

ECONOMETRIA I

ENDOGENEIDADE

Victor Oliveira

Núcleo de Economia Internacional e Desenvolvimento Econômico

PPGDE – 2023

- 1 Endogeneidade
- 2 Simultaneidade
- 3 Variável Omitida
- 4 Método de Variáveis Instrumentais

Endogeneidade

- Dizemos que existe **endogeneidade** no modelo linear

$$y = \mathbf{x}'\boldsymbol{\beta} + e \quad (1)$$

se $\boldsymbol{\beta}$ é um parâmetro de interesse e

$$\mathbb{E}(\mathbf{x}e) \neq 0 \quad (2)$$

- Para diferenciar a eq. (1) da regressão e modelos de projeção, chamaremos a eq. (1) de **equação estrutural** e o $\boldsymbol{\beta}$ de **parâmetro estrutural**.
- Quando $\mathbb{E}(\mathbf{x}e) \neq 0$ dizemos que \mathbf{x} é **endógeno** para $\boldsymbol{\beta}$.

- De fato, podemos definir um coeficiente de projeção linear $\beta^* = \mathbb{E}(\mathbf{x}'\mathbf{x})^{-1}\mathbb{E}(\mathbf{x}\mathbf{y})$ e uma equação de projeção linear

$$y = \mathbf{x}'\beta^* + e^* \quad (3)$$

$$\mathbb{E}(\mathbf{x}e^*) = 0 \quad (4)$$

- Contudo, sob endogeneidade dada pela eq. (2), o coeficiente de projeção β^* não se iguala ao parâmetro estrutural. Como vemos isso?

$$\begin{aligned} \beta^* &= (\mathbb{E}(\mathbf{x}'\mathbf{x}))^{-1} \mathbb{E}(\mathbf{x}\mathbf{y}) \\ &= (\mathbb{E}(\mathbf{x}'\mathbf{x}))^{-1} \mathbb{E}(\mathbf{x}(\mathbf{x}'\beta + e)) \\ &= \beta + (\mathbb{E}(\mathbf{x}'\mathbf{x}))^{-1} \mathbb{E}(\mathbf{x}e) \neq \beta \end{aligned} \quad (5)$$

dados que $\mathbb{E}(\mathbf{x}e) \neq 0$.

- Assim endogeneidade requer que o coeficiente seja definido de maneira diferente da projeção linear. Chamaremos essa definição de **estrutural**.
- **Implicações para o estimador:** endogeneidade implica que o estimador de mínimos quadrados é inconsistente para o parâmetro estrutural.
- Na verdade, sob uma amostra *i.i.d.*, mínimos quadrados é consistente para o coeficiente de projeção (β^*) e é inconsistente para β . Veja que

$$\hat{\beta} \xrightarrow{p} (\mathbb{E}(\mathbf{x}'\mathbf{x}))^{-1} \mathbb{E}(\mathbf{x}\mathbf{y}) = \beta^* \neq \beta \quad (6)$$

- A inconsistência dos mínimos quadrados é tipicamente chamada de **viés de endogeneidade** ou **viés de estimação** por conta da endogeneidade.
- Como β é um parâmetro estrutural e é um parâmetro de interesse, a endogeneidade requer o uso de métodos de estimação alternativos.
- **Principais causas da inconsistência do estimador de MQO:**
 - Erro de medida clássico
 - Variável omitida
 - Simultaneidade

Simultaneidade

- Considere que você tenha a seguinte equação:

$$y = \beta_0 + \beta_1 x + \gamma' \mathbf{v} + u \quad (7)$$

em que y é a incidência de AIDS por país (em %), x é a porcentagem de jovens que usam preservativos nas relações sexuais de “alto risco”, e \mathbf{v} é um vetor que inclui outras variáveis relevantes para explicar y , tal que $\text{cov}(\mathbf{v}, u) = 0$.

- Não seria razoável esperar que o “**modelo estrutural**” dado pela eq. (7) que relaciona as variáveis acima contivesse uma segunda equação,

$$x = \alpha_0 + \alpha_1 y + \delta' \mathbf{w} + e. \quad (8)$$

- Ou seja, que x também dependesse de y ?
- Suponha que estejamos interessados em estimar a eq. (7) que é mais interessante do ponto de vista de formulação de políticas públicas. Será que a estimação por mínimos quadrados é uma boa alternativa?
- A resposta é, em geral, **NÃO!**
- Como iremos ver, a condição $\text{cov}(x, u) = 0$ é violada. E, portanto, o estimador de mínimos quadrados é **inconsistente**.

- O fato de que x e u devem ser correlacionados na eq. (7) pode ser verificado facilmente. Observe que:
 - ① Quando u varia, y varia na mesma direção, pela eq. (7);
 - ② Quando y varia, x também varia na mesma direção, pela eq. (8);
 - ③ Logo, há correlação entre u e x : **quando u varia, x também varia!**
- Voltando ao exemplo: digamos que certo país tenha um u “alto” em decorrência de algum fator puramente aleatório (por exemplo, menor aversão ao risco), o que implica maior incidência de AIDS, *ceteris paribus*.

- Isso significa que mais jovens usarão preservativos para se proteger, pois a maior incidência de AIDS torna o sexo sem proteção mais arriscado.
- Logo, há uma correlação entre os fatores em u e a porcentagem de jovens que usam preservativos.
- Em termos formais, temos um sistema de duas equações e duas incógnitas y e x – **modelo estrutural**.

$$y = \beta_0 + \beta_1 x + \gamma' \mathbf{v} + u \quad (9)$$

$$x = \alpha_0 + \alpha_1 y + \delta' \mathbf{w} + e. \quad (10)$$

- Como resolvemos o sistema de y e x ?

- Resolvendo o sistema para y e x em função das variáveis exógenas v e w e dos distúrbios, obtemos a “**forma reduzida**”:

$$y = \frac{1}{1 - \alpha_1\beta_1} (\beta_0 + \beta_1\alpha_0 + \beta_1\delta'w + \gamma'v + \beta_1e + u) \quad (11)$$

$$x = \frac{1}{1 - \alpha_1\beta_1} (\alpha_0 + \alpha_1\beta_0 + \beta_1\gamma'v + \delta'w + \alpha_1u + e). \quad (12)$$

- Para que o estimador de mínimos quadrados seja consistente, é necessário que $\text{cov}(u, x) = 0$.

- Ou seja, a covariância entre u e cada termo que compõe x (na forma reduzida) deve ser nula. Por hipótese supõe-se que:

$$\text{cov}(\mathbf{w}, u) = \text{cov}(\mathbf{v}, u) = \text{cov}(e, u) = 0 \quad (13)$$

- Com essa hipótese, anula-se a maior parte dos termos. **Mas a forma reduzida do modelo mostra explicitamente que x também depende de u .**
- Logo é evidente que, em geral, há uma correlação entre x e u :

$$\text{cov}(x, u) = \mathbb{E}(xu) = \frac{\alpha_1 \sigma_u^2}{1 - \alpha_1 \beta_1} \neq 0 \quad (14)$$

- Portanto, o estimador de mínimos quadrados aplicado é **viesado e inconsistente!**
- Esse tipo de viés do estimador de mínimos quadrados é chamado de “**viés de equações simultâneas**” ou simplesmente “**viés de simultaneidade**”.
- Em geral, não é possível saber a **direção do viés**.

Exemplo

- Suponha que o modelo seja:

$$y = \beta_0 + \beta_1 x + u \quad (15)$$

$$x = \alpha_0 + \alpha_1 y + \delta' \mathbf{w} + e. \quad (16)$$

- Novamente, teremos:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) (\beta_0 + \beta_1 x_i + u_i)}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\begin{aligned}
\hat{\beta}_1 &= \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
\hat{\beta}_1 &\xrightarrow{p} \beta_1 + \frac{\text{cov}(x, u)}{\text{var}(x)}, \quad \text{quando } n \rightarrow \infty \\
\hat{\beta}_1 &\xrightarrow{p} \beta_1 + \underbrace{\frac{\text{cov}(x, u)}{\text{var}(x)}}_{\neq 0} \\
\hat{\beta}_1 &\xrightarrow{p} \beta_1 + \underbrace{\frac{\alpha_1 \sigma_u^2}{1 - \alpha_1 \beta_1}}_{\neq 0}
\end{aligned} \tag{17}$$

Exemplos

- Outros exemplos:

- ① Criminalidade *versus* número de policiais em determinada região.
- ② Horas trabalhadas *versus* salário médio em determinado setor da indústria (oferta e demanda).
- ③ Consumo de bebidas alcoólicas *versus* desempenho do aluno.
- ④ Abertura comercial *versus* crescimento econômico.
- ⑤ Democracia *versus* crescimento econômico.
- ⑥ Corrupção *versus* crescimento econômico.
- ⑦ Função de produção: os insumos capital e trabalho dependem de fatores não observáveis e esses, por sua vez, influenciam o nível de produção.

Variável Omitida

- Seja o seguinte exemplo

$$\underbrace{\log(\text{salario})_i}_{=y_i} = \beta_0 + \beta_1 \underbrace{\text{educ}_i}_{=x_i} + \underbrace{\beta_2 \text{habil}_i + \epsilon_i}_{=u_i} \quad (18)$$

- Assim,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) (\beta_0 + \beta_1 x_i + u_i)}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\begin{aligned}
\hat{\beta}_1 &= \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
\hat{\beta}_1 &\xrightarrow{p} \beta_1 + \frac{\text{cov}(x_i, u_i)}{\text{var}(x_i)}, \quad \text{quando } n \rightarrow \infty \\
\hat{\beta}_1 &\xrightarrow{p} \beta_1 + \beta_2 \frac{\text{cov}(x_i, \text{habil}_i)}{\text{var}(x_i)} + \underbrace{\frac{\text{cov}(x_i, \epsilon_i)}{\text{var}(x_i)}}_{=0}
\end{aligned} \tag{19}$$

- Considere um modelo linear populacional

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + u \quad (20)$$

em que

$$\mathbb{E}(u) = 0, \quad \text{cov}(x_j, u) = 0, \quad j = 1, 2, \dots, k-1 \quad (21)$$

e as variáveis x_1, x_2, \dots, x_{k-1} são exógenas, mas x_k é potencialmente endógena.

- A estimação por mínimos quadrados da eq. (20) gera estimadores inconsistentes para todos os β_j se $\text{cov}(x_k, u) \neq 0$.
- O **método de Variáveis Instrumentais** (VI) fornece uma solução geral para o problema de uma variável explicativa endógena.

- O uso de VI com x_k endógeno requer que a variável instrumental z_1 satisfaça duas condições. A primeira condição é a de **exogeneidade do instrumento**:

$$\text{cov}(z_1, u) = 0 \quad (22)$$

- A segunda condição é a de **relevância do instrumento** e requer a existência de uma relação entre z_1 e a variável endógena x_k dada por:

$$x_k = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \dots + \delta_{k-1} x_{k-1} + \theta_1 z_1 + r_k \quad (23)$$

- Por definição, $\mathbb{E}(r_k) = 0$ e $\mathbb{E}(r_k | x_1, x_2, \dots, x_{k-1}, z_1) = 0$.

- O pressuposto principal é que $\theta_1 \neq 0$. **Esta condição é descrita como**

$$\text{cov}(z_1, x_k) \neq 0 \quad (24)$$

- Se x_k for a única variável explicativa na eq. (23), então a projeção linear é:

$$x_k = \delta_0 + \theta_1 z_1 + r_k \quad (25)$$

em que $\theta_1 = \frac{\text{cov}(z_1, x_k)}{\text{var}(z_1)}$.

- Quando a variável z_1 satisfaz ambas as condições, eq. (22) e (23), dizemos que z_1 é um **instrumento válido** para x_k .

- Como x_1, x_2, \dots, x_{k-1} são não correlacionados com u , eles servem como seus próprios instrumentos.
- A lista de todos os instrumentos é a mesma que a lista de variáveis exógenas, embora frequentemente apenas nos referimos aos instrumentos da variável explicativa endógena.
- A projeção linear na eq. (23) é chamada de **forma reduzida** para a variável explicativa endógena x_k .
- Usamos o termo **forma reduzida** em todo contexto de VI porque ela é uma forma simples de *estabelecer que uma variável endógena foi linearmente projetada nas variáveis exógenas*. A terminologia também transmite que não há nada necessariamente estrutural.

- Obtemos a **forma reduzida** para y plugando as equações e rearranjando:

$$y = \alpha_0 + \alpha_1 x_1 + \dots + \alpha_{k-1} x_{k-1} + \lambda_1 z_1 + \nu \quad (26)$$

em que $\nu = u + \beta_k r_k$ é o erro da forma reduzida, $\alpha_j = \beta_j + \beta_k \delta_j$, e $\lambda_1 = \beta_k \theta_1$.

- Por pressuposto, $\mathbb{E}(\nu | x_1, x_2, \dots, z_1) = 0$.
- A estimação da eq. (26) por mínimos quadrados gera estimativas consistentes para os parâmetros da forma reduzida α_j e λ_1 .

- Com isso podemos verificar que os pressupostos realizados sobre a variável instrumental z_1 resolve o **problema de identificação** para β_j na eq. (20).
- **Identificação** significa que podemos escrever o β_j em termos dos momentos populacionais nas variáveis observáveis. Para vermos isso, vamos reescrever a eq. (20) como

$$y = \mathbf{x}\boldsymbol{\beta} + u \quad (27)$$

em que a constante está incluída dentro do vetor \mathbf{x} tal que $\mathbf{x} = (1, x_2, \dots, x_k)$. Temos o vetor de dimensão $(1 \times k)$ de todas as variáveis exógenas como

$$\mathbf{z} \equiv (1, x_2, \dots, x_{k-1}, z_1) \quad (28)$$

- O pressuposto dado pelas eq. (27) e (28) implica k condições de ortogonalidade

$$\mathbb{E}(\mathbf{z}'u) = 0 \quad (29)$$

- Multiplicando a eq. (27) por \mathbf{z}' , tirando a esperança e usando a eq. (29), temos

$$[\mathbb{E}(\mathbf{z}'\mathbf{x})] \boldsymbol{\beta} = \mathbb{E}(\mathbf{z}'y) \quad (30)$$

em que $\mathbb{E}(\mathbf{z}'\mathbf{x})$ tem dimensão $k \times k$ e $\mathbb{E}(\mathbf{z}'y)$ tem dimensão $k \times 1$.

- A eq. (30) representa um sistema de k equações lineares com k parâmetros desconhecidos $\beta_1, \beta_2, \dots, \beta_k$. O sistema tem solução única se e somente se a matriz $\mathbb{E}(\mathbf{z}'\mathbf{x})$ tem *rank* completo, isto é:

$$\text{rank} [\mathbb{E}(\mathbf{z}'\mathbf{x})] = k \quad (31)$$

cuja solução é

$$\beta = [\mathbb{E}(\mathbf{z}'\mathbf{x})]^{-1} \mathbb{E}(\mathbf{z}'\mathbf{y}) \quad (32)$$

- As esperanças de $\mathbb{E}(\mathbf{z}'\mathbf{x})$ e $\mathbb{E}(\mathbf{z}'\mathbf{y})$ podem ser estimadas consistentemente usando uma amostra aleatória de $(\mathbf{x}, y, \mathbf{z}_1)$, e assim identificar o vetor β .

- Dado uma amostra aleatória $\{(\mathbf{x}_i, y_i, z_{i1}) : i = 1, 2, \dots, N\}$ da população, o **estimador de variável instrumental** de β é:

$$\hat{\beta} = \left(\frac{1}{N} \sum_{i=1}^N z_i' \mathbf{x}_i \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^N z_i' y_i \right) = (\mathbf{Z}' \mathbf{X})^{-1} \mathbf{Z}' \mathbf{Y} \quad (33)$$

em que \mathbf{Z} e \mathbf{X} são matrizes de dados com dimensão $(N \times k)$ e \mathbf{Y} é um vetor de dados com dimensão $(N \times 1)$ em y_i .

- A consistência desse estimador é dada pela *lei dos grandes números*.

- Quando buscamos por instrumentos para uma variável explicativa endógena, as condições dada pela eq. (22) e (24) são igualmente importantes para identificar β .
- Duas observações importantes:
 - ① A condição de exogeneidade, $\text{cov}(z_1, u) = 0$, **não é testável**, pois refere-se à covariância entre z_1 e um erro **não observável**.
 - ② A condição de relevância, $\text{cov}(z_1, x) \neq 0$, **pode ser testada** em uma regressão de x em z com um teste de significância no coeficiente associado ao instrumento.

ECONOMETRIA I

ENDOGENEIDADE

Victor Oliveira

Núcleo de Economia Internacional e Desenvolvimento Econômico

PPGDE – 2023