Econometria II, Lic. em Economia

Luís Filipe Martins | Dpt. de Métodos Quantitativos | ISCTE - EG

luis.martins@iscte.pt | http://home.iscte.pt/~lfsm | Lisboa, 26/09/2005

ENDOGENEIDADE E VARIÁVEIS INSTRUMENTAIS

Wooldridge §15

0.1 Motivação e Consequências para o OLS

Uma das hipóteses clássicas ($M1 - M5$) do MRLM,

$$y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i = x_i^{'}\beta + u_i, i = 1, ..., n, \tag{1}$$

é $M2$ de exogeneidade estrita entre erros e regressores

$$E(u_i|X) = E(u_i|x_1, ..., x_n) = 0, i = 1, ..., n \Leftrightarrow E(u|X) = 0_{n\times 1}, \tag{2}$$

uma condição que implica $E(u_i) = 0$ e

$$Cov(u_i, x_{ij}) = E(u_i x_{ij}) = 0, j = 1, ..., k, i = 1, ..., n. \tag{3}$$

Portanto, quando derivámos as propriedades do OLS assumimos que, no modelo em causa, os erros e os regressores não estavam autocorrelacionados - eram exógenos. Na prática existem casos de modelizações nas quais esta hipótese pode não ser válida. Eis três exemplos de modelos com endogeneidade.

Example 1 *$Sal = \beta_1^* + \beta_2^* Exp + u$, onde Sal é salário, Exp é anos de experiência e u inclui a variável idade que está correlacionada com Exp.*

Este é um exemplo de um modelo em que se encontra omitida uma variável relevante. Quando se admite que a variável idade é estatisticamente significativa para explicar o nível de salários, o verdadeiro modelo é

$$Sal = \beta_1 + \beta_2 Exp + \beta_3 Id + u, \beta_3 \neq 0. \tag{4}$$

Como foi demonstrado num capítulo anterior,

$$\widehat{\beta}_2^* = \frac{\sum_{i=1}^n \left(Exp_i - \overline{Exp}\right) Sal_i}{\sum_{i=1}^n \left(Exp_i - \overline{Exp}\right)^2}, \tag{5}$$

em que, porque a verdadeira especificação é (4) ,

$$E\left(\widehat{\beta}_2^* | Exp, Id\right) = \beta_2 + \beta_3 \frac{\sum_{i=1}^n \left(Exp_i - \overline{Exp}\right) Id_i}{\sum_{i=1}^n \left(Exp_i - \overline{Exp}\right)^2} = \beta_2 + \beta_3 \widehat{\delta}_2 \neq \beta_2, \tag{6}$$

onde $Id = \delta_1 + \delta_2 Exp + u$. O estimador $\widehat{\beta}_2^*$ não é centrado porque $\beta_3 \neq 0$ (idade é relevante) e $\widehat{\delta}_2 \neq 0$. Este segundo resultado é verdadeiro pois $E\left(Id|Exp\right) = \delta_1 + \delta_2 Exp$, uma função de Exp! Apesar de ser enviesado, será que o estimador é consistente (corrige o bias assimptóticamente)? A resposta é negativa pois, como $Cov(Id, Exp) \neq 0$,

$$p\lim_{n\to\infty} \widehat{\beta}_2^* = \beta_2 + \beta_3 \frac{Cov(Id, Exp)}{V(Exp)} \neq \beta_2 \tag{7}$$

O sentido do bias (enviesamento) para amostras finitas e assimptóticamente depende do sinal da correlação na amostra e nas variáveis, respectivamente, para além do sinal de β_3. Este podemos tomar como positivo, $\beta_3 > 0$, em que maiores salários são, em média, pagos a trabalhadores de maior idade. Por outro lado podemos assumir que $\sum_{i=1}^n \left(Exp_i - \overline{Exp}\right) Id_i$ e $Cov(Id, Exp)$ são de sinal positivo - experiência aumenta com a idade. Consequentemente, os enviesamentos (finito ou assimptótico) são de sinal positivo,

$$\beta_3 \frac{\sum_{i=1}^n \left(Exp_i - \overline{Exp}\right) Id_i}{\sum_{i=1}^n \left(Exp_i - \overline{Exp}\right)^2} > 0; \beta_3 \frac{Cov(Id, Exp)}{V(Exp)} > 0, \tag{8}$$

o que significa que $\widehat{\beta}_2^*$ sobreavalia o verdadeiro valor $\widehat{\beta}_2$.

Example 2 *$Sav = \beta_1 + \beta_2 Inc^* + u$, onde Sav é poupança e Inc^* é rendimento disponível mas que é medido com erro: $e = Inc - Inc^*, E(e) = 0, V(e) = \sigma_e^2$, onde Inc é o verdadeiro valor.*

Este é um exemplo de erro de medida na variável independente em que o verdadeiro modelo é

$$Sav = \beta_1 + \beta_2 \left(Inc - e\right) + u = \beta_1 + \beta_2 Inc + \left(u - \beta_2 e\right), \tag{9}$$

onde

$$Cov(u - \beta_2 e, Inc) = Cov(u, Inc) + Cov(-\beta_2 e, Inc) = -\beta_2 \sigma_e^2 \neq 0, \tag{10}$$

assumindo que $Cov(u, Inc) = 0$ e $Cov(e, Inc^*) = 0$:

$$Cov(e, Inc) = E(e.Inc) = E(e^2) - E(e.Inc^*) = \sigma_e^2. \tag{11}$$

Foi provado anteriormente que $\widehat{\beta}_2$ é inconsistente:

$$p\lim_{n\to\infty} \widehat{\beta}_2 = \beta_2 + \frac{Cov(u - \beta_2 e, Inc)}{V(Inc)} = \beta_2 \left(\frac{\sigma_{Inc^*}^2}{\sigma_{Inc^*}^2 + \sigma_e^2}\right) < \beta_2. \tag{12}$$

Example 3 $\begin{cases} q^d = \beta p + u \\ q^s = \delta p + e \end{cases}$, *onde, para um dado mercado (bens, serviços, trabalho,...), p é preço, q^d é a quantidade procurada, q^s é a quantidade oferecida.*

Este é um exemplo de um modelo de equações simultâneas. Como a quantidade observada é determinada em equilibrio de procura e oferta, nenhuma das curvas (procura (β) e oferta (δ)) é identificada. Se $q^s = \delta p + \gamma z + e$, z é uma variável exógena, então a oferta não é identicada (não é possivel estimar δ, γ) mas a procura é identificável (é possível estimar β). De qualquer forma, a estimação de β em $q^d = q = \beta p + u$ por OLS é inconsistente:

$$p = \frac{q - \gamma z - e}{\delta} = \frac{\beta p + u - \gamma z - e}{\delta} \Leftrightarrow p = \pi z + \eta (u - e),$$

o que implica na equação de procura $E(u|p) = E(u|\pi z + \eta (u - e)) \neq 0$. De uma forma semelhante, $Cov(u, p) = E(up) = E(u[\pi z + \eta (u - e)]) \neq 0$. Portanto, o OLS é enviesado e inconsistente.

O que estes exemplos nos permitem concluir é que quando existe endogeneidade no modelo o OLS é enviesado e inconsistente. Este resultado não é de estranhar visto que a hipótese $M2$ (que implica $M2'$) é uma condição necessária para que o OLS seja centrado (e, consequentemente, consistente). Esta é uma situação mais "grave" para o OLS quando comparada com os casos de heterocedasticidade e autocorrelação dos erros. Para além destes problemas, e como consequência, a inferência com o OLS standard torna-se errónea; não faz sentido falar em estimação robusta de $V(\widehat{\beta})$ e o estimador não é BLUE (não é U - unbiased!), nem assimptóticamente. Nestas condições, importa definir um método de estimação alternativo ao OLS. O método é de variáveis instrumentais (ou 2SLS) e será discutido nas próximas secções.

0.2 Variáveis Instrumentais e o Estimador IV

Tal como o OLS e o MLE, o estimador IV (variáveis instrumentais) pertence à classe de estimadores GMM. O IV é um GMM linear e a sua utilização justifica-se principalmente em modelos com endogeneidade,

$$Cov(u_i, x_{ij}) = E(u_i x_{ij}) \neq 0, \text{ para algum } j = 1, ..., k \text{ e } i = 1, ..., n, \tag{13}$$

onde se mantém a hipótese de $E(u) = 0$. Na estimação IV é necessário recorrer ao uso de variáveis instrumentais a(s) qual(is) estão associadas a um particular regressor endogeno.

Definition 1 *z é uma variável instrumental para x se (i) z é exogeno em relação ao erro $u, Cov(u, z) = E(uz) = 0$; e (ii) z está (fortemente) correlacionado com o regressor $x, Cov(x, z) \neq 0$*

Example 4 *Anos de experiência do pai pode ser um instrumento para Exp. Nível de escolaridade dos pais e número de irmãos/irmãs podem ser instrumentos para Educ. Teste IQ pode ser*

um instrumento para talento/aptidão. Salário pode ser um instrumento para Inc^*. No mercado agrícola, o número de dias de chuva pode ser uma variável exogena na equação da oferta.

Para que um instrumento z para x seja estatisticamente válido, temos de verificar ambas condições $Cov(u,z)=0$ e $Cov(x,z)\neq 0$. Em relação à primeira, e como não observamos u, partimos apenas de pressupostos da teoria económica ou senso comum. A análise da correlação amostral entre z e os residuos $\widehat{u}$ é um tanto quanto simplista e não rigorosa. Para verificar que $Cov(x,z)\neq 0$ e qual a intensidade da relação entre eles, uma das possibilidades é usar um teste t na regressão auxiliar

$$x=\delta_1+\delta_2 z+u;\delta_2=\frac{Cov\left(x,z\right)}{V(z)}. \tag{14}$$

Após a definição de variável intrumental é muito simples a expressão do estimador IV. No MRLS,

$$y=\beta_1+\beta_2 x+u, Cov(u,x)=E(ux)\neq 0, \tag{15}$$

para uma amostra de dimensão n e um único instrumento para x,

$$\widehat{\beta}_{2,IV}=\frac{\sum_{i=1}^{n}\left(z_i-\overline{z}\right)\left(y_i-\overline{y}\right)}{\sum_{i=1}^{n}\left(z_i-\overline{z}\right)\left(x_i-\overline{x}\right)}=\beta_2+\frac{\sum_{i=1}^{n}\left(z_i-\overline{z}\right)u_i}{\sum_{i=1}^{n}\left(z_i-\overline{z}\right)\left(x_i-\overline{x}\right)}. \tag{16}$$

Naturalmente, $\widehat{\beta}_{1,IV}=\overline{y}-\widehat{\beta}_{2,IV}\overline{x}$. Ao contrário do OLS e GLS, o IV não resulta de minimização de quadrados de erros. A sua expressão pode no entanto ser justificada pelo método dos momentos: de (15),

$$\begin{aligned} Cov(y,z) &= Cov(\beta_1+\beta_2 x+u,z)=\beta_2 Cov(x,z)+Cov(u,z)=\beta_2 Cov(x,z)\Leftrightarrow \\ \beta_2 &= \frac{Cov(y,z)}{Cov(x,z)}. \end{aligned} \tag{17}$$

Alternativamente, resulta dos momentos $E\left(u\right)=0$ e $E(uz)=0$. Quando a variável instrumental é válida, o estimador IV é consistente, $p\lim\limits_{n\to\infty}\widehat{\beta}_{IV}=\beta$[1]. Mas vejamos o seguinte. Num modelo com endogeneidade, $Corr(u,x)\neq 0$,

$$p\lim_{n\to\infty}\widehat{\beta}_{2,OLS} = \beta_2+Corr(u,x)\frac{\sigma_u}{\sigma_x}\neq\beta_2, \tag{18}$$

$$p\lim_{n\to\infty}\widehat{\beta}_{2,IV} = \beta_2+\frac{Corr(u,z)}{Corr(x,z)}\frac{\sigma_u}{\sigma_x}=\beta_2, \tag{19}$$

porque o instrumento é válido, $Corr(u,z)=0$ e $Corr(x,z)\neq 0$. Mas, se o instrumento não é válido, no sentido de que $Corr(u,z)\neq 0$, então o IV também é inconsistente, $p\lim\limits_{n\to\infty}\widehat{\beta}_{IV}\neq\beta$. O IV tem um menor enviesamento assimptótico do que o OLS se $Corr(u,x)>\frac{Corr(u,z)}{Corr(x,z)}$, condição

[1] Nada pode ser dito em relação a amostras de dimensão finita pois a hipótese que assumimos é de $Cov(z,u)=0$. Estudos de Monte Carlo podem mostrar que existe um enviesamento para pequenas amostras.

esta que não é fácil de testar a não ser de uma forma intuitiva em que se calculam as correlações amostrais onde u é representado pelos residuos $\widehat{u}$.

Sob a hipótese de $E(u^2|z) = \sigma^2 = V(u)$,

$$V\left(\widehat{\beta}_{2,IV}|\mathbf{x},\mathbf{z}\right) = \frac{\sigma^2}{n\sigma_x^2 Corr(x,z)^2}, \tag{20}$$

que pode ser consistentemente estimada por

$$\widehat{V\left(\widehat{\beta}_{2,IV}|\mathbf{x},\mathbf{z}\right)} = \frac{\widehat{\sigma}^2}{n\widehat{\sigma}_x^2 R_{x,z}^2}, \tag{21}$$

onde $\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n \widehat{u}_i^2$ e $\widehat{u}_i$ são os residuos IV, $\widehat{u}_i = y_i - \widehat{\beta}_{1,IV} - \widehat{\beta}_{2,IV}x_i$, $\widehat{\sigma}_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2$, e $R_{x,z}^2$ resulta do modelo (14). Portanto, para n elevado,

$$se\left(\widehat{\beta}_{2,IV}\right) = \sqrt{\frac{\widehat{\sigma}^2}{SST_x R_{x,z}^2}} = \frac{\widehat{\sigma}\sqrt{\sum_{i=1}^n (z_i - \overline{z})^2}}{\sum_{i=1}^n (z_i - \overline{z})(x_i - \overline{x})}, \tag{22}$$

que coincide com a expressão para o estimador inconsistente OLS $\sqrt{\frac{\widehat{\sigma}^2}{SST_x}}$ se $R_{x,z}^2 = 1$, isto é, se x e z são linearmente dependentes na amostra! Em geral, isso não acontece, e porque $0 < R_{x,z}^2 < 1$, o IV é menos efficiente (maior *se* do que o OLS). Mas, não nos esqueçamos que, ao contrário do IV, o OLS é inconsistente sob a hipótese de endogeneidade. Quanto maior é a correlação entre x e z (melhor qualidade do instrumento), maior é a precisão na estimação IV do modelo.

Passamos agora à análise geral no MRLM,

$$y_i = x_i'\beta + u_i, i = 1, ..., n, \Leftrightarrow y = X\beta + u, \tag{23}$$

em que

$$Cov(u, X) = E(X'u) = p\lim_{n\to\infty}\frac{X'u}{n} = p\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^n x_i u_i\right) \neq 0_{k\times 1}. \tag{24}$$

Suponhamos em primeiro lugar que existem tantos instrumentos como regressores,

$$Z_{n\times k} = \begin{pmatrix} z_1' \\ ... \\ z_n' \end{pmatrix} = \begin{pmatrix} z_{11} & ... & z_{1k} \\ ... & & ... \\ z_{n1} & ... & z_{nk} \end{pmatrix}; Z_{k\times n}' = (z_1, ..., z_n); \tag{25}$$

$$p\lim_{n\to\infty}\frac{Z'u}{n} = p\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^n z_i u_i\right) = 0_{k\times 1}. \tag{26}$$

Nestas condições, (23) é equivalente a $Z^{'}y = Z^{'}X\beta + Z^{'}u$ e por isso, supondo que a matriz $Z^{'}X = \sum_{i=1}^{n} z_i x_i^{'}$ não é singular,

$$p\lim_{n\to\infty}\frac{Z^{'}y}{n} = \left(p\lim_{n\to\infty}\frac{Z^{'}X}{n}\right)\beta + p\lim_{n\to\infty}\frac{Z^{'}u}{n} \Leftrightarrow \beta_{k\times 1} = \left(p\lim_{n\to\infty}\frac{Z^{'}X}{n}\right)^{-1}.p\lim_{n\to\infty}\frac{Z^{'}y}{n}, \quad (27)$$

$$\widehat{\beta}_{IV} = \left(\frac{Z^{'}X}{n}\right)^{-1}\frac{Z^{'}y}{n} = \left(Z^{'}X\right)^{-1}Z^{'}y = \left(\sum_{i=1}^{n} z_i x_i^{'}\right)^{-1}\sum_{i=1}^{n} z_i y_i \quad (28)$$

$$= \beta + \left(Z^{'}X\right)^{-1}Z^{'}u = \beta + \left(\sum_{i=1}^{n} z_i x_i^{'}\right)^{-1}\sum_{i=1}^{n} z_i u_i. \quad (29)$$

O IV é consistente,

$$p\lim_{n\to\infty}\widehat{\beta}_{IV} = \beta + p\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i^{'}\right)^{-1}.p\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} z_i u_i = \beta, \quad (30)$$

e, sob $V(u|Z) = \sigma^2 I_n$, tem como matriz consistente de variâncias-covariâncias,

$$\widehat{\Sigma}_{k\times k} = V\left(\widehat{\widehat{\beta}_{IV}|X,Z}\right) = E\left(\left(\widehat{\beta}_{IV}-\beta\right)\widehat{\left(\widehat{\beta}_{IV}-\beta\right)}^{'}|X,Z\right) \quad (31)$$

$$= \widehat{\sigma}^2\left(Z^{'}X\right)^{-1}\left(Z^{'}Z\right)\left(X^{'}Z\right)^{-1} = \widehat{\sigma}^2\left(\sum_{i=1}^{n} z_i x_i^{'}\right)^{-1}\left(\sum_{i=1}^{n} z_i z_i^{'}\right)\left(\sum_{i=1}^{n} x_i z_i^{'}\right)^{-1}; \quad (32)$$

$$\widehat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{n}\widehat{u}_{i,IV}^2. \quad (33)$$

Note que se a amostra é *i.i.d.*, pela (U)LLN, $p\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i^{'}\right) = E\left(z_1 x_1^{'}\right)$. Se $V(u|Z) = \sigma^2\Omega$, então

$$\Sigma = \sigma^2\left(Z^{'}X\right)^{-1}\left(Z^{'}\Omega Z\right)\left(X^{'}Z\right)^{-1}. \quad (34)$$

Pode-se provar que $\widehat{\Sigma}$ é "menor" (maior eficiencia) quando "X e Z" estão mais correlacionados e que o IV é, em geral, menos eficiente do que o OLS. As matrizes de variâncias-covariâncias (e o próprio estimador) do IV e do OLS coincidem quando $Z = X$. Mas o OLS é inconsistente! Assimptóticamente, e sob algumas condições de regularidade, $\sqrt{n}\left(\widehat{\beta}_{IV}-\beta\right)$ é normalmente distribuido.

Na prática, no entanto, o modelo pode ser sobre-identificado: O número de intrumentos m é superior ao número de regressores k,

$$Z_{n\times m} = \begin{pmatrix} z_1^{'} \\ ... \\ z_n^{'} \end{pmatrix} = \begin{pmatrix} z_{11} & ... & z_{1m} \\ ... & & ... \\ z_{n1} & ... & z_{nm} \end{pmatrix}; Z_{m\times n}^{'} = (z_1, ..., z_n); m > k; \quad (35)$$

$$p\lim_{n\to\infty}\frac{Z^{'}u}{n} = p\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} z_i u_i\right) = 0_{m\times 1}. \quad (36)$$

Quando $m > k$, a matriz $Z'X$ de dimensão $m \times k$ não é quadrada e por isso não admite uma inversa. Neste caso, pré-multiplicamos (23) por $X'Z\left(Z'Z\right)^{-1}Z'$ para obter

$$\widehat{\beta}_{IV} = \left(X'Z\left(Z'Z\right)^{-1}Z'X\right)^{-1}X'Z\left(Z'Z\right)^{-1}Z'y; \tag{37}$$

$$\widehat{\Sigma} = \widehat{\sigma}^2\left[\left(X'Z\right)\left(Z'Z\right)^{-1}\left(Z'X\right)\right]^{-1}. \tag{38}$$

- Escolha do número de instrumentos. Escolha dos instrumentos.

0.3 Estimador 2SLS

Nesta secção apresentamos um método de estimação alternativo que se baseia em minimos quadrados a dois passos (2SLS). Para efeitos de simplicidade na exposição consideremos o modelo estrutural

$$y = \beta_1 + \beta_2 x + \beta_3 z + u; Cov(u,x) \neq 0, Cov(u,z) = 0, \tag{39}$$

em que x é endógeno, z é exógeno e z_1, z_2 são instrumentos válidos para x. Num primeiro passo, estima-se consistentemente por OLS o modelo (v satisfaz as hipóteses standard)

$$x = \delta_1 + \delta_2 z + \delta_3 z_1 + \delta_4 z_2 + v; \delta_3 \neq 0, \delta_4 \neq 0, \tag{40}$$

em que x tem como regressores todas as variáveis exógenas (instrumentos e do modelo estrutural). Como a combinação linear de todas as variáveis exógenas $E\left(x|z,z_1,z_2\right) = \delta_1 + \delta_2 z + \delta_3 z_1 + \delta_4 z_2$ é o "melhor" instrumento para x, este é estimado por

$$E\left(x|\widehat{z,z_1},z_2\right) = \widehat{x} = \widehat{\delta}_1 + \widehat{\delta}_2 z + \widehat{\delta}_3 z_1 + \widehat{\delta}_4 z_2. \tag{41}$$

No segundo passo, estima-se consistentemente por OLS o modelo estrutural (39)

$$y = \beta_1 + \beta_2 \widehat{x} + \beta_3 z + u, \tag{42}$$

onde $\widehat{x}$ é determinado no primeiro passo e o estimador resultante para β coincide com o IV. No entanto, os *se* para o 2SLS não são os mesmos que do IV (22) .

E para o MRLM, numa forma mais geral? Como sabemos, o estimador 2SLS coincide com o estimador IV (ver (27) e (37)). A ideia é a seguinte: No modelo

$$y = X_1\beta_1 + X_2\beta_2 + u, X = \left(X_1, X_2\right), \beta' = \left(\beta_1', \beta_2'\right), \tag{43}$$

em que X_1 é o conjunto de variáveis endógenas,

$$Cov(u, X_1) = E(X_1'u) = p\lim_{n\to\infty}\frac{X_1'u}{n} \neq 0. \tag{44}$$

No primeiro passo, do sistema $X_1 = Z\pi + v$, resulta o OLS $\widehat{\pi} = \left(Z'Z\right)^{-1} Z'X_1$ e $\widehat{X}_1 = Z\widehat{\pi} = Z\left(Z'Z\right)^{-1} Z'X_1$. Fazendo a substituição no modelo original, e após alguma cálculos, o 2SLS (37) é o OLS em

$$y = \left(\widehat{X}_1 + \widehat{v}\right)\beta_1 + X_2\beta_2 + u \Leftrightarrow y = Z\left(Z'Z\right)^{-1} Z'X_1\beta_1 + X_2\beta_2 + u^*. \tag{45}$$

A expressão do 2SLS é dada em (37). Uma condição necessária é que o número de instrumentos (colunas de Z) seja igual ou superior ao número total de regressores $(X_1, X_2), k$. Alternativamente, podia-se tomar $y = X\beta + u$ em que no primeiro passo $\widehat{\pi} = \left(Z'Z\right)^{-1} Z'X$ e $\widehat{X} = Z\left(Z'Z\right)^{-1} Z'X$ e o 2SLS era o OLS em $y = \widehat{X}\beta + v$, $\widehat{\beta} = \left(\widehat{X}'\widehat{X}\right)^{-1}\widehat{X}'y$.

0.4 Testes: Endogeneidade e Sobre-Identificação

Tal como apresentámos nos capítulos de heterocedasticidade e autocorrelação dos erros, torna-se necessário nesta discussão apresentar testes à exogeneidade dos regressores. No caso de mais do que um instrumento por regressor endogeno, também podemos testar a validade de algum dos instrumentos no sentido de não estarem correlacionados com os erros.

O teste de Hausman é o procedimento de referência para inferir sobre a exogeneidade dos regressores. Podendo ser utilizado noutro contexto, o teste de Hausman procura comparar estatisticamente dois estimadores $\widehat{\beta}_E$ e $\widehat{\beta}_C$ para o mesmo modelo e vector de parâmetros β. O estimador $\widehat{\beta}_C$ é consistente sob ambas as hipóteses H_0 e H_1 enquanto que o estimador $\widehat{\beta}_E$ apenas é consistente sob a nula H_0. Por outro lado, sob a hipótese nula $H_0, \widehat{\beta}_E$ é (assimptóticamente) mais eficiente do que $\widehat{\beta}_C$. No contexto deste capítulo, onde H_0 é exogeneidade e H_1 é endogeneidade, $\widehat{\beta}_C$ é o IV e $\widehat{\beta}_E$ é o OLS.

Vejamos o teste de Hausman com o recurso a regressão auxiliar. Porque $Cov(u, x) \neq 0$, os erros u e v dos modelos (39) e (40) estão autocorrelacionados, $Cov(u, v) \neq 0$. O teste à exogeneidade é muito simplesmente o teste rácio t ao coeficiente δ na regressão auxiliar (aumentada em relação ao modelo estrutural em $\delta\widehat{v}$)

$$y = \beta_1 + \beta_2 x + \beta_3 z + \delta\widehat{v} + u, \tag{46}$$

onde $\widehat{v}$ são os resíduos do primeiro passo. Se existirem mais do que uma variável endogena (por exemplo, 5) então usa-se o teste F à significância conjunta dos 5 coeficientes associados aos 5 resíduos do primeiro passo.

A fórmula geral da estatística de Hausman é dada por

$$H = n\left(\widehat{\beta}_C - \widehat{\beta}_E\right)'\left(\widehat{AV\left(\widehat{\beta}_C\right)} - \widehat{AV\left(\widehat{\beta}_E\right)}\right)^{-1}\left(\widehat{\beta}_C - \widehat{\beta}_E\right), \tag{47}$$

onde $\widehat{AV\left(\widehat{\beta}\right)}$ é a variância-covariância assimptótica, em que sob a hipótese nula é assimptóticamente

distribuida χ^2_k. No nosso caso de interesse,

$$H = \left(\widehat{\beta}_{IV} - \widehat{\beta}_{OLS}\right)' \left(V\left(\widehat{\widehat{\beta}_{IV}}\right) - V\left(\widehat{\widehat{\beta}_{OLS}}\right)\right)^{-1} \left(\widehat{\beta}_{IV} - \widehat{\beta}_{OLS}\right) \xrightarrow{d} \chi^2_k, \tag{48}$$

onde $\widehat{\beta}_{IV}, \widehat{\beta}_{OLS}, V\left(\widehat{\widehat{\beta}_{IV}}\right), V\left(\widehat{\widehat{\beta}_{OLS}}\right)$ foram definidos anteriormente. Para testar se um particular regressor x, que tem associado o parâmetro β_2 no modelo, não é endógeno

$$H_0 : Cov(u, x) = E(ux) = 0 \Leftrightarrow p\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} x_i u_i\right) = 0, \tag{49}$$

pode-se utilizar a estatística

$$\frac{\left(\widehat{\beta}_{2,IV} - \widehat{\beta}_{2,OLS}\right)^2}{V\left(\widehat{\widehat{\beta}_{2,IV}}\right) - V\left(\widehat{\widehat{\beta}_{2,OLS}}\right)} \xrightarrow{d} \chi^2_1. \tag{50}$$

Para finalizar, apresentamos um teste à não autocorrelação de erros e instrumentos onde o numero de instrumentos é pelo menos dois. Num contexto IV ou GMM linear, procura-se testar a hipótese nula de $H_0 : E(uZ) = 0, Z_{m\times 1}$ em que os $m > 1$ instrumentos para x não estão autocorrelacionados com u. Se H_0 não é aceite então concluimos que pelo menos um dos instrumentos não é válido e está correlacionado com os erros u. O modelo diz-se sobre-identificado pois $m > 1$. No caso de exacta identificação, $m = 1$, não é possivel desenvolver este teste[2].

O teste é similar ao anterior de Hausman em que se recorre a residuos. Neste caso, no entanto, a estatística de teste é nR^2 em que o R^2 resulta do modelo auxiliar,

$$\widehat{u} = \delta_1 + \delta_2 z + \pi_1 z_1 + \pi_2 z_2 + ... + \pi_m z_m + e, \tag{51}$$

onde $\widehat{u}$ são os resíduos IV do modelo estrutural e $z_1, z_2, ..., z_m$ são os (possiveis) instrumentos para x. A estatistica é assimptóticamente distribuida como χ^2_{m-1}. Na prática, pode-se estimar o modelo por IV com um instrumento existente e depois testar a introdução de **um** novo instrumento em que $m - 1 = 1$. De uma forma mais geral, pode-se testar para r regressores endogenos a introdução de q instrumentos adicionais no modelo. A estatistica seria distribuida como χ^2_q, onde q é a diferença entre o numero de instrumentos disponiveis (fora do modelo estrutural) e o numero de regressores endógenos r.

Aplicações

(...)

[2]Na verdade, o GMM linear considera $E(u(\beta).Z) = 0_{m\times 1}$ onde $m \geq k$ (numero de instrumentos é maior ou igual ao número de parâmetros).

Exercícios

1. Discuta a questão de endogeneidade e a escolha de instrumentos no modelo $Nota = \beta_1 + \beta_2 Faltas + u$, em que $Nota$ é a nota do exame e $Faltas$ é o numero de aulas que o aluno faltou.

2. Provar que o 2SLS no modelo $y = \beta_1 + \beta_2 x + u$ e apenas z é instrumento para x coincide com o IV (16) .

3. ...