Applied Panel Data Analysis – Lecture 7

Christopher F. Parmeter

AGRODEP
March 10-14th, 2015
Addis Ababa, Ethiopia

- Learn about the consequences of heteroskedasticity and serial correlation in the error term
- Discuss construction of variance-covariance matrices, as well as appropriate bootstrapping techniques
- Discuss recent work on general clustering in error processes

- With microdata it is likely that heteroskedasticity is present, either from clustering or impacts from the covariates in the model
- With long panels it is conceivable that serial correlation is present
- Need to account for these impacts to conduct robust inference

- When we think of the fixed effects framework for the unobserved effects model heteroskedasticity manifests itself through ε_{it}
- Can think of the Arellano (1987) approach to construct robust variance-covariance matrix discussed in Lecture 3
- However, if clustering is present then a more efficient construction of the variance-covariance matrix can be built

- In the fixed effects framework we can think of clustering as both a mean and variance effect from the unobserved effects
- That is, c_i or d_t impact both $E[y_{it}|x_{it}, c_i, d_t]$ and $Var(y_{it}|x_{it}, c_i, d_t)$
- In this case the heteroskedasticity is unchecked because we can have heteroskedasticity in the ith dimension, in the tth dimension and the it dimension
- Alternatively, consider clusters, which are (sub)sets of our panel dimenions
- If i represents an individual, a cluster could be the gender of the person or the race of the person
- If t represents a time period, a cluster could be years in which there is a drought
- Here the clusters are potentially smaller groupings that the full panel dimensions

- Recently Cameron, Gelbach and Miller (2011) proposed a robust variance-covariance estimator for two-way clustering; see also Thompson (2011)
- To help understand this estimator lets first think about the impact of clustering in a one-way setting
- Recall (15) from Lecture 3

$$Var(\tilde{\beta}) = (\tilde{X}'\tilde{X})^{-1}\left[\sum_{i=1}^{N}\tilde{X}_i'\hat{\tilde{\varepsilon}}_i\hat{\tilde{\varepsilon}}_i'\tilde{X}_i\right](\tilde{X}'\tilde{X})^{-1} \qquad (1)$$

- Here clustering is by the individual

- Now suppose we have clustering based on $g \in \{1, 2, \ldots, G\}$ groups
- Cameron, Gelbach and Miller (2011, eq 2.4) provide the one-way cluster robust variance-covariance matrix estimator

$$Var(\tilde{\beta}) = (\tilde{X}'\tilde{X})^{-1} \left[\sum_{g=1}^{G} \tilde{X}_g' \hat{\tilde{\varepsilon}}_g \hat{\tilde{\varepsilon}}_g' \tilde{X}_g \right] (\tilde{X}'\tilde{X})^{-1} \qquad (2)$$

- If $G = N$ then this is exactly the variance-covariance matrix for the random effects framework (clustering at the individual level)
- We can gain some intuition for the two-way cluster robust framework by thinking a bit more about the one-way setup

- Rewrite (2) as

$$Var(\tilde{\beta}) = (\tilde{X}'\tilde{X})^{-1}\widehat{B}(\tilde{X}'\tilde{X})^{-1} \tag{3}$$

  where $\widehat{B} = \tilde{X}'\left(\hat{\tilde{\varepsilon}}\hat{\tilde{\varepsilon}}' \odot S^G\right)\tilde{X}$

- Here $\odot$ represents Hadamard matrix multiplication (element-wise) and S^G is an $NT \times NT$ matrix with (i,j)th element equal to $1\{i,j \in$ same cluster$\}$

- Now lets assume there are two dimensions to grouping, $g \in \{1, 2, \ldots, G\}$ and $h \in \{1, 2, \ldots, H\}$
- Here we have

$$Var(\tilde{\beta}) = (\tilde{X}'\tilde{X})^{-1}\widehat{B}(\tilde{X}'\tilde{X})^{-1} \tag{4}$$

where $\widehat{B} = \tilde{X}' \left(\hat{\tilde{\varepsilon}}\hat{\tilde{\varepsilon}}' \odot S^{GH} \right) \tilde{X}$

- In the two-way clustering setup S^{GH} is an $NT \times NT$ matrix with (i, j)th element equal to $1\{i, j \in$ share at least one cluster$\}$

- An insightful decomposition of B in the two way setting is possible
- Note that $S^{GH} = S^G + S^H - S^{G \cap H}$, thus

$$\widehat{B} = \tilde{X}' \left(\hat{\tilde{\varepsilon}} \hat{\tilde{\varepsilon}}' \odot S^G \right) \tilde{X} + \tilde{X}' \left(\hat{\tilde{\varepsilon}} \hat{\tilde{\varepsilon}}' \odot S^H \right) \tilde{X} - \tilde{X}' \left(\hat{\tilde{\varepsilon}} \hat{\tilde{\varepsilon}}' \odot S^{G \cap H} \right) \tilde{X} \tag{5}$$

- Using (5) in (4) yields

$$Var(\tilde{\beta}) = Var^G(\tilde{\beta}) + Var^H(\tilde{\beta}) - Var^{G \cap H}(\tilde{\beta}) \tag{6}$$

- The decomposition in (7) is instructive
- The total two-way cluster-robust variance covariance matrix is made up of the individual one-way cluster robust variance covariance matrices, minus the joint clustering (we subtract to avoid double counting)
- This decomposition suggests a useful one-way implementation to construct $Var(\tilde{\beta})$:
  - Compute the within estimator and estimate the variance matrix clustering purely on G, purely on H and purely on $G \cap H$
  - Combine these three variance matrices to construct the full, two-way cluster robust variance-covariance matrix

- The clustering of standard errors, either one-way or two-way, hinges on the fact that the user know the direction of the clustering
- Currently no test exists for clustering in a given dimesion
- However, it is straightforward to think of exploiting the fact that clustering in a given dimension presents itself as a specific structure in the error terms (think serial correlation in the one-way random effects model) which could be tested
- For example, if clustering was based on gender, then we would expect correlation amongst the error terms for all men and all women

- As with heteroskedasticity robust standard errors, cluster robust standard errors are biased downwards
- Cameron, Gelbach and Miller (2008) present several small sample corrections that are useful in practice
- If you are using statistical software to construct your standard errors it is useful to know which (if any) correction is used
- For example, Stata uses $\sqrt{\frac{G(NT-1)}{(G-1)(NT-K)}}\hat{\tilde{\varepsilon}}_g$ in place of $\hat{\tilde{\varepsilon}}_g$ for one-way cluster robust construction

- Thompson (2011) provides a detailed discussion of when to use clustered standard errors based on individual and time
- There are three features of the data that are important to consider when deciding if clustered standard errors are useful: the distribution of the errors, the distribution of the regressors, and the number of observations in both clustering dimensions
- Using Cameron et al.'s (2011) decomposition of the clustered variance-covariance formula in (7), clustering on firm and time yields

$$Var(\tilde{\beta}) = Var^{firm}(\tilde{\beta}) + Var^{time}(\tilde{\beta}) - Var^{white}(\tilde{\beta}) \quad (7)$$

where $Var^{white}(\tilde{\beta})$ is the usual heteroskedasticity robust variance covariance matrix of White (1980)

- If the analyst only clusters on firm, then $Var^{time}(\tilde{\beta}) - Var^{white}(\tilde{\beta})$ is omitted from the variance estimate
- If the analyst only clusters on time, then $Var^{firm}(\tilde{\beta}) - Var^{white}(\tilde{\beta})$ is omitted from the variance estimate
- When these omitted terms are large, there will be meaningful consequences on the variance estimates
- For example, if, conditional on the regressors, the error terms are not correlated across firms, then there is no bias induced by omitting clustering on time

- Thompson also shows that it is more important to cluster on the smaller dimension, in micro panels this would be clustering based on time
- One can show that if the dimensions of the panel are severely distorted then we need not worry about clustering in both dimensions

$$\lim_{T\to\infty, N \text{ fixed}} \frac{Var^{firm}(\tilde{\beta}) + Var^{time}(\tilde{\beta}) - Var^{white}(\tilde{\beta})}{Var^{firm}(\tilde{\beta})} = 1 \tag{8}$$

- Intuition: as T becomes large, we average away noise due to variation across time, but we do not average away noise due to variation across firms
- The opposite results would hold if we held T fixed and let N grow large

- Within the random effects framework we can think of allowing each of the three error components to be heteroskedastic
- Mazodier and Trognon (1978) consider the modified random effects framework with $c_i \sim D(0, \sigma^2_{c_i})$ and $\varepsilon_{it} \sim IID(0, \sigma^2_{\varepsilon})$
- In this case the variance covariance matrix of the one-way error component $u_{it} = c_i + \varepsilon_{it}$ is

$$\Omega = diag(\sigma^2_{c_i}) \otimes J_T + \sigma^2_{\varepsilon} (I_N \otimes I_T) \tag{9}$$

where $diag(\sigma^2_{c_i})$ is an $N \times N$ diagonal matrix

- Baltagi and Griffin (1988) implemented this random effects framework by following Fuller and Battese (1974), replace J_T with $T\bar{J}_T$ and I_T with $E_T + \bar{J}_T$
- Doing so yields

$$\Omega = diag(T\sigma_{c_i}^2 + \sigma_\varepsilon^2) \otimes \bar{J}_T + \sigma_\varepsilon^2 (I_N \otimes E_T) \qquad (10)$$

- From this spectral decomposition we have

$$\Omega^r = diag[(\tau_i^2)^r] \otimes \bar{J}_T + (\sigma_\varepsilon^2)^r \left(I_N \otimes E_T\right) \tag{11}$$

  where $\tau_i^2 = T\sigma_{c_i}^2 + \sigma_\varepsilon^2$
- Premultiplication by $\sigma_\varepsilon \Omega^{-1/2}$ yields the variable transformation $\check{z} = z_{it} - \theta_i \bar{z}_{i\cdot}$ where $\theta_i = 1 - (\sigma_\varepsilon / \tau_i)$
- OLS estimation of $\check{y}$ on $\check{X}$ produces the GLS estimator

- Given that $\sigma^2_{c_i}$ is unknown it must be estimated
- However, as noted by Phillips (2003), this leads to an incidental parameters problem unless $T \longrightarrow \infty$
- An alternative is to parametrically specify $\sigma^2_{c_i}$ and use the residuals from within estimation to consistently estimate the unknown parameters of the individual skedastic functions

- We could switch the source of the heteroskedasticity in the one-way model
- $c_i \sim IID(0, \sigma_c^2)$ and $\varepsilon_{it} \sim D(0, \sigma_{\varepsilon_{it}}^2)$
- However, it is likely that heteroskedasticity occurs at all levels of the error component
- In this setting some restrictions must be made otherwise there are more parameters than observations $(N + NT)$
- When this happens a similar transformation of y and X will occur, but the form of $\check{z}$ is complicated; see Randolph (1988, page 352)

- Given that misspecification of the heteroskedasticity of the error term can cause problems, it is advised to construct robust standard errors; this is the common recommendation in cross-sectional settings
- However, in panel data settings, one must also contend with the assumptions implicit in the fixed and random effects framework
- Thus, in panel data settings, heteroskedasticity is a more important phenomena to understand that in the cross-sectional setting

- The extant serial correlation in the error term for the random effects framework of the unobserved effects model is quite simple
- When studying relationships such as capital investment, inflation or consumption this is not general enough
- Ignoring this form of serial correlation will result in consistent but inefficient estimators of the coefficients and biased estimators of the variance-covariance matrix
- When serial correlation is present a generalization of the Fuller and Battese (1973) transformation can be used to correct for the presence of serial correlation

- Our random effects framework is

$$y_{it} = x'_{it}\beta + c_i + \varepsilon_{it} \tag{12}$$

  where $c_i \sim IID(0, \sigma_c^2)$ and $\varepsilon_{it} = \rho\varepsilon_{i,t-1} + \nu_{it}$
- Using this setup, Baltagi and Li (1991) derive the GLS transformation following the pure time-series approach of Prais and Winsten

- The transformation makes use of the fact that $\nu_{i,0} \sim IID(0, \sigma_\nu^2/(1-\rho^2))$ which results in a transformation matrix for each individual of

$$\Gamma = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -\rho & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 & 0 \\ 0 & 0 & 0 & \cdots & -\rho & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & -\rho & 1 \end{bmatrix} \tag{13}$$

- The transformed regression error is

$$u = (I_n \otimes \Gamma)(c + \varepsilon) = (I_n \otimes \Gamma \imath_T)c + (I_n \otimes \Gamma)\varepsilon \tag{14}$$

- Noting that $\Gamma \imath_T = (1-\rho)\tilde{\imath}_T$ where $\tilde{\imath}_T = (\varpi, \imath'_{T-1})$ with $\varpi = \sqrt{(1+\rho)/(1-\rho)}$, u can be written as

$$u = (1-\rho)(I_n \otimes \tilde{\imath}_T)c + (I_n \otimes \Gamma)\varepsilon \tag{15}$$

- The variance-covariance matrix of the transformed errors is then

$$\Omega = \sigma_c^2(1-\rho)^2(I_N \otimes \tilde{\imath}_T\tilde{\imath}'_T) + \sigma_\nu^2(I_N \otimes I_T) \tag{16}$$

- Following Wansbeek and Kapteyn, notice that $d^2 = \tilde{\imath}_T' \imath_T = \varpi^2 + (T-1)$
- We can replace $\tilde{\imath}_T \tilde{\imath}_T'$ with $\varphi^2 \bar{J}_T^{\varpi}$ in (16) and note that $I_T = E_T^{\varpi} + \bar{J}_T^{\varpi}$, then the spectral decomposition of Ω is

$$\Omega = \sigma_{\varpi}^2 (I_N \otimes \bar{J}_T^{\varpi}) + \sigma_{\nu}^2 (I_N \otimes E_T^{\varpi}) \tag{17}$$

  where $\sigma_{\varpi}^2 = d^2 \sigma_c^2 (1-\rho)^2 + \sigma_{\nu}^2$
- Lastly, $\sigma_{\nu} \Omega^{-1/2} = (I_N \otimes I_T) - \theta_{\varpi} (I_N \otimes \bar{J}_T^{\varpi})$ with $\theta_{\varpi} = 1 - (\sigma_{\nu} / \sigma_{\varpi})$

- Using this transformation we have the transformation

$$\check{z}_i = \sigma_\nu \Omega^{-1/2} z_i = (z_{i1} - \theta_\varpi \varpi r_i, z_{i2} - \theta_\varpi r_i, \ldots, z_{iT} - \theta_\varpi r_i) \tag{18}$$

where

$$r_i = \frac{\varpi z_{i1}^* + \sum_{t=2}^{T} z_{it}^*}{d^2} \tag{19}$$

and $z^* = (I_N \otimes \Gamma) z$

- Notice that the first observation for each individual gets a different transformation than the remainder of the observations
- There is a transformation stemming from the Prais and Winsten transformation (this is so the first observations are not discarded as in a Cochrane and Orcutt type transformation)
- There is also a Fuller and Battese transformation on top of this transformation
- Notice that if $\varphi = 1$, then $d^2 = T$, $\sigma^2_\varphi = \sigma^2_1$ and $\theta_\varpi = \theta$, the classic one-way GLS transformation with no serial correlation arises

- Baltagi and Li (1991) suggesting estimating ρ via $\frac{\widehat{Q}_1-\widehat{Q}_2}{\widehat{Q}_0-\widehat{Q}_1}$ where

$$\widehat{Q}_s = \frac{\sum_{i=1}^{N}\sum_{t=s+1}^{T}\hat{u}_{it}\hat{u}_{i,t-s}}{N(T-s)} \tag{20}$$

  using pooled OLS residuals
- Once an estimate of ρ has been determined, the unknown variance components can be estimated using the pooled OLS residuals (with the corresponding Prais-Winsten transformation) with

$$\hat{\sigma}_\nu^2 = u'(I_N \otimes E^\varpi)u/N(T-1) \tag{21}$$

  and

$$\hat{\sigma}_\varpi^2 = u'(I_N \otimes \bar{J}_T^\varpi)u/N \tag{22}$$

- Similar transformations exist for AR (2) and AR(4) specifications of the error term as well
- Special care is needed when there exist unequal spacing across individuals, but a simple transformation still exists
- MA (1) processes can also be accomodated
- If concerned about serial correlation should test for its presence using methods described from lecture 2

- While cluster-robust and heteroskedasticity robust standard errors are simple to compute, in small sample settings they can be unreliable
- An alternative to asymptotically valid formulas is to deploy the bootstrap
- While resampling in a cross sectional setting is relatively straightforward, resampling from a panel requires additional care

- Cameron, Gelbach and Miller (2008) recently discussed bootstrapping in panel data models where clustering is present
- Robust standard errors are only valid asymptotically and typically require the number of observations or clusters goes to infinity
- In many economic settings this is untenable; A study where clustering is on the number of regions of the world will clearly not be able to argue convincingly that the number of regions is increasing
- Bias adjustment of clustered standard errors is possible (Angrist and Lavy, 2002) but this is not a panacea

- Bootstrap methods generate a number of pseudo-samples from the original sample
- For each pseudo-sample calculate the statistic of interest, and use the distribution of this statistic across pseudo-samples to infer the distribution of the original sample statistic (instead of using the asymptotic distribution)
- Many options for resampling the data available; which one to use?

- In panel dimension there is a further complication that does not exist when bootstrapping in cross-sectional settings – the unbalanced panel
- When we bootstrap we want the resamples to have the same size as the original sample, however, given the time dimension, we also want to sample individuals using all available observations
- From resample to resample the sample size could be bigger or smaller than the original sample depending on those individuals that are selected
- This would suggest that a pairs bootstrap would not work well in unbalanced panel data settings

- Can use the unbalanced wild bootstrap
- How it works:
  - Resample residuals for each individual based on standard wild bootstrap methodology
  - ALL residuals for an individual are multiplied by same rescaling
- This ensures that the sample size is constant across resamples and within person correlation is accounted for

- For the estimated unobserved effects model we have residuals $\hat{\varepsilon}_{it} = y_{it} - \hat{c}_i - x'_{it}\hat{\beta}$
- For the i^{th} individual's residuals, $\{\hat{\varepsilon}_{i1}, \ldots, \hat{\varepsilon}_{iT_i}\}$, form bootstrap residuals $\left\{\hat{\varepsilon}^*_{i1}, \ldots, \hat{\varepsilon}^*_{iT_i}\right\} = a \cdot \{\hat{\varepsilon}_{i1}, \ldots, \hat{\varepsilon}_{iT_i}\}$ with probability p_a and $= b \cdot \{\hat{\varepsilon}_{i1}, \ldots, \hat{\varepsilon}_{iT_i}\}$ with probability p_b
- The constants a and b and probabilities p_a and p_b depend on the type of wild bootstrap deployed
  - The Mammen (1993) version of the wild bootstrap uses $a = (1 - \sqrt{5})/2 \approx -0.6180$ and $b = 1 - a \approx= 0.3820$ with probabilities $p_a = (1 + \sqrt{5})/2\sqrt{5} \approx 0.7236$ and $p_b = 1 - p_a \approx 0.2764$
  - The Rademacher version of the wild bootstrap uses $a = 1$ and $b = -1$ with probabilities $p_a = p_b = 0.5$
- This could be generalized to resample on clusters different than the individual (region or time for example)

- The Mammen wild bootstrap ensures that the first three moments of the resampled residual distribution match the first three moments of the actual residual distribution while the Rademacher wild bootstrap can only match the first two moments
- The matching on the third moment is important theoretically
- Standard bootstrap theory dictates that the bootstrap provides asymptotic refinements because it captures departures from symmetry (the third moment)
- Thus, the bootstrap can provide finite sample improvements over asymptotic approxmiations based on the degree of asymmetry it can capture

- Wild bootstrap standard errors
  1. Estimate $\hat{\beta}$ and $\hat{c}_i$
  2. For a resample $\left\{(y^*_{11}, \hat{\varepsilon}^*_{11}), \ldots, (y^*_{NT_N}, \hat{\varepsilon}^*_{NT_N})\right\}$ using the wild bootstrap for the appropriate cluster
  3. Obtain estimates $\hat{\beta}^*_s$ and $\hat{c}^*_{i,s}$
  4. Repeat steps 2 and 3 B times
  5. Reject H_0 : at the α level if $t_j > z^B_{\alpha/2}$ where $t_j = \frac{\hat{\beta}_j - \beta^o_j}{s.e.(\hat{\beta}_{j,B})}$ and

$$s.e.(\hat{\beta}_{j,B}) = \left(\frac{1}{B-1} \sum_{s=1}^{B} \left(\hat{\beta}^*_{j,s} - \bar{\hat{\beta}}^*_j \right)^2 \right)^{1/2}$$

where $\bar{\hat{\beta}}^*_j = B^{-1} \sum_{s=1}^{B} \hat{\beta}^*_{j,s}$

- There are a total of $\binom{2n-1}{n}$ unique resamples of the data
- This gets large very quickly as n increases
- In practice want B large enough to obtain a good approximation, but small enough that the computational burden is minimal
- At a minimum B should be such that $\alpha(B+1)$ is an integer; this ensures that your bootstrap can produce a test with exact size
- B controls the power of your test, larger B equates to a smaller power loss
- Common in practice to see $B = 399$ and $B = 999$; quite interestingly, B should be selected with regarding to significance level being tested
- Smaller α, larger B should be

- Discussed importance of clustering, heteroskedasticity and serial correlation
- Easily remedied in applied settings using bootstrap, GLS type corrections and/or heteroskedasticity robust covariance matrices
- Bootstrap provides a simple finite sample approach to calculating robust statistics and standard errors