Causalidade e Variáveis Instrumentais

Prof. Alexandre Gori Maia

Universidade Estadual de Campinas

Disciplina: Econometria Aplicada II

Universidade Nacional Agraria La Molina

Ementa

Viés de Omissão

Mínimos Quadrados em Dois Estágios

Propensity Score Matching

Bibliografia

Angrist, J.; Pischke, J. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, Caps. 1-4, 2009.

Teorema de Gauss-Markov

- Para que os estimadores de MQO sejam os Melhores Estimadores Lineares Não Viesados (MELNV, ou *BLUE*):

Modelo Clássico de Regressão Linear

1) Relação Linear entre *X* e *Y*: ***Viés e Consistência***

O ajuste só é válido para relações lineares.

2) Os valores de X são fixos em repetidas amostras, e não aleatórios:

Quem varia é o regressando, o regressor é fixo e dado, qualquer que seja a amostra. Fazemos a pressuposição que dado um valor de X, Y irá variar segundo uma distribuição de probabilidade com valor esperado dado por $E(Y/X_i)$.

3) Os erros possuem esperança condicional zero, ou seja, $E(e|X_i)=0$:

É a mesma coisa afirmar que $E(Y/X_i)=\mathbf{X}\beta$.

4) A variabilidade dos erros é constante, qualquer que seja *X*: ***Eficiência***

Não há relação entre os erros e as variáveis independentes.

5) Os erros são não autocorrelacionados, ou seja, $E(e_{it}e_{js})=0$:

Não há relação entre valores ordenados dos erros segundo tempo ou espaço.

6) Os erros apresentam distribuição normal: ***Inferência***

Não é um pressuposto necessário para que os estimadores de MQO sejam MELNV, mas necessário para que estes tenham distribuição normal.

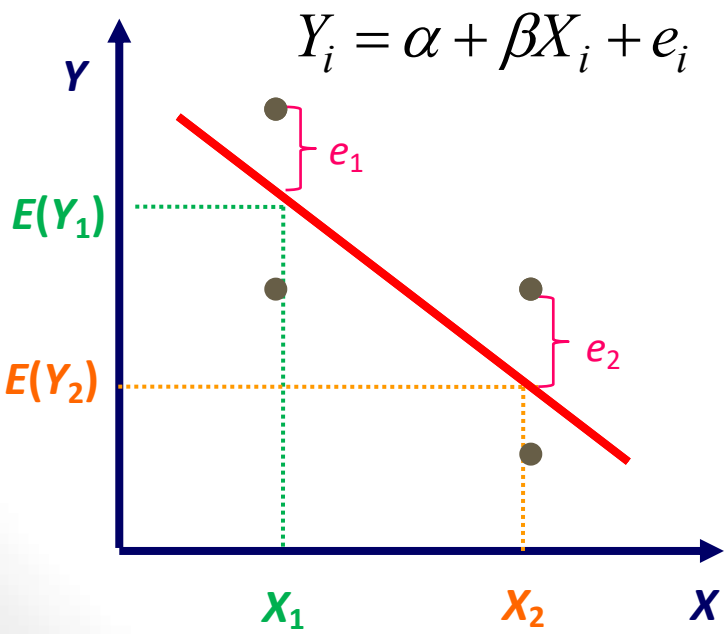
Endogeneidade

Caso os valores de *X* não sejam fixos, mas comportem-se como um v.a., esses não poderão estar relacionados aos erros *e* do modelo:

$$E(e|X) = 0$$

Dizemos que um regressor *X* é endógeno quando este apresenta relação com os erros *e* do modelo:

$$E(e|X) \neq 0$$



Controlando-se os valor de *X*, seria possível observar variações aleatórias de *Y* ou *e* (que representa a influência das variáveis omitidas).

Mas se, por exemplo, um efeito positivo de *e* sobre *Y* causar também efeito sobre *X*, a reta se deslocará para cima. Nessas condições, os estimadores de MQO serão tendenciosos e inconsistentes.

Causas - Viés de Omissão

- Suponha 6 propriedades agrícolas com 3 tamanhos distintos (*A* em hectares);
- A produção agrícola (*Y*) tenderá a ser maior propriedades maiores;
- Imagine agora que sejam disponibilizados crédito governamentais (*X*, em mil reais) em cada propriedade, sem que o crédito tenha qualquer impacto na produção (*Y*). Mas as propriedades maiores obtiveram mais crédito;

A=2	*A*=2	*A*=4	*A*=4	*A*=6	*A*=6
Y=2000	*Y*=2200	*Y*=4200	*Y*=4000	*Y*=6200	*Y*=6000
X=2	*X*=4	*X*=6	*X*=8	*X*=10	*X*=12

- Se relacionarmos a o volume de crédito (*X*) com a produção (*Y*), sem considerar o tamanho da propriedade, podemos enganosadamente pensar que seu volume influencia na produção:

Y=2000	*Y*=2200	*Y*=4200	*Y*=4000	*Y*=6200	*Y*=6000
X=2	*X*=4	*X*=6	*X*=8	*X*=10	*X*=12

Produções *Y* altas estão associadas a elevadas quantidades de *X*, o que não significa que *X* necessariamente determine *Y*.

Viés de Omissão - Definição

- Suponha que a real relação entre as variáveis na população seja:

$$Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + e_i$$

- Mas, erroneamente, se ajusta o modelo:

$$Y_i = \tilde{\alpha} + \tilde{\beta}_1 X_1 + e_i$$

- A omissão indevida do regressor X_2 no modelo causará viés na estimativa de $\tilde{\beta}_1$. Pode-se demonstrar que:

$$E(\tilde{\beta}_1) = \beta_1 + \beta_2 \tilde{\delta}_1 \qquad \text{onde} \quad X_2 = \tilde{\delta}_0 + \tilde{\delta}_1 X_1$$

- Assim, o viés em β_1 dependerá de β_2 e do sentido da relação entre X_1 e X_2. De maneira geral:

	Corr (X_1, X_2) > 0	Corr (X_1, X_2) < 0
$\beta_2 > 0$	Viés Positivo	Viés Negativo
$\beta_2 < 0$	Viés Negativo	Viés Positivo

Exercícios

1) O arquivo *Data_RelativeIncome.xls* contém uma amostra de domicílios com dados para renda relativa (média da vizinhança) e suficiência de renda (GORI MAIA, A. Relative Income, Inequality and Subjective Wellbeing: Evidence for Brazil. Social Indicators Research, v. 113, p. 1193-1204, n. 2013) :
   a) Analise a relação entre suficiência de renda e o log da renda média da vizinhança, sem controles;
   b) Analise a relação entre suficiência de renda e renda média da vizinhança, incorporando controles para renda per capita e outros controles que julgar necessário;

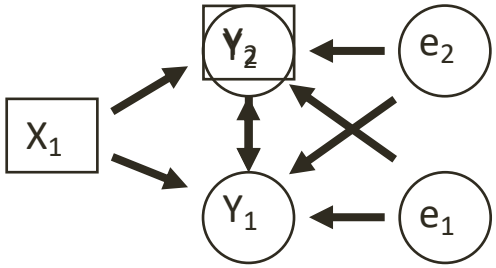
Causas - Simultaneidade

Sistema de Equações Simultâneas:

Seja o sistema de equações simultâneas:

$$Y_{1i} = \alpha_0 + \alpha_1 Y_{2i} + \alpha_2 X_{1i} + e_{1i}$$

$$Y_{2i} = \beta_0 + \beta_1 Y_{1i} + \beta_2 X_{1i} + e_{2i}$$



Y_1 e Y_2 são mutuamente dependentes, ou ditas **variáveis endógenas**, determinadas dentro do sistema. X_1 e X_2 são **variáveis exógenas,** definidas fora do sistema.

Conseqüencias da simultaneidade

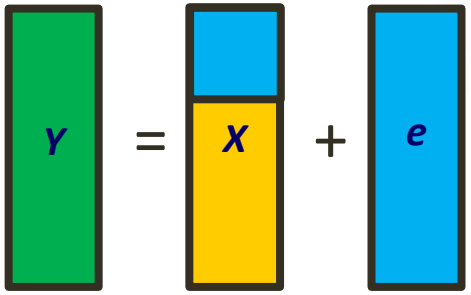
Como há uma mútua relação entre as variáveis endógenas Y_1 e Y_2, os erros e_1 da primeira equação afetarão, simultanemente, Y_1 e Y_2. Analogamente, os erros e_2 da segunda equação afetarão simultaneamente Y_2 e Y_1. A existência de relação entre erros e_1 e regressor Y_2 na primeira equação, assim como a relação entre erros e_2 e regressor Y_1 na segunda equação, tornam os estimadores de MQO **viesados** e **inconsistentes**.

Correção - Variáveis Instrumentais

Desejamos analisar: $Y_i = \alpha + \beta X_i + e_i$

Mas temos: $\mathrm{Cov}(X_i, e_i) \neq 0$

Os estimadores de MQO serão viesados e inconsistentes

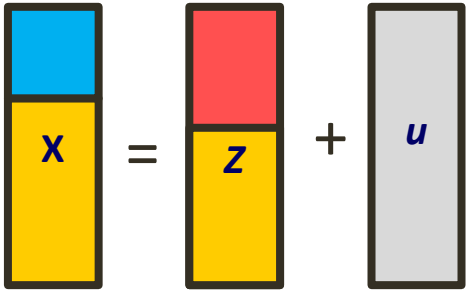


Desejamos encontrar um instrumento Z tal que:

$\mathrm{Cov}(Z_i, e_i) = 0$ e $Cov(X_i, Z_i) \neq 0$

A parcela de Z associada a X será estimada por:

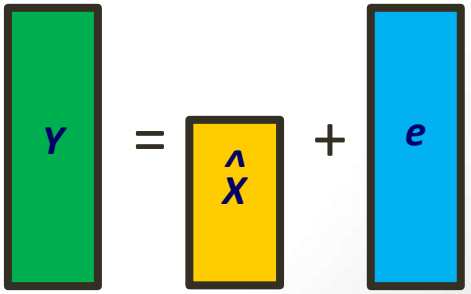
$$\hat{X}_i = \hat{\delta}_0 + \hat{\delta}_1 Z_i$$



Um estimador consistente pode ser obtido por:

$$Y_i = \alpha + \beta \hat{X}_i + e_i$$

Embora consistente, o estimador obtido com o uso de VI tende a ser viesado para amostras pequenas.



Mínimos Quadrados 2 Estágios

Dado um sistema de equações:

$$\begin{cases} Y_1 = \alpha_0 + \alpha_1 Y_2 + \alpha_2 X + e_1 \\ Y_2 = \beta_0 + \beta_1 Y_1 + e_2 \end{cases}$$

Subidentificada

Exatamente identificada

A forma reduzida será dada por:

$$\begin{cases} Y_1 = \underbrace{\frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1}}_{\pi_1} - \underbrace{\frac{\alpha_2}{\alpha_1 - \beta_1}}_{\pi_2} X + \underbrace{\frac{e_2 - e_1}{\alpha_1 - \beta_1}}_{u_1} \\ Y_2 = \underbrace{\frac{\alpha_1\beta_0 - \alpha_0\beta_1}{\alpha_1 - \beta_1}}_{\pi_3} - \underbrace{\frac{\alpha_2\beta_1}{\alpha_1 - \beta_1}}_{\pi_4} X + \underbrace{\frac{\alpha_1 e_2 - \beta_1 e_1}{\alpha_1 - \beta_1}}_{u_2} \end{cases}$$

Aplicando-se MQO...

$$\begin{cases} \hat{Y}_1 = \hat{\pi}_1 + \hat{\pi}_2 X \\ \hat{Y}_2 = \hat{\pi}_3 + \hat{\pi}_4 X \end{cases}$$

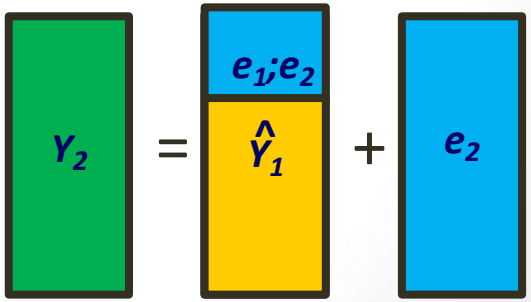
Para a equação identificável (Y_2):

Desejamos analisar: $Y_2 = \beta_0 + \beta_1 Y_1 + e_2$

Mas temos: $Cov(Y_1, e_2) \neq 0$

Então analisamos: $Y_2 = \beta_0 + \beta_1 \hat{Y}_1 + e_2$

Pois: $Cov(\hat{Y}_1, Y_1) \neq 0$ e $Cov(\hat{Y}_1, e_2) = 0$



Mínimos Quadrados 2 Estágios

Passos do MQ2E:

1) **Identifiação:** verificar quais equações do sistema são identificadas (possuem instrumentos exógenos);
2) **Forma Reduzida**: construir sistema de equações reduzidas tendo, em cada equação, uma variável endógena em função das variáveis exógenas do sistema;
3) **Variável instrumental**: obter valores previstos para variáveis endógenas nas equações da forma reduzida;
4) **Resolver forma estrutural**: aplicar MQO nas equações identificáveis da forma estrutural, substituindo os regressores endógenos pelas suas respectivas variáveis instrumentais;

Dado um sistema de equações na forma estrutural:

$$Y_1 = \alpha_0 + \alpha_1 Y_2 + \alpha_2 X_1 + e_1$$ Exatamente identificada

$$Y_2 = \beta_0 + \beta_1 Y_1 + \beta_2 X_2 + e_2$$ Exatamente identificada

1

2

Chega-se à forma reduzida:

$$\begin{cases} Y_1 = \pi_1 + \pi_2 X_1 + \pi_3 X_2 + u \\ Y_2 = \pi_4 + \pi_5 X_1 + \pi_6 X_2 + v \end{cases}$$

Aplicando-se MQO...

$$\begin{cases} \hat{Y}_1 = \hat{\pi}_1 + \hat{\pi}_2 X_1 + \hat{\pi}_3 X_2 \\ \hat{Y}_2 = \hat{\pi}_4 + \hat{\pi}_5 X_1 + \hat{\pi}_6 X_2 \end{cases}$$

3

$$\begin{cases} Y_1 = \alpha_0 + \alpha_1 \hat{Y}_2 + \alpha_2 X_1 + e_1' \\ Y_2 = \beta_0 + \beta_1 \hat{Y}_1 + \beta_2 X_2 + e_2' \end{cases}$$

4

Importante

Os estimadores de MQ2E são consistentes, embora tendam a ser viesados para amostras pequenas.

Exercícios

2) O arquivo *Data_HealthIncome.xls* contém uma amostra de domicílios com dados para saúde e rendimento do trabalho (MAIA, A. G. , RODRIGUES, C. G. . Saúde e mercado de trabalho no Brasil: diferenciais entre ocupados agrícolas e não agrícolas. Revista de Economia e Sociologia Rural (Impresso), v. 48, p. 737-765, n. 2010) :
   a) Analise a relação entre saúde e rendimento do trabalho usando MQO;
   b) Analise a relação entre saúde e rendimento do trabalho usando MQ2E;

Causa - Viés de Seleção

- A seleção de grupos ou indivíduos da população é definida por critérios não aleatórios, os quais não são observados ou controlados;

- A comparação entre o grupo de tratados (T=1) e o grupo de controle (T=0) em um modelo clássico de regressão implicaria em estimativas tendenciosas e inconsistentes, pois:

$$Y = \alpha + \mathbf{\beta x} + \rho T + e \qquad \text{e} \qquad E(e|T) \neq 0$$

- O ideal seria estimarmos o *Average Treatment Effect* (ATE) comparando os resultados para o mesmo indivíduos antes (Y_0) e depois do tratamento (Y_1). Se a seleção fosse totalmente aleatória teríamos:

$$ATE = E(Y_{1i} - Y_{0i}) = E(Y_i|T = 1) - E(Y_i|T = 0)$$

Correção - Pareamento

- Seja a comparação entre o grupo de tratados (T=1) e o grupo de controle (T=0) em um modelo clássico de regressão:

$$Y = \alpha + \boldsymbol{\beta}\mathbf{x} + \rho T + e$$

- Em que a seleção de T não seja aleatório e dependente de fatores não controlados:

$$E(e|T) \neq 0$$

- O método de *Propensity Score Matching* elimina o viés de seleção que se origina de fatores observáveis ($\mathbf{x}$), comparando indivíduos tratados e não tratados com características similares (*propensity scores* – p($\mathbf{x}$) – similares):

$$p(\mathbf{x}) = prob(T = 1) = \boldsymbol{\pi}\mathbf{x} + u$$

- O efeito do tratamento será então dado pelo *Average Effect of Treatment on the Treated* (ATT):

$$ATT = E[Y_{1i} - Y_{0i}|T_i = 1, p(\mathbf{x_i})] = E[Y_{1i}|T_i = 1, p(\mathbf{x_i})] - E[Y_{0i}|T_i = 0, p(\mathbf{x_i})]$$

Exercícios

3) O arquivo *Data_MFA.xls* contém uma amostra de domicílios com dados para o programa Mas Famílian en Accion (MFA) e percepção de pobreza (MORALES MARTINEZ, D.; GORI MAIA, A. The impacts of cash transfers on subjective wellbeing and poverty: The case of Colombia. International Journal of Family and Economic Issues (in press), 2018) :
   a) Analise o impacto do programa MAF sobre a percepção de pobreza usando MQO;
   b) Analise o impacto do programa MAF sobre a percepção de pobreza usando *propensity score matching*;

Exercícios

4) O arquivo *Data_AgriculturalCensus.xls* contém uma amostra do Censo Agropecuário Brasileiro de 2006 (SANTOS, G. E., GORI MAIA, A., SILVEIRA, R. L. Measuring the Farm Level Impact of Rural Credit: A Two-stage Approach. Annals of the Agricultural and Applied Economics Association 2017 Annual Meeting, July 30-August 1, Chicago) :
   a) Aplique MQO para analisar o impacto do acesso ao crédito no (log) valor total da produção agropecuária;
   b) Avalie a necessidade de controle por fatores observáveis;
   c) Aplique MQ2E utilizando o (log) valor da dívida total como instrumento para acesso ao crédito;
   d) Utilize o PSM para estimar o ATT;