Empirical Methods

Topic 2d:

Panel Data

Panel Data

Latex Color = “LightGoldenrodYellow”

Panel Data Intro I

- So far we have dealt with *cross-sectional data*:
  - Observations on economic agents (e.g. individuals, households, firms, etc.) collected at one point in time.
  - Model given by

$$y_i = x_i'\beta + \epsilon_i, \qquad i = 1, \ldots, N \tag{1}$$

  and variation is over individuals i and not over time t.

- *Panel data* includes multiple observations on agents *over time*:
  - Individuals indexed by i, $i = 1, \ldots, N$
  - Time indexed by t, $t = 1, \ldots, T$

Panel Data Intro II

- With both i's and t's, we can think about generalizing (1)
  - The most common generalization is to allow for separate effects for both individuals, i, and time periods, t:

$$y_{it} = x_{it}'\beta + \alpha_i + d_t + \epsilon_{it}$$

  - where...

Panel Data Notation I

$$y_{it} = x_{it}'\beta + \alpha_i + d_t + \epsilon_{it} \tag{2}$$

where

- x_{it}' is a $1 \times K$ row vector containing variables that vary
  - Across i only (e.g. gender, education) and/or
  - Across i and t (e.g. experience)
- α_i, $i = 2, \ldots, N$, discussed further in the coming slides
- d_t, $t = 2, \ldots, T$ is a vector of time intercepts
  - Excluding one to prevent multi-collinearity with the constant (inside x_{it})
  - (I'll show why this is necessary in a few slides)
- ϵ_{it} a time-specific deviation from α_i

Short Panels

- The typical focus of panel-data methods is *short panels*, i.e.
  - Large N but small T
  - Thus we rely on "cross-section-type" arguments for consistency and asymptotic normality
    - e.g. T fixed, but $N \to \infty$
    - (That's fine - these are what I've shown you so far)

Panel Data Notation II

- Notation gets more complicated once we introduce panel data
  - As there are now two dimensions of variation, *i and t*
  - (This very important to understanding panel data methods!)
- So far we have written our estimating equation with double subscripts:

$$y_{it} = x_{it}'\beta + \alpha_i + d_t + \epsilon_{it}$$

- Let's now put this into matrix notation

Panel Data Notation III

- Stack the T observations for each individual i into its own vector

$$y_i = X_i\beta + \alpha_i \iota_T + \epsilon_i$$

- where we've subsumed d_t into X_i, ι_T is a $T \times 1$ vector of ones, and

$$y_i = \underbrace{\begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{bmatrix}}_{T \times 1} \quad X_i = \underbrace{\begin{bmatrix} x_{i11} & x_{i21} & \cdots & x_{iK1} \\ x_{i12} & x_{i22} & \cdots & x_{iK2} \\ \vdots & & & \vdots \\ x_{i1T} & x_{i2T} & \cdots & x_{iKT} \end{bmatrix}}_{T \times K} \quad \alpha_i \iota_T = \underbrace{\begin{bmatrix} \alpha_i \\ \alpha_i \\ \vdots \\ \alpha_i \end{bmatrix}}_{T \times 1} \quad \epsilon_i = \underbrace{\begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{iT} \end{bmatrix}}_{T \times 1}$$

- Note:
  - If we had enough observations for each i, we could in principle run separate regressions for each person!
    - (In which case we would drop d_t and α_i would be i's constant term)
  - This is rare except with Big Data applications - there it's common
    - Assume this away in what follows...

Panel Data Notation IV

- And then stack each of the N individuals:

$$y = X\beta + \alpha + \epsilon$$

- where

$$y = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{NT \times 1} \quad X = \underbrace{\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}}_{NT \times K} \quad \alpha = \underbrace{\begin{bmatrix} \alpha_1 \iota_T \\ \alpha_2 \iota_T \\ \vdots \\ \alpha_N \iota_T \end{bmatrix}}_{NT \times 1} \quad \epsilon = \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}}_{NT \times 1}$$

- And the properties of ϵ, e.g. $E(\epsilon|X)$ and $V(\epsilon)$, will be described in detail in what follows.

Aside: Dummy Variable Multicollinearity I

- I said earlier that we only include
  - $N-1$ dummy variables for each i
  - $T-1$ dummy variables for each t
- Showing why this is also gives us practice with panel data notation...
- Suppose you had 3 individuals and 2 time periods
- Let
  - $x_{1i} = 1$ be the constant term (as always)
  - d_t be a dummy variable for each time period
    - ★ For our example, we'd have d_1 and d_2

Aside: Dummy Variable Multicollinearity II

- Suppose we were to include a constant and *both* time dummies in our regression
- These three covariates for $i = 1, 2, 3$ and $t = 1, 2$ are given by:

$$\text{For} \begin{bmatrix} \\ x_{it} \\ \\ \end{bmatrix} = \begin{bmatrix} x_{11} \\ x_{12} \\ x_{21} \\ x_{22} \\ x_{31} \\ x_{32} \end{bmatrix} \quad \text{constant} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad d_1 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad d_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

- Do you see why this would be problematic?*
  - ______________________________

Aside: Dummy Variable Multicollinearity III

- More generally, whenever you add a set of dummy variables that span your data
  - e.g. men and women, all the individuals or years (in panel data)
- You have two choices:
  1. Drop the constant term, and both...
     - ★ Interpret each dummy variable as $E(y_i|x_i = 0, \text{that dummy} = 1)$
  2. Drop one of the dummy variables, and...
     - ★ Interpret the constant as $E(y_i|x_i = 0, \text{the } \underline{\text{excluded}} \text{ dummy} = 1)$
     - ★ Interpret each dummy variable as *the difference between* $E(y_i|x_i, \text{that dummy} = 1)$ and $E(y_i|x_i, \text{the excluded dummy} = 1)$
- Introduced for the first time here, but an important general principle!

Panel Data Examples I

- There are many examples of well-known and much-used panel datasets:
  - PSID: Panel Study of Income Dynamics (USA)
  - BHPS: British Household Panel Survey (UK)
  - ECHP: European Community Household Panel
  - GSOEP: German Socioeconomic Panel
  - SHP: Swiss Household Panel

Panel Data Examples II

- Looking at the determinants of a wide variety of outcomes:
  - Individuals' earnings
  - Households' expenditures
  - Firms' investments
  - Firms' productivities
  - Regions' migration patterns
  - Countries' income per capita

Unobserved Heterogeneity (α_i)

Latex Color = "LightGoldenrodYellow"

Unobserved Heterogeneity (α_i)

$$y_{it} = x_{it}'\beta + \alpha_i + d_t + \epsilon_{it}$$

- α_i plays a very important role in panel data analysis
  - And in econometrics more generally
- Note:
  - α_i is *unobserved*
    - It's measures tastes or costs that are (or might be) different for each individual in the sample....
    - That yield different choices of y_{it}.
  - It is constant across time
    - Note no t attached to it

α_i is called **Unobserved (individual-specific) Heterogeneity**

Example of Unobserved Heterogeneity and its Effects

- Unobserved heterogeneity can be very important
- For example, when analyzing the impact of aggregate (country-level) investment on income per capita
  - α_i could measure "good institutions":
    - ★ Countries with good institutions (e.g. infrastructure, rule of law) are likely to have high per-capita income
    - ★ This characteristic is likely to (often) be unchanging over time.
    - ★ It's perhaps (positively) correlated with firms' decisions to invest.
    - ★ $\Rightarrow$ ignoring it could yield an (upwardly) biased estimate of the effect of investment on income.
    - ★ (We will show this shortly...)

Unobserved Heterogeneity Bias I

How do you know there is bias from Unobserved Heterogeneity?

- If we ignore α_i in our regression, it effectively becomes part of the error term:

$$\begin{aligned} y_{it} &= x_{it}'\beta + \underbrace{d_t + \alpha_i}_{\text{Switched order!}} + \epsilon_{it} \\ &= x_{it}'\beta + d_t + \nu_{it} \end{aligned}$$

where $\nu_{it} = \alpha_i + \epsilon_{it}$ is the *composite error term* that includes α_i

Unobserved Heterogeneity Bias II

$$
\begin{aligned}
y_{it} &= x_{it}'\beta + d_t + \alpha_i + \epsilon_{it} \\
&= x_{it}'\beta + d_t + \nu_{it}
\end{aligned}
$$

- Since
  - α_i is part of the error term...
  - It may also correlated with one of our x's...
  - If so, we've violated our most important assumption: $E(\nu_i|x_i) = 0$
  - $\Rightarrow$ our estimate of β is biased!

Unobserved Heterogeneity Bias III

- Unobserved heterogeneity in this case is just an omitted variable
- And we know the formula for omitted variable bias, which here is:

$$
\boxed{
\begin{array}{rrcl}
\text{Truth:} & y & = & X\beta + (\gamma)\alpha + \epsilon \\
\text{You estimate:} & y & = & X\beta + \epsilon \\
\Rightarrow & E(\hat{\beta}) & = & \beta + \gamma(X'X)^{-1}X'\alpha \\
 & & = & \beta + \gamma\hat{\beta}_{\alpha\_on\_X}
\end{array}
}
$$

where

- $\gamma = \{-1, 1\}$ is the (implicit) sign of impact of α_i on y_i
- $\hat{\beta}_{\alpha\_on\_X}$ measures the correlation between α_i and x_i

Unobserved Heterogeneity Bias IV

$$E(\hat{\beta}) = \beta + \gamma\hat{\beta}_{\alpha\_on\_X}$$

- For our investment-income example:
  - We worry α_i measures "good institutions" and better institutions increase income ($\gamma > 0$)
  - We further worry that firms in countries with good institutions invest more ($\hat{\beta}_{\alpha\_on\_X} > 0$)
  - $E(\hat{\beta}_{inv}) = \beta_{inv} + (+)(+)$
    - ★ Meaning that we think their may be *positive* bias on our estimate of the impact of investment on income.

Aside: Bias in Cross-Section Settings I

- You may be thinking...
  - **Wait a minute!**
  - **Isn't Heterogeneity Bias relevant for Cross-Section data???**
  - (Why are we hearing about it only now???)

Aside: Bias in Cross-Section Settings II

- Heterogeneity Bias *is* relevant for cross-section analysis - it's an *i*-specific correlated unobservable.
  - Q: Why do you think we don't teach it before now?
  - A: ____?
    - ★
    - ★

Aside: Bias in Cross-Section Settings III

- The good news: Panel Data *is* rich enough to let us handle problems of unobserved heterogeneity.
  - It is for this reason that it is our second tool (after IV) to help resolve endogeneity issues in econometrics
  - With...
    - ★ Panel data, and
    - ★ A time-constant correlated unobservable
  - ...we can use panel data methods (especially Fixed Effects and/or First Differences) to resolve the endogeneity problem
- The coming slides show you how this works.

Panel Data Assumptions and Alternative Estimators

Panel Data Assumptions

Latex Color = "LightGrey"

Panel Data Assumptions Intro

- Before showing you how it works, I need to describe a few assumptions needed by these methods
- These include assumptions about:
  1. (Seen before:) Correlation between ϵ_{it} and ϵ_{js}
     - ★ i.e. Panel-data versions of the CLRM Assumptions 3 and 4 (Homoskedasticity and No Correlation)
  2. (Seen before:) Correlation between x_{it} and ϵ_{it}
     - ★ i.e. Panel-data versions of the CLRM Assumption 2 (Mean-zero error)
  3. (New:) Correlation between x_{it} and α_i
     - ★ New with panel data because α_i is new
- There is a lot here, so I will try to simplify as much as possible!

Panel Data Assumptions (1): $Cov(\epsilon_{is}, \epsilon_{it})$ I

- And so: do the easiest first
- Our baseline assumptions on heteroscedasticity and autocorrelation in ϵ_{it} across i and t are the same as for the CLRM:

$$\begin{aligned} (A3, \text{Homoskedasticity}): \quad Var(\epsilon_{it}) &= \sigma_{\epsilon}^2 \qquad t = 1, \ldots, T \\ (A4, \text{No Correlation}): \quad Cov(\epsilon_{is}, \epsilon_{it}) &= 0 \qquad s \neq t \end{aligned}$$

Panel Data Assumptions (1): $Cov(\epsilon_{is}, \epsilon_{it})$ II

- That being said,
  - The standard estimation methods presented below are consistent (for fixed T and $N \to \infty$) even if ϵ_{it} has arbitrary heteroscedasticity and/or serial correlation (that isn't too strong)
    - ★ (Serial correlation in ϵ_{it} appears to be particularly common)
  - Though they will of course impact the efficiency of an estimator
    - ★ (i.e. standard error calculations)

The other two panel data assumptions

- The other two panel data assumptions are the key ones:
  - (2) Between x_{it} and ϵ_{it}
  - (3) Between x_{it} and α_i
- Some of the most common panel data methods rely differentially on these two assumptions (e.g. Pooled OLS v Fixed Effects):
  - Strong on one and weak on the other
  - (A common pattern in econometrics)
- We will discuss this tradeoff once we introduce the two simplest versions (weak and strong) of each assumption

Panel Data Assumptions (2): $Cov(x_{it}, \epsilon_{it})$ I

- The two most common assumptions about x_{it} and ϵ_{it}:
  1. Contemporaneous Exogeneity:

$$Cov(x_{it}, \epsilon_{it}) = 0 \qquad t = 1, \ldots, T$$

     - ★ This is just the panel-data version of our normal CLRM assumption (A2, Mean-zero error)
     - ★ (And the weaker of our two assumptions re: $Cov(x_{it}, \epsilon_{it})$)

Panel Data Assumptions (2): $Cov(x_{it}, \epsilon_{it})$ II

- Two assumptions about x_{it} and ϵ_{it}, cont:
  2. Strict Exogeneity:

$$Cov(x_{is}, \epsilon_{it}) = 0 \qquad s, t = 1, \ldots, T$$

  - ★ The covariates at any time s are uncorrelated with the idiosyncratic errors at any time t
  - ★ This is a strong assumption...
  - ★ ...that is used by *almost all* of the standard panel data estimation methods...

Panel Data Assumptions (2): $Cov(x_{it}, \epsilon_{it})$ III

- Implications of Strict Exogeneity include
  - One must correctly specify the dynamic structure of the model
    - (e.g., the *exactly* correct lag structure of covariates)
  - One cannot have lagged dependent variables
    - (e.g., if $s = t + 1$, then including $y_{i,t-1}$ as a covariate implies including $x_{is} = y_{it}$, which is necessarily correlated with ϵ_{it})
  - Shocks today cannot affect future values of the covariates
    - (e.g. A good outcome today cannot change my investment tomorrow)
- $\Rightarrow$ Must evaluate how reasonable is this assumption in your application!
  - (Tho note often you're stuck with it)
  - (In which case must evaluate how much you believe your results under this assumption)

Panel Data Assumptions (3): $Cov(x_{it}, \alpha_i)$

- There are two common assumptions invoked about the relationship between x_{it} and α_i:
  1. Arbitrary Effects:
     - ★ No restrictions are placed on the relationship between x_{it} and α_i
     - ★ Not really an assumption; more like "no assumption".
     - ★ This is Very Good.
  2. Uncorrelated Effects:

$$Cov(x_{it}, \alpha_i) = 0 \qquad t = 1, \ldots, T$$

     - ★ This is Very Strong.
     - ★ (The whole point of worrying about unobserved heterogeneity is because you think this assumption is violated!)
- We will discuss each of these in more detail when introducing the estimators that rely on them

Key Panel Data Assumptions Overview

- To summarize:

Table: Strength of two key Panel Data Assumptions

	Weak(er)	Strong(er)
(2) $Cov(x_{it}, \epsilon_{it})$	Contemporaneous Exogeneity	Strict Exogeneity
(3) $Cov(x_{it}, \alpha_i)$	Arbitrary Effects	Uncorrelated Effects

Four Estimation Methods

Latex Color = "LightGrey"

Panel Data Estimation Methods Intro

- There are four common ways to estimate panel data models:
  1. Pooled OLS
     - ★ (The thing you already know... applied to panel data)
  2. Fixed Effects
     - ★ (The really useful estimator that can resolve some kinds of endogeneity)
  3. Random Effects
     - ★ (The most efficient estimator - but the one that relies on the strongest assumptions)
  4. First Differences
     - ★ (An estimator very similar to Fixed Effects)
- We'll cover each in turn

Pooled OLS

Latex Color = "LightGrey"

Pooled OLS I

- Our estimating equation for Pooled OLS is

$$\begin{aligned} y_{it} &= x_{it}'\beta + \alpha_i + \epsilon_{it} \\ &= x_{it}'\beta + \nu_{it} \end{aligned}$$

  - where we've subsumed d_t into x_{it} and $\nu_{it} = \alpha_i + \epsilon_{it}$ is, as before, a composite error term.

- As you can see, we are ignoring the unobserved heterogeneity, α_i:
  - It is just part of the composite error, $\nu_{it} = \alpha_i + \epsilon_{it}$.

Pooled OLS II

- To estimate, we run a simple OLS regression on all the data

$$\begin{aligned} \hat{\beta}_{POLS} &= (X'X)^{-1}X'y \\ V(\hat{\beta}_{POLS}) &= \sigma^2(X'X)^{-1} \end{aligned}$$

  - For X and Y defined on Slide 9

Pooled OLS III

- We know for OLS that the key thing for consistency and asymptotic normality is that we satisfy Assumption 2 (Mean-zero error)
  - $E(\nu_{it}|x_{it}) = 0$
- Because $\nu_{it} = \alpha_i + \epsilon_{it}$, to do so requires...
  - *Contemporaneous Exogeneity*...

$$Cov(x_{it}, \epsilon_{it}) = 0 \qquad t = 1, \ldots, T$$

  - ... and *Uncorrelated Effects*:

$$Cov(x_{it}, \alpha_i) = 0 \qquad t = 1, \ldots, T$$

    - (We'll discuss later how reasonable are these assumptions)
    - (As well as how to test the latter)

Pooled OLS IV*

- We can allow arbitrary heteroskedasticity and serial correlation in ν_{it}
  - We just must be sure to use variance formulas that accommodate that
- For example, let $\hat{v}_{it} = y_{it} - x_{it}'\hat{\beta}_{POLS}$
  - Then we should calculate standard errors with the formula:

$$\hat{V}(\hat{\beta}_{POLS}) = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}x_{it}'\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T} \hat{v}_{is}\hat{v}_{it}x_{is}x_{it}'\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}x_{it}'\right)^{-1}$$

  - (The more econometrics you do the more you get comfortable with formulas like this!)

Pooled OLS: Intuition

- After introducing each estimator, I'll briefly provide some intuition for it.
  - And after introducing them all, I provide some intuition about the tradeoffs between them
- The intuition for Pooled OLS is the easiest: it's the same intuition as for "regular old OLS"...
  - ...when applied to panel data
- The key addition relative to regular old OLS? Covered 2 slides ago:
  - Assumption 2 implies the "regular" Assumption 2 (Contemporaneous Exogeneity)
  - As well as the new addition of Uncorrelated Effects
    - (Which, to be fair, is also there with OLS - we just didn't talk about it!)

Fixed Effects

Latex Color = "LightGrey"

Fixed Effects I

- Fixed Effect estimation starts by *de-meaning* the data
- Let

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}, \quad \bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it}, \text{ and } \bar{\epsilon}_i = \frac{1}{T}\sum_{t=1}^{T} \epsilon_{it}$$

- Then

$$\begin{aligned} y_{it} &= x_{it}'\beta + \alpha_i + \epsilon_{it} \\ \Rightarrow \bar{y}_i &= \bar{x}_i'\beta + \alpha_i + \bar{\epsilon}_i \end{aligned}$$

where $\frac{1}{T}\sum_{t=1}^{T} \alpha_i = \frac{1}{T}(T\alpha_i) = \alpha_i$ as α_i doesn't vary with t

Fixed Effects II

$$y_{it} = x_{it}'\beta + \alpha_i + \epsilon_{it}$$
$$\bar{y}_i = \bar{x}_i'\beta + \alpha_i + \bar{\epsilon}_i$$

- Subtracting the second equation from the first yields

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)'\beta + \epsilon_{it} - \bar{\epsilon}_i$$

- Which is normally written as

$$\boxed{\ddot{y}_{it} = \ddot{x}_{it}'\beta + \ddot{\epsilon}_{it}}$$

  where
  - $\ddot{y}_{it} \equiv y_{it} - \bar{y}_i$ (similarly for $\ddot{x}_{it}$ and $\ddot{\epsilon}_{it}$)...
  - ...and α_i has dropped out as it is constant across time

Fixed Effects III

$$\boxed{\ddot{y}_{it} = \ddot{x}_{it}'\beta + \ddot{\epsilon}_{it}}$$

- To estimate, we simply run OLS on the (transformed) equation above
- Let
  - $\ddot{y}$ be the $NT \times 1$ vector with typical element $\ddot{y}_{it}$
  - $\ddot{X}$ be the $NT \times K$ matrix with typical element $\ddot{x}_{ikt}$
  - $\ddot{\epsilon}$ be the $NT \times 1$ vector with typical element $\ddot{\epsilon}_{it}$
- Then

$$\boxed{\hat{\beta}_{FE} = (\ddot{X}'\ddot{X})^{-1}\ddot{X}'\ddot{y}}$$

Fixed Effects IV

$$\boxed{\ddot{y}_{it} = \ddot{x}_{it}'\beta + \ddot{\epsilon}_{it}}$$

- Looking at $\ddot{x}_{it}$ and $\ddot{\epsilon}_{it}$, it's clear why we need strict exogeneity for consistency
- As always for consistency, we need

$$Cov(\ddot{x}_{it}, \ddot{\epsilon}_{it}) = 0$$

- But

$$\ddot{x}_{it} \equiv x_{it} - \frac{1}{T}(x_{i1} + \ldots + x_{IT}) \quad \text{and} \quad \ddot{\epsilon}_{it} \equiv \epsilon_{it} - \frac{1}{T}(\epsilon_{i1} + \ldots + \epsilon_{IT})$$

- Thus $Cov(\ddot{x}_{it}, \ddot{\epsilon}_{it}) = 0 \Rightarrow$
  - We need $Cov(x_{is}, \epsilon_{it}) = 0$ for *each pair* of x_{is} and ϵ_{it}
  - And that's Strict Exogeneity

Fixed Effects V

$$\boxed{\ddot{y}_{it} = \ddot{x}_{it}'\beta + \ddot{\epsilon}_{it}}$$

- Note we also have *Arbitrary Effects*,
  - i.e. The relationship between α_i and x_{it} is completely unrestricted
    - (Because α_i doesn't enter anywhere into the estimating equation)
    - (As we've differenced it out of our estimating equation)
    - (This is really good)

Fixed Effects VI

- Under homoskedasticity and no serial correlation...

$$\begin{aligned} V(\epsilon_i | x_i, \alpha_i) &= \sigma_\epsilon^2 I_T \\ \Rightarrow V(\hat{\beta}_{FE}) &= \sigma_\epsilon^2 (\ddot{X}'\ddot{X})^{-1} \end{aligned}$$

  - ...we can estimate the asymptotic variance of $\hat{\beta}_{FE}$ as

$$\hat{V}(\hat{\beta}_{FE}) = \hat{\sigma}_\epsilon^2 \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \ddot{x}_{it} \ddot{x}_{it}' \right)^{-1}$$

  - where $\hat{\ddot{\epsilon}}_{it} = \ddot{y}_{it} - \ddot{x}_{it}' \hat{\beta}_{FE}$ and

$$\boxed{\hat{\sigma}_\epsilon^2 = \frac{1}{N(T-1)-K} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{\ddot{\epsilon}}_{it}^2}$$

    - ★ (This is an important input into something we'll need later)

Least Squares Dummy Variable Estimator I

$$\begin{aligned} y_{it} &= x_{it}'\beta + \alpha_i + \epsilon_{it} \\ \ddot{y}_{it} &= \ddot{x}_{it}'\beta + \ddot{\epsilon}_{it} \end{aligned}$$

- An equivalent way to estimate a model with fixed effects...
  - (tho it takes longer - and sometimes *much longer* - in Stata or R)
- ...is to include dummy variables, a_i, for each of the α_i

$$y_{it} = x_{it}'\beta + a_i + \epsilon_{it}$$

  - This means estimating as many extra parameters as you have *i*
    - (Could be a lot!)
- Aka the *Least Squares Dummy Variable* (LSDV) estimator

Least Squares Dummy Variable Estimator II

$$y_{it} = x_{it}'\beta + a_i + \epsilon_{it}$$

- There is an important limitation of the LSDV estimator with short panels
  - i.e. large N and small T
- Can you guess what it is?*
- Despite this problem, $\hat{\beta}_{FE}$ *is* consistent for β

Least Squares Dummy Variable Estimator III

$$y_{it} = x_{it}'\beta + a_i + \epsilon_{it}$$

- The a_i in the LSDV estimator are sometimes called *nuisance parameters* or *incidental parameters*
  - Nuisance parameters $\equiv$
    - Parameters in the model we're not inherently interested in
  - Incidental parameters $\equiv$
    - Nuisance parameters that grow with the sample size
- The lack of consistent estimation of α_i in panel data models is often called the "*incidental parameters problem*"

Fixed Effects VII

- Regardless of the representation of the FE model,

$$\begin{aligned} \ddot{y}_{it} &= \ddot{x}_{it}'\beta + \ddot{\epsilon}_{it} \qquad \text{or} \\ y_{it} &= x_{it}'\beta + a_i + \epsilon_{it} \end{aligned}$$

- Because we've included a separate effect for each i,
  - We only rely on variation *within individuals over time* to identify β
- For this reason, de-meaning is called the *within transformation*
  - And Fixed Effects is called the *Within Estimator*

The Between Estimator I

- The Fixed Effects estimator relies on de-meaning the data
- We could, however, also run a regression on the "meaned data":

$$\begin{aligned}\bar{y}_i &= \bar{x}_i'\beta + \alpha_i + \bar{\epsilon}_i \\ \bar{y}_i &= \bar{x}_i'\beta + \bar{\nu}_i\end{aligned}$$

  where $\bar{\nu}_i = \alpha_i + \bar{\epsilon}_i$

- This is called the *Between Estimator* for panel data
  - Because we are only relying on variation *between individuals*
    - (Note this is a single cross-section...
    - ...as we've averaged across time for each individual...
    - ...thus only i and no t subscripts)

The Between Estimator II

- The Between Estimator isn't often used as we still have α_i in the error term
  - And if we're going to ignore unobserved heterogeneity, we might as well use the Random Effects estimator introduced next.
  - (As it's the most efficient estimator that assumes away the effects of unobserved heterogeneity)
- It is still useful, however, for a few things:
  - Understanding the sources of variation in the data
  - The relationship between the Fixed Effects and Random Effects estimator
    - (Which I won't teach but you sometimes see in textbooks)
  - And...

The Between Estimator III

$$\bar{y}_i = \bar{x}_i'\beta + \bar{\nu}_i$$

- We can use the Between Estimator to help estimate σ_α^2
  - Which we're going to need momentarily
- To get there, note

$$\begin{aligned} \bar{\nu}_i &= \alpha_i + \bar{\epsilon}_i \\ &= \alpha_i + \frac{1}{T}\sum_i \epsilon_{it} \\ \Rightarrow \sigma_{\bar{\nu}}^2 &= \sigma_\alpha^2 + \frac{1}{T^2}\sum_i \sigma_\epsilon^2 \\ &= \sigma_\alpha^2 + \frac{1}{T}\sigma_\epsilon^2 \end{aligned}$$

The Between Estimator IV

$$\bar{y}_i = \bar{x}_i'\beta + \bar{\nu}_i$$
$$\sigma^2_{\bar{\nu}} = \sigma^2_\alpha + \frac{1}{T}\sigma^2_\epsilon$$

- We then use this to estimate σ^2_α by:
  - Estimating the first line of equation above (by OLS), $\Rightarrow \hat{\beta}_{Between}$
  - $\Rightarrow \hat{\bar{\nu}}_i = \bar{y}_i - \bar{x}_i'\hat{\beta}_{Between}$
  - $\Rightarrow \hat{\sigma}^2_{\bar{\nu}} = \frac{1}{N-K}\sum_i \hat{\bar{\nu}}_i^2$
  - And we can use our estimate of $\hat{\sigma}^2_\epsilon$ from Slide 49) to calculate

$$\boxed{\hat{\sigma}^2_\alpha = \hat{\sigma}^2_{\bar{\nu}} - \frac{1}{T}\hat{\sigma}^2_\epsilon}$$

Fixed Effects: Intuition I

- The fixed effects estimator is a very important estimator in econometrics
- It's primary advantage is that it resolves - *completely* - concerns about (time-constant) unobserved heterogeneity
  - (What's the intuition for how?)
  - ____
- The consequence of this very attractive feature is that estimation of β must therefore only come from the time-series variation in the data
  - In essence, all of the cross-sectional variation is "used up" to identify the fixed effects
- This is both Good *and* Bad
  - The Good is described above

Fixed Effects: Intuition II

- The Bad comes in two flavors
- First, including FEs means we can't estimate any covariate that doesn't vary across time. At All!
  - For example, suppose we wanted to estimate the effects of education on wages
    - ★ If education is constant across time for each i
    - ★ Then the FE estimator *cannot* estimate $\beta_{Education}$!
- Second, adding FEs may wipe out *much* of the variation in the data
  - If there isn't much time-series variation within each i, you'll have very imprecise estimates of β
  - Such is life! This is a common tradeoff between the econometric goals of consistency and efficiency

Random Effects

Latex Color = "LightGrey"

Random Effects I

- The assumptions for the Random Effects estimator are the same as Pooled OLS,
  - But it is more efficient
- The reasons is that α_i induces *serial correlation* in the composite error term, $\nu_{it} = \alpha_i + \epsilon_{it}$
  - Even if we assume there is no serial correlation in ϵ_{it}, i.e. $E(\epsilon_{is}\epsilon_{it}) = 0$
  - (...along with $E(\epsilon_{it}, \alpha_i) = 0$, an often-weak assumption)

$$
\begin{aligned}
Cov(\nu_{is}, \nu_{it}) &= Cov(\alpha_i + \epsilon_{is}, \alpha_i + \epsilon_{it}) \\
&= Cov(\alpha_i, \alpha_i) \\
&= Var(\alpha_i) \\
&= \sigma_\alpha^2
\end{aligned}
$$

Random Effects II

- With this panel-induced serial correlation, we can always use the general formula for $\hat{V}(\hat{\beta}_{POLS})$ that accommodates it
  - (As we showed on Slide 42)
- But we can also do better
  - And get a more efficient estimator by modeling the serial correlation
- This is just Generalized Least Squares (GLS)
  - (Q: I didn't teach you GLS this semester, so what is it?)
  - (A: A way to get a more efficient estimator than OLS when there is heteroskedasticity and/or serial correlation and/or clustering in the error term)
  - (Here we have serial correlation induced by α_i)

Random Effects III

- The Random Effects estimator is a particular version of (F)GLS
- We assume, for $\nu_{it} = \alpha_i + \epsilon_{it}$, that

$$\begin{aligned} V(\epsilon_{it}) &= \sigma_\epsilon^2 \\ Cov(\epsilon_{is}, \epsilon_{it}) &= 0 \\ Cov(\alpha_i, \epsilon_{it}) &= 0 \end{aligned}$$

- *And*
    - Strict Exogeneity (as for FE, pretty strong):

    $$Cov(x_{is}, \epsilon_{it}) = 0 \qquad s, t = 1, \ldots, T$$

    - Uncorrelated Effects (as for Pooled OLS, very strong):

    $$Cov(x_{it}, \alpha_i) = 0 \qquad t = 1, \ldots, T$$

    - (These are strong assumptions; hope the efficiency gains are worth it!)

Random Effects IV

- Given these assumptions, for ν_i a $T \times 1$ vector of composite errors for i,

$$E(\nu_i \nu_i') = \Omega_i = \begin{bmatrix} \sigma_\alpha^2 + \sigma_\epsilon^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_\epsilon^2 & \cdots & \sigma_\alpha^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 + \sigma_\epsilon^2 \end{bmatrix}$$
$$= \sigma_\alpha^2 \iota_T \iota_T' + \sigma_\epsilon^2 I_T$$

where ι_T is a $T \times 1$ vector of ones and I_T a $T \times T$ identity matrix.

Random Effects V

- And the overall variance-covariance matrix for ν, a $NT \times 1$ vector is

$$E(\nu\nu') = \Omega \quad = \begin{bmatrix} \Omega_i & 0 & \cdots & 0 \\ 0 & \Omega_i & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Omega_i \end{bmatrix}$$

Random Effects VI

- The GLS estimator in this setting is given by:

$$\boxed{\hat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y} \quad \boxed{V(\hat{\beta}_{GLS}) = (X'\Omega^{-1}X)^{-1}}$$

- And the Feasible GLS estimator by:

$$\boxed{\hat{\beta}_{FGLS} = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}y} \quad \boxed{V(\hat{\beta}_{FGLS}) = (X'\hat{\Omega}^{-1}X)^{-1}}$$

  - Where the only difference between the two is that we estimate Ω with $\hat{\Omega}$ in the latter

- To implement, we only need an estimate of Ω

Random Effects VII

- Estimating $\Omega_i = \sigma_\alpha^2 \iota_T \iota_T' + \sigma_\epsilon^2 I_T$ requires estimates of σ_α^2 and σ_ϵ^2
- We have an estimate of σ_ϵ^2 from our Fixed Effects (Within Estimator):

$$\boxed{\hat{\sigma}_\epsilon^2 = \frac{1}{N(T-1)-K} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{\ddot{\epsilon}}_{it}^2}$$

- And we have an estimate of σ_α^2 from our Between Estimator:

$$\boxed{\hat{\sigma}_\alpha^2 = \hat{\sigma}_{\bar{\nu}}^2 - \frac{1}{T} \hat{\sigma}_\epsilon^2}$$

- Therefore it's quite easy to estimate $\hat{\Omega}$

Random Effects VIII*

- It turns out that there is a way to transform (weight) observations to "correct" for the structure of the error variance-covariance matrix
  - That yields the Random Effects estimator that would come out of the matrix formula
- To do so, let

$$
\begin{aligned}
\lambda &= 1 - \sqrt{\frac{\sigma_\alpha^2}{\sigma_\alpha^2 + T\sigma_\epsilon^2}} \\
\Rightarrow \hat{\lambda} &= 1 - \sqrt{\frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_\alpha^2 + T\hat{\sigma}_\epsilon^2}}
\end{aligned}
$$

Random Effects IX*

- Given $\hat{\lambda}$ we first transform the data

$$\begin{aligned}\tilde{y}_{it} &= y_{it} - \hat{\lambda}\bar{y}_i \\ \tilde{x}_{it} &= x_{it} - \hat{\lambda}\bar{x}_i \\ \tilde{\nu}_{it} &= \nu_{it} - \hat{\lambda}\bar{\nu}_i\end{aligned}$$

  - (This transformation ensures that $\tilde{\nu}_{it}$ has nice properties)
    - (Given our assumptions about α_i and ϵ_{it})

- Then the Random Effects estimating equation can be written as

$$\tilde{y}_{it} = \tilde{x}_{it}'\beta + \tilde{v}_{it}$$

  - OLS estimation of which yields the random effects estimator, $\hat{\beta}_{RE}$

Random Effects X

$$\tilde{y}_{it} = \tilde{x}_{it}'\beta + \tilde{v}_{it}$$
$$\equiv (y_{it} - \hat{\lambda}\bar{y}_i) = (x_{it} - \hat{\lambda}\bar{x}_i)'\beta + \tilde{\nu}_{it}$$
$$\lambda = 1 - \sqrt{\frac{\sigma_\alpha^2}{\sigma_\alpha^2 + T\sigma_\epsilon^2}}$$

Note:

- As $\sigma_\epsilon^2 \to \infty$ (or $T \to \infty$), the square root term $\to 0$ and $\lambda \to 1$, and
  - $\hat{\beta}_{RE} \to \hat{\beta}_{FE}$
- As $\sigma_\epsilon^2 \to 0$, the square root term $\to 1$ and $\lambda \to 0$, and
  - $\hat{\beta}_{RE} \to \hat{\beta}_{POLS}$
- $\Rightarrow$ You can (loosely) think of RE as a weighted combination of the POLS and FE estimators

Random Effects XI*

- The Random Effects estimator can fail to be the true GLS estimator for (at least) two reasons:
  1. $V(\nu_i)$ may not have the specified form
     - ★ In particular, there could be further unmodelled serial correlation
  2. Maybe $V(\nu_i) \neq V(\nu_i|x_i)$
- These aren't horrible
  - As long as strict exogeneity holds, the RE estimator is consistent
    - ★ And is likely to "do better" than Pooled OLS
    - ★ (i.e. be more efficient)
  - Tho we should be sure to still use robust variance estimation
    - ★ Even though in principle we're estimating a GLS model!

Random Effects: Intuition I

- What's the intuition for the Random Effects estimator?
  - Basically that it's a more efficient version of Pooled OLS
  - Where the efficiency gain comes from there being a common element, α_i, in the variance-covariance matrix of the econometric error
    - ★ (Which is $v_{it} = \alpha_i + \epsilon_{it}$)
- Accounting for the serial correlation in v_{it} induced by the presence of the constant-across-time α_i means RE makes "better use" of the information contained in the data compared to Pooled OLS
  - Shows up as smaller standard errors
  - Tho note it *also* changes the coefficient estimates, β!

Random Effects: Intuition II

- Efficiency is good, right? Right!
  - So it must come with some strings attached...
  - And so it does
- The key downside of the RE estimator is that it relies on the strongest assumptions
  - And if these assumptions are wrong, then it is inconsistent
  - (And that's Very Bad)

First Differences

Latex Color = "LightGrey"

First Differences I

- Our final estimator can be handled quickly
- Like the Fixed Effects estimator, it eliminates α_i
  - But instead of doing it by de-meaning the data...
  - ...it does so by taking differences in adjacent observations, i.e.

$$
\begin{aligned}
y_{it} &= x_{it}'\beta + \alpha_i + \epsilon_{it} \qquad t = 1, \ldots, T \\
y_{i,t-1} &= x_{i,t-1}'\beta + \alpha_i + \epsilon_{i,t-1} \qquad t = 2, \ldots, T \\
\Rightarrow \Delta y_{it} &= \Delta x_{it}'\beta + \Delta\epsilon_{it} \qquad t = 2, \ldots, T
\end{aligned}
$$

where $\Delta y_{it} \equiv y_{it} - y_{i,t-1}$ (and similarly for Δx_{it} and $\Delta\epsilon_{it}$)

  - Note we've lost the first observation in our dataset
  - (Tho researchers often cheat and assume that we had a y_{i0})

First Differences II

- The assumptions for consistency are similar to those for the FE estimator
- The least restrictive condition is:

$$Cov(\Delta x_{it} \Delta \epsilon_{it}) = 0 \qquad t = 2, \ldots, T$$

  - Which itself holds if

$$\begin{aligned} E(x_{it}\epsilon_{it}) &= 0 \\ E(x_{i,t-1}\epsilon_{it}) &= 0, \quad \text{and} \\ E(x_{i,t+1}\epsilon_{it}) &= 0 \end{aligned}$$

    - (Note this is slightly-weaker-version of *Strict Exogeneity*)
    - (This fact underlies much of dynamic panel data estimation)

- As for the FE estimator, we again have *Arbitrary Effects*,
  - i.e. The relationship between α_i and x_{it} is completely unrestricted

First Differences III

$$\Delta y_{it} = \Delta x_{it}'\beta + \Delta\epsilon_{it} \qquad t = 2, \ldots, T$$

- We estimate the FD estimator by OLS on the equation above
- Since we have a differenced error term,

$$\begin{aligned} V(\Delta\epsilon_{it}) &= V(\epsilon_{it} - \epsilon_{i,t-1}) \\ &= 2\sigma_\epsilon^2 \\ Cov(\Delta\epsilon_{i,t-1}, \Delta\epsilon_{it}) &= E(\epsilon_{i,t-1} - \epsilon_{i,t-2})(\epsilon_{it} - \epsilon_{i,t-1}) \\ &= -\sigma_\epsilon^2 \\ \Rightarrow Corr(\Delta\epsilon_{i,t-1}, \Delta\epsilon_{it}) &= -0.5 \end{aligned}$$

  - ... and we should allow for heteroskedasticity and (especially) serial correlation when calculating standard errors

First Differences: Intuition

- The intuition for First Differences is the same as for Fixed Effects
  - You solve the problem of unobserved heterogeneity by differencing the data
  - For FE, you subtract the mean
  - For FD, you subtract the previous observation
    - (Six of one, half-a-dozen of the other...)
    - (i.e., this isn't a big conceptual differences)
- You'll be safe if you simply think of FD as a version of FE

Mapping Estimators to their Assumptions

Latex Color = "LightGrey"

Mapping Estimators to Assumptions I

- Recall our key panel data assumptions:

	Weak	Strong
(2) $Cov(x_{it}, \epsilon_{it})$	Contemporaneous Exogeneity	Strict Exogeneity
(3) $Cov(x_{it}, \alpha_i)$	Arbitrary Effects	Uncorrelated Effects

- How do each of our four panel data estimators rely on each of these?

	Assumption on $Cov(x_{it}, \epsilon_{it})$	Assumption on $Cov(x_{it}, \alpha_i)$
Pooled OLS	Contemporaneous Exogeneity	Uncorrelated Effects
Fixed Effects	Strict Exogeneity	Arbitrary Effects
Random Effects	Strict Exogeneity	Uncorrelated Effects
First Differences	(slightly weaker) Strict Exogeneity	Arbitrary Effects

Mapping Estimators to Assumptions II

- Let's re-order our list, ranking estimators by the strength of the assumptions on which they rely
  - Where it's always better to have ______________ assumptions!
- And list also their efficiency properties ...
  - (And combine FE with FD as they are so close conceptually)

Asumptions	Estimator	Assumption on $Cov(x_{it}, \epsilon_{it})$	Assumption on $Cov(x_{it}, \alpha_i)$	Efficiency
Stronger	Random Effects	Strict Exogeneity	Uncorrelated Effects	Most efficient
$\downarrow$	Pooled OLS	Contemporaneous Exogeneity	Uncorrelated Effects	$\uparrow$
Weaker	Fixed Effects / First Differences	Strict Exogeneity	Arbitrary Effects	Least efficient

Mapping Estimators to Assumptions III

Asumptions	Estimator	Assumption on $Cov(x_{it}, \epsilon_{it})$	Assumption on $Cov(x_{it}, \alpha_i)$	Efficiency
Stronger	Random Effects	Strict Exogeneity	Uncorrelated Effects	Most efficient
$\downarrow$	Pooled OLS	Contemporaneous Exogeneity	Uncorrelated Effects	$\uparrow$
Weaker	Fixed Effects / First Differences	Strict Exogeneity	Arbitrary Effects	Least efficient

- This table highlights a common lesson in econometrics:
  - There is a tradeoff between the strength of your assumptions and the efficiency of your estimator!
- And so...
  - Which panel data estimator should you use???
  - It turns out there is a test that can help you make the choice

Testing

Which Estimator to Use? I

- The previous table suggests that you should use one of only two (classes of) estimators
  - Fixed Effects / First Differences
    - (If you're worried about the bias from unobserved heterogeneity)
    - (As α_i is differenced out/estimated)
    - (And is thus not in the error term)
  - Random Effects
    - (If you're not worried about U.H. bias)
    - (As it's most efficient when not)
- (Usually don't bother with Pooled OLS...
  - ... unless very tworried about Strict Exogeneity of x_{it} and ϵ_{it})

Which Estimator to Use? II

- How shall we choose between them?
  - Using straightforward logic...
- We often care most about two different properties of estimators
  1. Unbiasedness/Consistency
  2. Efficiency
     - ★ In *that* order!
- So the goal is the most efficient consistent estimator.
  - Which is that between FE and RE?
    - ★ *

Testing Uncorrelated Effects I

- To determine the best estimator, we must test whether or not the uncorrelated effects assumption in the RE estimator is satisfied
  - i.e. whether $Cov(x_{it}, \alpha_i) = 0$
- We test the assumption by comparing the (time-varying) coefficients in the FE and RE models
  - Called *The Hausman Test*

Testing Uncorrelated Effects: Intuition

- OK, that's not obvious. What's the intuition of the Hausman Test?
- Simple! We know that (under its weaker assumptions) the FE estimator is always consistent
  - But the RE estimator is only consistent if the Uncorrelated Effects assumption is true
- Thus compare the two estimates!
  - If Uncorrelated Effects is true, the two sets of estimates shouldn't be "too different"
  - That's what the Hausman Test does

Testing Uncorrelated Effects II

- The "classic" Hausman Test computes a quadratic form in the difference in FE and RE coefficients:

$$H = (\hat{\beta}_{FE} - \hat{\beta}_{RE})'(\hat{V}_{FE} - \hat{V}_{RE})^{-}(\hat{\beta}_{FE} - \hat{\beta}_{RE})$$

where $(\cdot)^{-}$ is the generalized matrix inverse

  - This is distributed as a χ^2_K under the standard Random Effects assumptions
  - (where recall β is a $K \times 1$ column vector)

Testing Uncorrelated Effects III

- If we're only interested in one of the elements of β,
  - E.g., one of the x_{it} is a key policy variable while others are simply control variables
  - Then we can write the test focusing on that variable as:

$$H = \frac{\hat{\beta}_{k,FE} - \hat{\beta}_{k,RE}}{\{se(\hat{\beta}_{k,FE})^2 - se(\hat{\beta}_{k,RE})^2\}^{1/2}}$$

Testing Uncorrelated Effects IV

- This version of the test has fallen out of favor a bit,
  - In part because it requires homoskedasticity and no serial correlation in the errors
    - ★ This gives the simplified form of $(\hat{V}_{FE} - \hat{V}_{RE})$ without worrying about covariance terms
    - ★ But these assumptions are often violated in typical datasets

Testing Uncorrelated Effects V

- The current best method to run the Hausman Test is to use a regression approach...
  - ...that allows the nesting of both the FE and RE estimators
- Suppose we are interested in a subset of the elements in x'_{it}
  - Call these w'_{it}, a $1 \times M$ row vector
  - (Cannot include the time dummies)
    - (Which is fine - these usually aren't parameters of interest)

Testing Uncorrelated Effects VI

- Wooldridge (2012, Section 10.7.3) shows how to implement the Hausman test as an F-test
- The Restricted model is the Random Effects specification:

$$\tilde{y}_{it} = \tilde{x}_{it}'\beta + \tilde{\nu}_{it}$$

where

- $\tilde{y}_{it} \equiv (y_{it} - \hat{\lambda}\bar{y}_i)$,
- $\hat{\lambda} = 1 - \sqrt{\frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_\alpha^2 + T\hat{\sigma}_\epsilon^2}}$, and
- $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_\epsilon^2$ were defined on Slide 66

Testing Uncorrelated Effects VII

- The Unrestricted model is given by

$$y_{it} = x_{it}'\beta + \bar{w}_i'\xi + \nu_{it}^*$$

  where

  - $\bar{w}_i$ are the across-time mean values of w_{it} and
  - ν_{it}^* is an error term

- The clever algebraic trick is that the estimate of β when including $\bar{w}_i$ *is* the Fixed Effect estimator
  - (Not obvious - this is why econometricians get the big bucks!)

Testing Uncorrelated Effects VIII

- The nice thing is that we can test $\xi = 0$ allowing for arbitrary heteroskedasticity and serial correlation
- Simply run a t- or F-test of $H_0 : \xi = 0$ from the unrestricted model
  - Either with a Pooled OLS or RE approach
  - (With standard errors calculated appropriately)
  - (i.e. with robust and/or clustered standard errors)
    - (Each yield identical coefficients when $\bar{w}_i$ included)

Panel Data Example

Latex Color = "LightGoldenrodYellow"

Panel Data Example I

- Let's do an example
- It's a panel dataset of airfares in the U.S. market for airplane flights
  - $i = 1, \ldots, 1,149$ U.S. air routes
    - ★ 94 origin cities, 97 destination cities
  - $t = 1997, \ldots, 2000$
  - Dependent variable: $y_{it} = \log(fare_{it})$
    - ★ Measured as the average one-way fare on route i in period t
  - Key independent variable: $x_{it} = concentration_{it}$
    - ★ Measured as the passenger (market) share of the largest airline on i in t
    - ★ (Other measures are possible, indeed more common)
  - Other control variable: $distance_{it}$
    - ★ Measuring both cost and demand factors
    - ★ Include as a quadratic

Panel Data Example II

- Before we dig into the data, let's do some higher-level thinking:
  1. What does economic theory suggest about how prices should respond to increases in concentration?*
  2. Why might we be worried about unobserved heterogeneity, i.e....
     - ★ Why might we think air routes differ in important ways (α_i) that might be correlated with concentration (x_{it})?*
     - ★ What would be the sign of any bias in this case?*

Panel Data Example III

- OK, let's see whether the data confirm our prior beliefs

See Stata Output in Class*

  - (Basic Summary Statistics)
  - (Pooled OLS - interpret $\hat{\beta}_{conc}$?)
  - (Fixed Effects)
    - (What happened to $\hat{\beta}_{dist}$?)
    - (What happened to $\hat{\beta}_{conc}$?)
  - (Random Effects)
    - (What happened to $\hat{\beta}_{conc}$?)
    - (Compare POLS to RE - does GLS 'help'?)
    - (Run the Augmented Regression Hausman Test)

Additional Examples

- I think the best way to learn econometrics is to "do" it
- So let's go through more examples:
  - The student's thesis from the IV notes
    - Analyzing the impact of cross-border workers (CBW) on votes "against foreigners"
    - (as she also had panel data)
  - A recent paper by me, one of my PhD students, and a Marketing prof at UChicago
    - Analyzing the of "online word of mouth" (i.e. Twitter) on movies' success
    - (If time)

Student Thesis Redux I

$$v_{it} = \beta_1 + \beta_2 CBW_{it} + \tilde{x}_i' \tilde{\beta} + \epsilon_{it}$$

Recall where we left off with the IV results of this thesis

- We thought there could be either negative or positive bias on β_2:
  - Negative from reverse causality and/or a correlated unobservable measuring "local openness" that wasn't adequately captured by dummies for city/suburbs/rural areas
  - Positive from a correlated unobservable measuring local (Swiss-population) unemployment rates
- Tried a number of specfications and found a consistent story:
  - Positive/significant effects of CBW with OLS
  - Negative/insignificant effects of CBS with IV
    - Using distance to the border as an IV

Student Thesis Redux II

- Let's try using panel data methods - what do we get? See Stata results in class
  - Fixed effects: no effect of CBW share (like IV)
  - Random effects: pos and sig effects (like Pooled OLS)
- The Hausman test? Rejected!

Student Thesis Redux III

- And so the $64,000 Question: which results to “believe”?
  - Easy! All are consistent!
- Can never be sure, but based on the balance of evidence in this data...
  1. There are no causal effects of cross-border workers on “anti-foreign votes” in Ticino municipalities, and
  2. Looking at OLS results is likely to yield the false conclusion that there is a positive effect of cross-border workers on anti-foreign votes

Twitter on Movie demand

- [To come]
- (If time)

Table of Contents