

LECTURE NOTE 1

CORRELATED REGRESSORS AND INSTRUMENTAL VARIABLES

I. INTRODUCTION

In the neoclassical regression model:

$$y = X\beta + \epsilon, \quad E[\epsilon | X] = 0, \quad \text{Var}(\epsilon | X) = \sigma^2 I.$$

In the generalized NCRM, the disturbance covariance matrix has more general structure (“second order” departure from NCRM):

$$y = X\beta + \epsilon, \quad E[\epsilon | X] = 0, \quad \text{but } \text{Var}(\epsilon | X) = \sigma^2 \Omega.$$

When the regressors are endogenous, we have “first order” departure from NCRM:

$$y = X\beta + \epsilon, \quad \text{but } E[\epsilon X] \neq 0, \quad \implies \quad E[\epsilon | X] \neq 0.$$

II. EXAMPLES

1. AUTOCORRELATION WITH LAGGED DEPENDENT VARIABLE:

$$y_t = x_t' \beta + \alpha y_{t-1} + \epsilon_t, \quad \epsilon_t = \rho \epsilon_{t-1} + u_t, \quad u_t \sim \text{i.i.d.}(0, \sigma_u^2).$$

$$\begin{aligned} \implies E[y_{t-1} \epsilon_t] &= E[(x_{t-1}' \beta + \alpha y_{t-2} + \epsilon_{t-1})(\rho \epsilon_{t-1} + u_t)] \\ &= \alpha \rho E[y_{t-2} \epsilon_{t-1}] + \rho E[\epsilon_{t-1}^2]. \end{aligned}$$

If y_t and ϵ_t are stationary, then $E[y_{t-1} \epsilon_t] = E[y_{t-2} \epsilon_{t-1}]$ and

$$E[y_{t-1} \epsilon_t] = \frac{\rho \sigma_u^2}{(1 - \alpha \rho)(1 - \rho^2)}.$$

2. OMITTED VARIABLES:

True model (long regression):

$$y_i = x_i'\beta + z_i'\gamma + \epsilon_i, \quad \epsilon \sim \text{i.i.d.}(0, \sigma^2),$$

where ϵ_t is independent of x_i and z_i .

Suppose that z_i is unobserved, so that we consider the model (short regression):

$$y_i = x_i'\beta + u_i, \quad u_i \equiv \epsilon_i + z_i'\gamma.$$

Hence,

$$\begin{aligned} E[x_i u_i] &= E[x_i \epsilon_i] + E[x_i z_i' \gamma] \\ &= 0 + E[x_i z_i'] \gamma \\ &\neq 0, \quad \text{in general.} \end{aligned}$$

Therefore, the short regression does not satisfy the NCRM assumption.

3. MEASUREMENT ERRORS (PERMANENT INCOME HYPOTHESIS MODEL):

True model:

$$y_i = \alpha + \beta z_i + u_i,$$

but z_i is not observed. Instead, we observe x_i , where

$$x_i = z_i + v_i.$$

Assume that $(u_i, v_i) \sim \text{i.i.d.}(0, \Sigma)$, i.e., $\text{Var}(u_i) = \sigma_{11}$, $\text{Var}(v_i) = \sigma_{22}$, and $\text{Cov}(u_i, v_i) = \sigma_{12}$. Also, $E[u_i] = E[v_i] = 0$ and (u_i, v_i) are independent of z_i .

Substituting x_i for z_i in the first equation gives

$$\begin{aligned} y_i &= \alpha_i + \beta(x_i - v_i) + u_i \\ &= \alpha_i + \beta x_i + (u_i - \beta v_i) \\ &= \alpha_i + \beta x_i + \epsilon_i, \end{aligned}$$

where $\epsilon_i \equiv u_i - \beta v_i$. It follows that

$$\begin{aligned} E[x_i \epsilon_i] &= E[(z_i + v_i)(u_i - \beta v_i)] \\ &= E[v_i u_i] - \beta E[v_i^2] \\ &= \sigma_{12} - \beta \sigma_{22} \\ &\neq 0, \quad \text{in general.} \end{aligned}$$

4. SIMULTANEOUS EQUATIONS (KEYNESIAN MODEL):

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad \epsilon_i \sim \text{i.i.d.}(0, \sigma^2),$$

$$x_i = y_i + z_i, \quad z_i \text{ independent of } \epsilon_i,$$

where y_i is consumption, x_i is income (or output), z_i is investment. Note that the second equation is merely an income identity

$$\begin{aligned} E[x_i \epsilon_i] &= E[(\alpha + \beta x_i + \epsilon_i + z_i) \epsilon_i] \\ &= \beta E[x_i \epsilon_i] + E[\epsilon_i^2]. \\ \implies E[x_i \epsilon_i] &= \frac{\sigma^2}{1 - \beta} \neq 0. \end{aligned}$$

5. SIMULTANEOUS EQUATIONS (SUPPLY AND DEMAND):

Demand equation:

$$q_t = x_t' \beta + \alpha p_t + \epsilon_t, \quad (\alpha < 0),$$

Inverse supply equation:

$$p_t = z_t' \gamma + \delta q_t + \eta_t, \quad (\delta > 0).$$

Assume that $E[\epsilon_t] = E[\eta_t] = 0$, and (ϵ_t, η_t) are independent of x_t and z_t .

For the demand equation:

$$\begin{aligned}
E[p_t \epsilon_t] &= E[(z_t' \gamma + \delta q_t + \eta_t) \epsilon_t] \\
&= \delta E[q_t \epsilon_t] + E[\eta_t \epsilon_t] \\
&= \delta E[(x_t' \beta + \alpha p_t + \epsilon_t) \epsilon_t] + E[\eta_t \epsilon_t] \\
&= \delta \alpha E[p_t \epsilon_t] + \delta \sigma_\epsilon^2 + \sigma_{12}. \\
\implies E[p_t \epsilon_t] &= \frac{\delta \sigma_\epsilon^2 + \sigma_{12}}{(1 - \delta \alpha)} \neq 0.
\end{aligned}$$

III. LEAST-SQUARES ESTIMATOR WITH ENDOGENOUS REGRESSORS

$$\begin{aligned}
b &= (X'X)^{-1} X'y = \beta + (X'X)^{-1} X'\epsilon. \\
\text{plim}_{n \rightarrow \infty} b &= \beta + \left(\text{plim}_{n \rightarrow \infty} \frac{1}{n} X'X \right)^{-1} \text{plim}_{n \rightarrow \infty} \left(\frac{1}{n} X'\epsilon \right). \\
\text{plim}_{n \rightarrow \infty} \frac{1}{n} X'X &= \text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n x_i x_i' = E[x_i x_i'] = \Sigma_{xx},
\end{aligned}$$

by the WLLN.

$$\text{plim}_{n \rightarrow \infty} \frac{1}{n} X'\epsilon = \text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n x_i \epsilon_i = E[x_i \epsilon_i] = \Sigma_{x\epsilon}.$$

But, $E[x_i \epsilon_i] \neq 0$, therefore

$$\text{plim}_{n \rightarrow \infty} b = \beta + \Sigma_{xx}^{-1} \Sigma_{x\epsilon} = \beta^* \neq \beta.$$

This is a major problem, if the parameter of interest is β (the “structural parameter”), and not β^* (the “reduced form parameter”).

METHOD OF MOMENTS INTERPRETATION OF THE LS PROBLEM:

The LS solves:

$$0 = \frac{1}{n} X'(y - Xb) = \frac{1}{n} \sum_{i=1}^n x_i (y_i - x_i' b),$$

i.e., the normal equations. But, in the population

$$E[x_i (y_i - x_i' \beta)] = E[x_i \epsilon_i] \neq 0.$$

So, it is as if we were using the wrong moments conditions. That is, the analogy principle does not hold. The sample moments do not have population counterparts.

IV. INSTRUMENTAL VARIABLE APPROACH

Suppose that there exists some $Z = (z_1, \dots, z_n)'$, such that:

1.

$$\text{plim}_{n \rightarrow \infty} \frac{1}{n} Z'X = \text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n z_i x'_i = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n E[z_i x'_i] \equiv \Sigma_{zx},$$

where Σ_{zx} is a $k \times k$ matrix with $\det(\Sigma_{zx}) \neq 0$,

2.

$$\text{plim}_{n \rightarrow \infty} \frac{1}{n} Z'\epsilon = \text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n z_i \epsilon_i = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n E[z_i \epsilon_i] \equiv \Sigma_{z\epsilon} = \mathbf{0},$$

where $\Sigma_{z\epsilon}$ is a $k \times 1$ vector and $\mathbf{0}$ is a $k \times 1$ vector of zeros (a sufficient condition would be $E[z_i \epsilon_i] = 0, \forall i$). Then the variables z_i are called: *instrumental variables* for x_i .

INSTRUMENTAL VARIABLE (IV) ESTIMATOR:

$$\begin{aligned} \hat{\beta}_{IV} &= (Z'X)^{-1} Z'y \\ &= \beta + \left(\frac{1}{n} Z'X \right)^{-1} \left(\frac{1}{n} Z'\epsilon \right) \\ &\longrightarrow \beta + \Sigma_{zx}^{-1} \Sigma_{z\epsilon} \text{ in probability} \\ &= \beta. \end{aligned}$$

So, the IV estimator is a consistent estimator of β .

METHOD OF MOMENTS INTERPRETATION OF IV ESTIMATOR:

Use analogue samples for the population condition: If $E[x_i \epsilon_i] = 0$ then

$$\begin{aligned} E[z_i \epsilon_i] &= \Sigma_{z\epsilon} = 0. \\ \implies \hat{\beta} &\text{ solves } \frac{1}{n} \sum_{i=1}^n z_i (y_i - x'_i \hat{\beta}) = \frac{1}{n} \sum_{i=1}^n z_i \hat{\epsilon}_i = 0, \end{aligned}$$

where $\hat{\epsilon}_i = y_i - x_i' \hat{\beta}$.

V. ASYMPTOTIC PROPERTIES OF IV ESTIMATOR

Suppose that $E[z_i \epsilon_i] = 0$, then by the CLT

$$\frac{1}{\sqrt{n}} Z' \epsilon = \frac{1}{\sqrt{n}} \sum_{i=1}^n z_i \epsilon_i \xrightarrow{D} N(0, \Phi),$$

where $\Phi = E[\epsilon_i^2 z_i z_i']$.

Therefore,

$$\begin{aligned} \sqrt{n}(\hat{\beta} - \beta) &= \left(\frac{1}{n} Z' X \right)^{-1} \frac{1}{\sqrt{n}} Z' \epsilon \\ &= \left(\frac{1}{n} \sum_{i=1}^n z_i x_i' \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n z_i \epsilon_i \\ &\longrightarrow N(0, \Sigma_{zx}^{-1} \Phi \Sigma_{zx}^{-1}) \text{ in distribution.} \end{aligned}$$

CONSISTENT ESTIMATOR FOR $M_{zx}^{-1} \Phi M_{zx}^{-1}$:

By construction, a consistent estimator for Σ_{zx} is given by

$$\hat{\Sigma}_{zx} = \frac{1}{n} \sum_{i=1}^n z_i x_i' = \frac{1}{n} Z' X \xrightarrow{p} \Sigma_{zx}.$$

CONSISTENT ESTIMATOR FOR Φ :

Use Newey-West estimator (in the time series context):

$$\Phi = \Gamma_0 + \sum_{j=1}^{\infty} (\Gamma_j + \Gamma_j'),$$

where

$$\begin{aligned} \Gamma_j &= \text{plim}_{T \rightarrow \infty} \left(\frac{1}{T} \sum_{t=j+1}^T \epsilon_t \epsilon_{t-j} z_t z_{t-j}' \right) \\ &= E[\epsilon_t \epsilon_{t-j} z_t z_{t-j}'], \quad \text{if } \epsilon_t \text{ and } z_t \text{ are stationary.} \\ \implies \hat{\Gamma}_j &= \frac{1}{T} \sum_{t=j+1}^T \hat{\epsilon}_t \hat{\epsilon}_{t-j} z_t z_{t-j}' \xrightarrow{p} \Gamma_j \end{aligned}$$

and

$$\hat{\Phi} = \hat{\Gamma}_0 + \sum_{j=1}^M \left(1 - \frac{j}{M+1}\right) (\hat{\Gamma}_j + \hat{\Gamma}'_j) \xrightarrow{p} \Phi,$$

if $M \rightarrow \infty$, $M/T^{1/4} \rightarrow 0$, as $T \rightarrow \infty$.

REMARKS:

1. If data are sampled randomly, so that

$$E[\epsilon_i \epsilon_j z_i z_j] = 0, \quad \text{if } i \neq j,$$

then use the Eicker-White estimator:

$$\hat{\Phi} = \hat{\Gamma}_0 = \frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_i^2 z_i z_i' \xrightarrow{p} \Gamma_0 = \Phi.$$

2. If $\text{Var}(\epsilon | Z) = \sigma^2 I$, then $\hat{\Phi}$ simplifies to

$$\hat{\Phi} = \hat{\sigma}^2 \left(\frac{1}{n} Z' Z \right)^{-1},$$

where $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - x_i' \hat{\beta})^2$.

3. The generalized NCRM is a special case of IV estimator

$$E[\epsilon_t | x_t] = 0, \quad \text{Cov}(\epsilon_t, \epsilon_s | x_t, x_s) = \sigma^2 \omega_{ts}.$$

Therefore, can use $z_i = x_i$, i.e., x_i serves as instrument of itself:

$$\hat{\beta}_{IV} = (Z' X)^{-1} Z' y = (X' X)^{-1} X' y.$$