PPOL 503-03, PPOL 503-04, Fall 2016
Course Notes #15: Pooled Cross Section Models and Simple Panel Data Models

I. INDEPENDENT POOLED CROSS SECTIONS

- Pooled cross-sectional data are random samples drawn at different times. For example, many surveys are repeated at regular intervals.

- Pooled cross-sectional data differ from panel data (a.k.a. longitudinal data) in that they do not follow the same individuals (or schools, firms, counties, etc.) across time.

- Pooled cross-sectional data likely lead to observations that are not identically distributed. We address this by including time period dummy variables, which allows the intercept term to vary across the time periods.

- Advantages of pooled cross-sectional data:
  - Increases the sample size.
  - Allows us to examine mean shifts in the outcome variable over time. That is, we can estimate what is the change in the expected value of the outcome variable over time, holding all other observable factors constant. Similarly, we can examine whether the effect of an independent variable on the outcome variable changes over time.
- Example 1 3.1: Determinants of Women's Fertility

Independent Variable	Coefficient	Standard Errors
educ	-0.128	0.018
age	-0.532	0.138
age^2	-0.0058	0.0016
black	1.076	0.174
east	0.217	0.133
northcen	0.363	0.121
west	0.198	0.167
farm	-0.053	0.147
othrural	-0.163	0.175
town	0.084	0.124
smcity	0.212	0.16
y74	0.268	0.173
y76	-0.097	0.179
y78	-0.069	0.182
y80	-0.071	0.183
y82	-0.522	0.172
y84	-0.545	0.175
constant	-7.742	3.052
n=1129		
R^2=.1295		

Base year is 1972. The coefficients on the year dummies show a sharp drop in fertility in the early 1980s. The coefficient on y82 implies that holding education, age and other factors constant, a woman had about .52 less children or about one-half a child in 1982 compared to 1972. This is a very large drop: 100 women in 1982 are predicted to have 52 fewer than 100 comparable women in 1972. The coefficients on Y82 and y84 are statistically significant.

- Example 13.2: Changes in the Return to Education and the Gender Wage Gap

$$\log(wage) = \beta_0 + \delta_0 y85 + \beta_1 educ + \delta_1 y85 \cdot educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 union + \beta_5 female + \delta_2 y85 \cdot female + u \quad (13.1)$$

The intercept for 1978 is β_0 and the intercept for 1985 is $\beta_0 + \delta_0$. The return to education in 1978 is β_1 and the return to education in 1985 is $\beta_1 + \delta_1$. This means that δ_1 captures how the return to another year of education has changed over the seven-year period. In 1978 the log(wage) differential between women and men is β_5 while the differential in 1985 is $\beta_5 + \delta_5$.

Hourly wage rate is expressed in nominal (or current) dollars. Since nominal wages grow simply due to inflation, we are really interested in measuring the effect of each explanatory variable on real wages. Suppose we decide to measure wages in 1978 dollars. This requires deflating 1985 wages to 1978 wages. The deflation factor is 107.6 / 65.2 = 1.65. Although we can divide each 1985 wage by 1.65, this is not necessary, provided a 1985 dummy is included in the regression and log (wage) is used as the dependent variable. Using real or nominal wages in the logarithmic functional form only affects the coefficient on the year dummy y85. The bottom line is when using the log (wage) as the dependent variable to examine how the return to education or the gender gap has changed over time, one does not need transform nominal wages into real wages. If we forget to allow for different intercepts in 1978 and 1985, the use of nominal wages can produce seriously misleading results. If one uses wage rather than log (wage), one must use the real wage and include a year dummy.

$$\begin{aligned} \log(wage) = {} & .459 + .118 y85 + .0747 educ + .0185 y85 \cdot educ \\ & (.093)\ (.123) \qquad (.0067) \qquad (.0094) \\ & + .0296 exper - .00040 exper^2 - .202 union \\ & (.0036) \qquad (.00008)\ \ 2 \qquad (.030) \qquad (13.2) \\ & - .317\, female + .085 y85 \cdot female \\ & (.037) \qquad (.051) \end{aligned}$$

The return to education in 1978 is estimated to be 7.5%; the return in 1985 is about 1.85 percentage points higher, or about 9.35%. The t statistic is about 1.97 which is significant at the 5% level.

As regards the gender gap, the results show that in 1978 a woman earned 31.7% less than a man (27.2% less is the more accurate estimator). In 1985 the gap in log(wages) is -.317 + .085 =-.232. The gender gap has fallen by 8.5 percentage points. The t statistic on the interaction is 1.67 which is significant at the 5% level one tailed test.

- Recall that we can apply the Chow test to test for structural changes across time. That is, if we have data for two different years, we interact the year dummy with each of the independent variables and then conduct an F-test on the joint significance of the year dummy and the interaction terms.
    - Frequently, we only test the joint significance of the interaction terms.
    - You can apply a Chow test for more than one period, but it is quite cumbersome.

II. TWO-PERIOD PANEL DATA ANALYSIS

- Panel data contain information on the same individuals (or schools, firms, counties, etc.) over time.

- The big advantage of a panel data set is that it allows us to control for unobservable time-invariant (fixed) characteristics that affect the outcome variable. If these unobservable fixed characteristics are correlated with the independent variables, then not accounting for them will lead to biased estimates.

- Consider the following regression equation:

  $$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + v_{it}$$

  Where *i* denotes the individual and *t* denotes the time-period (e.g., year), $d2_t$ is a dummy variable that equals zero in the first year and one in the second year.

- We consider v_{it} as the *composite error term*, where $v_{it} = a_i + u_{it}$. The variable a_i (known as the fixed effect) captures all of the unobserved, time-invariant factors that affect y_{it}. The variable u_{it} is called the *idiosyncratic* (or *time-varying*) error term. It represents all unobserved individual factors that change over time and affect y_{it}.

- If we pool our data and run OLS, in order to get unbiased coefficient estimates, we must assume that the error term is uncorrelated with x_{it}. A panel data set allows us to control for the unobservable fixed effect that may be correlated with the explanatory variables.

- If we have two years of data, we can estimate a first-differenced equation:

$$y_{i2} = (\beta_0 + \delta_0) + \beta_1 x_{i2} + a_i + u_{i2} \qquad (t = 2)$$
$$y_{i1} = \beta_0 + \beta_1 x_{i1} + a_i + u_{i1} \qquad (t = 1)$$
$$(y_{i2} - y_{i1}) = \delta_0 + \beta_1 (x_{i2} - x_{i1}) + (u_{i2} - u_{i1}),$$
$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i$$

  where Δ denotes the change from period 1 to period 2.

- Note how the unobservable fixed effect drops out of the estimation equation. We must still assume that the change in u_{it} is uncorrelated with changes in the independent variables. This assumption holds if the idiosyncratic error term at each time period is uncorrelated with the explanatory variable in both time periods.

- Note also that the first-differenced equation requires that x varies across time for the individual cross-sections. First-differencing reduces the variation in the sample because it removes all cross-sectional variation.

- Suppose we want to estimate the impact of city unemployment on city crime rate and we have data for 1982 and 1987. The pooled OLS estimates are as follows:

$$\widehat{crmrte} = 93.42 + 7.94 d87 + .427 unem$$
$$\qquad\qquad (12.74)\ (7.98) \qquad (1.188)$$
$$n = 92, R^2 = .012$$

- The first-differenced estimates are as follows:

```
. gen crmrte_lag=crmrte[_n-1]
(1 missing value generated)

. gen unem_lag=unem[_n-1]
(1 missing value generated)

. gen diffy=crmrte-crmrte_lag
(1 missing value generated)

. gen diffx=unem-unem_lag
(1 missing value generated)

. drop if year==82
(46 observations deleted)
. reg diffy diffx
```

```
. reg diffy diffx

      Source |       SS       df       MS              Number of obs =      46
-------------+------------------------------           F(  1,    44) =    6.38
       Model |  2566.43744     1  2566.43744           Prob > F      =  0.0152
    Residual |  17689.5497    44  402.035219           R-squared     =  0.1267
-------------+------------------------------           Adj R-squared =  0.1069
       Total |  20255.9871    45  450.133047           Root MSE      =  20.051

------------------------------------------------------------------------------
       diffy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       diffx |   2.217999   .8778658     2.53   0.015     .4487772    3.987222
       _cons |    15.4022   4.702117     3.28   0.002      5.92571     24.8787
------------------------------------------------------------------------------
```

- The first-difference model removes the omitted-variables bias caused by time-invariant unobservable characteristics.

- Another example (example 13.5) (Sleeping versus Working):

$$slpnap_{it} = \beta_0 + \delta_0 d81 + \beta_1 totwrk_{it} + \beta_2 educ_{it} + \beta_3 marr_{it} + \beta_4 yngkid_{it} + \beta_5 gdhlth_{it} + a_{it} + u_{it}, \quad t = 1, 2$$

$$\Delta slpnap_i = \delta_0 + \beta_1 \Delta totwrk_i + \beta_2 \Delta educ_i + \beta_3 \Delta marr_i + \beta_4 \Delta yngkid_i + \beta_5 \Delta gdhlth_i + \Delta u_i$$

$$\Delta \widehat{slpnap}_i = -92.63 + .227 \Delta totwrk_i + .024 \Delta educ_i$$
$$\qquad\qquad (45.87)\ (.036) \qquad\quad (48.759)$$
$$+104.21 \Delta marr_i + 94.67 \Delta yngkid_i + 87.58 \Delta gdhlth_i$$
$$\quad (92.86) \qquad (87.65) \qquad\quad (76.60)$$
$$n = 239, R^2 = .15$$

The coefficient on $\Delta totwrk$ indicates a tradeoff between sleeping and working. One more hour of work is associated with .227 (60) = 13.62 fewer minutes of sleeping (t = 6.31). No other variable except the intercept is significantly different from zero.

III. POLICY ANALYSIS WITH TWO-PERIOD PANEL DATA

- Just as we did with the pooled cross-sectional data, we can use panel data for program evaluation. The set-up is to have a sample of individuals from a time period before treatment occurs. We also have data on the same individuals post-treatment, with some of these individuals having received the treatment. This is similar to pooled cross-sections except that the individuals are the same for the pre- and post-treatment samples.

- Example: Michigan job training program in which *scrap* measures the number of items per 100 that must be scrapped due to defects and *grant* is a dummy variable measuring whether a firm received a job training grant. We have data for 1987 and 1988.

$$\log scrap_{it} = \beta_0 + \delta_0 y88_t + \beta_1 grant_{it} + a_i + u_{it}, \; t = 1,2$$

- Unobserved firm effects might contain average employee ability, and capital and managerial skills. However, these should be fairly constant over a two-year period.

- Differencing to remove the unobserved fixed effects yields:

$$\Delta \log(\hat{scrap}) = -0.057 - 0.317\Delta grant$$
$$(0.097) \quad (0.164)$$
$$n = 54, R^2 = 0.067$$

Having a training grant is estimated to lower the scrap rate by about 27.2%

Exp (.317) -1 = -.272.

- If treatment only occurs in the second period, then the coefficient estimate measures the difference-in-differences:

$$\hat{\beta}_1 = \overline{\Delta y}_{treat} - \overline{\Delta y}_{control}$$

IV. DIFFERENCING WITH MORE THAN TWO TIME PERIODS

- It's a relatively straightforward extension to use differencing equations with more than two time periods. For example, suppose we have N individuals and three time periods (so we have 3N observations).

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

- We typically include time period dummy variables as well as an intercept term. This allows a separate intercept for each time period.

- A critical assumption needed to obtain unbiased estimates is that the idiosyncratic error term is uncorrelated with each explanatory variable in each time period:

$$\mathrm{cov}(x_{itj}, u_{is}) = 0, \ \forall t, s, j$$

- Given this assumption ,the explanatory variables are exogenous after we eliminate the fixed effects (a_i). We eliminate the fixed effects by differencing adjacent time periods. That is, we subtract time period one

from time period two, and we subtract time period two from time period three:

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

for t = 2 and 3.

- Note that it is important to organize your data appropriately before differencing. You should also be careful not to subtract the last observation from person i-1 from the first observation of person i.

- Again, the key assumption is that the change in the idiosyncratic error term is uncorrelated with the changes in the explanatory variables.

- Note that the equation above does not contain an intercept term. That is problematic for computing R^2, so the first-differenced equation is usually written as:

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

- For more than three time periods, the estimating equations looks like:

$$\Delta y_{it} = \alpha_0 + \alpha_3 d3_t + \alpha_4 d4_t + \mathrm{K} + \alpha_T dT_{t+} + \beta_1 \Delta x_{it1} + \mathrm{K} + \beta_k \Delta x_{itk} + \Delta u_{it} \quad t = 2,3,\mathrm{K}\ ,T$$

- Example 13.9 (County Crime Rates in North Carolina): 90 counties for the years 1981 through 1987.
  The quantities in parentheses are the usual OLS standard errors. The quantities in brackets are the standard error robust to serial correlation and heteroskedasticity.
  The three probability variables—of arrest, conviction and serving time in prison—all have the expected signs and are statistically significant. A 1% increase in the probability of arrest is predicted to lower the crime rate by .33%. The average sentence variable has the correct sign but it is not statistically significant. The coefficient on police per capita is contrary to expectations. It says that a 1% increase in police per capita increases crime rates by about .4%. This finding is hard to believe. Two possible explanations: 1) crime rate is calculated from reported crimes—it might be that when there are additional police, more crimes are reported; 2) the police variables might be endogenous—counties might enlarge the police force when they expect crime rates to increase.

$$\Delta \log(\hat{crmrte}) = .008 - .100 d83 - .048 d84 - .005 d85$$
$$(.017)\ (.024)\quad (.024)\quad (.023)$$
$$[.014]\ [.022]\quad [.020]\quad [.025]$$
$$+.028 d86 + .041 d87 - .327 \Delta \log(prbarr)$$
$$(.024)\quad (.024)\quad (.030)$$
$$[.021]\quad [.024]\quad [.056]$$
$$-.022 \Delta \log(avgsen) + .398 \Delta \log(polpc)$$
$$(.022)\quad (.026)$$
$$[.025]\quad [.101]$$
$$n = 540, R^2 = .433, \bar{R}^2 = .422$$