Mile stone IV

Endogeneity and IV estimation

- Exogeneity
- Endogeneity
- Instrumental variables
- IV-Estimation of simple regression model

Exogeneity

Revisit the standard regression model,

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_K X_K + u, \tag{21}$$

and recall the standard assumption about the conditional mean of u:

$$\mathrm{E}(u|X_1, \ldots, X_K) = \mathrm{E}(u) = 0. \tag{22}$$

If this assumption is violated, then the OLS estimator $\hat{\boldsymbol{\beta}}$ will generally be biased:

$$\mathrm{E}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}(\boldsymbol{u}|\mathbf{X}).$$

Exogeneity

For experimental data, the regressors X_j are often under the control of the econometrician, hence we are usually satisfied with showing unbiasedness conditional on a specific "design matrix" $\mathbf{X}$

For observational data, very often we have to deal with stochastic regressors X_j. In this case, we have to show for unbiasedness for all possible $\mathbf{X}$, i.e.:

$$\mathrm{E}_{\mathbf{X}}\left(\mathrm{E}(\hat{\boldsymbol{\beta}}|\mathbf{X})\right) = \boldsymbol{\beta} + \mathrm{E}_{\mathbf{X}}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}(\boldsymbol{u}|\mathbf{X})\right),$$

where $\mathrm{E}_{\mathbf{X}}(g(\mathbf{X}))$ is the expectation which respect to all possible values $\mathbf{X}$.

Exogeneity

Therefore, assumption (22) is often substituted by a weaker assumption:

All explanatory variables X_j are **exogenous**, meaning that X_j is uncorrelated with the unexplained disturbance term u, i.e.:

$$\text{Cov}(X_j, u) = 0, \qquad \forall j = 1, \ldots, K. \tag{23}$$

- (22) implies (27), but (27) is a weaker assumption than (22).
- Hence, (27) guarantees consistency of the OLS estimator, but not necessarily unbiasedness.

Endogeneity

If condition (27) does not hold for a particular explanatory variable X_j, meaning that X_j is correlated with the unexplained disturbance term u, then **endogeneity** is present:

$$\text{Cov}(X_j, u) \neq 0, \tag{24}$$

Endogeneity is a serious issue:

- (24) implies that both (22) and (27) are violated;
- the OLS estimator $\hat{\beta}_j$ is usually not only biased, but also inconsistent.

Endogeneity

Intuitive reason for biasedness also in the limit:

- The explanatory variable X_j in (21) cannot be changed independently from u;
- Only **part** of the change we observe in $\mathrm{E}(Y|X_1, \ldots, X_K)$, when we change X_j, is directly caused by X_j.
- An additional change in $\mathrm{E}(Y|X_1, \ldots, X_K)$ is caused by the unobserved factors summarized in the disturbance term u.
- Due to the correlation between X_j and u, changing X_j changes $\mathrm{E}(u|X_1, \ldots, X_K)$ and hence $\mathrm{E}(Y|X_1, \ldots, X_K)$.

Endogeneity

What is the partial (average) effect of changing X_1 to $X_1^\star$, say e.g. $X_1^\star = X_1 + 1$, *ceteris paribus*?

From (21) we obtain:

$$\mathrm{E}(Y|X_1^\star, \ldots, X_K) - \mathrm{E}(Y|X_1, \ldots, X_K) =$$
$$\beta_1(X_1^\star - X_1) + (\mathrm{E}(u|X_1, \ldots, X_K) - \mathrm{E}(u|X_1^\star, \ldots, X_K)).$$

Under the standard assumption (22), the second term disappears. The effect is equal to β_1.

Endogeneity

If u and X_1 are correlated, then regressing u on X_j on yields $\mathrm{E}(u|X_1,\ldots,X_K) = \gamma_1 X_1 + \gamma_0$ with $\gamma_1 \neq 0$.

Hence

$$\mathrm{E}(Y|X_1^\star,\ldots,X_K) - \mathrm{E}(Y|X_1,\ldots,X_K) =$$
$$\beta_1(X_1^\star - X_1) + \gamma_1(X_1^\star - X_1) = (\beta_1+\gamma_1)(X_1^\star - X_1).$$

Under endogeneity, i.e. $\gamma_1 \neq 0$, the change we observe in $\mathrm{E}(Y|X_1^\star,\ldots,X_K)$ in reaction to changing X_1, is the sum of the direct (causal) effect of changing X_1 on Y, β_1, and the indirect effect of changing additional (unobserved) factors in u, γ_1.

Endogeneity

Endogeneity is a serious issue:

- There is no way to separate the two factors, when endogeneity is ignored.

- OLS estimation applied to regression model (21), yields an estimator of $(\hat{\beta}_1 + \hat{\gamma}_1)$, rather than the causal effect $\hat{\beta}_1$.

- If u and X_1 are positively correlated, then $\gamma_1 > 0$, and the OLS estimator overrates the direct effect of X_1 on Y.

- If u and X_1 are negatively correlated, then $\gamma_1 < 0$, and the OLS estimator underrates the direct effect of X_1 on Y.

Endogeneity

There are three main reasons for endogeneity:

- A relevant variables is omitted from the model $\Rightarrow$ omitted variable bias.
- There is (classical) measurement error in one of the variables $\Rightarrow$ attenuation bias.
- There is reverse causality: X_j affects Y and Y simultaneously affects X_j $\Rightarrow$ simultaneous equation bias.

In general, all coefficients are biased even if there is only one endogenous variable.

Instrumental Variable

Find a suitable variable Z, the so-called instrumental variable, with the following properties:

- The instrument must be **relevant**, Z is correlated with X_j i.e.:

$$\text{Cov}(Z, X_j) \neq 0.$$

- The instrument Z must be **valid** (exogenous), Z is uncorrelated with u:

$$\text{Cov}(Z, u) = 0.$$

IV Estimation in the simple regression model

Because $\mathrm{Cov}(Z, X_j) \neq 0$, the instrument Z introduces **exogenous** change in X_j, without effecting U.

We can test whether $\mathrm{Cov}(Z, X_j) \neq 0$ using

$$H_0 : \pi_1 = 0$$

in the first-stage regression

$$X_j = \pi_0 + \pi_1 Z + V. \tag{25}$$

IV Estimation in the simple regression model

Consider the simple regression model,

$$Y = \beta_0 + \beta_1 X_1 + u. \tag{26}$$

Using the **methods of moment** approach, we have

$$\mathrm{Cov}(Z, Y) = \beta_1 \mathrm{Cov}(Z, X_1) + \mathrm{Cov}(Z, U).$$

Because $\mathrm{Cov}(Z, U) = 0$, we obtain:

$$\beta_1 = \mathrm{Cov}(Z, Y)/\mathrm{Cov}(Z, X_1).$$

IV Estimation in the simple regression model

Hence, in the simple regression model, the IV moment estimator is given by:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(z_i - \overline{z})(y_i - \overline{y})}{\sum_{i=1}^{N}(z_i - \overline{z})(x_i - \overline{x})}.$$

Alternative estimation method - two-stage least square:

- Use the first stage equation to estimate the part of X_1 that is explained by Z:

$$\hat{X}_1 = \pi_0 + \pi_1 Z.$$

IV Estimation in the simple regression model

- Use $\hat{X}_1$ instead of X_1 as predictor in regression model (26):

$$Y = \beta_0 + \beta_1 \hat{X}_1 + u. \tag{27}$$

- Both methods yield a consistent estimator of β_1.

- Standard errors of the IV estimator obtained from (27) are larger than for OLS estimation.

- For a weak instrument, where $\mathrm{Cov}(Z, X_1)$ is close to 0, standard errors may be extremely large.