Acknowledgement: These notes are adapted versions of those previously prepared by Daley Kutzman.

Related materials

- Wooldridge 4e, Ch. 13.3 through heterogenity bias (just a fancy name for a type of omitted variable bias)
- Handout #19 on two year and multi-year panel data
- Lecture on Tues 4/15
- Section on Wed 4/9 and Wed 4/16
- Question 4 on Daily Assignment #16, also see solutions posted on BSpace
- For a more advanced discussion of fixed effects you can review Wooldridge 4e, Ch. 14.1

1 The basics of panel data

We've now covered three types of data: cross section, pooled cross section, and panel (also called longitudinal). In a panel data set we track the unit of observation over time; this could be a state, city, individual, firm, etc.. To help you visualize these types of data we'll consider some sample data sets below.

Table 1.

indiv	year	wage	edu	exper	female
1	1990	3.10	11	2	1
2	1990	3.24	12	22	1
.	.	.	.	.	.
100	1990	5.30	12	7	0

Cross sectional data is a snapshot of a bunch of (randomly selected) individuals at one point in time. Table 1 provides an example of a cross sectional data set, because we only observe each house once and all of the observations are from the year 1990. Since we use i to index people, firms, cities, etc., the notation for cross sectional data is what you've seen before:

$$wage_i = \beta_0 + \beta_1 edu_i + \beta_2 exper_i + \beta_3 female_i + u_i$$

Table 2.

house	year	hprice	bdrms	bthrms	sqrft
1	2000	85,500	3	2.0	1600
2	2000	67,300	3	2.5	1400
.	.	.	.	.	.
100	2000	134,000	4	2.5	2000
101	2010	243,000	4	3.0	2600
102	2010	65,000	2	1.0	1250

In contrast, **pooled cross sectional data** is multiple snapshots of multiple bunches of (randomly selected) individuals (or states or firms or whatever) at many points in time. Table 2 is an example of a pooled cross-sectional data set because we only observe each house once (102 houses) but some of the observations are from the year 2000 while others are from the year 2010. We can use the same notation here as in cross section, indexing each person, firm, city, etc. by i. Suppose we have two cross sectional datasets from two different years; *pooling* the data means to treat them as one larger sample and control for the fact that some observations are from a different year:

$$hprice_i = \beta_0 + \beta_1 bdrms_i + \beta_2 bthrms_i + \beta_3 sqrft_i + \delta y2010_i + u_i$$

Table 3.

obs.	i	t	murder rate	pop density	police
1	1	2000	9.3	2.24	440
2	1	2001	11.6	2.38	471
3	2	2000	7.6	1.61	75
4	2	2001	10.3	1.73	75
.	.	.	.	.	
199	100	2000	11.1	11.1	520
200	100	2001	17.2	17.2	493

Finally, there is **panel data** which is more like a movie than a snapshot because it tracks particular people, firms, cities, etc. over time. Table 3 provides an example of a panel data set because we observe each city i in the data set at two points in time (the year 2000 and 2001). In summary, the data set has 100 cities but 200 observations. This particular panel data set is sometimes referenced as a 'balanced panel data set' because we observe every single city in both the year 2000 and 2001. However, if we observed some of the cities in the year 1999 but not all of them, then we would call it an 'unbalanced panel data set' (this distinction often isn't very important). With a panel data (balanced or unbalanced) we start indexing observations by t as well as i to distinguish between our observations of city i at various points in time:

$$murders_{it} = \beta_0 + \beta_1 pop_{it} + \beta_2 unemp_{it} + \beta_3 police_{it} + \alpha_i + \delta_t + u_{it}$$

where the α_i represents city fixed effects and the δ_t represents year fixed effects. In a nutshell, α_i can be thought of as shorthand for a set of binary (indicator) city variables each multiplied by their respective regression coefficients (that is, a binary variable for each city multiplied by its regression coefficient). Similarly, δ_t can be thought of as shorthand for a set of binary (indicator) year variables each multiplied by their respective regression coefficients (that is, a binary variable for each year multiplied by its regression coefficient). We'll consider this in more detail next.

Fixed Effects Regression

I suspect many of you may be confused about what this α_i term has to do with a dummy variable. It certainly looks strange, given that it's not attached to any variable! Let's consider a subset of our example panel data from Table 3, where the unit of observation is a city-year, and suppose we have data for 3 cities for 3 years—so 9 total observations in our dataset.

obs	i	t	murder rate	pop density	City1	City2	City3	Yr00	Yr01	Yr02
1	1	2000	9.3	2.24	1	0	0	1	0	0
2	1	2001	11.6	2.38	1	0	0	0	1	0
3	1	2002	11.8	2.42	1	0	0	0	0	1
4	2	2000	7.6	1.61	0	1	0	1	0	0
5	2	2001	10.3	1.73	0	1	0	0	1	0
6	2	2002	11.9	1.81	0	1	0	0	0	1
7	3	2000	11.1	6.00	0	0	1	1	0	0
8	3	2001	17.2	6.33	0	0	1	0	1	0
9	3	2002	20.3	6.42	0	0	1	0	0	1

Since we have multiple observations for each city, we can run the following regression:

$$murder_{it} = \beta_0 + \beta_1 popden_{it} + \alpha_2 City2 + \alpha_3 City3 + \delta_2 Yr01 + \delta_3 Yr02 + u_{it}$$

In this regression specification $City2$ and $City3$ are each indicator variables for cities 2 and 3 in the data set; notice I exclude an indicator variable for city 1 to avoid perfect multicollinearity. Likewise, $Yr01$ and $Yr02$ are indicator variables for the year 2001 and the year 2002.

How do we interpret β_1, α_2 or δ_2 here? To answer this question it is instructive to start with a different parameter, the intercept, β_0, which give us the average murder rate given zero values for all of the explanatory variables model. Note that is $City2 = 0$ and $City3 = 0$ then by process of elimination β_0 must be related to the murder rate in $City1$ (the city/category excluded from the regression). But that's not all, β_0 is also related to the murder rate in the base year 2000 because $Yr01 = 0$ and $Yr02 = 0$. Given this example, we have the following interpretations.

- δ_t estimates the common change/difference (to all cities) in the murder rate in year t relative to the year 2000, *controlling for population density and city-specific time-invariant characteristics (the city fixed effects)*. We call δ_t a year fixed effect precisely because the change is common to all cities in year t; in other words, the 'effect' of year t is 'fixed' across all cities.
- Similarly, α_i estimates the common change/difference (to all years) in the murder rate in city i relative to city 1, *controlling for population density and year-specific characteristics/shocks common to all cities (the year fixed effects)*. We call α_i a city fixed effect precisely because the difference is common to all years in city i; in other words, the 'effect' of city i is 'fixed' across all years.
- β_1 is the estimated effect of population density on crime, *controlling for city-specific time-invariant characteristics and year-specific shocks (the city and year fixed effects)*.

To see the interpretation of α_i more clearly, suppose we're *only* looking at observations from city 3 (i.e. $City2 = 0$ and $City3 = 1$):

$$murders_{3t} = \beta_0 + \beta_1 popden_{3t} + \alpha_2 \cdot 0 + \alpha_3 \cdot 1 + \delta_2 Yr01 + \delta_3 Yr02 + u_{3t}$$

This simplifies to the following:

$$murders_{3t} = \beta_0 + \beta_1 popden_{3t} + \alpha_3 + \delta_2 Yr01 + \delta_3 Yr02 + u_{3t}$$

This is where the α_i term comes from in a fixed effect regression! For any given cross sectional unit (i), which in this example is a city, the other terms with city dummies drop out and we only have the term with a dummy for that city, $\alpha_i City_i$ left. For fixed effect regressions, we simply save time by writing an α_i instead of writing out each dummy variable. You can imagine that if we had 85 cities instead of 3, writing out each dummy variable would get super tedious.

Now suppose we *only* look at observations from the year 2002 (i.e. $Yr01 = 0$ and $Yr02 = 1$):

$$murder_{i2} = \beta_0 + \beta_1 popden_{i2} + \alpha_2 City2 \ + \alpha_3 City3 \ + \delta_2 \cdot 0 \ + \delta_3 \cdot 1 \ + u_{it}$$

$$murder_{i2} = \beta_0 + \beta_1 popden_{i2} + \alpha_2 City2 \ + \alpha_3 City3 \ + \delta_3 + u_{it}$$

We can also write the time dummy variables in shorthand as δ_t. Note the subscripts on these variables: for a given city, its city dummy variable isn't going to vary by year, and for a given year, its year dummy variable isn't going to vary by city. So we often write this regression as:

$$murder_{it} = \beta_0 + \beta_1 popden_{it} + \alpha_i + \delta_t + u_{it}$$

To be consistent with the notation in Wooldridge and elsewhere we can also write:

$$murder_{it} = \beta_0 + \beta_1 popden_{it} + a_i + d_t + u_{it}$$

Because it's more conventional in the academic literature these days, I prefer reserving Greek for parameters (like regression coefficients which we typically estimate) and using the English alphabet to denote the outcome and explanatory variables. But it really doesn't matter.

Panel Regressions in STATA:

There are a few ways to implement a regression that controls for city and time effects (i.e. regression models with fixed effects). In these examples, I use a dataset about murder rates and unemployment rates across US states in the years 1987, 1990, and 1993.

1. $\widehat{mrdrte}_{it} = \hat{\beta}_0 + \hat{\beta}_1 unem_{it} + \underbrace{\alpha_2 State2 + ...\alpha_{50} State50}_{\text{Dummy for all but one state}} + \underbrace{\delta_1 Yr2001 + \delta_2 Yr2002}_{\text{Dummy for all but one year}} + u_{it}$

   In STATA (note that when we write `state_2 - state_51` STATA includes all variables appearing between `state_2` and `state_51` in the 'variable list'; be careful about ordering of your variable list when using this code):

```
reg mrdrte unem state_2 - state_51 year_2 year_3
```

```
      Source |       SS       df       MS              Number of obs =     153
-------------+------------------------------           F( 53,    99) =   17.75
       Model |  11622.5233    53  219.292892           Prob > F      =  0.0000
    Residual |  1222.81484    99   12.351665           R-squared     =  0.9048
-------------+------------------------------           Adj R-squared =  0.8538
       Total |  12845.3381   152  84.5088034           Root MSE      =  3.5145
------------------------------------------------------------------------------
      mrdrte |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        unem |   .2019432   .2947557     0.69   0.495    -.3829162    .7868025
     state_2 |   2.182073   2.886745     0.76   0.452    -3.545855    7.910001
     state_3 |   .7759888   2.897709     0.27   0.789    -4.973695    6.525672

----------------------Deleted some fixed effect results to save space---------
    state_51 |  -5.036179   2.927538    -1.72   0.089    -10.84505    .7726923
      year_2 |   1.577016   .7433858     2.12   0.036     .1019775    3.052055
      year_3 |   1.681938   .6959821     2.42   0.017     .3009584    3.062917
       _cons |   6.077295   3.300348     1.84   0.069    -.4713127     12.6259
------------------------------------------------------------------------------
```

2. $\widehat{mrdrte}_{it} = \hat{\beta}_1 unem_{it} + \underbrace{\alpha_1 State1 + ...\alpha_{50} State50}_{\text{Dummy for each state}} + \underbrace{\delta_1 Yr2001 + \delta_2 Yr2002}_{\text{Dummy for all but one year}} + u_{it}$

   In STATA (note that the 'noconstant' option tells STATA to note estimate an intercept; the idea is that if you don't exclude a state indicator variable then you can't estimate an intercept. Why? Because the intercept reflects an excluded state.):

```
reg mrdrte unem state_1 - state_51 year_2 year_3, noconstant
```

```
      Source |       SS       df       MS              Number of obs =     153
-------------+------------------------------           F( 54,    99) =   32.37
       Model |  21588.0857    54  399.779365           Prob > F      =  0.0000
    Residual |  1222.81484    99   12.351665           R-squared     =  0.9464
-------------+------------------------------           Adj R-squared =  0.9172
       Total |  22810.9006   153  149.090853           Root MSE      =  3.5145
------------------------------------------------------------------------------
      mrdrte |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        unem |   .2019432   .2947557     0.69   0.495    -.3829162    .7868025
     state_1 |   6.077295   3.300348     1.84   0.069    -.4713127     12.6259
     state_2 |   8.259368   3.061705     2.70   0.008     2.184281    14.33445
     state_3 |   6.853283   2.997107     2.29   0.024      .906374    12.80019

----------------------Deleted some fixed effect results to save space---------
    state_51 |   1.041116   2.871721     0.36   0.718    -4.657002    6.739234
      year_2 |   1.577016   .7433858     2.12   0.036     .1019775    3.052055
      year_3 |   1.681938   .6959821     2.42   0.017     .3009584    3.062917
------------------------------------------------------------------------------
```

3. $\widehat{mrdrte}_{it} = \hat{\beta}_0 + \hat{\beta}_1 unem_{it} + \underbrace{\delta_1 Yr2001 + \delta_2 Yr2002}_{\text{Dummy for all but one year}} + \underbrace{\alpha_i}_{\text{State "fixed effect"}} + u_{it}$

In STATA:

```
xtset state
xtreg mrdrte unem year_2 year_3, fe

Fixed-effects (within) regression               Number of obs      =       153
Group variable: id                              Number of groups   =        51

R-sq:  within  = 0.0676                         Obs per group: min =         3
       between = 0.1015                                        avg =       3.0
       overall = 0.0314                                        max =         3
                                                F(3,99)            =      2.39
corr(u_i, Xb)  = 0.0951                         Prob > F           =    0.0731
------------------------------------------------------------------------------
      mrdrte |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        unem |   .2019432   .2947557     0.69   0.495    -.3829162    .7868025
      year_2 |   1.577016   .7433858     2.12   0.036     .1019775    3.052055
      year_3 |   1.681938   .6959821     2.42   0.017     .3009584    3.062917
       _cons |   5.778023   1.911012     3.02   0.003     1.986161    9.569885
-------------+----------------------------------------------------------------
     sigma_u |  8.6877605
     sigma_e |  3.5144936
         rho |  .85936665   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(50, 99) =    17.33                 Prob > F = 0.0000
```

Notice how the estimated coefficient for unemployment did not change between the three regressions above! This is because we control for the same state and time effects in all regressions, just in different ways.