Intermediate Econometrics

Professor: Xiaojun Song

Timekeeper: Rui Zhou

Spring 2023

目录

9 Panel Data **2**

9.1 DiD 2

9.2 First Difference 4

9.3 Fixed Effects Estimation 5

9.4 Random Effects Estimation 8

9.5 Correlated Random Effects Estimation 9

9.6 General Policy Analysis with Panel Data 10

9 Panel Data

In panel data, each individual is assumed i.i.d. Specifically speaking, indicators within individual is allowed with cross-correlation, but correlations across individuals are not allowed.

9.1 DiD

Most of the time, we are interested in an event that happened in the timeline (often supposed to be exogenous) and see its influence. Naturally, we classify our data into the control and treatment group. We can construct a dummy as a cutoff to represent the happening of this event, such as year, province, country, etc. Usually, we assume parallel trend behind the control and the treatment group, and reason that any difference between groups are caused by the event. The key for DiD is to understand the "trend", for "parallel trend" and "difference in trend". Thus, what we are interested in is not the trend itself, but the difference between trends. Graphically speaking, differences in slopes. Hence, it is crystal clear that an interaction term is needed.

In the example, we are interested in the effect of building a new garbage incinerator on housing prices (price before and after the garbage incinerator was built). Variable *rprice* means the real housing price, *nearinc* is a dummy meaning whether this house is near the incinerator. A naive rookie may estimate such a simple model using the data in the year when the incinerator was built like $rprice = \gamma_0 + \gamma_1 nearinc + u$, and claim that γ_1 is the effect. However, the fact is that, regardless of whether or not the incinerator was built, those houses near the incinerator was far from center of the city and had a relatively lower price. This is the underlying "parallel trend". The trend itself is not of interest; an interaction term representing differences in slopes is indeed of need. (In order to detect slope differences, you need to first control for parallel trends, *nearinc* and the time dummy are indispensable.) Under such logic, it is smooth to understand difference-in-differences (DiD) estimator and why it is called that way.

Define a dummy $y81$. $y81 = 1$ if the observation comes from year of 1981 when construction of the incinerator started; $y81 = 0$ otherwise. Use the following equation to estimate difference-in-difference estimator and its standard error:

$$rprice = \beta_0 + \delta_0 y81 + \beta_1 nearinc + \delta_1 (y81 \cdot nearinc) + u$$

where δ_1 is what we want.

Generally speaking, DiD estimator can be used to estimate the effect of natural experiment or quasi-natural experiment. In a (quasi-) natural experiment, observations can be divided into control group and treatment group. (Quasi-) natural experiment is different from true experiment. In a true experiment, subjects are randomly assigned to either control or treatment group. However in a natural experiment, control and treatment group all come from a certain change in "policies". So, we have to control for the systematic difference between the two groups. To achieve that, we need data from (at least) two years, one before the

policy change, one after the policy change. We estimate the equation

$$y = \beta_0 + \delta_0 d2 + \beta_1 dT + \delta_1 (d2 \cdot dT) + o.f.$$

where C stands for control group, T for treatment group; $dT = 1$ if from a treatment group. $d2 = 1$ indicates the second period after the policy change.

Careful interpretation of the equation will help a lot to our understanding of DiD. The term $\beta_1 dT$ controls for fixed effect of the two groups. $\delta_0 d2$ controls for the time trend. $\delta_1 (d2 \cdot dT)$ detects any differences in slopes after controlling for fixed effect and time trend.

	$Before$	$After$	$After - Before$
$Control$	β_0	$\beta_0 + \delta_0$	δ_0
$Treatment$	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \delta_0 + \delta_1$	$\delta_0 + \delta_1$
$Treatment - Control$	β_1	$\beta_1 + \delta_1$	δ_1

Therefore, δ_1 measures the effect of policy, which is also the average treatment effect (ATE as covered before). From the table above, we can see there are two ways to understand δ_1:

1. $\hat{\delta}_1 = (\bar{y}_{1,T} - \bar{y}_{1,C}) - (\bar{y}_{0,T} - \bar{y}_{0,C})$
2. $\hat{\delta}_1 = (\bar{y}_{1,T} - \bar{y}_{0,T}) - (\bar{y}_{1,C} - \bar{y}_{0,C})$

The pratical meaning of $\hat{\delta}_1$ can be interpreted as it that, compare the difference in outcomes of the units that are affected by the policy change (= treatment group) and those not affected (= control group), before and after the policy was enacted. $\hat{\delta}_1$ represents before/after comparisons in "(quasi-)natural experiments", so DiD can be used to evaluate policy changes or other exogenous events. The substraction hopefully *controls for* any changes in *external factors* that are common to both the treated and control groups, which will be the case when we have *random* assignment. In this case, the DiD estimator can be interpreted as the average treatment effect.

Note that DiD only works if the difference in outcomes between two groups is not changed by other factors than the policy change (e.g. there must be no differential trends). Parallel trends assumes that any trends in the outcome y would trend at the same rate towards the same direction in the absence of the intervention. If the parallel assumption is violated, we cannot be sure that the DiD estimator is identifying the effects of the policy or simply other unaccounted factor causing different trends between these groups. However in such cases, we can add flexibility by adding an additional control group. In the regression equation, more interaction terms are considered, in order to account for possibly different trends in different levels of some other factors (e.g., difference-in-difference-in-differences estimators).

9.2 First Difference

In the example, we are interested in the effect of x_{it} on y_{it}. Assume that no other explanatory variables are available, we wonder whether it would be possible to estimate the causal effect of x_{it} on y_{it}. Luckily, if the individuals are observed for at least two periods, and other factors affecting y_{it} stay approximately *constant* over those periods.

$$y_{it} = \beta_0 + \delta_0 d_t + \beta_1 x_{it} + v_{it}, \text{where } v_{it} = a_i + u_{it}$$

where

- d_t: time dummy for the second second
- a_i: unobserved time-constant factors (= fixed effect)
- u_{it}: other unobserved factors (= idiosyncratic error)

However as has been emphasized intensively before, this simple regression may encounter issues of omitting variables. One possible way out is to control more factors as possible, but some are unobservable, some hard to control. Another way is to categorize the composite error v_{it} into two parts: the time-invariant a_i and time-variant part u_{it}, as we did above. The most direct and simple way is to pool the data and use OLS to estimate coefficients. By doing this, we must assume that v_{it} is uncorrelated with x_{it}. Even if we assume that idiosyncratic (time-varying) error is uncorrelated with x_{it}, as long as a_i correlates with x_{it}, pooling OLS will give us biased and inconsistent estimators. The error caused by this is also called heterogeneity bias. In most applications, we collect panel data mainly to consider the arbitrary correlation of unobserved fixed effect a_i and explanatory variables. Since a_i is time-invariant, we can take the difference of two adjacent years. For two distinct t such that $d_t = 0, 1$ respectively, we can estimate that

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i$$

which is called the first-differenced equation. As long as the assumption that Δu_i is uncorrelated with Δx_i , the regression comes back to the ordinary case. The key assumption holds if for any period t, u_i is uncorrelated with explanatory variables in the two periods (strict exogeneity).

Note again that there may be arbitrary correlation between the unoserved time-invariant characteristics and the included explanatory variables. If a_i correlates with x_{it}, the error term v_{it} will correlate with x_{it}, then both δ_0, β_1 will be biased. OLS in the original equation would therefore fail. The first-differenced panel estimator is thus a way to consistently estimate causal effects in the presence of time-invariant endogeneity. Because a_i will be eliminated, **any form of the correlation** of α_i and x_i is allowed, since all that will be offset in first difference. In other words, α_i is free to be general. However, "general" is no equivalent to "random", and α_i is fixed.

However, in order to eliminate a_i, first differencing may cut down the variation in explanatory variables greatly. Though x_{it} may vary a lot in each period, it is still possible that Δx_i does not vary much. Little variation in explanatory variables will lead to a relatively large standard error, and less precision of estimation. Another

issue with first differencing is that, it cannot be used to estimate time-invariant variable's effect, because all that will be eliminated once you do the first differencing (such as education level, gender, etc.). As was mentioned, strict exogeneity is a critical assumption. If such assumption is violated, first-differenced estimator will lose consistency.

One last note. First differenced equation will yield the same result as DiD if x_i here is a dummy variable. In DiD, we introduce the term $\beta_1 dT$ to control for fixed effect of the two groups, $\delta_0 d2$ to control for the time trend. $\delta_1(d2 \cdot dT)$ detects any differences in slopes after controlling for fixed effect and time trend. In first differenced equation, first differencing comes first to wipe out the fixed effect; the time trend is attributed to the intercept term, and the $\beta_1 \Delta x_i$ term by definition is equivalent to that interaction term.

First differencing can be applied to multi-period data. Take a concrete example of a panel data in which each person has three observations across three periods.

$$y_{it} = \delta_1 + \delta_2 d2_t + \delta_3 d3_t + \beta_1 x_{it} + u_{it}$$

where two time dummies are introduced to allow for different intercepts across periods. Take the first difference and we will get

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

where $\Delta d2_t = 1, \Delta d3_t = 0$ when $t = 2$; $\Delta d2_t = -1, \Delta d3_t = 1$ when $t = 3$. Note that the equation above does not include an intercept, which is not convenient for computing R^2. Better to estimate an equivalent equation like

$$\Delta y_{it} = \alpha_0 + \alpha_3 d3_t + \beta_1 \Delta x_{it} + \Delta u_{it}$$

First differenced equation requires no serial correlation. It is easy to check this requirement. Let $r_{it} = \Delta u_{it}$, if r_{it} follows $r_{it} = \rho r_{i,t-1} + e_{it}$, the requirement demands $\rho = 0$. We can check it by testing against the null hypothesis $H_0 : \rho = 0$, and draw the conclusion based on t statistics.

9.3 Fixed Effects Estimation

Same as before, start with the simple equation of

$$y_{it} = \beta_1 x_{it} + \alpha_i + u_{it}$$

where α_i is the fixed effect for individual i, u_{it} is the idiosyncratic error.

9.3.1 Within Estimator

Use **time variation** with cross-sectional units to get **within estimators**.

From the regression equation, first sum it up and then get the average.

$$\frac{1}{T}\sum_{i=1}^{T} y_{it} = \beta_1 \cdot \frac{1}{T}\sum_{i=1}^{T} x_{it} + \alpha_i + \frac{1}{T}\sum_{i=1}^{T} u_{it}$$

which for simplicity can be written as

$$\bar{y}_i = \beta_1 \bar{x}_i + \alpha_i + \bar{u}_i$$

Then, we can get the new regression equation under fixed effect.

$$y_{it} - \bar{y}_i = \beta_1(x_{it} - \bar{x}_i) + (u_{it} - \bar{u}_i)$$

Or equivalently,

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{it} + \ddot{u}_{it}$$

where $\ddot{y}_{it} \equiv y_{it} - \bar{y}_i$ (same for $\ddot{x}_{it}, \ddot{u}_{it}$, and the fixed effect α_i is by nature *removed*. $\ddot{y}_{it}, \ddot{x}_{it}$ are called **time-demeaned** data. Such transformation for fixed effect model is also called within transformation. The demeaned equation is without intercept term. If you wonder why I did not include an intercept term like β_0 in the original equation, you can do it yourself the demeaning process and see if β_0 will be eliminated. One comfortable justification is that, "intercept" is perfectly absorbed by individual fixed effect α_i.

However, the basic logic under within estimator happens to be its drawback. If some regressors of interest are **invariant** to time, only individual-dependent, they will be canceled out. Gender and education background are typical examples.

When it comes to the estimation of individual fixed effect, since

$$\bar{u}_i = \frac{1}{N}\sum_{i=1}^{N} u_{it} \xrightarrow{p} 0$$

We have

$$\hat{\alpha}_i = \bar{y}_i - \beta_1 \bar{x}_i$$

If the original regression equation has k-many explanatory variables, after within transformation, the demeaned equation for fixed effect will have k-many variables (But note that the equation does not have an intercept!). However, in order to estimate the average over time for each individual, we will lose 1 df for each average we estimated. Therefore, the degree of freedom for residual is $df = NT - N - k = N(T-1) - k$.

9.3.2 Between Estimator

On the other hand, **between estimator** does *not* consider time variation for each individual.

Use $\bar{y}_i = \beta_1 \bar{x}_i + \alpha_i + \bar{u}_i$ for $i = 1, 2, ..., N$. and regress $\bar{y}_i$ on $\bar{x}_i$ to get the regressed equation. The corresponding coefficients are called between estimator. Since α_i is allowed to have unknown correlation with x_i, slope

coefficients may be biased and inconsistent. If we think a_i is uncorrelated with x_i, using random effect estimator might be more powerful. Between estimator ignores the important information across time.

Note that, for unbiasedness and consistency, strict exogeneity is required under $\bar{y}_i = \beta_1 \bar{x}_i + \alpha_i + \bar{u}_i$, i.e., u_{it} is not correlated with x_{it} at any time period, including the past, present and future.

9.3.3 Discussions

Within Estimator

- The R^2 of the demeaned equation is inappropriate, since α_i is not considered here.
- The effect of time-invariant variables cannot be estimated.
  - However, the effect of interactions for time-variant variables with time-invariant ones can be estimated.
- If a **full** set of time dummies are included, the effect of variables whose change over time is constant cannot be estimated. (e.g., working experience by year)
  - Multicollinearity
- Degrees of freedom have to be adjusted, since the N time averages are estimated in addition.
  - Resulting degrees of freedom to be $NT - N - K$.

Alternative Interpretation of Fixed Effects

Another way to interpret fixed effects is to introduce a dummy for each individual in the original regression.

$$y_{it} = a_1 ind1_{it} + a_2 ind2_{it} + ... + a_N indN_{it} + \beta_1 x_{it} + u_{it}$$

where $indK_{it}$ indicates if the observation stems from individual K. If so, $indK_{it} = 1$; and 0 otherwise. The equation does not fall into dummy variable trap since intercept term is not included. However, in model with fixed effect α_i, the story would be reversed. This is because with the full set of N-many indicators and the intercept, multicollinearity problem will arise. When N is large, this method will be impratical.

Fixed Effect v.s. First Differencing

- In the case $T = 2$, fixed effects and first differencing are identical.
  - Note that time dummy is naturally required in first differencing model, but that is not the case for fixed effect model. For equivalence, we have to include a time dummy to represent the second period.
  - First differencing is more straightforward. Heteroskedasticity-robust standard errors are easier to compute.
- For $T > 2$, fixed effects is more efficient if classical assumptions hold.
- First differencing may work better in the case of severe serial correlation in the errors.

  - $u_{it} = \rho u_{i,t-1} + v_{it}$. The errors follow a random walk if $\rho = 1$. Then, $\Delta y_{it} = \beta_1 \Delta x_{it} + v_{it}$ makes you relieved.
  - If T is large relative to N, the panel has a pronounced time series character.
- In practice, better to compute both and check robustness.

Unbalanced Panel Data

A panel data is unbalanced when not all cross-sectional units has the same number of observations. If you are to pick between FE and FD to deal with unbalanced panel data,

- FE considers average value, and some missing values are tolerated.
- FD requires that each observation have available data for both t and $t-1$.

FE generally preserves more data than FD with unbalanced panels.

9.4 Random Effects Estimation

Consider a random effect model as

$$y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_i + u_{it}$$

where α_i is assumed to be random, and $Cov(x_{it}, \alpha_i) = 0$, i.e., completely unrelated to explanatory variable(s).

Then, the composite error, $v_{it} = \alpha_i + u_{it}$, is uncorrelated with the explanatory variables. But v_{it} is serially correlated for observations coming from the same individual i.

$$\begin{aligned} Cov(v_{it}, v_{is}) &= Cov(\alpha_i + u_{it}, \alpha_i + u_{is}) \\ &= Cov(\alpha_i, \alpha_i) + Cov(\alpha_i, u_{is}) + Cov(\alpha_i, u_{it}) + Cov(u_{it}, u_{is}) \\ &= \sigma_\alpha^2 + 0 + 0 + 0 \\ &= \sigma_\alpha^2 \neq 0 \end{aligned}, \forall t \neq s$$

under the assumption that idiosyncratic errors are serially uncorrelated, i.e., $Cov(u_{it}, u_{is}) = 0$.

$$Corr(v_{it}, v_{is}) = \frac{Cov(v_{it}, v_{is})}{\sqrt{Var(v_{it}) Var(v_{is})}} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_u^2} > 0$$

If OLS is used, standard errors have to be adjusted for the fact that errors are correlated over time for given i, i.e., take the serial correlation structure into account. Because of the serial correlation, OLS is not efficient. We're to find a transformation to the model so that GM-assumptions would hold. As in fixed effect model, time-demeaned equation might be a candidate.

$$y_{it} - \bar{y}_i = \beta_1 (x_{it} - \bar{x}_i) + (v_{it} - \bar{v}_i)$$

However, $(v_{it} - \bar{v}_i)$ is still serially-correlated.

We plug in a λ to do the transformation

$$y_{it} - \lambda\bar{y}_i = \beta_1(x_{it} - \lambda\bar{x}_i) + (v_{it} - \lambda\bar{v}_i)$$

in the hope that $e_{it} = v_{it} - \lambda\bar{v}_i$ would be serially-uncorrelated, i.e.

$$Cov(e_{it}, e_{is}) \neq 0, \forall t \neq s$$

λ is then called the quasi-demeaning parameter, and the regression equation $y_{it}-\lambda\bar{y}_i = \beta_1(x_{it}-\lambda\bar{x}_i)+(v_{it}-\lambda\bar{v}_i)$ is called the quasi-demeaned equation. It can be proved that

$$\lambda = 1 - \sqrt{\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}}, \quad \lambda \in [0, 1]$$

λ is theoretically unknown but can be estimated by a given sample.

- If the random effect is relatively unimportant compared to the idiosyncratic error, $\lambda \to 0$, FGLS will be close to pooled OLS.
- If the random effect is relatively important compared to the idiosyncratic error, $\lambda \to 1$, FGLS will be close to fixed effects.

Luckily, random effects estimation can be used to estimate the effect of time-invariant variables. (Compared to fixed-effect model's limitations) But in economics, unobserved individual effects are seldomly uncorrelated with explanatory variables so that fixed effects is more convincing.

Remarks

Leave yourself with proof of $Cov(e_{it}, e_{is}) \neq 0, \forall t \neq s$, with

1. $E(e_{it}) = 0$;
2. $Var(e_{it}) = \sigma_u^2$;
3. $Cov(e_{it}, e_{is}) = 0, \forall t \neq s$.

Meanwhile, you can prove that β_1 is **BLUE**.

9.5 Correlated Random Effects Estimation

When using CRE to choose between FE and RE, we must include any time-constant variables α_i that appear in RE estimation.

$$y_{it} = \beta_1 x_{it} + \alpha_i + u_{it}$$

We try to combine FE and RE. This time, α_i is not only allowed to correlate with x_{it}, but also assumed to follow a distribution (or a certain way of correlation with x_{it}).

$$\alpha_i = \gamma_0 + \gamma_1\bar{x}_i + \gamma_i$$

where γ_i is assumed to be random and uncorrelated with each x_{it}. Then, $Cov(\gamma_i, \bar{x}_i) = 0$.

The CRE equation is then

$$y_{it} = \gamma_0 + \beta_1 x_{it} + \gamma_1 \bar{x}_i + v_{it}$$

where $v_{it} = u_{it} + \gamma_i, E(v_{it}) = 0, Cov(v_{it}, x_{it}) = 0$. Compared to FE or RE, the differences in equations lies in $\gamma_1 \bar{x}_i$,

Estimating the equation by RE (or even just pooled OLS) yields

$$\hat{\beta}_{CRE,j} = \hat{\beta}_{FE,j}, \forall j = 1, ..., k$$
$$\hat{\alpha}_{CRE,j} = \hat{\alpha}_{FE,j}, \forall t = 1, ..., T$$

which indicates time-varying estimates will be the same as in FE.

This intriguing discovery tells us that, adding time average $\bar{x}_i$ and using RE estimates is equivalent to first demeaning time average and using pooled OLS. This provides a new way to interpret FE: when estimating partial effect of x_{it} on y_{it}, FE controls for time average $\bar{x}_i$. Moreover, CRE "visualizes" the difference between FE and RE. CRE itself will yield the same result as FE. When setting $\gamma_1 = 0$, the CRE equation is then reduced to RE equation. Obviously, when $\gamma_1 \neq 0$, since $\bar{x}_i$ and x_{it} appear in the same equation, problems of multicollinearity will arise, which will cause $\hat{\beta}_{FE}$ to have a larger standard error and less precision; especially when variation of x_{it} is small across time. Hence, generally speaking, FE estimators are less precise than those in RE. Another advantage of CRE is that it allows for estimation of the effects of time-constant explanatory variables, which is not possible using FE.

At the meantime, CRE provides a simple but formal way to choose between FE and RE. RE will set $\gamma_1 = 0$, while FE will estimate γ_1. We can conduct a hypothesis test where the null hypothesis is in support of RE.

$$H_0 : \gamma_1 = 0$$

Under H_0, RE is sufficient. If H_0 is rejected, FE is preferred.

9.6 General Policy Analysis with Panel Data

The two-period, before-after setting is a special case of a more general policy analysis framework when $T \geq 2$.

$$y_{it} = a_1 + a_2 d2_t + ... + a_T dT_t + \beta w_{it} + x_{it}\varphi + \alpha_i + u_{it}$$

where w_{it} is the binary policy variable and β estimates the average treatment effect (ATE) of the policy.

To allow w_{it} to be systematically related to the unobserved fixed effect a_i, such as self-selection problem, we estimate the regression with either FD or FE, using cluster-robust standard errors.

We need to be careful if the policy variable w_{it} reacts to past shocks, which is the feedback from the error term to the policy variable. In this case, we can introduce an extra $\delta w_{i,t+1}$ into the model and test for feedback, i.e., testing against $H_0 : \delta = 0$.

Moreover, if time trends are unique across individuals ($T \geq 3$), we can safely introduce a new term $g_i t$ into the model, which is a unit-specific time trend. This allows the policy intervention to not only be correlated with level differences among units (captured by α_i), but also by trend differences. The model is still estimated by FE, and $g_i t$ will be differenced to be the intercept term.