Panel Data Model

December, 2021

Panel Data

- Panel data is obtained by observing the same person, firm, county, etc. over several periods.
- Unlike the pooled cross sections, the observations for the same cross section unit (panel, entity, cluster) in general are dependent. Thus cluster-robust statistics that account for correlation within panel should be   used.

Panel Data

A double subscript distinguishes **entities** (states) and **time** periods (years)

i = entity (state), n = number of entities,
so $i = 1,\ldots,n$

t = time period (year), T = number of time periods
so $t =1,\ldots,T$

Data: Suppose we have 1 regressor. The data are:

$$(X_{it}, Y_{it}),\ i = 1,\ldots,n,\ t = 1,\ldots,T$$

Panel Data and Causality

- Panel data can be used to control for time invariant unobserved heterogeneity, and therefore is widely used for causality research.
- By contrast, cross sectional data cannot control for time invariant unobserved heterogeneity, so may suffer bigger omitted variable bias than panel data.
- The idea is simple. We take various forms of difference, and the time invariant unobserved heterogeneity is removed.
- Effectively, the panel data use the same panel as both treatment group and control group, and by invoking the before and after comparison, remove the time invariant omitted variables. The limitation of panel data is that time varying omitted variables are still present. But overall, the omitted variable bias gets smaller than cross sectional data.

Unobserved Effect Panel Data Model

- Consider a two-period unobserved effect model

$$y_{it} = \beta_0 + \delta_0 d_t + \beta_1 x_{it} + a_i + e_{it} \qquad (1)$$

- The subscript *i* indexes panels, while *t* indexes periods.

- a_i is time constant unobserved heterogeneity. e_{it} is the idiosyncratic error, or time-varying unobserved heterogeneity. $a_i + e_{it}$ is the composite error term.

- d_t is time dummy, so is panel constant and time varying; you can think of a_i as panel dummy, so is time constant and panel varying.

Endogeneity

- The main reason to use panel data is to correct for the endogeneity caused by unobserved time constant effect, i.e.,

$$\mathrm{cov}(x_{it}, a_i) \neq 0 \qquad (2)$$

- Given that nonzero covariance, the pooled OLS estimator applied to (1) is inconsistent.

First Difference (FD) Estimator I

- The repeated observations for the same panel make it possible to remove a_i via differencing
- First write down the regression for period 2 and period 1 explicitly as

$$
\begin{aligned}
y_{it=2} &= \beta_0 + \delta_0 * 1 + \beta_1 x_{it=2} + a_i + e_{it=2} \qquad (3) \\
y_{it=1} &= \beta_0 + \delta_0 * 0 + \beta_1 x_{it=1} + a_i + e_{it=1} \qquad (4)
\end{aligned}
$$

Now it is clear that a_i can be removed by subtracting the second equation from the first one.

First Difference (FD) Estimator II

- So we compute the first time difference for each panel

$$\Delta y_i = y_{i,t=2} - y_{i,t=1} \quad (5)$$

$$\Delta x_i = x_{i,t=2} - x_{i,t=1} \quad (6)$$

$$\Delta e_i = e_{i,t=2} - e_{i,t=1} \quad (7)$$

- Finally, run the regression using the first-differened data, called first difference equation:

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta e_i \quad (8)$$

  Notice that both a_i and β_0 disappear. In general, differencing removes all time constant variables (such as gender).

- OLS applied to the FD regression (8) yields the so called first-difference estimator. The FD estimator is consistent and has causal interpretation if the regressor in (8) is exogenous, i.e.,

$$E(\Delta x_i, \Delta e_i) = 0 \quad (9)$$

Serial Correlation

- In general the error term in the difference regression (8), Δ*e*$e_{i,}$ is negatively serially correlated when e_{it} is serially uncorrelated.
- For example, if data have three periods, then

$$E(\Delta e_{i,t}\Delta e_{i,t-1}) = E[(e_{i,t=3} - e_{i,t=2})(e_{i,t=2} - e_{i,t=1})] = -\sigma_e^2 < 0$$

- So cluster-robust statistics should be used.

Diminishing Variation

Typically, the variation in the differenced independent variable is much smaller than the variation in the original independent variable. Thus imprecise estimate can be expected from FD estimator. Like the IV estimator, here we face the same tradeoff of efficiency versus unbiasedness.

Gretl Command

The Gretl in order to obtain the FD estimates, first we need to make sure that the data is loaded as the panel data – identifying cross-sectional ID and the time ID.

Then we need to generate the difference of each variable

$$dvariable = variable_t - variable_{t-1}$$

We can do this by writing the command:
diff variablename

This command will difference each variable within the cross-sectional unit

FD Estimator can be used to control for time-constant unobserved heterogeneity. FD estimator cannot be used when the regressor of interest is time-constant. FD estimator is imprecise when the regressor changes little over time.

Fixed Effect (FE) Estimator I

For concreteness let $t = (1, 2, 3)$ in the following causal model

$$y_{it} = \beta_0 + \delta_1 d1_t + \delta_2 d2_t + \beta_1 x_{it} + a_i + e_{it} \tag{10}$$

Note that there are two time-dummies in (10) because there are three periods.

$$d1_t = \begin{cases} 1, & \text{period 1;} \\ 0, & \text{period 2 3.} \end{cases} \tag{11}$$

$$d2_t = \begin{cases} 1, & \text{period 2;} \\ 0, & \text{period 1 3.} \end{cases} \tag{12}$$

so period 3 is the base period.

Fixed Effect (FE) Estimator II

Averaging (10) across *i* leads to the so called between regression

$$\bar{y}_i = \beta_0 + \delta_1 \bar{d}1_t + \delta_2 \bar{d}2_t + \beta_1 \bar{x}_i + a_i + \bar{e}_i \qquad (13)$$

where the time averages are

$$\bar{y}_i = \frac{1}{3}\sum_{t=1}^{3} y_{it} \qquad (14)$$

$$\bar{x}_i = \frac{1}{3}\sum_{t=1}^{3} x_{it} \qquad (15)$$

$$\bar{e}_i = \frac{1}{3}\sum_{t=1}^{3} e_{it} \qquad (16)$$

The average of a_i is itself since it is time-invariant. Note that these averages are of variables across time by cross-sectional unit not by variable alone.

Between Regression

OLS estimator applied to the between regression is inconsistent since

$$\operatorname{cov}(\bar{x}_i a_i) = \frac{1}{n}\sum_t \operatorname{cov}(x_{it} a_i) \neq 0,$$

Fixed Effect (FE) Estimator III

Subtracting the between regression (13) from (10) leads to the so called within regression

$$y_{it}^{\texttt{demean}} = \delta_1 d1_t^{\texttt{demean}} + \delta_2 d2_t^{\texttt{demean}} + \beta_1 x_{it}^{\texttt{demean}} + e_{it}^{\texttt{demean}} \tag{18}$$

where

$$\begin{aligned} y_{it}^{\texttt{demean}} &= y_{it} - \bar{y}_i && (19) \\ x_{it}^{\texttt{demean}} &= x_{it} - \bar{x}_i && (20) \\ e_{it}^{\texttt{demean}} &= e_{it} - \bar{e}_i && (21) \end{aligned}$$

Note a_i is removed. Finally OLS applied to the within regression (18) is the FE estimator.

Gretl Command

The Gretl command:

Panel dep-var indep-var(s) --fixed effects

If you include a time constant independent variable such as gender, it will be dropped.

Fixed Effect

You can estimate the fixed effect $\hat{a}_i$ by replacing coefficient with its FE estimates in the between regression

$$\hat{a}_i = \bar{y}_i - \hat{\beta}_0 - \hat{\delta}_1 \bar{d}1_t - \hat{\delta}_2 \bar{d}2_t - \hat{\beta}_1 \bar{x}_i \tag{24}$$

STATA also reports the F test for the joint significance of fixed effects:

$$H_0 : \hat{a}_1 = \hat{a}_2 = \ldots = \hat{a}_{n-1} = 0 \tag{25}$$

Remarks

- FE estimator cannot be used when the regressor of interest x_{it} is time-invariant (such as gender).
- Cluster Robust Standard Error should be used in the within regression (18) since $e_{it}^{\texttt{demean}}$ is serially correlated.
- FE and FD estimators are inconsistent when $E(x_{it}e_{it}) \neq 0$ in the main model (10). In that case, IV variable is needed to completely resolve the endogeneity issue.

Three Questions

- Q: Why do we need time dummy?
- A: The time dummy $d1_t$ and $d2_t$ in (10) can control for time varying but panel constant unobserved effect. Example is national trend. It affects every panel and evolves over time.
- Q : Why do we need panel dummy?
- The panel dummy c_j in (22) can control for panel varying but time constant unobserved effect. Example is ability. It varies across persons but remains unchanged over time.
- Q: What if there are time-varying omitted variables?
- A: IV is still needed if there is time-varying omitted variable.

Summary

* Panel data model is useful when the omitted variable is time-invariant.
* Panel data model cannot be used when the key regressor is time-invariant.
* IV Estimator applied to the Within Regression should be considered when the omitted variable is time-varying.