The role of the conrolling variable in the regression

Hisahiro Naito

University of Tsukuba

The role of x2 in the regression

- In the regression, the role of x_1, which is the variable of main interest, and x_2 are not equal.
- This implies that the assumption needed for x_1 and x_2 are different.
- Also, the interpretation of the coefficient of x_1 and x_2 are different
- If the proper assumption is satisfied, we can give a causal interpretation on the coefficient of x_1.
- But not for the coefficient of x_2
- The same argument holds for z_1 and x_2 in the IV regression where z_1 is the IV variable and x_2 is control variable.

The model

- Consider the data generation process $y_i = \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i$
- Our main interest is β_1: the effect of x_1 on y
- Suppose that in the real data x_1 is generated by complete randomness such as random trial experiment or rolling a dice or something else.
- In this case, in order to know the effect of x_1 on y, we just need to run the regression y on x_1.
- We do not need to include x_2 since x_1 is completely random. Because of x_1 is completkt random, x_1 is not correlated with ε_i,neither x_2. Thus we can treat $\beta_2 x_2 + \varepsilon_i$ as the error term and this error term is not correlated with x_1

Basic

- Conditional mean zero expectation.
- If $E(\varepsilon_i|x_{1i}, x_{2i}) = 0$ for all x_{1i} and x_{2i}. This is called conditional mean zero assumption. In this case. OLS estimate of $\beta_1$and β_2 are not biased. We can give a causal interpretation.
- The above conditional mean zero expectation says that conditional mean of ε_i is always zero no matter what x_1 and x_2 are. Thus, we can think that x_1 and x_2 are completly random and they are uncorrected with ε_i. Thus OLS does a good job.

Basic(2)

- What happen if $E(\varepsilon_i|x_{1i}, x_{2i}) = 0$ is not satisfied. However, sometime it is possible to find a situation of conditional independence of x_1.
- Conditional independence of x_1 is $E(\varepsilon_i|x_{1i}, x_{2i}) = E(\varepsilon_i|x_{2i})$.
- Conditional independence of x_{1i} says if it is conditioned on x_{2i}, the the expectation of ε_i is the same no matter what x_{1i} is.

- In other words, if it is conditioned on x_{2i}, we can think that x_{1i} is a random variable. As a result, it is not correlated with ε.
- In many cases, we can think that the variation of x_1 is random if we give enough conditional variables x_2.
- For example, suppose that we are interested in the effect of college education on earning. If we use, IQ, parent's income, grade in high school, regional economic background as x_{2i}, then we can reasonably think that the variation of x_1 after controlling IQ, parent's income, grade in high school, regional economic back ground is almost random like a rolling a dice given the information on x_2. In this case, the OLS coefficient of x_1 has a causal interpretation.
- Note that we cannot give a causal interpretation on the coefficient of x_2.

Proof

- Assume that $E(\varepsilon_i|x_{1i}, x_{2i}) = 0$ is not satisfied but conditional independence of x_1 ,$E(\varepsilon_i|x_{1i}, x_{2i}) = E(\varepsilon_i|x_{2i})$ is satisfied.
- This implies that given x_{2i}, x_{1i} is random and not correlated with ε_i. But x_{2i} itself is correlated with ε_i.
- Thus, assume that $E(\varepsilon_i|x_{2i}) = \gamma_0 + \gamma_1 x_{2i}$ and $\gamma_1 \neq 0$.
- Now define $v_i = \varepsilon_i - E(\varepsilon_i|x_{1i}, x_{2i})$. If we take the conditional expectation of v_i on x_{1i} and x_{2i}, then
- $E[v_i|x_{1i}, x_{2i}] = E[\varepsilon_i|x_{1i,}x_{2i}] - E(\varepsilon_i|x_{1i}, x_{2i}).=0$

Proof(2)

- $y_i = \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i$
- $y_i = \beta_1 x_1 + \beta_2 x_2 + v_i + E(\varepsilon_i | x_{1i}, x_{2i})$
- $y_i = \beta_1 x_1 + \beta_2 x_2 + v_i + E(\varepsilon_i | x_{2i})$
- $y_i = \beta_1 x_1 + \beta_2 x_2 + v_i + \gamma_0 + \gamma_1 x_{2_i}$
- $y_i = \gamma_0 + \beta_1 x_1 + (\beta_2 + \gamma_1) x_2 + v_i$
- Note that by the way that v_i is constructed, $E(v_i | x_{1i}, x_{2i}) = 0$. Thus, conditional mean zero assumption on x_{1i} and x_{2i} holds. Thus, we can run the OLS and the OLS coefficient of x_{1i} will give β_1 and the OLS coefficient of x_{2i} will give $\beta_2 + \gamma_1$.

In the case of IV

- In the case of IV, the same argument holds. In this case, what we need is
- $E(\varepsilon_i|z_i, x_{2i}) = E(\varepsilon_i|x_2)$.
- This allow the correlation of x_2 with the error term. But since our interest is the effect of x_1 on y, it is OK.
- The above conditional independence assumption implies that the conditional expecation is the same not matter what z is given x_2.
- In other words, z is random given x_2.
- This is a weaker assumption than $E(\varepsilon|z) = 0$. $E(\varepsilon|z) = 0$ requirs that conditional expection is zero.no matter what x_2 while $E(\varepsilon_i|z_i, x_{2i}) = E(\varepsilon_i|x_2)$ require that given x_2, the z does not affect the conditional expectation.