A Step-by-Step Guide to Conducting Robust Hausman Tests for Random Effects vs. Fixed Effects Models Using Unbalanced Panel Data in Stata

By Zachariah Rutledge*

Working Draft: Please do not cite without permission from author.

Note: Supplementary Online Resources Can Be Found at:
https://www.zachrutledge.com/educational-resources.html

Introduction and Background

Assume that you want to estimate the following model:

$$y_{it} = \mathbf{x_{it}}\boldsymbol{\beta} + c_i + u_{it}, \tag{1}$$

where i denotes the panel variable, t denotes the time variable, $\mathbf{x}_{it}$ contains at least one variable of interest but may also include control variables, and c_i denotes the unobserved individual effects, but you are not sure whether you should use a random effects or fixed-effects (within) estimator. According to Wooldridge (2002), if the following two-part assumption (which implies that the error structure is homoskedastic, serially uncorrelated and that the unobserved individual effects c_i are homoskedastic) holds:[1]

$$E(u_i u_i' | x_i, c_i) = \sigma_u^2 I_T \tag{a}$$
$$E(c_i^2 | x_i) = \sigma_c^2 \tag{b}$$

where I_T is the identity matrix of size TxT, then the standard Hausman test built into Stata can be used.[2] However, if either (a) or (b) are violated, then the standard Hausman test is no longer valid; instead, a robust version of the Hausman test must be used.

Definitions

The random effects estimator is defined as follows:

$$\check{y}_{it} = \check{\mathbf{x}}_{it}\boldsymbol{\beta} + \check{v}_{it} \tag{2}$$

where $\check{y}_{it} = y_{it} - \lambda_i \bar{y}_i$, which is referred to as the quasi-demeaned variable y (and similarly for the vector $\mathbf{x}$ and the composite error $v_{it} = c_i + u_{it}$). λ_i is the panel-group-specific GLS weight used in the random effects

*Ph.D. Candidate, Agricultural and Resource Economics, University of California, Davis. Contact: zjrutledge@ucdavis.edu
[1]This assumption is referred to as Assumption RE.3 in Wooldridge (2002).
[2]See https://www.stata.com/manuals13/rhausman.pdf for details about how to conduct a standard Hausman test in Stata.

estimator defined by $\lambda_i = 1 - [\frac{\sigma_u^2}{\sigma_u^2 + T_i\sigma_c^2}]^{1/2}$ where T_i is number of time periods that panel group i appears in the data such that the random effects model can be written as:

$$y_{it} - \lambda_i \bar{y_i} = (\mathbf{x}_{it} - \lambda_i \bar{\mathbf{x}}_i)\boldsymbol{\beta} + (v_{it} - \lambda_i \bar{v_i}), \tag{3}$$

and $\bar{y_i} = \frac{\sum_{i=1}^{T_i} y_i}{T_i}$ (and similarly for $\bar{\mathbf{x}}_i$ and $\bar{v}_i$).[3] Borrowing from Wooldridge (2002), let $\mathbf{w}_{it}$ denote a subset (or the entire set) of the time varying elements of $\mathbf{x}_{it}$ from (1) (excluding any time dummy variables) that you want to run the robust Hausman test on. Let $\ddot{\mathbf{w}}_{it}$ denote the time-demeaned (or within-transformed) variables such that if $\ddot{w}_{it} \in \ddot{\mathbf{w}}_{it}$, then $\ddot{w}_{it} = w_{it} - \bar{w}_i$.

The Robust Hausman Test

Wooldridge (2002) explains that the easiest way to conduct the robust Hausman test is to conduct a Wald test on the null hypothesis that $H_0 : \boldsymbol{\xi} = 0$ using the following regression:[4]

$$\check{y}_{it} = \check{\mathbf{x}}_{it}\boldsymbol{\beta} + \ddot{\mathbf{w}}_{it}\boldsymbol{\xi} + error_{it} \tag{4}$$

where $\mathbf{w}_{it} \in \mathbf{x}_{it}$ represents your main variable(s) of interest from (1).[5]

Conducting the Test in Stata with an Unbalanced Panel

In order to conduct the robust Hausman test in Stata with an unbalanced panel using (4), the following steps may be used. You can click on this link to find an annotated do file and sample data to follow my example step by step.[6]

1. Run the random effects model with robust standard errors in Stata using the "theta" option then collect some of the scalars that are stored in the "e" matrix
2. Generate σ_u^2 and σ_e^2 by retrieving scalars e(sigma_u) and e(sigma_e) from Stata's e matrix. [7]
3. Generate a variable that assigns a value for T_i to each observation in each panel group
4. Generate a variable that assigns a value for λ_i to each observation in each panel group
5. Manually construct the arithmetic mean for the y and $\mathbf{x}$ variables by generating $\bar{y}_i$ and $\bar{\mathbf{x}}_i$
   (a) This is relatively straighforward for continuous variables, such as the dependent variable and main regressor in my online example, but for sets of dummy variables (such as what you might typically

[3] Identifying T_i for each panel group is the critical difference between conducting the Hausman test with balanced and unbalanced panels. With a balanced panel, $T_i = T \ \forall \ i$, which requires fewer steps to conduct the robust Hausman test in Stata. In fact, Cameron and Trivedi (2010) provide sample code to conduct the robust Hausman test with a balanced panel and explain that "The code becomes more complex in the unbalanced case, because we need to compute $[\lambda_i]$ for each observation." They also suggest using the xtoverid command to conduct the robust Hausman test after running the xtreg, re command; however, the xtoverid program does not work in certain settings when the "i." command is used to include factor variables as regressors. In applications where the researcher wants to make use of a series of dummy variables defined by the categories within a factor variable, especially when multiple factor variables are being used or if the analysis is being conducted on a subset of the categories in one or more of the factor variables, the method I outline in this paper may be less cumbersome and/or time consuming than the procedure that would be necessary to "fix" the xtoverid bug. My contributions are two-fold: (i) I extend Cameron and Trivedi's Stata example into the unbalanced panel context and (ii) I demonstrate how to conduct the robust Hausman test in Stata when using two-stage least squares regressions, which to the best of my knowledge has not been explained in the literature to date (part (ii) coming soon).

[4] This null hypothesis is equivalent to the null hypothesis that the random effects model is the appropriate model to use.

[5] Note that this procedure permits the test to be conducted on more than one variable of interest.

[6] The do file is titled "Robust Hausman Test Do File" and the Stata data file is titled "Robust Hausman Test Sample Data." These files can be downloaded from https://www.zachrutledge.com/educational-resources.html.

[7] Note: The notation from Wooldridge (2002) differs from that used by Stata (see https://www.stata.com/manuals13/xtxtreg.pdf). The scalar "sigma_u" retrieved from the e matrix in Stata represents σ_c in Wooldridge (2002), and the scalar "sigma_e" from Stata represents σ_u in Wooldridge (2002).

use with the "i." command), the process is more involved. A set of code must be run separately for each category contained in the categorical variable. This can be done by identifying the codes that identify each category and running a loop using these codes. For regressions with many factor variables, it is relatively simple to construct these loops by using the "tab" command in Stata to identify the codes for each of factor variable, highlighting the corresponding table that is generated from the "tab" command, copying the table with the "copy as table" option from Stata, pasting the table into Excel as "text," deleting any unnecessary information in excel, appending the table in Excel with code that will run in Stata, then copying it from Excel and pasting it into a do file.

6. Manually time-demean the y and $\mathbf{x}$ variables by generating $\ddot{y}_{it}$ and $\ddot{x}_{it}$

   (a) See note 5a

7. Manually quasi-demean the y the $\mathbf{x}$ variables by generating $\check{y}_{it}$ and $\check{\mathbf{x}}_{it}$

   (a) See Note 5a

8. Generate a quasi-demeaned constant (1 - λ_i)

9. Check to make sure you have properly time-demeaned your variables. You can do this by running the model using the "xtreg" command with the "fe" option and comparing the coefficient on your main regressor of interest with the coefficient from a pooled OLS model that is ran with the manually constructed time-demeaned variables that were generated in steps 5-7 (see online code for example). Your coefficients should be identical (or extremely close due to rounding differences).

10. Check to make sure you have properly quasi-demeaned your variables. You can do this running the model using the "xtreg" command with the "re" option and comparing the coefficient on your main regressor of interest with the coefficient from a pooled OLS model ran with the manually constructed quasi-demeaned variables that were generated in steps 5-7 (see online code for example). Your coefficients should be identical (or extremely close due to rounding differences).

11. Run equation (4) using pooled OLS regression with robust (or cluster-robust errors) standard errors while including the quasi-demeaned constant as a right-hand-side variable and specifying the "nocons" option

12. Run a Wald test on $\boldsymbol{\xi}$ using the "test" command in Stata

References

Cameron, A. C. and Trivedi, P. K. (2010). *Microeconometrics Using Stata*. Stata Press.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.