Panel Data Methods

- $y_{it} = \beta_0 + \beta_1 x_{it1} + \ldots \beta_k x_{itk} + u_{it}$

A True Panel vs. A Pooled Cross Section

- Often loosely use the term panel data to refer to any data set that has both a cross-sectional dimension and a time-series dimension
- More precisely it's only data following the same cross-section units over time
- Otherwise it's a pooled cross-section

Pooled Cross Sections

- We may want to pool cross sections just to get bigger sample sizes
- We may want to pool cross sections to investigate the effect of time
- We may want to pool cross sections to investigate whether relationships have changed over time

See CPS78_85.doc for example on time effects

Difference-in-Differences

- Say random assignment to treatment and control groups, like in a medical experiment
- One can then simply compare the change in outcomes across the treatment and control groups to estimate the treatment effect
- For time 1,2, groups A, B $(y_{2,B} – y_{2,A}) - (y_{1,B} – y_{1,A})$, or equivalently $(y_{2,B} – y_{1,B}) - (y_{2,A} – y_{1,A})$, is the difference-in-differences

See KIELMC.doc file for example on the diff-in-diff estimator.

Difference-in-Differences (cont)

- A regression framework using time and treatment dummy variables can calculate this difference-in-difference as well
- Consider the model: $lprice_{it} = \beta_0 + \beta_1 nearinc_{it} + \beta_2 y81_{it} + \beta_3 nearinc_{it} * y81_{it} + u_{it}$
- The estimated β_3 will be the difference-in-differences in the group means

Difference-in-Differences (cont)

- Write the model in model general terminology:

$$y_{it} = \beta_0 + \beta_1 treatment_{it} + \beta_2 after_{it} + \beta_3 treatment_{it} * after_{it} + u_{it}$$

- The estimated β_3 will be the difference-in-differences in the group means

Difference-in-Differences (cont)

- When don't truly have random assignment, the regression form becomes very useful
- Additional *x*'s can be added to the regression to control for differences across the treatment and control groups
- Sometimes referred to as a "natural experiment" especially when a policy change is being analyzed

Two-Period Panel Data

- It's possible to use a panel just like pooled cross-sections, but can do more than that
- Panel data can be used to address some kinds of omitted variable bias
- If can think of the omitted variables as being fixed over time, then can model as having a composite error

Unobserved Fixed Effects

- Suppose the population model is $y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}$
- Here we have added a time-constant component to the error, $\upsilon_{it} = a_i + u_{it}$
- If a_i is correlated with the x's, OLS will be biased, since we a_i is part of the error term
- With panel data, we can difference-out the unobserved fixed effect

First-differences

- We can subtract one period from the other, to obtain $\Delta y_{it} = \delta_0 + \beta_1 \Delta x_{it1} + \ldots + \beta_k \Delta x_{itk} + \Delta u_{it,}$ where $\Delta y_{it} = y_{it} - y_{i,(t-1)}$ etc.
- This model has no correlation between the *x*'s and the error term, so no bias
- Need to be careful about organization of the data to be sure compute correct change

Differencing w/ Multiple Periods

- Can extend this method to more periods
- Simply difference adjacent periods
- So if 3 periods, then subtract period 1 from period 2, period 2 from period 3 and have 2 observations per individual
- Simply estimate by OLS, assuming the Δu_{it} are uncorrelated over time