The instrumental variables model

Econometrics for minor Finance, Lecture 7

Tunga Kantarcı, Fall 2025

Linear regression model: Model assumption: Error is endogenous

We showed that

$$\mathsf{E}\left[\hat{\beta}_1 \,\middle|\, x\right] = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,\mathsf{E}\left[u_i \mid x\right]}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

If on average x is informative of u, that is, if

$$\mathsf{E}\left[u_i \mid x_i\right] \neq 0$$

the OLS estimator

$$\hat{\beta}$$

is biased and inconsistent since the second term does not disappear.

Linear regression model: Model assumption: Error is endogenous: What to do?

We need a new estimator that has desirable sampling distribution properties at least when the sample size is large. For example, a consistent but biased estimator is already better than the biased and inconsistent OLS estimator.

Linear regression model: Model assumption: Error is endogenous: What to do?

There are in fact different estimators that are consistent. The instrumental variables, IV, and LIML estimators estimate a single equation, and hence are called single-equation methods. The 3SLS, GMM, and FIML estimators jointly estimate an entire system of equations, and hence are called system of equations methods.

In this course we study the IV estimator.

The IV model

Consider the linear model

$$y_i = \beta x_i + u_i$$

where

$$x_i$$

is endogenous so that

$$\mathrm{E}\left[u_i \mid x_i\right] \neq 0$$

The IV Model: Assumptions: Linearity

Linearity. The model is linear in the parameters.

The IV Model: Assumptions: Errors can be heteroskedastic for the endogenous regressor

Homoskedasticity. Because x_i may be endogenous, we do not make any assumption for the conditional variance for x_i:

$$\text{Var}\left[u_i \mid x_i\right]$$

The IV Model: Assumptions

Suppose

$$z_i$$

is an instrumental variable.

$$z_i$$

satisfies two main assumptions.

The IV Model: Assumptions: Relevance

Relevance. In the population each instrument is correlated with the endogenous variable it is meant to explain:

$$\text{Cov}\left[z_i, x_i\right] \neq 0$$

The idea is that the instrument has genuine predictive information for the endogenous variable. Otherwise, it cannot help us isolate exogenous variation.

The IV Model: Assumptions: Relevance in expectation form

Because the model includes a constant, the intercept absorbs the average levels of the variables. This allows us to rewrite them in mean-zero form:

$$x_i = x_i - \mathrm{E}[x_i]$$

and

$$z_i = z_i - \mathrm{E}[z_i]$$

Centering the variables does not change any covariances or the IV estimator. After centering:

$$\mathrm{E}[x_i] = 0$$

and

$$\mathrm{E}[z_i] = 0$$

The IV Model: Assumptions: Relevance in expectation form

The covariance between two variables is defined as:

$$\text{Cov}\left[z_i, x_i\right] = \text{E}\left[z_i x_i\right] - \text{E}\left[z_i\right]\text{E}\left[x_i\right]$$

With

$$\text{E}[z_i] = 0$$

and

$$\text{E}[x_i] = 0$$

we have

$$\text{Cov}\left[z_i, x_i\right] = \text{E}\left[z_i x_i\right]$$

The IV Model: Assumptions: Relevance in expectation form

Then, the covariance condition

$$\text{Cov}\left[z_i, x_i\right] \neq 0$$

is equivalent to the moment condition

$$\text{E}\left[z_i x_i\right] \neq 0$$

This is the same relevance requirement: the instrument must contain predictive information for the endogenous regressor.

The IV Model: Assumptions: Exogeneity

Exogeneity. In the population, the error is uncorrelated with the instrument:

$$\text{Cov}\left[z_i, u_i\right] = 0$$

The idea is that the instrument contains no information about the unobserved determinants of the outcome. Otherwise, it would reintroduce the very endogeneity problem IV is designed to solve.

The IV Model: Assumptions: Exogeneity in expectation form

The covariance between two variables is defined as:

$$\mathrm{Cov}\left[z_i, u_i\right] = \mathrm{E}\left[z_i u_i\right] - \mathrm{E}\left[z_i\right]\mathrm{E}\left[u_i\right]$$

With

$$\mathrm{E}[z_i] = 0$$

and

$$\mathrm{E}[u_i] = 0$$

we have

$$\mathrm{Cov}\left[z_i, u_i\right] = \mathrm{E}\left[z_i u_i\right]$$

The IV Model: Assumptions: Exogeneity in expectation form

Then, the covariance condition

$$\text{Cov}\left[z_i, u_i\right] = 0$$

is equivalent to the moment condition

$$\text{E}\left[z_i u_i\right] = 0$$

This is the same exogeneity requirement: the instrument must contain no information about the unobserved determinants of the outcome.

The IV Model: Assumptions: Errors are homoskedastic

Homoskedasticity. That is,

$$\text{Var}\left[u_i \mid z_i\right] = \sigma^2$$

for all i.

The IV Model: Assumptions: Random sampling

Random sampling. The data we collect

$$\{(x_i, z_i, u_i) : i = 1, 2, \ldots, n\}$$

are independent and identically distributed draws from the population.