Applied Panel Data Analysis – Lecture 1

Christopher F. Parmeter

AGRODEP
September 9-13th, 2013
Dakar, Senegal

- Learn current methods for modeling with panel data
- Apply these methods with actual datasets for hands on learning
- Use the open source statistical software R
- Overall focus will be on the applications of the econometric models with brief overviews of the statistical underpinnings

- Access to panel data offers the analyst options not available with cross-section or time series data
- Can track individuals/families/regions/firms over time providing more dynamic analysis
- Unobserved heterogeneity easier to control for, allows for more robust conclusions from the econometric model

- Control for individual heterogeneity
- More informative data
- More variability in the data
- Less collinearity
- Higher degrees of freedom
- Dynamic adjustment
- Reduce aggregation bias
- Test more complicated models of behavior

- Consider a cross-section of women with 50% average annual participation rate in the labor force
- This 50% could arise because each woman has a 50% chance of participating in the labor market in any given year or 50% of the women work every year while 50% of the women never work
- These two cases are extremes, in one case there is high turnover while in the other this is no turnover
- We would need panel to distinguish between these two cases

- Data design and collection
- Distortion of measurement error
- Self-selection
- Nonresponse
- Attrition
- Short time-series dimension
- Cross-sectional dependence

- With access to panel data (and panel data models) comes more choices available to the analyst
- A strong background in each model is required to ensure proper application and interpretation of the results

- To begin, assume we have N cross-sectional units, observed over T time periods, for a total of NT observations
- x_{it} is a $1 \times K$ vector of covariates (or regressors) for $i = 1, \ldots, N$ and $t = 1, \ldots, T$
- The population model is

$$y_{it} = x_{it}\beta + \varepsilon_{it}, \tag{1}$$

  where β is a $K \times 1$ vector, y_{it} is our scalar response (regressand) variable and ε_{it} is the regression error
- The model in (1) is a linear panel data model

- If we ignore the double subscript on y, x and ε there is nothing that distinguishes the linear panel data model from a linear cross-sectional model
- How we use the i and t dimensions of the data will determine how much we exploit the panel structure afforded to us

- Use pooled ordinary least squares (OLS) to estimate linear panel data model in (1)
- Premultiply (1) by x'_{it} to obtain

$$x'_{it}y_{it} = x'_{it}x_{it}\beta + x'_{it}\varepsilon_{it} \tag{2}$$

- If we assume that $E(\varepsilon_{it}) = 0$ and $Cov(x_{it}, \varepsilon_{it}) = 0$ then we have

$$\beta = \left[E\left(x'_{it}x_{it}\right)\right]^{-1} E\left(x'_{it}y_{it}\right) \tag{3}$$

- Given data on x and y we can estimate both of these expectations to construct at estimator for β
- We replace $E\left(x_{it}'x_{it}\right)$ with $(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}'x_{it}$ and $E\left(x_{it}'y_{it}\right)$ with $(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}'y_{it}$
- Our pooled OLS estimator is

$$\hat{\beta} = \left((NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}'x_{it}\right)^{-1}\left((NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}'y_{it}\right) \tag{4}$$

- These summation signs can get cumbersome
- Can be easier to work with matrices
- Let $X_i = (x_{i1}, x_{i2}, \ldots, x_{iT})$, $y_i = (y_{i1}, y_{i2}, \ldots, y_{iT})$ and $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT})$
- Further, let $y = (y_1, y_2, \ldots, y_N)$, $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)$, which we refer to as stacked vectors; they both have dimension $NT \times 1$
- Finally, we have $X = (X_1, X_2, \ldots, X_N)$, which is the stacked matrix of covariates; this matrix has dimension $NT \times K$

- With this notation we can express our pooled OLS estimator of β as

$$\hat{\beta} = (X'X)^{-1}X'y \tag{5}$$

- To determine the limiting distribution of our pooled OLS estimator and the variance of this distribution we need to manipulate how our estimator looks
- Note that we can equivalently write $\hat{\beta}$ in (4) as

$$\hat{\beta} = \beta + \left((NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} x'_{it} x_{it} \right)^{-1} \left((NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} x'_{it} \varepsilon_{it} \right) \quad (6)$$

- This decomposition is useful because we can then bring β to the left hand side and multiply by $\sqrt{NT}$ to obtain

$$\sqrt{NT}\left(\hat{\beta}-\beta\right)=\left((NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}'x_{it}\right)^{-1}\left((NT)^{-1/2}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}'\varepsilon_{it}\right) \quad (7)$$

- Or in matrix form

$$\sqrt{NT}\left(\hat{\beta}-\beta\right)=\left((NT)^{-1}X'X\right)^{-1}\left((NT)^{-1/2}X'\varepsilon\right) \quad (8)$$

- Under very minimal assumptions we can show that

$$\left((NT)^{-1}X'X\right)^{-1} \xrightarrow{P} E\left[X'X\right]^{-1} = A^{-1}$$

- Further, by the central limit theorem we have

$$\left((NT)^{-1/2}X'\varepsilon\right)^{-1} \xrightarrow{D} N\left(0, E\left[X'\varepsilon\varepsilon'X\right]\right)$$

- Let $B = E\left[X'\varepsilon\varepsilon'X\right]$

- We combine these two results to obtain

$$\sqrt{NT}\left(\hat{\beta}-\beta\right) \xrightarrow{D} N\left(0, A^{-1}BA^{-1}\right) \tag{9}$$

- We refer to $A^{-1}BA^{-1}$ as the sandwich form of the asymptotic variance-covariance matrix
- Under a homoskedasticity assumption, $E\left[\varepsilon\varepsilon'|X\right]=\sigma^2 I$, we have that $E\left[X'\varepsilon\varepsilon'X\right]=\sigma^2 E\left[X'X\right]=\sigma^2 A$ so $A^{-1}BA^{-1}=\sigma^2 A^{-1}$, yielding

$$\sqrt{NT}\left(\hat{\beta}-\beta\right) \xrightarrow{D} N\left(0, \sigma^2 A^{-1}\right) \tag{10}$$

- So what do all of these results tell us?
- Under minimal assumptions, treating the panel data as a pooled sample, the pooled OLS estimator is unbiased, consistent and asymptotically normal with variance-covariance matrix $\sigma^2 A^{-1}$
- When might these minimal assumptions be violated?
- In micro datasets it is likely that heteroskedasticity is present
- If lagged variables are present as part of the covariate set then we may need a stronger assumption than contemporaneous exogeneity, $Cov(x_{it}, \varepsilon_{it}) = 0$

- $Cov(x_{it}, \varepsilon_{it})$ can be equivalently restated as $E\left[\varepsilon_{it}|x_{it}\right] = 0$
- A slightly stronger condition that can allow for some dynamics in the linear panel data model is sequential exogeneity, $E\left[\varepsilon_{it}|x_{it}, x_{it-1}, \ldots, x_{i1}\right] = 0$
- An even stronger restriction would be strict exogeneity, $E\left[\varepsilon_{it}|x_{i1}, x_{i2}, \ldots, x_{iT}\right] = 0$

- Contemporaneous exogeneity says nothing about the relationship between x_{is} and ε_{it} for $s \neq t$
- Sequential exogeneity says nothing about the relationship between x_{is} and ε_{it} for $s > t$
- Strict exogeneity rules out correlations across all time periods between x and ε
- It is crucial that we understand that these three different assumptions have different implications for the statistical properties of the estimators we will study

- Let $x_{it} = (1, y_{it-1})$ so our model is

$$y_{it} = \beta_0 + \beta_1 y_{it-1} + \varepsilon_{it}$$

- Contemporaneous exogeneity holds by construction if $E\left[y_{it}|y_{it-1}\right] = \beta_0 + \beta_1 y_{it-1}$ is the data generating process
- Sequential exogeneity holds if $E\left[y_{it}|y_{it-1}, y_{it-2}\ldots, y_{i0}\right] = E\left[y_{it}|y_{it-1}\right]$ which implies that only a single lag of y_{it} appears in the full dynamic expectation
- Strict exogeneity would fail because $E\left[\varepsilon_{it}|y_{i0}, y_{i1}\ldots, y_{iT-1}\right] =$ $E\left[y_{it} - \beta_0 - \beta_1 y_{it-1}|y_{i0}, y_{i1}\ldots, y_{iT-1}\right] = \varepsilon_{it} \neq 0$
- So strict exogeneity fails when there are lagged dynamics in a model, but sequential or contemporaneous exogeneity will still hold depending on the type of dynamics

- Consider the model of Holzer et al. (1993) who study the impact of job training grants on firm's scrap rates
- A generic linear panel data model for their setup would be

$$\log(scrap_{it}) = \beta_0 + \beta_1 grant_{it} + \varepsilon_{it} \tag{11}$$

- An overriding concern with this generic setup is that firms that receive grants may have high scrap rates to start with
- We could account for this by including the lagged scrap rate

$$\log(scrap_{it}) = \beta_0 + \beta_1 grant_{it} + \beta_2 \log(scrap_{it-1}) + \varepsilon_{it} \tag{12}$$

- Now we would need to worry about the different implications of contemporaneous and sequential exogeneity for our pooled OLS estimator

- Note that our assumption that $E[X'\varepsilon\varepsilon' X] = \sigma^2 E[X'X]$ is quite restrictive
- First, we assuming that the error term is constant with respect to our covariate
- Second, we are assuming that the unconditional error variance does not vary over time
- Third, we also have that $E[\varepsilon_{it}\varepsilon_{is}x'_{it}x_{is}] = 0$ for $t \neq s$
- In any application any of these assumptions may be seen as overly restrictive

- When heteroskedasticity is present it is no longer the case that $E[X'\varepsilon\varepsilon' X] = \sigma^2 A$
- Following White (1980), we can replace B in the sandwich form with a consistent estimator
- White (1980) proved that

$$(NT)^{-1} X'\hat{\varepsilon}\hat{\varepsilon}' X \xrightarrow{P} E[X'\varepsilon\varepsilon' X] = B, \qquad (13)$$

where $\hat{\varepsilon} = y - X'\hat{\beta}$

- Using this estimator for B will allow us to conduct heteroskedasticity robust inference

- Pooled panel data estimation works almost identical to OLS
- Care is required regarding the statistical assumptions placed on the error terms and the covariates
- Pooled OLS estimator is unbiased, consistent and asymptotically normal
- Heteroskedasticity robust inference can easily be undertaken