Applied Panel Data Analysis – Lecture 5

Christopher F. Parmeter

AGRODEP
September 9-13th, 2013
Dakar, Senegal

- Discuss tests of poolability and correlation between covariates and unobserved heterogeneity
- Both of these tests have important implications for how we think about the linear panel data model

- Having estimated the pooled panel data model, it is natural to question the existence of unobserved heterogeneity
- Having estimated the unobserved effects model under both the fixed and random effects framework, a natural question is which one is more appropriate for the data at hand?

- We will discuss the intuition for both of these tests
- Specific attention will be given to the fixed vs. random effects estimation frameworks
- Learn how to use Monte Carlo simulations to understand how these tests work

- Testing for poolability naturally arises when considering the unobserved effects model
- Now, there are a few conceptual issue to think about prior to testing for poolability
  - Is the appropriate unobserved effects model in the fixed or random effects framework?
  - Do only the intercepts vary or can other coefficients vary across individuals?

- Lets look at the simplest pooling setup, the fixed effects framework for the unobserved effects model
- Further, we will assume that only the intercepts vary across individuals
- In this case our null hypothesis is

$$H_0 : c_1 = c_2 = \cdots = c_N = 0 \tag{1}$$

- This is simply at $F - test$ in the mold of Chow (1961)
- Our test statistic in this case is

$$F = \frac{RSS_R - RSS_{UR}}{RSS_{UR}} \cdot \frac{N(T-1) - K}{N-1} \tag{2}$$

  which is distributed asymptotically as $F_{N_1, N(T-1)-K}$
- RSS_R would be the residual sum of squares from the pooled OLS model while RSS_{UR} would be the residual sum of squares from the fixed effects framework

- Failure to reject H_0 does not imply that the pooled OLS model is appropriate
- Keep in mind that there could be underlying sources of misspecification that contribute to what you learn from this test
- Perhaps the model is misspecified functional, perhaps the coefficients on some of the individual-time varying covariates differ across individuals, perhaps the random effects framework is appropriate, etc.

- We will ignore functional misspecification for the time being and focus on a model where all of the coefficients are allowed to vary across individuals
- Consider the unobserved effects model, but where β can now vary across individuals

$$y_{it} = x_{it}{\prime}\beta_i + c_i + \varepsilon_{it} \tag{3}$$

- For this setup we act as though we treat each cross section independently from all other cross sections

- Our hypothesis of interest is $H_0 : c_i = c, \beta_i = \beta \ \forall i$ so that our restricted model becomes the pooled model
- Keep in mind that T must be larger than K for this test to be implementable
- For many micro panels this will not be feasible as T may be on the order of 5-10 while K may be on the order of 15-30
- However, the discussion here can easily be modified to allow a subset of β to vary across individuals

- An important implicit assumption when testing H_0 under the fixed effects framework is that $\varepsilon \sim IID(0, \sigma_\varepsilon I_{NT})$
- If heteroskedasticity or autocorrelation is present (as in the random effects framework) then a robust test statistic will be needed
- If the constant variance assumption fails then the test for poolability will be grossly misleading

- How do we implement the test?
- Lets introduce some matrix notation
- For individual i our unrestricted regression model is

$$y_i = Z_i \delta_i + \varepsilon_i \tag{4}$$

where $y_i' = (y_{i1}, y_{i2}, \ldots, y_{iT})$, $Z_i = [\imath_T, X_i]$ and $\varepsilon_i' = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT})$

- Note that X_i is $T \times K$, δ_i' is $1 \times (K+1)$ while both y_i and ε_i are $T \times 1$

- With this unrestricted model our null hypothesis can be written as $H_0 : \delta_1 = \delta_2 = \cdots = \delta_N = \delta$
- Our restricted model is

$$y = Z\delta + \varepsilon \tag{5}$$

- Here $Z' = (Z'_1, Z'_2, \ldots, Z'_N)$ and $\varepsilon' = (\varepsilon'_1, \varepsilon'_2, \ldots, \varepsilon'_N)$

- Estimation of the restricted model in (5) produces the estimator $\hat{\delta}_{OLS} = (Z'Z)^{-1} Z'y$
- Estimation of each of the unrestricted models in (4) produces the estimator $\hat{\delta}_{i,OLS} = (Z_i'Z_i)^{-1} Z_i'y_i$
- Define $\hat{\varepsilon}_i^* = y_i - Z_i\hat{\delta}_{i,OLS}$ and $\hat{\varepsilon} = y - Z\hat{\delta}_{OLS}$
- Further, let $\hat{\varepsilon}^{*\prime} = (\varepsilon_1^{*\prime}, \varepsilon_2^{*\prime}, \ldots, \varepsilon_N^{*\prime})$

- The test statistic for the null hypothesis of poolability of the data is

$$F = \frac{\hat{\varepsilon}'\hat{\varepsilon} - \hat{\varepsilon}^{*\prime}\hat{\varepsilon}^{*}}{\hat{\varepsilon}^{*\prime}\hat{\varepsilon}^{*}} \frac{N(T-1-K)}{(N-1)(K+1)} \tag{6}$$

- $\hat{\varepsilon}'\hat{\varepsilon}$ is exactly RSS_R and $\hat{\varepsilon}^{*\prime}\hat{\varepsilon}^{*}$ is exactly RSS_{UR}, the only difference with the earlier setting is that now we allow more parameters to vary across individuals
- This is exactly a Chow test but for N groups instead of the common 2 groups that is routinely used in applied work

- If heteroskedasticity or autocorrelation was perceived to be relevant for the modeling situation then a robust form of the test statistic in (6) will be needed
- If instead of assuming that $\varepsilon \sim IID(0, \sigma_\varepsilon I_{NT})$ we assume $\varepsilon \sim D(0, \Sigma)$, then we would need to deploy FGLS when we estimated both the unrestricted and restricted models
- Under GLS, our restricted model is $\Sigma^{-1/2} y = \Sigma^{-1/2} Z\delta + \Sigma^{-1/2}\varepsilon$

- Under GLS our unrestricted model for individual i is $\Sigma^{-1/2}y = \Sigma^{-1/2}Z^*\delta^* + \Sigma^{-1/2}\varepsilon$ where

$$Z^* = \begin{bmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_N \end{bmatrix}$$

and $\delta^{*\prime} = (\delta_1', \delta_2', \ldots, \delta_N')$

- Use $\dot{\hat{\varepsilon}}$ to denote the residuals from the restricted model
- Use $\dot{\hat{\varepsilon}}^*$ to denote the residuals from the unrestricted model
- Then our robust F-statistic is

$$F = \frac{\dot{\hat{\varepsilon}}'\dot{\hat{\varepsilon}} - \dot{\hat{\varepsilon}}^{*\prime}\dot{\hat{\varepsilon}}^*}{\dot{\hat{\varepsilon}}^{*\prime}\dot{\hat{\varepsilon}}^*}\frac{N(T-1-K)}{(N-1)(K+1)} \tag{7}$$

- Imposing the restriction $\delta_i = \delta$, regardless if it is true or not, will reduce the variance of the pooled OLS estimator at the expense of introducing bias (due to misspecification)
- Toro-Vizcarrondo and Wallace (1968) noted "if one is willing to accept some bias in trade for a reduction in variance, . . . one might prefer the restricted estimator."
- See Baltagi (1995) for simulations results that explore the gains and losses in mean squared error relating to the pooling issue

- Testing for poolability in the random effects framework is different than the fixed effects framework
- Here the test is not about differences in parameters across individuals, but the presence of serial correlation
- Recall that in the random effects framework that $corr(\varepsilon_{it}, \varepsilon_{is}) = \sigma_c^2/(\sigma_c^2 + \sigma_\varepsilon^2)$ for $t \neq s$
- Thus, if we are in the random effects framework and we are testing the suitability of the pooled panel data model, we need to concern ourselves with serial correlation in the error term

- If the assumptions for the random effects framework to be valid are true but there does not exist an unobserved effect then pooled OLS is efficient
- Statistically, this is equivalent to testing $H_0 : \sigma_c^2 = 0$
- We can effectively ignore the panel structure of the data when testing this assumption given that under the null hypothesis the data can be treated as a pooled cross section
- A common test statistic for AR(1) serial correlation (in the time series setting) is to regress the residuals, $\hat{\varepsilon}_t$ on the lagged residuals $\hat{\varepsilon}_{t-1}$ and perform a standard t-test on the coefficient on the lagged residuals
- We can follow the same procedure here, except we must account for both individuals and time

- Formally, our test statistic can be formed by scaling the estimate of the variance of the unobserved effects

$$N^{-1/2}\sum_{i=1}^{N}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\hat{\varepsilon}_{it}\hat{\varepsilon}_{is} \tag{8}$$

- We only scale by N here instead of NT since our working assumption is that $N \longrightarrow \infty$ while T is fixed
- Regardless of the distribution of ε, this scaled variance estimator has a limiting normal distribution with variance

$$E\left[\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\hat{\varepsilon}_{it}\hat{\varepsilon}_{is}\right]^2 \tag{9}$$

- This suggests the test statistic

$$\frac{\sum_{i=1}^{N}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\hat{\varepsilon}_{it}\hat{\varepsilon}_{is}}{\left[\sum_{i=1}^{N}\left(\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\hat{\varepsilon}_{it}\hat{\varepsilon}_{is}\right)^{2}\right]^{1/2}} \tag{10}$$

- This test statistic has an asymptotically normal distribution
- Note that you will always reject the null hypothesis when the variance estimate of σ_c^2 is negative because the alternative hypothesis is one-sided; it makes no sense in this setting to have a two-sided hypothesis

- When implementing this test in practice it will not be uncommon to reject H_0
- The reason for this is that under the random effects framework, x_{it} is prevented from containing lagged dependent variables and this test detects myriad forms of serial correlation, not just the presence of the random effect
- Thus a rejection of the null should not be taken to mean the random effects framework is correct
- A more appropriate test for serial correlation should be based purely on the error term, ignoring the unobserved effect, however, this is not a test of the presence of the unobserved effect so we discuss this later

- The key distinction between the fixed and random effects frameworks is the working assumption of $E[c_i|x_{it}] = E[c_i]$ for all t
- It is important to test this assumption to ensure that the proper framework is exploited during estimation of the unobserved effects model
- Hausman (1978) proposed a test based upon a weighted squared difference statistic

- The Hausman test we will develop here will be used specifically for testing between the fixed and random effects framework
- However, this style of test works in other settings and we will discuss this intuition as well

- The null hypothesis is that the random effects specification is appropriate
- Notice that the fixed effects estimator is consistent when c_i and x_{it} are correlated, which includes 0
- The random effects estimator is consistent only when c_i and x_{it} are uncorrelated
- Thus, under H_0 : both estimators are consistent; this is the key piece of information that Hausman (1978) exploited to construct a test between these two frameworks

- A rejection of the null hypothesis is taken as evidence against the assumption that $E[c_i|x_{it}] = E[c_i]$
- The test still requires that $E[\varepsilon_{it}|x_{i1}, x_{i2}, \ldots, x_{iT}] = E[\varepsilon_{it}] = 0$, strict exogeneity of the covariates
- If this assumption fails then the estimators for the fixed and random effects frameworks are inconsistent and the test is uninformative

- Caveat Emptor: we must be careful to consider which coefficient estimates we can compare
- The fixed effects framework only allows identification of time-varying explanatory variables whereas the random effects framework allows identification of time-constant explanatory variables
- Further, we cannot compare coefficients on pure time effects either!
- The comparison is on variables that vary at both the individual and time level

- Hausman's test between the fixed and random effects framework is based on the difference in the estimates across the two frameworks
- Under the random effects framework our estimator is $\hat{\beta}_{GLS}$ while under the fixed effects framework our estimator is $\tilde{\beta}$
- Consider $\hat{q}_1 = \hat{\beta}_{GLS} - \tilde{\beta}$
- The Hausman test has the equivalent null hypothesis $H_0 : q_1 = 0$

- To build an appropriate test statistic we will need to derive the variance-covariance of $\hat{q}_1$
- Recall that $\hat{\beta}_{GLS} = \beta + \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}\varepsilon$ and $\tilde{\beta} = \beta + (X'QX)^{-1} X'Q\varepsilon$
- We have

$$\begin{aligned}
E[\hat{q}_1] =& E\left[\left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}\varepsilon - \left(X'QX\right)^{-1} X'Q\varepsilon\right] \\
=& E\left[\left(\left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1} - \left(X'QX\right)^{-1} X'Q\right)\varepsilon\right] \\
=& E\left[E\left[\left(\left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1} - \left(X'QX\right)^{-1} X'Q\right)\varepsilon|X\right]\right] \\
=& E\left[\left(\left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1} - \left(X'QX\right)^{-1} X'Q\right)E\left[\varepsilon|X\right]\right] \\
=& 0
\end{aligned}$$

- The variance-covariance matrix of $\hat{q}_1$ is

$$\begin{aligned} Var(\hat{q}_1) =& Var(\hat{\beta}_{GLS}) + Var(\tilde{\beta}) - 2Cov(\hat{\beta}_{GLS}, \tilde{\beta}) \\ =& \left(X'\Omega^{-1}X\right)^{-1} + \sigma_{\varepsilon}^2 \left(X'QX\right)^{-1} - 2Cov(\hat{\beta}_{GLS}, \tilde{\beta}) \end{aligned}$$

- We have that

$$\begin{aligned} Cov(\hat{\beta}_{GLS}, \tilde{\beta}) =& E\left[\left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}\varepsilon\varepsilon' QX \left(X'QX\right)^{-1}\right] \\ =& \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1} E\left[\varepsilon\varepsilon'|X\right] QX \left(X'QX\right)^{-1} \\ =& \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}\Omega QX \left(X'QX\right)^{-1} \\ =& \left(X'\Omega^{-1}X\right)^{-1} X'QX \left(X'QX\right)^{-1} \\ =& \left(X'\Omega^{-1}X\right)^{-1} = Var(\hat{\beta}_{GLS}) \end{aligned}$$

- Using all of these pieces we have

$$\begin{aligned} Var(\hat{q}_1) &= \left(X'\Omega^{-1}X\right)^{-1} + \sigma_\varepsilon^2\left(X'QX\right)^{-1} - 2Cov(\hat{\beta}_{GLS}, \tilde{\beta}) \\ &= \left(X'\Omega^{-1}X\right)^{-1} + \sigma_\varepsilon^2\left(X'QX\right)^{-1} - 2\left(X'\Omega^{-1}X\right)^{-1} \\ &= \sigma_\varepsilon^2\left(X'QX\right)^{-1} - \left(X'\Omega^{-1}X\right)^{-1} \end{aligned} \quad (11)$$

- The Hausman test statistic is given by

$$H = \hat{q}_1'\left(\widehat{Var(\hat{q}_1)}\right)^{-1}\hat{q}_1 \quad (12)$$

- Hausman (1978) showed that $H \sim \chi^2_K$

- Given that Ω contains σ_ε^2 it is necessary to use the same estimate of σ_ε^2 that appears in both Ω and the variance-covariance matrix of $\tilde{\beta}$
- The reason for this is that if one uses different estimates for σ_ε^2, there is no guarantee that $\widehat{Var(\hat{q}_1)}$ will be positive definite, resulting in a negative test statistic

- The Hausman test can be recast as an omitted variable test
- The intuition here is that the unobserved effect is not adequately captured in the random effects framework and so its inclusion leads to more variation in y being explained
- Consider the following regression

$$\check{y} = \check{X}\beta + QX\delta + \omega, \tag{13}$$

where the $\check{z}$ contains elements $\check{z}_{it} = z_{it} - \theta\bar{z}_{i.}$ and ω is an IID error term

- The Hausman test is equivalent to testing $H_0 : \delta = 0$

- Why does this work this way?
- If $\delta = 0$ then we have the regression of $\check{y}$ on $\check{X}$ which will produce the GLS estimator under the random effects framework as we discussed in Lecture 4
- QX has typical elements $\tilde{x}_{it} = x_{it} - \bar{x}_{i\cdot}$
- δ captures the further impact of time-demeaning that may be missed in the random effects framework
- To see this note that

$$\tilde{x}_{it} = \check{x}_{it} - (1 - \theta)\bar{x}_{i\cdot} \tag{14}$$

- Thus, the inclusion of both $\check{x}$ and $\tilde{x}$ captures all aspects of time demeaning

- What does the test statistic look like from this regression?
- The Hausman test statistic from regression (13) is

$$H = \hat{\delta}' \left[\widehat{Var(\hat{\delta})} \right]^{-1} \hat{\delta} \quad (15)$$

- An interesting aspect of (13) is that one can show using either partitioned regression or the Frisch-Waugh-Lovell theorem that

$$\hat{\delta} = \tilde{\beta} - \hat{\beta}_{Between} \quad (16)$$

- It also can be shown that

$$Var(\hat{\delta}) = Var(\tilde{\beta}) + Var(\hat{\beta}_{Between}) \tag{17}$$

- It may not be apparent but the Hausman statistics in (12) and (15) are identical
- Recall from Lecture 4 that we decomposed the random effects estimator as

$$\hat{\beta}_{GLS} = W_1 \tilde{\beta} + (I - W_1)\hat{\beta}_{Between} \tag{18}$$

- Thus, $\hat{q}_1 = \hat{\beta}_{GLS} - \tilde{\beta} = (I - W_1)\left(\tilde{\beta} - \hat{\beta}_{Between}\right) = \Gamma\hat{\delta}$, where Γ is invertible

- This yields the following equivalence:

$$H = \hat{q}_1' \left(\widehat{Var(\hat{q}_1)} \right)^{-1} \hat{q}_1 = \hat{\delta}'\Gamma' \left(\Gamma' \widehat{Var(\hat{\delta})} \Gamma \right)^{-1} \Gamma\hat{\delta}$$

$$= \hat{\delta}' \left(\widehat{Var(\hat{\delta})} \right)^{-1} \hat{\delta} = H$$

- The appeal of using the specification in (15) is that it is easier to construct a Hausman test that is robust to unspecified heteroskedasticity and serial correlation
- Recall that the Hausman test as derived in (15) has the working assumption that the error terms, ε are IID, if this assumption fails then the Hausman test is no longer valid

- Failing to reject the Hausman test should not directly be taken to imply that the random effects framework is appropriate for the unobserved effects model
- Baltagi and Griffin (1983) show that omitted dynamics can impact the Hausman test
- Amini, Delgado, Henderson and Parmeter (2012) provide detailed evidence that neglected nonlinearities can cause size distortions in the Hausman test

- Guggenberger (2010) develops a pre-test bias theory for the use of the Hausman test prior to standard specification test
- That is, it is common to estimate the unobserved effects model under both the fixed and random effects framework and then perform a Hausman test to determine which is more appropriate
- From there standard empirical practice is carried out: model evaluation, specification testing, significance testing, etc.
- Guggenberger's results suggest this approach will be misleading
- The Hausman test (even when H_0 : is correct) will suggest the wrong model sometimes (based on the size of the model)
- His suggestion is to base standard statistical significance off of the fixed effects framework given that this model is consistent under both frameworks

- Two tests unique to panel data are the test of poolability and the Hausman test
- The poolability test is a special version of the Chow test
- The Hausman test is a test for neglected time constant impacts
- Failure to reject either hypothesis needs to be considered carefully against the underlying assumptions of the unobserved effects model