Introductory Econometrics

Lecture 28: Pooling Cross Sections

Instructor: Ma, Jun

Renmin University of China

December 24, 2021

Pooled cross section

- An independently pooled cross section is obtained by sampling randomly from a large population at different points in time.
  - in each year, draw a random sample on hourly wages, education, experience etc. from the population of workers;
  - in every other year, draw a random sample on the selling price, square footage, number of bathrooms etc. of houses sold in a particular metropolitan area.
- Statistically, these data points are independently sampled observations.
- Sampling from the population at different points in time likely leads to observations that are not identically distributed. This can be dealt easily by allowing the intercept and slopes to change over time.
- We can simply include dummy variables for all but one year, where the earliest year is the base year.

Change in return to education

- A wage equation pooled across 1978 (base year) and 1985 is

$$\log(wage) = \beta_0 + \delta_0 y85 + \beta_1 educ + \delta_1 y85 \cdot educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 union + \beta_5 female + \delta_5 y85 \cdot female + u.$$

- The variable $y85$ is a dummy variable equal to one if the observation comes from 1985 and zero if it comes from 1978.
- The return to education in 1978 is β_1, and the return to education in 1985 is $\beta_1 + \delta_1$.
- δ_1 measures how the return to another year of education has changed over the seven-year period.
- Finally, in 1978, the $\log(wage)$ differential between women and men is β_5; the differential in 1985 is $\beta_5 + \delta_5$.
- We can test the null hypothesis that nothing has happened to the gender differential over this seven-year period by testing $H_0 : \delta_5 = 0$.

Policy evaluation

- Pooled cross sections can be very useful for evaluating the impact of a certain policy.
- Very often we get access to two cross-sectional data sets, collected before and after the occurrence of an event.
- The data can be used to determine the effect on economic outcomes.
- A control group is not affected by the policy change. A treatment group is thought to be affected by the policy change.
- To control for systematic differences between the control and treatment groups, we need two years of data, one before the policy change and one after the change.
- Thus, our sample is usefully broken down into four groups: the control group before the change, the control group after the change, the treatment group before the change, and the treatment group after the change.

Effect of a garbage incinerator's location on housing prices

- Kiel and McClain (1995) studied the effect that a new garbage incinerator had on housing values in North Andover, Massachusetts.
- The rumor that a new incinerator would be built in North Andover began after 1978, and construction began in 1981. We will use data on prices of houses that sold in 1978 and another sample on those that sold in 1981.
- The hypothesis is that the price of houses located near the incinerator would fall relative to the price of more distant houses. We define a house to be near the incinerator if it is within three miles.
- Let $rprice$ denote the house price. A naive analyst would use only the 1981 data and estimate a very simple model:

$$\begin{aligned} \widehat{rpice} &= 101307.5 - 30688.27 \cdot nearinc \\ &\quad (3093.0) \qquad\qquad (5827.71) \\ &\quad n = 142, R^2 = 0.165. \end{aligned}$$

- The coefficient on *nearinc* is the difference in the average selling price between homes near the incinerator and those that are not. The estimate shows that the average selling price for the former group was \$30688.27 less than for the latter group.
- This regression result only captures correlation. It does not imply that siting of the incinerator is causing the lower housing values. If we run the same regression for 1978 before the incinerator was not eve rumored, we get

$$\begin{aligned} \widehat{rpice} &= 82517.23 - 18824.37 \cdot nearinc \\ &\quad (2653.79) \qquad\qquad (4744.59) \\ &\quad n = 179,\ R^2 = 0.082. \end{aligned}$$

- Therefore, the average value of a home near the site was \$18824.37 less than the average value of a home not near the site. This is consistent with the view that the incinerator was built in an area with lower housing values.

The simple difference-in-differences (DID) model

- $g \in \{0,1\}$: control group ($g=0$), treatment group ($g=1$).
- $t \in \{0,1\}$: pre-intervention (policy) period ($t=0$), post-intervention (policy) period ($t=1$).
- For $i = 1,2,...,n_{gt}$, the outcome for the i-th individual in the (g,t) group is generated by

$$Y_i^{gt} = \gamma_t + \lambda_g + \delta \cdot d_{gt} + U_i^{gt},$$

  where n_{gt} is the sample size of the (g,t) group, γ_t is the time effect, λ_g is the group effect, d is the policy effect and U_i^{gt} is the effect from the unobserved factors.
- d_{gt} is the treatment status pattern: $d_{00} = d_{10} = d_{01} = 0$ and $d_{11} = 1$.
- We assume that

$$\left\{U_i^{gt} : i = 1,2,...,n_{gt}, g \in \{0,1\}, t \in \{0,1\}\right\}$$

  are i.i.d.

- We pool the data and use dummy variables to control for the time effect and group effect. Let $n = \sum_{g\in\{0,1\}}\sum_{t\in\{0,1\}} n_{gt}$. Then for $i = 1, 2, \ldots, n$,

$$\begin{aligned} Y_i &= \alpha + \gamma T_i + \lambda G_i + \delta D_i + U_i \\ D_i &= T_i \times G_i, \end{aligned}$$

  where D_i is the treatment status, $T_i \in \{0, 1\}$ indicates the period when the i-th individual is surveyed and $G_i \in \{0, 1\}$ indicates the group the i-th individual belongs to.
- We run OLS of Y_i against an intercept, T_i, G_i and D_i. If U_i is uncorrelated with T_i and G_i, OLS consistently estimates the policy effect δ.
- The OLS coefficient $\widehat{\delta}$ (the DID estimator) can be written as

$$\begin{aligned} \widehat{\delta} = & \left\{ \frac{\sum_{i:G_i=1,T_i=1} Y_i}{|i : G_i = 1, T_i = 1|} - \frac{\sum_{i:G_i=0,T_i=1} Y_i}{|i : G_i = 0, T_i = 1|} \right\} \\ & - \left\{ \frac{\sum_{i:G_i=1,T_i=0} Y_i}{|i : G_i = 1, T_i = 0|} - \frac{\sum_{i:G_i=0,T_i=0} Y_i}{|i : G_i = 0, T_i = 0|} \right\}, \end{aligned}$$

  where $|i : G_i = g, T_i = t|$ denotes number of observations in the (g, t) group.

- Simple rearrangement:

$$\begin{aligned}\widehat{\delta} &= \left\{\frac{\sum_{i:G_i=1,T_i=1} Y_i}{|i:G_i=1,T_i=1|} - \frac{\sum_{i:G_i=1,T_i=0} Y_i}{|i:G_i=1,T_i=0|}\right\} \\ &\quad - \left\{\frac{\sum_{i:G_i=0,T_i=1} Y_i}{|i:G_i=0,T_i=1|} - \frac{\sum_{i:G_i=0,T_i=0} Y_i}{|i:G_i=0,T_i=0|}\right\}.\end{aligned}$$

- The first term is the difference in means over time for the treated group. We compute the same trend in averages for the control group. By subtracting the second term from the first term, we hope to get a good estimator of the causal impact of the program or intervention.
- We can also augment the regression by incorporating covariates $(X_{i1}, ..., X_{ik})$ to avoid omitted variable bias:

$$Y_i = \alpha + \gamma T_i + \lambda G_i + \delta D_i + \sum_{j=1}^{k} \beta_j X_{ij} + U_i.$$

Effect of a garbage incinerator's location on housing prices

- The DID model formalizes the idea that we look at how the regression coefficient changes.
- For $T_i = 1$,

$$Y_i = \alpha + \gamma + (\lambda + \delta) G_i + U_i.$$

  We are not able to distinguish the effect of the policy (δ) from the effect of the location (λ), if we use only the post-intervention data.

- The DID estimator is the difference in the two OLS coefficients $-30688.27 - (-18824.37) = -11863.9$, which can be expressed as

$$\left(\overline{rprice}_{81,nr} - \overline{rprice}_{81,fr}\right) - \left(\overline{rprice}_{78,nr} - \overline{rprice}_{78,fr}\right),$$

  where nr stands for "near the incinerator site" and fr stands for "farther away from the site".

- We use the two-year pooled data and find the standard error by running the regression

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

- We also include various housing characteristics to avoid omitted variable bias. The kinds of homes selling near the incinerator in 1981 might have been systematically different than those selling near the incinerator in 1978.

TABLE 13.2 Effects of Incinerator Location on Housing Prices

Dependent Variable: *rprice*			
Independent Variable	**(1)**	**(2)**	**(3)**
constant	82,517.23 (2,726.91)	89,116.54 (2,406.05)	13,807.67 (11,166.59)
y81	18,790.29 (4,050.07)	21,321.04 (3,443.63)	13,928.48 (2,798.75)
nearinc	−18,824.37 (4,875.32)	9,397.94 (4,812.22)	3,780.34 (4,453.42)
y81·nearinc	−11,863.90 (7,456.65)	−21,920.27 (6,359.75)	−14,177.93 (4,987.27)
Other controls	No	*age*, age^2	Full Set
Observations	321	321	321
R-squared	.174	.414	.660

A general framework for policy analysis

- Another way to expand the basic DD methodology is to obtain multiple control and treatment groups as well as more than two time periods. We can create a very general framework for policy analysis by allowing a general pattern of interventions, where some units are never "treated" and others may be treated in different time periods. It is even possible that early in the study some units are subject to a policy but then later on the policy is dropped.
- In the general setting, we are interested in a policy intervention that applies at the group level. The model is

$$Y_i^{gt} = \gamma_t + \lambda_g + \delta \cdot d_{gt} + \sum_{j=1}^{k} \beta_j X_{ij}^{gt} + U_i^{gt}, i = 1, 2, ..., n_{gt}$$

$$g = 1, ..., G; t = 1, ..., T,$$

where d_{gt} is the treatment pattern: which is one if group g in year t is subject to the policy intervention, and zero otherwise. The group/time cell (g, t) has n_{gt} observations.

- When there are multiple periods, a full set of time-periods dummies is added to the regression. Similarly, when there are multiple groups, a full set of group dummies can be added.
- A policy dummy (which equals to one for observations in a specific group and period subject to the policy) replaces the interaction term.
- In practice, one includes an intercept and excludes one group and one time period. Then we estimate the model using pooled OLS, where the pooling is across all individuals across all (g,t) pairs.
- $\left(X_{i1}^{gt},...,X_{ik}^{gt}\right)$ can include measured variables that change only at the (g,t) level but also individual-specific covariates.
- The model can be applied to important problems such as studying the labor market impacts of minimum wages. Minimum wages can vary at the city level. The individual outcomes y_i^{gt} can be hourly wage. It could be very important to account for both time and city effects. In addition, we might have information on education, workforce experience, and background variables for individuals.

The DID model in the potential outcome framework

- We embed the DID model in the potential outcome framework. We do not impose the linear model assumption and study under what conditions a treatment effect parameter is identified.
- $(Y^t(0), Y^t(1))$: potential outcomes for an arbitrary individual in the population at time period t.
- $Y^t(1) - Y^t(0)$: individual treatment effect at t.
- $t = 0$: pre-intervention period; $t = 1$: post-intervention period; $d = 0$: no intervention; $d = 1$: intervention.
- $D^t \in \{0, 1\}$: a dummy variable, e.g., influenced ($D^t = 1$) or not influenced ($D^t = 0$) by the policy intervention.
- $G^t \in \{0, 1\}$: a dummy variable, e.g., treated group ($G^t = 1$) or non-treated group ($G^t = 0$).
- In the pre-intervention period, $D^0 = 0$. In the post-intervention period, some get influenced by the intervention. $D^0 = 0$ and $D^1 = G^1$.

- We observe: $Y^t = D^t Y^t(1) + (1 - D^t) Y^t(0)$. We have two populations: $\left(Y^0, G^0, D^0\right)$ and $\left(Y^1, G^1, D^1\right)$. Since $D^0 = 0$, $Y^0 = Y^0(0)$.
- Common trend assumption (CTA):

$$\begin{aligned} &\mathrm{E}\left[Y^1(0) \mid G^1 = 1\right] - \mathrm{E}\left[Y^0(0) \mid G^0 = 1\right] \\ &\qquad = \mathrm{E}\left[Y^1(0) \mid G^1 = 0\right] - \mathrm{E}\left[Y^0(0) \mid G^0 = 0\right]. \end{aligned}$$

- The CTA assumes that in the absence of the treatment, the average outcome for the treated group and the average outcome for the non-treated group would have experienced the same variation over time.
- $\mathrm{E}\left[Y^1(0) \mid G^1 = 1\right]$ is a counterfactual quantity: when $G^1 = 1$, $Y^1 = Y^1(1)$ and $Y^1(0)$ is unobserved.

Identification of the average treatment effect

- Under the CTA, the average treatment effect on the treated (ATT) at the post-intervention period is identified:

$$
\begin{aligned}
\text{ATT} &= \mathrm{E}\left[Y^1(1) - Y^1(0) \mid D^1 = 1\right] = \mathrm{E}\left[Y^1(1) - Y^1(0) \mid G^1 = 1\right] \\
&= \mathrm{E}\left[Y^1(1) \mid G^1 = 1\right] - \mathrm{E}\left[Y^0(0) \mid G^0 = 1\right] \\
&- \left(\mathrm{E}\left[Y^1(0) \mid G^1 = 0\right] - \mathrm{E}\left[Y^0(0) \mid G^0 = 0\right]\right) \\
&= \left(\mathrm{E}\left[Y^1(1) \mid G^1 = 1\right] - \mathrm{E}\left[Y^1(0) \mid G^1 = 0\right]\right) \\
&- \left(\mathrm{E}\left[Y^0(0) \mid G^0 = 1\right] - \mathrm{E}\left[Y^0(0) \mid G^0 = 0\right]\right) \\
&= \left(\mathrm{E}\left[Y^1 \mid G^1 = 1\right] - \mathrm{E}\left[Y^1 \mid G^1 = 0\right]\right) \\
&- \left(\mathrm{E}\left[Y^0 \mid G^0 = 1\right] - \mathrm{E}\left[Y^0 \mid G^0 = 0\right]\right),
\end{aligned}
$$

where we applied $G^1 = D^1$ and $Y^0 = Y^0(0)$.

- The quantity

$$\begin{aligned}
&\left(\mathrm{E}\left[Y^1 \mid G^1 = 1\right] - \mathrm{E}\left[Y^1 \mid G^1 = 0\right]\right) \\
&\qquad - \left(\mathrm{E}\left[Y^0 \mid G^0 = 1\right] - \mathrm{E}\left[Y^0 \mid G^0 = 0\right]\right) \\
&= \left(\frac{\mathrm{E}\left[Y^1 G^1\right]}{\mathrm{E}\left[G^1\right]} - \frac{\mathrm{E}\left[Y^1\left(1-G^1\right)\right]}{1-\mathrm{E}\left[G^1\right]}\right) - \left(\frac{\mathrm{E}\left[Y^0 G^0\right]}{\mathrm{E}\left[G^0\right]} - \frac{\mathrm{E}\left[Y^0\left(1-G^0\right)\right]}{1-\mathrm{E}\left[G^0\right]}\right)
\end{aligned}$$

  is a feature of the observed population. The equality follows from LIE.

- Suppose that we have repeated cross section data: $\left\{\left(Y_i^1, G_i^1\right) : i = 1, 2, \ldots, n_1\right\}$ and $\left\{\left(Y_i^0, G_i^0\right) : i = 1, 2, \ldots, n_0\right\}$. Then, the estimated ATT is a DID estimator:

$$\begin{aligned}
\widehat{\mathrm{ATT}} = &\left(\frac{\sum_{i=1}^{n_1} Y_i^1 G_i^1}{\sum_{i=1}^{n_1} G_i^1} - \frac{\sum_{i=1}^{n_1} Y_i^1\left(1-G_i^1\right)}{\sum_{i=1}^{n_1}\left(1-G_i^1\right)}\right) \\
&- \left(\frac{\sum_{i=1}^{n_0} Y_i^0 G_i^0}{\sum_{i=1}^{n_0} G_i^0} - \frac{\sum_{i=1}^{n_0} Y_i^0\left(1-G_i^0\right)}{\sum_{i=1}^{n_0}\left(1-G_i^0\right)}\right).
\end{aligned}$$

A linear model

- Suppose that

$$\begin{aligned} Y_i^1 &= \gamma^1 + \lambda G_i^1 + \delta D_i^1 + U_i^1, i = 1, 2, ..., n_1 \\ Y_i^0 &= \gamma^0 + \lambda G_i^0 + U_i^0, i = 1, 2, ..., n_0. \end{aligned}$$

  Then, $Y_i^1(1) = \gamma^1 + \lambda G_i^1 + \delta + U_i^1$ and $Y_i^1(0) = \gamma^1 + \lambda G_i^1 + U_i^1$. $\left(\gamma^1, \gamma^0\right)$ are the time effects.
- Assume that U_i^1 is uncorrelated with G_i^1 and U_i^0 is uncorrelated with G_i^0.
- Then, we can verify that the CTA is satisfied:

$$\begin{aligned} &\mathrm{E}\left[Y_i^1(0) \mid G_i^1 = 1\right] - \mathrm{E}\left[Y_i^0(0) \mid G_i^0 = 1\right] \\ &\qquad = \left(\gamma^1 + \mathrm{E}\left[U_i^1 \mid G_i^1 = 1\right]\right) - \left(\gamma^0 + \mathrm{E}\left[U_i^0 \mid G_i^0 = 1\right]\right) \end{aligned}$$

  and

$$\begin{aligned} &\mathrm{E}\left[Y_i^1(0) \mid G_i^1 = 0\right] - \mathrm{E}\left[Y_i^0(0) \mid G_i^0 = 0\right] \\ &\qquad = \left(\gamma^1 + \mathrm{E}\left[U_i^1 \mid G_i^1 = 0\right]\right) - \left(\gamma^0 + \mathrm{E}\left[U_i^0 \mid G_i^0 = 0\right]\right). \end{aligned}$$