Linear Panel Data Models

Michael R. Roberts

Department of Finance
The Wharton School
University of Pennsylvania

October 5, 2009

Example

- Link between crime and unemployment. Data for 46 cities in 1982 and 1987.
- Consider CS regression using 1987 data

$$\begin{array}{rcccl} \widehat{crimeRate} & = & \underset{(20.76)}{128.38} & - & \underset{(3.42)}{4.16 unem}, \quad R^2 = 0.033 \end{array}$$

- Higher unemployment *decreases* the crime rate (insignificantly)?!?!?!
- Problem = omitted variables
- Solution = add more variables (age distribution, gender distribution, education levels, law enforcement, etc.)
- Use like lagged crime rate to control for unobservables

Panel Data Approach

- Panel data approach to unobserved factors. 2 types:
  1. constant across time
  2. vary across time

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + a_i + u_{it}, \quad t = 1, 2$$

where $d2 = 1$ when $t = 2$ and 0 when $t = 1$

- Intercept for period 1 is β_0, for period 2 $\beta_0 + \delta_0$
- Allowing intercept to change over time is important to capture secular trends.
- a_i captures all variables that are constant over time but different across cross-sectional units. (a.k.a. **unobserved effect**, **unobserved heterogeneity**)
- u_{it} is **idiosyncratic error** or time-varying error and represents unobserved factors that change over time and effect y_{it}

Example (Cont)

- Panel approach to link between crime and unemployment.

$$crimeRate_{it} = \beta_0 + \delta_0 d78_t + \beta_1 unem_{it} + a_i + u_{it}$$

  where $d87 = 1$ if year is 1987, 0 otherwise, and a_i is an unobserved city effect that doesn't change over time or are roughly constant over the 5-year window.
- Examples:
  1. Geographical features of city
  2. Demographics (race, age, education)
  3. Crime reporting methods

Pooled OLS Estimation

- How do we estimate β_1 on the variable of interest?
- Pooled OLS. Ignore a_i. But we have to assume that a_i is $\perp$ to *unem* since it would fall in the error term.

$$crimeRate_{it} = \beta_0 + \delta_0 d78_t + \beta_1 unem_{it} + v_{it}$$

where $v_{it} = a_i + u_{it}$. SRF:

$$\begin{array}{cccccccc} \widehat{crimeRate} & = & 93.42 & + & 7.94 d87 & + & 0.427 unem, & R^2 = 0.012 \\ & & (12.74) & & (7.98) & & (1.188) & \end{array}$$

Positive coef on *unem* but insignificant

First Difference Estimation

- Difference the regression equation across time to get rid of fixed effect and estimate differenced equation via OLS.

$$\begin{aligned} y_{i2} &= (\beta_0 + \delta_0) + \beta_1 x_{i2} + a_i + u_{i2}, (t = 2) \\ y_{i1} &= \beta_0 + \beta_1 x_{i1} + a_i + u_{i1}, (t = 1) \end{aligned}$$

- Differencing yields

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i$$

  where Δ denotes period 2 minus period 1.
- Key assumption: $\Delta x_i \perp \Delta u_i$, which holds if at each time t, $u_{it} \perp x_{it} \forall t$. (i.e., strict exogeneity).
- This rules out lagged dependent variables.
- Key assumption: Δx_i must vary across some i

First Difference Example

- Reconsider crime example:

$$\begin{array}{rcccl} \widehat{crimeRate} & = & 15.40 & + & 2.22\Delta unem, \quad R^2 = 0.012 \\ & & (4.70) & & (0.88) \end{array}$$

- Now positive *and* significant effect of unemployment on crime
- Intercept $\implies$ crime expected to increase even if unemployment doesn't change!
- This reflects secular increase in crime rate from 1982 to 1987

Practical Issues

- Differencing can really reduce variation in x
- x may vary greatly in cross-section but Δx may not
- Less variation in explanatory variable means larger standard errors on corresponding coefficient
- Can combat by either
    1. Increasing size of cross-section (if possible)
    2. Taking longer differences (over several periods as opposed to adjacent periods)

Example

- Michigan job training program on worker productivity of manufacturing firms in 1987 and 1988

  $$scrap_i t = \beta_0 + \delta_0 y88_t + \beta_1 grant_i t + a_i + u_i t$$

  where i, t index firm-year, $scrap$ = scrap rate = # of items per 100 that must be tossed due to defects, $grant = 1$ if firm i in year t received job training grant.
- a_i is firm fixed effect and captures average employee ability, capital, and managerial skill...things constant over 2-year period.
- Difference to zap a_i and run 1st difference (FD) regression

  $$\begin{array}{rcccl} \Delta \widehat{scrap} & = & \underset{(0.405)}{-0.564} & - & \underset{(0.683)}{0.739}\Delta grant, \quad N = 54, R^2 = 0.022 \end{array}$$

- Job training grant lowered scrap rate but insignificantly

Example (Cont)

- Is level-level model correct?

$$\Delta \widehat{log(scrap)} = \underset{(0.097)}{-0.57} - \underset{(0.164)}{0.317 \Delta grant}, \quad N = 54, R^2 = 0.067$$

- Job training grant lowered scrap rate by 31.7% (or 27.2% = exp(-0.317) - 1).
- Pooled OLS estimate implies insignificant 5.7% reduction
- Large difference between pooled OLS and first difference suggests that firms with lower-ability workers (low a_i) are more likely to receive a grant.
- I.e., $Cov(a_i, grant_{it}) < 0$. Pooled OLS ignores a_i and we get a downward omitted variables bias

Program Evaluation Problem

- Let y = outcome variable, *prog* = program participation dummy.

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 prog_{it} + a_i + u_{it}$$

- Difference regression

$$\Delta y_{it} = \delta_0 + \beta_1 \Delta prog_{it} + \Delta u_{it}$$

- If program participation only occurs in the 2nd period then OLS estimator of β_1 in the differenced equation is just:

$$\hat{\beta}_1 = \overline{\Delta y_{treat}} - \overline{\Delta y_{control}} \tag{1}$$

- Intuition:
  1. $\Delta prog_{it} = prog_{i2}$ since participation in 2nd period only. (i.e., $\Delta prog_{it}$ is just an indicator identify the treatment group)
  2. Omitted group is non-participants.
  3. So β_1 measures the average outcome for the participants *relative* to the average outcome of the nonparticipants

Program Evaluation Problem (Cont)

- Note: This is just a difference-in-differences (dif-in-dif) estimator
- "Equivalent" model:

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 prog_{it} + \beta_2 d2_t \times prog_{it} + a_i + u_{it}$$

where β_2 has same interpretation as β_1 from above.

- If program participation can take place in both periods, we can't write the estimator as in (1) but it has the same interpretation: change in average value of y due to program participation
- Adding additional time-varying controls poses no problem. Just difference them as well. This allows us to control for variables that might be correlated with program designation.

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 prog_{it} + \gamma' X_{it} + a_i + u_{it}$$

Setup

- N individuals, $T = 3$ time periods per individual

$$y_{it} = \delta_1 + \delta_2 d2_t + \delta_3 d3_t + \beta_1 x_{it1} + ... + \beta_k x_{itk} + a_i + u_{it}$$

- Good idea to allow different intercept for each time period (assuming we have small T)
- Base period, $t = 1$, $t = 2$ intercept $= \delta_1 + \delta_2$, etc.
- If a_i correlated with any explanatory variables, OLS yields biased and inconsistent estimates. We need

$$Cov(x_{itj}, u_{is}) = 0 \forall t, s, j + ... + \beta_k x_{itk} + a_i + u_{it} \tag{2}$$

  (i.e., strict exogeneity after taking out a_i)
- Assumption (2) rules out cases where future explanatory variables react to current changes in idiosyncratic errors (i.e., lagged dependent variables)

Estimation

- If a_i is correlated with x_{itj} then x_{itj} will be correlated with composite error $a_i + u_{it}$
- Eliminate a_i via differencing

$$\Delta y_{it} = \delta_2 \Delta d2_t + \delta_3 \Delta d3_t + \beta_1 \Delta x_{it1} + \ldots + \beta_k \Delta x_{itk} + \Delta u_{it}$$

  for $t = 2, 3$
- Key assumptions is that $Cov(\Delta x_{itj}, \Delta u_{it}) = 0 \forall j$ and $t = 2, 3$.
- Note no intercept and time dummies have different meaning:

$$\begin{array}{l} t = 2 \implies \Delta d2_t = 1, \Delta d3_t = 0 \\ t = 3 \implies \Delta d2_t = -1, \Delta d3_t = 1 \end{array}$$

- Unless time dummies have a specific meaning, better to estimate

$$\Delta y_{it} = \alpha_0 + \alpha_3 \Delta d3_t + \beta_1 \Delta x_{it1} + \ldots + \beta_k \Delta x_{itk} + \Delta u_{it}$$

  for $t = 2, 3$ to help with R^2 interpretation

Setup

- N individuals, T time periods per individual

$$\begin{aligned} y_{it} &= \delta_1 + \delta_2 d2_t + \delta_3 d3_t + \ldots + \delta_T dT_t \\ &+ \ldots + \beta_1 x_{it1} + \ldots + \beta_2 x_{itk} + a_i + u_{it} \end{aligned}$$

- Differencing yields estimation equation

$$\begin{aligned} \Delta y_{it} &= \alpha_0 + \alpha_3 \Delta d3_t + \ldots + \alpha_T dT_t \\ &+ \beta_1 \Delta x_{it1} + \ldots + \beta_k \Delta x_{itk} + \Delta u_{it} \end{aligned}$$

for $t = 1, \ldots, T - 1$

Standard Errors

- With more than 2-periods, we must assume Δu_{it} is uncorrelated over time for the usual SEs and test statistics to be valid
- If u_{it} is uncorrelated over time & constant Var, then Δu_{it} is correlated over time

$$\begin{aligned} Cov(\Delta u_{i2}, \Delta u_{i3}) &= Cov(u_{i2} - u_{i1}, u_{i3} - u_{i2}) = -\sigma^2_{u_{i2}} \\ &\Longrightarrow Corr(\Delta u_{i2}, \Delta u_{i3}) = -0.5 \end{aligned}$$

- If u_{it} is stable AR(1), then Δu_{it} is serially correlated
- If u_{it} is random walk, then Δu_{it} is serially *un*correlated

Testing for Serial Correlation

- Test for serial correlation in the FD equation.
- Let $r_{it} = \Delta u_{it}$
- If r_{it} follows AR(1) model

$$r_{it} = \rho r_{i,t-1} + e_{it}$$

  we can test $H_0 : \rho = 0$ by
  1. Estimate FD model via pooled OLS and get residuals
  2. Run pooled OLS regression of $\hat{r}_{it}$ on $\hat{r}_{i,t-1}$
  3. $\hat{\rho}$ is consistent estimator of ρ so just test null on this estimate
  4. (Note we lose an additional time period because of lagged difference.)
- Depending on outcome, we can easily correct for serial correlation in error terms.

Chow Test

- Null: Do the slopes vary over time?
- Can answer this question by interacting slopes with period dummies.
- The run a Chow test as before.

Chow Test

- Can't estimate slopes on variables that don't change over time — they're differenced away.
- Can test whether partial effects of time-constant variables change over time.
- E.g., observe 3 years of wage and wage-related data

$$\begin{aligned} log(wage_{it}) &= \beta_0 + \delta_1 d2_t + \delta_2 d3_t + \beta_1 female_i + \gamma_1 d2_t \times female_i \\ &+ \gamma_2 d3_t \times female_i + \lambda X_{it} + a_i + u_{it} \end{aligned}$$

- First differenced equation

$$\begin{aligned} \Delta log(wage_{it}) &= \delta_1 \Delta d2_t + \delta_2 \Delta d3_t + \gamma_1 (\Delta d2_t) \times female_i \\ &+ \gamma_2 (\Delta d3_t) \times female_i + \lambda \Delta X_{it} + \Delta u_{it} \end{aligned}$$

- This means we can estimate how the wage gap has changed over time

Drawbacks

- First differencing isn't a panacea. Potential issues
  1. If level doesn't vary much over time, hard to identify coef in differenced equation.
  2. FD estimators subject to severe bias when strict exogeneity assumption fails.
     1. Having more time periods does *not* reduce inconsistency of FD estimator when regressors are not strictly exogenous (e.g., including lagged dep var)
  3. FD estimator can be worse than pooled OLS if 1 or more of explanatory variables is subject to measurement error
     1. Differencing a poorly measured regressor reduces its variation relative to its correlation with the differenced error caused by CEV.
     2. This results in potentially sizable bias

Fixed Effects Transformation

- Consider a univariate model

$$y_{it} = \beta_1 x_{it} + a_i + u_{it}, t = 1, 2, ..., T$$

- For each unit i, compute time-series mean.

$$\bar{y}_i \quad = \quad \beta_1 \bar{x}_i + a_i + \bar{u}_i, \text{ where} \bar{y}_i = (1/T)\Sigma y_{it}$$

- Subtract the averaged equation from the original model

$$\begin{aligned} (y_{it} - \bar{y}_i) &= \beta_1 (x_{it} - \bar{x}_i) + (u_{it} - \bar{u}_i), t = 1, 2, ..., T \\ \ddot{y}_{it} &= \beta_1 \ddot{x}_{it} + \ddot{u}_{it}, t = 1, 2, ..., T \end{aligned}$$

- $\ddot{z}$ represents **time-demeaned** data
- **Fixed Effect Transformation** $=$ **Within Transformation**

Fixed Effects Estimator

- We can estimate the transformed model using pooled OLS since it has eliminated the unobserved fixed effect a_i just like 1st differencing
- This is called **fixed effect estimator** or **within estimator**
- "within" comes from OLS using the time variation in y and x *within* each cross-sectional unit
- Consider general model

$$y_{it} = \beta_1 x_{it1} + ... + \beta_k x_{itk} + a_i + u_{it}, t = 1, 2, ..., T$$

- Same idea. Estimate time-demeaned model using pooled OLS

$$\ddot{y}_{it} \quad = \quad \beta_1 \ddot{x}_{it} + ... + \beta_k \ddot{x}_{itk} \ddot{u}_{it}, t = 1, 2, ..., T$$

Fixed Effects Estimator Assumptions

- We need strict exogeneity on the explanatory vars to get unbiased
- I.e., u_{it} is uncorrelated with each x across *all* periods.
- Fixed effect (FE) estimation, like FD, allows for arbitrary correlation between a_i and x in any time period
- FE estimation, like FD, precludes estimation of time-invariant effects that get killed by FE transformation. (e.g., gender)
- We need u_{it} to be homoskedastic and serially uncorrelated for valid OLS analysis.
- Degrees of Freedom is *not* $NT - k$, where $k = \#$ of xs.
    1. Degrees of Freedom $= NT - N - k$, since we lose one df for each cross-sectional obs from the time-demeaning.
    2. For each i, demeaned errors add up to 0 when summed across $t \implies$ 1 less df.
    3. This is like imposing a constraint for each cross-sectional unit. (There's no constraint on the original idiosyncratic errors.)

FE Implicit Constraints

- We can't include time-constant variables.
  1. Can interact them with time-varying variables to see how their effect varies over time.
- Including full set of time dummies (except one) $\Longrightarrow$ can't estimate effect of variables whose *change* across time is constant.
  1. E.g., years of experience will change by one for each person in each year. a_i accounts for average differences across people or differences across people in their experience in the initial time period.
  2. Conditional on a_i, the effect of a one-year increase in experience cannot be distinguished from the aggregate time effects because experience increases by the same amount for everyone!
  3. A linear time trend instead of year dummies would create a similar problem for experience

E.g., FE Implicit Constraints

- Consider an annual panel of 500 firms from 1990 to 2000
- Include full set of year indicators $\Longrightarrow$ can't include
    1. firm age
    2. macroeconomic variables
- These are all collinear with the year indicators and intercept.

Dummy Variable Regression

- We could treat a_i as parameters to be estimated, like intercept.
- Just create a dummy for each unit i.
- This is called **Dummy Variable Regression**
- This approach gives us estimates and standard errors that are identical to the within firm estimates.
- R^2 will be very high...lots of parameters.
- $\hat{a}_i$ may be of interest. Can compute from within estimates as:

$$\hat{a}_i = \bar{y}_i - \hat{\beta}_1 \bar{x}_{i1} - ... - \hat{\beta}_k \bar{x}_{ik}, i = 1, ..., N$$

  where $\bar{x}$ is time-average
- $\hat{a}_i$ are unbiased but inconsistent (**Incidental Parameter Problem**).
- Note: reported intercept estimate in FE estimation is just average of individual specific intercepts.

FE or FD?

- With $T = 2$, doesn't matter. They're identical
- With $T \geq 32$, FE$\neq$FD
- Both are unbiased under similar assumptions
- Both are consistent under similar assumptions
- Choice hinges on relative efficiency of the estimators (for large N and small T), which is determined by serial correlation in the idiosyncratic errors, u_{it}
  1. Serially uncorrelated $u_{it} \implies$ FE more efficient than FD and standard errors from FE are valid.
  2. Random walk $u_{it} \implies$ FD is better because transformed errors are serially uncorrelated.
  3. In between...efficiency differences not clear.
- When T is large and N is not too large, FE could be bad
- Bottom line: Try both and understand differences, if any.

FE with Unbalanced Panels

- **Unbalance Panel** refers to panel data where units have different number of time series obs (e.g., missing data)
- Key question: Why is panel unbalanced?
- If reason for missing data is uncorrelated with u_{it}, no problem.
- If reason for missing data is correlated with u_{it}, problem. This implies nonrandom sample. E.g.,
    1. Sample firms and follow over time to study investment
    2. Some firms leave sample because of bankruptcy, acquisition, LBO, etc. (**attrition**)
    3. Are these exit mechanisms likely correlated with unmeasured investment determinants (u_{it})? Probably.
    4. If so, then resulting **sample selection** causes biased estimators.
    5. Note, fixed effects allow attrition to be correlated with a_i. So if some units are more likely to drop out of the sample, this is captured by a_i.
    6. But, if this prob varies over time with unmeasured things affecting investment, problem.

Between Estimator

- **Between Estimator (BE)** is the OLS estimator on the cross-sectional equation:

$$\bar{y}_i = \beta_1 \bar{x}_{i1} + \ldots + \beta_k \bar{x}_{ik} + a_i + \bar{u}_i, \text{ where}$$

- I.e., run a cross-sectional OLS regression on the time-series averages
- This produces biased estimates when a_i is correlated with $\bar{x}_i$
- If a_i is uncorrelated with $\bar{x}_i$, we should use **random effects** estimator (see below)

R^2

- When estimating fixed effects model via FE, how do we interpret R^2?
- It is the amount of *time variation* in y_{it} explained by the *time variation* in X
- Demeaning removes all cross-sectional (between) variation prior to estimation

RE Assumption

- Same model as before

$$y_{it} = \beta_1 x_{it1} + ... + \beta_k x_{itk} + a_i + u_{it}, t = 1, 2, ..., T$$

- Only difference is that **Random Effects** assumes a_i is uncorrelated with each explanatory variable, $x_{itj}, j = 1, ..., k; t = 1, ..., T$

$$Cov(x_{itj}, a_i) = 0, t = 1, ..., T; j = 1, ..., k$$

- This is a very strong assumption in empirical corporate finance.

RE Cont.

- Under RE assumption:
  1. Using a transformation to eliminate a_i is *inefficient*
  2. Slopes β_j can be consistently estimated using a single cross-section...no need for panel data.
     1. This would be inefficient because we're throwing away info.
  3. Can use pooled OLS to get consistent estimators.
     1. This ignores serially correlation in composite error $(v_{it} = a_i + u_{it})$ term since

$$Corr(v_{it}, v_{is}) = \sigma_a^2/(\sigma_a^2 + \sigma_u^2), t \neq s$$

     2. Means OLS estimates give wrong SEs and test statistics.
     3. Use GLS to solve

RE and GLS Estimation

- Recall GLS under heteroskedasticity? Just transform data (e.g., divide by σ_{u_i}) and use OLS...same idea here
- Transformation to eliminate serial correlation is:

$$\lambda = 1 - [\sigma_u^2/(\sigma_u^2 + T\sigma_a^2)]^{1/2}$$

  which is $\in [0, 1]$
- Transformed equation is:

$$\begin{aligned} y_{it} - \lambda \bar{y}_i &= \beta_0(1 - \lambda) + \beta_1(x_{it1} - \lambda \bar{x}_{i1}) \\ &+ \ldots + \beta_k(x_{itk} - \lambda \bar{x}_{ik}) + (v_{it} - \lambda \bar{v}_i) \end{aligned}$$

  where $\bar{x}$ is time average.
- These are **quasi-demeaned data** for each variable...like within transformation but for λ

RE and GLS Comments

- Just run OLS on transformed data to get GLS estimator.
- FGLS estimator just uses a consistent estimate of λ. (Use pooled OLS or fixed effects residuals to estimate.)
- FGLS estimator is called **Random Effects Estimator**
- RE Estimator is biased, consistent, and anorm when N gets big and T is fixed.
- We can estimate coef's on time-invariant variables with RE.
- When $\lambda = 0$, we have pooled OLS
- When $\lambda = 1$, we have FE estimator.

RE or FE?

- Often hard to justify RE assumption ($a_i \perp x_{itj}$)
- If key explanatory variable is time-invariant, can't use FE!
- Hausman (1978) test:
  1. Use RE unless test rejects orthogonality condition between a_i and x_{itj}.
  2. Rejection means key RE assumption fails and FE should be used.
  3. Failure to reject means RE and FE are sufficiently close that it doesn't matter which is chosen.
  4. Intuition: Compare the estimates under efficient RE and consistent FE. If close, use RE, if not close, use FE.
- Bottom line: Use FE in empirical corporate applications.

Setup

- The model and approach in this section follows Bond 2002:

$$y_{it} = \rho y_{it-1} + a_i + u_{it}, |\rho| < 1; N = 1, ..., N; t = 2, ..., T$$

- Assume the first ob comes in $t = 1$
- Assume u_{it} is independent across i, serially uncorrelated, and uncorrelated with a_i.
    1. Within unit dependence captured by a_i
- Assume N is big, and T is small (typical in micro apps)
    1. Asymptotics are derived letting N get big and holding T fixed
- exogenous variables, x_{itk} and period fixed effects, v_t have no substantive impact on discussion

The Problem

- Fixed effects create endogeneity problem.
- Explanatory variable y_{it-1} is correlated with error $a_i + u_{it}$

$$\begin{aligned} Cov(y_{it-1}, a_i + u_{it}) &= Cov(a_i + u_{it-1}, a_i + u_{it}) \\ &= Var(a_i) > 0 \end{aligned}$$

- Correlation is $> 0 \implies$ OLS produces upward biased and inconsistent estimate of ρ (Recall omitted variables bias formula.)

$$Corr(y_{it-1}, a_i) > 0 \text{ and } Corr(y_{it}, y_{it-1}) > 0$$

- Bias does not go away as the number of time periods increases!

Within Estimator - Solve 1 Problem

- Within estimator eliminates this form of inconsistency by getting rid of fixed effect a_i

$$\ddot{y}_{it} \quad = \quad \beta_1 \ddot{y}_{it-1} + \ddot{u}_{it}, t = 2, ..., T$$

where

$$\ddot{y}_{it} = 1/T \sum_{i=2}^{T} y_{it}; \ddot{y}_{it-1} = 1/(T-1) \sum_{i=1}^{T-1} y_{it}; \ddot{u}_{it} = 1/T \sum_{i=2}^{T} u_{it}$$

Within Estimator - Create Another Problem

- Introduces another form of inconsistency since

$$Corr(\ddot{y}_{it-1}, \ddot{u}_{it}) \quad = \quad Corr(y_{it-1} - \frac{1}{T-1}\sum_{i=1}^{T-1} y_{it}, u_{it} - \frac{1}{T}\sum_{i=2}^{T} u_{it})$$

is not equal to zero. Specifically,

$$\begin{aligned} Corr(y_{it-1}, -\frac{1}{T-1}u_{it-1}) &< 0 \\ Corr(-\frac{1}{T-1}y_{it}, u_{it}) &< 0 \\ Corr(-\frac{1}{T-1}y_{it-1}, -\frac{1}{T-1}u_{it-1}) &> 0, t = 2, ..., T-1 \end{aligned}$$

- Negative corr dominate positive $\implies$ within estimator imparts negative bias on estimate of ρ. (Nickell (1981))
- Bias disappears with big T, but not big N

Bracketing Truth

- OLS estimate of ρ is biased up
- Within estimate of ρ is biased down
- $\implies$ true ρ will *likely* lie between these estimates. I.e., consistent estimator should be in these bounds.
- When model is well specified and this bracketing is *not* observed, then
    1. maybe inconsistency, or
    2. severe finite sample bias

  for consistent estimator

ML Estimators

- See Blundell and Smith (1991), Binder, Hsiao, and Pesaran (2000), and Hsiao (2003).
- Problem with ML in small T panels is that distribution of y_{it} for $t = 2, \ldots, T$ depends crucially on distribution of y_{i1}, initial condition.
- y_{i1} could be
  1. stochastic,
  2. non-stochastic,
  3. correlated with a_i,
  4. uncorrelated with a_i,
  5. specified so that the mean of the y_{it} series for each i is mean-stationary $(a_i/(1-\rho))$, or
  6. specified so that higher order stationarity properties are satisfied.
- Each assumption generates different likelihood functions, different estimates.
- Misspecification generates inconsistent estimates.

First Difference Estimator

- First-differencing eliminates fixed effects

$$\Delta y_{it} = \rho \Delta y_{it-1} + \Delta u_{it}, |\rho| < 1; i = 1, ..., N; t = 3, ..., T$$

  where $\Delta y_{it} = y_{it} - y_{it-1}$
- Key: first differencing doesn't introduce *all* of the realizations of the disturbance into the error term like within estimator. But,

$$Corr(\Delta y_{it-1}, \Delta u_{it}) = Corr(y_{it-1} - y_{it-2}, u_{it} - u_{it-1}) < 0$$

  $\Longrightarrow$ downward bias & typically greater than within estimator.
- When $T = 3$, within and first-difference estimators identical.
- Recall when $T = 2$ and no lagged dependent var, within and first-difference estimators identical.

IV Estimators 1

- Require weaker assumptions about initial conditions than ML
- Need **predetermined** initial conditions (i.e., y_{i1} uncorrelated with all future errors $u_{it}, t = 2, \dots T$.
- First-differenced 2SLS estimator (Anderson and Hsiao (1981, 1982)
- Need an instrument for Δy_{it} that is uncorrelated with Δu_{it}
- Predetermined initial condition + serially uncorrelated $u_{it} \implies$ lagged level y_{it-2} is uncorrelated with Δu_{it} and available as an instrument for Δy_{it-1}
- 2SLS estimator is consistent in large N, fixed T and identifies ρ as long as $T \geq 3$
- 2SLS is also consistent in large T, but so is within estimator

IV Estimators 2

- When $T > 3$, more instruments are available.
- y_{i1} is the only instrument when $T = 3$, y_{i1} and y_{i2} are instruments when $T = 4$, and so on.
- Generally, $(y_{i1}, ..., y_{t-2})$ can instrument Δy_{t-1}.
- With extra instruments, model is overidentified, and first differencing $\implies$ u_{it} is MA(1) if u_{it} serially uncorrelated.
- Thus, 2SLS is inefficient.

GMM Estimator

- Use GMM (Hansen (1982)) to obtain efficient estimates Hotz-Eakin, Newey, and Rosen (1988) and Arellano and Bond (1991).
- Instrument matrix:

$$Z_i = \begin{bmatrix} y_{i1} & 0 & 0 & \dots & 0 & \dots & 0 \\ 0 & y_{i1} & y_{i2} & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & 0 & \dots & \vdots \\ 0 & 0 & 0 & \dots & y_{i1} & \dots & y_{iT-2} \end{bmatrix}$$

  where rows correspond to first differenced equations for $t = 3, ..., T$ for individual i.
- Moment conditions

$$E(Z_i' \Delta u_i) = 0, i = 1, ..., N$$

  where $\Delta u_i = (\Delta u_{i3}, ..., \Delta u_{iT})'$

2-Step GMM Estimator

- GMM estimator minimizes

$$J_N = \left(\frac{1}{N} \sum_{i=1}^{N} \Delta u_i' Z_i \right) W_N \left(\frac{1}{N} \sum_{i=1}^{N} Z_i' \Delta u_i \right)$$

- Weight matrix W_N is

$$W_N = \left[\frac{1}{N} \sum_{i=1}^{N} \left(Z_i' \widehat{\Delta u_i} \widehat{\Delta u_i'} Z_i \right) \right]^{-1}$$

  where $\widehat{\Delta_i}$ is a consistent estimate of first-dif residuals from a preliminary consistent estimator.

- This is known as 2-step GMM.

1-Step GMM Estimator

- Under homoskedasticity of u_{it}, an asymptotically equivalent GMM estimator can be obtained in 1-step with

$$W_{1N} = \left[\frac{1}{N} \sum_{i=1}^{N} \left(Z_i' H Z_i \right) \right]^{-1}$$

  where H is $T-2$ square matrix with 2's on the diagonal, -1's on the first off-diagonals, and 0's everywhere else.
- Since W_{1N} doesn't depend on any unknowns, we can minimize the J_N in one step.
- Or, we can use this one step estimator to obtain starting values for the 2-step estimator.

GMM in Practice

- Most people use 1-step becase
  1. Modest efficiency gains from 2-step, even with heteroskedasticity
  2. Dependence of 2-step weight matrix on estimates makes asymptotic approximations suspect. (SEs too small). Windmeijer (2000) has finite sample correction for 2-step GMM estimator.
- $T > 3 \implies$ overidentification $\implies$ test of overidentifying restrictions, or Sargan test ($NJ_N \; \chi^2$).
- Key assumption of serially uncorrelated disturbances can also be tested for no 2nd order serial correlation in differenced residuals (Arellano and Bond (1991).
  - More instruments are not better because of IV bias
  - Negative 1st order serial correl expected in 1st differenced residuals if u_{it} is serially uncorr.
- See Bond and Windmeijer (2002) for more info on tests.

Extensions

- Intuition extends to higher order AR models & limited MA serial correlation of errors, provided sufficent # of time series obs. E.g,
    - u_{it} is MA(1) $\implies$ Δu_{it} is MA(2).
    - y_{it-2} is not a valid instrument, but y_{it-3} is.
    - Now we need $T \geq 4$ to identify ρ
- First-differencing isn't the only transformation that will work (Arellano and Bover (1995)).

Model

- The model now is

  $$y_{it} = \rho y_{it-1} + \beta x_{it} + a_i + u_{it}, |\rho| < 1; N = 1, ..., N; t = 2, ..., T$$

  where x is a vector of current and lagged additional explanatory variables.
- The new issue is what to assume about the correl between x and the error $a_i + u_{it}$.
- To make things simple, assume x is scalar and that the u_{it} are serially uncorrelated

Assumptions about x_{it} and $(a_i + u_{it})$

- If x_{it} is correl with a_i, we can fall back on transformations that eliminate a_i, e.g., first-differencing.
- Different assumptions about x and u
    1. x_{it} is endogenous because it is correlated with contemporaneous and past shocks, but uncorrelated with future shocks
    2. x_{it} is predetermined because it is correlated with past shocks, but uncorrelated with contemporaneous and future shocks
    3. x_{it} is strictly exogenous because it is uncorrelated with past, contemporaneous, and future shocks

Endogenous x_{it}

- In case 1, endogenous x_{it} then
  - x_{it} is treated just like y_{it-1}.
  - $x_{it-2}, x_{it-3}, \ldots$ are valid instruments for the first differenced equation for $t = 3, \ldots, T$
  - If y_{i1} is assumed predetermined, then we replace the vector $(y_{i1}, \ldots, y_{it-2})$ with $(y_{i1}, \ldots, y_{it-2}, x_{i1}, \ldots, x_{it-2})$ to form the instrument matrix Z_i
- In case 2, predetermined x_{it}
  - If y_{i1} is assumed predetermined, then we replace $(y_{i1}, \ldots, y_{it-2})$ with $(y_{i1}, \ldots, y_{it-2}, x_{i1}, \ldots, x_{it-1})$ to form instrument matrix Z_i
- In case 3, strictly exogenous x_{it}
  - Entire series, $(x_{i1}, \ldots x_{iT})$, are valid instruments
  - If y_{i1} is assumed predetermined, then we replace $(y_{i1}, \ldots, y_{it-2})$ with $(y_{i1}, \ldots, y_{it-2}, x_{i1}, \ldots, x_{iT})$ to form instrument matrix Z_i

In Practice

- Typically moment conditions will be overidentifying restrictions
- This means we can test the validity of a particular assumption about x_{it} (e.g., Difference Sargan tests)
- E.g., the moments assuming endogeneity of x_{it} are a strict subset of the moments assuming x_{it} is predetermined.
- We can look at difference in Sargan test statistics under these two assumptions, $(S - S') \ \chi^2$ to test validity of additional moment restrictions. (Arellano and Bond (1991))
- Additional moment conditions available if we assume x_{it} and a_i are uncorrelated. Hard to justify this assumption though.
- May assume that Δx_{it} is uncorrelated with a_i.
- Then Δx_{is} could be valid instrument for in levels equation for period t (Arellano and Bover (1995)

Difference Moments 1

- We could also used lagged differences, Δy_{it-1}, as instruments in the levels equation.
- Validity of this depends on stationarity assumption on initial conditions y_{i1} (Blundell and Bond (1998). Specifically,

$$E\left[\left(y_{i1} - \frac{a_i}{1-\rho}\right) a_i\right] = 0, i = 1, ..., N$$

- Intuitively, this means that the initial conditions don't deviate systematically from the long run mean of the time series.
- I.e., y_{it} converges to this value, $\frac{a_i}{1-\rho}$ from period 2 onward.

Difference Moments 2

- Mean stationarity implies $E(\Delta y_{i2}a_i)) = 0$ for $i = 1, ..., N$
- The autoregressive structure of the model and the assumption that $E(\Delta u_{it}a_i) = 0$ for $i = 1, ..., N$ and $t = 3, ..., T$ implies $T - 2$ non-redundant moment conditions

  $$E\left[\Delta y_{it-1}(a_i + u_{it})\right]$$

  for $i = 1, ..., N$ and $t = 3, ..., T$
- These moment conditions are in additon to those for the first-difference equations above, $E(Z_i'\Delta u_i) = 0$

Why extra moments are helpful 1

- Under additional assumptions, estimation no longer depends on just first-differenced equation and lagged level instruments.
- If the seris y_{it} is persistent (i.e., $\rho \approx 1$), then Δy_{it} is close to white noise
- This means the instruments, y_{it-2}, will be weak. i.e., weakly correlated with the endogenous variable Δy_{it-1}
- Alternatively, if $Var(a_i)/Var(u_{it}$ is large, then we will have a weak instrument problem as well.
- Consider

$$y_{it} = \rho y_{it-1} + a_i(1-\rho) + u_{it}$$

- As $\rho \rightarrow 1, y_{it}$ approaches a random walk and ρ is not identified using moment conditions for first-differenced equation, $E(Z_i \Delta u_i) = 0$