Journal of Statistical and Econometric Methods, Vol. 10, No. 4, 2021, 19-25
ISSN: 2241-0384 (print), 2241-0376 (online)
https://doi.org/10.47260/jsem/1043
Scientific Press International Limited

Random Effects Endogeneity: A Variable Pretest

Mitch Kunce[1]

Abstract

The appealing but complex Hausman and Taylor (1981) random effects (instrumental variable) estimator requires prior knowledge that certain explanatory variables in a panel are uncorrelated with the latent group effects. The purpose of this examination is to outline a tractable variable pretest that facilitates the initial sorting of regressors as likely exogenous or endogenous. The variable pretest proposed herein builds on the pretest estimator suggested by Baltagi et al (2003) by providing the necessary foundation for regressor identification. Extensions are suggested for the two-way error components construct.

Keywords: Panel data, Random effects, Variable pretest, Hausman-Taylor.
JEL Classification: C12, C13, C23.

[1] DouglasMitchell Econometric Consulting, Laramie, WY USA

Article Info: *Received*: October 28, 2021. *Revised*: November 14, 2021.
Published online: November 19, 2021.

1. Introduction

Consider the well known one-way time-series and cross-section model,

$$Y_{it} = X_{it}\beta + Z_i\gamma + \varepsilon_{it} \qquad (i = 1, \ldots, N;\ t = 1, \ldots, T;\ n = NT), \tag{1}$$

where Y_{it} is the dependent variable, X_{it} are observable variables that vary across groups *i* and over time *t*, Z_i are observable time-invariant variables, β and γ are *k* and *f* vectors of estimated coefficients and ε_{it} denotes the overall error term.[2]
The error term is comprised of two components,

$$\varepsilon_{it} = \mu_i + \nu_{it}, \tag{2}$$

where μ_i denotes the unobservable group specific effects and ν_{it} is the remainder stochastic disturbance. The component μ_i is time-invariant and will account for group specific effects not included in the right-hand-side (RHS). The remainder disturbance ν_{it} varies with groups and time and is assumed orthogonal to *X*, *Z* and μ with a mean of zero and a constant variance σ_ν^2.

Generally, two specifications of equation (1) are considered and differ based on their treatment of μ_i. First, 'fixed effects' (FE) treats μ_i as fixed but unknown constants differing across groups. This specification is easily estimated by including group dummy variables in the RHS (Least Squares Dummy Variable (LSDV) estimator). However, if *N* and/or *T* are large, LSDV suffers from the loss of precious degrees of freedom. Alternatively, estimates can be obtained by transforming the data into deviations from respective group means ('within' estimator). The two fixed effects estimation methods described reveal two crucial defects: (i) time-invariant variables are eliminated so γ cannot be estimated, and (ii) the estimator is not fully efficient because, in certain cases, it ignores variation across groups.

Second, 'random effects' (RE) assumes that the μ_i are random variables, distributed independently across groups with variance σ_μ^2. Estimates of this specification are based on transformations of the data into deviations from weighted respective group means where the weights are based on, generally, the estimated variances of the components in equation (2) and *T* (Feasible General Least Squares (FGLS) estimator). Specifically, the weight on the group means takes the form,

$$\hat{\theta} = 1 - \frac{\hat{\sigma}_\nu}{\sqrt{T\hat{\sigma}_\mu^2 + \hat{\sigma}_\nu^2}} \tag{3}$$

[2] Many presentations of this familiar model may include a scalar constant term, α.

Note, if $\hat{\theta} = 1$, random effects is the 'within' fixed effects estimator. Unbiased robust estimates of the variance components are best obtained from pooled ordinary least squares (OLS) and LSDV estimators. The potential correlation of μ_i with the variables in X_{it} and Z_i is a defect of the random effects construct. If these correlations are present, random effects estimation yields biased and inconsistent estimates of β. Conversely, by transforming the data into deviations from the simple group means the fixed effects estimator is not impacted by this lack of orthogonality.

Hausman (1978) outlines a specification test of the null hypothesis of orthogonality between μ_i and X_{it} , Z_i where $H_0 : E(\mu_i | X_{it}, Z_i) = 0$. By failing to reject the null, both fixed effects and random effects are unbiased and consistent, but fixed effects is less efficient. When the null is rejected, fixed effects is unbiased and consistent but random effects is not. Accordingly, if the null is not rejected the two estimates should not differ systematically. A likely test of the null should consider the difference between the two estimators, $\hat{g} = \hat{\beta}_{FE} - \hat{\beta}_{RE}$, within the sampling error. Hausman (1978) formally derives the chi-squared test statistic based on the Wald criterion,

$$\chi_K^2 = \hat{g}'[Var(\hat{g})]^{-1}\hat{g}\ , \tag{4}$$

where K degrees of freedom equals the number of estimated slope coefficients. The center positive definite matrix should be based on robust covariance estimates.

The random effects specification requires exogeneity of all regressors and the components in equation (2). Conversely, the fixed effects model allows for endogeneity of all the regressors and μ_i , but ignores observable variables Z_i. In order to avoid this all or nothing choice of exogeneity and accommodate the estimation of γ, Hausman and Taylor (1981) (HT) propose a third specification for estimating equation (1) where the RHS is split into two main categories of variables, those assumed uncorrelated (exogenous) with μ_i and v_{it}, and those correlated (endogenous) with μ_i, but not v_{it}. Table 1 shows the four possible sets of observable variables for equation (1).

Table 1: Hausman-Taylor variable sets

	Exogenous	**Endogenous**
Time varying	X_1 is n x k_1	X_2 is n x k_2
Time invariant	Z_1 is n x f_1	Z_2 is n x f_2

The exogenous category identified serves two functions, (i) using group mean deviations, unbiased estimates of the respective elements of β are produced, and (ii) the exogenous set and group means provide valid instruments for the unbiased and efficient estimation of β and γ. An advantage of panel data is the formulation of instruments from *within* the model construct. The order condition for identification requires that k_1 (the number of variables in X_1) is greater than or equal to f_2 (the number of variables in Z_2). When $k_1 > f_2$, the model is over-identified and HT is more efficient than 'within' fixed effects.

The appealing but complex HT estimator requires prior knowledge that certain RHS variables in equation (1) are uncorrelated with μ_i . The purpose of this paper is to outline a tractable variable pretest that facilitates the initial sorting of regressors as categorized in Table 1. The variable pretest proposed herein builds on the pretest estimator suggested by Baltagi et al (2003) by providing the necessary foundation for regressor identification. The balance of this examination is divided into four sections. Section 2 outlines the motivation and steps of the pretest. Section 3 provides an example estimation and interprets the empirical inference with conclusions and extension suggestions drawn in section 4.

2. Pretest

The Hausman (1978) specification test described above provides a natural starting point. Forms of the Hausman test are routinely used as pretests in applied work (see Guggenberger (2010) for an extensive review).[3] As a first step, estimate equation (1) with both fixed ('within') and random effects (FGLS) specifications and subsequently construct the chi-squared statistic given in equation (4).[4] This initial statistic becomes the base of comparison. If the Hausman null hypothesis is not rejected, FGLS is unbiased, efficient and commonly the correct specification. If the null is rejected, identifying the RHS variables that contribute to the size of the statistic estimated from equation (4) is the primary focus herein. Second, estimate succeeding chi-squared statistics from re-specified models by dropping one sequential regressor each iteration. For example, assume the RHS includes 3 observable variables - drop variable 1 and estimate the model with 2 and 3 - drop variable 2 and estimate the model with 1 and 3 - drop variable 3 and estimate the model with 1 and 2. A formal sorting tenet for the vector of resulting chi-squared statistics may, ostensibly, appear arbitrary and perhaps best illustrated by example.

[3] Cornwell et al (1992) suggests that remainder noise correlations should be tested as well where, $H_0: \quad E\left(v_{it} \middle| X_{it}, Z_i\right) = 0$.

[4] A computational note, while there is no miraculous software written for all panel data estimation and testing, packages like LIMDEP, SAS, STATA, and TSP are generally sufficient. Use robust variance/covariance corrections when needed.

3. Example

Data for this example was obtained from Kunce (2021). The data are a balanced panel of all 23 counties (N = 23) within the state of Wyoming (USA) spanning the years 2010 - 2020 (T = 11; n = 253). The cited analysis examines the contention that observable socioeconomic factors matter in explaining variation in age-adjusted mortality within the state of Wyoming. Covariates include, percent of a county's population classified as non-white, percent with a bachelor's degree or higher, percent whose income is below the national poverty level, median income in 1,000s of 2020 dollars, a county's unemployment rate, the number of licensed hospital beds in the county (time-invariant) and percent of the population with no health insurance. In order to identify Table 1 variable sets, iterative one-way fixed and random effects regressions were performed varying 7 sets of variables by dropping the regressor indicated in the first column of Table 2 with the resulting Hausman statistic in the second column. For example, the fifth row depicts the resulting test statistic when the UNEMPLOYMENT variable is dropped from the right-hand-side. Note that the Hausman test statistic reduces to 18.88 from 33.56. The UNEMPLOYMENT variable appears to be a significant 'correlation contributor' therefore pretests as likely endogenous. The BEDS variable cannot be estimated with the FE specification and its inclusion in the pretest routine affects the RE coefficient estimates and variance/covariance estimates. Accordingly, the BEDS variable pretests to be a correlation contributor.

Table 2: $\boldsymbol{\mu_i}$ correlation tests*

	χ^2_6
NONWHITE	32.05
EDUCATION	20.94
POVERTY	28.37
INCOME	23.38
UNEMPLOYMENT	18.88
BEDS	25.84
UNINSURED	29.25

*All RHS variables χ^2_7 = 33.56, base of comparison.

Recall that the necessary condition for identification and efficient estimation of β and γ is that $k_1 > f_2$ and f_2 may be empty (Hausman and Taylor (1981) pp. 1385-1387). Thus, a natural sorting from Table 2 follows,

X_1: NONWHITE, POVERTY, UNINSURED
X_2: EDUCATION, INCOME, UNEMPLOYMENT
Z_1: SCALAR CONSTANT
Z_2: BEDS

Table 3 shows the results of three error component models for the Kunce (2021) example. The first column depicts the RE estimates which assume no correlation between the RHS and the μ_i . The sizeable LM statistic confirms the importance of controlling for specific county level effects. The second column of Table 3 presents the 'within' FE estimates. The F-test, null hypothesis of county homogeneity is rejected at the < 1% level. The time-invariant BEDS variable is eliminated by the county mean differencing data transformation. Comparing the FE and RE estimates using the Hausman test rejects the null hypothesis of orthogonality confirming that the RE model is misspecified. This initial Hausman test outcome justifies the use of the HT instrumental variable method. Given the variable pretest results from above, the last column of Table 3 shows the estimates using the HT routine in *LIMDEP 11*®. Interestingly, the coefficient of BEDS is twice the coefficient estimated using RE (-0.01 vs -0.005). A Hausman test based on the difference between FE and the HT estimator fails to reject the null hypothesis of orthogonality. There are two degrees of freedom in this chi-squared test since there are two over-identifying conditions (the number of X_1 variables minus the number of Z_2 variables, see Baltagi et al (2003)). The variable pretest herein is shown to be valid, we cannot reject that the set of instruments X_1 and Z_1 are appropriate. Other combinations of variable sorting are certainly possible, but subsequent results do not improve upon those found in Table 3.

Table 3: Example estimation results

	RE	**FE**	**HT**
NONWHITE (t)	0.05 (1.43)	0.26 (1.53)	0.18 (1.22)
EDUCATION (t)	-0.07 (-4.63)	-0.03 (-0.61)	-0.03 (-0.76)
POVERTY (t)	0.03 (0.78)	0.06 (1.45)	0.07 (1.44)
INCOME (t)	-0.01 (-0.68)	-0.04 (-1.65)	-0.04 (-1.69)
UNEMPLOYMENT (t)	0.16 (2.00)	0.23 (2.74)	0.21 (2.62)
UNINSURED (t)	-0.02 (-0.76)	0.05 (1.26)	0.04 (1.10)
BEDS (t)	-0.005 (-2.03)	-	-0.01 (-1.68)
LM test (p value)	69.35 (0.00)	-	-
Hausman test (p value)	33.56 (0.00)	-	2.47 (0.29)
Homogeneity F(22, 223)(p value)	-	5.35 (0.00)	-

4. Conclusion

In this paper, we have developed a tractable pretest method (for use with panel data) which treats the problem of identifying explanatory variables that are likely correlated with latent group effects. The proposed variable pretest builds on the work of Baltagi et al (2003) by outlining a foundational sorting mechanism as discussed in Hausman and Taylor (1981). Extending the variable pretest to the two-way error components specification continues to be based on the difference between

FE (with both group and time effects) and the two-way FGLS estimator. Kang (1985) derives the conditions to be considered regarding the relevant Hausman testable hypothesis. Wyhowski (1994) extends the HT specification to include latent time effects.[5]

References

[1] Hausman, J. (1978). Specification tests in econometrics. Econometrica 46(6), pp. 1251-1271.

[2] Hausman, J. and Taylor, W. (1981). Panel data and unobservable individual effects. Econometrica 49(6), pp. 1377-1398.

[3] Baltagi, B., Bresson, G. and Pirotte, A. (2003). Fixed effects, random effects or Hausman-Taylor? A pretest estimator. Economics Letters, 79(3), pp. 361-369.

[4] Guggenberger, P. (2010). The impact of a Hausman pretest on the size of a hypothesis test. Econometric Theory 26(2), pp. 369-382.

[5] Kunce, M. (2021). Socioeconomic effects on all-cause mortality: Evidence from the American rural west. International Journal of Social Science and Economic Research 6(10), pp. 4172-4187.

[6] Cornwell, C., Schmidt, P. and Wyhowski, D. (1992). Simultaneous equations and panel data. Journal of Econometrics 51, pp. 151-181.

[7] Kang, S. (1985). A note on the equivalence of specification tests in the two-factor multivariate variance components model. Journal of Econometrics 28, pp. 193-203.

[8] Greene, W. (2018). Econometric Analysis. Eighth Edition. Pearson Education Inc. New York, NY.

[9] Wyhowski, D. (1994). Estimation of a panel data model in the presence of correlation between regressors and a two-way error component. Econometric Theory 10, pp. 130 - 139.

[5] Depending on the econometric software used, augmentations may be required to manage a two-way HT estimator (see Greene (2018) pp. 427-445 and Wyhowski (1994)).