Econ 312

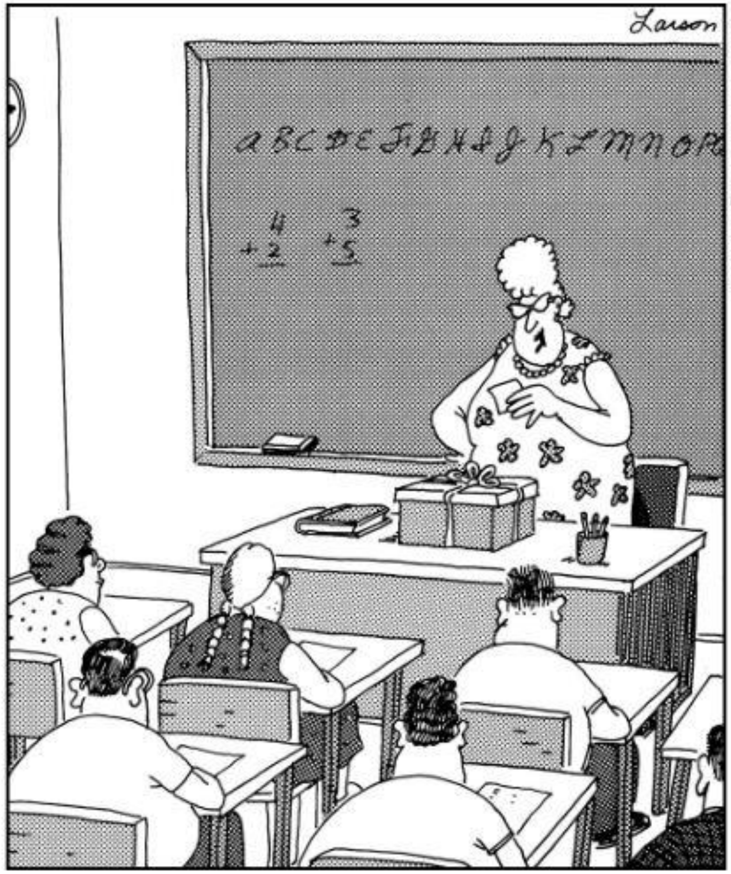
Wednesday, April 8

Pooled Samples and Fixed-Effects Models

Reading: Wooldridge, Chapter 13 and Section 14.1

Class notes: Pages 124 to 129

Today’s Far Side offering



“And the note says: ‘Dear classmates and Ms. Kilgore: Now that my family has moved away, I feel bad that I whined so much about being mistreated. Hope the contents of this box will set things right. Love, Pandora.’... How sweet.”

There’s no one in our class named Pandora, right? Right?

Context and overview

- Now that we know the principles of regressions with time series, we examine models for samples with both time and cross-sections
- We distinguish **pooled samples** from the more restrictive class of **panel samples**
- In this class we consider estimation methods appropriate for pooled samples and the **fixed-effects estimator** for panel data

Definitions

- **Pooled data** are a "time series of cross sections"
  - Cannot necessarily track same unit over time
  - Example: government surveys that sample different people each time
- **Panel data** are when we can identify individual units from one period to the next
  - Example: time-series data at the country, state, county, or city level
  - Some surveys (NLS-YM, PSID, etc.) track same individuals over time
- **Balanced vs. unbalanced panels**: Are there missing cells?
- **Asymptotic issues**: Is n getting large or is T getting large?

Pooled-data models

- Do the samples from each time period follow the **same model**?
- If so, we might just be able to do OLS on the pooled sample
- If not, we need to decide what is likely to vary
  - Just the level of the series? **Time dummies**
  - Slope coefficient of a variable *x*? **Interaction terms** with time dummies
  - Variance of the error? **Weighted LS** (or robust SEs)
  - Everything? **Separate regressions** for each period

Time dummies

- Include time dummies for all but one time period
- Constant is intercept for the omitted period
- Coefficients on dummy for any period is the difference between intercept in that period and in the omitted period
- Can test the dummies jointly to see if the intercept varies over time
- Test of individual dummy is whether that period’s intercept differs *from the omitted period*

Three-period time dummy example

$$y_{it} = \alpha_0 + \alpha_1 d_1 + \alpha_2 d_2 + \beta_1 x_{1it} + \beta_2 x_{2it} + u_{it}$$

- *d* variables are dummies for periods 1 and 2 (period 3 omitted)
- Intercepts for the periods are

  $\alpha_0 + \alpha_1$ for period 1

  $\alpha_0 + \alpha_2$ for period 2

  α_0 for period 3
- *F* test of all same is $H_0:\ \alpha_1 = 0,\ \alpha_2 = 0$
- *t* test of $H_0:\ \alpha_1 = 0$ tests whether intercept for periods 1 and 3 are equal

Interactions

- Interactions of time dummies with slope coefficient work the same way
- Coefficient on un-interacted variable is for omitted period
- Coefficients on interactions are differences between period and omitted period
- *F* test of all interactions tests null that all periods have same slope coefficient
- *t* test of individual interaction tests whether that period's slope is same as omitted period

Variables that change only over time

- **Aggregated variables** are same for each unit and vary only over time
- These would be **perfectly collinear** with time dummies, so cannot use such variables *and* time dummies
- Daily problem #31: GDP has one value for 2000 and one for 1990
  - Dummy for 2000 predicts this variable perfectly
  - Cannot include both GDP and dummy
- If we have T time series observations, can *at most* include only $T-1$ non-cross-sectionally-varying variables (without time dummies)

Panel data

- **Balanced panel** has data for each of n cross-sectional units for each of T periods (assume this today)
- Can we estimate by OLS with nT observations?
  - Yes, but potential issues with error term
  - And issues with constancy of coefficients across n and T
- Error term might have **different variance** in different periods t
- Error term might be **correlated between observations** with same unit i and different t
- Clustered standard errors will correct standard errors for these problems: Option **vce(cluster)** in Stata

Modeling differences across units *i*

- Most general model: all coefficients are different for each unit

$$y_{it} = \beta_{0i} + \beta_{1i} x_{1it} + \beta_{2i} x_{2it} + u_{it}$$

- There are 3*n* coefficients and *nT* observations
- Could just estimate *n* equations separately for each *i*
- Might be impractical for small *T*

Fixed-effects model

- What if only the intercept terms vary by unit?
- **Unit fixed-effects model**

$$y_{it} = \beta_{0i} + \beta_1 x_{1it} + \beta_2 x_{2it} + u_{it}$$

- Allows regression line for each unit (state, country, individual, etc.) to be shifted higher or lower, but no difference in slope
- Can be estimated two basically equivalent ways
  - LSDV: include **dummy variables** for each unit
  - **De-meaning** the variables for each unit (subtract unit mean from each observation)
- Dummy coefficients only get more precise as T increases (not n)

Least-squares with dummy variables

$$y_{it} = \sum_{j=1}^{n} \beta_{0j} D_{ji} + \beta_1 x_{1it} + \beta_2 x_{2it} + u_{it}$$

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

- β_{0j} is intercept term for unit j (constant omitted to avoid collinearity)
- $n + k = n + 2$ coefficients and nT observations
  - Not reliable if T is small
- Computationally difficult if n is large: need inverse of $n + k \times n + k$ matrix

Fixed effects via de-meaned data

$$y_{it} = \beta_{0i} + \beta_1 x_{1it} + \beta_2 x_{2it} + u_{it}$$

- Average across T periods for each i:

$$\frac{1}{T}\sum_{t=1}^{T} y_{it} = \beta_{0i} + \beta_1 \frac{1}{T}\sum_{i=1}^{T} x_{1it} + \beta_2 \frac{1}{T}\sum_{t=1}^{T} x_{2it} + \frac{1}{T}\sum_{t=1}^{T} u_{it}$$

$$\bar{y}_i = \beta_{0i} + \beta_1 \bar{x}_{1i} + \beta_2 \bar{x}_{2i} + \bar{u}_i$$

- Subtracting unit means from original equation gives “within-unit estimator”:

$$y_{it} - \bar{y}_i = \beta_1 \left(x_{1it} - \bar{x}_{1i} \right) + \beta_2 \left(x_{2it} - \bar{x}_{2i} \right) + \left(u_{it} - \bar{u}_i \right) \text{ or}$$

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{1it} + \beta_2 \ddot{x}_{2it} + \ddot{u}_{it}$$

Issues with fixed-effects estimator

- **Only uses within-unit variation** over time to estimate coefficients
- If an x varies *only* across units (sex or ethnicity), then we cannot estimate its coefficient in FE model
  - All of the de-meaned values will be zero
  - In LSDV, variable is collinear with unit dummies
- No constant term because all de-meaned variables have zero mean
  - Could estimate using "between-unit estimator" of unit means if we need a constant
- Degrees of freedom: Only $n(T-1)$ observations are independent
  - Should use correct denominator under SSR to estimate σ^2

Panel data in Stata

- Commands for panel data are prefixed by **xt**
- Need to define structure of data: **xtset unitvar timevar**
  - Both unitvar and timevar MUST be numeric
  - Cannot use state or country names, for example
- Basic regression with fixed effects
  - **xtreg y x1 x2 x3 , fe**
  - (Default is random effects, so need the fe option)
- Many other statistical commands are available in “xt” versions

Time fixed effects

- Can also allow **intercept to vary over time**, as in pooled-data model
- Usually easiest to do this with time dummies, but could also do by double-de-meaning
- This is the **differences-in-differences** model that we studied earlier
- **Three sources of variation** in sample:
    1. Within units over time
    2. Within time period across units
    3. Variation in differences over time between units
- The differences-in-differences estimator uses only #3

Review and summary

- **Pooled samples** have different cross-sections at multiple times
- Regressions with time dummies are most common estimator
  - Variables that do not vary across units cannot be used
- (Balanced) **panel samples** have the same sample units observed at multiple times
- Most common estimator for panel data is **unit fixed-effects** model
  - Variables that do not vary over time cannot be used
- Can also have **time fixed effects** with or without unit effects
  - **Differences-in-differences** estimator has both

Challenge for today

Take a common phrase and change one letter to make a new phrase that is meaningful. For example, I avoid the free samples at Costco under the principle of:

“**Taste not, want not.**”

Send me one that you come up with, or just add it to the conversation at the end of our conference.

What's next?

- In the next class, we will finish our brief analysis of panel data by considering the **random-effects model**
- We will also walk through an **example** of a panel-data application