Variáveis Instrumentais e MQ2E

Capítulo 15 (Wooldridge, 2011)

Luiz Guilherme Scorzafave

USP-RP

Maio 2020

Introdução

Queremos enfrentar o problema da endogeneidade

Roteiro

Variáveis omitidas - motivação

- Além do uso da variável *proxy* para resolver o problema da variável omitida, podemos usar o método de variáveis instrumentais

Exemplo

$$ln(sal) = \beta_0 + \beta_1 educa + \beta_2 \underbrace{\text{aptidão}}_{\text{não observado}} + e$$

Se não houver *proxy* para aptidão, ela vai para o erro e:

$$ln(sal) = \beta_0 + \beta_1 educa + v$$

E se aptidão e educa forem correlacionados, EMQO será viesado e inconsistente.

- Genericamente, seja

$$y = \beta_0 + \beta_1 x + u \qquad , \qquad cov(x,u) \neq 0$$

- Para obter estimadores consistentes de β_0 e β_1 precisamos de informação adicional que virá por uma variável z, satisfazendo:

$$\left.\begin{array}{ll} (15.4) \qquad & cov(z,u) = 0 \\ (15.5) \qquad & cov(z,x) \neq 0 \end{array}\right\} \text{z é variável instrumental de } x$$

- É difícil testar (15.4), mas pode-se testar (15.5):

$$x = \pi_0 + \pi_1 z + v \qquad \text{logo: } \pi_1 = \frac{cov(z,x)}{var(z)}$$

$$\text{Então se } \pi_1 \neq 0, \text{vale } (15.5)$$

- No nosso exemplo, a variável instrumental deve ser **não correlacionada** com aptidão, e **correlacionada** com educação.
- Por exemplo, a variável 'escolaridade da mãe' é correlacionada com a escolaridade dos filhos, mas pode também ser correlacionada com a aptidão dos filhos, logo cov(z,u)=0 pode não ser válida.

Exemplo 2

$$nota = \beta_0 + \beta_1 faltas + u$$

- u é a 'qualidade' do aluno (aptidão, motivação, etc), e então $cov(u, faltas) \neq 0$
- Um instrumento factível é distância casa-campus ($dist$).
- Alunos pobres podem morar longe. Logo, se renda afeta desempenho, $cov(dist, u) \neq 0$
- Mas se incluirmos a renda na regressão, $cov(dist, u) = 0$

Identificação

- A identificação de um parâmetro significa que posso escrever β_1 em termos de momentos populacionais.
- Pode-se escrever:

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

- Logo:

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

- Dado uma amostra aleatória, podemos obter o estimador de variáveis instrumentais (VI) de β_1:

$$\hat{\beta}_{1_{VI}} = \frac{\sum_{i=1}^{n}(z_i - \overline{z})(y_i - \overline{y})}{\sum_{i=1}^{n}(z_i - \overline{z})(x_i - \overline{x})} \qquad e \qquad \hat{\beta}_{0_{VI}} = \overline{y} - \hat{\beta}_{1_{VI}}\overline{x}$$

Observações

- 1) Se z=x, então VI= MQO
- 2) $plim(\widehat{\beta_{1_{VI}}}) = \beta_1$, se cov(z,u) $= 0$ e cov(z,x) $\neq 0$

No entanto, o estimador sempre será viesado e o tamanho do viés pode ser grande em pequenas amostras

Inferência com o Estimador de VI

- O estimador de variáveis instrumentais (EVI) tem uma distribuição aproximadamente normal em grandes amostras (EVI $\sim^{ap}$ *Normal*)
- Para fazer inferência, é necessário homocedasticidade:

$$E(u^2 \mid z) = \sigma^2 = var(u)$$

- Pode ser mostrado, que[1] :

$$(15.12) \qquad \text{Avar}(\hat{\beta}_1) = \frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2} \qquad \uparrow corr(x,z), \downarrow var(\hat{\beta}_1)$$

- Dada uma amostra aleatória podemos recuperar todos fatores de (15.12):

$$\sigma_x^2 \rightarrow \text{var. amostral de x } (SQT_x); \quad \hat{\sigma^2} = \frac{1}{n-2}RSS; \quad \rho_{x,z}^2 = R_{x,z}^2$$

- Logo:

$$\widehat{\text{Avar}(\hat{\beta}_1)} = \frac{\hat{\sigma^2}}{SQT_x R_{x,z}^2}$$

[1] Avar = Variância Assintótica

- Note que como $R^2 < 1, var(\widehat{\beta^{OLS}}) < var(\widehat{\beta^{VI}})$

Exemplo -Wooldridge 15.1

$$ln(sal) = \beta_0 + \beta_1 educa + u$$

	Estimativa β_0	Estimativa β_1
MQO	-0,185 (0,185)	0,109 (0,014)
VI (educação pai como VI)	-0,441 (0,446)	0,059 (0,035)

$$\widehat{educ} = 10,24 + 0,269 educp$$

Também podem ser usadas variáveis binárias como instrumentos.

Exemplo

- Angrist (1990): efeito de ser veterano no Vietnã sobre ganhos

$$ln(ganhos) = \beta_0 + \beta_1 veterano + u$$

- Pode existir problema de auto seleção: a decisão de alistar-se pode estar correlacionada com outras características que afetem os ganhos.
- Angrist (1991) diz que o sorteio militar é um **experimento natural**, criando, portanto, um instrumento.
- Os números do sorteio não são correlacionados com u, mas são correlacionado com o fato de ser veterano, portanto $corr(sort, vet) \neq 0 \quad e \quad corr(sort, u) = 0$

Problemas do Instrumento Pobre ($cov(z,x)$ pequena)

$$plim(\widehat{\beta_{1VI}}) = \beta_1 + \frac{corr(z,u)}{corr(z,x)}\frac{\sigma_u}{\sigma_x}$$

- σ_u=Desvio Padrão de u na população
- σ_x=Desvio Padrão de x na população
- Note que mesmo se corr(z,u) seja pequena, a inconsistência pode ser grande se corr(z,x) for ainda menor. Do capítulo 5, temos:

$$plim(\widehat{\beta_{1MQO}}) = \beta_1 + corr(x,u)\frac{\sigma_u}{\sigma_x}$$

- Assim, VI é preferível a MQO com relação a consistência quando:

$$\frac{corr(z,u)}{corr(z,x)} < corr(x,u)$$

- Não precisamos preocupar com R^2 com VI, pois o objetivo de VI não é este

Roteiro

Estimação de VI no MRLM

O modelo é facilmente estendido para regressão múltipla:

Equação Estrutural

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1 \qquad (15.22)$$

- $z_1 \rightarrow$ variável exógena ($cov(z_1, u_1) = 0$)
- $y_2 \rightarrow$ variável explicativa 'suspeita' de ser endógena
- $y_1 \rightarrow$ variável endógena

Exemplo

- $z_1 =$ experiência
- $y_1 =$ salário
- $y_2 =$ educação

- Precisamos de um instrumento para $y_2 \rightarrow$ vamos chamá-la de z_2. Assim, assumimos:

$$E(u_1) = 0 \qquad cov(z_1, u_1) = 0 \qquad cov(z_2, u_1) = 0$$

Podemos obter os correspondentes amostrais desta equação

$$\sum_{i=1}^{n} (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

$$\sum_{i=1}^{n} z_{i1}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

$$\sum_{i=1}^{n} z_{i2}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0$$

- OBS: Se y_2 é exógeno, e $z_2 = y_2$, VI=OLS

- Para definir a $corr(z_2, y_2)$, escrevemos a **forma reduzida** (variável endógena em função das exógenas):

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2 \qquad (15.26)$$

- $E(v_2) = cov(z_1, v_2) = cov(z_2, v2) = 0 \qquad , \pi_2 \neq 0$

Exemplo 15.4

Card (1995) usou dummy de proximidade de uma faculdade como instrumento para educação na equação de salário e obteve $\pi_2 \neq 0$

Roteiro

1. Variáveis omitidas - motivação
   - Identificação
   - Inferência com o Estimador de VI
   - Problemas do Instrumento Pobre
2. Estimação de VI no MRLM
3. Mínimos Quadrados em dois Estágios
   - Hipóteses para que MQ2E tenha propriedades assintóticas
   - Multicolinearidade e MQ2E
4. Solução de VI para erros nas variáveis
5. Teste de Endogeneidade e de Restrições Sobreidentificadoras
   - Teste de endogeneidade de Hausman
   - Teste de restrições sobreidentificadoras
6. Forma matricial

Mínimos Quadrados em dois Estágios

- Até agora, vimos os casos em que usamos um instrumento apenas para tratar de uma variável endógena. Quando há mais de um instrumento usamos o Mínimos Quadrados em dois Estágios (MQ2E)

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1 \qquad (*)$$

- Suponha que temos duas variáveis exógenas excluídas: z_2 e z_3
- A ideia é que, a princípio, poderíamos usar z_2 ou z_3 como o instrumento para y_2 se eles fossem correlacionados com este último (e não com u_1).
- Entretanto, podemos fazer melhor: criar uma **combinação linear** de z_1, z_2 e z_3 que seja mais correlacionada com y_2

- **Forma reduzida**:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2 \qquad (15.26)$$

- $E(v_2) = cov(z_i, v_2) = 0, i = 1, 2, 3$
- A melhor VI de y_1 é y_2^*:

$$y_2^* = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3$$

$\pi_2 \neq 0$ ou $\pi_3 \neq 0$

- y_2^* é a parte de y_2 não correlacionada com u
- v_2 é a parte de y_2 possivelmente correlacionada com u

Deve-se estimar π_0, π_1, e π_2:

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

OBS: Teste F em z_2 e z_3

- Substitua $\hat{y}_2$ em y_2 em (*)
- MQ2E :
    - 1º estágio $\rightarrow$ (+)
    - 2º estágio $\rightarrow$ (*)
- Quando só há um instrumento MQ2E=VI

Exemplo 15.5

$$\underset{(0,400)}{ln(sal) = 0,048} + \underset{(0,031)}{0,061educ} + \underset{(0,013)}{0,044exper} - \underset{(0,0004)}{0,0009exper^2}$$

$$educ = \pi_0 + \pi_1 exper + \pi_2 exper^2 + \pi_3 educmae + \pi_4 educpai$$

$$\left.\begin{array}{ll} \text{H0:} & \pi_3 = \pi_4 = 0 \\ \text{H1:} & \pi_3, \pi_4 \neq 0 \end{array}\right\} \text{Teste F (est. F=55,40 p valor=0)}$$

Hipóteses para que MQ2E tenha propriedades assintóticas

- Seja
$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + ...\beta_k z_{k-1} + u_1 \qquad (15.42)$$
- Para a consistência precisamos:
  - Cada variável instrumental é não correlacionada com u_1
  - Pelo menos uma variável exógena fora de (15.42) seja correlacionada com y_2
- Para que erros padrão e estatísticas t serem válidas, precisamos de homoscedasticidade

Multicolinearidade e MQ2E

- A variância assimptótica do estimador MQ2E de β_1 pode ser aproximada como:

$$var(\widehat{\beta_{1_{MQ2E}}}) = \frac{\sigma^2}{[\widehat{SQT_2}(1 - \hat{R}_2^2)]}$$

- $\sigma^2 = var(u_1)$
- $\widehat{SQT_2}$ é a variação total em $\hat{y}_2$
- $\hat{R}_2^2$ é o R-quadrado de uma regressão de $\hat{y}_2$ sobre todas as outras variáveis exógenas que aparecem na equação estrutural.
- Por isso, uma amostra da tamanho grande ($\uparrow n$) é uma boa idéia para compensar um alto $\hat{R}_2^2$

Condições de identificação de uma equação

- Condições para identificação de uma equação:
  - **Necessária** (condição de ordem) : nº de variáveis endógenas da equação estrutural = nº de variáveis exógenas excluídas
  - **Suficiente** (condição de classificação ou de posto): quando há apenas 1 variável endógena implica que o coeficiente de uma exógena da forma reduzida deve ser $\neq 0$

Condição de posto: π_2 ou $\pi_3 \neq 0$

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

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

Roteiro

Solução de VI para erros nas variáveis

- Genericamente, seja

$$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + u \qquad y_2 \text{ e } x_2 \text{ observados e } x_1^* \text{ não observado}$$

- Seja e_1 o erro de medida. Temos que: $x_1 = x_1^* + e_1$
- Logo, caso e_1 seja correlacionado com x_1, o MQO é viesado e inconsistente:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + (u - \beta_1 e_1)$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + (u - \beta_1 e_1)$$

- Em alguns casos, podemos usar VI para resolver o problema de erro de medida
- Hipótese: e_1 é correlacionado com x_1 e **não** é com x_2 e x_1^*
- Preciso de uma VI para x_1 (que seja correlacionada com x_1 e não com u)
- Possibilidade: obter nova estimação de x_1^*, por exemplo, z_1
  - $z_1 = x_1^* + a_1 \qquad corr(a_1, e_1) = 0$
  - Como é x_1^* que afeta y, é razoável supor que $cov(z_1, u) = 0$
- Exemplo: Ashenfelter e Kruger (94): idade do irmão gêmeo

Roteiro

Teste de endogeneidade de Hausman

- MQ2E é menos eficiente que MQO quando variáveis explicativas são exógenas. Assim, é útil realizar um teste de endogeneidade de uma variável explicativa para ver se MQ2E é necessário.
- Suponha que temos y_2 como suspeita de ser endógena:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1 \qquad (15.49)$$

- z_1 e z_2 são exógenas, e z_3 e z_4 também são exógenas mas não aparecem na regressão acima.
- Se y_2 é não correlacionado com u_1, devemos estimar por MQO. Caso contrário, por VI.
- Hausman (78) sugeriu comparar MQO e MQ2E , e determinar se as diferenças são estatisticamente significantes, pois tanto MQO quanto MQ2E serão consistentes se todas as variáveis forem exógenas. Se MQO $\neq$ MQ2E, então y_2 deve ser endógeno.

- Estimação na forma reduzida:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v_2 \qquad (15.50)$$

- Como $cov(z_j, u_1) = 0$ (para todo j), então y_2 será não correlacionado com $u_1 \Leftrightarrow corr(v_2, u_1) = 0$. Queremos testar isso!
- Escreva: $u_1 = \delta_1 v_2 + e_1$
- $corr(u_1, v_2) \neq 0 \Leftrightarrow \delta_1 = 0$ (teste t)
- A maneira mais fácil para testar é incluir $\hat{v}_2$ de (15.50) como regressor em (15.49):

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

- Teste t: H0 $\delta_1 = 0$ (robusta a heterocedasticidade)
- Se rejeita H0, y_2 é endógeno pois $cov(v_2, u_1) \neq 0$

Teste de restrições sobreidentificadoras

- Se tivermos mais de uma variável instrumental, podemos testar se alguma delas é não correlacionada com o erro da equação estrutural
- Suponha que :

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

- Usamos somente z_3 como instrumento para y_2 e calculamos $\hat{u}_1$ (MQ2E).
- Verificamos se $cov(z_4, \hat{u}_1) \neq 0$. Se for, z_4 não é bom instrumento para y_2, **assumindo-se** que $cov(z_3, u_1) = 0$

Passos:

- 1) Estime modelo estrutural por MQ2E e obtenha os resíduos ($\hat{u}_1$)
- 2) Regrida $\hat{u}_1$ em todas exógenas e obtenha o R_1^2
- 3) H0: Todas variáveis instrumentais não correlacionadas com u_1, $nR_1^2 \sim^a \mathcal{X}^2_{(q)}$, onde q é o nº de VI fora do modelo - nº endógenas
- 4) Se rejeita H0, conclui-se que pelo menos alguns instrumentos **não são exógenos**

Roteiro

Forma matricial

- Seja o modelo $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$, em que $\boldsymbol{E}(\mathbf{u} \mid \mathbf{X}) \neq \mathbf{0}$ com k-1 variáveis explicativas. **X** é uma matriz (n x k), onde ***n*** é o tamanho da amostra e ***k*** o número de parâmetros a serem estimados. $\boldsymbol{\beta}$ é um vetor de dimensão (k x 1); ***y*** e ***u*** são vetores de dimensão (n x 1).
- Dispõe-se de ***m*** instrumentos, descrito pela matriz **Z**, de dimensão (n x m).
- Os instrumentos devem ser correlacionados com **X** e não correlacionados com ***u***.
- Pode-se utilizar mais de um instrumento por regressor. Além disso, quando uma variável não precisa ser instrumentalizada, ela mesma será seu próprio instrumento de modo que $\boldsymbol{m} \geq \boldsymbol{k}$

- O primeiro estágio consiste em regredir as variáveis explicativas em seus instrumentos, isto é :

$$\mathbf{X} = \mathbf{Z}\gamma + \mathbf{v} \qquad (**)$$

- Em que γ é uma matriz de parâmetros de dimensão (m x k) e $\boldsymbol{v}$ representa o termo de erro.A estimação de $(**)$ por MQO gera $\hat{\gamma} = \mathbf{(Z'Z)^{-1}Z'X}$, de onde se obtém os valores previstos $\hat{\boldsymbol{X}} = \boldsymbol{Z}\hat{\gamma} = \boldsymbol{Z(Z'Z)^{-1}Z'X}$.
- No segundo estágio, substituímos $\boldsymbol{X}$ por $\hat{\boldsymbol{X}}$ na equação estrutural e se estima essa última por MQO:

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

- Sendo que o estimador de MQ2E será dado por:

$$\hat{\beta}^{\mathbf{MQ2E}} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}$$

- Para chegar aos resultados abaixo, deve-se usar algumas regras de álgebra matricial:
- $(A')' = A$
- $(A')^{-1} = (A^{-1})'$
- $(ABC)' = C'B'A'$
- $\mathbf{A}.\mathbf{A}^{-1} = A^{-1}.A = I$
- A partir de onde é possível deduzir que:
- $[(\mathbf{Z}'\mathbf{Z})^{-1}]' = [(\mathbf{Z}'\mathbf{Z})']^{-1} = (\mathbf{Z}'\mathbf{Z})^{-1}$
- $(\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})' = \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$
- $(\mathbf{Z}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1} = (\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{Z}) = \mathbf{I}$

$$\hat{\beta}^{\mathbf{MQ2E}} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}$$

- Note que:

$$\begin{aligned}\hat{\mathbf{X}}'\hat{\mathbf{X}} = (\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})'(\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}) &= \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} \\ &= \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\end{aligned} \tag{3}$$

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

- Chega-se ao formato final do estimador de MQ2E:

$$\beta^{MQ2E} = (X'Z(Z'Z)^{-1}Z'X)^{-1}(X'Z(Z'Z)^{-1}Z'y)$$

$$\beta^{MQ2E} = (X'Z(Z'Z)^{-1}Z'X)^{-1}(X'Z(Z'Z)^{-1}Z'y)$$

- Note que se dim(Z)=dim (X) (ou k = m), ou seja, se tenho o mesmo número de variáveis explicativas endógenas e de instrumentos (caso exatamente identificado), as matrizes $(X'Z)$ e $(Z'X)$ são quadradas de dimensão (n x n) e, portanto, é verdade que:

$$(X'Z(Z'Z)^{-1}Z'X)^{-1} = (Z'X)^{-1}(Z'Z)(X'Z)^{-1}$$

- Resultado que decorre da propriedade $(\mathbf{ABC})^{-1} = C^{-1}B^{-1}A^{-1}$, fazendo $\mathbf{A} = X'Z, B = (Z'Z)^{-1}, C = (Z'X)$

$$(X'Z(Z'Z)^{-1}Z'X)^{-1} = (Z'X)^{-1}(Z'Z)(X'Z)^{-1}$$

Substituindo a equação acima em:

$$\beta^{MQ2E} = (X'Z(Z'Z)^{-1}Z'X)^{-1}(X'Z(Z'Z)^{-1}Z'y)$$

Chega-se em:

$$\beta^{VI} = (Z'X)^{-1}(Z'Z)(X'Z)^{-1}(X'Z)(Z'Z)^{-1}Z'y$$

$$\beta^{VI} = (Z'X)^{-1}(Z'Z)(Z'Z)^{-1}Z'y$$

$$\beta^{VI} = (Z'X)^{-1}(Z'y)$$

Esse é o estimador de variáveis instrumentais no caso exatamente identificado (m=k)

- Ao contrário de MQO, o estimador de VI é consistente quando há correlação entre regressores e erro.
- Contudo, não se garante não viés, de modo que tal estimador não apresenta bom desempenho em amostras pequenas.