OBJETIVOS

O grupo de *slides*, a seguir, objetiva responder às seguintes perguntas:

- ✓ **Qual é a natureza dos regressores endógenos?**
- ✓ **O que acontece com as propriedades dos estimadores de MQO quando incluímos regressores endógenos ao modelo de regressão de interesse?**
- ✓ **O que é uma variável instrumental e qual a sua utilidade?**

Insper | Instituto de Ensino e Pesquisa

ENDOGENEIDADE

Aula 15

Wooldridge, 2011 – Seção 9.4 e Capítulo 15

Endogeneidade

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

Problema:

$$\operatorname{plim}\left(\frac{1}{n}\underset{\sim}{\mathbf{X}}'\underset{\sim}{\varepsilon}\right) \neq \underset{\sim}{\mathbf{0}}$$

Endogeneidade

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

- ***Forma funcional especificada incorretamente;***
- ***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 II)**
- ***Simultaneidade entre y e*** x_1, x_2, ...***ou*** x_k.

Endogeneidade

A presença de regressores endógenos num modelo de regressão viola a MLR.4.

Ou seja, viola a suposição de que

$$\mathrm{E}\left(\underset{\sim}{\varepsilon} \,\middle|\, \underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_k \right) = \underset{\sim}{\mathbf{0}}$$

Sob 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

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:

$$nota_i = \beta_0 + \beta_1 faltas_i + \varepsilon_i$$

Qual motivo nos levaria a desconfiar da violação da premissa:

$$\text{plim}\left(\frac{1}{n}\underset{\sim}{\mathbf{faltas}}'\underset{\sim}{\boldsymbol{\varepsilon}}\right) \neq 0 \text{ ?}$$

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

VARIÁVEIS INSTRUMENTAIS (IV)

Wooldridge, 2011 – 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

Considere o modelo

$$salario_i = \beta_0 + \beta_1 educ_i + \varepsilon_i \qquad (1)$$

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).***

Variáveis Instrumentais

Todavia, 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 *educ*, isto é,

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

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

Variáveis Instrumentais

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

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

Pergunta: Quais variáveis poderiam ser instrumentos para *educ*? Justifique a sua resposta.

Voltando ao Exemplo

Considere o seguinte modelo de regressão linear simples:

$$nota_i = \beta_0 + \beta_1 faltas_i + \varepsilon_i$$

Pergunta: **Liste ao menos um instrumento para *faltas*? Justifique a sua resposta.**

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

Wooldridge, 2011 – 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) 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.

Observação: Essa nomenclatura é derivada dos modelos de equações simultâneas, que serão estudados em breve.

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 (quando é possível obter valores exatos dos parâmetros estruturais) ou superidentificada (quando mais de uma valor numérico puder ser obtido para alguns dos parâmetros estruturais).

A Estimação

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

$$\underset{\sim}{\mathbf{y}} = \underset{\sim}{\mathbf{X}}\,\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\boldsymbol{\varepsilon}}$$

Considere $\mathbf{Z}$ uma matriz de instrumentos (a matriz $\mathbf{Z}$ é construída de forma análoga à matriz $\mathbf{X}$, entrando no lugar dos regressores endógenos os respectivos instrumentos).

Observação: ***Os regressores exógenos que aparecem na matriz de explicação serão usados como instrumentos deles mesmos na matriz de instrumentos.***

SUPOSIÇÕES ADICIONAIS

Lembrando que para obtenção dos resultados a seguir, faremos uso das seguintes suposições adicionais:

(a.1) $\operatorname{plim}\left(\frac{1}{n}\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\boldsymbol{\varepsilon}}\right) = \underset{\sim}{\mathbf{0}}$

= 0 Pq?

(b.1) $\operatorname{plim}\left(\frac{1}{n}\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\mathbf{X}}\right) = \underset{\sim}{\mathbf{Q}}_{\mathbf{ZX}}$

$\underset{\sim}{\mathbf{Z}}$ - Matriz de Instrumentos

$\underset{\sim}{\mathbf{X}}$ - Matriz de Explicação

A Estimação

Pré-multiplicando

$$\underset{\sim}{\mathbf{y}} = \underset{\sim}{\mathbf{X}}\,\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\boldsymbol{\varepsilon}}$$

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

$$\underset{\sim}{\mathbf{Z}}'\,\underset{\sim}{\mathbf{y}} = \underset{\sim}{\mathbf{Z}}'\,\underset{\sim}{\mathbf{X}}\,\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\mathbf{Z}}'\,\underset{\sim}{\boldsymbol{\varepsilon}}$$

Ainda, multiplicando a equação anterior por n^{-1}, vem que:

$$\frac{1}{n}\,\underset{\sim}{\mathbf{Z}}'\,\underset{\sim}{\mathbf{y}} = \frac{1}{n}\,\underset{\sim}{\mathbf{Z}}'\,\underset{\sim}{\mathbf{X}}\,\underset{\sim}{\boldsymbol{\beta}} + \frac{1}{n}\,\underset{\sim}{\mathbf{Z}}'\,\underset{\sim}{\boldsymbol{\varepsilon}}$$

A Estimação

Tomando o limite de probabilidade em ambos os lados da igualdade, temos que:

$$\operatorname{plim}\left(\frac{1}{n}\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\mathbf{y}}\right)=\operatorname{plim}\left(\frac{1}{n}\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\mathbf{X}}\underset{\sim}{\boldsymbol{\beta}}+\frac{1}{n}\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\boldsymbol{\varepsilon}}\right)$$

Ainda,

$$\operatorname{plim}\left(\frac{1}{n}\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\mathbf{y}}\right)=\operatorname{plim}\left(\frac{1}{n}\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\mathbf{X}}\underset{\sim}{\boldsymbol{\beta}}\right)+\operatorname{plim}\left(\frac{1}{n}\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\boldsymbol{\varepsilon}}\right)$$

= 0

A Estimação

Também,

$$\operatorname{plim}\left(\frac{1}{n}\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\mathbf{y}}\right)=\operatorname{plim}\left(\frac{1}{n}\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\mathbf{X}}\right)\underset{\sim}{\boldsymbol{\beta}}$$

Da expressão anterior, podemos isolar o vetor de parâmetros, obtendo

$$\underset{\sim}{\boldsymbol{\beta}}=\left[\operatorname{plim}\left(\frac{1}{n}\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\mathbf{X}}\right)\right]^{-1}\operatorname{plim}\left(\frac{1}{n}\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\mathbf{y}}\right)$$

A Estimação

Das propriedades do plim(·), vem que

$$\underset{\sim}{\boldsymbol{\beta}} = \operatorname{plim}\left[\left(\frac{1}{n}\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\mathbf{X}}\right)^{-1}\right]\operatorname{plim}\left[\left(\frac{1}{n}\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\mathbf{y}}\right)\right] = \operatorname{plim}\left[\left(\frac{1}{n}\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\mathbf{X}}\right)^{-1}\left(\frac{1}{n}\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\mathbf{y}}\right)\right]$$

O que resulta

$$\underset{\sim}{\boldsymbol{\beta}} = \operatorname{plim}\left[\left(\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\mathbf{X}}\right)^{-1}\left(\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\mathbf{y}}\right)\right]$$

Logo,

$$\underset{\sim}{\boldsymbol{\beta}} = \operatorname{plim}\left(\underset{\sim}{\hat{\boldsymbol{\beta}}}_{\mathbf{IV}}\right)$$

A Estimação

Finalmente,

$$\underset{\sim}{\hat{\beta}}_{IV} = \left(\underset{\sim}{Z}'\underset{\sim}{X}\right)^{-1}\underset{\sim}{Z}'\underset{\sim}{y}$$

Que é um vetor de estimadores consistente!

Observações

- **O método de estimação com o uso de variáveis instrumentais (IV) é mais geral do que MQO;**

- **MQO é um caso particular de IV, uso as variáveis explicativas como instrumentos delas mesmas.**

Variância do Vetor de Estimadores

Não é difícil demonstrar que, sob a suposição de homocedasticidade do vetor de erros, além das demais suposições usuais, a variância do vetor de estimadores de IV é dada por:

$$\mathrm{Var}\left(\underset{\sim}{\hat{\boldsymbol{\beta}}}_{\mathbf{IV}}\right) = \sigma^2 \left(\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\mathbf{X}}\right)^{-1} \underset{\sim}{\mathbf{Z}}'\underset{\sim}{\mathbf{Z}} \left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{Z}}\right)^{-1}$$

Todavia, como σ^2 é um parâmetro desconhecido, precisaremos propor um estimador para tal quantidade.

Variância do Vetor de Estimadores

Um estimador usual para σ^2 é dado por:

$$\hat{\sigma}^2 = \frac{1}{n}\underset{\sim}{\hat{\boldsymbol{\varepsilon}}}'\underset{\sim}{\hat{\boldsymbol{\varepsilon}}} = \frac{1}{n}\left(\underset{\sim}{\mathbf{y}} - \underset{\sim}{\mathbf{X}}\underset{\sim}{\hat{\boldsymbol{\beta}}}_{\mathbf{IV}}\right)'\left(\underset{\sim}{\mathbf{y}} - \underset{\sim}{\mathbf{X}}\underset{\sim}{\hat{\boldsymbol{\beta}}}_{\mathbf{IV}}\right)$$

Dessa forma,

$$\widehat{\mathrm{Var}\left(\underset{\sim}{\hat{\boldsymbol{\beta}}}_{\mathbf{IV}}\right)} = \hat{\sigma}^2\left(\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\mathbf{X}}\right)^{-1}\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\mathbf{Z}}\left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{Z}}\right)^{-1}$$

Propriedade do Vetor de Estimadores

Ainda, é possível provar que

$$\sqrt{n}\left(\underset{\sim}{\hat{\boldsymbol{\beta}}}_{\mathbf{IV}} - \underset{\sim}{\boldsymbol{\beta}}\right) \overset{a}{\sim} N\left(\underset{\sim}{\mathbf{0}}, \sigma^2 \underset{\sim}{\mathbf{Q}}_{\mathbf{ZX}}^{-1} \underset{\sim}{\mathbf{Q}}_{\mathbf{ZZ}} \underset{\sim}{\mathbf{Q}}_{\mathbf{XZ}}^{-1}\right)$$

<u>Observação</u>: Caso a suposição de homocedasticidade não seja válida, então

$$\mathrm{Var}\left(\underset{\sim}{\hat{\boldsymbol{\beta}}}_{\mathbf{IV}}\right) = \sigma^2 \left(\underset{\sim}{\mathbf{Z}}' \underset{\sim}{\mathbf{X}}\right)^{-1} \underset{\sim}{\mathbf{Z}}' \underset{\sim}{\boldsymbol{\Omega}} \underset{\sim}{\mathbf{Z}} \left(\underset{\sim}{\mathbf{X}}' \underset{\sim}{\mathbf{Z}}\right)^{-1}$$

Observações

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

(ii) Para estimar o vetor de parâmetros, precisamos garantir que a matriz Z'X admite inversa.

(iii) Ainda, se Z for uma matriz de dimensão *n* x *L* e X for uma matriz de dimensão *n* x *k*, precisaremos que *L* = *k*. (equação exatamente identificada!)

Observações

(iv) Faremos, ainda, uma suposição adicional de que a variável instrumental *z* seja fortemente correlacionada com a variável endógena *x*. (suposição ligada ao fato do uso de instrumentos fortes, em detrimento aos instrumentos fracos)

(v) De (iv) conseguimos garantir que o método de estimação proposto apresenta bom desempenho com amostras finitas.

Leitura: Wooldridge (2003, p. 493-94)

Exercício

Considere o modelo de regressão linear geral

$$\underset{\sim}{\mathbf{y}} = \underset{\sim}{\mathbf{X}}\,\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\boldsymbol{\varepsilon}}$$

em que

$$\underset{\sim}{\mathbf{X}} = \left[\begin{array}{cccc} \underset{\sim}{i} & \underset{\sim}{x_1} & \underset{\sim}{x_2} & \cdots \; \underset{\sim}{x_k} \end{array}\right]$$

Desconfia-se que as variáveis x_1 e x_2 sejam endógenas.

Exercício (cont.)

Perguntas:

a) **Seria razoável propor o mesmo instrumento para ambas as variáveis? Discuta as implicações no método de estimação.**

b) **Haveria algum problema no caso em que fossem propostos exatamente um instrumento diferente para cada variável?**

Exercício (cont.)

Perguntas: (cont.)

c) Seria razoável propor mais de um instrumento para cada variável endógena?

d) Admitindo a validade de (c), como ficariam as dimensões das matrizes Z e X?

e) Admitindo a validade de (c), seria possível gerar diretamente o vetor de estimadores?

Exercício (cont.)

Solução para o que foi discutido em (c), (d) e (e):

2SLS

(mínimos quadrados em dois estágios)

Método de estimação utilizado quando

a equação estrutural encontra-se sobreidentificada!

MÍNIMOS QUADRADOS EM 2 ESTÁGIOS (2SLS)

Wooldridge, 2011 – Capítulo 15

2SLS

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

$$\underset{\sim}{\mathbf{y}} = \underset{\sim}{\mathbf{X}}\,\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\boldsymbol{\varepsilon}}$$

Inicialmente devemos construir a matriz de instrumentos Z, lembrando que os regressores exógenos da equação estrutural são considerados instrumentos deles mesmos.

2SLS

Como as matrizes X e Z não apresentam as mesmas dimensões, o procedimento adotado é o seguinte:

1o. Estágio: Regredir cada variável explicativa do modelo original em função dos instrumentos (estimação das formas reduzidas) e gerar uma matriz de valores ajustados;

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

2SLS

Em notação matricial:

1o. Estágio:

Estimar os parâmetros da equação auxiliar

$$\underset{\sim}{\mathbf{X}} = \underset{\sim}{\mathbf{Z}}\,\underset{\sim}{\boldsymbol{\pi}} + \underset{\sim}{\boldsymbol{\upsilon}}$$

via por MQO, obtendo

$$\underset{\sim}{\hat{\boldsymbol{\pi}}} = \left(\underset{\sim}{\mathbf{Z}}'\,\underset{\sim}{\mathbf{Z}}\right)^{-1}\underset{\sim}{\mathbf{Z}}'\,\underset{\sim}{\mathbf{X}}$$

Gerar a matriz de valores ajustados $\underset{\sim}{\hat{\mathbf{X}}} = \underset{\sim}{\mathbf{Z}}\,\underset{\sim}{\hat{\boldsymbol{\pi}}}$

2SLS

Em notação matricial: (cont.)

2o. Estágio:

Estimar os parâmetros da equação de interesse, via MQO, substituindo a matriz de explicação X pela a matriz de valores ajustados, obtida na etapa anterior. Ou seja, estimar os parâmetros da equação:

$$\underset{\sim}{\mathbf{y}} = \underset{\sim}{\hat{\mathbf{X}}}\,\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\boldsymbol{\varepsilon}}$$

obtendo

$$\underset{\sim}{\hat{\boldsymbol{\beta}}}^{(2SLS)} = \left(\underset{\sim}{\hat{\mathbf{X}}}'\,\underset{\sim}{\hat{\mathbf{X}}}\right)^{-1}\underset{\sim}{\hat{\mathbf{X}}}'\,\underset{\sim}{\mathbf{y}}$$

2SLS

Do primeiro estágio, vale observar que:

$$\underset{\sim}{\hat{\boldsymbol{\pi}}} = \left(\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\mathbf{Z}}\right)^{-1}\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\mathbf{X}}$$

assim,

$$\underset{\sim}{\hat{\mathbf{X}}} = \underset{\sim}{\mathbf{Z}}\,\underset{\sim}{\hat{\boldsymbol{\pi}}} = \underset{\sim}{\mathbf{Z}}\left(\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\mathbf{Z}}\right)^{-1}\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\mathbf{X}} = \underset{\sim}{\mathbf{P}_{\mathbf{Z}}}\,\underset{\sim}{\mathbf{X}}$$

em que,

$$\underset{\sim}{\mathbf{P}_{\mathbf{Z}}} = \underset{\sim}{\mathbf{Z}}\left(\underset{\sim}{\mathbf{Z}}'\underset{\sim}{\mathbf{Z}}\right)^{-1}\underset{\sim}{\mathbf{Z}}'$$

2SLS

Dessa forma,

$$\underset{\sim}{\hat{\beta}}^{(2SLS)} = \left(\underset{\sim}{\hat{X}}'\underset{\sim}{\hat{X}}\right)^{-1}\underset{\sim}{\hat{X}}'\underset{\sim}{y} = \left(\underset{\sim}{X}'\underset{\sim}{P_Z}'\underset{\sim}{P_Z}\underset{\sim}{X}\right)^{-1}\underset{\sim}{X}'\underset{\sim}{P_Z}'\underset{\sim}{y} = \left(\underset{\sim}{X}'\underset{\sim}{P_Z}\underset{\sim}{X}\right)^{-1}\underset{\sim}{X}'\underset{\sim}{P_Z}'\underset{\sim}{y}$$

Ou seja,

$$\underset{\sim}{\hat{\beta}}^{(2SLS)} = \left(\underset{\sim}{X}'\underset{\sim}{P_Z}\underset{\sim}{X}\right)^{-1}\underset{\sim}{X}'\underset{\sim}{P_Z}'\underset{\sim}{y}$$

Variância do Vetor de Estimadores

Não é difícil demonstrar que, sob a suposição de homocedasticidade do vetor de erros, além das demais suposições usuais, a variância do vetor de estimadores de 2SLS será dada por:

$$\mathrm{Var}\left(\underset{\sim}{\hat{\boldsymbol{\beta}}}^{(2SLS)}\right) = \sigma^2\left(\underset{\sim}{\mathbf{X}}'\,\underset{\sim}{\mathbf{P}_Z}\,\underset{\sim}{\mathbf{X}}\right)^{-1}$$

Todavia, como σ^2 é um parâmetro desconhecido, precisaremos propor um estimador para tal quantidade.

Variância do Vetor de Estimadores

Um estimador usual para σ^2 é dado por:

$$\hat{\sigma}^2 = \frac{1}{n}\underset{\sim}{\hat{\boldsymbol{\varepsilon}}}'\underset{\sim}{\hat{\boldsymbol{\varepsilon}}} = \frac{1}{n}\left(\underset{\sim}{\mathbf{y}} - \underset{\sim}{\mathbf{X}}\underset{\sim}{\hat{\boldsymbol{\beta}}}^{(2SLS)}\right)'\left(\underset{\sim}{\mathbf{y}} - \underset{\sim}{\mathbf{X}}\underset{\sim}{\hat{\boldsymbol{\beta}}}^{(2SLS)}\right)$$

Dessa forma,

$$\widehat{\mathrm{Var}\left(\underset{\sim}{\hat{\boldsymbol{\beta}}}^{(2SLS)}\right)} = \hat{\sigma}^2\left(\underset{\sim}{\mathbf{X}}'\mathbf{P}_{\mathbf{Z}}\underset{\sim}{\mathbf{X}}\right)^{-1}$$

Propriedade do Vetor de Estimadores

Sob certas condições, é possível provar que

$$\underset{\sim}{\hat{\boldsymbol{\beta}}}^{(2SLS)} \overset{a}{\sim} N\left(\underset{\sim}{\boldsymbol{\beta}}, Var\left(\underset{\sim}{\hat{\boldsymbol{\beta}}}^{(2SLS)}\right)\right)$$

<u>Observação</u>: Caso a suposição de homocedasticidade não seja válida, então

$$\text{Var}\left(\underset{\sim}{\hat{\boldsymbol{\beta}}}^{(2SLS)}\right) = \sigma^2\left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{P}_Z}\underset{\sim}{\mathbf{X}}\right)^{-1}\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{P}_Z}\underset{\sim}{\boldsymbol{\Omega}}\underset{\sim}{\mathbf{P}_Z}\underset{\sim}{\mathbf{X}}\left(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{P}_Z}\underset{\sim}{\mathbf{X}}\right)^{-1}$$

Aplicação

Exercício sobre IV – MROZ (1987)

Considere o seguinte modelo de regressão linear múltipla:

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

(para as mulheres que trabalham)

Utilizando o arquivo *MROZ.xls*:

a) **Você diria que MQO é um método de estimação adequado para estimar os parâmetros deste modelo? Justifique.**

b) **Estime os parâmetros do modelo, via MQO. Comente.**

c) **Use *educação da mãe* como instrumento para *educ* e reestime os parâmetros do modelo de interesse, por IV. Comente os resultados obtidos.**

Aplicação

Exercício sobre IV – MROZ (1987)

Considere o seguinte modelo de regressão linear múltipla:

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

(para as mulheres que trabalham)

Utilizando o arquivo *MROZ.xls*: (cont.)

d) Usando *educação do pai*, *educação da mãe* e *educação do marido* como instrumentos para *educ*, estime os parâmetros do modelo por 2SLS. Comente.

e) As estimativas dos parâmetros, em (b), diferiram muito quando comparadas àquelas obtidas por IV e 2SLS? Discuta as implicações do resultado observado.

TESTE DE ENDOGENEIDADE

Wooldridge, 2011 – Capítulo 15

Teste de Endogeneidade

O estimador de 2SLS é menos eficiente que o de MQO quando as variáveis explicativas são exógenas. Assim sendo, se torna útil fazer um teste de endogeneidade de uma variável explicativa que mostre se a utilização de 2SLS é necessária.

Teste de Endogeneidade

Considere o modelo

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \varepsilon_1 \qquad (1)$$

em que

y_2 – variável endógena

z_1 e z_2 – variáveis exógenas

Ainda, considere que *z_3* seja um instrumento para *y_2*.

Teste de Endogeneidade

Fatos

1. Se y_2 for não correlacionada com ε_1, então, devemos estimar os parâmetros do modelo por MQO (mais eficiente).
2. MQO e 2SLS fornecem estimadores consistentes se a condição de exogeneidade estiver satisfeita.

Teste de Endogeneidade

HAUSMAN (1978), sugeriu fazer uma comparação direta das estimativas de MQO e 2SLS e determinar se as diferenças são estatisticamente significantes.

Se as estimativas geradas por MQO e 2SLS diferirem de forma significante, concluímos que y_2 deve ser endógena (supondo z_1 e z_2 exógenas).

Teste de Endogeneidade

Procedimento para aplicação do Teste de Hausman:

i. Estime a forma reduzida de y_2, regredindo y_2 sobre todas as variáveis exógenas (inclusive aquelas da equação estrutural e as IVs adicionais).

ii. Obtenha os resíduos.

iii. Estime a equação estrutural, por MQO, utilizando os resíduos, obtidos em (ii), como variável explicativa.

iv. Se o parâmetro associado ao resíduo for estatisticamente significante, concluiremos que y_2 é endógena.

Aplicação

Exercício sobre IV – MROZ (1987)

Considere o seguinte modelo de regressão linear múltipla:

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

(para as mulheres que trabalham)

Usando *educação do pai*, *educação da mãe* e *educação do marido* como instrumentos para *educ*, estime os parâmetros do modelo por 2SLS e verifique se a variável *educ* é endógena. Ainda, discuta os resultados obtidos, levando em consideração as propriedades dos estimadores, para cada um dos métodos utilizados anteriormente. Para tanto, utilize o arquivo *MROZ.xls*.

TESTE DE SARGAN

VERIFICAÇÃO DA VALIDADE DOS INSTRUMENTOS

Gujarati, 2011 – página 663

TESTE DE SARGAN

Qual a validade do instrumento, ou seja, como sabemos se os instrumentos escolhidos são independentes do termo de erro?

Para responder à pergunta anterior, Sargan (1964) desenvolveu um teste estatístico, chamado de SARG, para testar a validade dos instrumentos.

TESTE DE SARGAN

O procedimento é o seguinte:

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.**

3) **Gere os resíduos de (2) e regrida-os em função de uma constante, todas as variáveis exógenas da equação estrutural e de todos os instrumentos.**

4) **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.**

5) **Rejeite H_0 (instrumentos válidos), se SARG > $\chi^2_{(crítico)}$.**

Exercício

Considere o seguinte modelo de regressão linear múltipla:

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

(para as mulheres que trabalham)

Utilizando o arquivo *MROZ.xls*:

a) Estime os parâmetros do modelo por 2SLS, usando *educação do pai*, *educação da mãe* e *educação do marido* como instrumentos para *educ* e conduza o teste de SARG. Comente.

b) Seria possível realizar o teste anteriormente proposto se tivéssemos apenas um instrumento? Justifique a sua resposta.

LEITURA COMPLEMENTAR I

TESTE

DE

RESTRIÇÕES SOBREIDENTIFICADORAS

(análogo ao Teste de SARGAN – validade dos instrumentos)

Wooldridge, 2011 – Capítulo 15

Teste de Restrições Sobreidentificadoras

Suponha que na equação estrutural de interesse apareça somente uma variável explicativa endógena.

Nesse caso:

- Se tivermos somente uma única IV, não teremos restrições sobreidentificadoras. Ou seja, não haverá nada que possa ser testado.
- Se tivermos duas IVs, teremos uma restrição sobreidentificadora. Se tivermos três IVs, teremos duas restrições sobreidentificadoras, e assim por diante.

Teste de Restrições Sobreidentificadoras

Procedimento para aplicação do teste:

i. Estime os parâmetros da equação estrutural por 2SLS.

ii. Obtenha os resíduos.

iii. Regrida os resíduos em função de todas as variáveis exógenas.

iv. Obtenha o R^2 (coeficiente de explicação).

v. Sob a hipótese nula de que todas as IVs são não correlacionadas com o erro da equação estrutural,

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

em que

q – é o número de IVs menos o número de regressores endógenos presentes no modelo.

Teste de Restrições Sobreidentificadoras

Rejeitar a hipótese nula significa que pelo menos uma das IVs não é exógena.

Aplicação

Exercício sobre IV – MROZ (1987)

Considere o seguinte modelo de regressão linear múltipla:

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

(para as mulheres que trabalham)

Para responder o item, a seguir, utilize o arquivo *MROZ.xls*:

a) **Usando *educação do pai*, *educação da mãe* e *educação do marido* como instrumentos para *educ*, estime os parâmetros do modelo por 2SLS.**

b) **Teste se o uso de três instrumentos (duas restrições sobreidentificadoras) gera viés no estimador de 2SLS (se algum dos instrumentos é correlacionado com o erro do modelo).**

Aplicação (cont.)

Exercício sobre IV – MROZ (1987)

Considere o seguinte modelo de regressão linear múltipla:

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

(para as mulheres que trabalham)

Para responder o item, a seguir, utilize o arquivo *MROZ.xls*:

c) Seria possível realizar o teste anteriormente proposto se tivéssemos apenas um instrumento? Justifique a sua resposta.

Referência da Aplicação

MROZ, T. A. (1987). The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions, *Econometrica,* 55, 765-799.

LEITURA COMPLEMENTAR II

(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

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

em que

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

Erros de Medição em y

Por simplicidade, vamos admitir que:

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