

## The instrumental variables model

Econometrics for minor Finance, Lecture 7

Tunga Kantacı, Fall 2025

Linear regression model: Model assumption: Error is endogenous

We showed that

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

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

$$E[u_i | x_i] \neq 0$$

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

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + u_i$$

where

$$\mathbf{x}_i$$

is  $K \times 1$  vector of regressors.  $K$  can include only one regressor though. Suppose they are all endogenous so that

$$\text{E}[u_i | \mathbf{x}_i] \neq 0$$

## The IV Model: Assumptions: Linearity

**Linearity.** The model is linear in the parameters.

# The IV Model: Assumptions

Suppose

$$\mathbf{z}_i$$

is a  $L \times 1$  vector of instrumental variables.

$$\mathbf{z}_i$$

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:

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

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:

$$x_i = x_i - E[x_i]$$

and

$$z_i = z_i - E[z_i]$$

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

$$E[x_i] = 0$$

and

$$E[z_i] = 0$$

## The IV Model: Assumptions: Relevance in expectation form

The covariance between two variables is defined as:

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

With

$$E[z_i] = 0$$

and

$$E[x_i] = 0$$

we have

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

## The IV Model: Assumptions: Relevance in expectation form

Then, the covariance condition

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

is equivalent to the moment condition

$$E[z_i x_i] \neq 0$$

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 each instrument:

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

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:

$$\text{Cov}[z_i, u_i] = E[z_i u_i] - E[z_i] E[u_i]$$

With

$$E[z_i] = 0$$

and

$$E[u_i] = 0$$

we have

$$\text{Cov}[z_i, u_i] = E[z_i u_i]$$

## The IV Model: Assumptions: Exogeneity in expectation form

Then, the covariance condition

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

is equivalent to the moment condition

$$E[z_i u_i] = 0$$

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 and non-autocorrelated

Homoskedasticity. That is,

$$\text{Var}[u_i | \mathbf{z}_i] = \sigma^2$$

for all  $i$ .

## The IV Model: Assumptions: Random sampling

**Random sampling.** The data we collect

$$\{(\mathbf{x}_i, \mathbf{z}_i, u_i) : i = 1, 2, \dots, n\}$$

are independent and identically distributed draws from the population.