Panel Data

Introduction

Definition

Panel data refers to data with multiple observations of each basic unit of observation.

- Does not necessarily involve time dimension
- Examples
  - $y_{it} = \text{outcome of observation } i \text{ at time } t$ ($i = 1, \ldots, N; t = 1, \ldots, T$)
    - Exs: countries over time; firms over time; individuals over time
  - $y_{ih} = \text{outcome of individual } i \text{ in household } h$ ($i = 1, \ldots, N; h = 1, \ldots, H$)
  - $y_{im} = \text{outcome of employee } i \text{ in firm } m$ ($i = 1, \ldots, N; m = 1, \ldots, M$)
  - $y_{is} = \text{test score of student } i \text{ in subject } s$ ($i = 1, \ldots, N; s = 1, \ldots, S$)
Preceding examples are two-dimensional panels

- More complex panels involve more dimensions
- Examples
  - \( y_{ist} = \text{outcome of county } i \text{ in state } s \text{ at time } t \)
  - \( y_{iht} = \text{outcome of individual } i \text{ in household } h \text{ at time } t \)
  - \( y_{ijst} = \text{outcome of firm } i \text{ in industry } j \text{ in state } s \text{ at time } t \)

- For exposition, we will stick to repeated observations of the same unit \( i \) over time \( t \)
Types of panel data sets ...

Definition
Balanced panel data set refers to situation where each unit of observation $i$ is observed the same number of time periods, $T$. Thus, the total sample size is $NT$.

Definition
Unbalanced panel data set refers to situation where each unit of observation $i$ is observed an unequal number of time periods, $T_i$. Thus, the total sample size is $\sum_i T_i$.

Micro vs macro panels
- Micro: Typically large $N$, small $T$
- Macro: Typically small $N$, large $T$
  - Some convergence over time as data become plentiful
Why does panel data require special attention?

Panel data is useful due to

1. Efficiency gains from large samples
2. Potential to analyze dynamic behavior (relative to CS)
3. Potential to address correlation between covariates and the error
   - Panel data methods provide a solution to endogeneity in only certain situations
   - Requires ‘new’ estimation techniques

Asymptotics with panel data requires more precision

- Large $N$, small $T$
- Small $N$, large $T$
- Large $N$, large $T$ (jointly or sequentially)

Panel Data
Model & Terminology

- Population regression fn given by $E[y | x_1, ..., x_k, c]$
  - Assuming linearity: $E[y | x_1, ..., x_k, c] = \beta_0 + x \beta + c$
  - $x_k$, $k = 1, ..., K$, are observable (to the econometrician)
  - $c$ is an unobservable (to the econometrician) variable

- Error form of the model

$$y = \beta_0 + x \beta + c + \epsilon$$

where $c$ is the unobserved effect and $\epsilon$ is the idiosyncratic error

- $c$ also referred to as individual heterogeneity, unobserved component, individual effect, latent variable
Some history

- \( c \) may be thought of as either a random variable, whose dbn may be estimated, or as a fixed parameter to be estimated
  - \( c \) is called a \textit{random effect} in the first view, a \textit{fixed effect} in the second view (Balestra & Nerlove 1966)
  - This terminology is not the common usage of these terms today
- Terminology has become linked to assumption concerning correlation between \( x \) and \( c \)
  - Random effect: \( \text{Cov}(x, c) = 0 \)
  - Fixed effect: \( \text{Cov}(x, c) \neq 0 \)
- The ‘modern’ usage of the terms are now universal
Formally, model is given by

\[ y_{it} = \beta_0 + x_{it} \beta + \tilde{e}_{it} \]

where \( \tilde{e}_{it} \) is the composite error and captures all unobserved determinants of \( y_{it} \)

- The composite error is decomposed into parts
- In a one-way error components model, \( \tilde{e}_{it} = c_i + \varepsilon_{it} \)
  - Unobserved effect captures time invariant unobserved attributes specific to \( i \)
  - Idiosyncratic error captures the remainder
- In a two-way error components model, \( \tilde{e}_{it} = c_i + \lambda_t + \varepsilon_{it} \)
  - Unobserved time effect captures individual invariant unobserved attributes specific to \( t \)
  - Idiosyncratic error captures the remainder

Problem: Given presence of \( c_i \) and \( \lambda_t \) and access to panel data, how can we recover consistent estimates of \( \beta \)?
Roadmap

1. Estimation techniques
   1. Pooled OLS (POLM)
   2. Random effects (RE)
   3. Least squares dummy variable model (LSDV)
   4. Fixed effects (FE)
   5. First-differencing (FD)
   6. Long differences (LD)

2. Dynamic panel models
3. LDV models revisited
4. Time series concerns revisited
Panel Data
Pooled OLS

- Estimate by OLS
  \[ y_{it} = \beta_0 + x_{it}\beta + \tilde{\epsilon}_{it} \]

- Assumptions

(POLS.i) Linearity

(POLS.ii) (Contemporaneous) Exogeneity: \( E[x_{it}'\tilde{\epsilon}_{it}] = 0, \ t = 1, \ldots, T \)

- If \( E[\tilde{\epsilon}\tilde{\epsilon}'] \neq \sigma^2 I_{NT} \), POLS
  1. Is inefficient relative to GLS
  2. Provides incorrect standard errors
  3. Requires large \( N \), small \( T \) asymptotics
Two options

- Estimate parameters using POLS, but report cluster-robust standard errors to allow for arbitrary correlation within $i$
  
  - See Cameron & Miller (JHR, 2015)
  - More details later
  - Stata: -reg, vce(cluster $i$)-

- Place structure on $E[\tilde{\varepsilon}\tilde{\varepsilon}'] = \Omega(\theta)$ and use pooled FGLS (PFGLS)

  - Equicorrelated errors: $\rho_{ts} = \text{Corr}(\tilde{\varepsilon}_{it}, \tilde{\varepsilon}_{is}) = \rho \ \forall i, t \neq s$
  - Autoregressive errors: $AR(p)$ s.t. $\tilde{\varepsilon}_{it} = \sum_{s=1}^{p} \rho_s \tilde{\varepsilon}_{it-s} + u_{it}$
  - Moving average errors: $MA(p)$ s.t. $\tilde{\varepsilon}_{it} = u_{it} - \sum_{s=1}^{p} \theta_s u_{it-s}$
  - Unstructured, except equal across $i$:
    \[
    \hat{\rho}_{ts} = \frac{1}{N} \sum_i \left( \tilde{\varepsilon}_{it} - \tilde{\varepsilon}_t \right) \left( \tilde{\varepsilon}_{is} - \tilde{\varepsilon}_s \right)
    \]

  - Stata: -xtreg, pa-
Panel Data
Random Effects

- Estimate by FGLS

\[ y_{it} = \beta_0 + x_{it}\beta + \tilde{\epsilon}_{it}, \quad \tilde{\epsilon}_{it} = c_i + \epsilon_{it} \]

- This is identical to POLS except explicit structure is placed on the composite error

- Assumptions

  (RE.i) Strict exogeneity: \( E[\epsilon_{it}|x_i, c_i] = 0, \ t = 1, \ldots, T \) (\( x_i \) is a vector)
  (RE.ii) Independence: \( E[c_i|x_i] = E[c_i] = 0 \)
  (RE.iii) Full rank: \( E[x_i'\Omega_i^{-1}x_i] = K \)
    where \( \Omega_i = E[\tilde{\epsilon}_i\tilde{\epsilon}_i'] \) is a \( T \times T \) matrix
  (RE.iv) Conditional homoskedasticity and zero covariances:
    \( E[\epsilon_i\epsilon_i'|x_i, c_i] = \sigma^2_\epsilon I_T \)
  (RE.v) Homoskedasticity: \( E[c_i^2|x_i] = \sigma^2_c \)
Estimation by Feasible GLS (FGLS)

- Under (RE.i) – (RE.iii), FGLS uses an unrestricted estimate of $\hat{\Omega}_i$

$$\hat{\beta}_{FGLS} = \left( \sum_i x'_i \hat{\Omega}_i^{-1} x_i \right)^{-1} \left( \sum_i x'_i \hat{\Omega}_i^{-1} y_i \right)$$

- Using (RE.iv), (RE.v), $\Omega_i = \Omega(\theta)$ has a convenient form

$$\Omega(\theta) = \begin{bmatrix}
\sigma^2_{\varepsilon} + \sigma^2_c & \sigma^2_c & \cdots & \sigma^2_c \\
\sigma^2_c & \ddots & \ddots & \vdots \\
\vdots & \ddots & \sigma^2_c \\
\sigma^2_c & \cdots & \sigma^2_c & \sigma^2_{\varepsilon} + \sigma^2_c
\end{bmatrix} \quad \text{(RE structure)}$$

- RE structure $\Rightarrow$

$$\text{corr}(\tilde{\varepsilon}_{it}, \tilde{\varepsilon}_{is}) = \frac{\sigma^2_c}{\sigma^2_c + \sigma^2_{\varepsilon}} = \frac{\text{Var}(c)}{\text{Var}(\tilde{\varepsilon})} \geq 0 \quad \forall \ t \neq s$$

which measures the fraction of the variance of the composite error due to the unobserved effect
Given estimates of $\hat{\sigma}_c^2$ and $\hat{\sigma}_\varepsilon^2$, estimator given by

$$\hat{\beta}_{RE} = \left( \sum_i x_i' \hat{\Omega}^{-1} x_i \right)^{-1} \left( \sum_i x_i' \hat{\Omega}^{-1} y_i \right)$$

where

$$\hat{\Omega} = \hat{\sigma}_\varepsilon^2 I_T + \hat{\sigma}_c^2 j_T j_T'$$

$I_T = T \times T$ identity matrix

$j_T = T \times 1$ vector of 1s

Notes

- Consistency does not require (RE.iv), (RE.v) holding – just like OLS is consistent even when $\Omega \neq I_T$
- Consistency of FGLS does require strict exogeneity such that

$$\text{plim} \left( \sum_i x_i' \hat{\Omega}^{-1} x_i \right)^{-1} \left( \sum_i x_i' \hat{\Omega}^{-1} \varepsilon_i \right) = 0$$
To estimate $\hat{\sigma}_c^2$ and $\hat{\sigma}_\varepsilon^2$

- Do POLS $\Rightarrow \hat{\beta}_{POLS}$
- Obtain residuals $\Rightarrow \hat{\varepsilon}_{it}$
- Obtain estimates

\[
\hat{\sigma}_\varepsilon^2 = \frac{1}{NT - K} \sum_i \sum_t \hat{\varepsilon}_{it}^2
\]
\[
\hat{\sigma}_c^2 = \frac{1}{\left[ \frac{NT(T-1)}{2} - K \right]} \sum_i \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \hat{\varepsilon}_{it} \hat{\varepsilon}_{is}
\]
\[
\hat{\sigma}_\varepsilon^2 = \hat{\sigma}_\varepsilon^2 - \hat{\sigma}_c^2
\]

- If $\varepsilon_{it}$ is heteroskedastic and/or serially correlated, then

\[
\hat{\Omega} = \frac{1}{N} \sum_i \hat{\varepsilon}_i \hat{\varepsilon}_i
\]

is more efficient asymptotically, but improved finite sample performance requires $N \ggg T$

- Note: RE estimator is asymptotically equivalent to PFGLS imposing equicorrelated errors
Can show that the RE estimator is equivalent to a quasi-time demeaned estimator

- Define
  \[ z_{it} \equiv z_{it} - \lambda \bar{z}_i. \]
  for any variable \( z_{it} \) for some value of \( \lambda \) and \( \bar{z}_i. \) is the average of \( z \) within \( i \)

- The RE estimator may be obtained by POLS estimation of
  \[ y_{it} = x_{it} \beta + \varepsilon_{it} \]
  with
  \[ \lambda = 1 - \left[ \frac{1}{1 + T \left( \frac{\sigma_x^2}{\sigma_{\varepsilon}^2} \right)} \right]^{1/2} \]

- In contrast, POLS sets \( \lambda = 0 \)
- \( \lambda \neq 0 \Rightarrow \) strict exogeneity is needed instead of exogeneity
Specification test: RE vs. POLS

- Hypothesis

\[ H_0 : \sigma^2_c = 0 \]
\[ H_a : \sigma^2_c \neq 0 \]

- Devising a test is difficult because the null lies on the edge of the feasible parameter space

- Breusch-Pagan (1980) test

\[
\lambda_{LM} = \frac{NT}{2(T-1)} \left[ \frac{\sum_{i=1}^{N} \left( \sum_{t=1}^{T} \hat{\varepsilon}_{it} \right)^2}{\sum_{i=1}^{N} \sum_{t=1}^{T} \hat{\varepsilon}_{it}^2} - 1 \right] \sim \chi^2_1
\]

where \( \hat{\varepsilon}_{it} \) is the residual from POLS and \( \sum_{t=1}^{T} \hat{\varepsilon}_{it} = T \hat{\varepsilon}_{it} \)

- Rejection of \( H_0 \) is indicative of serial correlation and not necessarily the RE structure per se
A more robust test which is valid for any dbn of $\tilde{\epsilon}_{it}$ is based on

$$\frac{\sum_{i=1}^{N} \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \tilde{\epsilon}_{it} \tilde{\epsilon}_{is}}{\left[ \sum_{i=1}^{N} \left( \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \tilde{\epsilon}_{it} \tilde{\epsilon}_{is} \right)^{2} \right]^{1/2}} \sim \mathcal{N}(0, 1)$$

where rejection of $H_{o}$ again is indicative of serial correlation and not necessarily the RE structure per se.

Wood (Biometrika, 2013) develops an alternative test.

Baltagi & Li (1995) devise a test of serial correlation of $\epsilon_{it}$ assuming $c$ and $\epsilon$ are normally distributed; this is a test of the RE structure as $\epsilon$ should be serially uncorrelated.

Extension to two-way error component model most easily handled by adding $T - 1$ time dummies, although there is also a two-way RE error components model as well estimable via FGLS.

Stata: -xtreg, re-, -xttest0-
Panel Data

Fixed Effects

- **Model**
  \[ y_{it} = \beta_0 + x_{it}\beta + \tilde{\epsilon}_{it}, \quad \tilde{\epsilon}_{it} = c_i + \epsilon_{it} \]

- **Alternative strategies** transform the model to eliminate \( c_i \) from the model
  - Several possible transformations
  - One commonly used approach is FE; also known as *within estimator*, or mean-differencing

- **Assumptions**
  - (FE.i) (RE.i) **Strict exogeneity**
  - (FE.ii) **Full rank**: \( \text{E}[\check{x}_i'\check{x}_i] = K \)
    where \( \check{x}_i = x_i - \bar{x}_i \).
  - (FE.iii) (RE.iv) **Conditional homoskedasticity and zero covariances**
Estimation...

- Structural model

\[ y_{it} = c_i + x_{it}\beta + \varepsilon_{it} \]

where \( \beta_0 \) is subsumed into \( c_i \)

- Averaging over \( T \) time periods implies

\[ \bar{y}_i = c_i + \bar{x}_i\beta + \bar{\varepsilon}_i \]

and

\[ \bar{y} = \bar{c} + \bar{x}\beta + \bar{\varepsilon} \]

where bars indicate average over \( T \) obs within group; double bars indicate average over entire sample
Taking differences yields

\[ y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)\beta + (\varepsilon_{it} - \bar{\varepsilon}_i) \]
\[ \ddot{y}_{it} = \ddot{x}_{it}\beta + \ddot{\varepsilon}_{it} \]

- Estimable by POLS \textit{without a constant}
- Equivalent to our prior notation with \( \lambda = 1 \)

Alternative estimating equation

\[ \ddot{y}_{it} + \bar{y} = \bar{c} + (\ddot{x}_{it} + \bar{x})\beta + (\ddot{\varepsilon}_{it} + \bar{\varepsilon}) \]
\[ \dddot{y}_{it} = \bar{c} + \dddot{x}_{it}\beta + \dddot{\varepsilon}_{it} \]

- Estimable by POLS \textit{with a constant}
- Constant is the mean unobserved effect
Notes

- Although estimating equation differs from structural equation, interpret $\hat{\beta}_{FE}$ as original $\beta$
- Transformation eliminates $c_i$ and any time invariant $x$’s; time invariant $x$’s invalidate (FE.ii)
- Estimates of unobserved effects are recoverable from FE estimation

$$\hat{c}_i = \bar{y}_i - \bar{x}_i \hat{\beta}_{FE}$$

- The dbn of $\hat{c}$ can be examined to assess the amount of heterogeneity in the population
- $\hat{c}_i$ is the mean of $\hat{\varepsilon}_{it}$ over $T$ for each $i$
- $\hat{c}_i$ is unbiased, but consistency requires $T \to \infty$
- Stata: -predict, u-

- Between estimator is obtained from OLS estimation of

$$\bar{y}_i = \bar{x}_i \beta + \bar{\varepsilon}_i$$

which is only consistent under (RE.i) and (RE.ii)
Inference

\[
\text{Var} \left( \hat{\beta}_{FE} \right) = \hat{\sigma}_\varepsilon^2 \left( \sum_i \sum_t \bar{x}'_{it} \bar{x}_{it} \right)^{-1}
\]

\[
\hat{\sigma}_\varepsilon^2 = \frac{1}{N(T-1) - K} \sum_i \sum_t \varepsilon_{it}^2
\]

- Differs from the usual POLS std errors since that would use \(NT - K\) dof rather than \(N(T-1) - K\)
- Difference arises since std errors must account for fact that \(N\) dof are lost due to estimation of means
- Solution: Multiply std errors by correction factor \(\sqrt{\frac{NT-K}{N(T-1)-K}}\)

- Stata: -xtreg, fe-, -areg-
Extension to two-way error component model most easily handled by adding $T - 1$ time dummies

If $T$ is too large, an alternative is to transform the model to eliminate the time effects, $\lambda_t$

- Model given by
  
  \[ \tilde{y}_{it} = \tilde{x}_{it} \beta + \tilde{e}_{it} \]

  where

  \[ \tilde{z}_{it} = z_{it} - \bar{z}_i - \bar{z}_t + \bar{z} \]

  for $z = \{y, x, \varepsilon\}$ estimable by POLS with dof correction

- Additional parameters estimated as

  \[ \hat{c} = \tilde{y}_{it} - \tilde{x}_{it} \hat{\beta} \]
  \[ \hat{c}_i = (\bar{y}_i - \bar{y}) - (\bar{x}_i - \bar{x}) \beta \]
  \[ \hat{\lambda}_t = (\bar{y}_t - \bar{y}) - (\bar{x}_t - \bar{x}) \beta \]

- Stata: -felsdvreg-, -reghdfe-
Specification tests

- $\varepsilon_{it}$ are serially correlated
  - Not obvious since FE yields estimates of $\bar{\varepsilon}_{it}$, not $\varepsilon_{it}$
  - Moreover, $\bar{\varepsilon}_{it}$ are negatively serially correlated under $H_0: \varepsilon_{it}$ is serially uncorrelated
    \[
    \text{Corr}(\bar{\varepsilon}_{it}, \bar{\varepsilon}_{is}) = -\frac{1}{T-1} \quad \forall s \neq t
    \]
  - Test implemented by estimating via OLS
    \[
    \bar{\varepsilon}_{iT} = \alpha_0 + \alpha_1 \bar{\varepsilon}_{iT-1} + \eta_i, \quad i = 1, \ldots, N
    \]
    and testing $H_0: \alpha_1 = -1/(T-1)$ using a standard $t$-test
  - Presence of serial correlation invalidates the usual FE standard errors; cluster-robust std errors given by
    \[
    \text{Var} \left( \hat{\beta}_{FE} \right) = \left( \bar{x}' \bar{x} \right)^{-1} \left( \sum_i \bar{x}_i' \bar{\varepsilon}_i \bar{\varepsilon}_i \bar{x}_i \right)^{-1} \left( \bar{x}' \bar{x} \right)^{-1}
    \]

  - See Cameron & Miller (JHR, 2015)
Alternative to cluster-robust standard errors is FE-FGLS estimator may be used

- **AR(1) model setup**

\[
\begin{align*}
    y_{it} &= c_i + \beta x_{it} + \varepsilon_{it} \\
    \varepsilon_{it} &= \rho \varepsilon_{it-1} + u_{it}
\end{align*}
\]

where \(|\rho| < 1\) and \(u_{it} \sim WN(0, \sigma_u^2)\)

- Estimation proceeds by estimating \(\hat{\rho}\) and transforming the model to yield \(iid\) errors, then the FE or RE estimator is applied to the transformed model


- Stata: -xtregar-
• $\varepsilon_{it}$ are cross-sectionally correlated (aka cross-sectional dependence)

\[ H_0 : \text{Cov}(\varepsilon_{it}, \varepsilon_{jt}) = 0 \quad \forall i \neq j \]
\[ H_a : \text{Cov}(\varepsilon_{it}, \varepsilon_{jt}) \neq 0 \text{ for some } i \neq j \]

- $T > N$: Breusch-Pagan (1980) test

\[ \lambda_{LM} = T \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \hat{\rho}_{ij}^2 \sim \chi_d^2 \]

where $d = N(N - 1)/2$ and

\[ \hat{\rho}_{ij} = \hat{\rho}_{ji} = \frac{\sum_{t=1}^{T} \hat{\varepsilon}_{it} \hat{\varepsilon}_{jt}}{\sqrt{\left(\sum_{t=1}^{T} \hat{\varepsilon}_{it}^2\right) \left(\sum_{t=1}^{T} \hat{\varepsilon}_{jt}^2\right)}} \]

★ Intuition: compute $N \times N$ correlation matrix; test significance of off-diagonal terms
★ Test does not have good statistical properties when $T < N$, and likely to do worse as $N \to \infty$
★ See -xttest2- in Stata
(cont.)

- A scaled version is appropriate asymptotically if $T \to \infty$ and then $N \to \infty$

\[ \lambda_{SCLM} = \sqrt{\frac{1}{N(N - 1)}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (T \hat{\rho}_{ij}^2 - 1) \sim \mathcal{N}(0, 1) \]

with an improved version developed in Peasaran et al. (2008)


\[ \lambda_{CD} = \sqrt{\frac{2T}{N(N - 1)}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \hat{\rho}_{ij} \sim \mathcal{N}(0, 1) \]

as $N \to \infty$ for sufficiently large $T$

- Stata: -xtcsd, pes-
Groupwise heteroskedasticity

- Errors are homoskedastic within groups, heteroskedastic across groups
- For example, errors for a given individual have same variance in each period, but each individual has a unique variance
- Structure:

\[
\Sigma_i = \begin{bmatrix}
\sigma_i^2 \\
0 & \ddots & \\
0 & \ddots & 0 \\
0 & \ddots & \sigma_i^2
\end{bmatrix}, \quad \Sigma = \begin{bmatrix}
\Sigma_1 \\
0 & \ddots & \\
0 & \ddots & 0 \\
0 & \ddots & \Sigma_N
\end{bmatrix}
\]

- Hypothesis

\[H_0 : \sigma_i^2 = \sigma^2 \forall i \]
\[H_a : \sigma_i^2 \neq \sigma^2 \text{ for some } i\]
(cont.)

- Modified Wald test statistic

\[ W' = \sum_{i=1}^{N} \frac{(\hat{\sigma}_i^2 - \bar{\sigma}^2)^2}{V_i} \sim \chi_N^2 \]

where

\[ V_i = \frac{1}{T-1} \sum_{t=1}^{T} \left( \hat{\varepsilon}_{it}^2 - \hat{\sigma}_i^2 \right)^2 \]

\[ \hat{\sigma}_i^2 = \frac{1}{T} \sum_{t=1}^{T} \hat{\varepsilon}_{it}^2 \]

- Notes
  - Valid in presence of non-normality
  - Lower power in ‘large \( N \), small \( T \)’ FE models

- Solution: FGLS or robust standard errors
- Stata: -xttest3-
Hausman test: FE vs. RE ($H_o : \beta_{FE} = \beta_{RE}$)

- Intuition
  - If $\text{Cov}(c_i, x_{it}) = 0$, then RE and FE are both consistent, but RE is more efficient $\Rightarrow \hat{\beta}_{RE} \approx \hat{\beta}_{FE}$
  - If $\text{Cov}(c_i, x_{it}) \neq 0$, then RE is inconsistent, but FE is consistent $\Rightarrow \hat{\beta}_{RE} \neq \hat{\beta}_{FE}$

- Test statistic based on difference $\hat{\beta}_{FE} - \hat{\beta}_{RE}$

\[
H = T \left( \hat{\beta}_{FE} - \hat{\beta}_{RE} \right)' \left( \Sigma_{FE} - \Sigma_{RE} \right)^{-1} \left( \hat{\beta}_{FE} - \hat{\beta}_{RE} \right) \sim \chi^2_K
\]

- If test statistic is too large, then reject $\text{Cov}(c_i, x_{it}) = 0$

- Issues:
  1. Does not perform well in small samples
  2. $\Sigma_{FE} - \Sigma_{RE}$ need not be PD

- Stata: -hausman-
Alternative test: FE vs. RE

- Mundlak (1978) approach offers an alternative approach to FE; referred to as *correlated random effects* (CRE) approach as dependence between $c$ and $x$ is directly modeled
- Structural model

$$y_{it} = c_i + x_{it} \beta + \varepsilon_{it}$$
$$E[c_i|x_{it}] = \bar{x}_i \gamma$$

where we can write $c_i = \bar{x}_i \gamma + a_i$ in error form

★ Substitution implies

$$y_{it} = x_{it} \beta + \bar{x}_i \gamma + \{c_i - E[c_i|x_{it}]\} + \varepsilon_{it}$$
$$= x_{it} \beta + \bar{x}_i \gamma + a_i + \varepsilon_{it}$$

which is estimable by RE since $E[a_i + \varepsilon_{it}|x_i] = 0$

★ Time invariant regressors do not drop out, but coeff is $\beta + \gamma$

★ An equivalent Hausman test of FE vs. RE is given by $H_0 : \gamma = 0$

- Approach is very useful in panel data LDV models (discussed later)
- Alternative to this alternative: replace $\bar{x}_i$ with $\{x_{is}\}_{s=1}^T$ (i.e., $\bar{x}_i \gamma$ becomes $\sum_s x_{is} \gamma_s$) [Chamberlain approach]
Panel Data
LSDV

- Rather than eliminate $c_i$, treats $c_i$ as a parameter to be estimated.
- Not included as part of the error term.
- Assumptions are identical to (FE.i) – (FE.iii).
- Estimation

$$y_{it} = \sum_{j=1}^{N} c_j D_{ji} + x_{it}\beta + \epsilon_{it}$$

where

$$D_{ji} = \begin{cases} 
1 & \text{if } j = i \\
0 & \text{otherwise}
\end{cases}$$

- Amounts to including $N$ dummy vars (and no constant), 1 for each unit of observation.
- Estimated by POLS.
\( \hat{\beta}_{LSDV} \) is consistent even if \( \text{Cov}(c_i, x_{it}) \neq 0 \) (regressors can always be correlated)

- Time invariant regressors still drop out since they are perfectly collinear with the observation dummies

\( \hat{c}_{i,LSDV} \) is unbiased (and \( = \hat{c}_{i,FE} \)), but consistency requires \( T \to \infty \)

- Only feasible computationally if \( N \) is of reasonable size

- Stata: -areg- (or use Stata’s factor vars syntax)
Panel Data
First-Differencing

- Alternative transformation than FE to eliminate $c_i$
- Assumptions
  
  (FD.i) (FE.i)
  
  (FD.ii) Full rank: $\sum_{t=2}^{T} E[\Delta x'_{it} \Delta x_{it}] = K$
  where $\Delta$ represents the change from the preceding year
  
  (FD.iii) Conditional homoskedasticity and zero covariances:
  $E[\Delta \varepsilon_i \Delta \varepsilon_i' | x_1, ..., x_T, c_i] = \sigma_{\Delta \varepsilon}^2 I_{T-1}$
  
  - Final assumption holds if $\varepsilon_{it}$ follows a random walk

$$\varepsilon_{it} = \varepsilon_{it-1} + \eta_{it}$$

where $E[\eta_i \eta_i' | x_1, ..., x_T, c_i] = \sigma_{\eta}^2 I_{T-1}$, or if $\varepsilon_{it}$ is white noise
Estimation

- Structural model

\[ y_{it} = c_i + x_{it} \beta + \varepsilon_{it} \]

- Implies

\[ y_{i1} = c_i + x_{i1} \beta + \varepsilon_{i1} \]

\[ \vdots \]

\[ y_{iT} = c_i + x_{iT} \beta + \varepsilon_{iT} \]
Taking differences between consecutive periods yields

\[ y_{i2} - y_{i1} = (x_{i2} - x_{i1})\beta + (\varepsilon_{i2} - \varepsilon_{i1}) \]

\[ \vdots \]

\[ y_{iT} - y_{iT-1} = (x_{iT} - x_{iT-1})\beta + (\varepsilon_{iT} - \varepsilon_{iT-1}) \]

or,

\[ \Delta y_{i2} = \Delta x_{i2} \beta + \Delta \varepsilon_{i2} \]

\[ \vdots \]

\[ \Delta y_{iT} = \Delta x_{iT} \beta + \Delta \varepsilon_{iT} \]

Estimating equation

\[ \Delta y_{it} = \Delta x_{it} \beta + \Delta \varepsilon_{it}, \ i = 1, \ldots, N; \ t = 2, \ldots, T \]

which is estimable by POLS

Inference does not require dof adjustment since estimation based on only \( N(T - 1) \) observations

Cluster-robust std errors obtained in a similar fashion as in FE

Stata: -reg- and the time series operators
Notes

- Interpretation of $\hat{\beta}_{FD}$ is same as original $\beta$
- Differencing eliminates $c_i$ and any time invariant $x$'s
- Estimates of unobserved effects are recoverable from FD estimation

$$\hat{c}_{i,FD} = \bar{y}_i - \bar{x}_i \hat{\beta}_{FD}$$

but consistency requires $T \to \infty$

Testing for serial correlation

- Estimate FD model $\Rightarrow \hat{\Delta \epsilon}_{it}$
- Estimate via OLS

$$\hat{\Delta \epsilon}_{it} = \rho \hat{\Delta \epsilon}_{it-1} + v_{it}, \quad t = 3, ..., T; \quad i = 1, ..., N$$

- If $\epsilon_{it}$ is serially uncorrelated, then $\Delta \epsilon_{it}$ will be serially correlated with $\text{Corr}(\Delta \epsilon_{it}, \Delta \epsilon_{it-1}) = -0.5$
Panel Data
Comparison of Estimators

- **POLLS vs RE vs FE/FD/LSDV**
  - POLLS/RE requires $E[x'_{it} c_i] = 0 \forall t$, FE/FD/LSDV do not
  - RE/FE/FD/LSDV requires $E[x'_{is} \epsilon_{it}] = 0 \forall s$, POLLS does not
    $\Rightarrow$ Little apparent advantage to RE

- $T = 2 \Rightarrow$ LSDV, FE, and FD are identical

- $T > 2 \Rightarrow$ FD differs, but all unbiased
  - Under (FE.i) – (FE.iii), FE is efficient
  - Under (FD.i) – (FD.iii), FD is efficient
    - Note: FE.iii assumes errors are iid in levels, FD.iii assumes errors are iid in differences (and hence may be serially correlated in levels) ... very different assumptions

- Long differences (LD) used in some instances; less efficient, but may be less susceptible to mis-specification bias

- Mis-specification may be indicated if FE, FD, and LD yield large discrepancies (ME, timing issues)
Panel Data

Strict Exogeneity

- Assumption of strict exogeneity is crucial in FE, FD models
- Specification test
  - FE model
    - Estimate via FE
      \[ y_{it} = c_i + x_{it} \beta + z_{it+1} \delta + \varepsilon_{it} \]
      where \( z \subseteq x \)
    - \( H_o : \text{strict exogeneity} \iff H_o : \delta = 0 \)
  - FD model
    - Estimate via POLS
      \[ \Delta y_{it} = \Delta x_{it} \beta + z_{it} \delta + \Delta \varepsilon_{it} \]
      where \( z \subseteq x \)
    - \( H_o : \text{strict exogeneity} \iff H_o : \delta = 0 \)
Panel Data

Goodness of Fit

- Three measures of goodness of fit are commonly reported:
  
  Within $R^2$ : $\text{Corr}[(y_{it} - \bar{y}_i), (x_{it}\hat{\beta} - \bar{x}_i\hat{\beta})]^2$
  
  Between $R^2$ : $\text{Corr}[\bar{y}_i, \bar{x}_i\hat{\beta}]^2$
  
  Overall $R^2$ : $\text{Corr}[y_{it}, x_{it}\hat{\beta}]^2$

- Each measure is based on the squared correlation between the actual and fitted values.
Panel Data

Long Panels

- Long panels entail $T > N$ and asymptotics rely on $T \to \infty$ for fixed $N$
- Estimation can easily be handled using the LSDV approach
- Thus, focus is on specification of the form of $\Omega \Rightarrow$ POLS or PFGLS
  - Arbitrary serial correlation in errors is not possible $\Rightarrow$ need to specify the form of $\Omega$
  - In addition, one can relax the assumption of independence across $i$
- Stata: -xtpcse-, -xtgls-, -xtsc-
Table 1: Selection of Stata commands and options that produce robust standard error estimates for linear panel models

<table>
<thead>
<tr>
<th>Command</th>
<th>Option</th>
<th>SE estimates are robust to disturbances that are</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>reg, xtregr</td>
<td>vce(robust)</td>
<td>heteroskedastic</td>
<td></td>
</tr>
<tr>
<td>reg, xtregr</td>
<td>cluster()</td>
<td>heteroskedastic and autocorrelated</td>
<td></td>
</tr>
<tr>
<td>xtregr</td>
<td></td>
<td>autocorrelated with AR(1)(^a)</td>
<td></td>
</tr>
<tr>
<td>newey</td>
<td></td>
<td>heteroskedastic and autocorrelated of type MA(q)(^b)</td>
<td></td>
</tr>
<tr>
<td>xtgls</td>
<td>panels(),</td>
<td>heteroskedastic, contemporaneously cross-sectionally correlated, and autocorrelated of type AR(1)</td>
<td>$N &lt; T$ required for feasibility; tends to produce optimistic SE estimates</td>
</tr>
<tr>
<td></td>
<td>corr()</td>
<td></td>
<td></td>
</tr>
<tr>
<td>xtpcse</td>
<td>correlation()</td>
<td>heteroskedastic, contemporaneously cross-sectionally correlated, and autocorrelated of type AR(1)</td>
<td>large-scale panel regressions with xtpcse take a lot of time</td>
</tr>
<tr>
<td>xtscc</td>
<td></td>
<td>heteroskedastic, autocorrelated with MA(q), and cross-sectionally dependent</td>
<td></td>
</tr>
</tbody>
</table>
Panel Data

Measurement Error

- Panel introduces several complications as it relates to measurement error in covariates
  1. In POLS, ME and OVB may partially offset
  2. In FE and FD,
     - bias from ME may be large if covariates are highly persistent
     - bias from ME affects estimators differently
  3. ME may be nonclassical in that it is also persistent
- Do not ignore ME in panel models
Model setup

\[ y_{it} = c_i + \beta x^*_{it} + \varepsilon_{it} \]
\[ x_{it} = x^*_{it} + v_{it} \]

Assuming classical ME and \( x^* \) to be stationary, one can show

- For POLS

\[
\text{plim } \hat{\beta} = \beta \left[ 1 - \frac{\sigma^2_v}{\sigma^2_x} \right] + \frac{\sigma_{xc}}{\sigma^2_x}
\]

\( \{\text{Attenuation Bias} \} + \{\text{OV Bias} \} \)

- If \( \text{sgn}(\beta) = \text{sgn}(\sigma_{xc}) \), then attenuation and omitted variable bias at least partially offset
Assuming classical ME and $x^*$ to be stationary, one can show (McKinnish 2008)

For differenced estimators

$$\text{plim} \hat{\beta} = \beta \left[ \frac{\sigma_{x^*}^2(1 - \rho_j)}{\sigma_{x^*}^2(1 - \rho_j) + \sigma_v^2} \right]$$

where $\rho_j = \text{Corr}(x_{it}^*, x_{it-j}^*)$

For FE estimator

$$\text{plim} \hat{\beta} = \beta \left\{ \frac{\sigma_{x^*}^2 - \frac{1}{T^2} \left[ T + 2 \sum_{j=1}^{T-1} (T - j - 1) \rho_j \right]}{\sigma_{x^*}^2 - \frac{1}{T^2} \left[ T + 2 \sum_{j=1}^{T-1} (T - j - 1) \rho_j \right] + \left( \frac{T-1}{T} \right) \sigma_v^2} \right\}$$
Griliches & Hausman (1986) verify that if $x_{it}^*$ has a declining correlogram, then $\rho_1 > \rho_2 > ...$ and

- FD has the largest bias
- Longer differences reduce the attenuation bias
- $T - 1$ long difference has the smallest bias
- Relative bias of long differences shorter than $T - 1$ and the FE estimator is ambiguous, depends on specifics of the correlation structure
- Relative bias of POLS and FE/FD/LD is unknown
- Longer differences result in a loss of dof, efficiency loss
Serial correlation in ME

\[ y_{it} = c_i + \beta x_{it} + \varepsilon_{it} \]
\[ x_{it} = x_{it}^* + \nu_{it} \]
\[ x_{it}^* = \rho x_{it-1}^* + \eta_{it} \]
\[ \nu_{it} = \delta \nu_{it-1} + \zeta_{it} \]

where \( \rho \) is the degree of persistence in \( x^* \) and \( \delta \) is the degree of persistence in \( \nu \)

- For differenced estimators

\[
\text{plim} \hat{\beta} = \beta \left[ \frac{\sigma^2_{x^*}(1 - \rho_j)}{\sigma^2_{x^*}(1 - \rho_j) + \sigma^2_\nu(1 - \delta_j)} \right]
\]

where \( \rho_j = \text{Corr}(x_{it}^*, x_{it-j}^*) \) and \( \delta_j = \text{Corr}(\nu_{it}, \nu_{it-j}) \)

- If \( \delta < \rho \), then the pattern for bias remains the same: longer differences \( \Rightarrow \) smaller bias
- If \( \delta > \rho \), then the pattern for bias reverses: FD \( \Rightarrow \) smallest bias
Consistent estimation in the presence of ME requires IV

- Assuming
  - No serial correlation in the ME,
  - Strict exogeneity of $x_{it}$ and $x_{it}^{*}$ conditional on $c_i$ with respect to $\varepsilon_{it}$, and
  - Strict exogeneity of $x_{it}^{*}$ conditional on $c_i$ with respect to $\upsilon_{it}$,

we can substitute for $x_{it}^{*}$ and FD to eliminate $c_i$:

$$\Delta y_{it} = \Delta x_{it} \beta + \Delta \varepsilon_{it} + \Delta \upsilon_{it}$$

- Possible instruments for $\Delta x_{it}$ include $x_{it-2}$ or $x_{it+1}$ or external instruments
Panel Data

Contemporaneous Correlation Between Covariates and the Error

- Measurement error or omitted variables can lead to contemporaneous correlation between $x$ and $\varepsilon$
- If $c$ is also correlated with $x$, begin by first-differencing to eliminate $c$

$$\Delta y_{it} = \Delta x_{it} \beta + \Delta \varepsilon_{it}$$

- POLS applied to this eqtn is biased if $\text{Corr}(\Delta x, \Delta \varepsilon) \neq 0$
- IV methods remain consistent; in addition to finding typical instruments, $z_{it}$, from outside the model, other possibilities include
  - $x_{it-2}$, $x_{it+1}$
  - $y_{it-2}$

- While popular, such instruments may be weak, invalid in the presence of serial correlation, or hard to justify (Reed 2015)
- Stata: -xtivreg2-
Panel data can lead to other ‘tricks’ that may generate instruments

Example: Pitt and Rosenzweig (1990) siblings FE model

- Sample restricted to hh’s with 1 boy, 1 girl
- Interested in effect of an endogenous hh level variable, $D_h$
- No outside instruments available
- Solution

\[
y_{bh} = x_{1h} \beta_b + x_{2h} \gamma + \tau_b D_h + \varepsilon_{bh} \\
y_{gh} = x_{1h} \beta_g + x_{2h} \gamma + \tau_g D_h + \varepsilon_{gh}
\]

\[\Rightarrow \Delta y_h = x_{1h} \Delta \beta + \Delta \tau D_h + \Delta \varepsilon_h\]

and now $x_{2h}$ are available as instruments for $D_h$, but model only identifies $\Delta \tau$ (the differential impact of $D$ on boys relative to girls)
Panel Data

Time Invariant Regressors

- FE, FD estimators preclude estimation of coeffs on time invariant regressors
- One solution is to use a two-step estimator
  - Model setup
    \[ y_{it} = c_i + x_{it}\beta + z_i\alpha + \epsilon_{it} \]
    where \( z_i \) is a vector of time invariant regressors
  - Estimation
    - Step #1: Estimate via FE or FD
      \[ y_{it} = \tilde{c}_i + x_{it}\beta + \epsilon_{it} \]
      yielding a consistent estimate of \( \beta \) and an unbiased estimate of \( \tilde{c}_i \)
    - Step #2: Estimate via OLS (or WLS)
      \[ \tilde{c}_i = \alpha_0 + z_i\alpha + v_i \]
      (Hornstein & Greene 2012)
  - \( \hat{\alpha} \) is consistent if \( E[v|z] = E[v] \); strong assumption on \( z \)
**Alternative solution:** Hausman & Taylor (1981)

- Re-write above model as

  \[ y_{it} = c_i + x_{1it}\beta_1 + x_{2it}\beta_2 + z_{1i}\alpha_1 + z_{2i}\alpha_2 + \epsilon_{it} \]

  such that

  - \( x_{1it} = \) vector of \( K_1 \) vars *uncorrelated* with \( c_i \)
  - \( x_{2it} = \) vector of \( K_2 \) vars *correlated* with \( c_i \)
  - \( z_{1i} = \) vector of \( L_1 \) vars *uncorrelated* with \( c_i \)
  - \( z_{2i} = \) vector of \( L_2 \) vars *correlated* with \( c_i \)

- Assumption about the unobservables

  \[
  \begin{align*}
  \mathbb{E}[c_i|x_{1it}, z_{1i}] &= 0; \quad \mathbb{E}[c_i|x_{2it}, z_{2i}] \neq 0 \\
  \text{Var}(c_i|x_{1it}, z_{1i}, x_{2it}, z_{2i}) &= \sigma_c^2 \\
  \text{Cov}(\epsilon_{it}, c_i|x_{1it}, z_{1i}, x_{2it}, z_{2i}) &= 0 \\
  \text{Var}(\epsilon_{it} + c_i|x_{1it}, z_{1i}, x_{2it}, z_{2i}) &= \sigma^2 = \sigma_{\epsilon}^2 + \sigma_c^2 \\
  \text{Corr}(\epsilon_{it} + c_i, \epsilon_{is} + c_i|x_{1it}, z_{1i}, x_{2it}, z_{2i}) &= \rho = \sigma_c^2 / \sigma^2
  \end{align*}
  \]
Estimation

- POLS or GLS is biased and inconsistent
- FE or FD precludes estimation of $\alpha$
- Solution is to use an IV estimator based only on the data at hand
  - Mean-differencing yields
    \[
    \dot{y}_{it} = \dot{x}_{1it}\beta_1 + \dot{x}_{2it}\beta_2 + \epsilon_{it}
    \]
    which could be estimated
  - This shows that $\dot{x}_1$ and $\dot{x}_2$ are uncorrelated with $c$ and are thus available as instruments for the eqtn in *levels*
  - $z_1$ serves as instruments for itself
  - $\bar{x}_1$ serves as instruments for $z_2$
  - $\bar{x}$ serve as instruments for $x$

- Identification requires $K_1 \geq L_2$
- IV regression must account for the treatment of $c$ as a RE $\Rightarrow$ IV-FGLS solution
- Weak IVs may still be a problem

Stata: -xthtaylor-
Panel Data
Dynamic Panel Model (DPD)

- Structural model

\[ y_{it} = c_i + x_{it}\beta + \gamma y_{it-1} + \varepsilon_{it}, \quad |\gamma| < 1 \]

where \( i = 1, \ldots, N; \ t = 2, \ldots, T; \) and \( T \geq 3 \)

- Assumptions

  (DPD.i) (FD.i) Strict exogeneity
  (DPD.ii) (FD.ii) Full rank
  (DPD.iii) (FD.iii) Conditional homoskedasticity and zero covariances
  (DPD.iv) Initial condition: \( E[y_{i1}\varepsilon_{it}] = 0 \ \forall t > 1 \)

- Interpretation

  - \( \beta = \) short-run (one-period) effect of a unit change in \( x \)
  - \( \beta / (1 - \gamma) = \) long-run effect of a unit change in \( x \)
Comments

- Even if $\text{Cov}(c_i, x_{it}) = 0$, POLS/RE not applicable since $\text{Cov}(c_i, y_{it-1}) \neq 0$
- Even if $\text{Cov}(c_i, x_{it}) = 0$, FE/FD not applicable since $E[y_{it-1} \varepsilon_{it-1}] \neq 0 \Rightarrow y_{it-1}$ is not strictly exogenous
  - Fixed Effects
    - Nickel (1981) derived the formula for the bias of $\hat{\gamma}_{FE}$
    - Bias $\to 0$ as $T \to \infty$ since $\text{Cov}(\hat{y}_{it-1}, \hat{\varepsilon}_{it}) \to 0$ as $T \to \infty$
  - FD
    - Model
      \[
      \Delta y_{it} = \Delta x_{it} \beta + \gamma \Delta y_{it-1} + \Delta \varepsilon_{it}, \ i = 1, ..., N; \ t = 3, ..., T
      \]
      which is not estimable by POLS since $\text{Cov}(\Delta y_{it-1}, \Delta \varepsilon_{it}) \neq 0$ since $\text{Cov}(y_{it-1}, \varepsilon_{it-1}) \neq 0$
    - Bias $\to 0$ as $T \to \infty$
- For both FE/FD, bias $\to 0$ as $N \to \infty \Rightarrow$ FE/FD are inconsistent under large $N$ asymptotics
- If $\text{Cov}(x_{it}, c_{i}) = \text{Cov}(x_{it}, \varepsilon_{it}) = 0 \ \forall x$, then
  \[ \hat{\gamma}_{POLS} > \gamma > \hat{\gamma}_{FE} > \hat{\gamma}_{FD} \]
  since
  \[ \text{Cov}(y_{it-1}, c_{i} + \varepsilon_{it}) > 0 > \text{Cov}(\bar{y}_{it-1}, \bar{\varepsilon}_{it}) > \text{Cov}(\Delta y_{it-1}, \Delta \varepsilon_{it}) \]
- Thus, estimators *bound* true value
- Not overly useful though since we usually are interested in $\beta$ as well
Solution #1 (Anderson & Hsiao 1981, 1982)

- FD, then estimate via instrumental variables (IV), treating $\Delta y_{it-1}$ as endogenous

- Instruments: $y_{it-2}$, $x_{it-2}$
  - $\text{Cov}(\Delta y_{it-1}, y_{it-2}) = \text{Cov}(y_{it-1}, y_{it-2}) - \text{Cov}(y_{it-2}, y_{it-2}) \neq 0$
  - $\text{Cov}(\Delta y_{it-1}, x_{it-2}) = \text{Cov}(y_{it-1}, x_{it-2}) - \text{Cov}(y_{it-2}, x_{it-2}) \neq 0$
  - $\text{Cov}(\Delta \varepsilon_{it}, y_{it-2}) = \text{Cov}(\varepsilon_{it}, y_{it-2}) - \text{Cov}(\varepsilon_{it-1}, y_{it-2}) = 0$ if $\varepsilon$ is not serially correlated
  - $\text{Cov}(\Delta \varepsilon_{it}, x_{it-2}) = \text{Cov}(\varepsilon_{it}, x_{it-2}) - \text{Cov}(\varepsilon_{it-1}, x_{it-2}) = 0$ if $x$ is predetermined (weaker than strictly exogenous)

- Model is overidentified

- Anderson & Hsiao (1981) also suggest $\Delta y_{it-3}$ as an instrument

- $\hat{\gamma}_{IV}$ is consistent as $N \to \infty$ for fixed $T$ as long as $T \geq 3$ under (DPD.i) – (DPD.iv)

- Note: As $\gamma \to 1$, $y_{it-2}$, $x_{it-2}$ become very weak instruments
Solution #2 (Arellano & Bond 1991)

- FD, then estimate via IV, treating $\Delta y_{it-1}$ as endogenous as above
- However, there are LOTS of potential instruments not considered in A&H method (if $T \geq 3$)
- What are potential IVs?
  - Need vars that are correlated with $\Delta y_{it-1}$, uncorrelated with $\Delta \epsilon_{it}$
  - Suitable candidates
    - $x_{it-2}$ (through $y_{it-2}$)
    - $y_{it-2}$ (through $y_{it-2}$)
    - $y_{it-3}, y_{it-4}, y_{it-5}, \ldots$ (through autoregressive process) ... e.g.,
      \[
      \text{Cov}(\Delta y_{it-1}, y_{it-3}) = \text{Cov}(y_{it-1}, y_{it-3}) - \text{Cov}(y_{it-2}, y_{it-1}) \neq 0
      \]
  - Lots of instruments (beware of weak IVs)
  - Not standard IV/TLSLS setup since # of instruments varies by observation (by $t$)
Estimation by GMM to utilize more instruments

- Efficient given (DPD.ii) – (DPD.iv)
- Writing out model for each period yields

\[
\Delta y_{i3} = \Delta x_{i3}\beta + \gamma \Delta y_{i2} + \Delta \varepsilon_{i3} \\
\Delta y_{i4} = \Delta x_{i4}\beta + \gamma \Delta y_{i3} + \Delta \varepsilon_{i4} \\
\vdots \\
\Delta y_{iT} = \Delta x_{iT}\beta + \gamma \Delta y_{iT-1} + \Delta \varepsilon_{iT}
\]

where IVs for \(\Delta y_{i2}\) are \(x_{i1}, y_{i1}\); IVs for \(\Delta y_{i3}\) are \(x_{i2}, y_{i2}, y_{i1}\); ... ; IVs for \(\Delta y_{iT-1}\) are \(x_{iT-2}, y_{iT-2}, \ldots, y_{i2}, y_{i1}\)

- GMM allows moment conditions to be derived using as many IVs as desired
- Requires \(\varepsilon_{it}\) to be serially uncorrelated; or, equivalently, \(\Delta \varepsilon_{it}\) should be \(AR(1)\)
ASIDE: GMM/MDE details

- MLE is efficient among CAN estimators \textit{in the context of the specified parametric model}
- GMM estimators are robust to some variations in the underlying DGP; less efficient than MLE if parametric assumptions hold
- Both are examples of M-estimators
- Method of Moments (MM) intuition
  - With random sampling, a sample statistic – moment – converges in probability to a constant
  - This constant is a fn of unknown population parameters
  - Thus, to estimate $K$ parameters ($\theta_k, k = 1, \ldots, K$) requires $K$ statistics with known plims
  - Equating the $K$ moments to the $K$ plims, these are inverted to solve for $\theta$ as fns of the moments
  - Moments are CAN $\Rightarrow$ estimator of $\theta$ will also be CAN
Example: $y_i \sim \mathcal{N}(\mu, \sigma^2)$

- Given a random sample, two moment conditions given by

$$
\bar{m}_1 = \text{plim} \frac{1}{N} \sum_i y_i = \mu \\
\bar{m}_2 = \text{plim} \frac{1}{N} \sum_i y_i^2 = \sigma^2 + \mu^2
$$

- Inversion yields

$$
\hat{\mu} = \bar{m}_1 = \bar{y} \\
\hat{\sigma}^2 = \bar{m}_2 - \bar{m}_1^2 = \left( \frac{1}{N} \sum_i y_i^2 \right) - \left( \frac{1}{N} \sum_i y_i \right)^2 = \frac{1}{N} \sum_i (y_i - \bar{y})^2
$$
General case for MM estimator

- **K** moment conditions

\[
\overline{m}_1 = \mu_1(\theta_1, \ldots, \theta_K) \\
\vdots \\
\overline{m}_K = \mu_K(\theta_1, \ldots, \theta_K)
\]

- Assuming the **K** eqtns are functionally independent, inversion yields

\[
\hat{\theta}_1 = \hat{\theta}_1(\overline{m}_1, \ldots, \overline{m}_K) \\
\vdots \\
\hat{\theta}_K = \hat{\theta}_K(\overline{m}_1, \ldots, \overline{m}_K)
\]

- Except for the case of random sampling from the exponential family of dbns, $\hat{\theta}_{MM}$ is not efficient
Asymptotic properties

\[ \text{Asy.Var}(\hat{\theta}) = \frac{1}{N} \left[ \overline{G}'(\hat{\theta}) \Phi^{-1} \overline{G}(\hat{\theta}) \right]^{-1} \]

where \( \overline{G}(\theta) \) is a \( K \times K \) matrix with row \( k \) given by

\[ \overline{G}(\theta) = \frac{\partial \overline{m}_k}{\partial \theta'} \]

and \( \Phi \) is the \( K \times K \) estimated asymptotic covariance matrix of the moment conditions, given by

\[ \frac{1}{N} \Phi_{jk} = \frac{1}{N} \left\{ \frac{1}{N} \sum_i [(m_j(y_i) - \overline{m}_j)(m_k(y_i) - \overline{m}_k)] \right\}, \quad j, k = 1, \ldots, K \]

and the moment conditions are expressed as

\[ \overline{m}_k = \frac{1}{N} \sum_i m_k(y_i) \]
Estimation based on orthogonality conditions

- CLRM given by
  \[ y_i = x_i \beta + \varepsilon_i \]
  where \( E[x_i \varepsilon_i] = 0 \)

  - Condition implies \( E[x_i (y_i - x_i \beta)] = 0 \)
  - Sample analog given by
    \[ \frac{1}{N} \sum_i x_i (y_i - x_i \hat{\beta}) = 0 \]
    which is a set of \( K \) moment conditions as a fn of \( K \) parameters
  - These moment conditions are identical to the LS normal eqtns
  - Thus, OLS is a MM estimator

- IV estimator solves
  \[ \frac{1}{N} \sum_i z_i (y_i - x_i \hat{\beta}) = 0 \]

- TSLS estimator solves
  \[ \frac{1}{N} \sum_i \hat{x}_i (y_i - x_i \hat{\beta}) = 0 \]
- ML can also be recast as a MM estimator
  - The scaled log-likelihood fn is
    $$\frac{1}{N} \ln \left[ \mathcal{L}(\theta) \right] = \frac{1}{N} \sum_i \ln[f(y_i|x_i, \theta)]$$
  - MLE solves
    $$\frac{1}{N} \partial \ln \left[ \mathcal{L}(\hat{\theta}|y) \right] = \frac{1}{N} \sum_i \partial \ln[f(y_i|x_i, \hat{\theta})] = 0$$
- Many estimators can be viewed as a MM estimator
Minimum Distance Estimation (MDE) extends the MM estimator to the case of over-identified models

### Setup

\[ m_1 = \mu_1(\theta_1, ..., \theta_K) \]
\[ \vdots \]
\[ m_L = \mu_L(\theta_1, ..., \theta_K) \]

where \( L \geq K \)

### MDE estimator obtained as

\[
\hat{\theta}_{MDE} = \arg \min_{\theta} \left[ m - \mu(\theta) \right]' W \left[ m - \mu(\theta) \right]
\]

where \( m \) is the vector of moment conditions, \( \mu(\theta) \) is the vector of plims, and \( W \) is a PD weighting matrix
Example:

- Two consistent estimates of the same parameter, $\theta$

$$\overline{m}_1 = \text{plim} \hat{\theta}_1 = \theta$$
$$\overline{m}_2 = \text{plim} \hat{\theta}_2 = \theta$$

- Estimator is

$$\hat{\theta}_{MDE} = \arg \min_{\theta} \left[ \begin{array}{c} \hat{\theta}_1 - \theta \\ \hat{\theta}_2 - \theta \end{array} \right]^T W \left[ \begin{array}{c} \hat{\theta}_1 - \theta \\ \hat{\theta}_2 - \theta \end{array} \right]$$

where $W$ might be the covariance matrix of $\hat{\theta}_1, \hat{\theta}_2$
Asymptotic properties: CAN with

\[
\text{Asy.Var}(\hat{\theta}) = \frac{1}{N} \left[ \overline{G}'(\theta) W \overline{G}(\theta) \right]^{-1} \left[ \overline{G}'(\theta) W \Phi W \overline{G}(\theta) \right] \left[ \overline{G}'(\theta) W \overline{G}(\theta) \right]^{-1}
\]

where \( \overline{G}(\theta) \) is a \( L \times K \) matrix with row \( \ell \) given by

\[
\overline{G}(\theta) = \text{plim} \frac{\partial m_{\ell}}{\partial \theta'}
\]

and \( \Phi \) is the \( L \times L \) estimated asymptotic covariance matrix of the moment conditions.

Optimal weight matrix minimizes the asymptotic variance: \( W = \Phi^{-1} \)

\[
\text{Asy.Var}(\hat{\theta}) = \frac{1}{N} \left[ \overline{G}'(\theta) \Phi^{-1} \overline{G}(\theta) \right]^{-1}
\]

which is equivalent to the MM estimator.
Generalized Method of Moments (GMM) extends the MDE

- Estimator is based on set of population orthogonality conditions
  \[ E[m(y_i, x_i, \theta_0)] = 0 \]
  where \( \theta_0 \) is a \( K \)-dimensional vector of true values and \( m(\cdot) \) is a \( L \)-dimensional vector \((L > K)\)
- Averaging over obs produces the sample moment conditions
  \[ E[\overline{m}(y_i, x_i, \theta_0)] = 0 \]
  where
  \[ \overline{m}(y_i, x_i, \theta_0) = \frac{1}{N} \sum_i m(y_i, x_i, \theta_0) \]
- GMM estimator obtained as
  \[ \hat{\theta}_{GMM} = \arg \min_{\theta} \overline{m}(y_i, x_i, \theta)' W \overline{m}(y_i, x_i, \theta) \]
  where \( W \) is a PD weighting matrix
- Optimal weight matrix minimizes the asymptotic variance: \( W = \Phi^{-1} \)

\[ \text{Asy.Var}(\hat{\theta}) = \frac{1}{N} \left[ \overline{G}'(\theta) \Phi^{-1} \overline{G}(\theta) \right]^{-1} \]

which is equivalent to the MDE estimator
Example:

- IV estimator solves
  \[
  \frac{1}{N} \sum_i z_i (y_i - x_i \hat{\beta}) = 0
  \]
  which yields \( K \) moments as fn of \( K \) parameters

- TSLS estimator can be written as
  \[
  \frac{1}{N} \sum_i z_i (y_i - x_i \hat{\beta}) = 0
  \]
  where now there are \( L \) moments as fn of \( K \) parameters
Inference in GMM

- Wald tests can be conducted as in ML
- LR tests can be conducted based on the minimized GMM criteria

\[ N(q_R - q) \sim \chi^2_J \]

where \( q \) (\( q_R \)) is the value of the criterion in the unrestricted (restricted) model and \( J = \# \) of restrictions

- LM tests can be conducted based on the derivatives of the unrestricted GMM criterion evaluated at the restricted estimator

See -gmm- in Stata
RETURNING: DPD Estimation Specifics

- Recall, the model

$$y_{it} = c_i + x_{it}\beta + \gamma y_{it-1} + \varepsilon_{it}, \quad t = 2, \ldots, T$$

- FD to eliminate $c_i$ yields

$$\Delta y_{i3} = \Delta x_{i3}\beta + \gamma \Delta y_{i2} + \Delta \varepsilon_{i3}$$
$$\Delta y_{i4} = \Delta x_{i4}\beta + \gamma \Delta y_{i3} + \Delta \varepsilon_{i4}$$

\vdots

$$\Delta y_{iT} = \Delta x_{iT}\beta + \gamma \Delta y_{iT-1} + \Delta \varepsilon_{iT}$$

where IVs for

- $\Delta y_{i2}$ are $x_{i1}, y_{i1}$
- $\Delta y_{i3}$ are $x_{i2}, y_{i2}, y_{i1}$
  \vdots
- $\Delta y_{iT-1}$ are $x_{iT-2}, y_{iT-2}, \ldots, y_{i2}, y_{i1}$
Define matrix of instruments (assuming no $x$’s)

\[
\begin{bmatrix}
  y_{i1} & 0 & 0 & \cdots & 0 & \cdots & 0 \\
  0 & y_{i1} & y_{i2} & \cdots & 0 & \cdots & 0 \\
  \vdots & 0 & 0 & \cdots & 0 & \cdots & 0 \\
  \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\
  0 & 0 & 0 & \cdots & y_{i1} & \cdots & y_{iT-2}
\end{bmatrix}
\]

 Moment conditions ($p \times 1$)

\[
E[z_i' \Delta \varepsilon_i] = 0
\]

\[
\Delta \varepsilon_i = [\Delta \varepsilon_{i3} \, \Delta \varepsilon_{i4} \, \cdots \, \Delta \varepsilon_{iT}]'
\]
Estimator

\[
\min_{\gamma} \left( \frac{1}{N} \sum_i \Delta \epsilon'_i z_i \right) \mathcal{W}_N \left( \frac{1}{N} \sum_i z'_i \Delta \epsilon_i \right)
\]

where

\[
\mathcal{W}_N = \left[ \left( \frac{1}{N} \sum_i z'_i \Delta \hat{\epsilon}_i \Delta \hat{\epsilon}'_i z_i \right) \right]^{-1}
\]

and \( \Delta \hat{\epsilon}_i \) are estimated residuals from consistent first-step estimates

- Known as two-step differenced GMM
Consistent first-step obtained using

\[ W_{1N} = \left[ \left( \frac{1}{N} \sum_i z_i' H z_i \right) \right]^{-1} \]

where

\[ H_{(T-2) \times (T-2)} = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{bmatrix} \]

- Known as one-step differenced GMM
- Two-step estimator is asymptotically efficient if \( \varepsilon_{it} \) are heteroskedastic; one-step if homoskedastic
- Not much gain in practice to two-step
Specification tests

1. Overidentification test (Sargan test)

\[
\Delta \epsilon' z \left[ \sum_i z_i' \Delta \hat{\epsilon}_i \Delta \hat{\epsilon}_i' z_i \right]^{-1} z' \Delta \epsilon \sim \chi^2_{p-1}
\]

where \( p \) is the \# of columns in \( z \)

2. No serial correlation in \( \epsilon_{it} \) \( \iff \Delta \epsilon_{it} \) is AR(1)

- Adding \( x \)'s into the model requires assumptions about their correlations with the errors
  - Assumptions about \( x \)'s
    - Strict exogeneity: \( E[x_{it} \epsilon_{is}] = 0 \ \forall t, s \)
    - Predetermined (sequential exogeneity): \( E[x_{it} \epsilon_{is}] = 0 \ \forall t \leq s \)
    - Endogenous: \( E[x_{it} \epsilon_{is}] = 0 \ \forall t < s \)
Incorporating \( x \)'s into moment conditions

- Moment conditions \((p \times 1)\)

\[
E[z'_i \Delta \epsilon_i] = 0, \quad \Delta \epsilon_i = [\Delta \epsilon_{i3} \Delta \epsilon_{i4} \cdots \Delta \epsilon_{iT}]'
\]

- \( x \)'s strictly exogenous

\[
z_i = \begin{bmatrix}
  x_{i1} & \cdots & x_{iT} & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\
  0 & \cdots & 0 & x_{i1} & \cdots & x_{iT} & \cdots & \cdots & \cdots & 0 \\
  \vdots & & & \vdots & & & \vdots & & & \vdots \\
  0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & x_{i1} & \cdots & x_{iT}
\end{bmatrix}
\]

where

\[
\tilde{z}_i = \begin{bmatrix}
y_{i1} & 0 & 0 & \cdots & 0 & \cdots & 0 \\
0 & y_{i1} & y_{i2} & \cdots & 0 & \cdots & 0 \\
\vdots & 0 & 0 & \cdots & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\
0 & 0 & 0 & \cdots & y_{i1} & \cdots & y_{iT-2}
\end{bmatrix}
\]
\( x \)'s predetermined

\[
\begin{bmatrix}
  x_{i1} & x_{i2} & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\
  0 & 0 & x_{i1} & x_{i2} & x_{i3} & \cdots & \cdots & \cdots & \cdots & 0 \\
  \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
  0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & x_{i1} & \cdots & x_{iT-1}
\end{bmatrix}
\]
• $x$’s endogenous

$$z_i = \begin{bmatrix}
    x_{i1} & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\
    0 & x_{i1} & x_{i2} & \cdots & \cdots & \cdots & 0 \\
    \vdots & \vdots & \vdots & \ddots & \cdots & \cdots & \vdots \\
    \vdots & \vdots & \vdots & \cdots & \ddots & \cdots & \vdots \\
    0 & \cdots & \cdots & 0 & x_{i1} & \cdots & x_{iT-2}
\end{bmatrix}$$

• Notes:
  ▶ Ahn & Schmidt (1995) utilize $T - 2$ additional moment conditions

$$E[\varepsilon_{it}\Delta\varepsilon_{it-1}] = 0, \quad t = 3, \ldots, T$$

▶ All three versions assume $E[x_{it}c_i] \neq 0$; if $E[x_{it}c_i] = 0$, then there are even more moment conditions

• Stata: -xtabond-, -xtdpd-; -xtdpd- allows for greater user-controlled flexibility
Solution #3 (Arellano & Bover 1995; Blundell & Bond 1998)

- A&B (1995), B&B (1998) show that as $|\gamma| \to 1$, A&H and A&B estimator is downward biased.
- In MC study, B&B find if $|\gamma| > 0.8$, bias is fairly severe.
- Problem arises due to the fact that the IVs become weak as persistence grows.
- Solution
  - Add additional moment conditions derived from the model in levels
    \[
    y_{it} = x_{it}^{\beta} + \gamma y_{it-1} + c_i + \epsilon_{it} \\
    = x_{it}^{\beta} + \gamma y_{it-1} + \tilde{\epsilon}_{it}
    \]
  - What are IVs for $y_{it-1}$?
    - $\Delta y_{it-1}$ (independent of $c_i, \epsilon_{it}$)
    - $\Delta y_{it-2}, \Delta y_{it-3}, \ldots$ (through autoregressive process)
Estimation

- Define

\[ z_{i}^{sys} = \begin{bmatrix}
    z_i & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\
    0 & \Delta y_{i2} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\
    \vdots & \Delta y_{i2} & \Delta y_{i3} & \vdots & \vdots & \vdots & \vdots \\
    \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
    0 & \cdots & \cdots & \cdots & 0 & \Delta y_{i2} & \cdots & \Delta y_{iT-1} \\
\end{bmatrix} \]

where \( z_i \) is the matrix of instruments from A&B

- Complete set of moment conditions

\[
E[z_{i}^{sys} \epsilon_{i}^{sys}] = 0 \\
\epsilon_{i}^{sys} = [\Delta \epsilon_{i3} \Delta \epsilon_{i4} \cdots \Delta \epsilon_{iT} \tilde{\epsilon}_{i3} \cdots \tilde{\epsilon}_{iT}]' 
\]
- Known as **system GMM**
- Requires additional assumption

$\text{(DPD.v)} \quad \mathbb{E}[c_i \Delta y_{i2}] = 0 \quad \forall i$

- **Stata**: `-xtdpdsys-`
Solution #4 (Hahn et al. 2007)

- An alternative approach to circumvent the weak IV issue when $|\gamma| \rightarrow 1$ is to utilize long differences (LD) instead of first-differencing to eliminate $c_i$

- Structural model

$$y_{it} = c_i + x_{it}\beta + \gamma y_{i(t-1)} + \epsilon_{it}, \quad t = 2, ..., T$$

- LD yields

$$y_{iT} - y_{i2} = (x_{iT} - x_{i2})\beta + \gamma(y_{iT-1} - y_{i1}) + \epsilon_{iT} - \epsilon_{i2}$$

where $y_{i1}$ is available as an IV for $y_{iT-1} - y_{i1}$

- Even if $|\gamma|$ is close to unity,

$$\text{Cov}(y_{iT-1}, y_{i1}) - \text{Cov}(y_{i1}, y_{i1}) \neq 0$$

if $T$ is sufficiently large
Hahn et al. (2007) also utilize additional IVs

- After initial IV estimation, obtain estimated residuals
  \[ \tilde{\varepsilon}_{is} \equiv y_{is} - x_{is} \hat{\beta} - \hat{\gamma} y_{is-1}, \ s = 3, \ldots, T - 1 \]
- Re-estimate the LD model using \( y_{i1}, \tilde{\varepsilon}_{is} \) as instruments
- Iterate this procedure; typically only a few iterations are needed

Hahn et al. (2007) demonstrate the superior performance of this estimator when \( |\gamma| \to 1 \)
Solution #5 (Kiviet 1995, 1999; Bun & Kiviet 2003; Bruno 2005; Hausman & Pinkovskiy 2017)

- Revert back to Nickel (1981) who derived bias to FE/LSDV estimate of DPD model
- Strategy
  - Obtain an estimate of the approximate bias of the LSDV estimator
  - Adjust the estimate $\hat{\gamma}_{LSDV}$ by the estimated bias to obtain the LSDVC (corrected) estimate
  - Implies $E[\hat{\gamma}_{LSDV} + BIAS - \gamma] = 0$
- MC evidence indicates good performance for moderate size $N$
- See -xtlsdvc-, -xtbcfe- in Stata
Solution #6 (Everaert 2013)

- Estimation by IV, but in levels using orthogonal to backward mean transformation to instrument for $y_{it-1}$

\[ y_{it} = c_i + x_{it} \beta + \gamma y_{it-1} + \varepsilon_{it}, \quad t = 2, \ldots, T \]

- Instrument for $y_{it-1}$ is the residual from the first-stage regression of $y_{it-1}$ on its ‘backward mean’ defined as

\[ \bar{y}^b_{it-1} = \frac{1}{t-1} \sum_{s=1}^{t-1} y_{is} \]

- If $\text{Cov}(x_{it}, c_i) \neq 0$, then
  - Instruments for $x_{it}$ are derived based on Hausman & Taylor (1981)
  - Specifically, deviations from individual-specific means, $x_{it} - \bar{x}_i$

- Consistency requires $T \to \infty$
Solution #7 (Bhargava and Sargan 1983; Hsiao et al. 2002; Kripfganz 2016)

- QML estimator for large $N$, small $T$ panels
- Based on ML estimator after first-differencing
- Stata: -xtdpdqml-
Final comment

- Preceding estimators require $\varepsilon_{it}$ to be serially uncorrelated
  - Can relax this assumption and allow $\varepsilon$ to follow a low order MA process
  - Stata: -xtdpd-

- Preceding estimators are no longer consistent if the data are \textit{irregularly spaced}
  - Irregular spacing occurs when the interval between periods in the data does not align with the interval between periods in the DGP
  - This can occur even if the data are \textit{evenly spaced}
Panel Data
LDV Models: Overview

• LDV panel models are complicated due to four main issues
  1. Estimation by MLE requires iid obs to compute a tractable likelihood; unlikely with panels
     ★ Solution ⇒ partial maximum likelihood
  2. Transformations to eliminate FEs are generally not available since problems are nonlinear
     ★ Solution (in some models) ⇒ conditional maximum likelihood
  3. DV approach to FE models are generally inconsistent unless $N, T \to \infty$
     ★ Known as the incidental parameters problem
     ★ Solutions ⇒ correlated random effects, bias-correction
  4. Marginal effects generally depend on $c$ which is only estimable if one places more structure on the model

• Research on-going in many areas, many unresolved questions
Panel Data
LDV Models: Binary Choice Models

- Case of a binary dependent variable with panel data
- Often useful to start with a LPM and use the FE or FD estimator
  - Same shortcomings as LPM with cross-section data
    - Heteroskedasticity
    - Predictions outside unit interval
  - Plus, implies restrictions on the feasible values of $c_i$
  - Nonetheless, very common in applied research
Instead, suppose the model is given by

\[ \Pr(y_{it} = 1|x_{it}) = G(x_{it}\beta), \quad t = 1, \ldots, T \]

where \( G(\cdot) \in (0, 1) \) is known and \( x_{it} \) must be exogenous (but not strictly).

- Recall, the likelihood fn gives the total probability of the data conditional on the parameters
  - We simplify this if we assume observations are iid
  - With panel data, we usually do not want to assume this; rather assume independent panels, but not independence across time within a panel
  - So, we want to model the joint \( \Pr(y_i|x_i) \)
  - However, without further assumptions, we cannot derive the dbn of \( y_i \equiv (y_{i1}, \ldots, y_{iT}) \) given \( x_i \equiv (x_{i1}, \ldots, x_{iT}) \)

Could assume

\[ \Pr(y_i = 1|x_i) = G_T(x_i\beta, \Sigma) \]

but then likelihood fn would entail \( T \) integrals if, say, \( G_T \) is a \( T \)-dimensional normal dbn
We can still obtain consistent estimates using the *partial likelihood fn*

\[
\ln[L(\theta)] = \sum_i \sum_t y_{it} \ln[G(x_{it}\beta)] + (1 - y_{it}) \ln[1 - G(x_{it}\beta)]
\]

which is different from ML in that we do not assume that

\[
f(y_i|x_i) = \prod_t f_t(y_{it}|x_{it})
\]

- Known as the PMLE of the *pooled probit/logit model* when \( G = \Phi, \Lambda \) (std normal CDF or logistic CDF)
- Under partial likelihood, it is assumed that \( G(x_{it}\beta) \) is the correct density of \( f(y_{it}|x_i) = f(y_{it}|x_{it}) \) for each \( t \), but the joint density \( G(y_i|x_i) \) is not the product of the marginals
Robust std errors should be used unless one assumes dynamic completeness, that is

$$\Pr(y_{it} = 1|x_{it}, y_{it-1}, x_{it-1}, ...) = \Pr(y_{it} = 1|x_{it})$$

which implies that lagged $x$ and $y$ do not enter the density for $y_{it}$ given $x_{it}$

- Specification test for dynamic completeness
  - Estimate the pooled probit/logit and obtain $\hat{\varepsilon}_{it} = y_{it} - G(x_{it}\hat{\beta})$
  - Estimate a new pooled probit/logit where $\hat{\varepsilon}_{it-1}$ is included as a regressor ($t > 1$); rejection of its coeff as zero implies rejection of dynamic completeness
Models With Unobserved Effects...

- Model given in latent form is

\[ y_{it}^* = x_{it}\beta + c_i + \varepsilon_{it}, \quad \varepsilon_{it} \sim \mathcal{N}/\mathcal{L} \]

\[ y_{it} = I(x_{it}\beta + c_i + \varepsilon_{it} > 0) \]

- RE specification assumes

\( f(c_i|x_{it}) \) is not dependent on \( x_{it} \)

- FE specification leaves \( f(c_i|x_{it}) \) unrestricted s.t. \( c_i \) and \( x_{it} \) may be correlated
RE Specification

- Assumptions

\[
E[\varepsilon_{it} | x] = E[c_i | x] = \text{Cov}(\varepsilon_{it}, c_j | x) = 0 \quad \forall i, j, t
\]

\[
\text{Cov}(\varepsilon_{it}, \varepsilon_{js} | x) = \text{Var}(\varepsilon_{it} | x) = \begin{cases} 
1 & i = j, t = s \\
0 & \text{otherwise}
\end{cases}
\]

\[
\text{Cov}(c_i, c_j | x) = \text{Var}(c_i | x) = \begin{cases} 
\sigma^2_c & i = j \\
0 & \text{otherwise}
\end{cases}
\]

where \( x \) includes \( x_{it} \) \( \forall i, t \) and assumptions imply that \( x_{it} \) is strictly exogenous conditional on \( c_i \)

- Implies

\[
E[\tilde{\varepsilon}_{it} | x] = 0
\]

\[
\text{Var}(\tilde{\varepsilon}_{it} | x) = \sigma^2_{\tilde{\varepsilon}} = \sigma^2_\varepsilon + \sigma^2_c = 1 + \sigma^2_c
\]

\[
\text{Corr}(\tilde{\varepsilon}_{it}, \tilde{\varepsilon}_{is} | x) = \rho = \frac{\sigma^2_c}{1 + \sigma^2_c}
\]

where \( \tilde{\varepsilon} \) is the composite error
In general, the likelihood involves high-dimensional integration. However, the RE structure simplifies estimation by factoring the density into \( f(\varepsilon_{i1}, ..., \varepsilon_{iT} | c_i) f(c_i) = f(\varepsilon_{i1} | c_i) \cdot \cdot \cdot f(\varepsilon_{iT} | c_i) f(c_i) \).

Log-likelihood becomes

\[
\ln[\mathcal{L}(\theta)] = \sum_i \ln \left[ \int_{-\infty}^{\infty} \prod_{t=1}^{T} \Pr(y_{it} | x_{it} \beta + c_i) \right] \cdot f(c_i) \, dc
\]

which requires integrating over the dbn of \( c \). Assuming \( c_i | x_i \sim \mathcal{N}(0, \sigma_c^2) \), the likelihood fn becomes

\[
\ln[\mathcal{L}(\theta)] = \sum_i \ln \left[ \int_{-\infty}^{\infty} \prod_{t=1}^{T_i} \Pr(y_{it} | x_{it} \beta + c_i) \right] \cdot \frac{1}{\sigma_c} \phi \left( \frac{c_i}{\sigma_c} \right) \, dc
\]

Butler & Moffitt (1982) show how to approximate the integral.

- Known as \textit{RE probit estimator}
- Estimate of \( \hat{\rho} \) allows testing for presence of unobserved heterogeneity
Notes...

- Maximum Simulated Likelihood (MSL) estimation is an alternative that allows for $c$ to follow some non-normal dbns
- RE logit model also exists and assumes $c \sim \mathcal{N}$
- Integrals approximated using various quadrature procedures; results may be sensitive to number of quadrature points
  - Stata: `-quadchk-`
Marginal effects are more complex since they depend on $c$

$$\frac{\partial \Pr(y_{it} = 1)}{\partial x_{kit}} = \beta_k \phi(x_{it}\beta + c_i)$$

- Since $E[c|x] = 0$, the ME at the population average of $c$ in the population is $\beta_k \phi(x_{it}\beta)$
- The relative ME of any two covariates, $x_k$ and $x_{k'}$, is

$$\frac{\frac{\partial \Pr(y_{it} = 1)}{\partial x_{kit}}}{\frac{\partial \Pr(y_{it} = 1)}{\partial x_{k'it}}} = \frac{\beta_k \phi(x_{it}\beta + c_i)}{\beta_{k'} \phi(x_{it}\beta + c_i)} = \frac{\beta_k}{\beta_{k'}}$$

which is not observation-specific
The average marginal effect (AME) is the value of $\beta_k \phi(x_{it} \beta + c_i)$ averaged over the dbn of $c$, which in the RE probit is

$$\frac{\beta_k}{\sigma_{\epsilon}} \phi \left( x_{it} \frac{\beta}{\sigma_{\epsilon}} \right)$$

which only requires knowledge of $\beta_c = \beta / \sigma_{\epsilon}$, known as the population-averaged parameters.

- Interestingly, $\beta_c$ is what is estimated by the pooled probit estimator if $c$ is ignored.
- Thus, one may estimate $\beta_c$ via pooled probit with robust std errors (which does not assume iid observations over time), or impose the full error structure and estimate the RE probit.
FE Specification

- Model re-written as

\[ y_{it} = I(c_i d_{it} + x_{it} \beta + \varepsilon_{it} > 0), \quad \varepsilon_{it} \sim \mathbb{N}/\mathbb{L} \]

where \( d_{it} \) is a dummy variable equal to 1 for obs \( i \), 0 otherwise

- Log-likelihood is

\[ \ln [\mathcal{L} (\theta)] = \sum_i \sum_{t=1}^{T} \Pr(y_{it} \mid c_i d_{it} + x_{it} \beta) \]

where

\[ \Pr(\cdot) = \begin{cases} 
  \Phi[q_{it} (c_i d_{it} + x_{it} \beta)] & \text{if } \varepsilon_{it} \sim \mathbb{N} \\
  \Lambda[q_{it} (c_i d_{it} + x_{it} \beta)] & \text{if } \varepsilon_{it} \sim \mathbb{L} 
\end{cases} \]

and \( q_{it} = 2y_{it} - 1 \)
The FE estimator is flawed

- As with continuous FE models, \( c_i \) estimates are inconsistent under fixed-\( T \) asymptotics
  - Unlike continuous FE models, the inconsistency also causes \( \beta \) to be inconsistent
  - There is also a small sample bias, although the magnitude remains an open question
  - Problem known as the *incidental parameters problem*

- Thus, there is a trade-off between the restrictiveness of the RE assumption vs. the bias/inconsistency due to the incidental parameters problem
- Research continues on bias-corrected estimators
Question: Why does the inconsistency of $c_i$ in the continuous FE model not affect $\beta$?

- In the continuous $y$ case, the model is transformed into deviations from obs-specific means
- Formally, while $f(y_{it}|X)$ depends on $c_i$, $f(y_{it}|X, \bar{y}_i)$ does not
  - $\bar{y}_i$ is a minimal sufficient statistic for $c_i$
  - $\bar{y}_i$ is then used in the estimation of $\beta$
- In the probit model, there is no sufficient statistic for $c_i$
- There is in the logit model
Before discussing the FE logit model, there is Chamberlain’s *correlated RE* probit estimator based on Mundlak (1978). Assume

\[ c_i | x_i \sim \mathcal{N}(\xi_0 + \bar{x}_i \xi_1, \sigma_a^2) \]

where \( a_i = c_i - \xi_0 - \bar{x}_i \xi_1 \) and \( \sigma_a^2 \) does not depend on \( x \).

Under the prior assumptions for the RE probit estimator, apply the RE probit estimator to the model

\[
\Pr(y_{it} | x_{it}, c_i) = \Phi(\xi_0 + x_{it} \beta + \bar{x}_i \xi_1 + a_i)
\]

where \( a_i | x_i \sim \mathcal{N}(0, \sigma_a^2) \) and \( x \) does not contain an intercept.

- This yields estimates of \( \sigma_a^2 \), not \( \sigma_c^2 \).
- Testing \( H_0 : \xi_1 = 0 \iff \text{RE model vs. Chamberlain’s RE model} \)

The AMEs can be obtained by estimating the pooled probit model

\[
\Pr(y_{it} | x_{it}) = \Phi(\tilde{\xi}_0 + x_{it} \tilde{\beta} + \bar{x}_i \tilde{\xi}_1)
\]

or after the RE probit model as

\[
\tilde{\beta}_k \phi \left( \tilde{\xi}_0 + x_{it} \tilde{\beta} + \bar{x}_i \tilde{\xi}_1 \right)
\]

where \( \tilde{\beta} = \beta / \sqrt{1 + \sigma_a^2} \) (and similarly for \( \tilde{\xi} \)).
Now, turn to the FE logit model

$$\Pr(y_{it} = 1|x_{it}) = G_{it} = \frac{\exp\{x_{it}\beta + c_i\}}{1 + \exp\{x_{it}\beta + c_i\}}$$

with the likelihood given as

$$\mathcal{L}(\theta) = \prod_i \prod_t (G_{it})^{y_{it}} (1 - G_{it})^{1-y_{it}}$$

and no assumptions are made about dependence between $x_{it}$ and $c_i$

Trick is to derive the joint dbn of $y_i|x_i, c_i, n_i$, where $n_i \equiv \sum_t y_{it}$ (# of periods obs $i$ has $y = 1$)

- Turns out that this dbn is independent of $c_i$
- Thus, $n_i$ is a *minimal sufficient statistic* for $c_i$

Probit model does not give rise to such a sufficient statistic such that the conditional dbn does not depend on $c_i$
Chamberlain (1980) observed that the *conditional likelihood* does not depend on $c_i$

$$L(\theta) = \prod_i \Pr(y_{i1}, \ldots, y_{iT} | x_i, n_i)$$

where the contribution of obs $i$ to the likelihood is

$$L_i(\theta) = \Pr(y_{i1}, \ldots, y_{iT} | x_i, n_i) = \frac{\Pr(y_{i1}, \ldots, y_{iT}, n_i | x_i)}{\Pr(n_i | x_i)} = \frac{\exp \{ \sum_t y_{it} (x_{it} \beta) \}}{\sum_{a \in R_i} \exp \{ \sum_t (a_t x_{it} \beta) \}}$$

and $R_i \subset \mathbb{R}^T$ defined as $\{a \in \mathbb{R}^T : a_t \in \{0, 1\} \text{ and } \sum_t a_t = n_i\}$

- $R_i$ represents all sequences of 1s and 0s across the $T$ periods such that $\sum_t a_t = n_i$
- The denominator is summed over the set of all $\binom{T}{n_i}$ unique sequences of $a_t$ that sum to $n_i$

Only obs for which $n_i \notin \{0, T\}$ contribute information about $\beta$ since $y_i$ is perfectly determined in these two cases
Test of homogeneity given by $H_o : c_i = c \ \forall i$

- Likelihood ratio test is not possible since the FE estimates are based on the conditional likelihood fn
- Hausman test is applicable since both are consistent under the null, but the FE model is inefficient, while only the FE model is consistent under the alternative

$$
\left(\hat{\beta}_{FE} - \hat{\beta}\right)' \left(\hat{\Sigma}_{FE} - \hat{\Sigma}\right)^{-1} \left(\hat{\beta}_{FE} - \hat{\beta}\right) \sim \chi^2
$$

Interpretation

- $\beta$ gives the ME of $x$ on the log odds ratio

$$
\ln \left[ \frac{\Pr(y_{it} = 1)}{\Pr(y_{it} = 0)} \right] = x_{it}\beta + c_i
$$

- MEs of $x$ on $\Pr(y_{it} = 1)$, and estimation of the AMEs, is not feasible without structure on the dbn of $c$, which is what the FE logit model avoids
Final notes...

- All RE, FE estimators require strict exogeneity of \( x_{it} \) conditional on \( c_i \)
  - Straightforward to test by adding leads, \( x_{it+1} \), to the model and testing for significance
  - Requires omitting the final period from the estimation

- FE logit vs. Chamberlain’s RE probit model
  - FE logit leaves the dbn of \( c \) unspecified, but does not yield estimates of the AMEs
  - Correlated RE probit based on the augmented pooled probit allows for serial correlation and estimates the AMEs, but forces one to place some structure on the dbn of \( c|x \)

- Additional work on FE binary choice models is on-going (e.g., Ai & Gan 2010)

- Stata: -xtprobit-, -xtlogit-
Dynamic Models

- Extensions to dynamic binary choice models are in progress
  - Dynamic binary probit models first considered in Heckman (1981a,b)
  - Orme (1997, 2001) and Wooldridge (2005) propose alternatives that are easier to estimate
  - Arulampalam and Stewart (2009) propose an easier to estimate version of Heckman's estimator; estimable in Stata in a straightforward manner

- Akay (2011) compares Heckman and Wooldridge's approaches
Model

\[ \Pr(y_{it} = 1|x_i, y_{it-1}, y_{it-2}, \ldots, c_i) = G(x_{it} \beta + \gamma y_{it-1} + c_i), \quad t = 2, \ldots, T \]

Initial conditions problem

- General estimation strategy is to model \( f(y_1, \ldots, y_T | x, c) \), specify a dbn \( f(c | x) \), and then integrate wrt this density
- To figure out \( f(y_1, \ldots, y_T | x, c) \), we can factor this as

\[
    f(y_1, \ldots, y_T | x, c) = f(y_2, \ldots, y_T | x, y_1, c) f(y_1 | x, c)
\]

where \( y_{i1} \) is the initial condition
- Thus, maximizing the likelihood fn requires

1. Specification of \( f(y_1 | x, c) \), or
2. Assumption that \( f(y_1 | x, c) = f(y_1 | x) \); i.e., \( y_{i1} \) is independent of \( c_i \) conditional on \( x_i \)
Under (2) the above model can be estimated by conditional MLE using the RE probit or correlated RE probit with only data from periods 2, .., \( T \) and assuming \( c_i | x_i \sim \mathcal{N}(0, \sigma_c^2) \)

\[
\ln[\mathcal{L}(\theta)] = \sum_i \ln \left[ \int_{-\infty}^{\infty} \left[ \prod_{t=2}^{T} \Pr(y_{it} | x_{it} \beta + \gamma y_{it-1} + c_i) \right] \frac{1}{\sigma_c} \phi \left( \frac{c_i}{\sigma_c} \right) dc \right]
\]

or assuming \( c_i | x_i \sim \mathcal{N}(\xi_0 + \bar{x}_i \xi_1, \sigma_a^2) \) and integrating wrt to the density of \( a \)
However, $y_{i1}$ is likely not to be conditionally independent of $c_i$.

- Even if period 1 corresponds to the ‘true’ initial period, in most economic applications it is difficult to justify the assumption of $y_{i1}$ being unrelated to $c_i$.
- Thus, the likelihood fn must account for the initial period.

\[
\ln[L(\theta)] = \sum_i \ln \left[ \int_{-\infty}^{\infty} \left[ \frac{\Pr(y_{i1}|x_i, c_i) \times \prod_{t=2}^{T_i} \Pr(y_{it}|x_{it}\beta + \gamma y_{it-1} + c_i)}{\sigma_c} \right] \frac{1}{\sigma_c} \phi \left( \frac{c_i}{\sigma_c} \right) dc \right]
\]

- Different solutions based on different specifications of $\Pr(y_{i1} = 1|x_i, c_i)$
Heckman approach

\[
\Pr(y_{i1} = 1|x_i, z_i, c_i) = G(x_i \beta + z_i \pi + \lambda c_i)
\]

where \( z \) may include additional determinants of the initial period if available and \( \lambda \) allows for differences in the distribution of the composite error in the initial period.

- Likelihood fn

\[
\ln[\mathcal{L}(\theta)] = \sum_i \ln \left[ \int_{-\infty}^{\infty} \frac{\Pr(y_{i1}|x_i \beta + z_i \pi + \lambda c_i) \times \prod_{t=2}^{T} \Pr(y_{it}|x_{it} \beta + \gamma y_{it-1} + c_i)}{\Pi_{t=2}^{T} \Pr(y_{it}|x_{it} \beta + \gamma y_{it-1} + c_i)} \right] \frac{1}{\sigma_c} \phi \left( \frac{c_i}{\sigma_c} \right) dc
\]

- Stewart (2007) extends this to allow for serial correlation in the idiosyncratic errors
- Correlation between \( x_i \) and \( c_i \) can be addressed using the Mundlak (1978) approach (i.e., including \( \bar{x}_i \) in \( x_{it} \) in which case \( c_i \) becomes \( a_i \))
- Stata: -redprob-, -redpace-
  (http://www2.warwick.ac.uk/fac/soc/economics/staff/academic/stewart/stata/)
Wooldridge approach

- Rather than model the initial condition, this approach specifies a dbn for \( f(c_i|x_i, y_{i1}) \) and integrates wrt this dbn.
- Extending the Mundlak (1978) approach, assume:

\[
c_i|x_i, y_{i1} \sim \mathcal{N}(\xi_0 + \bar{x}_i \xi_1 + \xi_2 y_{i1}, \sigma_a^2)
\]

where \( a_i = c_i - \xi_0 - \bar{x}_i \xi_1 - \xi_2 y_{i1} \) and \( \sigma_a^2 \) does not depend on \( x \) or \( y_{i1} \).
- Now, the conditional likelihood fn is given by:

\[
\ln[\mathcal{L}(\theta)] = \sum_i \ln \left[ \int_{-\infty}^{\infty} \prod_{t=2}^{T} \Pr(y_{it}|\xi_0 + x_{it} \beta + \bar{x}_i \xi_1 + \gamma(y_{it-1} + \xi_2 y_{i1} + a_i) \right] \frac{1}{\sigma_a} \phi \left( \frac{a_i}{\sigma_a} \right) da
\]

which can be estimated using the usual RE probit model.
- Akay (2011) shows this approach performs better than the Heckman approach when \( T > 5 \).
Orme approach

- Not as commonly used
- Write the model for the initial period as
  \[ y_{i1} = I(x_{i1} \beta + z_{i1} \pi + \varepsilon_{i1} > 0) \]

where \( \varepsilon_{i1}, c_i \sim N_2(0, 0, \sigma^2_{\varepsilon}, \sigma^2_c, r) \)

- Utilizing properties of the bivariate normal density, we have
  \[ c_i = r \frac{\sigma_c}{\sigma_\varepsilon} \varepsilon_{i1} + \sigma_c \sqrt{1 - r^2} w_i \]

where \( w_i \sim N(0, 1) \) and is independent of \( \varepsilon_{i1} \)

- Substituting in for \( c_i \) yields

\[
\Pr(y_{it} = 1|x_i, y_{it-1}, c_i) = G \left( x_{it} \hat{\beta} + \gamma y_{it-1} + r \frac{\sigma_c}{\sigma_\varepsilon} \varepsilon_{i1} + \sigma_c \sqrt{1 - r^2} w_i \right), \quad t = 2, \ldots
\]

which can be estimated by conditional ML using the usual RE probit model where \( w_i \) is the RE and \( \varepsilon_{i1} \) is replaced with

\[
\hat{\varepsilon}_{i1} = \frac{(2y_{i1} - 1) \phi(x_{i1} \hat{\beta} + z_{i1} \hat{\pi})}{\Phi[(2y_{i1} - 1)(x_{i1} \hat{\beta} + z_{i1} \hat{\pi})]}
\]

where \( \hat{\beta} \) and \( \hat{\pi} \) are estimated via probit on data from initial period
Panel Data
LDV Models: Tobit Model

- Start with the pooled tobit model

\[ y_{it}^* = x_{it} \beta + \varepsilon_{it}, \quad \varepsilon_{it} | x_{it} \sim \mathcal{N}(0, \sigma^2) \]
\[ y_{it} = \max(0, y_{it}^*) \]

which is estimable as a tobit model with \( NT \) obs assuming \( x_{it} \) are exogenous (but not strictly)

- As with the pooled probit, this a Partial ML estimator
- Dynamic completeness implies usual inference is correct
  - Tested by estimating the following pooled tobit

\[ y_{it}^* = x_{it} \beta + \gamma_1 r_{it-1} + \gamma_2 (1 - r_{it-1}) \hat{\varepsilon}_{it-1} + \eta_{it} \]

where \( r_{it-1} = I(y_{it-1} = 0) \) and \( \hat{\varepsilon}_{it-1} = y_{it-1} - x_{it-1} \hat{\beta} \) if \( y_{it-1} > 0 \) and \( \hat{\varepsilon}_{it-1} = 0 \) if \( y_{it-1} = 0 \) using \( t > 1 \)

- Absent this assumption, robust std errors are needed
Model with unobserved effects is given in latent form as

\[ y_{it}^* = x_{it}\beta + c_i + \varepsilon_{it}, \quad \varepsilon_{it} | x_i, c_i \sim \mathcal{N}(0, \sigma_{\varepsilon}^2) \]

\[ y_{it} = \max(0, y_{it}^*) \]

Skipping a simple RE version, move to a Chamberlain-like model

- Assume

\[ c_i | x_i \sim \mathcal{N}(\bar{\xi}_0 + \bar{x}_i\bar{\xi}_1, \sigma_a^2) \]
\[ a_i | x_i \sim \mathcal{N}(0, \sigma_a^2) \]
\[ \varepsilon_{it} | x_i, a_i \sim \mathcal{N}(0, \sigma_{\varepsilon}^2) \]

where \( a_i = c_i - \bar{\xi}_0 - \bar{x}_i\bar{\xi}_1 \)

- Assumption of \( \varepsilon_{it} | x_i, c_i \sim \mathcal{N}(0, \sigma_{\varepsilon}^2) \) implies \( x_{it} \) are strictly exogenous conditional on \( c_i \)
- Also assume that \( \varepsilon_{it} \) are serially uncorrelated

Under this setup, a RE tobit estimator with \( x_{it} \) and \( \bar{x}_i \) as covariates provides consistent estimates of the parameters

Testing \( H_0 : \bar{\xi}_1 = 0 \Leftrightarrow \) RE tobit model vs. Chamberlain’s RE model
Final notes...

- AMEs discussed in Wooldridge (2002)
- No firm results on the behavior of MLE including obs-specific dummies
  - Prior research suggests $\beta$ is not biased, but $\sigma$ is biased down (Greene 2003)
  - Even if $\beta$ is unbiased, inference and computation of marginal effects are difficult given the bias in $\sigma$
  - Research continues
- Extensions to dynamic tobit models are available (e.g., following the Wooldridge (2005) approach)
- Stata: -xttobit-
Panel Data

LDV Models: Count Models

- Start with the pooled Poisson model
  - Model expected number of events conditional on $x$
    \[ E[y_{it} | x_{it}] = F(x_{it} \beta) \]
    where $F(\cdot) \geq 0$
  - Common functional form of $F(\cdot) = \exp\{x_{it} \beta\}$
  - Recall, Poisson dbn depends only on the mean given by $\lambda_{it} = \exp\{x_{it} \beta\}$
  - Log-likelihood fn given by
    \[ \ln[\mathcal{L}(\theta)] = \sum_i \sum_t [-\lambda_{it} + y_{it} \ln \lambda_{it} - \ln(y_{it}!)] \]
    - Estimator is CAN assuming conditional mean is correctly specified regardless of Poisson dbn
    - Further assumptions needed for inference
RE model given by

$$E[y_{it} \mid x_{it}, c_i] = F(x_{it} \beta + \alpha_i)$$

where $F(\cdot) = \exp\{x_{it} \beta + \alpha_i\}$

Assumptions

1. Functional form arises if one assumes that $c_i = \exp(\alpha_i)$ and unobserved effect enters multiplicatively
2. $x_{it}$ is strictly exogenous
3. $c_i | x_i \sim \Gamma(\delta_0, 1 / \delta_0)$ dbn
   - Implies $E[c_i] = 1$, $\text{Var}(c_i) = 1 / \delta_0 \equiv \eta_0^2$
4. No serial correlation in $y_{it}$ conditional on $x_i, c_i$

Poisson assumption implies equi-dispersion in $y_{it} | x_i, c_i$

However, $\text{Var}(y_{it} | x_i) = E[y_{it} | x_i](1 + \eta_0^2 E[y_{it} | x_i]) \Rightarrow$ over-dispersion conditional on $x_i$ alone
Estimation entails forming the likelihood fn, based on the density
\( f(y_{i1}, \ldots, y_{iT} | x_i) \), which requires integrating over the dbn of \( c \)

- Assumption concerning the gamma dbn allows this density to be tractable
  - Assuming normality is also possible
  - Under maintained assumptions, MLE is efficient
  - Estimates are sensitive to violations

- A QMLE RE estimator relaxes some of these assumptions and is equivalent to a pooled NB model
- A Mundlak (1978) correlated RE model is feasible
Hausman et al. (1984) develop a FE version, allowing dependence between $x_i$ and $c_i$

Requires same assumptions as RE model above except those concerning the dbn of $c_i$

The *conditional likelihood* does not depend on $c_i$

$$L(\theta) = \prod_i \Pr(y_{i1}, ..., y_{iT} | x_i, n_i, c_i)$$

if we condition on $n_i$, $n_i = \sum_{t=1}^{T} y_{it}$ ($n_i$ is a minimal sufficient statistic)

The contribution of obs $i$ to the conditional log-likelihood is

$$\ln[L_i(\theta)] = \ln[\Gamma (1 + \sum_t y_{it})] - \sum_t \ln \left[ \Gamma (y_{it} + 1) \right] + \sum_t y_{it} \ln \left[ \frac{\exp\{x_{it} \beta\}}{\sum_s \exp\{x_{is} \beta\}} \right]$$

Known as the FE Poisson estimator

- Consistency only requires the conditional mean to be correctly specified
- As such, estimator is applicable to *any* model where this is the case (e.g., $y_{it}$ could be continuous, contain negative values, binary, etc.)
  - Example: gravity models in levels
Notes:

- If $n_i = 0$, then obs does not contribute any information
- FE-NB model also feasible, based on

$$\ln[\mathcal{L}_i(\theta)] = \ln \left[ \Gamma \left( \sum_t \lambda_{it} \right) \right] + \ln \left[ \Gamma \left( 1 + \sum_t y_{it} \right) \right] - \ln \left[ \Gamma \left( \sum_t y_{it} + \sum_t \lambda_{it} \right) \right]$$

$$+ \sum_t \left\{ \ln \left[ \Gamma \left( y_{it} + \lambda_{it} \right) \right] - \ln \left[ \Gamma(\lambda_{it}) \right] - \ln \left[ \Gamma(1 + y_{it}) \right] \right\}$$

where $\lambda_{it} = \exp\{x_{it}\beta\}$

- FE-NB allows estimation of coeffs on time invariant regressors and an intercept
- FE-Poisson does not

- Difference follows from the fact that the Poisson specifies
  $$E[y_{it}|x_{it}] = F(x_{it}\beta + \alpha_i),$$
  whereas NB specifies
  $$E[y_{it}|x_{it}] = \theta_i \exp\{x_{it}\beta\},$$
  where $\theta$ is a scaling parameter rather than a shifter of the conditional mean function

- Since the conditional mean is homogeneous in the FE-NB, an intercept and effects of time invariant regressors are estimable

- Dynamic count models considered in Wooldridge (2005)

Stata: -xtpoisson-, -xtpqml-, -xtnbreg-
Panel Data
LDV Models: Qualitative Choice Models

- Pg. 653-4 in Wooldridge
Panel Data
LDV Models: Ordered Models

- FE ordered probit
  - Suffers from incidental parameters problem, as in the binary probit model
  - Evidence suggests that finite sample bias appears to be of similar magnitude

- FE ordered logit is consistent, but has limited usefulness
  - Model setup
    \[ y_{it}^* = x_{it} \beta + c_i + \epsilon_{it}, \quad \epsilon_{it} \sim \mathcal{N} \]
    \[ y_{it} = j \quad \text{if} \quad y_{it}^* \in (\mu_{j-1}, \mu_j) \]

where

- ★ \( j = 0, 1, \ldots, J \)
- ★ \( \mu_{-1} = -\infty, \mu_0 = 0, \text{ and } \mu_J = \infty \)

- Probabilities given by
  \[ \Pr(y_{it} = j) = \Lambda(\mu_j - x_{it} \beta - c_i) - \Lambda(\mu_{j-1} - x_{it} \beta - c_i) \]
Define a new sequence of dummy variables

\[ w_{it,j} = \begin{cases} 
1 & \text{if } y_{it} > j \\
0 & \text{otherwise} 
\end{cases}, \quad j = 0, 1, \ldots, J - 1 \]

It follows that

\[
\Pr(w_{it,j} = 1|X) = \Lambda(-\mu_j + x_{it}\beta + c_i)
\]

\[
= \Lambda(\theta_i + x_{it}\beta)
\]

which reduces to a series of \( J - 1 \) binary, FE logits estimable as before
Notes:

- Procedure produces $J - 1$ consistent estimates of $\beta$; these can be combined into a single vector $\beta$ using MDE

$$\hat{\beta}_{MDE} = \arg\min_{\beta} \sum_{j=0}^{J-1} \sum_{m=0}^{J-1} (\hat{\beta}_j - \beta)' \left[ V_{jm}^{-1} \right] (\hat{\beta}_m - \beta)$$

where $V_{jm}^{-1}$ is the $j, m$ block of the inverse of the $(J - 1)K \times (J - 1)K$ matrix $V$ that contains the Asy.Cov($\hat{\beta}_j, \hat{\beta}_m$), but estimates of the off-diagonal blocks are not obvious.

- While $\beta$ is consistent, the threshold parameters are not identified, implying no estimates of marginal effects or predicted probabilities.

- RE version available
Panel Data
Heterogeneous Time Trends

- All estimators to this point are known as *homogeneous* estimators as they assume constant parameters across $i$.
- Allowing for individual-specific time trends yields

$$ y_{it} = c_i + \lambda_i t + x_{it} \beta + \varepsilon_{it} $$

- Estimation proceeds by FD

$$ \Delta y_{it} = \lambda_i + \Delta x_{it} \beta + \Delta \varepsilon_{it} $$

and then estimating this model by RE, FE, FD, or LSDV given the assumptions one wishes to invoke.

- Note: This requires $T \geq 3$
Panel Data
Heterogeneous Slopes

- Model allowing for heterogeneous slopes more generally is given by

\[ y_{it} = c_i + \gamma_i y_{it-1} + x_{it} \beta_i + \varepsilon_{it} \]
\[ \gamma_i = \gamma + \eta_{1i} \]
\[ \beta_i = \beta + \eta_{2i} \]
\[ \theta_i = \frac{\beta_i}{1 - \gamma_i} \]

- Perhaps more realistic in many applications as assumption of identical effects of covariates may not be justified
- Estimation depends on what one is interested in estimating: population average parameters and/or panel-specific parameters
- Even if one is only interested in population average parameters, acknowledgement of panel-specific parameters complicates estimation
- Estimation now typically requires \( N, T \to \infty \)
Swamy (1970) FGLS estimator

- Restricted to models where $c_i = c$ and $\gamma_i = 0 \ \forall i$

$$y_{it} = x_{it}\beta_i + \varepsilon_{it}$$

and $x$ includes an intercept

- Assumptions:
  
  (S.i) $E[\varepsilon] = 0$
  
  (S.ii) $E[\varepsilon_i \varepsilon_j'] = \begin{cases} \sigma^2_i I_T & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$
  
  (S.iii) $\beta_i = \beta + \eta_i$, where $E[\eta] = 0$
  
  (S.iv) $\eta_i$ and $\varepsilon_j$ are independent
  
  (S.v) $E[\eta_i \eta_j'] = \begin{cases} \Omega & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$

  where $\Omega$ is a $K \times K$ matrix
Substituting for $\beta_i$ implies

$$y_{it} = x_{it} \beta + (\varepsilon_{it} + x_{it} \eta_i) = x_{it} \beta + v_{it}$$

where

$$E[vv'] = \begin{bmatrix}
  x_1 \Omega x_1' + \sigma_1^2 I_T & 0 & \cdots & 0 \\
  0 & x_2 \Omega x_2' + \sigma_2^2 I_T & \cdots & 0 \\
  \vdots & \vdots & \ddots & \vdots \\
  0 & 0 & \cdots & x_N \Omega x_N' + \sigma_N^2 I_T
\end{bmatrix}$$

Thus, there are two sources of heteroskedasticity, as well as fact that $v$ is serially correlated within panels

Estimation could ignore panel-specific parameters and instead estimate the model using pooled OLS with clustered standard errors

Alternative is FGLS
GLS (Aitken) estimator given by

$$\hat{\beta}_{GLS} = (x'\Sigma^{-1}x)^{-1}x'\Sigma^{-1}y$$

$$= \left[ \sum_i x_i' (x_i\Omega x_i' + \sigma_i^2 I_T)^{-1} x_j \right]^{-1} \sum_i x_i' (x_i\Omega x_i' + \sigma_i^2 I_T)^{-1} y_i$$

$$= \sum_i \omega_i \hat{\beta}_i$$

where

$$\omega_i = \left\{ \sum_j \left[ \Omega + \sigma_j^2 (x_j'x_j)^{-1} \right]^{-1} \right\}^{-1} \left[ \Omega + \sigma_i^2 (x_i'x_i)^{-1} \right]^{-1}$$

$$\hat{\beta}_i = (x_i'x_i)^{-1} x_i' y_i$$

which is a weighted average of $\hat{\beta}_i$. 

FGLS estimator

- Requires estimates of $\Omega$ and $\sigma_i^2 \ orall i$
- Estimates given by

$$
\hat{\sigma}_i^2 = \frac{y_i'M_iy_i}{T-K} = \frac{\hat{\varepsilon}_i\hat{\varepsilon}_i}{T-K}
$$

$$
\hat{\Omega} = \frac{S_b}{N-1} - \frac{1}{N} \sum_i \hat{\sigma}_i^2 (x_i'x_i)^{-1}
$$

where

$$
M_i = I - x_i(x_i'x_i)^{-1}x_i'
$$

$$
S_b = \sum_i \hat{\beta}_i\hat{\beta}_i' - \frac{1}{N} \sum_i \hat{\beta}_i \sum_i \hat{\beta}_i'
$$

and $\hat{\beta}_i$ and $\hat{\varepsilon}_i$ are obtained via panel-specific OLS (\therefore large $T$)

- Can test $H_0: \beta_1 = \cdots = \beta_N$ after
- Stata: -xtrc-
Pesaran & Smith (1995) estimators: $c_i \neq c, \gamma_i \neq \gamma \ \forall i$

- Mean Group (MG) Estimator
  - Estimate separate time series regressions for each $i$
  - Obtain an average estimate of $\gamma$ and $\beta$ across the $N$ regressions

- Pooled (P) Estimator
  - Pool data, but allow parameters to have a random component

- Aggregate (A) Estimator
  - Averaging the data across $i$ (for each $t$)
  - Estimate a single, aggregate time series regression

- Cross-Section (CS) Estimator
  - Average the data across $t$ (for each $i$)
  - Estimate a single, cross-section regression

$\Rightarrow$ MG provides estimates of individual-specific coeffs, as well as average coeffs; remaining estimators estimate only the average coeffs

- See -xtmg- in Stata
Properties

- In a static model \((\gamma_i = 0 \ \forall i)\) with \(x\) strictly exogenous and \(\beta_i\) differ randomly and are distributed independently across \(i\), all four estimators are unbiased and consistent for \(\beta\).
- In a dynamic model \((\gamma_i \neq 0 \ \forall i)\), this is not the case.
  - MG estimator remains consistent as \(N, T \to \infty\).
  - P and A estimators are inconsistent due to serial correlation in the regressors (at least \(y_{it-1}\) is serially correlated due to \(c_i\)) which, when coeff heterogeneity is ignored, induces serial correlation in the error term which leads to inconsistent estimates in models with lagged dependent vars (proofs in paper).
  - CS estimator remains consistent for the average long-run coeffs, \(\bar{\theta}\), where \(\theta_i = \beta_i / (1 - \gamma_i)\), but is biased and inconsistent if based on short \(T\).
Pooled estimator

- Estimating equation

\[ y_{it} = c_i + \gamma y_{it-1} + x_{it} \beta + [\varepsilon_{it} + \eta_{1i} y_{it-1} + x_{it} \eta_{2i}] \]

\[ = c_i + \gamma y_{it-1} + x_{it} \beta + u_{it} \]

where \( \mathbb{E}[y_{it-1} u_{it}] \neq 0 \) and \( \mathbb{E}[x_{it} u_{it}] \neq 0 \) if \( x \) is serially correlated; \( \mathbb{E}[y_{it-1} u_{it}] \neq 0 \) even if \( x \) is not serially correlated since \( \mathbb{E}[y_{it-1} \eta_{1i}] \neq 0 \)

- IV solution does not exist since any strong IV will also be correlated with \( u_{it} \)

Aggregate estimator

- Estimating equation

\[ \bar{y}_t = \bar{c} + \gamma \bar{y}_{t-1} + \bar{x}_t \beta + [\bar{\varepsilon}_t + \frac{1}{N} \sum_i (\eta_{1i} y_{it-1} + x_{it} \eta_{2i})] \]

\[ = \bar{c} + \gamma \bar{y}_{t-1} + \bar{x}_t \beta + \bar{u}_t \]

- OLS will be biased due to correlation between regressors and \( \bar{u} \)
- Unlikely that an IV solution exists
Cross-section estimator

- Estimating equation

\[ \bar{y}_i = c_i + \gamma \bar{y}_{i,-1} + \bar{x}_i \beta + [\bar{\varepsilon}_i + (\eta_{1i} \bar{y}_{i,-1} + \bar{x}_i \eta_{2i})] = c_i + \gamma \bar{y}_{i,-1} + \bar{x}_i \beta + \bar{\nu}_i \]

which is the *between* estimator of the dynamic model

- OLS will be biased even if \( \gamma \) and \( \beta \) are homogeneous, but \( c \) is heterogeneous (due to correlation with \( \bar{y}_{i,-1} \))
Alternative cross-section estimator

- Estimating equation obtained by substituting

\[
\bar{y}_{i,-1} = \bar{y}_i - \left(\frac{1}{T}\right)(y_{iT} - y_{i0}) = \bar{y}_i - \Delta_i
\]

since \( \bar{y}_{i,-1} = \left(\frac{1}{T}\right) \sum_{t=1}^{T} y_{it-1} \) and solving for \( \bar{y}_i \)

\[
\bar{y}_i = \frac{1}{1 - \gamma_i} c_i - \frac{\gamma_i}{1 - \gamma_i} \Delta_i + \bar{x}_i \theta_i + \frac{1}{1 - \gamma_i} \bar{\epsilon}_i
\]

- Introduce heterogeneity as

\[
\frac{\gamma_i}{1 - \gamma_i} = \frac{\gamma}{1 - \gamma} + \xi_{1i}
\]

\[
\theta_i = \theta + \xi_{2i}
\]

yielding

\[
\bar{y}_i = \frac{1}{1 - \gamma_i} c_i - \frac{\gamma}{1 - \gamma} \Delta_i + \bar{x}_i \theta + \left[ \frac{1}{1 - \gamma_i} \bar{\epsilon}_i - \xi_{1i} \Delta_i + \bar{x}_i \xi_{2i} \right]
\]

which can be estimated by OLS omitting \( \Delta_i \) since it is asymptotically uncorrelated with \( \bar{x}_i \)

- Provides biased but consistent estimates of the average LR coeff, \( \bar{\theta} \), if \( x \) is uncorrelated with \( c \); bias vanishes only as \( T \to \infty \)
Pesaran, Shin & Smith (1997, 1999) estimator

- MG estimator allows all parameters to vary by $i$
- FE estimator allows only the intercept, $c$, to vary with $i$
- Pooled Mean Group (PMG) estimator is a compromise
  - Model written as a ECM
  - The long-run coeffs in the cointegrating relationship, $\theta$, are constrained to be equal across $i$
  - The short-run coeffs and speed of adjustment parameters vary by $i$
  - Stata: -xtmg-

- Fomby & Balke (1997): threshold cointegration allows the speed of adjustment parameter to vary between two regimes
Nonstationary Panels

- The increased use of panel data by more macro-oriented economists led to study of asymptotics as both $N, T \to \infty$
- Allowing $T \to \infty$ led to concerns over nonstationarity, spurious regressions, cointegration
- Estimation with nonstationary data is easier in some respects with panel data than with pure time series
  - Many test statistics and estimators are asymptotically normal whereas they are not with time series
  - Spurious regression gives consistent estimates as $N, T \to \infty$
- Brief overview of these topics; see Baltagi for details and additional references
Nonstationary Panels
Panel Unit Roots

- Asymptotic properties of test statistics and estimators depend crucially on the way in which \( N \) and \( T \) go to \( \infty \)
- Three possibilities
  1. Sequential Limit: one, say \( N \), goes to \( \infty \) followed by the other, say \( T \)
  2. Diagonal Path Limit: \( N, T \to \infty \) at same rate, but along a specific, diagonal path s.t. \( T = T(N) \)
  3. Joint Limit: \( N, T \to \infty \) at same rate without placing diagonal path restrictions

- Joint limit theory is the most robust, but is more difficult and requires stronger conditions (e.g., existence of higher moments)
- Stata: -xtunitroot- (replaces previous, user-written commands)
Model

- Given by

\[ y_{it} = \rho_i y_{i(t-1)} + z_{it} \beta_i + \epsilon_{it} \]

where

\[ z_{it} = \begin{cases} 
1 & \Rightarrow \text{FE model} \\
0 & \Rightarrow \text{FE model with panel-specific time trend} \\
[1 \ t] & \Rightarrow \text{FE model with panel-specific time trend}
\end{cases} \]
Test Procedures

1. Levin, Lin, & Chu (LLC, 2002)
2. Harris & Tsavalis (1999)
4. Im, Pesaran, & Shin (IPS, 2003)
5. Fisher-type tests

- Hlouskova & Wagner (2006) conduct a large-scale MC study of the different tests
Tests differ in terms of

1. Whether $\rho_i = \rho \ \forall i$ is assumed
2. Rate at which $N, T \to \infty$

<table>
<thead>
<tr>
<th>Test</th>
<th>Options</th>
<th>Asymptotics</th>
<th>$\rho$ under $H_a$</th>
<th>Panels</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLC</td>
<td>noconstant</td>
<td>$\sqrt{N}/T \to 0$</td>
<td>common</td>
<td>balanced</td>
</tr>
<tr>
<td>LLC</td>
<td>$N/T \to 0$</td>
<td>common</td>
<td>balanced</td>
<td></td>
</tr>
<tr>
<td>LLC</td>
<td>trend</td>
<td>$N/T \to 0$</td>
<td>common</td>
<td>balanced</td>
</tr>
<tr>
<td>HT</td>
<td>noconstant</td>
<td>$N \to \infty, T$ fixed</td>
<td>common</td>
<td>balanced</td>
</tr>
<tr>
<td>HT</td>
<td>trend</td>
<td>$N \to \infty, T$ fixed</td>
<td>common</td>
<td>balanced</td>
</tr>
<tr>
<td>Breitung</td>
<td>noconstant</td>
<td>$(T, N)_{\text{seq}} \to \infty$</td>
<td>common</td>
<td>balanced</td>
</tr>
<tr>
<td>Breitung</td>
<td>trend</td>
<td>$(T, N)_{\text{seq}} \to \infty$</td>
<td>common</td>
<td>balanced</td>
</tr>
<tr>
<td>IPS</td>
<td>noconstant</td>
<td>$N \to \infty, T$ fixed</td>
<td>panel-specific</td>
<td>unbalanced</td>
</tr>
<tr>
<td>IPS</td>
<td>trend</td>
<td>$N \to \infty, T$ fixed</td>
<td>panel-specific</td>
<td>unbalanced</td>
</tr>
<tr>
<td>IPS</td>
<td>lags()</td>
<td>$(T, N)_{\text{seq}} \to \infty$</td>
<td>panel-specific</td>
<td>unbalanced</td>
</tr>
<tr>
<td>IPS</td>
<td>trend lags()</td>
<td>$(T, N)_{\text{seq}} \to \infty$</td>
<td>panel-specific</td>
<td>unbalanced</td>
</tr>
<tr>
<td>Fisher-type</td>
<td></td>
<td>$T \to \infty, N$ finite</td>
<td>panel-specific</td>
<td>unbalanced</td>
</tr>
<tr>
<td>Hadri LM</td>
<td>trend</td>
<td>$(T, N)_{\text{seq}} \to \infty$</td>
<td>(not applicable)</td>
<td>balanced</td>
</tr>
<tr>
<td>Hadri LM</td>
<td>trend</td>
<td>$(T, N)_{\text{seq}} \to \infty$</td>
<td>(not applicable)</td>
<td>balanced</td>
</tr>
</tbody>
</table>

- no constant $\Rightarrow z = 0$; trend $\Rightarrow z = [1 \ t]$
Before examining the panel tests, recall the ADF tests from a pure time series

- Model

\[ y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{s=1}^{p} \gamma_s \Delta y_{t-s} + \epsilon_t, \text{ or} \]
\[ \Delta y_t = \alpha + \beta t + \delta y_{t-1} + \sum_{s=1}^{p} \gamma_s \Delta y_{t-s} + \epsilon_t \]

where the random walk with and without drift are obtained imposing the proper restrictions

- Test statistics are

\[ DF_\tau = \frac{\hat{\gamma} - 1}{\text{s.e.}(\hat{\gamma})} \]

or the usual \( t \)-statistic for \( \hat{\alpha} \)

- Critical values depend on \( T \) and whether or not the intercept and/or the time trend are included
Harris & Tsavalis (1999) Test

- Model
  \[ y_{it} = \rho y_{it-1} + z_{it} \beta_i + \varepsilon_{it} \]
  where \(\varepsilon\) is iid and normally distributed

- Hypotheses
  \[
  \begin{align*}
  H_0 : & \quad \text{all panels contain a unit root} \\
  H_a : & \quad \text{no panels contain a unit root}
  \end{align*}
  \]

- Test statistic is based on the derivation of the asymptotic dbn of \(\hat{\rho}_{ols}\) under \(H_0\) (and each choice of \(z\))
  \[
  \sqrt{N}(\hat{\rho}_{ols} - \mu) \sim N(0, \sigma^2)
  \]
  where \(\mu\) and \(\sigma^2\) depend on the choice of \(z\)

- Asymptotic dbn is valid as \(N \to \infty\) (even with fixed \(T\))

- Test requires a balanced panel
Levin, Lin & Chu (LLC, 2002) Test

- **Model**
  \[ \Delta y_{it} = \phi y_{it-1} + z_i \beta_j + \sum_{s=1}^{p} \gamma_{is} \Delta y_{it-s} + u_{it} \]
  where \( p \) is chosen based on some model selection criteria

- **Hypotheses**
  \[ H_0: \text{all panels contain a unit root} \]
  \[ H_a: \text{no panels contain a unit root} \]

- **Test requires**
  1. \( \varepsilon_{it} \) to be independent across panels (zero cross-sectional dependence);
     \( u_{it} \) to be white noise
  2. \( \sqrt{N/T} \) or \( N/T \to 0 \) depending on choice of \( z \)
  3. balanced panel

- **Test is based on a ‘bias-adjusted’ \( t \)-statistic for \( \phi \) which has a \( N(0,1) \) dbn

- **However, procedure starts with panel-specific OLS estimates of the model**
Breitung (2000) Test

- Amended version of LLC that performs better with fixed effects
- Hypotheses
  \[
  \begin{align*}
  H_0 & : \text{all panels contain a unit root} \\
  H_a & : \text{no panels contain a unit root}
  \end{align*}
  \]

- Test requires
  1. a balanced panel
  2. \(\epsilon\) is iid

- A robust version is available that allows for contemporaneous cross-sectional dependence
Im, Pesaran & Shin (IPS, 2003) Test

- Model
  \[ \Delta y_{it} = \phi_i y_{it-1} + z_i \beta_i + \sum_{s=1}^{p} \gamma_{is} \Delta y_{it-s} + \epsilon_{it} \]
  where \( \epsilon \) is iid normal across panels, but may be groupwise heteroskedastic

- Hypotheses
  \[ H_o : \text{all panels contain a unit root (} \phi_i = 0 \quad \forall i) \]
  \[ H_a : \text{at least some panels do not contain a unit root} \]

- Formally, \( H_a \) entails that \( \lim_{N \to \infty} (N_1 / N) \to \delta, \delta \in (0, 1] \), where \( N_1 \) is the # of panels without a unit root

- Test statistic based on average of the ADF statistics for \( \hat{\phi} \) obtained from each panel separately

- Critical values based on fixed \( T \) and either fixed \( N \) or \( N \to \infty \)
Combining $p$-values (Fisher-type) Test

- Developed in Madala & Wu (1999)
- Hypotheses

$$H_o : \text{all panels contain a unit root}$$
$$H_a : \text{at least some panels do not contain a unit root}$$

- Similar to IPS, but rather than averaging test statistics across individual panels, test is based on aggregating $p$-values from individual time series unit root tests
- $P$-values obtained using either ADF of Phillips-Perron unit root tests applied to each panel
- Choi (2001) offers some extensions
Residual-Based LM Test

- Developed in Hadri (2000); flips the null and alternative hypotheses to address bias toward the null in classical statistical tests
- Hypotheses

  \[ H_0 : \text{no panels contain a unit root} \]
  \[ H_a : \text{at least one panel contains a unit root} \]

- Test designed for large \( T \), moderate \( N \) panels; asymptotically test requires \( T \to \infty \) followed by \( N \to \infty \)
- Test statistic is based on the OLS residuals
- Test requires a balanced panel
Pesaran (2003) Test

- Extends IPS to allow for cross-sectional dependence; referred to as CIPS
- Test statistic based on average of the cross-sectionally augmented DF (CADF) statistics obtained from each panel separately
- CADF regressions include the lagged cross-section mean, $\bar{y}_{t-1}$, and its FD, $\Delta \bar{y}_t$, in the ADF regressions
- Stata: -pescadf-
Nonstationary Panels

Spurious Regression

- Entorf (1997)
  - FE regressions involving independent random walks with and without drift
  - For $T \to \infty$ and $N$ fixed, standard FE results are highly misleading; significant, high $R^2$

- Phillips & Moon (1999)
  - Analyzed four cases:
    1. Panel spurious regression with no cointegration
    2. Heterogeneous panel cointegration (vectors vary across $i$)
    3. Homogeneous panel cointegration
    4. Near-homogeneous panel cointegration
  - Show that pooled OLS is consistent in each case using both sequential and joint limits
  - Intuition: information from independent cross-section data is valuable

- Choi (2013)

- My own simulations follow ...
Simulations

- DGP

\[
\begin{align*}
  y_{it} &= c_i + y_{it-1} + \varepsilon_{yt}, & t = 2, \ldots, T \\
  x_{it} &= c_i + x_{it-1} + \varepsilon_{xt}, & t = 2, \ldots, T \\
  y_{i1} &= c_i \\
  x_{i1} &= c_i \\
  \varepsilon_{xt}, \varepsilon_{yt} &\sim N(0, 1)
\end{align*}
\]

- Reps = 20,000

- Sample sizes

1. \(N = 10, \ T = 101\)
2. \(N = 10, \ T = 1001\)
3. \(N = 1000, \ T = 101\)
4. \(N = 1000, \ T = 1001\)

- Estimating equation

\[
y_{it} = c_i + \beta x_{it} + \varepsilon_{it}
\]
Note: $y = c + L.y + e_y$; $x = c + L.x + e_x$; $e_y, e_x \sim N(0,1)$. $N = 10$; $T = 101$; Reps = 20000.
Empirical Distribution of t-Statistics

Spurious Regression

Note: $y=L.y+ey; x=L.x+ex; \text{ey,ex} \sim N(0,1)$. $N = 10; T = 101; \text{Reps} = 20000.$
Empirical Distribution of Beta

Spurious Regression

Note: \( y = c + L.y + \epsilon_y; \quad x = c + L.x + \epsilon_x; \quad \epsilon_y, \epsilon_x \sim N(0,1). \quad N = 10; \quad T = 1001; \quad \text{Reps} = 20000. \)
Note: $y = \text{L.y} + \text{ey}; \quad x = \text{L.x} + \text{ex}; \quad \text{ey}, \text{ex} \sim \text{N(0,1)}. \quad N = 10; \quad T = 1001; \quad \text{Reps} = 20000.$
Note: $y = c + Ly + ey$; $x = c + Lx + ex$; $ey, ex \sim N(0,1)$. $N = 1000$; $T = 101$; Reps = 20000.
Note: $y = \text{L}y + \text{ey}; x = \text{L}x + \text{ex}; \text{ey}, \text{ex} \sim N(0, 1)$. $N = 1000; T = 101; \text{Reps} = 20000.$
Empirical Distribution of Beta

Spurious Regression

Note: $y = c + L.y + e_y; x = c + L.x + e_x; e_y, e_x \sim N(0, 1)$. $N = 1000; T = 1001; Reps = 20000$. 

Density

Beta
Note: $y = \mu + \epsilon_y; x = \mu + \epsilon_x; \epsilon_y, \epsilon_x \sim N(0,1)$. $N = 1000; T = 1001; \text{Reps} = 20000$. 

Spurious Regression
Nonstationary Panels

Cointegration

- Model

\[ y_{it} = \beta_i x_{it} + z_{it} \gamma_i + \varepsilon_{it} \]

where \( y \) and \( x \) are \( I(d) \), \( d > 0 \), variables

\[ z_{it} = \begin{cases} 
1 & \Rightarrow \text{FE model} \\
0 & \\
[1 \ t] & \Rightarrow \text{FE model with panel-specific time trend} 
\end{cases} \]

and \( \varepsilon_{it} \) is \( I(d - 1) \) iff \( y \) and \( x \) are cointegrated

- Homogeneity or poolability refer to whether the cointegrating vector is common across panels, \( \beta_i = \beta \)

- Tests differ depending on whether they allow for cross-sectional dependence, structural breaks, etc.

- Error correction models allow for estimation of adjustment speed conditional on cointegration
Test procedures

- **Kao (1999) Tests**
  - Based on DF- or ADF-type tests of the FE residuals
  - Different test statistics depending on assumptions concerning strict exogeneity
  - $H_0$: no cointegration

- **Residual-Based LM Test**
  - Developed in McCoskey & Kao (1998)
  - Based on residuals from Fully Modified OLS (FMOLS) or dynamic OLS (DOLS)
  - $H_0$: cointegration
  - Stata: `-xtdolshm-`

  - $H_0$: no cointegration
  - Stata: `-xtpedroni-`

- **Westerlund (2007)**
  - Stata: `-xtwest-`