\ddot{\mathbf{v}}_{it} \mid \hat{\ddot{\mathbf{x}}}_{it})$, where $\hat{\hat{\ddot{x}}}_{it}$ are the fitted values from the first-stage regression of $\ddot{\mathbf{x}}_{it}$ on $\ddot{\mathbf{z}}_{it}$.

Procedure 4.2

(1) Obtain $\hat{u}_{it}$ from pooled 2SLS on the time-demeaned data using the full set of instruments $\ddot{\mathbf{z}}_{it}$.

(2) Define $\ddot{\mathbf{v}}_{it}$ as a strict subset of $\ddot{\mathbf{z}}_{it}$, regress $\ddot{\mathbf{v}}_{it}$ on $\hat{\hat{\ddot{x}}}_{it}$, the fitted values from the first-stage regression in step one; save the residuals as $\hat{\mathbf{r}}_{it}$.

(3) Same as step four in procedure (4.1)

Example 4.3 (Testing for nonlinearities). As in test for endogeneity, partition (2.1) as

(4.3) $$y_{it} = \mathbf{x}_{it1}\beta_1 + \mathbf{x}_{it2}\beta_2 + c_i + u_{it} \quad (i = 1,........N \ ; t = 1,.......T)$$

where $\mathbf{x}_{it1}$ exogenous and $\mathbf{x}_{it2}$ is endogenous. Therefore, $\mathbf{x}_{it1}$ is contained in $\mathbf{z}_{it}$ the set of instruments. Note that the test for nonlinearities assumes the more restrictive zero conditional mean assumption (2.4) under the null hypothesis. To obtain a RESET-type (Ramsey 1969) we augment the (2.1)

(4.4) $$y_{it} = \mathbf{x}_{it1}\beta_1 + \mathbf{x}_{it2}\beta_2 + c_i + \alpha_1(\mathbf{x}_{it1}\beta_1 + \mathbf{x}_{it2}\beta_2 + c_i)^2 + \alpha_2(\mathbf{x}_{it1}\beta_1 + \mathbf{x}_{it2}\beta_2 + c_i)^3 + u_{it}$$

$$(i = 1,........N \ ; t = 1,.......T).$$

The null hypothesis is H_o: $\alpha_1 = \alpha_2 = 0$.

Note that the time-demeaning of (4.4) eliminates any interaction between the time-invariant unobservables and time-varying explanatory variables since the correlation between them is usually thought to be though the time average on the explanatory variables. That is, in general we think that the time-demeaned explanatory variables are unrelated to the time-invariant unobservables. This results in the following augmentation of the estimating equation (2.2)

(4.5) $\ddot{y}_{it} = \ddot{\mathbf{x}}_{it1}\beta_1 + \ddot{\mathbf{x}}_{it2}\beta_2 + \alpha_1(\ddot{\mathbf{x}}_{it1}\beta_1 + \ddot{\mathbf{x}}_{it2}\beta_2)^2 + \alpha_2(\ddot{\mathbf{x}}_{it1}\beta_1 + \ddot{\mathbf{x}}_{it2}\beta_2)^3 + \ddot{u}_{it}$

$(i = 1, \ldots\ldots N\ ;\ t = 1, \ldots\ldots T)$.

However, if $E(\ddot{\mathbf{x}}'_{it}c_i) \neq 0$ then the following test can be viewed as a test for possible interactions between the time-varying explanatory variables and the time-invariant unobservables.

Procedure 4.3

(1) Obtain $\hat{u}_{it}$ from the pooled 2SLS regression on the time-demeaned data using instruments $\ddot{\mathbf{z}}_{it}$.

(2) Define the 1×2 vector $\hat{\mathbf{h}}_{it} \equiv \{(\ddot{\mathbf{x}}_{it1}\hat{\beta}_1 + \ddot{\mathbf{x}}_{it2}\hat{\beta}_2)^2, (\ddot{\mathbf{x}}_{it1}\hat{\beta}_1 + \ddot{\mathbf{x}}_{it2}\hat{\beta}_2)^3\}$ where we use the coefficient estimates obtained in step (1).

(3) Augment the instruments $\mathbf{z}_{it}$ with some non-linear functions of $\mathbf{z}_{it}$, that is, define $\mathbf{g}_{it} \equiv g(\mathbf{z}_{it})$. Generally you would include the squares, cubes and cross products of $\mathbf{z}_{it}$ in $\mathbf{g}_{it}$. Obtain $\hat{\ddot{\mathbf{x}}}_{it2}^*$ as the fitted values from the pooled OLS regression

$$\ddot{\mathbf{x}}_{it2} \text{ on } \ddot{\mathbf{g}}_{it}, \ddot{\mathbf{z}}_{it}; \qquad i = 1, \ldots\ldots N\ ;\ t = 1, \ldots\ldots T.$$

Define $\hat{\ddot{\mathbf{x}}}_{it}^* \equiv (\ddot{\mathbf{x}}_{it1}, \hat{\ddot{\mathbf{x}}}_{it2}^*)$.

Note that step three can be omitted if $E(\ddot{\mathbf{x}}_{it2} \mid \ddot{\mathbf{z}}_{it})$ is linear in $\ddot{\mathbf{z}}_{it}$. Then $\hat{\ddot{\mathbf{x}}}_{it}^{*} \equiv (\ddot{\mathbf{x}}_{it1}, \hat{\hat{\ddot{\mathbf{x}}}}_{it2})$, where $\hat{\hat{\ddot{\mathbf{x}}}}_{it2}$ are the fitted values from the first-stage regression of $\ddot{\mathbf{x}}_{it2}$ on $\ddot{\mathbf{z}}_{it}$ from step one.

(4) Obtain $\hat{\ddot{\mathbf{v}}}_{it}$ as the 1×2 vector of fitted values from the regression

$$\hat{\mathbf{h}}_{it} \text{ on } \ddot{\mathbf{g}}_{it}, \ddot{\mathbf{z}}_{it}; \qquad i = 1,........N \; ; t = 1,\ldots\ldots T.$$

(5) Obtain $\hat{\mathbf{r}}_{it}$ as the residuals from the regression

$$\hat{\ddot{\mathbf{v}}}_{it} \text{ on } \hat{\ddot{\mathbf{x}}}_{it}^{*}; \quad i = 1,........N \; ; t = 1,\ldots\ldots T.$$

(6) Same as step four in procedure (4.1)

5. Conclusion

The above procedures offer simple methods for testing hypotheses about the conditional mean in linear panel data models estimated by pooled 2SLS using time-demeaned data. We derived test statistics that are robust to heteroskedasticity of unknown form as well as cluster serial correlation. However, we have not address the presence of Arch errors. That is even the robust form the tests assumes the variance is constant across time. If the time dimension is relatively small such an assumption is minor. However, as the time dimension grows one would need to consider the potential problems associated with time-series models. Since we assume a fixed T we have not addressed such issues.