Panel Data Analysis

Prf. José Fajardo
FGV/EBAPE

Regression with Time Fixed Effects

An omitted variable might vary over time but not across states:

- Safer cars (air bags, etc.); changes in national laws
- These produce intercepts that change over time
- Let S_t denote the combined effect of variables which changes over time but not states ("safer cars").
- The resulting population regression model is:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + \beta_3 S_t + u_{it}$$

Time fixed effects only

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_3 S_t + u_{it}$$

This model can be recast as having an intercept that varies from one year to the next:

$$\begin{aligned} Y_{i,1982} &= \beta_0 + \beta_1 X_{i,1982} + \beta_3 S_{1982} + u_{i,1982} \\ &= (\beta_0 + \beta_3 S_{1982}) + \beta_1 X_{i,1982} + u_{i,1982} \\ &= \lambda_{1982} + \beta_1 X_{i,1982} + u_{i,1982}, \end{aligned}$$

where $\lambda_{1982} = \beta_0 + \beta_3 S_{1982}$ Similarly,

$$Y_{i,1983} = \lambda_{1983} + \beta_1 X_{i,1983} + u_{i,1983},$$

where $\lambda_{1983} = \beta_0 + \beta 3 S_{1983}$, etc.

Two formulations of regression with time fixed effects

1. "T-1 binary regressor" formulation:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \delta_2 B2_t + \ldots \delta_T BT_t + u_{it}$$

where $B2_t = \begin{cases} 1 \text{ when } t=2 \text{ (year \#2)} \\ 0 \text{ otherwise} \end{cases}$, etc.

2. "Time effects" formulation:

$$Y_{it} = \beta_1 X_{it} + \lambda_t + u_{it}$$

Time fixed effects: estimation methods

1. "T-1 binary regressor" OLS regression

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \delta_2 B2_{it} + \ldots \delta_T BT_{it} + u_{it}$$

   - Create binary variables $B2,\ldots,BT$
   - $B2$ = 1 if t = year #2, = 0 otherwise
   - Regress Y on X, $B2,\ldots,BT$ using OLS
   - Where's $B1$?

2. "Year-demeaned" OLS regression
   - Deviate Y_{it}, X_{it} from *year* (not state) averages
   - Estimate by OLS using "year-demeaned" data

Estimation with both entity and time fixed effects

$$Y_{it} = \beta_1 X_{it} + \alpha_i + \lambda_t + u_{it}$$

- When T = 2, computing the first difference and including an intercept is equivalent to (gives exactly the same regression as) including entity and time fixed effects.
- When T > 2, there are various equivalent ways to incorporate both entity and time fixed effects:
  - entity demeaning & T – 1 time indicators (this is done in the following STATA example)
  - time demeaning & n – 1 entity indicators
  - T – 1 time indicators & n – 1 entity indicators
  - entity & time demeaning

```
. gen y83=(year==1983);                    First generate all the time binary variables
. gen y84=(year==1984);
. gen y85=(year==1985);
. gen y86=(year==1986);
. gen y87=(year==1987);
. gen y88=(year==1988);
. global yeardum "y83 y84 y85 y86 y87 y88";
. xtreg vfrall beertax $yeardum, fe vce(cluster state);

Fixed-effects (within) regression               Number of obs      =       336
Group variable: state                           Number of groups   =        48
R-sq:  within  = 0.0803                         Obs per group: min =         7
       between = 0.1101                                        avg =       7.0
       overall = 0.0876                                        max =         7
corr(u_i, Xb)  = -0.6781                        Prob > F           =    0.0009
                                  (Std. Err. adjusted for 48 clusters in state)
------------------------------------------------------------------------------
             |               Robust
      vfrall |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     beertax |  -.6399799   .3570783    -1.79   0.080    -1.358329    .0783691
          y83 |  -.0799029   .0350861    -2.28   0.027    -.1504869   -.0093188
          y84 |  -.0724206   .0438809    -1.65   0.106    -.1606975    .0158564
          y85 |  -.1239763   .0460559    -2.69   0.010    -.2166288   -.0313238
          y86 |  -.0378645   .0570604    -0.66   0.510    -.1526552    .0769262
          y87 |  -.0509021   .0636084    -0.80   0.428    -.1788656    .0770615
          y88 |  -.0518038   .0644023    -0.80   0.425    -.1813645    .0777568
        _cons |    2.42847   .2016885    12.04   0.000     2.022725    2.834215
-------------+----------------------------------------------------------------
```

Are the time effects jointly statistically significant?

```
.   test $yeardum;

 ( 1)  y83 = 0
 ( 2)  y84 = 0
 ( 3)  y85 = 0
 ( 4)  y86 = 0
 ( 5)  y87 = 0
 ( 6)  y88 = 0

       F(  6,    47) =    4.22
            Prob > F =    0.0018
```

Yes

The Fixed Effects Regression Assumptions and Standard Errors for Fixed Effects Regression

Under a panel data version of the least squares assumptions, the OLS fixed effects estimator of β_1 is normally distributed. However, a new standard error formula needs to be introduced: the "clustered" standard error formula. This new formula is needed because observations for the same entity are not independent (it's the same entity!), even though observations across entities are independent if entities are drawn by simple random sampling.

Here we consider the case of entity fixed effects. Time fixed effects can simply be included as additional binary regressors.

LS Assumptions for Panel Data

Consider a single X:

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}, \quad i = 1,\ldots,n,\ t = 1,\ldots,T$$

1. $E(u_{it}|X_{i1},\ldots,X_{iT},\alpha_i) = 0$.
2. $(X_{i1},\ldots,X_{iT},u_{i1},\ldots,u_{iT})$, $i = 1,\ldots,n$, are i.i.d. draws from their joint distribution.
3. (X_{it}, u_{it}) have finite fourth moments.
4. There is no perfect multicollinearity (multiple X's)

Assumptions 3&4 are least squares assumptions 3&4
Assumptions 1&2 differ

Assumption #1: $E(u_{it}|X_{i1},\ldots,X_{iT},\alpha_i) = 0$

- u_{it} has mean zero, given the entity fixed effect *and* the entire history of the X's for that entity
- This is an extension of the previous multiple regression Assumption #1
- This means there are no omitted lagged effects (any lagged effects of X must enter explicitly)
- Also, there is not feedback from u to future X:
  - Whether a state has a particularly high fatality rate this year doesn't subsequently affect whether it increases the beer tax.
  - Sometimes this "no feedback" assumption is plausible, sometimes it isn't. We'll return to it when we take up time series data.

Assumption #2: $(X_{i1},\ldots,X_{iT},u_{i1},\ldots,u_{iT})$, $i = 1,\ldots,n$, are i.i.d. draws from their joint distribution.

- This is an extension of Assumption #2 for multiple regression with cross-section data
- This is satisfied if entities are randomly sampled from their population by simple random sampling.
- This does ***not*** require observations to be i.i.d. *over time* for the same entity – that would be unrealistic. Whether a state has a high beer tax this year is a good predictor of (correlated with) whether it will have a high beer tax next year. Similarly, the error term for an entity in one year is plausibly correlated with its value in the year, that is, corr(u_{it}, u_{it+1}) is often plausibly nonzero.

Autocorrelation (serial correlation)

Suppose a variable Z is observed at different dates t, so observations are on Z_t, $t = 1,\ldots, T$. (Think of there being only one entity.) Then Z_t is said to be ***autocorrelated*** or ***serially correlated*** if corr(Z_t, Z_{t+j}) ≠ 0 for some dates $j \neq 0$.

- "Autocorrelation" means correlation with itself.
- cov(Z_t, Z_{t+j}) is called the ***j^{th} autocovariance*** of Z_t.
- In the drunk driving example, u_{it} includes the omitted variable of annual weather conditions for state i. If snowy winters come in clusters (one follows another) then u_{it} will be autocorrelated (*why*?)
- In many panel data applications, u_{it} is plausibly autocorrelated.

Independence and autocorrelation in panel data in a picture:

	$i=1$	$i=2$	$i=3$	L	$i=n$
$t=1$	u_{11}	u_{21}	u_{31}	L	u_{n1}
M	M	M	M	L	M
$t=T$	u_{1T}	u_{2T}	u_{3T}	L	u_{nT}

← Sampling is i.i.d. across entities →

- If entities are sampled by simple random sampling, then ($u_{i1},\ldots, u_{iT}$) is independent of ($u_{j1},\ldots, u_{jT}$) for different entities $i \neq j$.
- But if the omitted factors comprising u_{it} are serially correlated, then u_{it} is serially correlated.

Under the LS assumptions for panel data:

- The OLS fixed effect estimator $\hat{\beta}_1$ is unbiased, consistent, and asymptotically normally distributed
- However, the usual OLS standard errors (both homoskedasticity-only and heteroskedasticity-robust) will in general be wrong because they assume that u_{it} is serially uncorrelated.
  - In practice, the OLS standard errors often understate the true sampling uncertainty: if u_{it} is correlated over time, you don't have as much information (as much random variation) as you would if u_{it} were uncorrelated.
  - This problem is solved by using "clustered" standard errors.

Clustered Standard Errors

- Clustered standard errors estimate the variance of $\hat{\beta}_1$ when the variables are i.i.d. across entities but are potentially autocorrelated within an entity.
- Clustered SEs are easiest to understand if we first consider the simpler problem of estimating the mean of *Y* using panel data…

Clustered SEs for the mean estimated using panel data

$$Y_{it} = \mu + u_{it},\ i = 1,\ldots, n,\ t = 1,\ldots, T$$

The estimator of μ mean is $\bar{Y} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} Y_{it}$.

It is useful to write $\bar{Y}$ as the average across entities of the mean value for each entity:

$$\bar{Y} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} Y_{it} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{T}\sum_{t=1}^{T} Y_{it}\right) = \frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i \quad ,$$

where $\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}$ is the sample mean for entity *i*.

Because observations are i.i.d. across entities, $(\bar{Y}_1,\ldots \bar{Y}_n)$ are i.i.d. Thus, if *n* is large, the CLT applies and

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i \xrightarrow{d} N(0, \sigma^2_{\bar{Y}_i}/n), \text{ where } \sigma^2_{\bar{Y}_i} = \text{var}(\bar{Y}_i).$$

- The *SE* of $\bar{Y}$ is the square root of an estimator of $\sigma^2_{\bar{Y}_i}/n$.
- The natural estimator of $\sigma^2_{\bar{Y}_i}$ is the sample variance of $\bar{Y}_1$, $s^2_{\bar{Y}_i}$. This delivers the clustered standard error formula for $\bar{Y}$ computed using panel data:

Clustered SE of $\bar{Y} = \sqrt{\frac{s^2_{\bar{Y}_i}}{n}}$, where $s^2_{\bar{Y}_i} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\bar{Y}_i - \bar{Y}\right)^2$