14.32 Econometric Data Science
Professor Anna Mikusheva
September, 2019

Lectures 12-13

Internal and external validity of a study

We define the validity of an econometric study by three characteristics: (1) estimates of the parameter of interest is unbiased and consistent; (2) tested hypothesis has desired significance level; (3) confidence sets have correct coverage.

We distingusish external and internal validity. External validity asks if the results of a study can be applied to other settings and/or other populations. For example, imagine that we estimated the average return to education using a sample of US workers from 2002. External validity asks whether the results will be similar for US workers in 2019? Or for workers in Germany? Or for current graduates? Or if a new minimal wage law will be passed. An assessment of external validity relies on the institutional setting and population. Also, the same study conducted on several different populations may bring us an understanding of how externally valid the results are.

Internal validity answers the question of whether a study is valid for a population, from which we have sample. If Assumptions 1-4 are satisfied, then our analysis is valid. The main discussion in this lecture is about what can go wrong, the so-called threats to internal validity.

Threat 1: Omitted variable bias

We have discussed this problem previously. An omitted variable bias problem arises when we can argue that there is a variable which is (1) a determinant of Y and is (2) correlated with X. In light of our updated assumption $E[e|X,W] = E[e|W]$, we need correlation with X even on the sub-population for which W is fixed (that is the essence of the correlation between X and Z is not captured by W). Notice that the correlation between Z and W is NOT a valid argument for omitted variable bias for the coefficient on X.

Possible solutions of the problem:

(1) Include Z in the regression as a control, if Z is available;

(2) Panel data (will be discussed later in the course);

(3) Instrumental Variable regression (will be discussed later);

(4) Run a controlled experiment (will be discussed later).

Threat 2: Functional form

The most common problem is that the effect we are trying to estimate is not constant, but rather depends either on the variable itself or level of other variables; that is, there is a need for non-linear modelling (polynomials, interactions, etc). Or controls enter the regression in a non-linear way. This threat was discussed in detail in lectures 10-12.

Threat 3: Error-in-variable bias

This problem arises when data we have contains mistakes or mis-reporting. We usually distinguish between systematic and measurement mistakes. The latter is the case when instead of variable Y we observe $Y^* =$

$Y + \epsilon$, where ϵ is mean zero and is independent from Y and any other variable in the data. A systematic mistake typically will lead to biases and is hard to correct, and for the correction we need to fully model how the mistake occured. Below we will be talking about measurement mistakes only.

First imagine that we have mistake in the dependent variable. That is, the ideal regression that satisfies assumptions 1-4 is

$$Y = \alpha + \beta X + \gamma W + e,$$

but instead of observing Y we observe $Y^* = Y + \epsilon$. Aparently, we still will end up with valid results if we run the regression:

$$Y^* = \alpha + \beta X + \gamma W + e^*,$$

where $e^* = e + \epsilon$ as long as ϵ is a measurement error with finite 4th moment. Indeed, independence of ϵ from all other variables leads to $E[e^*|X, W] = E[e|X, W] = E[e|W] = E[e^*|W]$. Thus, a measurement mistake in the dependent variable does not cause problems.

A measurement mistake to the regressor does cause bias. Assume that the ideal regression is

$$Y = \alpha + \beta X + e, \quad E[e|X] = 0,$$

and we observe $X^* = X + \epsilon$. The claim is that the OLS estimate from regressing Y on X^* has bias. Indeed,

$$Y = \alpha + \beta X + e = \alpha + \beta(X^* - \epsilon) + e = \alpha + \beta X^* + e^*,$$

where $e^* = e - \beta\epsilon$. Notice that the new error does not satisfy the exogeneity restriction:

$$cov(e^*, X^*) = cov(e - \beta\epsilon, X + \epsilon) = -\beta Var(\epsilon).$$

In the derivation of consistency we showed that

$$\widehat{\beta} \to^p \beta + \frac{cov(X^*, e^*)}{Var(X^*)} = \beta\left(1 - \frac{Var(\epsilon)}{Var(X) + Var(\epsilon)}\right)$$

Thus we see that the OLS is asymptotically biased (converges to a wrong value), but we also know the direction of the bias (toward zero), thus we estimate a smaller effect than it truly exists. This is called an *attenuation bias*.

Possible solutions: (1) IV regression (will study later) or (2) find the size of the measurement mistake and correct the bias.

Threat 4: sample selection

Sample selection is a violation of an i.i.d. assumption, when the sampling process influences the availability of data (not all observations available). We distinguish three cases:

- Data is missing at random
- Data is missing based on values of X's (exogenous selection).
- Data is missing based on the value of Y or the error term e (endogenous selection).

We will argue that the first two sample selections do not lead to biases while the last one ends in 'sample selection bias'.

For the first type of missing data imagine that you have collected at random data on 200 individuals, but then your computer crashed and the data on randomly selected 100 of them is missing. This would be exactly equivalent to initially collecting the data on randomly chosen 100 individuals. Thus, the computer crash does not introduce bias, it just reduces the accuracy of your estimation.

An example of the second type of selection process: selection with a higher probability of one of the studied group (overselecting poor schools in class size study). An example of the third type is an evaluation of the performance of mutual funds based on a sample of survivors.

Claim is: exogenous sample selection does not lead to bias, while endogenous sample selection results in bias.

Assume that the ideal regression is

$$Y_i = \alpha + \beta X_i + e_i,$$

which satisfies all assumptions, that is, $E[e|X] = 0$. Let the selection process be described by variable s_i which is a dummy equal to 1 for observations selected for the sample and 0 for not selected. Then the regression we have is

$$s_i Y_i = \alpha s_i + \beta(s_i X_i) + (s_i e_i),$$

where observations with $s_i = 0$ has no information. We would run regression only on non-zero draws. Denote those variables as Y_i^*, X_i^*, e_i^*. For validity, we need $E[e^*|X^*] = 0$.

Consider the second type of selection. Assume that exogenous sample selection happens, in particular $s = \mathbb{I}\{\gamma x + \delta Z \geq 0\}$, where Z is a variable (which maybe unavailable to us) that is independent from e. Then exogeneity the assumption implies $E(e|Z, X) = E(e|X) = 0$. Thus by the law of conditional expectations:

$$E(se|X, Z) = E(\mathbb{I}\{\gamma x + \delta Z \geq 0\}e|X, Z) = \mathbb{I}\{\gamma x + \delta Z \geq 0\}E(e|X, Z) = 0$$

and

$$E(e^*|X^*) = E(se|sX) = E\left(E[se|X, Z]|sX\right) = 0.$$

Here we have used the law of iterated expectations. Thus exogeneity is preserved for the selected regression.

Sometimes researchers intentionally create the selection of the second type in order to improve the accuracy of the estimation, and it is called stratification.

Sample selection bias arises when a selection process:

(i) influences the availability of data and

(ii) is related to the dependent variable.

An example of sample selection bias is estimating the average return on managed hedge funds. Unfortunately, most available data sets have data only on the mutual fund that survived for a few years, but do not have information on the failed one. You may imagine that the model for all funds is

$$returns_i = \mu + e_i.$$

But you have in your data set only observations for those funds that were successful enough: $s_i = \mathbb{I}\{e_i > C\}$. It is obvious that the average of observed returns will overestimate μ.

Threat 5: Simulteneous causality

This situation arises when not only X causes Y, but also Y causes X. For example, in the class-size-effect case it may be not just the lower student-teacher ratio that leads to better test scores, but there is a reverse relation as well: suppose districts with low test scores are given extra resources and as a result of a political process they also have a low student-teacher ratio. That is, assume that there exists two causal effects:

- a causal effect of X on Y: $Y_i = \alpha + \beta X_i + e_i$;
- a causal effect of Y on X: $X_i = \gamma + \delta Y_i + v_i$.

Assume for a moment that $\delta > 0$. This would mean that 'large values of Y would imply large values of X' and thus there is a non-trivial correlation $cov(X_i, e_i) \neq 0$. This would lead to the OLS bias.

Solutions to this problem:

- Run a randomized control trial.
- Instrumental Variable regression (will be covered in this course).
- Develop a full causal model (hard to do).