

Model Validity

Erick F Molina

Summer, 2024

ITAM

This presentation is based on the references listed in the course syllabus and on notes by Arturo Aguilar, Mauricio Romero, Laura Schechter, Shaun McRae, Kirill Borusyak, among others. Any remaining errors are my own.

External Validity

What is external validity?

- It refers to a study's ability to generalize its results to other populations, contexts, or historical periods.
- It allows the conclusions of a study to be useful across different settings.

For example:

- Mexico was a pioneer in implementing social policies aimed at reducing poverty (Prospera, Oportunidades, Progresa).
- Due to their design, results from these programs can potentially be generalized to other developing countries with similar characteristics.
- However, cultural, economic, and political differences must be taken into account, as they may affect external validity.

Common problems and how to strengthen external validity

Common problems:

- Differences across populations.
- Differences in context.

How to address external validity concerns:

- In an **experimental design**, researchers can include diverse and representative samples.
- By conducting a **sensitivity analysis**, researchers can assess how results change under different assumptions.
- **Replication:** Conducting studies in different contexts and populations.

Tips on external validity

- External validity is essential for the applicability and relevance of econometric studies.
- Identifying and mitigating common external validity issues strengthens the robustness of results.
- Proper experimental designs, sensitivity analyses, and replication are key to improving external validity.

Internal Validity

Internal Validity

- Internal validity refers to whether the results obtained from a given sample are valid and generalizable to the population from which the sample was drawn.
- Achieving internal validity requires careful attention to potential problems in the analysis.
- In econometrics, some of the most common issues are:
 - Model misspecification
 - Omitted variable bias
 - Measurement error

Omitted Variable Bias (OVB)

Omitted variable bias (OVB) occurs when a regression model excludes a relevant variable that is correlated with one or more included regressors. This can lead to biased estimates of the regression coefficients.

- Biased results can lead to incorrect conclusions and policy decisions. For example, when evaluating a public policy, biased estimates may result in poor decision-making.
- Identifying and correcting OVB is essential for obtaining accurate and reliable estimates.

Omitted Variable Bias (OVB)

Suppose we have the following regression with many variables:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{K-1} X_{K-1} + \beta_K X_K + U \quad (1)$$

Now consider another regression omitting X_K :

$$Y = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_{K-1} X_{K-1} + V \quad (2)$$

What is the relationship between β_j and α_j for $j = \{0, 1, \dots, K-1\}$?

Define the residual regression of X_K as:

$$X_K = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_{K-1} X_{K-1} + W \quad (3)$$

Omitted Variable Bias (OVB)

Substituting yields:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{K-1} X_{K-1} + \beta_K (\gamma_0 + \gamma_1 X_1 + \cdots + \gamma_{K-1} X_{K-1} + W) + U$$

Rearranging terms:

$$Y = \tilde{\beta}_0 + \tilde{\beta}_1 X_1 + \cdots + \tilde{\beta}_{K-1} X_{K-1} + \beta_K W + U \quad (4)$$

where $\tilde{\beta}_j = \beta_j + \beta_K \gamma_j$ for $j = \{0, 1, \dots, K-1\}$.

Defining the error term as $V = \beta_K W + U$, equations (2) and (4) are identical. Therefore, the first-order conditions for both problems are the same, implying $\tilde{\beta}_j = \alpha_j$. Hence, the bias is given by:

$$\alpha_j - \beta_j = \beta_K \gamma_j$$

The signs of β_K and γ_j determine the direction of the bias.

Omitted Variable Bias (OVB)

Example:

Consider a regression of wages on years of education:

$$\log(\text{wage}) = \alpha + \beta(\text{educ}) + \varepsilon$$

- If ability, an **omitted variable**, is positively correlated with education and has a positive effect on wages, the estimated coefficient on education (β) will be biased upward.

Omitted Variable Bias (OVB)

To reduce omitted variable bias, we can:

- Include as many relevant variables as possible in the regression model. *Note: caution is required.*
- Use econometric methods such as instrumental variables to address omitted variable bias.
- Implement control techniques such as propensity score matching.

Measurement error occurs when a variable available in a dataset contains errors. These errors may arise for several reasons, most commonly due to:

- Data entry errors. For example, an interviewer may omit a zero when recording a value.
- Precision errors. For instance, respondents may not recall a value accurately, or the survey questions may be poorly designed.

Suppose we want to observe X , but instead we observe $\tilde{X}$, which is X measured with error. That is, we observe $\tilde{X}_i = X_i + w_i$, where w_i denotes measurement error.

For example, suppose we want to estimate the coefficient β_1 in the following regression:

$$Y_i = \beta_0 + \beta_1 X_i + U_i \quad (5)$$

Since our dataset only contains $\tilde{X}_i$, we actually estimate:

$$Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 \tilde{X}_i + V_i \quad (6)$$

In regression (6), the estimated coefficient is:

$$\tilde{\beta}_1 = \frac{\text{Cov}(\tilde{X}, Y)}{\text{Var}(\tilde{X})}$$

We are interested in the relationship between $\tilde{\beta}_1$ and β_1 , where:

$$\beta_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

To establish this relationship:

$$\tilde{\beta}_1 = \frac{\text{Cov}(X + w, Y)}{\text{Var}(X + w)} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \frac{\text{Var}(X)}{\text{Var}(X + w)} + \frac{\text{Cov}(w, Y)}{\text{Var}(X + w)} \quad (7)$$

Assuming the measurement error is uncorrelated with the regression error, $\text{Cov}(w, U) = 0$:

$$\tilde{\beta}_1 = \beta_1 \left(\frac{\sigma_X^2}{\text{Var}(X + w)} \right) + \frac{\text{Cov}(w, \beta_0 + \beta_1 X + U)}{\text{Var}(X + w)} \quad (8)$$

Measurement Error

A common assumption is that measurement error (w) is uncorrelated with the true value of X . This is known as **classical measurement error**. A typical example is random data entry error.

In this case, $Cov(w, X) = 0$, and therefore:

$$\tilde{\beta}_1 = \beta_1 \left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2} \right) \quad (9)$$

This implies **attenuation bias**: the estimated coefficient using mismeasured variables is smaller in absolute value than the true coefficient.

Some final takeaways:

- The larger the measurement error, the more severe the attenuation bias.
- Some measurement errors are systematic rather than random, making them harder to address.
- In such cases, the direction of bias is also influenced by $\text{Cov}(w, X)$.