# Identification, IV, Difference of Differences, Differnce of Difference of Differences

Hisahiro Naito

University of Tsukba

## Identification Issues

- Consider the equilibrium of a particular market. Assume that there are many market for this goods cross sectionally across countries.
- You are interested in estimating the demand function
- Estimating the demand function is equivalent to recovering information on indifference curve of the utility
- From the price effect and income effect and using the Slutsky equation, you can get the price elasticity of the Hicksian demand
- The price elasticity of the Hicksian demand recovers the information on the shape of the indifference curve
- Similarly, estimating the supply curve is equivalent to recovering information on the production function.
- In the case of the labor supply, the case is opposite.
- The estimating the labor supply curve will recover the information on the utility function. Estimating the labor demand recover the information on the production function.

## Identification

- To estimate the demand curve, you need the exogenous variation of the supply curve
- To estimate the supply curve, you need the exogenous variation of the demand curve
- In other words, to estimate the consumption function, or utility function or consumer behavior, you need the exogenous variation of the supply side
- To estimate the production function, you need the variation of the preference or the utility or consumers' behavior
- Again, the case of labor supply is opposite.
- To estimate the labor supply, you need to have the exogenous variation of the production function.

## IV

- In general, if you have one endogenous variable, you need at least one exogenous variable that is not included in the target equation.
- This is called the Rank condition.
- Consider the following equation. All variables are calculated from the deviation of the mean. Thus, $\sum x_i = 0, \sum y_{1i} = 0, \sum y_{2i} = 0$
- You are interested in estimating the following equation:
  $y_{2i} = \beta_1 y_{1i} + \varepsilon_i$
- Assume that $E[\varepsilon | y_{1i}] \neq 0$.Thus, $y_{1i}$ is endogenous variable.
- In this case, you need an instrument variable that will affect $y_{1i}$ but does not enter the main equation.

- Let's assume that there is another variable $z_i$ and assume that $z_i$ affect $y_{1i}$ and it is not correlated with $\varepsilon_i$ and does not enter the main equation.
- Then we can write $y_{1i} = \gamma z_{1i} + u_i$.
- TSL can be done as follows. First estimate $y_{1i} = \gamma z_{1i} + u_i$ by OLS and get the estimate of $\gamma$, $\widehat{\gamma}$. Then, get the predicted value of $y_{1i}$, $\widehat{y}_{1i}$.
- Then regress $y_{2i}$ on $\widehat{y}_{1i}$. Then, you will get the consistent estimate of $\beta_1$ and $\beta_2$.

- $z_{1i}$ is the variable that is not included in the main equation. It is called the excluded variable and the equation $y_{1i} = \gamma z_{1i} + u_i$ is called the first stage equation.
- Note that

$$\widehat{\gamma} = \frac{\sum z_i y_{1i}}{\sum z_i^2} \ \widehat{y}_{1i} = z_i \frac{\sum z_i y_{1i}}{\sum z_i^2}.$$

$$\begin{aligned} TSL &= \frac{\sum y_2 \widehat{y}_{1i}}{\sum \widehat{y}_{1i}^2} \\ &= \frac{\frac{\sum z_i y_{1i}}{\sum z_i^2} \sum z_i y_{2i}}{\sum z_i^2 \{\frac{\sum z_i y_{1i}}{\sum z_i^2}\}^2} \\ &= \frac{\sum z_i y_{2i}}{\sum z_i y_{1i}} \equiv IV \end{aligned}$$

# IV (3)

- Rank condition implies that you have some variable $z_{1i}$ that is not included in the main equation.
- If the Rank condition is not satisfied, then you have multicollinear problem. You cannot identify the equation at all.
- The Rank condition is only necessary condition.
- The order condition say that this variable must affect the endogenous variable. That is $\gamma \neq 0$ in a statistical significant way.
- If the Order condition is not satisfied, your standard error becomes too big, your estimate becomes so unstable. In such a case, the cure make the patient worse. OLS is much better actually although your paper is not published by using just OLS.
- The degree of the Rank condition is quite strong contrary to your perception in the IV or TSL.
- If there is one endogenous variable and one exogenous instrumental variable, the required F statistics is 10. In one endogenous variable and one instrumental variable, F statistics is equal to square of t statistics. Thus, you need to have an instrumental variable whose t statistics in the first stage regression is greater than 3.2!

## IV(4)

- To understand the intuition of IV, it is useful to consider the Wald estimator.
- To analyze the Wald Estimator, consider the simplest case where $y_{2i} = \beta_1 y_{1i} + \varepsilon_i, \mathrm{E}[\varepsilon|\mathrm{y}_{1i}] = \neq 0, y_{1i} = \gamma z_{1i} + u_i$
- Now suppose that $z_i$ take only two values, $z_i = 1$or 0.z=1 means that experiment status. Let $N_e$ be the number of the sample where $z_i = 1$ (experiment status)
- Let $N_c$ be the number of the sample where $z_i = 0$ (control status)
- First notice that IV estimator is $\beta_{Iv} = \frac{\sum y_{2i} z_i}{\sum y_{1i} z_i} = \frac{cov^s(y_{2i}, z_i)}{cov^s(y_{1i}, z_i)}$

- Note that $y_{2i}$ and $y_{1i}$ is the deviation from the mean. So, $Y_{2i} = y_{2i} + \overline{Y_2}$ and $Y_{1i} = y_{1i} + \overline{Y_2}$.
- Now calculate $\{\sum_{z_i=1} Y_{2i}\}/N_e - \{\sum_{z=0} Y_{2i}\}/N_c$
- This number show how much $Y_2$ change on average when z change from 0 to 1.
- Also calculate $\{\sum_{z_i=1} Y_{1i}\}/N_e - \{\sum_{z=0} Y_{1i}\}/N_c$. This number sows how much $y_1$ change on average when z change from 0 to 1.
- Notice that by the assumption, the movement of $z_i$ is completely random.

- Since we are interested in the effect of $y_1$ on $y_2$, it is natural to consider the following Wald estimator

$$Wald = \frac{\{\sum_{z_i=1} Y_{2i}\}/N_e - \{\sum_{z=0} Y_{2i}\}/N_c}{\{\sum_{z_i=1} Y_{1i}\}/N_e - \{\sum_{z=0} Y_{1i}\}/N_c}$$

- The numerator calculate on average how much $Y_{2i}$ changes when $z$ exogenously changes from 0 to 1.
- The denominator calculates on average how much $Y_{1i}$ change when z exogenously change from 0to 1.
- Thus, the above division calculate the effect of $Y_{1i}$ on $Y_{2i}$, which is $\widehat{\beta}$.

- Now calculate

$$
\begin{aligned}
cov^s(y_{2i}, z_i) &= \frac{1}{N-1}\sum(Y_{2i} - \overline{Y_2})(z_i - \overline{z}) \\
&= \frac{1}{N-1}\sum Y_{2i}z_i - N\overline{Y2}\,\overline{z} \\
&= \frac{1}{N-1}\{N_e\overline{Y_2}^e - \frac{N_e}{N}\{N_e\overline{Y_2}^e + N_c\overline{Y_2}^c\} \\
&= \frac{N_e}{N-1}\{\overline{Y_2}^e - \frac{N_e}{N}\overline{Y_2}^e - \frac{N_c}{N}\overline{Y_2}^c\} \\
&= \frac{N_eN_c}{N-1}\{\overline{Y_2}^e - \overline{Y_2}^c\}
\end{aligned}
$$

- Similarly for $cov^s(y_{1i}, z_i)$, we have

$$
cov^s(y_{2i}, z_i) = \frac{N_eN_c}{N-1}\{\overline{Y_1}^e - \overline{Y_1}^c\}
$$

Thus the Wald estimator is IV.

- Thus, what IV is doing can be summarized as follows. Graphically it can be explained as follows:
- First, for numerator, calculate how much the dependent variable changes when the instrumental variable changes. For denominator, calculate how much the endogenous variable change when the instrumental variable changes. Then, calculate the ratio between the numerator and the denominator. This is what IV is doing.
- Thus, it is clear that we need IV that moves the endogenous variable significantly when the instrumental variable changes. Otherwise, we cannot calculate the IV.

- When you have one endogenous variable, but you two or more than two instrumental variable, you can run the over-identification test.
- The idea of the over-identification test is as follows.
- Suppose that you run the IV regression with one instrument and get $\beta^1$
- Then you run the another IV regression with another instrument and get $\beta^2$.
- If both instrument variable are valid instruments, then $\beta^1$ should be very close to $\beta^2$. If they are not so close, it indicates that one of those two instruments are bad instrument.
- Over-identification test is very useful to convince reader that what you are doing is the right thing and the instrumental variables are valid instrument.

- When you run the IV regression, it is important to plot the relationship between endogenous variable and the instrumental variable.
- Sometime, a high correlation at the first stage regression between instrumental variables and endogenous variables are generated the outlier variable. In such a case, the validity of the instrument is questionable
- You instrumental variable take more than two values, you can check at which point, the strong correlations between the instrumental variables and the endogenous variables are generated.
- For example, suppose that you find the years of schooling is a good instrumental variable. But in reality yeas of schooling take from 6 to 18-20. Then, you can examine graphically at which point the strong relationship between the instrumental variable and the endogenous variable is generated.
- Also, you can make many dummy instrumental variable in this case and check the over-identification test.

Difference of Differences

- In the IV, I emphasize that you need to find the exogenous variation of z that move affect the endogenous variable substantially.
- This consideration naturally leads to you the Difference of Deferences estimator.
- Let $y_{tgi}$ be the outcome of the time $t$, group g, of individual $i$.Define $x_{tgi}$ in this way.
- Let t=A, B. and g be tr or cr. A means after. B mean Before. tr mean the treatment group and cr mean a control group.
- First consider $\frac{\overline{Y}_{A,tr}-\overline{Y}_{B,tr}}{\overline{x}_{A,tr}-\overline{x}_{B,tr}}Y$
- This is the effect of the treatment for the treated group.

- But it can include the effect of business cycle, time trend and other macro-economic shocks
- Thus, we can consider

$$\frac{\overline{Y}_{A,tr} - \overline{Y}_{B,tr} - \{\overline{Y}_{A,cr} - \overline{Y}_{B,cr}\}}{\overline{x}_{A,tr} - \overline{x}_{B,tr} - \{\overline{x}_{A,cr} - \overline{x}_{B,cr}}$$

- This is called the difference of differences estimator.
- You can estimate the difference of difference by running the following regress in stata

$$y_{tgi} = \alpha + \beta_1 D_t + \beta_2 D_g + \beta_3 x_{tgi}$$

- where $D_t$ is the time dummy and $D_g$ is the class dummy. In this case, $\beta_3$ will become the difference of differences estimator.
- $\beta_1$ control the macroeconomic shock that are common to both treatment group and control group. $\beta_2$ is the time-invariant group fixed effect.
- The key identifying assumption of the D-D estimator is that the time trend effect are same for two groups. Of this assumption is called parallel assumption.

- For conducting the difference of differences, you can calculate

$$\frac{\overline{Y}_{A,tr} - \overline{Y}_{B,tr} - \{\overline{Y}_{A,cr} - \overline{Y}_{B,cr}\}}{\overline{x}_{A,tr} - \overline{x}_{B,tr} - \{\overline{x}_{A,cr} - \overline{x}_{B,cr}}$$

  or you can run the regression

$$y_{tgi} = \alpha + \beta_1 D_t + \beta_2 D_g + \beta_3 x_{tgi}$$

- However, the running the regression has several advantages.
- First, you can calculate the standard error correctly and directly test the null hypothesis.
- Second, you can put more demographic control variables in the regression equation.

$$y_{tgi} = \alpha + \beta_1 D_t + \beta_2 D_g + \beta_3 x_{tgi} + \beta_4 W_{tgi}$$

- This will reduce the variance of the error term and increase the statistic precision which means that you have a higher t-statistics.

## Grouping Estimator

- Another useful estimator is the grouping estimator which was used by Blundell etc.
- Let $D_t$ be the dummy variable for the cohort group
- Let $D_g$ be the dummy variable for the educational group (high school graduate, college graduate)
- Assume that different cohort$\times$education group experience the different wage increases due to technological progress. i.e. wage inequality between high school graduate and college graduate. (Labor demand shock)
- Then, at the first stage run the following regression
  - $$w_{it}(1-\tau_{it}) = \gamma_1 D_t + \gamma_2 D_g + \gamma_3 D_t \times D_g + \varepsilon_{it}$$
- Get the predicted value of $w_{it}(1-\tau_{it})$ as$\widehat{w_{it}(1-\tau_{it})}$.
- Then, run the following regression

$$l_{it} = \beta_1 D_t + \beta_2 D_g + \beta_3 \widehat{w_{it}(1-\tau_{it})}$$

- This is similar to Eissa's difference of differences. But they are different in many dimension.