Instrumental Variables

Paolo Zacchia

Econometric Theory

Lecture 10

Solving Endogeneity: Instrumental Variables

- This lecture is centered on the most characteristic concept of econometrics: **instrumental variables**.

- Instrumental variables have been developed as the classical solution to the problem of **endogeneity**, that is when:

$$\mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{x}_i\right] \neq 0$$

  i.e. White's "exogeneity" assumption in linear models fails.

- The term "endogeneity" originally comes from the analysis of Simultaneous Equations Models; it now applies broadly.

- Before proceeding, it is useful to elaborate a **taxonomy** of endogeneity, i.e. 1. omitted variable bias; 2. simultaneity; 3. measurement error; 4. structural endogeneity.

Omitted Variable Bias

- The Omitted Variable Bias was analyzed in Lecture 7. To recapitulate, if the CEF is $\mathbb{E}\left[Y_i \middle| \boldsymbol{x}_i, S_i\right] = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \delta_0 S_i$,

$$\widehat{\boldsymbol{\beta}}_{OLS} \xrightarrow{p} \boldsymbol{\beta}_0 + \delta_0 \, \mathbb{E}\left[\boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right]^{-1} \mathbb{E}\left[\boldsymbol{x}_i S_i\right]$$

  is the probability limit of the OLS estimator that is based on $\boldsymbol{x}_i$ and that "omits" the variable S_i.

- The bias term depends on *i.* δ_0, the coefficient of S_i in the CEF, and *ii.* the population linear projection of S_i on $\boldsymbol{x}_i$.

- If either component of the bias is zero, omitting S_i is then innocuous!

- Lecture 7 generalized this to multiple omitted variables.

Fixed effects

- A particular instance of Omitted Variable Bias occurs with **panel data**. Let the true model be:

$$y_{it} = \alpha_i + \mathbf{x}_{it}^{\mathrm{T}} \boldsymbol{\beta}_0 + \varepsilon_{it}$$

  where $\mathbb{E}\left[\varepsilon_{it} \middle| \boldsymbol{x}_{it}\right] = 0$, while α_i is an unobserved **individual fixed effect** (an error term constant over time for each i).

- Endogeneity due to **unobserved heterogeneity** occurs if:

$$\mathbb{E}\left[\alpha_i \middle| \boldsymbol{x}_{it}\right] \neq 0$$

  that is, unobserved idiosyncratic factors that are **constant** over time (like the ability of workers – or the know-how of firms) may correlate with the explanatory variables $\boldsymbol{x}_{it}$.

- Instrumental variables may not be necessary in this case.

"Within" and "between" transformations

- Some model **transformations** help address fixed effects.

- The "**within" transformation** is:

$$y_{it} - \bar{y}_i = \left(\mathbf{x}_{it} - \bar{\mathbf{x}}_i\right)^{\mathrm{T}} \boldsymbol{\beta}_0 + \left(\varepsilon_{it} - \bar{\varepsilon}_i\right)$$

  where $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, $\bar{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{it}$, & $\bar{\varepsilon}_i = \frac{1}{T}\sum_{t=1}^{T} \varepsilon_{it}$.

- The "**between" transformation** is:

$$\Delta y_{it} = \Delta \mathbf{x}_{it}^{\mathrm{T}} \boldsymbol{\beta}_0 + \Delta \varepsilon_{it}$$

  where $\Delta a_{it} = a_{it} - a_{i(t-1)}$ ($\Delta \mathbf{a}_{it} = \mathbf{a}_{it} - \mathbf{a}_{i(t-1)}$ for vectors).

- Both transformations mechanically remove fixed effects and are asymptotically equivalent. However, $\boldsymbol{\beta}_0$ is inferred from **time-varying** explanatory variables $\boldsymbol{x}_{it}$ only.

Simultaneity

- Simultaneity is the classical instance of endogeneity.

- Consider a SEM $\mathbf{\Gamma y}_i = \mathbf{\Phi z}_i + \boldsymbol{\varepsilon}_i$ whose exogenous variables $\boldsymbol{z}_i$ are mean-independent of the P error terms $\boldsymbol{\varepsilon}_i$ as:

$$\mathbb{E}\left[\boldsymbol{\varepsilon}_i \middle| \boldsymbol{z}_i\right] = \mathbf{0}$$

  implying $\mathbb{E}\left[\boldsymbol{z}_i \varepsilon_{pi}\right] = \mathbf{0}$ for $p = 1, \ldots, P$ (or: $\mathbb{E}\left[\boldsymbol{z}_i \boldsymbol{\varepsilon}_i^{\mathrm{T}}\right] = \mathbf{0}$).

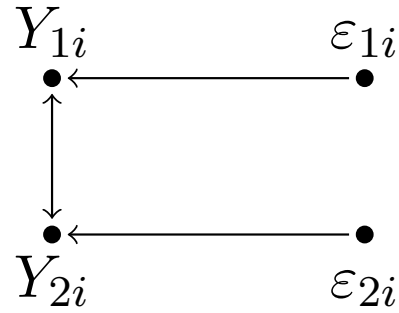
- **By construction** this does not apply to the P endogenous variables $\boldsymbol{y}_i = \mathbf{\Pi} \boldsymbol{z}_i + \mathbf{\Gamma}^{-1} \boldsymbol{\varepsilon}_i$ because:

$$\mathbb{E}\left[\boldsymbol{y}_i \boldsymbol{\varepsilon}_i^{\mathrm{T}}\right] = \mathbf{\Pi} \cdot \underbrace{\mathbb{E}\left[\boldsymbol{z}_i \boldsymbol{\varepsilon}_i^{\mathrm{T}}\right]}_{=\mathbf{0}} + \mathbf{\Gamma}^{-1} \cdot \mathbb{E}\left[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^{\mathrm{T}}\right] = \mathbf{\Gamma}^{-1} \cdot \mathbb{V}\mathrm{ar}\left[\boldsymbol{\varepsilon}_i\right] \neq \mathbf{0}$$

  hence the expression “endogeneity.”

Simultaneity: graphical intuition

- To build intuition, consider a simple SEM of two equations.
- There are two endogenous variables Y_{1i} and Y_{2i}, as well as two error terms ε_{1i} and ε_{2i}
- Structural relationships between them are displayed in the next graph.



- All error terms indirectly affect all endogenous variables.
- Endogeneity follows as $\mathbb{E}\left[\varepsilon_{1i} \middle| Y_{2i}\right] \neq 0$ and $\mathbb{E}\left[\varepsilon_{2i} \middle| Y_{1i}\right] \neq 0$.

Measurement error: bivariate case (1/3)

- Failure of mean independence $\mathbb{E}[\varepsilon_i | \boldsymbol{x}_i] = 0$ can also occur when the explanatory variables $\boldsymbol{x}_i$ are *observed with error*.

- Consider a bivariate linear model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ with "exogenous" X_i ($\mathbb{E}[\varepsilon_i | X_i] = 0$).

- Consider researchers unable to observe the true variable X_i but only able to observe an *error-ridden version* of it:

$$X_i^* = X_i + \upsilon_i$$

  where υ_i denotes the *error in the measurement* of X_i.

- Let $\mathbb{E}[\upsilon_i] = 0$. Also let υ_i be **completely random**: that is, independent from both the "true" X_i and the error term ε_i.

$$\mathbb{E}[X_i \upsilon_i] = \mathbb{E}[\varepsilon_i \upsilon_i] = 0$$

  This scenario is known as **classical measurement error**.

Measurement error: bivariate case (2/3)

- Given actual data $\{(y_i, x_i^*)\}_{i=1}^N$ (with $\bar{x}^* = \frac{1}{N}\sum_{i=1}^N x_i^*$), the OLS estimator of the regression slope is inconsistent since:

$$\widehat{\beta}_{1,OLS} = \frac{\sum_{i=1}^N (x_i^* - \bar{x}^*)\, y_i}{\sum_{i=1}^N (x_i^* - \bar{x}^*)^2} \xrightarrow{p} \frac{\mathbb{C}\text{ov}\left[X_i + \upsilon_i, \beta_0 + \beta_1 X_i + \varepsilon\right]}{\mathbb{V}\text{ar}\left[X_i + \upsilon_i\right]}$$

$$= \frac{\beta_1 \mathbb{V}\text{ar}\left[X_i\right]}{\mathbb{V}\text{ar}\left[X_i\right] + \mathbb{V}\text{ar}\left[\upsilon_i\right]} = \beta_1 \left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_\upsilon^2}\right)$$

  where $\sigma_x^2 \equiv \mathbb{V}\text{ar}\left[X_i\right]$ and $\sigma_\upsilon^2 \equiv \mathbb{V}\text{ar}\left[\upsilon_i\right]$.

- The probability limit is *smaller* than β_1 to an extent that depends upon the **noise-to-signal ratio** $\sigma_\upsilon^2 \sigma_x^{-2}$.

$$\frac{\sigma_x^2}{\sigma_x^2 + \sigma_\upsilon^2} = \frac{1}{1 + \sigma_\upsilon^2 \sigma_x^{-2}} = 1 - \frac{\sigma_\upsilon^2 \sigma_x^{-2}}{1 + \sigma_\upsilon^2 \sigma_x^{-2}} \in (0, 1)$$

- This is called **attenuation bias**: intuitively, measurement error makes the relationship between Y_i and X_i^* weaker.

Measurement error: bivariate case (3/3)

- To better see the link with endogeneity, consider the model:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \varepsilon_i \\ &= \beta_0 + \beta_1 X_i^* + \varepsilon_i - \beta_1 \upsilon_i \end{aligned}$$

  where the *actual regressor* is X_i^*, and the actual error term is consequently $\varepsilon_i - \beta_1 \upsilon_i$.

- Clearly, by construction it is:

$$\mathbb{E}\left[\varepsilon_i - \beta_1 \upsilon_i \middle| X_i^*\right] \neq 0$$

  because X_i^* incorporates the error υ_i.

- No similar problem affects the dependent variable Y_i: if the researcher observes it with error ($Y_i^* = Y_i + \upsilon_i$) the model:

$$Y_i^* = \beta_0 + \beta_1 X_i + \varepsilon_i + \upsilon_i$$

  is still estimated consistently if υ_i is "completely random."

Measurement error: multivariate case (1/3)

- This generalizes to the multivariate model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}$.

- Each explanatory variable X_{ki} may or may not be affected by measurement error: the actually observed variables are:

$$X_{ki}^* = X_{ki} + \upsilon_{ki}$$

  for $k = 1, \ldots, K$ and $i = 1, \ldots, N$.

- In **compact** matrix notation, the actual realizations of the explanatory variables as well as the measurement errors are collected by two $N \times K$ matrices.

$$\mathbf{X}^* = \begin{bmatrix} x_{11}^* & x_{21}^* & \cdots & x_{K1}^* \\ x_{12}^* & x_{22}^* & \cdots & x_{K2}^* \\ \vdots & \vdots & \ddots & \vdots \\ x_{1N}^* & x_{2N}^* & \cdots & x_{KN}^* \end{bmatrix}; \; \mathbf{U}^* = \begin{bmatrix} \upsilon_{11}^* & \upsilon_{21}^* & \cdots & \upsilon_{K1}^* \\ \upsilon_{12}^* & \upsilon_{22}^* & \cdots & \upsilon_{K2}^* \\ \vdots & \vdots & \ddots & \vdots \\ \upsilon_{1N}^* & \upsilon_{2N}^* & \cdots & \upsilon_{KN}^* \end{bmatrix}$$

Measurement error: multivariate case (2/3)

- The "true estimated model" is:

$$\mathbf{y} = \mathbf{X}^* \boldsymbol{\beta}_0 + \boldsymbol{\varepsilon} - \mathbf{U}\boldsymbol{\beta}_0$$

  and the "actual" error term is now $\varepsilon_i - \sum_{k=1}^{K} \beta_{k,0} \upsilon_{ki}$.

- Again by construction, it is:

$$\mathbb{E}\left[\varepsilon_i - \sum_{k=1}^{K} \beta_{k,0} \upsilon_{ki} \middle| \boldsymbol{x}_i^* = \mathbf{x}_i^* \right] \neq 0$$

  where $\mathbf{x}_i^*$ is the i-th row of $\mathbf{X}^*$.

- This holds even under "complete randomness" of υ_{ki}, that is:

$$\mathbb{E}\left[X_{k'i} \upsilon_{ki} \right] = \mathbb{E}\left[\varepsilon_{ki} \upsilon_i \right] = 0$$

  for $k, k' = 1, \ldots, K$.

Measurement error: multivariate case (3/3)

- Define for convenience the following probability limits.

$$\frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{X} \xrightarrow{p} \boldsymbol{\Sigma}_{\boldsymbol{x}}$$
$$\frac{1}{N}\mathbf{U}^{\mathrm{T}}\mathbf{U} \xrightarrow{p} \boldsymbol{\Sigma}_{\boldsymbol{v}}$$

- Thus, the "actual" OLS estimator can be written as follows.

$$\widehat{\boldsymbol{\beta}}_{OLS} = \left(\mathbf{X}^{*\mathrm{T}}\mathbf{X}^{*}\right)^{-1}\mathbf{X}^{*\mathrm{T}}\mathbf{y} =$$
$$\boldsymbol{\beta}_0 + \left[\frac{1}{N}\left(\mathbf{X}+\mathbf{U}\right)^{\mathrm{T}}\left(\mathbf{X}+\mathbf{U}\right)\right]^{-1}\frac{1}{N}\left(\mathbf{X}+\mathbf{U}\right)^{\mathrm{T}}\left(\boldsymbol{\varepsilon}-\mathbf{U}\boldsymbol{\beta}_0\right)$$

- As $\frac{1}{N}\mathbf{U}^{\mathrm{T}}\boldsymbol{\varepsilon} \xrightarrow{p} \mathbf{0}$ and $\frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{U} \xrightarrow{p} \mathbf{0}$ (a matrix of zeroes), it is:

$$\widehat{\boldsymbol{\beta}}_{OLS} \xrightarrow{p} \left[\mathbf{I} - \left(\boldsymbol{\Sigma}_{\boldsymbol{x}} + \boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{v}}\right]\boldsymbol{\beta}_0 \leq \boldsymbol{\beta}_0$$

  thus generalizing the attenuation bias formula.

Structural endogeneity (1/2)

- Endogeneity is said "structural" when built in the model.
- Consider for example the following time-series model:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \mathbf{x}_t^{\mathrm{T}} \boldsymbol{\gamma}_0 + \mathbf{x}_{t-1}^{\mathrm{T}} \boldsymbol{\gamma}_1 + \varepsilon_t$$
$$\varepsilon_t = \rho \varepsilon_{t-1} + \xi_t$$

  where ε_t features an AR(1) structure with **autoregressive** parameter $\rho \in (0, 1)$ and a white noise "innovation" ξ_t.

- It is obvious that:

$$\mathbb{E}\left[\varepsilon_t \middle| Y_{t-1}\right] = \mathbb{E}\left[\rho \varepsilon_{t-1} + \xi_t \middle| Y_{t-1}\right] = \\ = \rho\, \mathbb{E}\left[\varepsilon_{t-1} \middle| Y_{t-1}\right] + \mathbb{E}\left[\xi_t \middle| Y_{t-1}\right] \neq 0$$

  and even if ξ_t is conditionally mean-independent, the grand error term ε_t is not, because its lag ε_{t-1} also affects the lag of the dependent variable Y_t *by construction*.

Structural endogeneity (2/2)

- Another example is a **spatial model** like:

$$\mathbf{y} = \beta_0 \boldsymbol{\iota} + \beta_1 \mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\gamma}_0 + \mathbf{W}\mathbf{X}\boldsymbol{\gamma}_1 + \boldsymbol{\varepsilon}$$

  with a $N \times N$ non-stochastic **spatial weighting matrix** written as $\mathbf{W}$ having a zero diagonal and that collects the w_{ij} **distances** between any two distinct units i and j.

- This model, common in say urban/regional economics, can be rewritten in terms of its solution for $\mathbf{y}$ as follows.

$$\mathbf{y} = (\mathbf{I} - \beta_1 \mathbf{W})^{-1} (\beta_0 \boldsymbol{\iota} + \mathbf{X}\boldsymbol{\gamma}_0 + \mathbf{W}\mathbf{X}\boldsymbol{\gamma}_1 + \boldsymbol{\varepsilon})$$

- Endogeneity arises due to feedback between the dependent variables (and the error terms) of different units.

$$\mathbb{E}\left[\varepsilon_i \middle| Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_N\right] \neq 0$$

- This is analogous to the problem of simultaneity!

Instrumental variables: introduction

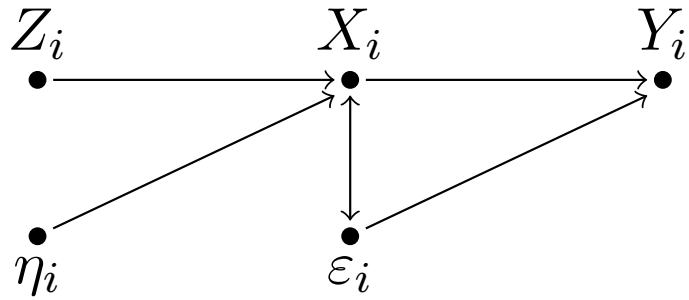
- Endogeneity can be solved through suitable **instrumental variables** (IVs), or **instruments**.
- Consider a bivariate linear model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ with X_i possibly "endogenous."
- IVs are variables Z_i that are mean-independent of the error:

$$\mathbb{E}\left[\varepsilon_i \middle| Z_i\right] = 0$$

  and *do not show up directly in the model* that explains Y_i.
- Moreover, IVs are "related" in a statistical and – possibly – "structural" sense to the endogenous regressor X_i.
- The "role" of IVs is best represented through an augmented structural model that explicitly features them.

Instrumental variables: graphical intuition

- Such an augmented model is best represented via a graph.



- Note that here X_i is co-dependent with ε_i (endogeneity).

- Moreover, X_i is determined by the IV Z_i as well as another error term η_i.

- However, the IV Z_i is unrelated to both error terms (ε_i, η_i) and does not itself "explain" Y_i, at least *not directly*.

- Any "effect" that Z_i has on Y_i is thus mediated through X_i.

Instrumental variables and triangular models

- Let the error term for X_i: η_i, also be mean-independent of the IV Z_i.

$$\mathbb{E}\left[\eta_i \middle| Z_i\right] = 0$$

- Under these hypotheses, the graph is the representation of a **restricted** version of a particular **triangular** SEM.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$X_i = \pi_0 + \pi_1 Z_i + \eta_i$$

- In such a SEM there are two endogenous variables: Y_i and X_i, and one exogenous variable: the IV Z_i.

- The model is **restricted** as Z_i does not affect Y_i directly.

- In IV parlance, the equation for Y_i is called the **structural form** and that for X_i, the **first stage**.

IV estimation of the triangular model (1/4)

- What if X_i is endogenous, that is $\mathbb{E}[\varepsilon_i | X_i] \neq 0$?
- Since the IV is exogenous, it is $\mathbb{E}[\varepsilon_i, \eta_i | Z_i] = 0$, so (π_0, π_1) are consistently estimated via OLS on the first stage.
- Even (β_0, β_1) are identified! Consider the **reduced form**:

$$Y_i = \beta_0 + \pi_0\beta_1 + \beta_1\pi_1 Z_i + \varepsilon_i + \beta_1\eta_i$$
$$X_i = \pi_0 + \pi_1 Z_i + \eta_i$$

  the first equation can be consistently estimated via OLS.
- There is here a unique mapping from the reduced form to the structural parameters.
- With a sample $\{(y_i, x_i, z_i)\}_{i=1}^N$, a **consistent** estimate of β_1 is obtained as the ratio between the two OLS coefficients of Z_i in the first equation of the reduced form and in the first stage, respectively.

IV estimation of the triangular model (2/4)

- To see this, write:

$$\widehat{\beta}_{1,IV} = \frac{\widehat{\beta_1 \pi_1}_{OLS}}{\widehat{\pi}_{1OLS}} = \frac{\frac{\sum_{i=1}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}(z_i - \bar{z})^2}}{\frac{\sum_{i=1}(z_i - \bar{z})(x_i - \bar{x})}{\sum_{i=1}(z_i - \bar{z})^2}}$$

$$= \frac{\sum_{i=1}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}(z_i - \bar{z})(x_i - \bar{x})}$$

  where $\bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i$, while $\bar{y}$ and $\bar{x}$ are analogous.

- This the **IV estimator** for β_1 in this triangular model.

- The IV estimator of the structural form's intercept β_0 is:

$$\widehat{\beta}_{0,IV} = \bar{y} - \widehat{\pi}_{0,OLS}\widehat{\beta}_{1,IV} - \widehat{\pi}_{1,OLS}\widehat{\beta}_{1,IV} \cdot \bar{z}$$

  that is, it is obtained by plugging the appropriate estimates for (π_0, π_1) and β_1 in the "average" reduced form for Y_i.

IV estimation of the triangular model (3/4)

- IV estimators are **consistent**: they are functions of other, consistent estimators of the model's reduced form.

- To see this more explicitly:

$$\widehat{\beta}_{1,IV} = \frac{\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(y_i - \bar{y})}{\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(x_i - \bar{x})} \xrightarrow{p} \frac{\mathbb{C}\text{ov}[Z_i, Y_i]}{\mathbb{C}\text{ov}[Z_i, X_i]}$$
$$= \beta_1 + \frac{\mathbb{C}\text{ov}[Z_i, \varepsilon_i]}{\mathbb{C}\text{ov}[Z_i, X_i]} = \beta_1$$

  where $\mathbb{C}\text{ov}[Z_i, \varepsilon_i] = \mathbb{E}[Z_i \varepsilon_i] = 0$ as the IV is "exogenous."

- The second-to-last equality follows from:

$$\mathbb{C}\text{ov}[Z_i, Y_i] = \beta_1 \mathbb{C}\text{ov}[Z_i, X_i] + \mathbb{C}\text{ov}[Z_i, \varepsilon_i]$$

  by expanding the expression for Y_i on the left-hand side.

IV estimation of the triangular model (4/4)

- The decomposition in the second line and an analysis akin to the application of Central Limit Theorems in the linear model show the following in the i.n.i.d. case.

$$\sqrt{N}\left(\widehat{\beta}_{1,IV} - \beta_1\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{E}\left[\varepsilon_i^2 \left(Z_i - \mathbb{E}\left[Z_i\right]\right)^2\right]}{\mathbb{C}\text{ov}\left[Z_i, X_i\right]^2}\right)$$

- Moreover, if the IV Z_i is independent of the error term ε_i (*homoscedasticity*) the previous expression reduces to:

$$\sqrt{N}\left(\widehat{\beta}_{1,IV} - \beta_1\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_0^2 \cdot \frac{\mathbb{V}\text{ar}\left[Z_i\right]^2}{\mathbb{C}\text{ov}\left[Z_i, X_i\right]^2}\right)$$

  where $\sigma_0^2 = \mathbb{E}\left[\varepsilon_i^2\right]$.

- The limiting variance of the IV estimator of β_1 is therefore *inversely proportional to the covariance between the IV: Z_i and the endogenous variable X_i.*

Failures of identification

- Why is the **restriction** that the IV Z_i does not appear in the structural form for Y_i key?

- The reduced form of the unrestricted triangular model is:

$$Y_i = \beta_0 + \pi_0\beta_1 + (\beta_1\pi_1 + \beta_2)\, Z_i + \varepsilon_i + \beta_1\eta_i$$
$$X_i = \pi_0 + \pi_1 Z_i + \eta_i$$

  which is not identified due to the extra parameter β_2 (the parameter for Z_i in the structural form).

- A similar issue occurs if the model is not triangular and Y_i enters the equation for X_i.

- This helps build intuition: the effect of X_i on Y_i, that is β_1, is identified by the variation in X_i predicted by the IV Z_i.

Requirements of a "good" instrument: summary

1. **Exogeneity**: the IV Z_i is mean-independent of both error terms i.e. $\mathbb{E}\left[\varepsilon_i, \eta_i \middle| Z_i\right] = 0$.

2. **Exclusion Restriction**: the IV Z_i does not directly affect the main endogenous variable Y_i of the structural form.

3. **No Reverse Causality**: the structural relationship that links the two endogenous variables is unidirectional: Y_i does not affect X_i directly (the system is indeed triangular).

4. **Relevance**: the covariance between the IV Z_i and the key endogenous explanatory variable X_i: $\mathbb{C}\text{ov}\left[Z_i, X_i\right]$, must be "sufficiently strong," as otherwise the IV estimates may be imprecise (a problem referred to as **weak instruments**).

This terminology is extensively used in econometrics.

Multiple Instruments (1/2)

- The analysis is now moved to the multivariate linear model.
- Let the model be $y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \varepsilon_i$ where $K \geq 2$; **partition** the explanatory variables as follows.

$$\mathbf{x}_i = \begin{bmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \end{bmatrix}$$

- Here $\mathbf{x}_{i1}$ is a subset of K_1 **exogenous** regressors...

$$\mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{x}_{i1}\right] = 0$$

- ...and $\mathbf{x}_{i2}$ is a subset of K_2 possibly **endogenous** ones.

$$\mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{x}_{i2}\right] \neq 0$$

- Clearly, $K_1 + K_2 = K$.

Multiple Instruments (2/2)

- Let a vector of K_2 **instrumental variables**, written as $\mathbf{z}_{i2}$, be available; these IVs are exogenous in the following sense.

$$\mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{z}_{i2}\right] = 0$$

- Group these IVs along the original K_1 exogenous regressors and write their realizations as follows.

$$\mathbf{z}_i = \begin{bmatrix} \mathbf{x}_{i1} \\ \mathbf{z}_{i2} \end{bmatrix}$$

- The error term is mean-independent of all the K exogenous variables.

$$\mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{z}_i\right] = 0$$

- Thus, the usual zero covariance implication also applies.

$$\mathbb{C}\text{ov}\left[\boldsymbol{z}_i, \varepsilon_i\right] = \mathbb{E}\left[\boldsymbol{z}_i \varepsilon_i\right] = \mathbb{E}_{\boldsymbol{z}}\left[\boldsymbol{z}_i \cdot \mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{z}_i\right]\right] = \mathbf{0}$$

The Multivariate IV Estimator (1/2)

- The (**just-identified**) **IV estimator** is defined as follows.

$$\widehat{\boldsymbol{\beta}}_{IV} = \left(\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i^{\mathrm{T}} \right)^{-1} \sum_{i=1}^{N} \mathbf{z}_i y_i$$

- By writing the $N \times K$ matrix that collects the $\mathbf{z}_i$ vectors of exogenous variables as:

$$\mathbf{Z} \equiv \begin{bmatrix} \mathbf{z}_1^{\mathrm{T}} \\ \mathbf{z}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{z}_N^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1K_1} & z_{11} & \dots & z_{1K_2} \\ x_{21} & \dots & x_{2K_1} & z_{21} & \dots & z_{2K_2} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{N1} & \dots & x_{NK_1} & z_{N1} & \dots & z_{NK_2} \end{bmatrix}$$

- . . . the above IV estimator can be more elegantly written in compact matrix notation.

$$\widehat{\boldsymbol{\beta}}_{IV} = \left(\mathbf{Z}^{\mathrm{T}} \mathbf{X} \right)^{-1} \mathbf{Z}^{\mathrm{T}} \mathbf{y}$$

The Multivariate IV Estimator (2/2)

- Clearly, the IV estimator is well-defined so long as matrix $\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i^{\mathrm{T}} = \mathbf{Z}^{\mathrm{T}}\mathbf{X}$ is invertible.

- It also admits a decomposition, analogous to that of OLS, which singles out the true value of the parameters $\boldsymbol{\beta}_0$.

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{IV} &= \boldsymbol{\beta}_0 + \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i^{\mathrm{T}}\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i \varepsilon_i \\ &= \boldsymbol{\beta}_0 + \left(\frac{1}{N}\mathbf{Z}^{\mathrm{T}}\mathbf{X}\right)^{-1} \frac{1}{N}\mathbf{Z}^{\mathrm{T}}\boldsymbol{\varepsilon}\end{aligned}$$

- If the variables in $\boldsymbol{z}_i$ are exogenous, the remainder term in the decomposition converges in probability to $\mathbb{E}\left[\boldsymbol{z}_i \varepsilon_i\right] = \mathbf{0}$.

- Hence, **the IV estimator is consistent**.

$$\widehat{\boldsymbol{\beta}}_{IV} \xrightarrow{p} \boldsymbol{\beta}_0$$

Example: Mincer, Revisited. (1/2)

- Return to the Mincer Equation from Lecture 7.

$$\log W_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 S_i + \varepsilon_i$$

- The error term $\varepsilon_i = \alpha_i + \epsilon_i$ is here the sum of **unobserved** "ability" α_i and some additional residual error ϵ_i.

- It is difficult to justify $\mathbb{E}[\alpha_i | S_i] = 0$ (implying $\mathbb{E}[\varepsilon_i | S_i] = 0$ too): education and individual ability are correlated.

- If a **suitable** IV Z_i is available, a consistent estimate of the parameters $\boldsymbol{\beta}_0 = (\beta_0, \beta_1, \beta_2, \beta_3)$ can be obtained with:

$$\mathbf{x}_i = \begin{bmatrix} 1 \\ x_i \\ x_i^2 \\ s_i \end{bmatrix} \quad \text{and} \quad \mathbf{z}_i = \begin{bmatrix} 1 \\ x_i \\ x_i^2 \\ z_i \end{bmatrix}$$

(x_i, s_i, z_i are the realizations of X_i, S_i, Z_i respectively).

Example: Mincer, Revisited. (2/2)

In this setting, a suitable IV Z_i is such that:

1. it is *exogenous*, in the sense that it does not correlate with unobserved ability and $\mathbb{E}[\varepsilon_i | Z_i] = 0$ holds;

2. it satisfies the *exclusion restriction*: it does not affect wages directly;

3. it fits a setting that rules out *reverse causality*: education is itself hardly affected by future individual wages;

4. it is *relevant*, that is, it correlates with education.

A famous example of IV Z_i for S_i is the "distance of one's home from a college" from a celebrated study by Card (1995). Others (many) have been proposed in the literature.

IVs in extended triangular models (1/2)

- The IV estimator in a multivariate environment can also be interpreted through the lenses of a triangular model.

- Let $K_1 = K - 1$, $K_2 = 1$, $\mathbf{x}_{i2} = x_{Ki} = s_i$, $\mathbf{z}_{i0} = z_i$. Write:

$$y_i = \mathbf{x}_{i1}^{\mathrm{T}} \boldsymbol{\beta}_{0\backslash K} + \delta_0 s_i + \varepsilon_i$$
$$s_i = \mathbf{x}_{i1}^{\mathrm{T}} \boldsymbol{\pi}_{0\backslash K} + \tau_0 z_i + \eta_i$$

  where $\boldsymbol{\beta}_0 = \begin{bmatrix} \boldsymbol{\beta}_{0\backslash K}^{\mathrm{T}} & \delta_0 \end{bmatrix}^{\mathrm{T}}$.

- Note the exclusion restriction: the IV z_i does not enter the equation for y_i.

- The reduced form of the model is as follows.

$$y_i = \mathbf{x}_{i1}^{\mathrm{T}} \left(\boldsymbol{\beta}_{0\backslash K} + \delta_0 \boldsymbol{\pi}_{0\backslash K} \right) + \delta_0 \tau_0 z_i + \varepsilon_i + \delta_0 \eta_i$$
$$s_i = \mathbf{x}_{i1}^{\mathrm{T}} \boldsymbol{\pi}_{0\backslash K} + \tau_0 z_i + \eta_i$$

IVs in extended triangular models (2/2)

- Parameter δ_0 is identified and – similarly as in the simpler, earlier triangular model – can be consistently estimated as:

$$\widehat{\delta}_{IV} = \frac{\widehat{\delta\tau}_{OLS}}{\widehat{\tau}_{OLS}} = \frac{\mathbf{z}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{y}}{\mathbf{z}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{z}} \left(\frac{\mathbf{z}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{s}}{\mathbf{z}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{z}} \right)^{-1} = \frac{\mathbf{z}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{y}}{\mathbf{z}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}1}\mathbf{s}}$$

- ...where $\mathbf{y}$, $\mathbf{s}$ and $\mathbf{z}$ are vectors of length N which collect, respectively, realizations y_i, s_i and z_i; while...

- ...$\mathbf{M}_{\mathbf{X}1} = \mathbf{I} - \mathbf{X}_1 \left(\mathbf{X}_1^{\mathrm{T}} \mathbf{X}_1 \right)^{-1} \mathbf{X}_1^{\mathrm{T}}$ is the **residual-maker matrix** that is obtained from the first $K-1$ (exogenous) explanatory regressors $\mathbf{x}_{i1}$.

- This estimator for δ_0 is numerically equivalent to the one obtained via the IV estimator; this result is ultimately an application of the Frisch-Waugh-Lovell algebra.

The Multivariate IV Estimator: Inference (1/2)

- How to perform statistical inference for the multivariate IV estimator?
- Let matrix $\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i^{\mathrm{T}} = \mathbf{Z}^{\mathrm{T}}\mathbf{X}$ be invertible.
- Also let the following probability limits be finite and of full rank.

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i^{\mathrm{T}} \xrightarrow{p} \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\boldsymbol{z}_i \boldsymbol{x}_i^{\mathrm{T}}\right] \equiv \widetilde{\mathbf{P}}_0$$

$$\frac{1}{N}\sum_{i=1}^{N} \varepsilon_i^2 \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}} \xrightarrow{p} \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\varepsilon_i^2 \boldsymbol{z}_i \boldsymbol{z}_i^{\mathrm{T}}\right] \equiv \widetilde{\boldsymbol{\Psi}}_0$$

- Finally, suppose that *the observations are independent* and that a suitable Central Limit Theorem can be extended to the random sequence $\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \mathbf{z}_i \varepsilon_i$.

The Multivariate IV Estimator: Inference (2/2)

- The limiting distribution of the IV estimator is as follows.

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_{IV}-\boldsymbol{\beta}_0\right)\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\widetilde{\mathbf{P}}_0^{-1}\widetilde{\boldsymbol{\Psi}}_0\widetilde{\mathbf{P}}_0^{\mathrm{T}-1}\right)$$

- The variance-covariance of the asymptotic distribution can be estimated as follows, for $e_i=y_i-\mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{IV}$.

$$\widehat{\mathbb{A}\mathrm{var}}_{HC}\left(\widehat{\boldsymbol{\beta}}_{IV}\right)=\left[\sum_{i=1}^{N}\mathbf{z}_i\mathbf{x}_i^{\mathrm{T}}\right]^{-1}\left[\sum_{i=1}^{N}e_i^2\mathbf{z}_i\mathbf{z}_i^{\mathrm{T}}\right]\left[\sum_{i=1}^{N}\mathbf{x}_i\mathbf{z}_i^{\mathrm{T}}\right]^{-1}$$

- In the clustering case (CCE), with $\mathbf{e}_c\equiv\mathbf{y}_c-\mathbf{X}_c\widehat{\boldsymbol{\beta}}_{IV}$ and $\mathbf{Z}_c$ being the cluster-specific sub-matrix of $\mathbf{Z}$, it is as follows.

$$\widehat{\mathbb{A}\mathrm{var}}_{CCE}\left(\widehat{\boldsymbol{\beta}}_{IV}\right)=$$

$$=\left[\sum_{c=1}^{C}\mathbf{Z}_c^{\mathrm{T}}\mathbf{X}_c\right]^{-1}\left[\sum_{c=1}^{C}\mathbf{Z}_c^{\mathrm{T}}\mathbf{e}_c\mathbf{e}_c^{\mathrm{T}}\mathbf{Z}_c\right]\left[\sum_{c=1}^{C}\mathbf{X}_c^{\mathrm{T}}\mathbf{Z}_c\right]^{-1}$$

- The homoscedastic and HAC cases are similarly obtained.

Two-Stages Least Squares: Introduction

- The multivariate IV estimator is a special case of the more general **Two-Stages Least Squares** (2SLS) estimator.
- The latter applies when the researcher has available **more instruments than endogenous variables**.
- Formally, it is $|\mathbf{z}_{i2}| = J_2 > K_2$; equivalently, $|\mathbf{z}_i| = J > K$.
- In principle, one can obtain different IV estimators for each appropriate subset of $\mathbf{z}_i$: the moder is **overidentified**.
- It is generally more efficient to **simultaneously** exploit all the information provided by the J_2 IVs.
- The 2SLS estimator does that via an abstract "procedure" in two stages (hence the name) that is described next.

Two-Stages Least Squares: Procedure (1/2)

- In the **first stage**, one performs a set of linear projections, one for **each** (endogenous) regressors $\mathbf{x}_i$ of the main model of interest, onto the set of exogenous variables $\mathbf{z}_i$.

- This results in a set of "projection" vectors $\widehat{\mathbf{x}}_i$ equal to:

$$\widehat{\mathbf{x}}_i^{\mathrm{T}} = \mathbf{z}_i^{\mathrm{T}} \left(\sum_{i=1}^{N} \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}} \right)^{-1} \left(\sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i^{\mathrm{T}} \right)$$

- . . . collected, in compact notation, as the $N \times K$ matrix:

$$\widehat{\mathbf{X}} = \mathbf{Z} \left(\mathbf{Z}^{\mathrm{T}} \mathbf{Z} \right)^{-1} \mathbf{Z}^{\mathrm{T}} \mathbf{X} = \mathbf{P}_{\mathbf{Z}} \mathbf{X}$$

  where $\mathbf{P}_{\mathbf{Z}} \equiv \mathbf{Z} \left(\mathbf{Z}^{\mathrm{T}} \mathbf{Z} \right)^{-1} \mathbf{Z}^{\mathrm{T}}$ is here the **projection matrix** based on the exogenous variables collected by the matrix $\mathbf{Z}$ of dimension $N \times J$.

Two-Stages Least Squares: Procedure (2/2)

- This is equivalent to running K **first stage regressions**:

$$x_{ki} = \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\pi}_{k0} + \eta_{ki}$$

  and calculating the fitted values $\widehat{x}_{ki} = \mathbf{z}_i^{\mathrm{T}} \widehat{\boldsymbol{\pi}}_{kOLS}$.

- Observe that if x_{ki} is contained in $\mathbf{z}_i$, it must be $\widehat{x}_{ki} = x_{ki}$, since a vector projected onto itself returns the input vector.

- Notice the similarity with the "first stage" equations of the previous triangular models: they are so named for a reason!

- In the **second stage**, the 2SLS estimator of $\boldsymbol{\beta}_0$ is obtained by an OLS regression of y_i onto the *projected* regressors $\widehat{\mathbf{x}}_i$: as if the model of interest were the following.

$$y_i = \widehat{\mathbf{x}}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \varepsilon_i$$

Two-Stages Least Squares: Expressions

- A 2SLS estimator can be thus written as follows.

$$\widehat{\boldsymbol{\beta}}_{2SLS} = \left(\sum_{i=1}^{N} \widehat{\mathbf{x}}_i \widehat{\mathbf{x}}_i^{\mathrm{T}} \right)^{-1} \sum_{i=1}^{N} \widehat{\mathbf{x}}_i y_i$$

- In is generally more convenient to express it using compact matrix notation:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{2SLS} &= \left(\widehat{\mathbf{X}}^{\mathrm{T}} \widehat{\mathbf{X}} \right)^{-1} \widehat{\mathbf{X}}^{\mathrm{T}} \mathbf{y} \\ &= \left(\mathbf{X}^{\mathrm{T}} \mathbf{P}_{\mathbf{Z}} \mathbf{X} \right)^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{P}_{\mathbf{Z}} \mathbf{y} \end{aligned}$$

  and the two expressions are equivalent since the projection matrix is symmetric and idempotent.

- Clearly, the 2SLS estimator is generally calculated directly, without explicitly following the procedure in two stages.

The Multivariate IV Estimator and 2SLS

- In the **just-identified** case ($J = K$), the 2SLS and the IV estimators coincide.

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{2SLS} &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{y} \\
&= \left[\mathbf{X}^{\mathrm{T}}\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{X}\right]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{y} \\
&= \left(\mathbf{Z}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\left(\mathbf{X}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{y} \\
&= \left(\mathbf{Z}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{y} \\
&= \widehat{\boldsymbol{\beta}}_{IV}
\end{aligned}$$

- In the derivation above, the third line is only possible as $\mathbf{X}$ and $\mathbf{Z}$ have the same (column) dimensions.

- Otherwise, some properties about the inversion of products of matrices would not be applicable.

Consistency and geometry of 2SLS (1/2)

- Also this estimator can be decomposed so as to isolate $\boldsymbol{\beta}_0$.

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{2SLS} &= \boldsymbol{\beta}_0 + \left(\frac{1}{N}\sum_{i=1}^{N}\widehat{\mathbf{x}}_i\widehat{\mathbf{x}}_i^{\mathrm{T}}\right)^{-1}\frac{1}{N}\sum_{i=1}^{N}\widehat{\mathbf{x}}_i\varepsilon_i \\ &= \boldsymbol{\beta}_0 + \left(\frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{X}\right)^{-1}\frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\boldsymbol{\varepsilon}\end{aligned}$$

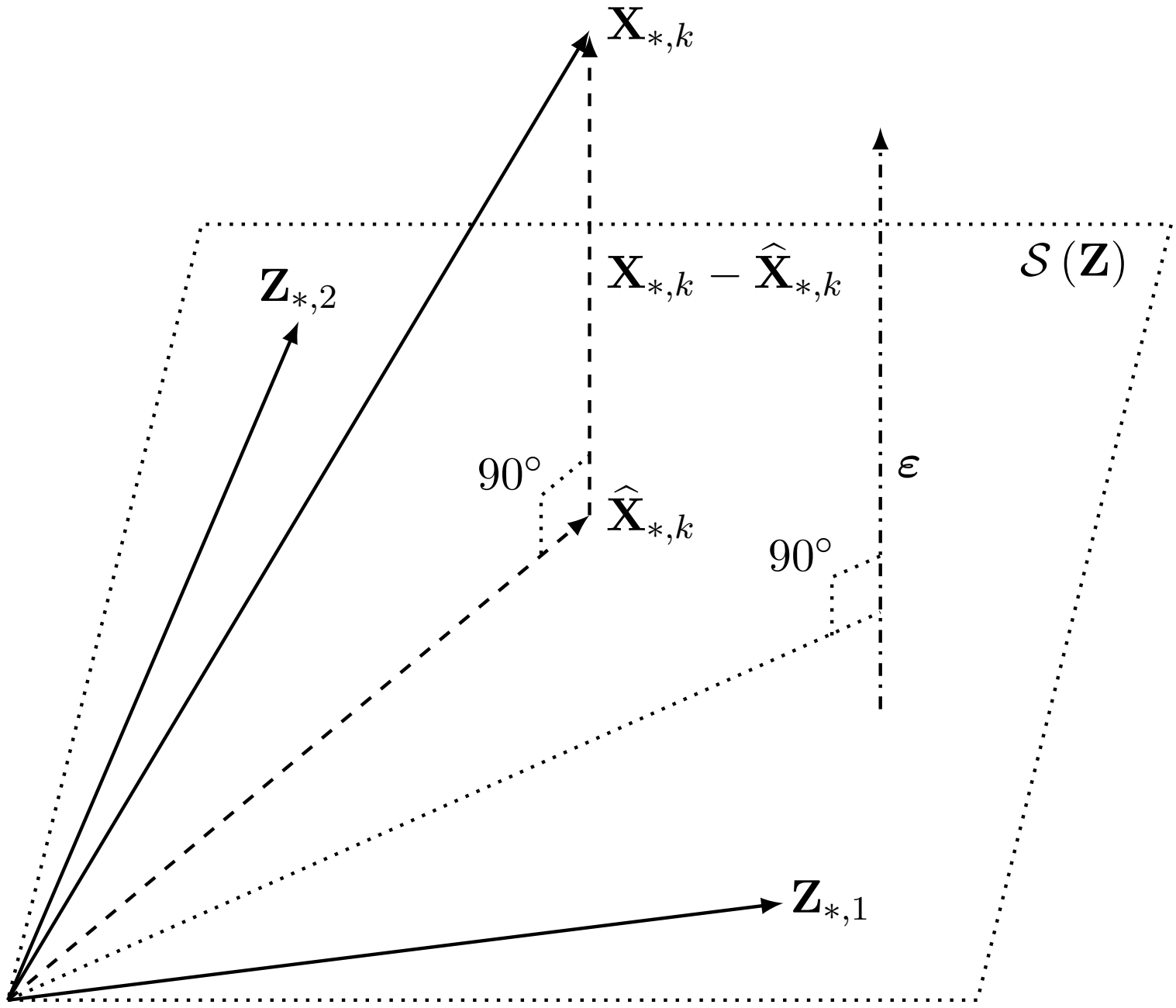
- Since $\mathbb{E}\left[\widehat{\boldsymbol{x}}_i\varepsilon_i\right] = \mathbf{0}$, it is:

$$\frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\boldsymbol{\varepsilon} = \frac{1}{N}\sum_{i=1}^{N}\widehat{\mathbf{x}}_i\varepsilon_i \xrightarrow{p} \mathbf{0}$$

  implying **consistency** of 2SLS estimator, i.e. $\widehat{\boldsymbol{\beta}}_{2SLS} \xrightarrow{p} \boldsymbol{\beta}_0$.

- This has a geometric interpretation: the first stage projects $\mathbf{x}_i$ onto $\mathcal{S}(\mathbf{Z})$, the space which is spanned by the exogenous variables $\mathbf{z}_i$; hence the "fitted" regressors **by construction** lie on $\mathcal{S}(\mathbf{Z})$ and so are orthogonal to $\boldsymbol{\varepsilon}$ (see the next figure).

Consistency and geometry of 2SLS (2/2)



Example: Mincer, Revisited again. (1/2)

- Return to the Mincer example. The IV estimator can there be obtained in two ways; both require the estimation of the following first stage model (note the squared experience).

$$S_i = \pi_0 + \pi_1 X_i + \pi_2 X_i^2 + \pi_3 Z_i + \eta_i$$

- The **reduced form** is identified if $\mathbb{E}\left[\alpha_i, \epsilon_i, \eta_i \middle| X_i, Z_i\right] = 0$.

$$\begin{aligned}\log W_i = \beta_0 + \beta_3\pi_0 + (\beta_1 + \beta_3\pi_1) X_i + (\beta_2 + \beta_3\pi_2) X_i^2 \\ + \beta_3\pi_3 Z_i + \alpha_i + \epsilon_i + \beta_3\eta_i \\ S_i = \pi_0 + \pi_1 X_i + \pi_2 X_i^2 + \pi_3 Z_i + \eta_i\end{aligned}$$

- By the Frisch-Waugh-Lovell theorem, a consistent estimate of β_3, that is numerically identical to the IV estimator, can be obtained as follows (the other estimates are analogous).

$$\widehat{\beta}_3 = \frac{\widehat{\beta_3\pi_3}}{\widehat{\pi}_3} = \frac{\mathbf{z}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}}\mathbf{y}}{\mathbf{z}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}}\mathbf{s}}$$

Example: Mincer, Revisited again. (2/2)

- Alternatively, one can compute the first stage projection:

$$\widehat{S}_i = \widehat{\pi}_0 + \widehat{\pi}_1 X_i + \widehat{\pi}_2 X_i^2 + \widehat{\pi}_3 Z_i$$

- . . . and estimate the following model by OLS; this delivers a just-identified IV-2SLS estimator.

$$\log W_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 \widehat{S}_i + \varepsilon_i$$

- The 2SLS approach also works in the **overidentified** case. Let there be two more IVs for S_i, named G_i and F_i.

- The first stage thus becomes:

$$S_i = \pi_0 + \pi_1 X_i + \pi_2 X_i^2 + \pi_3 Z_i + \pi_4 G_i + \pi_5 F_i + \eta_i$$

  and the "second" stage proceeds as above.

Asymptotics of 2SLS: assumptions (1/3)

- Next, the **asymptotic** properties of the more general 2SLS estimator are laid out.

- To this end, it is useful to formally state the **assumptions** on which these properties are grounded – they "substitute" assumptions 2-6 from Lecture 8.

Assumption 9

Independently but not identically distributed IVs. The observations in the sample $\{(y_i, \mathbf{x}_i, \mathbf{z}_i)\}_{i=1}^{N}$ are *independently*, but *not* necessarily *identically*, distributed (i.n.i.d.).

Assumption 10

Exogeneity of the Instruments. Conditional on the $J \geq K$ regressors $\boldsymbol{z}_i$, the error term ε_i has mean zero.

$$\mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{z}_i\right] = 0$$

Asymptotics of 2SLS: assumptions (2/3)

Assumption 11

Asymptotics of the Projected Regressors. The following probability limit exists, is finite, and has full rank.

$$\frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{X} \xrightarrow{p} \mathbf{P}_0$$

Assumption 12

Heteroscedastic, Independent Errors. The variance of the error term ε_i, conditional on the instruments $\boldsymbol{z}_i$, is unrestricted (*heteroscedasticity*), and the conditional covariance between two error terms of two different observations $i, j = 1, \ldots, N$ is zero.

$$\mathbb{E}\left[\varepsilon_i^2 \middle| \boldsymbol{z}_i\right] = \sigma^2(\boldsymbol{z}_i) \equiv \sigma_i^2$$
$$\mathbb{E}\left[\varepsilon_i \varepsilon_j \middle| \boldsymbol{z}_i, \boldsymbol{z}_j\right] = 0$$

Asymptotics of 2SLS: assumptions (3/3)

Assumption 13

Asymptotics of Projected Regressors interacted with the Errors. Given a diagonal matrix of squared errors:

$$\mathbf{E} \equiv \begin{bmatrix} \varepsilon_1^2 & 0 & \cdots & 0 \\ 0 & \varepsilon_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \varepsilon_N^2 \end{bmatrix}$$

the following probability limit exists, is finite, and has full rank.

$$\frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{E}\mathbf{P}_{\mathbf{Z}}\mathbf{X} \xrightarrow{p} \mathbf{\Psi}_0$$

In addition, conditions hold so that the following Central Limit Theorem result applies.

$$\frac{1}{\sqrt{N}}\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\boldsymbol{\varepsilon} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{\Psi}_0\right)$$

Asymptotics of 2SLS

Theorem 1

Large Sample properties of the 2SLS Estimator. *Under Assumptions 1, 3 and 9-13, the 2SLS estimator is consistent:*

$$\widehat{\boldsymbol{\beta}}_{2SLS} \xrightarrow{p} \boldsymbol{\beta}_0$$

and it is asymptotically normal, that is:

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_{2SLS} - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{P}_0^{-1}\boldsymbol{\Psi}_0\mathbf{P}_0^{-1}\right)$$

hence its asymptotic distribution, for a given N, is as follows.

$$\widehat{\boldsymbol{\beta}}_{2SLS} \overset{A}{\sim} \mathcal{N}\left(\boldsymbol{\beta}_0, \frac{1}{N}\mathbf{P}_0^{-1}\boldsymbol{\Psi}_0\mathbf{P}_0^{-1}\right)$$

Proof.

The proof is analogous to the one for the OLS case, as it exploits the estimator decomposition, the given assumptions, as well as Slutskij's Theorem and the Cramér-Wold device. □

Estimation of the 2SLS variance-covariance (1/2)

The asymptotic variance-covariance of 2SLS is estimated as:

$$\widehat{\mathbb{A}\text{var}}\left(\widehat{\boldsymbol{\beta}}_{2SLS}\right) = \frac{1}{N}\widehat{\mathbf{P}}_N^{-1}\widehat{\boldsymbol{\Psi}}_N\widehat{\mathbf{P}}_N^{-1}$$

where:

$$\widehat{\mathbf{P}}_N \equiv \frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{X}$$

$$\widehat{\boldsymbol{\Psi}}_N \equiv \frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\widehat{\mathbf{E}}_N\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{X}$$

with $\widehat{\mathbf{P}}_N \xrightarrow{p} \mathbf{P}_0$, $\widehat{\boldsymbol{\Psi}}_N \xrightarrow{p} \boldsymbol{\Psi}_0$; while $\widehat{\mathbf{E}}_N$ is the following matrix:

$$\widehat{\mathbf{E}}_N \equiv \begin{bmatrix} e_1^2 & 0 & \cdots & 0 \\ 0 & e_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e_N^2 \end{bmatrix}$$

where $e_i \equiv y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{2SLS}$ for $i = 1, \ldots, N$.

Estimation of the 2SLS variance-covariance (2/2)

With clustered observations (CCE), $\widehat{\mathbf{\Psi}}_{CCE} \xrightarrow{p} \mathbf{\Psi}_0$ is required:

$$\widehat{\mathbf{\Psi}}_{CCE} \equiv \frac{1}{N}\left[\sum_{c=1}^{C} \mathbf{X}_c^{\mathrm{T}}\mathbf{Z}_c\left(\mathbf{Z}_c^{\mathrm{T}}\mathbf{Z}_c\right)^{-1}\mathbf{Z}_c^{\mathrm{T}}\mathbf{e}_c\mathbf{e}_c^{\mathrm{T}}\mathbf{Z}_c\left(\mathbf{Z}_c^{\mathrm{T}}\mathbf{Z}_c\right)^{-1}\mathbf{Z}_c^{\mathrm{T}}\mathbf{X}_c\right]$$

where $\mathbf{e}_c \equiv \mathbf{y}_c - \mathbf{X}_c\widehat{\boldsymbol{\beta}}_{2SLS}$. HAC cases are treated analogously.

In the homoscedastic case, it is $\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\boldsymbol{Z}\right] = \sigma_0^2\mathbf{I}$ and thus:

$$\frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{E}\mathbf{P}_{\mathbf{Z}}\mathbf{X} \xrightarrow{p} \sigma_0^2 \cdot \lim_{N\to\infty}\frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\mathbf{X} = \sigma_0^2\mathbf{P}_0$$

hence, $\mathbf{\Psi}_0 = \sigma_0^2\mathbf{P}_0$ and:

$$\widehat{\boldsymbol{\beta}}_{2SLS} \overset{A}{\sim} \mathcal{N}\left(\boldsymbol{\beta}_0, \frac{\sigma_0^2}{N}\mathbf{P}_0^{-1}\right)$$

which, for $\mathbf{e} \equiv \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{2SLS}$, is (efficiently) estimated as follows.

$$\widehat{\mathbb{A}\mathrm{var}}\left(\widehat{\boldsymbol{\beta}}_{2SLS}\right) = \frac{\mathbf{e}^{\mathrm{T}}\mathbf{e}}{N}\left[\mathbf{X}^{\mathrm{T}}\mathbf{Z}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{X}\right]^{-1}$$

Example: 2SLS for structural endogeneity

- Recall the time-series example of structural endogeneity. It can be recast via iterated substitution as:

$$y_t = \beta_0 + \sum_{s=0}^{t-1} \beta_1^s \left(\mathbf{x}_{t-s}^{\mathrm{T}} \boldsymbol{\gamma}_0 + \mathbf{x}_{t-1-s}^{\mathrm{T}} \boldsymbol{\gamma}_1 \right) + \sum_{s=0}^{t-1} \sum_{z=0}^{t-s} \beta_1^s \rho^z \xi_{t-s-z}$$

  with $(\mathbf{x}_0, \xi_0)$ at $t = 0$.

- Hence **all valid lags** $\mathbf{x}_{t-s}$, for $s \geq 2$, can be combined into the instruments vector $\mathbf{z}_t$ in a 2SLS framework.

- Similarly, the spatial model can be rephrased as follows.

$$\mathbf{y} = \sum_{s=0}^{\infty} \beta_1^s \mathbf{W}^s \left(\beta_0 \boldsymbol{\iota} + \mathbf{X} \boldsymbol{\gamma}_0 + \mathbf{W} \mathbf{X} \boldsymbol{\gamma}_1 + \boldsymbol{\varepsilon} \right)$$

- Hence, **all linearly independent** columns of the matrices in $\{\mathbf{W}^s \mathbf{X}\}_{s=2}^{\infty}$ can enter the instruments matrix $\mathbf{Z}$.

Instrumental variables in practice: overview

Thus far, the discussion about the IV/2SLS estimator has been theoretical, but there are practical considerations that must be taken take into account. These are reviewed next.

Here is an overview of the topics to be discussed.

- **Control function approaches**, that is methods that are complementary to IV/2SLS estimators for exploiting IVs.
- Statistical **tests for endogeneity** (or lack thereof).
- The issue of IV/2SLS **small sample bias**.
- The problem of **weak instruments**.

Some of these issues, especially control function approaches and tests for endogeneity, are intertwined.

Control function methods: a summary

Control function methods use IVs to "treat" endogeneity in the **error term** itself. In a linear model, they work as follows.

1. In a first stage, like in 2SLS, K_2 regressions are run – one for each endogenous variables; **residuals are calculated** consequently for $k = 1, \ldots, K_2$.

$$\widehat{\eta}_{ki} \equiv x_{ki} - \mathbf{z}_i^{\mathrm{T}} \widehat{\boldsymbol{\pi}}_{kOLS} = x_{ki} - \widehat{x}_{ki}$$

2. In a second stage an OLS regression of the structural form **augmented with the estimated residuals** is run:

$$y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \widehat{\boldsymbol{\eta}}_i^{\mathrm{T}} \boldsymbol{\rho}_0 + \varsigma_i$$

where $\widehat{\boldsymbol{\eta}}_i = \begin{bmatrix} \widehat{\eta}_{1i} & \widehat{\eta}_{2i} & \ldots & \widehat{\eta}_{K_2 i} \end{bmatrix}^{\mathrm{T}}$ while $\boldsymbol{\rho}_0$ collects the K_2 parameters associated with each set of residuals, while ς_i is some new error term.

Consistency of control function methods (1/3)

The OLS estimates are actually consistent for both $\boldsymbol{\beta}_0$ and $\boldsymbol{\rho}_0$. A semi-formal argument is provided next.

- Consider the k-th first stage for $k = 1, \ldots, K_2$; note that:

$$\begin{aligned}\mathbb{E}\left[\eta_{ki}\varepsilon_i\right] &= \mathbb{E}\left[\left(X_{ki} - \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\pi}_0\right)\varepsilon_i\right] \\ &= \underbrace{\mathbb{E}\left[X_{ki}\varepsilon_i\right]}_{\neq 0} - \underbrace{\mathbb{E}\left[\boldsymbol{z}_i\varepsilon_i\right]^{\mathrm{T}}}_{=0}\boldsymbol{\pi}_0 \neq 0\end{aligned}$$

  thus, η_{ki} "contains information" about X_{ki}'s endogeneity.

- Specify a **statistical model for the error term**, which is also called a **control function**. Let it be linear.

$$\varepsilon_i = \boldsymbol{\eta}_i^{\mathrm{T}}\boldsymbol{\rho}_0 + \xi_i$$

- Here, $\boldsymbol{\eta}_i = \begin{bmatrix}\eta_{1i} & \eta_{2i} & \ldots & \eta_{J_2 i}\end{bmatrix}^{\mathrm{T}}$ and ξ_i is a residual error.

Consistency of control function methods (2/3)

- The **population linear projection** ε_i onto $\boldsymbol{\eta}_i$ is:

$$\boldsymbol{\rho}_0 \equiv \mathbb{E}\left[\boldsymbol{\eta}_i \boldsymbol{\eta}_i^{\mathrm{T}}\right]^{-1} \mathbb{E}\left[\boldsymbol{\eta}_i \varepsilon_i\right]$$

- . . . and $\boldsymbol{\eta}_i^{\mathrm{T}} \boldsymbol{\rho}_0$ is the **extent of endogeneity** in the model.

- Note that, since the IVs are exogenous:

$$\mathbb{E}\left[\boldsymbol{z}_i \xi_i\right] = \underbrace{\mathbb{E}\left[\boldsymbol{z}_i \varepsilon_i\right]}_{=\mathbf{0}} - \underbrace{\mathbb{E}\left[\boldsymbol{z}_i \boldsymbol{\eta}_i^{\mathrm{T}}\right]}_{=\mathbf{0}} \boldsymbol{\rho}_0 = \mathbf{0}$$

- . . . and by definition of linear projection, for $k = 1, \ldots, J_2$:

$$\mathbb{E}\left[\eta_{ki} \xi_i\right] = 0$$

- . . . therefore the following implication must hold.

$$\mathbb{E}\left[X_{ki} \xi_i\right] = \underbrace{\mathbb{E}\left[\boldsymbol{z}_i \xi_i\right]^{\mathrm{T}}}_{=\mathbf{0}} \boldsymbol{\pi}_0 + \underbrace{\mathbb{E}\left[\eta_{ki} \xi_i\right]}_{=0} = 0$$

Consistency of control function methods (3/3)

- By reworking the original structural equation one gets:

$$y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \boldsymbol{\eta}_i^{\mathrm{T}} \boldsymbol{\rho}_0 + \xi_i$$

  which could be estimated by OLS if $\boldsymbol{\eta}_i$ were observable.

- While $\boldsymbol{\eta}_i$ can hardly be observed it can be estimated in the first stage, hence everything is reconciled with:

$$\varsigma_i \equiv \left(\boldsymbol{\eta}_i - \widehat{\boldsymbol{\eta}}_i\right)^{\mathrm{T}} \boldsymbol{\rho}_0 + \xi_i = \mathbf{z}_i^{\mathrm{T}} \cdot \sum_{k=0}^{K_0} \left(\boldsymbol{\pi}_{k0} - \widehat{\boldsymbol{\pi}}_{k,OLS}\right) \rho_{k0} + \xi_i$$

  and moreover one can show that the first component of the above expression is mean-independent of $(\boldsymbol{x}_i, \widehat{\boldsymbol{\eta}}_i)$.

- The consistency result sought after thus holds since:

$$\mathbb{E}\left[\boldsymbol{x}_i^{\mathrm{T}} \varsigma_i\right] = \mathbf{0} \quad \text{and} \quad \mathbb{E}\left[\widehat{\boldsymbol{\eta}}_i^{\mathrm{T}} \varsigma_i\right] = \mathbf{0}$$

  themselves hold.

Control functions: additional considerations

- Control function estimates of $\boldsymbol{\beta}_0$ are numerically equivalent to IV-2SLS. The proper proof is based on a variation of the Frisch-Waugh-Lovell Theorem and the projections' algebra.

- In practice, control function approaches are seldom used in linear models. They are impractical if endogenous variables enter the structural form non-linearly.

- More importantly, they deliver **larger standard errors** in comparison with the competing IV-2SLS approach.

- They are often used in **tests for endogeneity** (see next).

- They are prevalently used to incorporate IVs in endogenous **non-linear** models where they might be more practical, as the competing GMM approaches (see Lecture 12) are often more cumbersome.

Tests for endogeneity (1/2)

How can one "test for endogeneity"?

$$\mathrm{H}_0 : \mathbb{E}\left[\boldsymbol{x}_i^{\mathrm{T}} \varepsilon_i\right] = \mathbf{0}$$

The above null hypothesis implies the following.

$$\mathrm{H}_0 : \operatorname{plim} \widehat{\boldsymbol{\beta}}_{OLS} - \operatorname{plim} \widehat{\boldsymbol{\beta}}_{2SLS} = \mathbf{0}$$

Thus, a natural test statistic is:

$$\widetilde{\mathcal{H}}_{\mathrm{H}_0} = \left[\widehat{\boldsymbol{\beta}}_{OLS} - \widehat{\boldsymbol{\beta}}_{2SLS}\right]^{\mathrm{T}} \widehat{\mathbb{A}\mathrm{var}} \left[\widehat{\boldsymbol{\beta}}_{OLS} - \widehat{\boldsymbol{\beta}}_{2SLS}\right] \left[\widehat{\boldsymbol{\beta}}_{OLS} - \widehat{\boldsymbol{\beta}}_{2SLS}\right]$$

where $\widetilde{\mathcal{H}}_{\mathrm{H}_0} \xrightarrow{p} \chi^2_K$; $\widetilde{\mathcal{H}}_{\mathrm{H}_0}$ should be close to zero under the null.

A problem with this test is that the following quantity is hard to compute, as the covariance is generally unknown.

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\widehat{\boldsymbol{\beta}}_{OLS} - \widehat{\boldsymbol{\beta}}_{2SLS}\right] = \\ = \mathbb{V}\mathrm{ar}\left[\widehat{\boldsymbol{\beta}}_{OLS}\right] + \mathbb{V}\mathrm{ar}\left[\widehat{\boldsymbol{\beta}}_{2SLS}\right] - 2\,\mathbb{C}\mathrm{ov}\left[\widehat{\boldsymbol{\beta}}_{OLS}, \widehat{\boldsymbol{\beta}}_{2SLS}\right]\end{aligned}$$

Tests for endogeneity (2/2)

Hausman (1978) showed that since OLS is efficient under i.i.d. errors (per the Gauss-Markov theorem), it is:

$$\mathbb{C}\text{ov}\left[\widehat{\boldsymbol{\beta}}_{OLS}, \widehat{\boldsymbol{\beta}}_{2SLS}\right] = \mathbb{V}\text{ar}\left[\widehat{\boldsymbol{\beta}}_{2SLS}\right]$$

which allows to formulate the **Hausman test statistic**:

$$\mathcal{H}_{\mathrm{H}_0} = \left[\widehat{\boldsymbol{\beta}}_{OLS} - \widehat{\boldsymbol{\beta}}_{2SLS}\right]^{\mathrm{T}} \left\{\widehat{\mathbb{A}\text{var}}\left[\widehat{\boldsymbol{\beta}}_{OLS}\right] - \widehat{\mathbb{A}\text{var}}\left[\widehat{\boldsymbol{\beta}}_{2SLS}\right]\right\} \left[\widehat{\boldsymbol{\beta}}_{OLS} - \widehat{\boldsymbol{\beta}}_{2SLS}\right]$$

with $\mathcal{H} \xrightarrow{p} \chi^2_K$ asymptotically. $\mathcal{H}_{\mathrm{H}_0}$ is easy to compute.

Beyond the i.i.d. scenario, more general tests are based on Wald statistics about $\widehat{\boldsymbol{\rho}}_{OLS}$ from the control function second stage, as

$$\mathrm{H}_0 : \boldsymbol{\rho}_0 = \mathbf{0}$$

implies $\mathbb{E}\left[\boldsymbol{\eta}_i^{\mathrm{T}} \varepsilon_i\right] = \mathbb{E}\left[\boldsymbol{x}_i^{\mathrm{T}} \varepsilon_i\right] = \mathbf{0}$ and thus "no endogeneity."

Small sample bias of IV-2SLS

While the IV-2SLS estimator is consistent, it is biased in small samples (unlike OLS). In fact:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\boldsymbol{\beta}}_{2SLS}\right] &= \boldsymbol{\beta}_0 + \mathbb{E}\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{\varepsilon}\right] \\
&= \boldsymbol{\beta}_0 + \mathbb{E}_{\boldsymbol{X},\boldsymbol{Z}}\left[\mathbb{E}\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{\varepsilon}\middle|\,\boldsymbol{X},\boldsymbol{Z}\right]\right] \\
&= \boldsymbol{\beta}_0 + \mathbb{E}_{\boldsymbol{X},\boldsymbol{Z}}\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}_{\boldsymbol{Z}}\cdot\mathbb{E}\left[\boldsymbol{\varepsilon}|\,\boldsymbol{X},\boldsymbol{Z}\right]\right]
\end{aligned}$$

and it is impossible to simplify the expression further, since due to endogeneity, it is $\mathbb{E}\left[\boldsymbol{\varepsilon}|\,\boldsymbol{X},\boldsymbol{Z}\right] \neq \mathbf{0}$.

The bias goes in the direction of the inconsistent OLS estimator (a fact that shall not be overlooked in actual practice): IV-2SLS requires large samples.

Weak instruments (1/4)

It has been observed that the limiting variance of IV-2SLS is in general inversely proportional to the covariance between $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$. What if it is small, and there are **weak instruments**?

1. First, IV-2SLS might deliver worse predictions than OLS in the MSE sense: a higher variance may offset a lower bias.

$$\operatorname{plim} \mathrm{MSE}_{IV-2SLS} = \\ = \underbrace{\left(\operatorname{plim} \widehat{\boldsymbol{\beta}}_{IV-2SLS} - \boldsymbol{\beta}_0\right)^2}_{= \textit{squared asymptotic bias}} + \underbrace{\mathbb{A}\mathrm{var}\left(\widehat{\boldsymbol{\beta}}_{IV-2SLS}\right)}_{= \textit{asymptotic variance}}$$

2. If the instruments are weak and **slightly endogenous**, the "cure" to endogeneity may even be worse then the disease! An even small bias may be amplified by the "low precision" that estimation based on weak instruments entails.

Weak instruments (2/4)

The second consideration especially is best illustrated through a simple triangular model with variables Y_i, X_i and Z_i. There:

$$\begin{aligned}\frac{\operatorname{plim}\beta_{1,IV}-\beta_1}{\operatorname{plim}\beta_{1,OLS}-\beta_1} &= \frac{\mathbb{C}\mathrm{ov}\left(Z_i,\varepsilon_i\right)}{\mathbb{C}\mathrm{ov}\left(Z_i,X_i\right)}\cdot\frac{\mathbb{V}\mathrm{ar}\left(X_i\right)}{\mathbb{C}\mathrm{ov}\left(X_i,\varepsilon_i\right)}\\ &= \frac{\mathbb{C}\mathrm{orr}\left(Z_i,\varepsilon_i\right)}{\mathbb{C}\mathrm{orr}\left(X_i,\varepsilon_i\right)}\cdot\frac{1}{\mathbb{C}\mathrm{orr}\left(Z_i,X_i\right)}\end{aligned}$$

which is derived by elaborating the probability limit of the two estimators under the assumptions that $\mathbb{C}\mathrm{ov}\left(Z_i,\varepsilon_i\right)\neq 0$, as well as $\mathbb{C}\mathrm{ov}\left(X_i,\varepsilon_i\right)\neq 0$.

Thus even if the instrument is, while not completely exogenous, somewhat "less endogenous" – that is $\mathbb{C}\mathrm{ov}\left(Z_i,\varepsilon_i\right) < \mathbb{C}\mathrm{ov}\left(X_i,\varepsilon_i\right)$ – a weak instrument might actually exacerbate the endogeneity problem. This is dangerous in applied research.

Weak instruments (3/4)

This can be generalized to overidentified triangular models:

$$y_i = \mathbf{x}_{i1}^{\mathrm{T}} \boldsymbol{\beta}_{0\setminus K} + \delta_0 s_i + \varepsilon_i$$
$$s_i = \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\pi}_0 + \eta_i$$

where errors are assumed i.i.d. for simplicity. As one can show:

$$\frac{\operatorname{plim} \delta_{IV} - \delta_0}{\operatorname{plim} \delta_{OLS} - \delta_0} = \frac{\mathbb{C}\mathrm{orr}\left(\widehat{S}_i, \varepsilon_i\right)}{\mathbb{C}\mathrm{orr}\left(S_i, \varepsilon_i\right)} \cdot \frac{1}{\operatorname{plim} \mathcal{R}^2_{s,\mathbf{z}|\mathbf{x}}}$$

where $\mathcal{R}^2_{s,\mathbf{z}|\mathbf{x}}$ is the **partialed out R-squared coefficient**.

$$\mathcal{R}^2_{s,\mathbf{z}|\mathbf{x}} = \frac{\mathbf{s}^{\mathrm{T}} \mathbf{P}_{\mathbf{Z}} \mathbf{M}_{\mathbf{X}1} \mathbf{P}_{\mathbf{Z}} \mathbf{s}}{\mathbf{s}^{\mathrm{T}} \mathbf{M}_{\mathbf{X}1} \mathbf{s}}$$

By the Frisch-Waugh-Lovell Theorem, this is the R^2 coefficient obtained from a regression of s_i on $\mathbf{z}_i$ after **partialing out** the exogenous regressors $\mathbf{x}_{i1}$, as follows (see Lecture 7).

$$\mathbf{M}_{\mathbf{X}1} \mathbf{s} = \mathbf{M}_{\mathbf{X}1} \mathbf{Z} \boldsymbol{\pi}_0 + \mathbf{M}_{\mathbf{X}1} \boldsymbol{\eta}$$

Weak instruments (4/4)

How to cope with the risk of weak instruments? Some practical considerations apply.

1. It is always useful to **test** the **statistical power** of the IVs via estimates of the first stage models. Some rules of thumb apply: t-statistics of the exogenous instruments higher than 3, or model-wide F-statistics higher than 10, are considered signs that the instruments are "satisfactorily strong."

2. Even if *in theory* 2SLS using many IVs is more likely to hit the efficiency bound, *in practice* this also increases the risk of using weak instruments. It is often more desirable to use fewer, but safer IVs.

The best practices evolve with time: it is advisable to follow the latest theoretical developments for better guidance on the use of IV-2SLS and more accurate resulting empirical analyses.

2SLS estimation of simultaneous equations

- 2SLS was originally devised to address simultaneity issues in Simultaneous Equations Models (SEMs).

- Rewrite the p-th equation of a SEM $\mathbf{\Gamma y}_i = \mathbf{\Phi z}_i + \boldsymbol{\varepsilon}_i$ as:

$$y_{pi} = \mathbf{x}_{pi}^{\mathrm{T}} \boldsymbol{\beta}_{p0} + \varepsilon_{pi}$$

  where y_{pi} is the realization of the p-th endogenous variable, while $\mathbf{x}_{pi}$ collects the realizations of all the variables, **both exogenous** and **endogenous**, that are **not** excluded from the structural form (their associated parameters are $\boldsymbol{\beta}_{p0}$).

- With $K_2 \leq P - 1$ endogenous variables in $\mathbf{x}_{pi}$, there are K_2 **first stage** equations from the **reduced form** of the SEM.

$$x_{ki} = \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\pi}_{k0} + \eta_{ki}$$

- If the p-th equation is just-identified or overidentified it can be easily estimated via 2SLS.

Joint estimation of simultaneous equations

- Although 2SLS is well suited to estimation of SEMs, it may not be the most efficient approach.

- If the SEM's P error terms $\boldsymbol{\varepsilon}_i$ are **stochastically related**, a **joint estimation approach** that takes this into account can deliver more efficient estimates.

- The **Three-Stages Least Squares** (3SLS) estimator was originally devised precisely as an extension of 2SLS for the sake of performing joint estimation of a SEM.

- In **Seemingly Unrelated Regressions** (SURs) – special cases of SEMs with $\mathbf{\Gamma} = \mathbf{I}$ – a similar issue occurs.

- Following a brief motivating example that makes a case for joint estimation, the 3SLS estimator is illustrated.

Example: Household labor supply

Consider the two-equations SEM:

$$H_{hi} = \alpha_0 + \alpha_1 H_{wi} + \alpha_2 S_{hi} + \alpha_3 S_{wi} + \alpha_4 \log W_{hi} + \alpha_5 \log W_{wi} + \varepsilon_{hi}$$
$$H_{wi} = \beta_0 + \beta_1 H_{hi} + \beta_2 S_{hi} + \beta_3 S_{wi} + \beta_4 \log W_{hi} + \beta_5 \log W_{wi} + \varepsilon_{wi}$$

where:

- subscript i denotes a household (the unit of observation),
- w denotes variables about to the wife, h about the husband,
- and for $s \in \{h, w\}$, H_{si}, S_{si}, $\log W_{hi}$ denote "hours worked," education and the logarithm of the wage respectively.

This model illustrates co-dependence in **labor supply** decisions in a household, and is identified with at least one restriction per equation (e.g. $\alpha_3 = \beta_2 = 0$, or $\alpha_1 = \beta_1 = 0$ leading to a SUR).

Presumably, however, the unobservable factors affecting the two spouses are correlated, calling for **joint estimation**.

$$\mathbb{C}\text{ov}\left(\varepsilon_{hi}, \varepsilon_{wi}\right) \neq 0$$

Three-Stages Least Squares (1/6)

The 3SLS estimator is best written in compact matrix notation. Express any SEM whose equations are at least exactly identified as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}$$

or:

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_P \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}_P \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{10} \\ \boldsymbol{\beta}_{20} \\ \vdots \\ \boldsymbol{\beta}_{P0} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_P \end{bmatrix}$$

where $\mathbf{y}_p$ is the N-dimensional vector obtained from stacking all the observations y_{pi} for $i = 1, \ldots, N$; $\boldsymbol{\varepsilon}_p$ is constructed similarly.

Analogously, matrix $\mathbf{X}_p$ results from vertically stacking vectors $\mathbf{x}_{pi}^{\mathrm{T}}$ for $i = 1, \ldots, N$; thus the above can be rephrased as follows.

$$\mathbf{y}_p = \mathbf{X}\boldsymbol{\beta}_{p0} + \boldsymbol{\varepsilon}_p$$

Three-Stages Least Squares (2/6)

Furthermore, consider the stacked **instruments** matrix $\mathbf{Z}$ as in 2SLS; the associated projection matrix $\mathbf{P_Z}$, and construct the P equation-specific matrices of projected regressors as:

$$\widehat{\mathbf{X}}_p = \mathbf{P_Z}\mathbf{X}_p$$

while

$$\widehat{\mathbf{X}} \equiv \begin{bmatrix} \widehat{\mathbf{X}}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{X}}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \widehat{\mathbf{X}}_P \end{bmatrix}$$

results from diagonally stacking block-by-block all $\widehat{\mathbf{X}}_p$ matrices. Given this notation, the 2SLS estimator for **all** P equations can be written more compactly as follows.

$$\widehat{\boldsymbol{\beta}}_{2SLS} = \left(\widehat{\mathbf{X}}^{\mathrm{T}}\widehat{\mathbf{X}}\right)^{-1}\widehat{\mathbf{X}}^{\mathrm{T}}\mathbf{y}$$

Three-Stages Least Squares (3/6)

Assume for simplicity that the error terms $\boldsymbol{\varepsilon}$ are i.i.d. **within equations**, but are correlated **across equations**:

$$\mathbb{E}\left[\boldsymbol{\varepsilon} \middle| \boldsymbol{Z}\right] = 0$$
$$\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}} \middle| \boldsymbol{Z}\right] = \boldsymbol{\Sigma} \otimes \mathbf{I}$$

where $\boldsymbol{\Sigma}$ is the symmetric $P \times P$ matrix that contains:

- the equation-specific variance of each equation p along the diagonal; and
- the cross-equation covariance terms outside the diagonal.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1P} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{P1} & \sigma_{P2} & \cdots & \sigma_{PP} \end{bmatrix}$$

Three-Stages Least Squares (4/6)

Therefore, by the definition of Kronecker product:

$$\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\boldsymbol{Z}\right] = \boldsymbol{\Sigma}\otimes\mathbf{I} = \begin{bmatrix} \sigma_{11}\mathbf{I} & \sigma_{12}\mathbf{I} & \cdots & \sigma_{1P}\mathbf{I} \\ \sigma_{21}\mathbf{I} & \sigma_{22}\mathbf{I} & \cdots & \sigma_{2P}\mathbf{I} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{P1}\mathbf{I} & \sigma_{P2}\mathbf{I} & \cdots & \sigma_{PP}\mathbf{I} \end{bmatrix}$$

hence, in addition to the equation-specific variance terms σ_{pp} on the diagonal, for any observation i the conditional covariance of the two shocks ε_{pi} and ε_{qi} – for the p-th and the q-th equations respectively – is equal to σ_{pq}.

All other cross-equation covariance terms (that is, those relative to two different observations) are assumed to be equal to zero.

This exhausts all the elements that are necessary to "construct" the 3SLS estimator, whose "stages" are outlined next.

Three-Stages Least Squares (5/6)

1. Project $\mathbf{x}_{pi}$ ($\mathbf{X}_p$) onto $\mathbf{z}_i$ ($\mathbf{Z}$) equation-by-equation, for all $p = 1, \ldots, P$;

2. Compute the 2SLS estimator $\widehat{\boldsymbol{\beta}}_{2SLS}$, as well as $P\,(P-1)$ estimates for the parameters contained in matrix $\boldsymbol{\Sigma}$:

$$\widehat{\sigma}_{pq} = \frac{\left(\mathbf{y}_p - \mathbf{X}_p \widehat{\boldsymbol{\beta}}_{p2SLS}\right)^{\mathrm{T}} \left(\mathbf{y}_q - \mathbf{X}_q \widehat{\boldsymbol{\beta}}_{q2SLS}\right)}{N}$$

   for $p, q = 1, \ldots, P$, delivering an estimate $\widehat{\boldsymbol{\Sigma}}_N$ of matrix $\boldsymbol{\Sigma}$;

3. Finally, compute the 3SLS estimator as:

$$\widehat{\boldsymbol{\beta}}_{3SLS} = \left[\widehat{\mathbf{X}}^{\mathrm{T}} \left(\widehat{\boldsymbol{\Sigma}}_N^{-1} \otimes \mathbf{I}\right) \widehat{\mathbf{X}}\right]^{-1} \widehat{\mathbf{X}}^{\mathrm{T}} \left(\widehat{\boldsymbol{\Sigma}}_N^{-1} \otimes \mathbf{I}\right) \mathbf{y}$$

   and its asymptotic variance as follows.

$$\widehat{\mathbb{A}\mathrm{var}}\left(\widehat{\boldsymbol{\beta}}_{3SLS}\right) = \left[\widehat{\mathbf{X}}^{\mathrm{T}} \left(\widehat{\boldsymbol{\Sigma}}_N^{-1} \otimes \mathbf{I}\right) \widehat{\mathbf{X}}\right]^{-1}$$

Three-Stages Least Squares (6/6)

Some additional considerations about the 3SLS are due.

- The 3SLS estimator can be seen as the GLS generalization of the 2SLS estimator of an identified SEM.
- In Lecture 12, 2SLS and 3SLS are revisited as special cases of GMM.
- Although 3SLS might provide efficiency improvements over 2SLS, it is not the most efficient estimator of a SEM, as the theory of GMM offers additional avenues for improvement.
- In fully parametric environments more methods to estimate SEMs: **LIML** (Limited Information Maximum Likelihood), and **FIML** (Full Information Maximum Likelihood) exist.