Empirical Example: Difference-in-Difference Estimator

Card and Krueger (1992) investigate the important question of how minimum wage affecting employment. On April 1, 1992, New Jersey's minimum wage rose from \$4.25 to \$5.05 per hour. Meanwhile the minimum wage in Pennsylvania remained unchanged. So NJ is the treatment group, while PA is the control group.

The authors focus on fast-food restaurants where the minimum wage is most likely binding, and tip is rare so that this omitted variable becomes irrelevant. Two waves of survey were conducted before and after the change in minimum wage. Some restaurants failed to respond once or twice. The authors carefully discuss the possible self-selection bias.

The authors use panel data. That means the same restaurant was interviewed twice, before and after the minimum wage changed. So we need two subscripts to index observations, one for restaurant (panel), and the other for time. We let $y_{i,t}$ denote the full time equivalent employment at the i-th restaurant at time $t, t = 1$ before the change and $t = 2$ after the change.

A naive research may use NJ data (treatment group) only. Let $D1$ be a time dummy defined as

$$D1 = \begin{cases} 1, & \text{after minimum wage changes;} \\ 0, & \text{before minimum wage changes.} \end{cases} \tag{1}$$

We can run the following regression using OLS

$$y_{i,t} = \beta_0 + \beta_1 D1_t + error \tag{2}$$

If using NJ data only, then

$$\hat{\beta}_1 = \bar{y}^{\texttt{NJ}}_{\texttt{after}} - \bar{y}^{\texttt{NJ}}_{\texttt{before}} \tag{3}$$

The benefit of using dummy variable regression to obtain difference in mean is that heteroskedasticity can be easily accounted for. In this case $\bar{y}^{\texttt{NJ}}_{\texttt{after}} = 21.03$ (with standard error of 0.52), $\bar{y}^{\texttt{NJ}}_{\texttt{before}} = 20.44$, and $\hat{\beta}_1 = 21.03 - 20.44 = 0.59$ (a POSITIVE number!!) with standard error of 0.54. The t-value is $0.59/0.54 = 1.09$. So the conclusion is minimum wage has *no* effect on employment, or at least, rising minimum wage does not reduce employment, contradicting the micro theory.

There are many reasons why $\hat{\beta}_1$ may be biased. For instance, the error in (2) may contain unobserved factor (such as macro-economy). The difference-in-difference (DID) estimator <u>assumes</u> the same unobserved factor also affects PA, so that PA is comparable to NJ (i.e., the only difference is PA has no change in minimum wage). Then we can compare the change in NJ to change in PA, or in other words, look at the difference in difference. Mathematically,

$$\hat{\beta}_1^{\texttt{NJ}} \to \beta_1 + \texttt{omitted variable bias} \tag{4}$$

$$\hat{\beta}_1^{\texttt{PA}} \to 0 + \texttt{same omitted variable bias} \tag{5}$$

$$\hat{\beta}_1^{\texttt{NJ}} - \hat{\beta}_1^{\texttt{PA}} \to \beta \tag{6}$$

In words, the time-difference reported by $\hat{\beta}_1$ converges to the true causal effect plus omitted variable bias. Then difference-in-difference removes the omitted variable bias since the bias appears in both NJ and PA regressions.

The difference-in-difference estimator can be obtained in several equivalent ways. Here we discuss three of them. First we can run regression (2) for NJ and PA, respectively. Denote the coefficient of time dummy by $\hat{\beta}_1^{\texttt{NJ}}$ and $\hat{\beta}_1^{\texttt{PA}}$. Then the DID estimator is the difference between two coefficient estimates

$$\texttt{DID Estimator} = \hat{\beta}_1^{\texttt{NJ}} - \hat{\beta}_1^{\texttt{PA}} \tag{7}$$

$$= (\bar{y}^{\texttt{NJ}}_{\texttt{after}} - \bar{y}^{\texttt{NJ}}_{\texttt{before}}) - (\bar{y}^{\texttt{PA}}_{\texttt{after}} - \bar{y}^{\texttt{PA}}_{\texttt{before}}) \tag{8}$$

In this case, $\texttt{DID Estimator} = 0.59 - (-2.16) = 2.76$ with standard error of 1.36 (how to get it?) and t-value of 2.03. So DID estimator implies that rising minimum wage causes employment to go up.

The second way to obtain the DID estimator is to run the following regression using OLS

$$y_{i,t} = \alpha_0 + \alpha_1 D1_t + \alpha_2 D2_i + \alpha_3 (D1_t D2_i) + error \tag{9}$$

where $D1$ is the time dummy specified in (1), and $D2$ is the state dummy (or treatment group dummy) defined as

$$D2 = \begin{cases} 1, & \text{NJ (treatment group);} \\ 0, & \text{PA (control group).} \end{cases} \tag{10}$$

Note that $D1$ has subscript t, so is time-varying. $D2$ is time-invariant. The DID estimator is just $\hat{\alpha}_3$ in (9). In other words, the DID estimator is the coefficient of the interaction term of time and state dummies. If we have pooled cross sections instead of panel data, then we need to replace $y_{i,t}$ with y_i in (9).

The third approach works for panel data only. For each restaurant we can compute first difference as

$$\Delta y_i = y_{i,t=2} - y_{i,t=1} \tag{11}$$

The stata command is

```
sort storeid;
by storeid: gen dy = y[_n]-y[_n-1];
```

where storeid is the unique id for each store (panel). If there are n stores, then n missing values will be generated. Next run the regression of

$$\Delta y_i = \delta_0 + \delta_1 D2_i + error \tag{12}$$

The DID estimator is just $\hat{\delta}_1$ in (12). The equation (1a) on page 779 of Card and Krueger (1992) is effectively (12) augmented with additional regressor x. We can show that

$$\begin{aligned}
\texttt{DID Estimator} &= \hat{\delta}_1 && (13)\\
&= \overline{\Delta y_i}^{\texttt{NJ}} - \overline{\Delta y_i}^{\texttt{PA}} && (14)\\
&= (\bar{y}^{\texttt{NJ}}_{\texttt{after}} - \bar{y}^{\texttt{NJ}}_{\texttt{before}}) - (\bar{y}^{\texttt{PA}}_{\texttt{after}} - \bar{y}^{\texttt{PA}}_{\texttt{before}}) && (15)\\
&= \hat{\beta}^{\texttt{NJ}}_1 - \hat{\beta}^{\texttt{PA}}_1 && (16)
\end{aligned}$$

So we should get the same estimate as (7).

There are multiple ways to define the treatment and control groups. For instance, within NJ, the restaurants that offer low wage are subject to change in minimum wage, so are in the treatment group. The NJ restaurants that offer high wage are in control group since they are unlikely to be affected by change in minimum wage. Then we need to redefine $D2$ as

$$D2 = \begin{cases} 1, & \text{low wage restaurant (treatment group);} \\ 0, & \text{high wage restaurant (control group).} \end{cases} \tag{17}$$

By using the NJ data only, the DID estimator is (see Table 3, columns under stores in New Jersey)

$$\begin{aligned}
\texttt{DID Estimator} &= \hat{\beta}^{\texttt{low wage}}_1 - \hat{\beta}^{\texttt{high wage}}_1 && (18)\\
&= (\bar{y}^{\texttt{low wage}}_{\texttt{after}} - \bar{y}^{\texttt{low wage}}_{\texttt{before}}) - (\bar{y}^{\texttt{high wage}}_{\texttt{after}} - \bar{y}^{\texttt{high wage}}_{\texttt{before}}) && (19)\\
&= 1.32 - (-2.04) = 3.36 && (20)
\end{aligned}$$

very close to 2.76, the estimate that uses PA stores as control group.

The authors also discuss how to test the validity of a specific control group. The basic idea is to show the control group with questionable validity is comparable to the control group with confirmed validity. The PA is a questionable control group, while the high-wage store in NJ is a valid control group for sure. Since the DID estimator using the high-wage restaurant in NJ as control group, 3.03, is close to the DID estimator using the PA restaurants as control group, 2.76, we conclude that using PA as control group is valid.

Questions:

- Why do the authors compare NJ to PA, not to, say, Alabama?
- Smart guy A says that the rising employment in NJ is part of a national upward trend, so has nothing to do with change in minimum wage. Is he right?
- Smart guy B says that the increase in minimum wage in NJ may be legislated for endogenous reason (e.g., as a measure to fight recession). Is he right?
- Some restaurants offer high wage (so minimum wage is not binding) while others offer low wage (so minimum wage is binding). Why do the authors also investigate the high wage restaurant?
- Smart guys C says that the reported result overestimates the true effect since the restaurants that were out of business are excluded, and those "bad" restaurant tended to hire falling number of employees. Is he right?