Objetivos

Ao final desse grupo de *slides* os alunos deverão ser capazes de:

- ✓ **Explicar a natureza dos regressores endógenos;**
- ✓ **Discutir as implicações nas propriedades dos estimadores de MQO quando regressores endógenos são incluídos ao modelo de regressão de interesse;**
- ✓ **Entender o significado e a utilidade de uma variável instrumental;**
- ✓ **Para o contexto que está sendo estudado, propor uma variável instrumental.**

Insper

Endogeneidade

Aula 15a

***Leitura DETALHADA*: Wooldridge, 2011 – Seção 9.4 e Capítulo 15**

Endogeneidade

Definição. Qualquer variável explicativa, num modelo de regressão linear múltipla do tipo

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

que seja correlacionada com o termo de erro estocástico é dita variável explicativa endógena.

Endogeneidade

Problema:

A presença de regressores endógenos num modelo de regressão linear faz com que

$$\mathrm{E}\left(\varepsilon_i \mid x_{1i}, x_{2i}, ..., x_{ki}\right) = 0, \;\; i = 1, \; 2, \; ..., \; n$$

Ou seja, viola a viola a MLR.4.

Endogeneidade

Recordação:

MLR.4

Todos os fatores contidos em ε devem ser não correlacionados com as variáveis explicativas e deve ter sido usada a forma funcional correta.

Endogeneidade

Quais fatos podem estar ligados à violação da MLR.4?

- ***Omissão de regressor relevante*, correlacionado com x_1, x_2, ... ou x_k;**
- ***Erro de medida* em x_1, x_2, ... ou x_k; (Leitura Complementar)**
- ***Simultaneidade* entre y e x_1, x_2, ...ou x_k;**
- ***Forma funcional especificada incorretamente*.**

Endogeneidade

Caso a suposição MLR.4 seja violada:

- **os estimadores de MQO dos parâmetros do modelo de regressão linear serão viesados, inconsistentes e ineficientes;**
- **o estimador da variância do termo de erro aleatório também será viesado e inconsistente;**
- **toda a análise inferencial estará comprometida.**

Exemplo

Considere o seguinte modelo de regressão linear simples:

$$nota_i = \beta_0 + \beta_1 faltas_i + \varepsilon_i \qquad (1)$$

Qual motivo nos levaria a desconfiar que:

$$\mathrm{E}\left(\varepsilon_i \mid faltas_i\right) \neq 0, \;\; i = 1, 2, ..., n$$ **?**

Resposta: *o regressor faltas pode estar correlacionado com o regressor motivação (que está no termo de erro, é não observável diretamente e certamente afeta a variável resposta nota).*

Voltando ao Exemplo

Ao calcularmos a covariância entre as variáveis *nota* e *faltas*, vem que:

$$\mathrm{Cov}(\text{faltas}, \text{nota}) = \mathrm{Cov}(\text{faltas}, \beta_0 + \beta_1 \text{faltas} + \varepsilon)$$

$$\mathrm{Cov}(\text{faltas}, \text{nota}) = \beta_1 \mathrm{Cov}(\text{faltas}, \text{faltas}) + \mathrm{Cov}(\text{faltas}, \varepsilon)$$

$$\mathrm{Cov}(\text{faltas}, \text{nota}) = \beta_1 \mathrm{Var}(\text{faltas}) + \mathrm{Cov}(\text{faltas}, \varepsilon)$$

$$\beta_1 = \frac{\mathrm{Cov}(\text{faltas}, \text{nota}) - \mathrm{Cov}(\text{faltas}, \varepsilon)}{\mathrm{Var}(\text{faltas})}$$

Notamos que o parâmetro β_1 não está refletindo o efeito de interesse. Ou seja, β_1 não está identificado.

Observação

Identificação: Podemos escrever o parâmetro β_1 em termos de momentos populacionais que podem ser estimados usando uma amostra de dados. (Wooldridge, 2010, p. 474)

VARIÁVEIS INSTRUMENTAIS (IV)

Wooldridge, 2010 (4ª edição) – Capítulo 15

Variáveis Instrumentais

Pergunta 1: Qual a utilidade das variáveis instrumentais?

O uso das variáveis instrumentais (IV) nos auxiliará na busca de estimadores consistentes, quando tivermos regressores endógenos presentes no modelo de regressão.

Pergunta 2: O que são variáveis instrumentais?

Resposta nos slides, a seguir!

Variáveis Instrumentais

Para o exemplo das notas, suponha que tenha sido observada uma variável explicativa *z* que satisfaça a duas suposições:

(a) *z* é não-correlacionada com ε, isto é,

$$\text{Cov}(z,\varepsilon) = 0$$

z é exógena em (5)

(b) *z* é correlacionada com *faltas*, isto é,

$$\text{Cov}(z, \textit{faltas}) \neq 0$$

Como verificar a validade de (a) e (b)?

Variáveis Instrumentais

Do *slide* anterior, chamaremos *z* de variável instrumental para *faltas* ou, simplesmente, instrumento para *faltas*.

A exigência que o instrumento *z* satisfaça (a) é resumida dizendo-se "*z* é exógena na equação (1)".

pergunta: Em (1), proponha ao menos uma variável que poderia ser usada como instrumento para *faltas*. Justifique a sua resposta.

Voltando ao Exemplo

Ao calcularmos a covariância entre as variáveis *nota* e *distância*, vem que:

$$\text{Cov}(\text{distância}, \text{nota}) = \text{Cov}(\text{distância}, \beta_0 + \beta_1 \text{faltas} + \varepsilon)$$

$$\text{Cov}(\text{distância}, \text{nota}) = \beta_1 \text{Cov}(\text{distância}, \text{faltas}) + \text{Cov}(\text{distância}, \varepsilon)$$

$$\text{Cov}(\text{distância}, \text{nota}) = \beta_1 \text{Cov}(\text{distância}, \text{faltas})$$

$$\beta_1 = \frac{\text{Cov}(\text{distância}, \text{nota})}{\text{Cov}(\text{distância}, \text{faltas})}$$

Notamos, agora, que o parâmetro β_1 está refletindo o efeito de interesse.

Ou seja, β_1 está identificado.

Voltando ao Exemplo

Dessa forma, do *slide* anterior, chamando nota de *y*, faltas de *x* e distância de *z*, não é difícil provar que o estimador para o parâmetro de interesse é dado por:

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

Já o intercepto, será estimado por:

$$\hat{\beta}_0^{(IV)} = \bar{y} - \hat{\beta}_1^{(IV)} \bar{x}$$

Exercício

Considere o modelo

$$salario_i = \beta_0 + \beta_1 educ_i + \varepsilon_i \qquad (2)$$

com

$$Cov(educ, \varepsilon) \neq 0.$$

Pergunta: qual razão estaria nos levando à violação desta premissa?

Resposta: ***educ deve estar correlacionada com habilidade (que certamente afeta salário e encontra-se no termo de erro e, além de tudo, é não observável diretamente).***

Insper

LEITURA COMPLEMENTAR

ERROS NAS VARIÁVEIS

Wooldridge, 2011 – Seção 9.4

Erros de Medição

Os erros de medição são potencialmente um problema sério, pois constituem mais um exemplo de viés de especificação com as consequências que serão dadas a seguir.

Erros de Medição em y

Suponha o seguinte modelo de regressão:

$$y_i^* = \beta_1 + \beta_2 x_{2i} + \varepsilon_i \qquad (1)$$

em que

y_i^* não é medida diretamente.

Entretanto, observamos

$$y_i = y_i^* + u_i$$

em que

u_i denota erros de medição em y_i^*.

Erros de Medição

Por exemplo, y pode representar a poupança anual registrada pelas famílias.

Infelizmente, muitas famílias podem não declarar com perfeição suas poupanças anuais; ou seja, em muitos casos, é fácil que algumas famílias deixem algumas categorias de fora ou superestimem o montante contribuído para determinado fundo.

Assim, geralmente podemos esperar que y e y^* sejam diferentes, pelo menos em alguns subconjuntos de famílias da população.

Erros de Medição em y

Dessa forma, ao invés de estimarmos os parâmetros de **(1)**, estimamos

$$y_i = \beta_1 + \beta_2 x_{2i} + \nu_i \qquad \textbf{(2)}$$

em que

$$\nu_i = \varepsilon_i + u_i.$$

Erros de Medição em y

Por simplicidade, vamos admitir que:

- $E(\varepsilon_i) = E(u_i) = 0$;
- $Cov(x_{2i}, \varepsilon_i) = 0$ (que é uma das premissas clássicas);
- $Cov(x_{2i}, u_i) = 0$; isto é, o erro de medição de y_i^* não está correlacionado com x_{2i}; e
- $Cov(\varepsilon_i, u_i) = 0$; isto é, o termo de erro de **(1)** e o termo de erro de medição não estão correlacionados.

Erros de Medição em y

Dessa forma, não é difícil ver que os parâmetros de (1) ou (2), estimados por MQO, serão não viesados.

Contudo, as variâncias dos estimadores de (1) e (2) serão diferentes, sendo que em (2) teremos estimadores menos eficientes (vale lembrar que o estimador da variância continua não viesado).

Erros de Medição em x

Suponha o seguinte modelo de regressão:

$$y_i = \beta_1 + \beta_2 x_{2i}^* + \varepsilon_i \qquad (3)$$

em que

x_{2i}^* não é medida diretamente.

Entretanto, observamos

$$x_{2i} = x_{2i}^* + \xi_i$$

em que

ξ_i denota erros de medição em x_{2i}^*.

Erros de Medição em x

Por exemplo, x_2 pode representar a renda familiar informada pelos estudantes, num estudo onde objetiva-se estimar o efeito renda familiar na nota média da graduação.

Em nosso exemplo, x_2^* representa a renda familiar efetiva.

Assim, a renda familiar informada pelos estudantes pode, facilmente, ter sido incorretamente medida.

Erros de Medição em x

Dessa forma, ao invés de estimarmos os parâmetros de **(3)**, estimamos

$$y_i = \beta_1 + \beta_2 (x_{2i} - \xi_i) + \varepsilon_i$$

$$y_i = \beta_1 + \beta_2 x_{2i} + (\varepsilon_i - \beta_2 \xi_i)$$

$$y_i = \beta_1 + \beta_2 x_{2i} + \zeta_i \qquad (4)$$

em que

$$\zeta_i = \varepsilon_i - \beta_2 \xi_i .$$

Erros de Medição em x

Mesmo supondo que ξ_i tenha média zero, que seja serialmente não correlacionado e não esteja correlacionado com ε_i, não podemos admitir que o termo composto ζ_i seja independente da variável explicativa do modelo de interesse, uma vez que

$$\mathrm{Cov}(x_{2i}, \zeta_i) = E\{[(x_{2i} - E(x_{2i}))][(\zeta_i - E(\zeta_i))]\}$$

$$\mathrm{Cov}(x_{2i}, \zeta_i) = E[\xi_i(\varepsilon_i - \beta_2 \xi_i)]$$

$$\mathrm{Cov}(x_{2i}, \zeta_i) = E[\xi_i(\varepsilon_i - \beta_2 \xi_i)] = -\beta_2 E[\xi_i^2] = -\beta_2 Var[\xi_i]$$

Erros de Medição em x

Dessa forma, a variável explicativa e o termo de erro, em (4), são correlacionados, o que viola a suposição de que a variável explicativa e o termo de erro estocástico sejam não correlacionados.

Assim sendo, não é difícil demonstrar que os estimadores de MQO dos parâmetros do modelo de regressão são tendenciosos e inconsistentes.

Erros de Medição em x

Wooldridge (2011, p. 301) mostra que

$$\text{plim}\left(\hat{\beta}_2\right) = \beta_2 \left[\frac{\sigma^2_{x_2^*}}{\sigma^2_{x_2^*} + \sigma^2_{\xi}} \right]$$

(viés de atenuação)

Como é esperado que o termo entre colchetes seja menor que 1, isso mostra que o estimador nunca convergirá para o parâmetro.