Objetivos

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

- ✓ **Escolher e utilizar um método de estimação adequado para os parâmetros de um modelo de regressão que apresente regressores endógenos.**

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

Aula 15b

Leitura DETALHADA: Wooldridge, 2010 (4ª edição) – Capítulo 15

Introdução

Ao longo dos próximos *slides* será mostrado como a disponibilidade de uma variável instrumental poderá ser utilizada para estimar de forma consistente os parâmetros do modelo de regressão de interesse, na presença de regressor endógeno.

Particularmente, mostraremos que sob as suposições (a) e (b), vistas na aula passada, conseguiremos identificar os parâmetros da equação estrutural de interesse.

Introdução

Problema de identificação

Por problema de identificação entendemos a possibilidade de recuperar, ou não, os parâmetros da equação estrutural (ou seja, aquela que retrata a estrutura de uma economia ou o comportamento de um agente econômico) a partir dos coeficientes estimados na forma reduzida.

Introdução

Forma Reduzida

Uma equação na forma reduzida é aquela que expressa uma variável endógena apenas em termos das variáveis exógenas e dos termos de erros estocásticos.

<u>Observação</u>: Essa nomenclatura é derivada dos modelos de equações simultâneas.

Introdução

Problema de identificação (cont.)

- **Se a recuperação dos parâmetros estruturais puder ser feita, com base nos parâmetros da forma reduzida, então dizemos que a equação estrutural em pauta é identificada.**

- **Caso a recuperação não possa ser concretizada, então a equação estrutural em pauta é dita não identificada (ou subidentificada).**

Introdução

Problema de identificação (cont.)

Quando identificada, uma equação estrutural pode ser exatamente identificada ou superidentificada.

Exemplo

Considere o seguinte modelo de regressão linear múltipla:

$$\ln(salario) = \beta_0 + \beta_1 \mathrm{i}dade + \beta_2 educ + \varepsilon \qquad \textbf{(a)}$$

Aqui, admita que *idade* seja exógena e *educ* endógena;

A equação anterior é chamada de equação estrutural, para enfatizar que estamos interessados nos β_j, o que simplesmente significa que a equação supostamente mede uma relação causal.

Ainda, é sabido que se os parâmetros de (a) forem estimados por MQO, então teremos problemas de viés e inconsistência nos estimadores.

Exemplo

Todavia, admita que tenha sido coletada a informação sobre a escolaridade das mães desses indivíduos (*educM*).

Também, admita que *educM* seja exógena em (a) e que *educM* seja correlacionada com *educ*.

Ou seja, *educM* é um instrumento para *educ*.

A maneira mais fácil de definir a condição de existência de correlação parcial entre *educM* e *educ* é escrevendo a variável explicativa endógena como uma função linear das variáveis exógenas e de um termo de erro.

Ou seja,

$$educ = \pi_0 + \pi_1 \text{idade} + \pi_2 educM + \upsilon \qquad \textbf{(b)}$$

Exemplo

A equação descrita em (b) é chamada forma reduzida, uma vez que expressa uma variável endógena apenas em termos das variáveis exógenas e do termo de erro estocástico.

A condição de identificação fundamental é que

$$\pi_2 \neq 0 \qquad \textbf{(c)}$$

Para testar (c), basta estimar os parâmetros de (b) por MQO e aplicar um teste t (possivelmente robusto à heterocedasticidade).

Como não podemos testar se *educM* é não correlacionado com ε em (a), então uma argumentação bem fundamentada deverá ser pensada.

Exemplo

Qual a utilidade de (b)?

Uma maneira de pensar (b) é notando que ela divide a variável endógena em duas partes e, uma vez obtida a parte ajustada do regressor endógeno, que é justamente a sua porção exógena, usamos tal informação para estimarmos adequadamente os parâmetros da equação estrutural, que é a equação que guarda as relações de interesse.

Voltando ao Exemplo

Admita, agora, que tenham sido coletadas as informações sobre a escolaridade das mães desses indivíduos (*educM*) e sobre a escolaridade dos pais desses indivíduos (*educP*).

Ainda, admita que *educM* e *educP* sejam exógenas em (a) e que *educM* e *educP* sejam correlacionadas com *educ*. Ou seja, *educM* e *educP* são instrumentos para *educ*.

Os regressores *educM* e *educP* são também conhecidos como restrições de exclusão.

Voltando ao Exemplo

A princípio poderíamos usar *educM* ou *educP* para instrumentalizar o regressor endógeno e, nesse caso, teríamos dois estimadores de IV, todavia nenhum deles seria, de forma geral, o mais eficiente. Assim sendo, o melhor seria usarmos as duas variáveis instrumentais.

Dessa forma, a condição de existência de correlação parcial poderia ser investigada pela forma reduzida dada por:

$$educ = \pi_0 + \pi_1 idade + \pi_2 educM + \pi_3 educP + \upsilon \qquad \text{(d)}$$

A condição de identificação fundamental, agora, é dada por

$$\pi_2 \neq 0 \quad \text{ou} \quad \pi_3 \neq 0 \qquad \text{(e)}$$

Voltando ao Exemplo

Do *slide* anterior, a equação estrutural (a):

- ✓ **não estará identificada se**

$$\pi_2 = \pi_3 = 0$$

- ✓ **estará exatamente identificada se, apenas,**

$$\pi_2 \neq 0 \quad \textbf{ou} \quad \pi_3 \neq 0$$

- ✓ **estará sobreidentificada se**

$$\pi_2 \neq 0 \qquad \textbf{e} \qquad \pi_3 \neq 0$$

Voltando ao Exemplo

Para testar (e), basta estimar os parâmetros de (d) por MQO e aplicar um teste F (possivelmente robusto à heterocedasticidade).

Como não podemos testar se *educM* e *educP* são não correlacionados com ε em (a), então uma argumentação bem fundamentada deverá ser pensada.

Qual a utilidade de (d)?

Uma maneira de pensar (d) é notando que ela divide a variável endógena em duas partes e, uma vez obtida a parte ajustada do regressor endógeno, que é justamente a sua porção exógena, usamos tal informação para estimarmos adequadamente os parâmetros da equação estrutural.

MÍNIMOS QUADRADOS EM 2 ESTÁGIOS (2SLS)

2SLS

Considere o modelo de regressão linear múltipla escrito na forma linear geral

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

- ✓ **1o. Estágio: Regredir cada regressor endógeno do modelo original em função dos instrumentos (estimação das formas reduzidas);**

- ✓ **2o. Estágio: Estimar os parâmetros do modelo de interesse, utilizando os regressores ajustados obtidos no estágio anterior.**

MROZ (1987)

Aplicação

Olhando apenas as mulheres que encontram-se no mercado de trabalho, considere o seguinte modelo de regressão linear múltipla:

$$\ln(wage) = \beta_0 + \beta_1 \text{exper} + \beta_2 \text{exper}^2 + \beta_3 educ + \varepsilon$$

a) A equação de interesse encontra-se identificada? Justifique adequadamente a sua resposta.

b) Você diria que MQO é um método de estimação adequado para estimar os parâmetros da equação de interesse? Justifique.

c) Utilizando a base de dados *MROZ.xls*, estime os parâmetros do modelo de interesse, via MQO. Interprete a estimativa associada a β_3.

Aplicação

d) **Usando** ***educação da mãe*** **como instrumento para** ***educ*****, escreva a forma reduzida para** ***educ*****.**

e) **Estime os parâmetros da forma reduzida, via MQO, e teste a relevância da variável** ***educação da mãe*****. Ainda, discuta as implicações do resultado obtido, em termos da identificação (exata ou sobre) ou não da equação de interesse. Adote α = 5%.**

f) **Usando** ***educação do pai*****,** ***educação da mãe*** **e** ***educação do marido*** **como instrumentos para** ***educ*****, escreva a forma reduzida para** ***educ*****.**

Aplicação

g) **Estime os parâmetros da forma reduzida, via MQO, e teste a relevância da *educação do pai*, *educação da mãe* e *educação do marido*. Ainda, discuta as implicações do resultado obtido, em termos da identificação (exata ou sobre) ou não da equação de interesse. Adote $\alpha = 5\%$.**

h) **Baseando-se nos resultados obtidos no item (g), você estimaria os parâmetros da equação de interesse por MQO, IV ou 2SLS? Justifique.**

i) **Levando em consideração o que foi discutido em (h), estime os parâmetros da equação de interesse. Comente os resultados obtidos.**

j) **As estimativas dos parâmetros em (i) diferiram muito quando comparadas àquelas obtidas em (c)? Comente.**