Variáveis Instrumentais

Profa. R. Ballini

Bibliografia Básica:

Wooldridge, J. (2002) *Introductory Econometric*, Cap. 15.

Greene, W. (2012). *Econometric Analysis*, Cap. 20.

Teorema de Gauss-Markov

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

1) O modelo de regessão é linear nos parâmetros;
2) Os valores de X são fixos em repetidas amostras;
3) Ausência de colinearidade perfeita;
4) $E(u_j|X_1, \ldots, X_k) = 0$;
5) $Var(u_j|X_1, \ldots, X_k) = \sigma^2$ constante;
6) $corr(u_i, u_j) = 0, \; i \neq j$;
7) O termo estocástico u_j se distribui normalmente.

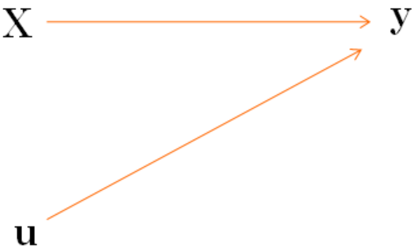
Exogeneidade

Podemos usar mínimos quadrados ordinários (MQO) para estimar consistentemente o seguinte modelo:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} \tag{1}$$

Nenhuma associação entre **X** e **u**; MQO é consistente.

- Suposição: $E(\mathbf{u}|\mathbf{X}) = 0$.

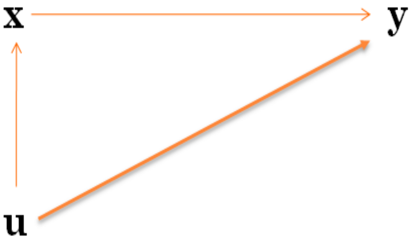


Endogeneidade

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

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + u_i$$

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



Endogeneidade

Quais poderiam ser as razões ligadas à ocorrência de tal fenômeno?

1. Omissão de variáveis relevantes, correlacionadas com $x_1, \ldots, x_k$;
2. Erros de medição em $x_1, \ldots, x_k$;
3. Simultaneidade entre y e uma ou mais variáveis explicativas.

Endogeneidade

Uma variável explicativa, em um modelo de regressão linear que é correlacionado com o termo de erro é denominada variável explicativa endógena, ou seja,

$$plim\left(\frac{1}{n}\mathbf{X}'\mathbf{u}\right) \neq 0$$

em que *plim* é a limite de probabilidade.

- Estimadores de MQO serão viesados, inconsistentes e ineficientes;
- Estimador da variância do erro será viesado e inconsistente;

Solução: empregar variáveis instrumentais, com o intuito de nos auxiliar na busca de estimadores consistentes.

Exemplo

Considere o modelo:

$$\ln(rendimento) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u$$

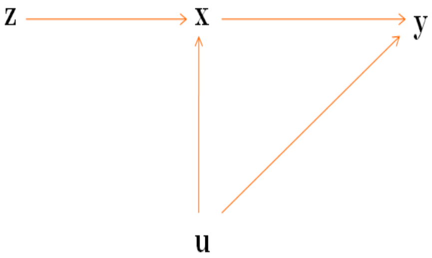
em que $Cov(educ, u) \neq 0$.

As variáveis *rendimento* e *educ* são endógenas e *exper* é exógena.

Variáveis Instrumentais

A solução deste problema por variáveis instrumentais:
Regressão de variáveis instrumentais: $\mathbf{y} = \mathbf{x}\beta + \mathbf{u}$

z é não correlacionado com u, mas correlacionado com x



A variável adicional z é chamada de instrumento para x.

Variáveis Instrumentais

Considere o modelo:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

em que $Cov(x, u) \neq 0$. Para termos uma estimativa consistente precisamos de uma nova informação z que satisfaça as seguintes hipóteses:

1. z é não correlacionada com u, isto é,

$$Corr(z, u) = 0 \tag{2}$$

2. z é correlacionada com x, ou seja,

$$Corr(z, x) \neq 0 \tag{3}$$

z é chamada de **variável instrumental (VI)** de x.

Variáveis Instrumentais

A hipótese (2) não pode ser validada.

A hipótese (3) pode ser testada a partir de uma amostra aleatória. Ou seja, dada uma amostra aleatória, é possível obter a regressão:

$$x = \pi_0 + \pi_1 z + v$$

em que

$$\pi_1 = \frac{Cov(x, z)}{Var(z)}$$

Se π_1 for significativamente diferente de zero, (3) é válida.

ESTIMAÇÃO DOS PARÂMETROS DO MODELO DE REGRESSÃO VIA USO DAS VARIÁVEIS INSTRUMENTAIS

Sob as suposições (2) e (3) conseguiremos identificar os parâmetros da equação estrutural de interesse.

Estimação

Considere o modelo de regressão linear múltipla:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$$

Seja $\mathbf{Z}$ a matriz de instrumentos ($\mathbf{Z}$ é construída de forma análoga à matriz $\mathbf{X}$, substituindo os regressores endógenos pelos respectivos instrumentos).

Observação: Regressores exógenos da matriz $\mathbf{X}$ são usados como instrumentos deles mesmos na matriz de instrumentos $\mathbf{Z}$.

Para obtenção dos resultados consideramos as seguintes suposições:

$$plim\left(\frac{1}{n}\mathbf{Z}'\mathbf{u}\right) = 0 \qquad plim\left(\frac{1}{n}\mathbf{Z}'\mathbf{X}\right) \neq 0$$

Estimação

Multiplicando a equação

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$$

pela transposta da matriz de instrumentos, $\mathbf{Z}$, temos:

$$\mathbf{Z}'\mathbf{y} = \mathbf{Z}'\mathbf{X}\beta + \mathbf{Z}'\mathbf{u}$$

Os estimadores de VI são obtidos por:

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

Que é um vetor de estimadores consistente.

Sob a suposição de homocedasticidade, a variância dos estimadores de IV é dada por:

$$Var(\hat{\beta}_{\mathbf{IV}}) = \sigma^2(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{X}'\mathbf{Z})^{-1}$$

Estimador usual para σ^2 é dado por:

$$\hat{\sigma}^2 = \frac{1}{n}e'e = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\beta}_{\mathbf{IV}})'(\mathbf{y} - \mathbf{X}\hat{\beta}_{\mathbf{IV}})$$

Regressão linear simples

Considerando o modelo de regressão:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

A covariância entre z e y é dada por:

$$Cov(z, y) = \beta_1 Cov(z, x) + Cov(z, u)$$

Sob as hipóteses (2) e (3), temos:

$$\beta_1 = \frac{Cov(z, y)}{Cov(z, x)}$$

Dada uma amostra aleatória, podemos estimar β_1, ou seja,

$$\hat{\beta}_1 = \frac{\sum(z_i - \bar{z})(y_i - \bar{y})}{\sum(z_i - \bar{z})(x_i - \bar{x})}$$

O estimador de β_0 é dado por:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Regressão Linear Simples

Sob as hipóteses (2), (3) e a hipótese de homocedasticidade:

$$E(u^2|z) = Var(u) = \sigma^2$$

a variância assintótica de $\hat{\beta}_1$ é dada por:

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2 R^2_{x,z}}$$

Dado que $0 < R^2_{x,z} < 1$, a variância de VI é sempre maior que a variância de MQO para o estimador de β_1.

Exemplo: Estimação do Retorno da Educação para Mulheres Casadas

Considere o seguinte modelo de regressão linear múltipla:

$$\ln(wage) = \beta_0 + \beta_1 exper + \beta_2 exper^2 + \beta_3 educ + u$$

Utilize os dados do arquivo MROZ.xlsx para estimar este modelo por MQO.

Use a variável educação da mãe (*motheduc*) como instrumento para *educ* e reestime os parâmetros do modelo de interesse por IV.

Observações

1. Admitindo a validade das suposições (2) e (3), o vetor de estimadores gerado com o uso de variáveis instrumentais é consistente (equação identificada).

2. Para estimar o vetor de parâmetros, precisamos garantir que a matriz $\mathbf{Z}'\mathbf{X}$ admita inversa.

3. Se $\mathbf{Z}$ for uma matriz de dimensão $n \times L$ e $\mathbf{X}$ for uma matriz de dimensão $n \times k$, precisaremos que $L = k$ (equação exatamente identificada).

4. Suposição adicional: a variável instrumental z seja fortemente correlacionada com a variável endógena x. **Suposição relacionada ao fato do uso de instrumentos fortes, em detrimento aos instrumentos fracos.**

Mínimos Quadrados de Dois Estágios - MQ2E

- Suponha mais de um instrumento para cada variável endógena

- Admitindo mais de um instrumento para cada variável endógena, como ficariam as dimensões das matrizes **Z** e **X**?

- Admitindo mais de um instrumento para cada variável endógena, seria possível gerar diretamente o vetor de estimadores?

Empregar o Método de Mínimos Quadrados em Dois Estágios - MQ2E

Método de estimação utilizado quando a equação estrutural encontra-se **sobreidentificada**

Mínimos Quadrados de Dois Estágios - MQ2E

Seja o modelo estrutural:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1 \tag{4}$$

em que y_1 é endógena, $E(u_1) = 0$, z_1 é exógena e y_2 é supostamente endógena.

Suponha que z_2 e z_3 instrumentos correlacionados com y_2.

As variáveis z_1, z_2 e z_3 são não correlacionadas com u_1: qualquer combinação linear será uma VI válida.

1o. Estágio

Uma possível combinação linear é:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2 \tag{5}$$

em que $E(v_2) = 0$, $Cov(z_1, v_2) = 0$, $Cov(z_2, v_2) = 0$ e $Cov(z_3, v_2) = 0$.

Equação (5) é denominada de **forma reduzida**.

Para que esta VI não seja perfeitamente correlacionada com z_1, precisamos que π_2 ou π_3 seja diferentes de zero:

$$\pi_2 \neq 0 \quad \pi_3 \neq 0$$

Se $\pi_2 = 0$ e $\pi_3 = 0$ a equação (4) não será identificada.

Para isso podemos usar teste de Wald (ou teste de restrição).

Instrumentos Fracos

Alguns autores (ver Baum, 2006)[1], formalizaram a definição de instrumentos fracos:

> Caso a estatística F da equação de primeiro estágio exceder 10, o(s) instrumento(s) é (são) cosiderado(s) forte(s).

[1]Baum, C. F. (2006). An Introduction to Modern Econometrics Using Stata. College Station, TX: Stata Press.

2o. Estágio

A forma reduzida pode ser estimada por MQO:

$$\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 z_3$$

Após verificação se π_2 e/ou π_3 são significativamente diferentes de zero, podemos usar $\hat{y}_2$ como VI de y_2 e estimar a regressão:

$$y_1 \text{ sobre } \hat{y}_2 \text{ e } z_1$$

por MQO.

Esta etapa é denominado **2o. estágio**.

Variância dos estimadores em MQ2E

A variância de β_1 em MQ2E é dada por:

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{SQT_2(1 - R_2^2)}$$

em que $\sigma^2 = Var(u_1)$, SQT_2 é a variação total de $\hat{y}_2$ e R_2^2 é o R-quadrado de $\hat{y}_2$ sobre todas as variáveis exógenas.

Considere o modelo de regressão linear múltipla:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$$

Construir a matriz de instrumentos $\mathbf{Z}$.

1o. Estágio: Estimar os parâmetros da equação auxiliar

$$\mathbf{X} = \mathbf{Z}\pi + \epsilon$$

por MQO, obtendo:

$$\hat{\pi} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$$

Gerar a matriz de valores ajustados

$$\hat{\mathbf{X}} = \mathbf{Z}\hat{\pi} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} = \mathbf{P_Z X}$$

em que $\mathbf{P_Z} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$.

2o. Estágio:
Estimar os parâmetros da equação:

$$\mathbf{y} = \hat{\mathbf{X}}\beta + \mathbf{u}$$

obtendo:

$$\hat{\beta}_{\mathbf{2SLS}} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y} = \left(\mathbf{X}'\mathbf{P_Z}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}'\mathbf{y}$$

Variância do Vetor de Estimadores

Sob a suposição de homocedasticidade do vetor de erros, a variância do vetor de estimadores de MQ2E será dada por:

$$Var(\hat{\beta}_{2SLS}) = \sigma^{\mathbf{2}}\left(\mathbf{X}'\mathbf{P_Z}\mathbf{X}\right)^{-1}$$

Como σ^2 é um parâmetro desconhecido, estimamos por:

$$\hat{\sigma}^2 = \frac{1}{n}e'e = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\beta}_{\mathbf{2SLS}})'(\mathbf{y} - \mathbf{X}\hat{\beta}_{\mathbf{2SLS}})$$

Exemplo: Estimação do Retorno da Educação para Mulheres Casadas

Considere o seguinte modelo de regressão linear múltipla:

$$\ln(wage) = \beta_0 + \beta_1 exper + \beta_2 exper^2 + \beta_3 educ + u$$

Utilize os dados do arquivo MROZ.xlsx para estimar este modelo por MQO.

Usando educação do pai (*fatheduc*), educação da mãe (*motheduc*) e educação do marido (*huseduc*) como instrumentos para educ, estime os parâmetros do modelo por MQ2S.

Observações

- Estimador MQ2E é menos eficiente que MQO quando as variáveis explicativas são exógenas.

- MQO e MQ2E fornecem estimadores consistentes se a codição de exogeneidade estiver satisfeita.

- Fazer teste de endogeneidade de uma variável explicativa para verificar se é necessário usar MQ2E.

Teste de Endogeneidade

Hausman (1978)[2] propôs o teste em que é baseado na comparação das estimativas de MQO e MQ2E, para determinar se as diferenças são significativamente diferentes de zero.

Considere o modelo:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1$$

em que y_2 é endógeno e z_1 e z_2 são exógenas.

Se as estimativas geradas por MQO e MQ2E forem significativamente diferentes, y_2 é endógena, supondo z_1 e z_2 exógenas.

[2] Hausman, J. A. *Specification Tests in Econometrics*, **Econometrica** n. 46, p. 1251-1271, 1978

Passos do Teste de Endogeneidade - Teste de Hausman

1. Estime a forma reduzida de y_2 sobre todas as variáveis exógenas:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v_2$$

   Obtenha os resíduos $\hat{v}_2$.

2. Adicione $\hat{v}_2$ à equação na forma estrutural:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 \hat{v}_2 + erro$$

   e estime o modelo por MQO. Se o coeficiente de $\hat{v}_2$ for significativamente diferente de zero, concluímos que y_2 é endógeno. Podemos usar um teste t robusto em relação à heterocedasticidade.

Verificação da Validade dos Instrumentos – Teste de Sargan

Qual a validade do instrumento?

Ou seja, como sabemos que os instrumentos escolhidos são independentes do termo de erro?

Sargan (1964)[3] desenvolveu um teste estatístico para validar os instrumentos

[3] Sargan(1964), "Wages and Prices in the United Kingdom: A Study in Econometric Methodology," in Econometric Analysis for National Economic Planning, eds. P. E. Hart, G. Mills, and J. K. Whitaker, London: Butterworths.

Teste de Sargan

Procedimento do teste:

1. Divida os regressores da equação estrutural em dois conjuntos: (a) conjunto dos regressores exógenos e (b) conjunto dos regressores endógenos;
2. Estime os parâmetros da equação estrutural, instrumentalizando adequadamente os regressores endógenos e obtenha os resíduos.
3. Regrida os resíduos em função de uma constante, todas as variáveis exógenas da equação estrutural e de todos os instrumentos e calcule a estatística:
$$SARG = (n - (k+1))R^2 \sim \chi^2_{(p-q)}$$
em que p é o número de instrumentos e q é o número de regressores endógenos.
4. Rejeite H_0 (instrumentos válidos), se $SARG > \chi^2_{crit}$.

- Teste similar ao de Sargan

Suponha um modelo com somente uma variável explicativa endógena.

1. Se houver somente uma VI, não teremos restrições sobreidentificadoras. Neste caso, nada pode ser testado.

2. Se houver duas VIs, teremos uma restrição sobreidentificadora; três VIs, teremos duas restrições sobreidentificadoras...

Hipótese nula do teste:

$$H_0: \text{ Todas as VIs são não correlacionadas com o erro}$$

Passos do Teste de Restrições Sobreidenticadoras

1. Estime a equação estrutural por MQ2E.
2. Obtenha os resíduos $\hat{u}_1$;
3. Regrida $\hat{u}_1$ sobre todas as variáveis exógenas, por MQO.
4. Obtenha o R-quadrado;
5. Sob a hipótese nula de que todas as VIs são não correlacionadas com u_1:

$$nR^2 \sim \chi^2_q$$

em que q é o número de VIs menos o número de regressores endógenos presentes no modelo.