Lecture notes VIII

Assessing Studies Based on Multiple Regression

- SW Ch. 9

Definitions: Internal validity means that the (statistical) inferences are valid for the population actually studied.

External validity means that the inferences can be generalized beyond the population actually studied, i.e. to other populations and studies

Examples: - Time series analysis and prediction
- Country specific studies

What is actually the underlying population in non-experimental data?

Focus on <u>internal</u> validity:

Internal validity specifically has to do with the following issues:

1. Is the estimator unbiased and (more importantly) consistent?
   - can we interpret the (internal-) parameter as representing a partial effect (i.e. causal effect of a partial change in the variable of interest)

2. Are the standard errors correct?
   - robustness standard errors

3. Do confidence intervals have the correct coverage and do hypothesis tests have the correct level of significance?

5 main reasons for violation of internal validity:

1. Omitted variable bias
2. Misspecification of functional form
3. Errors in variables
4. Sample selection
5. Simultaneous causality

1. Omitted variable bias:

- one variable determining Y and which is correlated with an included variable is excluded

example: "public sector" dummy in estimates of returns to schooling.

Remedy: Test whether additional variables change estimates of interest parameter. Are they significant determinants of Y? $\bar{R}^2$-adj may be helpful.

Try out different specifications and disclose the results!

2. Misspecification of functional form

- Example 1: The marginal returns to schooling is not constant $\neq 0$ log-log model is misspecified.

- Example 2: Relation between R&D subsidies obtained under Slaatte Funn and number of employees in the firm not a linear-log model!

3. Errors-in-variables

The regressor is measured with error:

$$X^{obs} = X + \varepsilon \quad , E(\varepsilon | X) = 0$$

where X true value of X and ε measurement error.

Regression:

$$Y = \beta_0 + \beta_1 X + u \quad , \quad E(u|X) = 0$$

$$= \beta_0 + \beta_1 (X^{obs} - \varepsilon) + u$$

$$= \beta_0 + \beta_1 X^{obs} + \tilde{u} \quad , \quad \tilde{u} = u - \beta_1 \varepsilon$$

Then

$$\text{bias } \hat{\beta}_1 = \frac{Cov(X^{obs}, \tilde{u})}{Var(X^{obs})} = \frac{Cov(X+\varepsilon, u - \beta_1 \varepsilon)}{Var(X+\varepsilon)}$$

$$= \frac{-\beta_1 Var(\varepsilon)}{Var(X) + Var(\varepsilon)} = -\beta_1 \frac{\sigma_\varepsilon^2}{\sigma_x^2 + \sigma_\varepsilon^2}$$

- bias does not vanish when $n \to \infty$

4. Sample selection

Whether or not a potential observation unit is included in the sample depends on the <u>error term</u> in the regression equation.

- but no problem if this depends on <u>observed</u> variables

[redacted]

<u>Example</u>: Truncation: Unit i is observed only if $Y_i \geq a$, where a is some threshold.

I.e. units with low values of Y_i are never observed

Assume:

$Y_i = \beta_0 + \beta_1 X_i + u_i$, $E(u_i | X_i) = 0$

If unit i is observed, then $Y_i \gtrless a$

$\Leftrightarrow \beta_0 + \beta_1 X_i + u_i \gtrless a$ (depends on u_i)

$\Rightarrow E(u_i | X_i, \text{"}i \text{ is observed"}) =$

$E(u_i | X_i, Y_i \gtrless a) = E(u_i | X_i, \beta_0 + \beta_1 X_i + u_i \gtrless a)$

$= E(u_i | u_i \gtrless a - \beta_0 - \beta_1 X_i) > 0$

5. Simultaneous causality

- X causes Y, but also Y causes X

Example: Supply and Demand

(1) $X^S = \beta_0 + \beta_1 P + \varepsilon$, $X^S \equiv$ supply

P is price, ε error term

(2) $X^D = \lambda_0 + \lambda_1 P + u$, $X^D \equiv$ demand, u is error

$Cov(\varepsilon, u) = 0$

In equilibrium: $X^S = X^D$

$\Leftrightarrow \quad \beta_0 + \beta_1 P + \varepsilon = \lambda_0 + \lambda_1 P + u$

$\Leftrightarrow \quad (\beta_1 - \lambda_1) P = \lambda_0 - \beta_0 + u - \varepsilon$

$\Leftrightarrow$

$$P = \frac{\lambda_0 - \beta_0}{\beta_1 - \lambda_1} + \frac{u - \varepsilon}{\beta_1 - \lambda_1}$$

If we estimate (1) by OLS, we get biased estimates, since

$$Cov(\varepsilon, P) = Cov\left(\varepsilon, \frac{u - \varepsilon}{\beta_1 - \lambda_1}\right) = -\frac{Var(\varepsilon)}{\beta_1 - \lambda_1}$$

even if $Cov(\varepsilon, u) = 0$

Similarly, if we estimate (2) by OLS, the regressor and error is also correlated:

$$Cov(u, P) = Cov\left(u, \frac{u - \varepsilon}{\beta - \lambda_1}\right) = \frac{Var(u)}{\beta_1 - \lambda_1}$$