Introdução à Econometria

Uma Abordagem Moderna

Jeffrey M. Wooldridge
Michigan State University

Tradução da quarta edição norte-americana

Tradução
José Antônio Ferreira

Revisão Técnica
Galo Carlos Lopez Noriega, MSc.
Docente de métodos quantitativos no MBA do Insper Ibmec São Paulo
e coordenador acadêmico de Educação Executiva do Insper Ibmec São Paulo

CENGAGE Learning

Austrália Brasil Canadá Cingapura Espanha Estados Unidos México Reino Unido

Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Wooldridge, Jeffrey M.
Introdução à econometria : uma abordagem moderna / Jeffrey M. Wooldridge ; tradução José Antônio Ferreira ; revisão técnica Galo Carlos Lopez Noriega. -- São Paulo : Cengage Learning, 2010.

Título original: Introductory econometrics : a modern approach
4. ed. norte-americana
Bibliografia.
ISBN 978-85-221-0446-8

1. Econometria II. Título.

10-11298 CDD-330.015195

Índices para catálogo sistemático:

1. Econometria 330.015195

15

Estimação de Variáveis Instrumentais e Mínimos Quadrados em Dois Estágios

Neste capítulo, estudaremos com mais profundidade o problema das **variáveis explicativas endógenas** em modelos de regressão múltipla. No Capítulo 3, derivamos o viés nos estimadores MQO quando uma variável importante era omitida; no Capítulo 5, mostramos que os estimadores MQO são, em geral, inconsistentes sob **variáveis omitidas**. O Capítulo 9 demonstrou que o viés de variáveis omitidas pode ser eliminado (ou pelo menos suavizado) quando uma variável *proxy* adequada é escolhida para representar uma variável explicativa não observada. Infelizmente, variáveis *proxy* adequadas nem sempre estão disponíveis.

Nos dois capítulos anteriores, explicamos como a estimação por efeitos fixos ou por primeira diferença pode ser usada com dados em painel para estimar os efeitos de variáveis independentes que variam no tempo, na presença de variáveis omitidas *constantes no tempo*. Embora tais métodos sejam bastante úteis, nem sempre temos acesso aos dados em painel. Mesmo que possamos obter dados em painel, eles serão de pouca utilidade se estivermos interessados no efeito de uma variável que não se altera ao longo do tempo: a estimação por primeira diferença ou por efeitos fixos elimina as variáveis explicativas constantes no tempo. Além disso, os métodos de dados em painel que estudamos até agora não solucionam o problema de as variáveis omitidas que variam no tempo serem correlacionadas com as variáveis explicativas.

Neste capítulo, consideramos uma abordagem diferente do problema da endogeneidade. Você verá como o método das variáveis instrumentais (VI) pode ser usado para solucionar o problema da endogeneidade de uma ou de mais variáveis explicativas. O método de mínimos quadrados em dois estágios (MQ2E ou MQDE) só é superado em popularidade pelo método de mínimos quadrados ordinários usado para estimar equações lineares em econometria aplicada.

Começaremos mostrando como os métodos VI podem ser usados para obter estimadores consistentes na presença de variáveis omitidas. Os métodos VI também podem ser usados para solucionar o problema de **erros nas variáveis**, pelo menos sob certas hipóteses. O próximo capítulo demonstrará como estimar modelos de equações simultâneas usando os métodos VI.

Nossa abordagem da estimação de variáveis instrumentais acompanha de perto nosso desenvolvimento dos mínimos quadrados ordinários na Parte 1, na qual presumimos que tínhamos uma amostra aleatória de uma população básica. Esse é um ponto de partida desejável, pois, além de simplificar a notação, enfatiza que as hipóteses importantes da estimação por VI são definidas em termos da população básica (como acontece com o MQO). Conforme mostramos na Parte 2, o MQO pode ser aplicado a dados de séries temporais, e o mesmo é verdadeiro para os métodos de variáveis instrumentais. A Seção 15.7 discute alguns problemas especiais que surgem quando os métodos VI são aplicados a dados de séries temporais. Na Seção 15.8, tratamos de aplicações a cortes transversais agrupados e dados em painel.

15.1 MOTIVAÇÃO: VARIÁVEIS OMITIDAS EM UM MODELO DE REGRESSÃO SIMPLES

Quando defrontados com a possibilidade de viés de variáveis omitidas (ou heterogeneidade não observada), até agora examinamos três opções: (1) podemos ignorar o problema e sofrer as consequências de estimadores viesados e inconsistentes; (2) podemos tentar encontrar e usar uma variável *proxy* adequada da variável não observada; ou (3) podemos presumir que a variável omitida não se altera ao longo do tempo e utilizar os métodos de efeitos fixos ou de primeira diferença vistos nos Capítulos 13 e 14. A primeira opção poderá ser satisfatória se as estimativas estiverem acopladas com a direção dos vieses dos parâmetros importantes. Por exemplo, se pudermos dizer que o estimador de um parâmetro positivo, digamos, o efeito do treinamento de pessoal sobre os salários subsequentes, é viesado para zero e tenhamos constatado uma estimativa positiva estatisticamente significante, ainda teremos descoberto alguma coisa: o treinamento de pessoal tem efeito positivo sobre os salários, e é provável que tenhamos subestimado o efeito. Infelizmente, o caso oposto, no qual nossas estimativas podem ter uma magnitude grande demais, ocorre com frequência, o que torna muito difícil para se obter qualquer conclusão útil.

A solução da variável *proxy* discutida na Seção 9.2 do Capítulo 9 também pode produzir resultados satisfatórios, mas nem sempre é possível encontrar uma boa *proxy*. Essa abordagem tenta resolver o problema da variável omitida substituindo a variável não observada por uma variável *proxy*.

Outra abordagem deixa a variável não observada no termo de erro, mas, em vez de estimar o modelo por MQO, ela usa um método de estimação que reconhece a presença da variável omitida. É isso que o método das variáveis instrumentais faz.

A título ilustrativo considere o problema da aptidão não observada em uma equação de salários-hora de trabalhadores adultos. Um modelo simples é

$$\log(salário\text{h}) = \beta_0 + \beta_1 educ + \beta_2 aptid + e,$$

em que e é o termo de erro. No Capítulo 9, mostramos como, sob certas hipóteses, uma variável *proxy* como QI pode substituir a aptidão, e assim um estimador consistente de β_1 será obtido a partir da regressão de

$$\log(salário\text{h}) \text{ sobre } educ, QI.$$

Suponha, porém, que uma variável *proxy* não esteja disponível (ou não tenha as propriedades necessárias para produzir um estimador consistente de β_1). Então, colocamos *aptid* no termo de erro e ficamos com o modelo de regressão simples

$$\log(salário\text{h}) = \beta_0 + \beta_1 educ + u, \tag{15.1}$$

em que u contém *aptid*. Naturalmente, se a equação (15.1) for estimada por MQO, o resultado será um estimador viesado e inconsistente de β_1 se *educ* e *aptid* forem correlacionados.

Constata-se ainda que podemos usar a equação (15.1) como a base da estimação, desde que possamos encontrar uma variável instrumental de *educ*. Para descrever essa abordagem, o modelo de regressão simples é escrito como

$$y = \beta_0 + \beta_1 x + u, \tag{15.2}$$

em que acreditamos que x e u são correlacionados:

$$\text{Cov}(x,u) \neq 0, \tag{15.3}$$

O método das variáveis instrumentais funciona sejam x e u correlacionados ou não, mas, por razões que veremos mais tarde, o MQO deverá ser usado se x for não correlacionado com u.

Para obter estimadores consistentes de β_0 e β_1 quando x e u forem correlacionados, necessitaremos de alguma informação adicional. A informação virá por meio de uma nova variável que satisfaça certas propriedades. Suponha que temos uma variável observável z que satisfaz as seguintes duas hipóteses: (1) z é não correlacionado com u, isto é,

$$\text{Cov}(z,u) = 0; \tag{15.4}$$

(2) z é correlacionado com x, isto é,

$$\text{Cov}(z,x) \neq 0. \tag{15.5}$$

Então, dizemos que z é uma **variável instrumental** de x, ou algumas vezes simplesmente um **instrumento** para x.

A exigência que o instrumento z satisfaça a (15.4) é resumida dizendo-se "z é exógena na equação (15.2)", e assim frequentemente nos referimos a (15.4) como **exogeneidade dos instrumentos**. No contexto de variáveis omitidas, exogeneidade instrumental significa que z não deve ter efeito parcial em y (após x e as variáveis omitidas terem sido controladas), e z deve ser não correlacionada com as variáveis omitidas. A equação (15.5) significa que z deve ser relacionado, positiva ou negativamente, com a variável explicativa endógena x. Esta condição algumas vezes é referida como uma **relevância dos instrumentos** (como "z é relevante para explicar a variação em x").

Existe uma diferença bastante importante entre os dois requisitos de uma variável instrumental. Como (15.4) envolve a covariância entre z e o erro não observado u, não podemos geralmente ter esperança de testar essa hipótese: na maioria dos casos, temos que manter $\text{Cov}(z,u) = 0$, recorrendo ao comportamento ou à introspecção econômica. (Em casos menos usuais, é possível que tenhamos uma variável *proxy* observável de algum fator contido em u, caso em que poderemos verificar se z e a variável *proxy* são mais ou menos não correlacionadas. Evidentemente, se tivermos uma boa *proxy* de um elemento importante de u, poderemos simplesmente adicionar a *proxy* como uma variável explicativa e estimar a equação expandida por mínimos quadrados ordinários. Veja a Seção 9.2.)

Em contraposição, a condição de que z seja correlacionado com x (na população) pode ser testada, dada uma amostra aleatória da população. A maneira mais fácil de fazer isso é estimar uma regressão simples entre x e z. Na população, temos

$$x = \pi_0 + \pi_1 z + v, \tag{15.6}$$

Então, como $\pi_1 = \text{Cov}(z,x)/\text{Var}(z)$, a Hipótese (15.5) será válida se, e somente se, $\pi_1 \neq 0$. Assim, deveremos ser capazes de rejeitar a hipótese nula

$$H_0: \pi_1 = 0 \tag{15.7}$$

contra a alternativa bilateral H_1: $\pi_1 \neq 0$, em um nível de significância suficientemente pequeno (digamos, 5% ou 1%). Se esse for o caso, podemos ter uma razoável confiança em que (15.5) se mantém.

Para a equação log(*salário*h) em (15.1), uma variável instrumental z de *educ* deve ser (1) não correlacionada com a aptidão (e com quaisquer outros fatores não observáveis que afetem o salário) e (2) correlacionada com educação. Algo como o último dígito do número da previdência social de um indivíduo, quase que com certeza satisfará o primeiro requisito: ele será não correlacionado com a aptidão, por ser determinado de forma aleatória. Porém, essa variável não será correlacionada com educação, e assim será uma variável instrumental muito pobre de *educ*.

O que chamamos de *variável proxy* da variável omitida transforma-se em uma VI pobre pelo motivo oposto. No exemplo de log(*salário*h) com a aptidão omitida, uma variável *proxy* de *aptid* deverá ser tão altamente correlacionada quanto possível com *aptid*, por exemplo. Uma variável instrumental deverá ser *não correlacionada* com *aptid*. Portanto, embora a variável *QI* seja uma boa candidata para ser uma variável *proxy* de *aptid*, não será uma boa variável instrumental de *educ*.

Se outras possíveis candidatas a variável instrumental satisfazem as exigências de exogeneidade em (15.4) é menos claro. Em equações de salários, os economistas trabalhistas usam variáveis do perfil familiar como VIs da educação. Por exemplo, a escolaridade da mãe (*educm*) é positivamente correlacionada com a educação dos filhos, como poderá ser verificado coletando uma amostra de dados sobre trabalhadores e computando uma regressão simples de *educ* sobre *educm*. Portanto, *educm* satisfará a equação (15.5). O problema é que a escolaridade da mãe também poderá estar correlacionada com a aptidão dos filhos (por meio da aptidão da mãe e talvez da qualidade da nutrição em tenra idade). Nesse caso, (15.4) falha.

Outra possível VI de *educ* em (15.1) é o número de irmãos durante o crescimento (*irms*). Normalmente, ter mais irmãos está associado a níveis médios mais baixos de educação. Assim, se o número de irmãos for não correlacionado com a aptidão, ele pode agir como uma variável instrumental de *educ*.

Como segundo exemplo, considere o problema de estimar o efeito causal de faltar às aulas sobre as notas do exame final. Em uma estrutura de regressão simples, temos

$$nota = \beta_0 + \beta_1 faltas + u, \tag{15.8}$$

em que *nota* é a nota no exame final e *faltas* é o número total de faltas às aulas durante o semestre. Certamente devemos estar preocupados se *faltas* está correlacionado com outros fatores em u: alunos mais aptos, altamente motivados, devem ter um menor número de faltas. Assim, uma regressão simples de *nota* sobre *faltas* pode não produzir uma boa estimativa do efeito causal de faltas às aulas.

Qual poderia ser uma boa VI de *faltas*? Necessitamos de algo que não tenha efeito direto sobre *nota* e que não seja correlacionado com a aptidão e motivação do aluno. Ao mesmo tempo, a VI deve ser correlacionada com *faltas*. Uma opção é usar a distância entre os alojamentos e o *campus*. Alguns alunos em grandes universidades se deslocam constantemente para o *campus*, o que pode aumentar a possibilidade de eles faltarem às aulas (em razão do mau tempo, por terem dormido demais etc.). Assim, *faltas* pode estar positivamente correlacionado com *distância*; isso pode ser verificado regredindo *faltas* sobre *distância* e fazendo-se um teste t, como descrito anteriormente.

Será *distância* correlacionado com u? No modelo de regressão simples (15.8), alguns fatores em u poderão ser correlacionados com *distância*. Por exemplo, alunos de família de baixa renda

provavelmente residem fora do *campus*; se a renda afetar o desempenho dos alunos, isso pode fazer com que *distância* seja correlacionado com u. A Seção 15.2 mostra como usar VI no contexto de regressão múltipla, de forma que outros fatores que afetem *nota* possam ser diretamente incluídos no modelo. Assim, *distância* pode ser uma boa VI de *faltas*. Uma abordagem VI pode não ser necessária se houver uma boa *proxy* da aptidão do aluno, como a nota média acumulada anterior ao semestre corrente (GPA).

Agora demonstramos que a disponibilidade de uma variável instrumental pode ser usada para estimar com consistência os parâmetros na equação (15.2). Particularmente, mostramos que as hipóteses (15.4) e (15.5) servem para *identificar* o parâmetro β_1. **Identificação** de um parâmetro nesse contexto significa podermos escrever β_1 em termos de momentos populacionais que possam ser estimados usando uma amostra de dados. Para escrever β_1 em termos de covariâncias populacionais, usamos a equação (15.2): a covariância entre z e y é

$$\text{Cov}(z,y) = \beta_1\text{Cov}(z,x) + \text{Cov}(z,u).$$

Visto que, sob a hipótese (15.4), $\text{Cov}(z,u) = 0$, e sob a hipótese (15.5), $\text{Cov}(z,x) \neq 0$, podemos resolver β_1 como

$$\beta_1 = \frac{\text{Cov}(z,y)}{\text{Cov}(z,x)}. \tag{15.9}$$

[Observe como essa álgebra simples falhará se z e x forem não correlacionados, isto é, se $\text{Cov}(z,x) = 0$.] A equação (15.9) mostra que β_1 é a covariância populacional entre z e y, dividida pela covariância populacional entre z e x, o que mostra que β_1 é identificada. Dada uma amostra aleatória, estimamos as quantidades populacionais pelos análogos da amostra. Após cancelar os tamanhos das amostras no numerador e no denominador, obtemos o **estimador de variáveis instrumentais (VI)** de β_1:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}, \tag{15.10}$$

Dada uma amostra de dados de x, y e z, é simples obter o estimador de VI em (15.10). O estimador de VI de β_0 é simplesmente $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$, que é bastante parecido com o estimador MQO do intercepto, exceto pelo fato de que o estimador de inclinação, $\hat{\beta}_1$, agora é o estimador de VI.

Não é por acaso que quando $z = x$ obtemos o estimador MQO de β_1. Em outras palavras, quando x é exógeno, ele pode ser usado como seu próprio VI, e o estimador de VI será, então, idêntico ao estimador MQO.

Uma aplicação simples da lei dos grandes números mostra que o estimador VI é consistente para β_1: $\text{plim}(\hat{\beta}_1) = \beta_1$, desde que as hipóteses (15.4) e (15.5) sejam satisfeitas. Se qualquer uma dessas hipóteses falhar, os estimadores VI não serão consistentes (veremos mais sobre isso adiante). Uma das características do estimador VI é que, quando x e u forem de fato correlacionados — de forma que a estimação por variáveis instrumentais será realmente necessária —, essencialmente ele sempre será viesado. Isso significa que, em amostras pequenas, o estimador VI pode ter um viés substancial, que é uma das razões pela qual amostras grandes são preferidas.

Inferência Estatística com o Estimador de VI

Dadas as estruturas similares dos estimadores de VI e MQO, não surpreende que o estimador de VI tenha uma distribuição aproximadamente normal em amostras de tamanhos grandes. Para fazer inferência sobre β_1, precisamos de um erro-padrão que possa ser usado para calcular estatísticas t e intervalos de confiança. A abordagem habitual é impor uma hipótese de homoscedasticidade, exatamente como no caso de MQO. Agora, a hipótese de homoscedasticidade é declarada condicional à variável instrumental, z, e não à variável explicativa endógena, x. Com as hipóteses anteriores sobre u, x e z, adicionamos

$$\text{E}(u^2|z) = \sigma^2 = \text{Var}(u). \tag{15.11}$$

Pode ser mostrado que, sob (15.4), (15.5) e (15.11), a variância assimptótica de $\hat{\beta}_1$ é

$$\frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2}, \tag{15.12}$$

em que σ_x^2 é a variância populacional de x, σ^2 é a variância populacional de u, e $\rho_{x,z}^2$ é o quadrado da correlação populacional entre x e z. Isso nos informa o quanto x e z são altamente correlacionados na população. Igualmente como o estimador de MQO, a variância assimptótica do estimador de VI decresce para zero à taxa de $1/n$, em que n é o tamanho da amostra.

A equação (15.12) é interessante por duas razões. A primeira fornece uma maneira de obter um erro-padrão do estimador de VI. Todas as quantidades em (15.12) podem ser consistentemente estimadas, dada uma amostra aleatória. Para estimar σ_x^2, simplesmente calculamos a variância amostral de x_i; para estimar $\rho_{x,z}^2$, podemos executar a regressão de x_i sobre z_i para obter o R-quadrado, digamos, $R_{x,z}^2$. Finalmente, para estimar σ^2, podemos usar os resíduos de VI,

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \; i = 1, 2, \ldots, n,$$

em que $\hat{\beta}_0$ e $\hat{\beta}_1$ são as estimativas de VI. Um estimador consistente de σ^2 parece exatamente igual ao estimador de σ^2 de uma regressão simples de MQO.

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2,$$

em que é padrão usar a correção dos graus de liberdade (embora isso tenha pouco efeito conforme o tamanho da amostra cresce).

O erro-padrão (assimptótico) de $\hat{\beta}_1$ é a raiz quadrada da variância assimptótica estimada, e esta última é dada por

$$\frac{\hat{\sigma}^2}{\text{SQT}_x \cdot R_{x,z}^2}, \tag{15.13}$$

em que SQT_x é a soma dos quadrados total de x_i. [Lembre-se que a variância amostral de x_i é SQT_x/n, e assim os tamanhos das amostras são cancelados para nos dar (15.13).] O erro-padrão resultante pode ser usado para construir estatísticas t de hipóteses que envolvam β_1 ou intervalos de confiança de β_1.

$\hat{\beta}_0$ que também tem um erro-padrão que não apresentamos aqui. Qualquer programa moderno de econometria calcula o erro-padrão após qualquer estimação por VI.

A segunda razão (15.12) é interessante porque nos permite comparar as variâncias assimptóticas dos estimadores de VI e MQO (quando x e u são não correlacionados). Sob as hipóteses de Gauss-Markov, a variância do estimador MQO é σ^2/SQT_x, enquanto a fórmula comparável do estimador de VI é $\sigma^2/(\text{SQT}_x \cdot R^2_{x,z})$; elas diferem apenas no fato de $R^2_{x,z}$ aparecer no denominador da variância da VI. Como o R-quadrado é sempre menor que um, a variância de VI é sempre maior que a variância de MQO (quando MQO é válido). Se $R^2_{x,z}$ for pequeno, então, a variância da VI poderá ser muito maior do que a variância do MQO. Lembre-se, $R^2_{x,z}$ mede a intensidade da relação linear entre x e z na amostra. Se x e z forem apenas levemente correlacionados, é possível que $R^2_{x,z}$ seja pequeno, e isso poderá ser traduzido em uma variância amostral muito grande do estimador de VI. Quanto mais altamente correlacionado for z com x, mais próximo de um será $R^2_{x,z}$, e menor será a variância do estimador de VI. No caso em que $z = x$, $R^2_{x,z} = 1$, e obtemos a variância de MQO, como esperado.

A discussão anterior destaca um importante preço a pagar ao executarmos uma estimação de VI quando x e u são não correlacionados: a variância assimptótica do estimador de VI é sempre maior, e algumas vezes muito maior, que a variância assimptótica do estimador de MQO.

EXEMPLO 15.1

(A Estimação do Retorno da Educação para Mulheres Casadas)

Utilizamos os dados sobre mulheres casadas que trabalham contidos no arquivo MROZ.RAW para estimar o retorno da educação no modelo de regressão simples

$$\log(salárioh) = \beta_0 + \beta_1 educ + u. \tag{15.14}$$

Para comparação, primeiro obtemos as estimativas por MQO:

$$\widehat{\log(salárioh)} = -0,185 + 0,109\ educ$$
$$(0,185)\quad (0,014) \tag{15.15}$$
$$n = 428,\ R^2 = 0,118.$$

A estimativa de β_1 implica um retorno de quase 11% para um ano a mais de educação.

Em seguida, usamos a educação do pai (*educp*) como uma variável instrumental de *educ*. Temos que sustentar que *educp* é não correlacionado com u. O segundo requisito é que *educ* e *educp* sejam correlacionados. Podemos verificar isso facilmente, usando uma regressão simples de *educ* sobre *educp* (utilizando somente as mulheres que trabalham da amostra):

$$\widehat{educ} = 10,24 + 0,269\ educp$$
$$(0,28)\quad (0,029) \tag{15.16}$$
$$n = 428,\ R^2 = 0,173.$$

EXEMPLO 15.1 (continuação)

A estatística t de *educp* é 9,28, e indica que *educ* e *educp* têm uma correlação positiva estatisticamente significante. (Aliás, *educp* explica cerca de 17% da variação em *educ* na amostra.) A utilização de *educp* como uma VI de *educ* produz

$$\widehat{\log(salárioh)} = -0,441 + 0,059\ educ$$
$$(0,446)\quad (0,035) \tag{15.17}$$
$$n = 428,\ R^2 = 0,093.$$

A estimativa de VI do retorno da educação é 5,9%, o que é um pouco mais que a metade da estimativa pelos MQO. Isso *sugere* que a estimativa de MQO é alta demais e é consistente com o viés de aptidão omitida. Entretanto, devemos lembrar que essas estimativas são de apenas uma amostra: nunca poderemos saber se 0,109 está acima do verdadeiro retorno da educação, ou se 0,059 está mais próximo do verdadeiro retorno da educação. Além disso, o erro-padrão da estimativa de VI é duas vezes e meia maior que o erro-padrão de MQO (isso era esperado, pelas razões dadas anteriormente). O intervalo de confiança de 95% de β_1, utilizando MQO, é muito mais apertado do que utilizando VI; de fato, o intervalo de confiança da VI, na realidade, contém a estimativa de MQO. Portanto, embora as diferenças entre (15.15) e (15.17) sejam grandes na prática, não podemos dizer se a diferença é *estatisticamente* significante. Mostraremos como testar isso na Seção 15.5.

No exemplo anterior, o retorno estimado da educação, usando VI, foi menor que usando MQO, o que corresponde às nossas expectativas. Contudo, esse poderia não ter sido o caso, como demonstra o exemplo a seguir.

EXEMPLO 15.2

(A Estimação do Retorno da Educação para Homens)

Agora usamos o arquivo WAGE2.RAW para estimar o retorno da educação para homens. Utilizamos a variável *irms* (número de irmãos) como uma instrumental de *educ*. Elas são negativamente correlacionadas, como podemos verificar com uma regressão simples:

$$\widehat{educ} = 14,14 - 0,228\ irms$$
$$(0,11)\quad (0,030)$$
$$n = 935,\ R^2 = 0,057.$$

Essa equação implica em cada irmão estar associado, na média, com cerca de menos 0,23 ano de educação. Se presumirmos que *irms* é não correlacionado com o termo de erro em (15.14), o estimador de VI será consistente. A estimação da equação (15.14) usando *irms* como uma VI de *educ* produz

$$\widehat{\log(salárioh)} = 5,13 + 0,122\ educ$$
$$(0,36)\quad (0,026)$$
$$n = 935.$$

EXEMPLO 15.2 (continuação)

(O *R*-quadrado foi calculado como negativo, de modo que não o descrevemos. Apresentamos a seguir uma discussão sobre o *R*-quadrado no contexto da estimação de VI.) A título de comparação, a estimativa por MQO de β_1 é 0,059 com um erro-padrão de 0,006. Diferentemente do exemplo anterior, a estimativa de VI agora é muito mais alta que a do MQO. Embora não saibamos se a diferença é estatisticamente significante, isso não interage com o viés da aptidão omitida do MQO. Pode ser que *irms* também seja correlacionado com a aptidão: mais irmãos significa, em média, menos atenção dos pais, o que pode resultar em menor aptidão. Outra interpretação seria que o estimador MQO é viesado para zero em razão de um erro de medida em *educ*. Isso não é inteiramente convincente, pois, como discutimos na Seção 9.3, não é provável que *educ* satisfaça o modelo clássico de erros nas variáveis.

Nos exemplos anteriores, a variável explicativa endógena (*educ*) e as variáveis instrumentais (*educp* e *irms*) tinham significados quantitativos. Entretanto, nada impede que a variável explicativa ou a VI sejam variáveis binárias. Angrist e Krueger (1991), em sua análise mais simples, propuseram uma engenhosa variável instrumental binária de *educ*, utilizando dados do censo, sobre homens nos Estados Unidos. Os autores definiram *pritrim* igual a um se o homem nasceu no primeiro trimestre do ano, e zero, caso contrário. Parece que o termo de erro em (15.14) — e, particularmente, a aptidão — deveria não ser relacionado com o trimestre de nascimento. Contudo, *pritrim* também precisa ser correlacionado com *educ*. Acontece que os anos de estudo *realmente* diferem sistematicamente na população, com base em trimestres de nascimento. Angrist e Krueger argumentaram de forma persuavisa que isso é em razão das leis de estudo obrigatório em vigor em todos os estados. Em resumo, os alunos nascidos no início do ano, em geral, começam a estudar com mais idade. Portanto, eles atingem o tempo de estudo obrigatório (16 anos, na maioria dos estados) com um pouco menos de educação do que os alunos que começaram a estudar com menos idade. Sobre os alunos que completaram o ensino médio, Angrist e Krueger verificaram não existir relação entre os anos de estudo e o trimestre de nascimento.

Como anos de estudo varia apenas levemente entre os trimestres de nascimento — o que significa que $R^2_{x,z}$ em (15.13) é muito pequeno — Angrist e Krueger precisaram de uma amostra de tamanho muito grande para obter uma estimativa VI razoavelmente precisa. Utilizando 247.199 homens nascidos entre 1920 e 1929, a estimativa por MQO do retorno da educação foi 0,0801 (erro-padrão de 0,0004), e a estimativa VI foi 0,0715 (0,0219); esses resultados estão registrados na Tabela III do trabalho de Angrist e Krueger. Observe como é elevada a estatística *t* da estimativa por MQO (próxima de 200), enquanto a estatística *t* da estimativa VI é somente 3,26. Assim, a estimativa VI é estatisticamente diferente de zero, mas seu intervalo de confiança é muito mais amplo do que aquele com base na estimativa por MQO.

Uma constatação interessante feita por Angrist e Krueger é que a estimativa VI não difere muito daquela feita por MQO. De fato, usando os homens nascidos na década seguinte, a estimativa VI é um pouco mais alta que a feita por MQO. Seria possível interpretar isso como uma demonstração de que não existe viés de aptidão omitida quando equações de salários são estimadas por MQO. Porém, o trabalho de Angrist e Krueger foi criticado em seus fundamentos econométricos. Como discutido por Bound, Jaeger & Baker (1995), não é óbvio que a época de nascimento seja não relacionada com fatores não observados que afetem o salário. Como explicaremos na próxima subseção, mesmo uma pequena correlação entre *z* e *u* pode causar sérios problemas para o estimador de VI.

Para análises de decisões de políticas, a variável explicativa endógena frequentemente é binária. Por exemplo, Angrist (1990) estudou o efeito que o fato de ser um veterano da Guerra do Vietnã tinha sobre os ganhos de aposentadoria. Um modelo simples é o seguinte

$$\log(ganhos) = \beta_0 + \beta_1 veterano + u, \qquad (15.18)$$

em que *veterano* é uma variável binária. A questão de estimar essa equação por MQO é que pode haver um problema de *autosseleção*, como mencionamos no Capítulo 7: talvez aqueles que procurem oportunidades na carreira militar decidam se alistar, ou a decisão de se alistar esteja correlacionada com outras características que afetam os ganhos. Isso fará com que *veterano* e *u* sejam correlacionados.

Angrist salientou que o sorteio militar do Vietnã fornecia um **experimento natural** (veja também o Capítulo 13) que criava uma variável instrumental de *veterano*. Foram entregues aos jovens números para sorteio que determinavam se eles seriam chamados para servir no Vietnã. Como os números fornecidos eram aleatoriamente atribuídos, parece possível que os números do sorteio militar fossem não correlacionados com o termo de erro *u*. Entretanto, aqueles que possuíam números baixos teriam que servir no Vietnã, de forma que a probabilidade de ser um veterano estaria correlacionada com os números do sorteio. Se essas duas premissas forem verdadeiras, o número do sorteio militar seria um bom candidato a VI de *veterano*.

QUESTÃO 15.1

Se alguns dos homens que receberam números baixos no sorteio militar obtivessem maior escolaridade para reduzir a probabilidade de serem selecionados, o número do sorteio seria uma boa variável instrumental de *veterano* em (15.18)?

Também é possível ter uma variável explicativa endógena binária e uma variável instrumental binária. Veja o Problema 15.1 no final deste capítulo como exemplo.

Propriedades da VI com uma Variável Instrumental Fraca

Já vimos que, embora a VI seja consistente quando *z* e *u* são não correlacionados e *z* e *x* têm qualquer correlação, positiva ou negativa, as estimativas de VI podem ter grandes erros-padrão, especialmente se *z* e *x* forem apenas fracamente correlacionados. A fraca correlação entre *z* e *x* pode ter consequências ainda mais sérias: o estimador VI pode ter um grande viés assimptótico mesmo se *z* e *u* forem só moderadamente correlacionados.

Podemos verificar isso estudando o limite de probabilidade do estimador VI quando *z* e *u* forem possivelmente correlacionados. Permitindo que $\hat{\beta}_{1,\text{VI}}$ denote o estimador de VI, podemos escrever

$$\text{plim}\,\hat{\beta}_{1,\text{VI}} = \beta_1 + \frac{\text{Corr}(z,u)}{\text{Corr}(z,x)} \cdot \frac{\sigma_u}{\sigma_x}, \qquad (15.19)$$

em que σ_u e σ_x são, respectivamente, os desvios-padrão de *u* e *x* na população. A parte interessante dessa equação envolve os termos de correlação. Ela mostra que, mesmo se Corr(*z*,*u*) for pequena, a inconsistência no estimador VI pode ser muito grande se Corr(*z*,*x*) também for pequena. Assim, mesmo se nos concentrarmos apenas na consistência, não será necessariamente melhor usar VI em lugar de MQO se a correlação entre *z* e *u* for menor que aquela entre *x* e *u*. Utilizando o fato de que Corr(*x*,*u*) = $= \text{Cov}(x,u)/(\sigma_x\sigma_u)$ com a equação (5.3) do Capítulo 5, podemos escrever o plim do estimador MQO — chamando-o de $\hat{\beta}_{1,\text{MQO}}$ — como

$$\text{plim}\,\beta_{1,\text{MQO}} = \beta_1 + \text{Corr}(x,u)\cdot\frac{\sigma_u}{\sigma_x}. \qquad \textbf{(15.20)}$$

A comparação dessas fórmulas mostra que é possível para a direção dos vieses assimptóticos serem diferentes em VI e MQO. Por exemplo, suponha $\text{Corr}(x,u) > 0$, $\text{Corr}(z,x) > 0$, e $\text{Corr}(z,u) < 0$. Então o estimador de VI tem um viés de baixa, enquanto o estimador de MQO tem um viés de alta (assimptoticamente). Na prática, esta situação provavelmente será rara. Mais problemático é quando a direção do viés é a mesma e a correlação entre z e x é pequena. Para melhor clareza, suponha que x e z sejam ambas positivamente correlacionadas com u e $\text{Corr}(z,x) > 0$. Então o viés assimptótico no estimador de VI será menor que o dos MQO somente se $\text{Corr}(z,u)/\text{Corr}(z,x) < \text{Corr}(x, u)$. Se $\text{Corr}(z,x)$ for pequena, então uma correlação aparentemente pequena entre z e u pode ser magnificada e tornar as VI pior que os MQO, mesmo se restringirmos a atenção no viés. Por exemplo, se $\text{Corr}(z,x) = 0{,}2$, $\text{Corr}(z,u)$ deverá ser menor que um quinto de $\text{Corr}(x, u)$ antes que as VI tenha menos viés assimptótico que os MQO. Em muitas aplicações, a correlação entre o instrumento e x é menor que 0,2. Infelizmente, como raramente temos uma ideia sobre as magnitudes relativas de $\text{Corr}(z,u)$ e $\text{Corr}(x,u)$, nunca podemos saber com certeza qual estimador tem o maior viés assimptótico [a menos, claro, que presumamos $\text{Corr}(z,u) = 0$].

No exemplo de Angrist e Krueger (1991) mencionado anteriormente, no qual x representa anos de escolaridade e z é uma variável binária indicando o trimestre de nascimento, a correlação entre z e x é bastante pequena. Bound, Jaeger & Baker (1995) examinaram as razões de o trimestre de nascimento e u terem alguma correlação. Pela equação (15.19), vemos que isso pode levar a um viés substancial no estimador VI.

Quando z e x não têm nenhuma correlação, as coisas ficam particularmente ruins, seja z correlacionado ou não com u. O exemplo seguinte ilustra porque devemos sempre verificar se a variável explicativa endógena é correlacionada com a candidata a VI.

EXEMPLO 15.3

(A Estimação do Efeito do Hábito de Fumar sobre o Peso de Nascimento)

No Capítulo 6, estimamos o efeito do hábito de fumar sobre o peso dos recém-nascidos. Sem outras variáveis explicativas, o modelo é

$$\log(pesonas) = \beta_0 + \beta_1 maços + u, \qquad \textbf{(15.21)}$$

em que *maços* é a quantidade de maços de cigarros fumados pela mãe por dia. Poderíamos suspeitar que *maços* estivesse correlacionado com outros fatores relativos à saúde ou à existência de um bom procedimento pré-natal, de forma que *maços* e u pudessem ser correlacionados. Uma possível variável instrumental de *maços* seria o preço médio dos cigarros no estado de residência, *precig*. Consideraremos que *precig* e u não correlacionados (embora o sistema de saúde estadual possa ser correlacionado com os impostos sobre cigarros).

Se cigarros são um típico produto de consumo, a teoria econômica básica sugere que *maços* e *precig* são negativamente correlacionados, de forma que *precig* pode ser usado como uma VI de *maços*. Para verificar isso, regredimos *maços* sobre *precig*, usando os dados contidos no arquivo BWGHT.RAW:

EXEMPLO 15.3 (continuação)

$$\widehat{maços} = 0{,}067 + 0{,}0003\ precig$$
$$(0{,}103)\quad(0{,}0008)$$
$$n = 1.388,\ R^2 = 0{,}0000,\ \bar{R}^2 = -0{,}0006.$$

Isso não indica qualquer relação entre o hábito de fumar durante a gravidez e o preço dos cigarros, o que talvez não seja tão surpreendente devido à natureza dependente do hábito de fumar.

Como *maços* e *precig* não são correlacionados, não devemos usar *precig* como uma VI de *maços* em (15.21). Mas o que acontece se o fizermos? Os resultados da VI seriam

$$\widehat{\log(pesonas)} = 4{,}45 + 2{,}99\ maços$$
$$(0{,}91)\quad(8{,}70)$$
$$n = 1.388$$

(o R-quadrado obtido é negativo). O coeficiente de *maços* é enorme e tem um sinal inesperado. O erro-padrão também é muito grande, de modo que *maços* não é significante. Entretanto, as estimativas não têm significado, pois *precig* não atende o único requisito de um VI que sempre podemos testar: a Hipótese (15.5).

O exemplo anterior mostra que a estimação de VI pode produzir resultados estranhos quando a condição de relevância instrumental, $\text{Corr}(z,x) \neq 0$, falha. De maior interesse prático é o assim chamado problema de **instrumento fraco**, que é livremente definido como o problema de "baixa" (mas não zero) correlação entre z e x. Numa aplicação particular, é difícil definir quão baixa será baixa demais, mas pesquisas teóricas recentes, suplementadas por estudos de simulação, têm esclarecido consideravelmente este problema. Staiger e Stock (1997) formalizaram o problema dos instrumentos fracos modelando a correlação entre z e x como uma função do tamanho da amostra; em particular, considera-se que a correlação encolhe para zero à razão $1/\sqrt{n}$. Sem surpresa, a distribuição assimptótica do estimador de variáveis instrumentais será diferente quando comparada com os assimptóticos habituais, em que a correlação é presumida como fixa e não zero. Uma das implicações do trabalho de Stock-Staiger é que a inferência estatística habitual, baseada em estatísticas t e na distribuição normal padrão, pode ser seriamente enganosa. [Veja Imbens e Wooldridge (2007) para mais informações.]

O Cálculo do *R*-Quadrado após a Estimação de VI

A maioria dos programas de regressão calcula o R-quadrado após a estimação de VI, utilizando a fórmula padrão: $R^2 = 1 - \text{SQR/SQT}$, em que SQR é a soma dos quadrados dos resíduos da VI e SQT é a soma dos quadrados total de y. Diferentemente do caso MQO, o R-quadrado da estimação de VI pode ser negativo, pois a SQR da VI pode, na realidade, ser maior que a SQT. Embora não faça mal algum descrever o R-quadrado da estimação de VI, ele não é muito útil. Quando x e u são correlacionados, não podemos decompor a variância de y em $\beta_1^2\,\text{Var}(x) + \text{Var}(u)$ e, assim, o R-quadrado não possui interpretação natural. Além disso, como veremos na Seção 15.3, esses R-quadrados *não podem* ser usados da maneira habitual para calcular testes F de restrições conjuntas.

Se nossa meta for produzir o maior R-quadrado, sempre usaremos MQO. Os métodos das VI são destinados a produzir melhores estimativas do efeito *ceteris paribus* de x sobre y quando x e u forem

correlacionados; a qualidade de ajuste não é um fator importante. Um alto *R*-quadrado resultante do MQO é de pouca ajuda se não pudermos estimar consistentemente β_1.

15.2 ESTIMAÇÃO DE VI DO MODELO DE REGRESSÃO MÚLTIPLA

O estimador de VI para o modelo de regressão simples é facilmente estendido para o caso da regressão múltipla. Começamos com o caso no qual somente uma das variáveis explicativas é correlacionada com o erro. Considere um modelo linear padrão com duas variáveis explicativas:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1, \tag{15.22}$$

Chamamos essa equação de **equação estrutural**, para enfatizar que estamos interessados em β_j, o que simplesmente significa que a equação supostamente mede uma relação causal. Nesse caso, usamos uma nova notação para distinguir as variáveis endógenas das **variáveis exógenas**. A variável dependente y_1 é claramente endógena, já que é correlacionada com u_1. As variáveis y_2 e z_1 são as variáveis explicativas, e u_1 é o erro. Como sempre, presumimos que o valor esperado de u_1 é zero: $E(u_1) = 0$. Usamos z_1 para indicar que essa variável é exógena em (15.22) (z_1 é não correlacionado com u_1). Usamos y_2 para indicar que esta variável é suspeita de ser correlacionada com u_1. Não especificamos porque y_2 e u_1 são correlacionados, mas por enquanto é melhor pensar em u_1 contendo uma variável omitida correlacionada com y_2. A notação na equação (15.22) tem origem em modelos de equações simultâneas (que trataremos no Capítulo 16), mas a usamos de forma mais genérica para facilmente distinguir variáveis explicativas exógenas de endógenas, em um modelo de regressão múltipla.

Um exemplo de (15.22) é

$$\log(salário) = \beta_0 + \beta_1 educ + \beta_2 exper + u_1, \tag{15.23}$$

em que $y_1 = \log(salário)$, $y_2 = educ$, e $z_1 = exper$. Em outras palavras, presumimos que *exper* é exógeno em (15.23), mas permitimos que *educ* — pelas razões habituais — seja correlacionado com u_1.

Sabemos que, se (15.22) for estimada por MQO, *todos* os estimadores serão viesados e inconsistentes. Assim, seguimos a estratégia sugerida na seção anterior e procuramos uma variável instrumental de y_2. Como consideramos z_1 não correlacionado com u_1, podemos usar z_1 como instrumento de y_2, presumindo que y_2 e z_1 sejam correlacionados? A resposta é não. Como a própria z_1 aparece como uma variável explicativa em (15.22), ela não pode servir como uma variável instrumental de y_2. Precisamos de outra variável exógena — vamos chamá-la z_2 — que *não* apareça em (15.22). Portanto, as hipóteses fundamentais são que z_1 e z_2 são não correlacionados com u_1; também presumimos que u_1 tem zero como valor esperado, o que não provoca perda de generalidade quando a equação contém um intercepto:

$$E(u_1) = 0,\ \mathrm{Cov}(z_1,u_1) = 0,\ \text{e}\ \mathrm{Cov}(z_2,u_1) = 0. \tag{15.24}$$

Em razão da hipótese de média zero, as últimas duas hipóteses são equivalentes a $E(z_1,u_1) = E(z_2,u_1) = 0$, e assim a abordagem do método dos momentos sugere a obtenção dos estimadores $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_2$, resolvendo os correspondentes amostrais de (15.24):

$$\begin{aligned}
\sum_{i=1}^{n} (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) &= 0 \\
\sum_{i=1}^{n} z_{i1}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) &= 0 \\
\sum_{i=1}^{n} z_{i2}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) &= 0.
\end{aligned} \tag{15.25}$$

Esse é um conjunto de três equações lineares com três incógnitas $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_2$, e é facilmente resolvido considerando os dados de y_1, y_2, z_1 e z_2. Os estimadores são chamados *estimadores de variáveis instrumentais*. Se entendermos que y_2 é exógeno e escolhermos $z_2 = y_2$, as equações em (15.25) serão exatamente as condições de primeira ordem dos estimadores MQO; veja as equações em (3.13) no Capítulo 3.

Ainda necessitamos que a variável instrumental z_2 seja correlacionada com y_2, mas o sentido como essas duas variáveis devem ser correlacionadas é complicado pela presença de z_1 na equação (15.22). Agora precisamos estabelecer a hipótese em termos de correlação *parcial*. A maneira mais fácil de definir a condição é escrever a variável explicativa endógena como uma função linear das variáveis exógenas e um termo de erro:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2, \tag{15.26}$$

em que, por definição, $E(v_2) = 0$, $\mathrm{Cov}(z_1,v_2) = 0$ e $\mathrm{Cov}(z_2,v_2) = 0$, e os π_j são parâmetros desconhecidos. A condição de identificação fundamental [com (15.24)] é que

$$\pi_2 \neq 0 \tag{15.27}$$

QUESTÃO 15.2

Suponha que queiramos estimar o efeito do uso de maconha na nota média de graduação. Em uma população de alunos veteranos de uma universidade, defina *diasusados* como o número de dias no mês anterior em que um aluno fumou maconha e considere a equação estrutural

$$supGPA = \beta_0 + \beta_1 diasusados + \beta_2 SAT + u.$$

(i) Defina *percento* como a percentagem de uma sala do ensino médio que informou uso regular de maconha. Se ela for uma candidata a VI de *diasusados*, escreva a forma reduzida de *diasusados*. Você acha que é possível (15.27) ser verdadeira?

(ii) Você acha que *percento* é verdadeiramente exógena na equação estrutural? Que problemas podem surgir nesse caso?

Em outras palavras, após considerar os efeitos parciais, z_1, y_2 e z_2 ainda são correlacionados. Essa correlação pode ser positiva ou negativa, mas não pode ser zero. Testar (15.27) é fácil: estimamos (15.26) por MQO e usamos um teste *t* (possivelmente tornando-o robusto quanto à heteroscedasticidade). Devemos sempre testar essa hipótese. Infelizmente, não podemos testar se z_1 e z_2 são não correlacionados com u_1; mas, confiantemente, podemos realizar o processo com base no raciocínio econômico ou intuição.

A equação (15.26) é um exemplo de uma **equação na forma reduzida**, significando que escrevemos uma variável endógena em termos de variáveis exógenas. Esse nome deriva de modelos de equações simultâneas — que estudaremos no próximo capítulo —, e é um conceito útil sempre que tivermos uma variável explicativa endógena. O nome ajuda a distingui-la da equação estrutural (15.22).

A adição de mais **variáveis explicativas exógenas** ao modelo é direta. Escreva o modelo estrutural como

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \ldots + \beta_k z_{k-1} + u_1, \qquad \textbf{(15.28)}$$

em que y_2 é pensado para ser correlacionado com u_1. Defina z_k como uma variável não pertencente a (15.28) que também seja exógena. Portanto, presumimos que

$$\mathrm{E}(u_1) = 0, \ \mathrm{Cov}(z_j, u_1) = 0, \ j = 1, \ldots, k. \qquad \textbf{(15.29)}$$

Sob a (15.29), $z_1, \ldots, z_{k-1}$ são variáveis exógenas que aparecem na (15.28). Na verdade, elas agem como suas próprias variáveis instrumentais na estimativa da β_j na (15.28). O caso especial de $k = 2$ é dado na equação na (15.25); juntamente com z_2, z_1 aparece no conjunto de condições de momento usado para se obter as estimativas das VI. De forma mais geral, $z_1, \ldots, z_{k-1}$ são usadas nas condições de momento juntamente com a variável instrumental da y_2, z_k.

A forma reduzida de y_2 é

$$y_2 = \pi_0 + \pi_1 z_1 + \ldots + \pi_{k-1} z_{k-1} + \pi_k z_k + v_2, \qquad \textbf{(15.30)}$$

e necessitamos de alguma correlação parcial entre z_k e y_2:

$$\pi_k \neq 0. \qquad \textbf{(15.31)}$$

Sob (15.29) e (15.31), z_k é uma VI válida de y_2. [Não nos importamos com os restantes π_j em (15.30); alguns ou todos eles podem ser zero.] Uma hipótese secundária adicional é que não há relações lineares perfeitas entre as variáveis exógenas; isso é análogo à hipótese da não existência de colinearidade perfeita no contexto de MQO.

Para a inferência estatística padrão, precisamos presumir a homoscedasticidade de u_1. Faremos uma descrição mais cuidadosa dessas hipóteses, em um cenário mais geral, na Seção 15.3.

EXEMPLO 15.4

(Utilizando a Proximidade da Faculdade como uma VI da Educação)

Card (1995) usou dados sobre salários e educação de uma amostra de homens em 1976 para estimar o retorno da educação. Ele usou uma variável *dummy* para o caso de alguém que tenha crescido perto de uma faculdade com cursos de graduação de quatro anos (*proxf4*) ser uma variável instrumental da educação. Em uma equação log(*salárioh*), ele incluiu outros controles padrão: experiência, uma variável *dummy* para negro, variáveis *dummy* para o caso de a pessoa residir em área metropolitana (EPRM) e residir no sul, e um conjunto completo de variáveis *dummy* regionais e uma *dummy* para o fato de residir em área metropolitana em 1966. Para *proxf4* ser uma variável instrumental válida, deve ser não correlacionada com o termo de erro na equação de salários-hora — o que presumimos — e deve ser parcialmente correlacionada com *educ*.

EXEMPLO 15.4 (continuação)

Para verificar o último requisito, regredimos *educ* sobre *proxf4* e todas as variáveis exógenas que aparecem na equação. (Isto é, estimamos a forma reduzida de *educ*.) Utilizando os dados contidos no arquivo CARD.RAW, obtemos, de forma condensada,

$$\widehat{educ} = 16{,}64 + 0{,}320\, proxf4 - 0{,}413\, exper + \ldots \qquad \textbf{(15.32)}$$
$$(0{,}24) \quad (0{,}088) \qquad (0{,}034)$$
$$n = 3.010, \ R^2 = 0{,}477.$$

Tabela 15.1

Variável dependente: log(*salárioh*).

Variáveis explicativas	MQO	VI
educ	0,075 (0,003)	0,132 (0,055)
exper	0,085 (0,007)	0,108 (0,024)
$exper^2$	−0,0023 (0,0003)	−0,0023 (0,0003)
negro	−0,199 (0,018)	−0,147 (0,054)
eprm	0,136 (0,020)	0,112 (0,032)
sul	−0,148 (0,026)	−0,145 (0,027)
Observações *R*-Quadrado	3.010 0,300	3.010 0,238
Outros controles: *eprm66*, *reg662*, ..., *reg669*		

Estamos interessados no coeficiente e na estatística *t* de *proxf4*. O coeficiente implica em 1976, com os outros fatores fixos (experiência, etnia, região etc.), as pessoas que residiam próximas de uma faculdade em 1966 tinham, em média, quatro meses a mais de estudo do que os que cresceram em áreas distantes de faculdades. A estatística *t* de *proxf4* é 3,64, o que produz um *p*-valor que é zero nas primeiras três casas decimais. Portanto, se *proxf4* for não correlacionado com fatores não observados no termo de erro, poderemos usar *proxf4* como uma VI de *educ*.

EXEMPLO 15.4 (continuação)

As estimativas de MQO e VI são mostradas na Tabela 15.1. Curiosamente, a estimativa de VI do retorno da educação é duas vezes maior que a do MQO, mas o erro-padrão da estimativa de VI é 18 vezes maior que do MQO. O intervalo de confiança de 95% da estimativa de VI é de 0,024 e 0,239, o que é uma faixa muito ampla. Intervalos de confiança maiores representam um preço que temos de pagar para obter um estimador consistente do retorno da educação quando acreditamos que *educ* é uma variável endógena.

Como discutido anteriormente, não devemos nos importar com o *R*-quadrado menor na estimação de VI: por definição, o *R*-quadrado do MQO será sempre maior, pois o MQO minimiza a soma dos quadrados dos resíduos.

15.3 MÍNIMOS QUADRADOS EM DOIS ESTÁGIOS

Na seção anterior, presumimos que tínhamos uma única variável explicativa endógena (y_2), juntamente com uma variável instrumental de y_2. Acontece com frequência que temos mais de uma variável exógena que é excluída do modelo estrutural e que pode ser correlacionada com y_2, e significa que são VIs válidas de y_2. Nesta seção, discutiremos como usar variáveis instrumentais múltiplas.

Uma Única Variável Explicativa Endógena

Considere novamente o modelo estrutural (15.22), que tem uma variável explicativa endógena e uma exógena. Suponha agora que temos *duas* variáveis exógenas excluídas de (15.22): z_2 e z_3. Nossas hipóteses de z_2 e z_3 não aparecerem em (15.22) e por serem não correlacionadas com o erro u_1 são conhecidas como **restrições de exclusão**.

Se z_2 e z_3 forem ambas correlacionadas com y_2, poderemos simplesmente usar cada uma delas como uma VI, como na seção anterior. Entretanto, nesse caso, teríamos dois estimadores de VI, e nenhum deles seria, de forma geral, eficiente. Como z_1, z_2 e z_3 são não correlacionadas com u_1, qualquer combinação linear também será não correlacionada com u_1, e, portanto, qualquer combinação linear das variáveis exógenas será uma VI válida. Para encontrar a melhor VI, escolhemos a combinação linear que seja mais altamente correlacionada com y_2. Isso acaba sendo fornecido pela equação na forma reduzida de y_2. Escreva

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2, \qquad \textbf{(15.33)}$$

em que

$$\mathrm{E}(v_2) = 0, \mathrm{Cov}(z_1, v_2) = 0, \mathrm{Cov}(z_2, v_2) = 0, \text{ e } \mathrm{Cov}(z_3, v_2) = 0.$$

Portanto, a melhor VI de y_2 (sob as hipóteses dadas no apêndice deste capítulo) é a combinação linear dos z_j em (15.33) que chamamos y_2^*:

$$y_2^* = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3. \qquad \textbf{(15.34)}$$

Para que esta VI não seja perfeitamente correlacionada com z_1, precisamos que pelo menos um π_2 ou π_3 seja diferente de zero:

$$\pi_2 \neq 0 \text{ ou } \pi_3 \neq 0. \qquad \textbf{(15.35)}$$

Essa é a hipótese fundamental de identificação, tão logo presumamos que todos os z_j sejam exógenos. (O valor de π_1 é irrelevante.) A equação estrutural (15.22) não será identificada se $\pi_2 = 0$ e $\pi_3 = 0$. Podemos testar H_0: $\pi_2 = 0$ e $\pi_3 = 0$ contra (15.35), usando uma estatística F.

Uma maneira útil de pensar em (15.33) é que ela divide y_2 em duas partes. A primeira é y_2^*; esta é a parte de y_2 que é não correlacionada com o termo de erro u_1. A segunda é v_2, e esta parte é possivelmente correlacionada com u_1, razão — pela qual y_2 é possivelmente endógeno.

Graças aos dados de z_j, podemos calcular y_2^*para cada observação, desde que conheçamos os parâmetros populacionais π_j. Na prática, isso nunca é real. No entanto, como vimos na seção anterior, sempre podemos estimar a forma reduzida por MQO. Assim, usando a amostra, regredimos y_2 sobre z_1, z_2 e z_3 e obtemos os valores estimados:

$$\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 z_3 \qquad \textbf{(15.36)}$$

(isto é, temos $\hat{y}_{i2}$ para cada i). Neste ponto, devemos verificar se z_2 e z_3 são conjuntamente significantes em (15.33) a um nível de significância razoavelmente pequeno (não mais que 5%). Se z_2 e z_3 não forem conjuntamente significantes em (15.33), estaremos perdendo nosso tempo com a estimação de VI.

Uma vez obtido $\hat{y}_2$, podemos usá-lo como VI de y_2. As três equações para estimar β_0, β_1 e β_2 são as duas primeiras de (15.25) e a terceira substituída por

$$\sum_{i=1}^{n} \hat{y}_{i2}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0. \qquad \textbf{(15.37)}$$

A solução das três equações com três incógnitas nos fornece os estimadores de VI.

Com instrumentos múltiplos, o estimador de VI que usa $\hat{y}_{i2}$ como instrumento também é chamado **estimador em mínimos quadrados em dois estágios (MQ2E)**. A razão é simples. Usando a álgebra do MQO, pode ser demonstrado que, quando usamos $\hat{y}_2$ como VI de y_2, as estimativas VI de $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_2$ são *idênticas* às estimadas por MQO da regressão de

$$\hat{y}_1 \text{ sobre } \hat{y}_2 \text{ e } z_1. \qquad \textbf{(15.38)}$$

Em outras palavras, podemos obter o estimador MQ2E em dois estágios. O primeiro estágio é executar a regressão em (15.36), no qual obteremos os valores estimados $\hat{y}_2$. O segundo estágio é a regressão MQO (15.38). Como usamos $\hat{y}_2$ em lugar de y_2, as estimativas MQ2E podem diferir substancialmente das estimativas MQO.

Alguns economistas gostam de interpretar a regressão em (15.38) da seguinte maneira: o valor estimado, $\hat{y}_2$, é a versão estimada de y_2^*, e y_2^* é não correlacionado com u_1. Portanto, o MQ2E primeiro "expurga" y_2 de sua correlação com u_1 antes de fazer a regressão MQO em (15.38). Podemos mostrar isso inserindo $y_2 = y_2^* + v_2$ em (15.22):

$$y_1 = \beta_0 + \beta_1 y_2^* + \beta_2 z_1 + u_1 + \beta_1 v_2. \qquad \textbf{(15.39)}$$

Agora o erro composto $u_1 + \beta_1 v_2$ tem média zero e é não correlacionado com y_2^* e z_1, motivo pelo qual a regressão MQO em (15.38) funciona corretamente.

A maioria dos programas econométricos possui comandos especiais de MQ2E, de modo que não há necessidade de fazer explicitamente as duas etapas. Aliás, na maioria dos casos, devemos evitar fazer o segundo estágio manualmente, já que os erros-padrão e as estatísticas de testes obtidos dessa maneira *não* são válidos. [A razão é que o termo de erro em (15.39) inclui v_2, mas o erro-padrão envolve somente a variância de u_1.] Qualquer programa de regressão que suporte MQ2E solicita a variável dependente, a relação das variáveis explicativas (tanto exógenas como endógenas) e a relação total das variáveis instrumentais (isto é, todas as variáveis exógenas). A descrição dos resultados em geral é bastante semelhante à do MQO.

No modelo (15.28) com uma única VI de y_2, o estimador de VI da Seção 15.2 é idêntico ao estimador MQ2E. Portanto, quando temos uma VI de cada variável explicativa endógena, podemos chamar o método de estimação de VI ou MQ2E.

A adição de mais variáveis exógenas altera muito pouco o processo. Por exemplo, suponha que a equação de salários seja

$$\log(salário\text{h}) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u_1, \qquad \textbf{(15.40)}$$

em que u_1 é não correlacionado tanto com *exper* como com $exper^2$. Suponha que também entendemos que a escolaridade tanto da mãe como do pai seja não correlacionada com u_1. Então, podemos usar as duas variáveis como VIs de *educ*. A forma reduzida da equação de *educ* é

$$educ = \pi_0 + \pi_1 exper + \pi_2 exper^2 + \pi_3 educm + \pi_4 educp + v_2, \qquad \textbf{(15.41)}$$

e a identificação exige que $\pi_3 \neq 0$ ou $\pi_4 \neq 0$ (naturalmente, ou ambas).

EXEMPLO 15.5

(Retorno da Educação para Mulheres que Trabalham)

Estimamos a equação (15.40) utilizando os dados contidos no arquivo MROZ.RAW. Primeiro, testamos H_0: $\pi_3 = 0$, $\pi_4 = 0$ em (15.41) usando um teste *F*. O resultado é $F = 55{,}40$, e *p*-valor = 0,0000. Como esperado, *educ* é (parcialmente) correlacionado com a educação dos pais.

Quando estimamos (15.40) por MQ2E, obtemos:

$$\widehat{\log(salário\text{h})} = 0{,}048 + 0{,}061\ educ + 0{,}044\ exper - 0{,}0009\ exper^2$$
$$(0{,}400)\quad (0{,}031)\qquad (0{,}013)\qquad (0{,}0004)$$
$$n = 428,\ R^2 = 0{,}136.$$

O retorno da educação estimado está em torno de 6,1%, comparado com uma estimativa por MQO de cerca de 10,8%. Em razão de seu erro-padrão relativamente grande, a estimativa MQ2E é pouco significante no nível de 5% contra uma alternativa bilateral.

As hipóteses necessárias para que o MQ2E tenha as desejadas propriedades de amostras grandes são fornecidas no Apêndice deste capítulo, mas é útil fazermos aqui um breve resumo delas. Se escrevermos a equação estrutural como em (15.28),

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \ldots + \beta_k z_{k-1} + u_1, \qquad \textbf{(15.42)}$$

então, presumimos que cada z_j é não correlacionado com u_1. Além disso, precisamos de pelo menos uma variável exógena *fora* de (15.42) que seja parcialmente correlacionada com y_2. Isso garantirá a consistência. Para que os habituais erros-padrão e estatísticas *t* do MQ2E sejam assimptoticamente válidos, também necessitamos de uma hipótese de homoscedasticidade: a variância do erro estrutural, u_1, não pode depender de qualquer das variáveis exógenas. Para aplicações de séries temporais, precisaremos de mais hipóteses, como veremos na Seção 15.7.

Multicolinearidade e MQ2E

No Capítulo 3, apresentamos o problema da multicolinearidade e mostramos como a correlação entre os regressores pode levar a erros-padrão grandes das estimativas MQO. A multicolinearidade pode ser ainda mais grave com o MQ2E. Para verificar a razão disso, a variância (assimptótica) do estimador MQ2E de β_1 pode ser aproximada como

$$\frac{\sigma^2}{[\widehat{\text{SQT}}_2(1 - \hat{R}_2^2)]}, \qquad \textbf{(15.43)}$$

em que $\sigma^2 = \text{Var}(u_1)$, $\widehat{\text{SQT}}_2$ é a variação total em $\hat{y}_2$, e $\hat{R}_2^2$ é o *R*-quadrado de uma regressão de $\hat{y}_2$ sobre todas as outras variáveis exógenas que aparecem na equação estrutural. Há duas razões para a variância do estimador MQ2E ser maior do que a do MQO. Primeiro, $\hat{y}_2$, por construção, tem menos variação que y_2. (Lembre-se: soma dos quadrados total = soma dos quadrados explicada + soma dos quadrados dos resíduos; a variação em y_2 é a soma dos quadrados total, enquanto a variação em $\hat{y}_2$ é a soma dos quadrados explicada da primeira etapa da regressão.) Segundo, a correlação entre $\hat{y}_2$ e as variáveis exógenas em (15.42) frequentemente é muito mais elevada que a correlação entre y_2 e aquelas variáveis. Esse fato define essencialmente o problema da multicolinearidade no MQ2E.

A título de ilustração, considere o Exemplo 15.4. Quando *educ* é regredido sobre as variáveis exógenas na Tabela 15.1 (sem a inclusão de *proxf4*), *R*-quadrado = 0,475; isso reflete um grau moderado de multicolinearidade, mas o importante é que o erro-padrão MQO em $\hat{\beta}_{educ}$ é bem pequeno. Quando obtemos os valores estimados do primeiro estágio, $\widehat{educ}$, e regredimos esses valores sobre as variáveis exógenas na Tabela 15.1, *R*-quadrado = 0,995, o que indica um grau bastante alto de multicolinearidade entre $\widehat{educ}$ e as demais variáveis exógenas na tabela. (Esse elevado *R*-quadrado não surpreende, pois $\widehat{educ}$ é uma função de todas as variáveis exógenas da Tabela 15.1, mais *proxf4*.) A equação (15.43) mostra que um $\hat{R}_2^2$ próximo de um pode resultar num erro-padrão bastante grande do estimador MQ2E. Entretanto, como no caso do MQO, uma amostra de tamanho grande pode ajudar a compensar um grande $\hat{R}_2^2$.

Variáveis Explicativas Endógenas Múltiplas

O método de mínimos quadrados em dois estágios também pode ser usado em modelos com mais de uma variável explicativa endógena. Por exemplo, considere o modelo

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 y_3 + \beta_3 z_1 + \beta_4 z_2 + \beta_5 z_3 + u_1, \quad \textbf{(15.44)}$$

em que $E(u_1) = 0$ e u_1 é não correlacionado com z_1, z_2 e z_3. As variáveis y_2 e y_3 são variáveis explicativas endógenas: cada uma delas pode ser correlacionada com u_1.

Para estimar (15.44) por MQ2E, precisamos de *pelo menos duas* variáveis exógenas que não apareçam em (15.44), mas que sejam correlacionadas com y_2 e y_3. Suponha que temos duas variáveis exógenas excluídas, digamos z_4 e z_5. Então, a partir de nossa análise de uma única variável explicativa endógena, precisaremos que z_4 ou z_5 apareça em cada uma das formas reduzidas de y_2 e y_3. (Como antes, podemos usar estatísticas F para fazer o teste.) Embora isso seja necessário para a identificação, infelizmente não é suficiente. Suponha que z_4 apareça em cada uma das formas reduzidas, mas que z_5 não apareça em nenhuma. Então, na realidade, não teremos duas variáveis exógenas parcialmente correlacionadas com y_2 e y_3. O estimador de mínimos quadrados em dois estágios não produzirá estimadores consistentes dos β_j.

De forma geral, quando temos mais de uma variável explicativa endógena em um modelo de regressão, a identificação pode falhar de várias e complicadas maneiras. Contudo, podemos facilmente estabelecer uma condição necessária para a identificação, que é chamada de **condição de ordem**.

Condição de Ordem para Identificação de uma Equação. Necessitamos ao menos tantas variáveis exógenas excluídas quantas forem as variáveis explicativas endógenas incluídas na equação estrutural. A condição de ordem é fácil de ser verificada, já que somente consiste em contar as variáveis endógenas e exógenas. A condição suficiente para a identificação é chamada de **condição de classificação**. Já vimos casos especiais da condição de classificação, por exemplo, na discussão em torno da equação (15.35). Uma expressão geral da condição de classificação exige álgebra matricial e está além do escopo deste texto. [Veja Wooldridge (2002, Capítulo 5).]

QUESTÃO 15.3

O modelo seguinte explica a taxa de crimes violentos, ao nível de cidades, em termos de uma variável binária que indica se existem leis de controle de armas e outros controles:

$$violento = \beta_0 + \beta_1 controledearmas + \beta_2 desemp + \beta_3 popul + \beta_4 porcnegro + \beta_5 idade18\_21 + \ldots$$

Alguns pesquisadores estimaram equações semelhantes usando variáveis como o número de membros da *National Rifle Association* (Associação Nacional do Rifle) na cidade e o número de assinantes de revistas sobre armas como variáveis instrumentais de *controledearmas* [veja, por exemplo, Kleck e Patterson (1993)]. Essas variáveis instrumentais são convincentes?

O Teste de Hipóteses Múltiplas após a Estimação por MQ2E

Devemos ser cuidadosos ao testar hipóteses múltiplas em um modelo estimado por MQ2E. É tentador usar a soma dos quadrados dos resíduos ou a forma R-quadrado da estatística F, como aprendemos ao estudar o MQO no Capítulo 4. O fato de o R-quadrado do MQ2E poder ser negativo sugere que a maneira habitual de calcular estatísticas F pode não ser apropriada; esse é precisamente o caso. De fato, se usarmos os resíduos do MQ2E para calcular os SQRs de ambos os modelos, o restrito e o sem restrições, não haverá garantia de que $SQR_r \geq SQR_{ir}$; se o inverso for verdadeiro, a estatística F será negativa.

É possível combinar a soma dos quadrados dos resíduos da regressão do segundo estágio [tal como (15.38)] com SQR_{ir} para obter uma estatística com uma distribuição aproximadamente F em amostras grandes. Como muitos programas econométricos incorporam comandos de testes de uso fácil, que podem ser empregados para testar hipóteses múltiplas após a estimação por MQ2E, omitimos os detalhes. Davidson e MacKinnon (1993) e Wooldridge (2002, Capítulo 5) oferecem discussões sobre como calcular estatísticas do tipo F do MQ2E.

15.4 SOLUÇÕES DE VI PARA PROBLEMAS DE ERROS NAS VARIÁVEIS

Nas seções anteriores, apresentamos o uso de variáveis instrumentais como um meio de solucionar o problema de variáveis omitidas, mas elas também podem ser usadas para tratar o problema de erro de medida. A título de ilustração, considere o modelo

$$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + u, \quad \textbf{(15.45)}$$

em que y e x_2 são observados, mas x_1^*, não. Seja x_1 uma medida observada de x_1^*: $x_1 = x_1^* + e_1$, sendo e_1 o erro de medida. No Capítulo 9, mostramos que a correlação entre x_1 e e_1 faz com que o MQO, onde x_1 é usado em lugar de x_1^*, seja viesado e inconsistente. Podemos verificar isso escrevendo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + (u - \beta_1 e_1). \quad \textbf{(15.46)}$$

Se as hipóteses dos erros clássicos nas variáveis (ECV) se mantiverem, o viés no estimador MQO de β_1 tenderá a zero. Sem hipóteses adicionais, não podemos fazer nada a esse respeito.

Em alguns casos, é possível usar um procedimento de VI para solucionar o problema de erro de medida. Em (15.45), presumimos que u é não correlacionado com x_1^*, x_1 e x_2; no caso de ECV, presumimos que e_1 é não correlacionado com x_1^* e x_2. Assim x_2 é exógeno em (15.46), mas x_1 é correlacionado com e_1. O que precisamos é de uma VI de x_1. Tal VI deve ser correlacionada com x_1, não correlacionada com u — de forma que possa ser excluída de (15.45) — e não correlacionada com o erro de estimação, e_1.

Uma possibilidade é obter uma segunda estimação de x_1^*, digamos z_1. Como é x_1^* que afeta y, é natural pressupor que z_1 é não correlacionado com u. Se escrevermos $z_1 = x_1^* + a_1$, em que a_1 é o erro de medida em z_1, devemos presumir que a_1 e e_1 são não correlacionados. Em outras palavras, tanto x_1 como z_1 medem incorretamente x_1^*, mas seus erros de medida são não correlacionados. Certamente, x_1 e z_1 são correlacionados por suas dependências de x_1^*, de modo que podemos usar z_1 como uma VI de x_1.

Onde podemos obter duas medidas de uma variável? Algumas vezes, quando um grupo de trabalhadores questiona sobre salário anual, seus empregadores poderão fazer uma contraoferta. No caso de pessoas casadas, cada cônjuge pode informar independentemente o nível de poupança ou renda familiar. No estudo de Ashenfelter e Krueger (1994) citado na Seção 14.3, foi solicitado a cada gêmeo o número de anos de educação formal de seu(ua) irmão(ã); isso fornece uma segunda medida que pode ser usada como uma VI da educação formal autorrelatada em equação de salários. (Ashenfelter e Krueger combinaram diferença e VI para explicar também o problema da aptidão omitida; veja mais sobre esse assunto na Seção 15.8.) Geralmente, no entanto, é raro ter duas medidas de uma variável explicativa.

Uma alternativa é usar outras variáveis exógenas como VIs de uma variável potencialmente mal medida. Por exemplo, nosso uso de *educm* e *educp* como VIs de *educ* no Exemplo 15.5 pode servir a esse propósito. Se pensarmos que $educ = educ^* + e_1$, as estimativas de VI no Exemplo 15.5 não sofrerão do erro de medida se *educm* e *educp* forem não correlacionados com o erro de medida, e_1. Isso provavelmente é mais razoável do que presumir serem *educm* e *educp* não correlacionados com a aptidão, o que está contido em u, em (15.45).

Os métodos VI também podem ser adotados quando usamos fatores como notas de testes para controlar características não observadas. Na Seção 9.2 do Capítulo 9, mostramos que, sob certas hipóteses, variáveis *proxy* podem ser usadas para solucionar o problema de variáveis omitidas. No Exemplo 9.3, usamos o QI como uma variável *proxy* da aptidão não observada. Isso simplesmente envolve a adição do QI ao modelo e a execução de uma regressão MQO. Entretanto, existe uma alternativa que funciona quando o QI não satisfaz plenamente as hipóteses da variável *proxy*. Como ilustração, escreva uma equação de salários como

$$\log(salário\text{h}) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + aptid + u, \qquad \textbf{(15.47)}$$

em que novamente temos o problema da aptidão omitida. Contudo, temos duas notas de testes que são *indicadores* da aptidão. Presumimos que as notas possam ser escritas como

$$teste_1 = \gamma_1 aptid + e_1$$

e

$$teste_2 = \delta_1 aptid + e_2,$$

em que $\gamma_1 > 0$, $\delta_1 > 0$. Como a aptidão é que afeta o salário, podemos supor que $teste_1$ e $teste_2$ são não correlacionados com u. Se escrevermos *aptid* em termos da nota do primeiro teste e inserirmos o resultado em (15.47), obteremos

$$\log(salário\text{h}) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \alpha_1 teste_1 + (u - \alpha_1 e_1), \qquad \textbf{(15.48)}$$

em que $\alpha_1 = 1/\gamma_1$. Agora, se presumirmos que e_1 é não correlacionado com todas as variáveis explicativas em (15.47), incluindo *aptid*, e_1 e $teste_1$ *devem* ser correlacionados. [Observe que *educ não é* endógeno em (15.48); porém, $teste_1$ o é.] Isso significa que estimar (15.48) por MQO produzirá estimadores inconsistentes dos β_j (e α_1). Sob as hipóteses que levantamos, $teste_1$ não satisfaz as hipóteses da variável *proxy*.

Se presumirmos que e_2 também é não correlacionado com todas as variáveis explicativas em (15.47) *e* que e_1 e e_2 são não correlacionados, e_1 será não correlacionado com a segunda nota de testes, $teste_2$. Portanto, $teste_2$ pode ser usado como uma VI de $teste_1$.

EXEMPLO 15.6

(O Uso de Duas Notas de Testes como Indicadores de Aptidão)

Usamos os dados contidos no arquivo WAGE2.RAW para implementar o procedimento precedente, no qual *QI* desempenha o papel de primeira nota de teste e *KWW* (conhecimento do mundo do trabalho) é a segunda nota de teste. As variáveis explicativas são as mesmas do Exemplo 9.3: *educ, exper, perm, casado, sul, urban* e *negro*. Em vez de adicionar *QI* e fazer o MQO, como na coluna (2) da Tabela 9.2, adicionamos *QI* e usamos *KWW* como sua variável instrumental. O coeficiente de *educ* é 0,025 ($ep = 0{,}017$). Essa é uma estimativa baixa, e não é estatisticamente diferente de zero. É um resultado problemático, sugerindo que uma de nossas hipóteses não se sustenta; talvez e_1 e e_2 sejam correlacionados.

15.5 O TESTE DE ENDOGENEIDADE E O TESTE DE RESTRIÇÕES SOBREIDENTIFICADORAS

Nesta seção, descreveremos dois importantes testes no contexto de estimação de variáveis instrumentais.

O Teste de Endogeneidade

O estimador MQ2E é menos eficiente que o MQO quando as variáveis explicativas são exógenas; como vimos, as estimativas MQ2E podem ter erros-padrão elevados. Portanto, é útil fazer um teste de endogeneidade de uma variável explicativa que mostre se o MQ2E é ainda necessário. Obter tal teste é bastante simples.

Para ilustrar, suponha que temos uma única variável suspeita de ser endógena,

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

em que z_1 e z_2 são exógenos. Temos duas outras variáveis exógenas, z_3 e z_4, que não aparecem em (15.49). Se y_2 for não correlacionado com u_1, devemos estimar (15.49) por MQO. Como podemos testar isso? Hausman (1978) sugeriu fazer uma comparação direta das estimativas MQO e MQ2E e determinar se as diferenças são estatisticamente significantes. Afinal de contas, tanto MQO como MQ2E serão consistentes se todas as variáveis forem exógenas. Se MQ2E e MQO diferirem de forma significante, concluímos que y_2 deve ser endógeno (supondo que os z_j são exógenos).

É uma boa ideia calcular MQO e MQ2E para verificar se as estimativas são diferentes na prática. Para determinar se as diferenças são estatisticamente significantes, é mais fácil usar um teste de regressão. Isso é feito com base na estimação da forma reduzida de y_2, que nesse caso é

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

Agora, como cada z_j é não correlacionado com u_1, y_2 será não correlacionado com u_1 se, e somente se, v_2 for não correlacionado com u_1; isso é o que queremos testar. Escreva $u_1 = \delta_1 v_2 + e_1$, em que e_1 é não correlacionado com v_2 e tem média zero. Então, u_1 e v_2 serão não correlacionados se, e somente se, $\delta_1 = 0$. A maneira mais fácil de verificar esse valor é incluir v_2 como um regressor adicional em (15.49) e fazer um teste t. Só existe um problema com a implementação desse procedimento: v_2 não é observado, porque ele é o termo de erro em (15.50). Como podemos estimar a forma reduzida de y_2 por MQO, podemos obter os resíduos da forma reduzida, $\hat{v}_2$. Portanto, estimamos

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 \hat{v}_2 + erro \qquad \textbf{(15.51)}$$

por MQO e testamos H_0: $\delta_1 = 0$, usando uma estatística t. Se rejeitarmos H_0 a um nível pequeno de significância, concluímos que y_2 é endógeno porque v_2 e u_1 são correlacionados.

O TESTE DE ENDOGENEIDADE DE UMA ÚNICA VARIÁVEL EXPLICATIVA:

(i) Estime a forma reduzida de y_2, regredindo y_2 sobre *todas* as variáveis exógenas (inclusive aquelas da equação estrutural e as VIs adicionais). Obtenha os resíduos, $\hat{v}_2$.

(ii) Adicione $\hat{v}_2$ à equação estrutural (que inclui y_2) e verifique a significância de $\hat{v}_2$, usando uma regressão MQO. Se o coeficiente de $\hat{v}_2$ for estatisticamente diferente de zero, concluiremos que y_2 é endógeno. Podemos usar um teste t robusto em relação à heteroscedasticidade.

EXEMPLO 15.7

(Retorno da Educação para Mulheres que Trabalham)

Podemos testar a endogeneidade de *educ* em (15.40), obtendo os resíduos $\hat{v}_2$ da estimação da forma reduzida (15.41) — usando somente mulheres que trabalham — e incluindo-os em (15.40). Quando fazemos isso, o coeficiente de $\hat{v}_2$ é $\hat{\delta}_1 = 0{,}058$, e $t = 1{,}67$. Essa é uma evidência moderada de correlação positiva entre u_1 e v_2. Provavelmente é uma boa ideia descrever ambas as estimativas porque a estimativa MQ2E do retorno da educação (6,1%) está bem abaixo da estimativa MQO (10,8%).

Uma característica interessante da regressão da parte (ii) do teste de endogeneidade é que as estimativas de todas as variáveis (exceto $\hat{v}_2$) são idênticas às estimativas MQ2E. Por exemplo, estimando a (15.51) pelos MQO produzirá as mesmas $\hat{\beta}_j$ que estimando a (15.49) pelos MQ2E. Um benefício dessa equivalência é que ela propicia uma fácil verificação de se você fez a regressão apropriada no teste de endogeneidade. Mas ela também fornece uma interpretação dos MQ2E diferente e proveitosa: adicionando v_2 à equação original como uma variável explicativa, e aplicando os OLS resolve a endogeneidade da y_2. Assim, quando começamos por estimar a (15.49) pelos OLS, podemos quantificar a importância de se permitir que a y_2 seja endógena, pela observação de quanto a $\hat{\beta}_1$ altera quando $\hat{v}_2$ é adicionada à equação. Independentemente do resultado dos testes estatísticos, podemos verificar se a alteração na $\hat{\beta}_1$ é esperada e é praticamente significante.

Também podemos testar a endogeneidade de múltiplas variáveis explicativas. Para cada variável suspeita de ser endógena, obtemos os resíduos da forma reduzida, como na parte (i). Depois, verificamos a significância conjunta desses resíduos na equação estrutural, usando um teste F. A significância conjunta indica que pelo menos uma variável explicativa suspeita é endógena. O número de restrições de exclusão testadas é o número de variáveis explicativas suspeitas de serem endógenas.

O Teste de Restrições Sobreidentificadoras

Quando apresentamos o estimador simples de variáveis instrumentais na Seção 15.1, enfatizamos que o instrumento deve satisfazer duas condições: ele deve ser não correlacionado com o erro (exogeneidade) e correlacionado com a variável explicativa endógena (relevância). Vimos agora que, mesmo em modelos com variáveis explicativas adicionais, a segunda condição pode ser testada usando um teste t (com só um instrumento) ou um teste F (quando existem múltiplos instrumentos). No contexto do estimador simples de VI, notamos que a exigência de exogeneidade não pode ser testada. Porém, se tivermos mais instrumentos do que necessitamos, poderemos efetivamente testar se algumas delas são não correlacionadas com o erro estrutural.

Como um exemplo específico, novamente considere a equação (15.49) com duas variáveis instrumentais para y_2, z_3 e z_4. Lembre-se, z_1 e z_2 agem basicamente como seus próprios instrumentos. Como temos dois instrumentos para y_2, podemos estimar (15.49) usando, digamos, somente a z_3 como uma VI da y_2; que $\check{\beta}_1$ designe o estimador de VI resultante de β_1. Então poderemos estimar a (15.49) usando somente a z_4 como uma VI da y_2; chame este estimador VI de $\tilde{\beta}_1$. Se todas as z_j forem exógenas, e se z_3 e z_4 forem cada uma delas parcialmente correlacionadas com y_2, então $\check{\beta}_1$ e $\tilde{\beta}_1$ serão ambas consistentes com β_1. Portanto, se nossa lógica na escolha de instrumentos for correta, $\check{\beta}_1$ e $\tilde{\beta}_1$ deverão diferir somente no erro de amostragem. Hausman (1978) propôs basear um teste de se z_3 e z_4 forem ambas exógenas, na diferença, $\check{\beta}_1 - \tilde{\beta}_1$. Em breve, forneceremos uma maneira mais simples de se obter um teste válido, mas, antes de fazermos isso, devemos entender como interpretar o resultado do teste.

Se concluirmos que $\check{\beta}_1$ e $\tilde{\beta}_1$ são estatisticamente diferentes um do outro, então não temos opção senão concluirmos que ou a z_3 ou a z_4, ou ambas falharam quanto ao requisito de exogeneidade. Infelizmente, não temos como saber qual foi a causa (a menos que declaremos desde o início que, digamos, z_3 é exógena). Por exemplo, se y_2 denotar anos de educação formal numa equação log de salário, z_3 for a escolaridade da mãe, e z_4 for a escolaridade do pai, uma diferença estatisticamente significante nos dois estimadores de VI indicará que a variável educacional de um ou ambos os pais é correlacionada com u_1 (15.54).

Certamente a rejeição de instrumentos como sendo exógenos é coisa séria e exige uma nova abordagem. Mas o problema mais sério, e sutil, na comparação de estimativas VI, é que elas podem ser semelhantes embora ambos os instrumentos falhem quanto ao requisito de exogeneidade. No exemplo anterior, parece ser provável que, se a escolaridade da mãe for positivamente correlacionada com a u_1, então também será a escolaridade do pai. Portanto, as duas estimativas VI podem ser semelhantes embora cada uma delas seja inconsistente. Na realidade, como VIs neste exemplo são escolhidas com o uso de raciocínio semelhante, o uso separado delas em procedimentos VI pode muito bem levar a estimativas similares que serão mesmo assim, tanto uma como a outra, inconsistentes. A questão é que não devemos nos sentir particularmente confortáveis se nossos procedimentos VI passarem no teste de Hausman.

Outro problema com a comparação de duas estimativas VI é que muitas vezes elas podem parecer de fato diferentes embora, estatisticamente, não possamos rejeitar a hipótese nula de que elas são consistentes no mesmo parâmetro populacional. Por exemplo, ao estimarmos (15.40) pelas VI usando *educm* como o único instrumento, o coeficiente na *educ* será 0,049 (0,037). Se usarmos somente *educp* como VI de *educ*, o coeficiente na *educ* será 0,070 (0,034). [Possivelmente sem causar espanto, a estimativa usando o nível de escolaridade de ambos os pais como VIs estará entre essas duas, 0,061 (0,031).] Taticamente, a diferença entre 5% e 7% no retorno de um ano de educação formal estimado é substancial. Porém, como mostrado no Exemplo 15.8, a diferença não é estatisticamente significante.

O procedimento de comparar diferentes estimativas de VI do mesmo parâmetro é um exemplo para testar as **restrições sobreidentificadoras**. A ideia geral é que temos mais instrumentos do que precisamos para estimar consistentemente os parâmetros. No exemplo anterior, tínhamos um instrumento a mais do que precisávamos, e isto resultou em uma restrição sobreidentificadora que pode ser testada. No caso geral, suponha que temos q instrumentos a mais do que necessitamos. Por exemplo, com uma variável explicativa endógena, y_2, e três instrumentos propostos para y_2, teremos $q = 3 - 1 = 2$ restrições sobreidentificadoras. Quando q é dois ou mais, comparar várias estimativas VI é complicado. Em vez disso, podemos facilmente calcular um teste estatístico baseado nos resíduos dos MQ2E. A ideia é que, se todos os instrumentos forem exógenos, os resíduos dos MQ2E deverão ser não correlacionados com os instrumentos, até o erro amostral. Mas se houver $k + 1$ parâmetros e $k + 1 + q$ instrumentos, os resíduos dos MQ2E terão uma média zero e serão identicamente não correlacionados com k combinações lineares dos instrumentos. (Este fato algébrico contém, como causa especial, o fato de que os resíduos dos MQO têm média zero e são não correlacionados com k variáveis explicativas.) Portanto,

o teste verifica se os resíduos dos MQ2E são correlacionados com q funções lineares dos instrumentos, e não precisamos decidir sobre qual função; o teste faz isso automaticamente.

O teste baseado em regressão a seguir será válido quando a hipótese de homoscedasticidade, listada como Hipótese MQ2E.5 no Apêndice deste capítulo, for válida.

O Teste de Restrições Sobreidentificadoras:

(i) Estime a equação estrutural por MQ2E e obtenha os resíduos MQ2E, $\hat{u}_1$.

(ii) Regrida $\hat{u}_1$ sobre *todas as variáveis exógenas*. Obtenha o R-quadrado, digamos R_1^2.

(iii) Sob a hipótese nula de que todas as VIs são não correlacionadas com u_1, $nR_1^2 \overset{a}{\sim} \chi_q^2$, em que q é o número de variáveis instrumentais fora do modelo menos o número total de variáveis explicativas endógenas. Se nR_1^2 exceder (digamos) o valor crítico de 5% na distribuição χ_q^2, rejeitamos H_0 e concluímos que pelo menos algumas das VIs não são exógenas.

EXEMPLO 15.8

(Retorno da Educação para Mulheres que Trabalham)

Quando usamos *educm* e *educp* como VIs de *educ* em (15.40), temos uma única restrição sobreidentificadora. A regressão dos resíduos do MQ2E $\hat{u}_1$ sobre *exper*, *exper*^{2}, *educm* e *educp* produz $R_1^2 = 0{,}0009$. Portanto, $nR_1^2 = 428(0{,}0009) = 0{,}3852$, que é um valor muito pequeno em uma distribuição χ_1^2 (p-valor = = 0,535). Portanto, as variáveis da educação dos pais passam no teste de sobreidentificação. Quando adicionamos a educação do marido à lista de VI, obtemos duas restrições sobreidentificadoras, e $nR_1^2 = 1{,}11$ (p-valor = 0,574). Sujeito às precauções anteriores, parece razoável adicionarmos *educmar* à lista de VI, pois isso reduz o erro-padrão da estimativa MQ2E: a estimativa MQ2E de *educ* usando as três variáveis instrumentais é 0,080 ($ep = 0{,}022$), de modo que isso torna *educ* muito mais significante do que quando *educmar* não é usada como uma VI ($\hat{\beta}_{educ} = 0{,}061$, $ep = 0{,}031$).

Quando $q = 1$, uma pergunta natural é: Como o teste obtido do procedimento baseado em regressão se compara com um teste-base sobre comparar diretamente as estimativas? Na realidade, os dois procedimentos são assimptoticamente o mesmo. Como um tópico prático, faz sentido calcular as duas estimativas VI para verificar como elas diferem entre si. De forma mais geral, quando $q \geq 2$, pode-se comparar as estimativas MQ2E com todas as VIs da estimativa VI usando instrumentos únicos. Fazendo isso, pode-se verificar se as várias estimativas VI são, na prática, diferentes, independentemente de o teste de sobreidentificação rejeitar ou falhar em rejeitar.

No exemplo anterior, aludimos a um fato geral sobre o MQ2E: sob as hipóteses-padrão do MQ2E, a adição de variáveis instrumentais à lista melhora a eficiência assimptótica do MQ2E. Entretanto, isso requer que quaisquer novas variáveis instrumentais sejam de fato exógenas — caso contrário, o MQ2E não será sequer consistente —, e isso será apenas um resultado assimptótico. Dados os tamanhos típicos das amostras disponíveis, a adição de variáveis instrumentais em demasia — isto é, o aumento do número de restrições sobreidentificadoras — pode causar vieses severos no MQ2E. Uma discussão detalhada nos desviaria muito do assunto. Uma boa ilustração é dada por Bound, Jaeger & Baker (1995), que argumentam que as estimativas MQ2E do retorno da educação obtidas por Angrist e Krueger (1991), usando muitas variáveis instrumentais, são propensas a ser seriamente viesadas (mesmo com centenas de milhares de observações!).

O teste de sobreidentificação pode ser usado sempre que tivermos mais variáveis instrumentais do que necessitamos. Se tivermos quantidade exata de variáveis instrumentais, o modelo é considerado *exatamente identificado*, e o R-quadrado na parte (ii) será zero, identicamente. Como mencionamos anteriormente, não podemos testar a exogeneidade dos instrumentos no caso do modelo exatamente identificado.

O teste pode se tornar robusto quanto à heteroscedasticidade de forma arbitrária; para detalhes, veja Wooldridge (2002, Capítulo 5).

15.6 O MQ2E COM HETEROSCEDASTICIDADE

A heteroscedasticidade no contexto do MQ2E suscita essencialmente os mesmos problemas do MQO. O mais importante é a possibilidade de obter erros-padrão e estatísticas de testes que são (assimptoticamente) robustos quanto à heteroscedasticidade de forma arbitrária e desconhecida. Na verdade, a equação (8.4) do Capítulo 8 continua a ser válida se os $\hat{r}_{ij}$ forem obtidos como os resíduos da regressão de $\hat{x}_{ij}$ sobre os outros $\hat{x}_{il}$, onde o símbolo "^" representa valores estimados das regressões do primeiro estágio (das variáveis explicativas endógenas). Wooldridge (2002, Capítulo 5) contém mais detalhes. Alguns programas econométricos fazem isso rotineiramente.

Também podemos testar a heteroscedasticidade, usando um teste análogo ao de Breusch-Pagan que apresentamos no Capítulo 8. Sejam $\hat{u}$ os resíduos MQ2E e $z_1, z_2, \ldots, z_m$ todas as variáveis exógenas (inclusive as usadas como VIs das variáveis explicativas endógenas). Assim, sob hipóteses razoáveis [explicadas, por exemplo, em Wooldridge (2002, Capítulo 5)], uma estatística assimptoticamente válida será a habitual estatística F da significância conjunta em uma regressão de $\hat{u}^2$ sobre $z_1, z_2, \ldots, z_m$. A hipótese nula de homoscedasticidade será rejeitada se os z_j forem conjuntamente significantes.

Se aplicarmos esse teste no Exemplo 15.8, usando *educm*, *educp* e *educmar* como instrumentos de *educ*, obteremos $F_{5,422} = 2{,}53$, e p-valor = 0,029. Isso é evidência de heteroscedasticidade no nível de 5%. Podemos calcular erros-padrão robustos em relação à heteroscedasticidade para explicar isso.

Se soubermos como a variância do erro depende das variáveis exógenas, poderemos usar um procedimento de MQ2E ponderado, essencialmente o mesmo da Seção 8.4. Após estimar um modelo para $\text{Var}(u|z_1, z_2, \ldots, z_m)$, dividimos a variável dependente, as variáveis explicativas e todas as variáveis instrumentais da observação i por $\sqrt{\hat{h}_i}$ em que $\hat{h}_i$ representa a variância estimada. (A constante, que é tanto uma variável explicativa como uma VI, é dividida por $\sqrt{\hat{h}_i}$; veja a Seção 8.4.) Em seguida, aplicamos o MQ2E na equação transformada usando as variáveis instrumentais transformadas.

15.7 A APLICAÇÃO DO MQ2E A EQUAÇÕES DE SÉRIES TEMPORAIS

Quando aplicamos o MQ2E a dados de séries temporais, muitas das considerações que surgiram sobre o MQO nos Capítulos 10, 11 e 12 são pertinentes. Escreva a equação estrutural de cada período de tempo como

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + b_k x_{tk} + u_t, \qquad \textbf{(15.52)}$$

em que uma ou mais das variáveis explicativas x_{tj} possam ser correlacionadas com u_t. Seja o conjunto de variáveis exógenas representado por $z_{t1}, \ldots, z_{tm}$:

$$\text{E}(u_t) = 0, \text{Cov}(z_{tj}, u_t) = 0, \; j = 1, \ldots, m.$$

Qualquer variável explicativa exógena também é uma z_{tj}. Para a identificação, é necessário que $m \geq k$ (temos tantas variáveis exógenas quanto variáveis explicativas).

A mecânica do MQ2E é idêntica para dados de séries temporais ou de corte transversal, mas para dados de séries temporais as propriedades estatísticas do MQ2E dependem das propriedades de tendência e de correlação das sequências básicas. Em particular, devemos ser cuidadosos ao decidir incluir uma variável de tendência se tivermos a variável dependente ou variáveis explicativas com tendência. Como uma tendência temporal é exógena, ela pode sempre servir como sua própria variável

instrumental. O mesmo é verdade em relação a variáveis *dummy* sazonais, se forem usados dados mensais ou trimestrais.

QUESTÃO 15.4

Um modelo para testar o efeito do crescimento dos gastos governamentais sobre o crescimento da produção é

$$cPIB_t = \beta_0 + \beta_1 cGOV_t + \beta_2 RAZINV_t + \beta_3 cTRAB_t + u_t,$$

em que *c* indica crescimento, *PIB* é o produto interno bruto real, *GOV* é o gasto governamental real, *RAZINV* é a razão do investimento interno sobre o PIB e *TRAB* é o tamanho da força de trabalho. [Veja equação (6) em Ram (1986)] Sob quais hipóteses uma variável *dummy* indicando se o presidente no ano $t - 1$ era um Republicano seria uma VI adequada de c*GOV*?

Séries que possuem forte persistência (têm raízes unitárias) devem ser usadas com cuidado, assim como no MQO. Frequentemente, diferenciar a equação é necessário antes da estimação, e isso se aplica também às variáveis instrumentais.

Sob hipóteses análogas às do Capítulo 11 para as propriedades assimptóticas do MQO, o MQ2E usando dados de séries temporais é consistente e distribuído normal e assimptoticamente. Na verdade, se substituirmos as variáveis explicativas pelas variáveis instrumentais ao estabelecer as hipóteses, somente precisaremos adicionar as hipóteses de identificação do MQ2E. Por exemplo, a hipótese de homoscedasticidade é definida como

$$E(u_t^2|z_{t1}, \ldots, z_{tm}) = \sigma^2, \qquad \textbf{(15.53)}$$

e a hipótese de ausência de correlação serial é estabelecida como

$$E(u_t u_s|\mathbf{z}_t, \mathbf{z}_s) = 0, \text{ para todo } t \neq s, \qquad \textbf{(15.54)}$$

em que $\mathbf{z}_t$ representa todas as variáveis exógenas no tempo *t*. Uma lista completa das hipóteses é apresentada no Apêndice deste capítulo. Forneceremos exemplos do MQ2E para problemas de séries temporais no Capítulo 16; veja também o Exercício em Computador 15.4, no site da Cengage.

Como no caso do MQO, a hipótese de ausência de correlação serial pode com frequência ser violada com dados de séries temporais. Felizmente, é bastante fácil testar a existência de correlação serial AR(1). Se escrevermos $u_t = \rho u_{t-1} + e_t$ e inserirmos essa expressão na equação (15.52), obteremos

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + \rho u_{t-1} + e_t, t \geq 2. \qquad \textbf{(15.55)}$$

Para testar H_0: $\rho_1 = 0$, devemos substituir u_{t-1} pelos resíduos do MQ2E, $\hat{u}_{t-1}$. Além disso, se os x_{tj} forem endógenos em (15.52), eles serão endógenos em (15.55), de modo que ainda necessitaremos usar uma VI. Como e_t é não correlacionado com todos os valores passados de u_t, $\hat{u}_{t-1}$ pode ser usado como seu próprio instrumento.

O Teste da Correlação Serial AR(1) Após MQ2E:

(i) Estime (15.52) por MQ2E e obtenha os resíduos do MQ2E, $\hat{u}_t$.
(ii) Estime

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + \rho \hat{u}_{t-1} + erro_t, t = 2, \ldots, n$$

por MQ2E, usando os mesmos instrumentos da parte (i), em adição a $\hat{u}_{t-1}$. Use a estatística *t* de $\hat{\rho}$ para testar H_0: $\rho = 0$.

Assim como na versão MQO desse teste no Capítulo 12, a estatística *t* somente tem justificação assimptótica, mas na prática ela tende a funcionar bem. Uma versão do teste robusto em relação à heteroscedasticidade pode ser usada para proteção contra a heteroscedasticidade. Além disso, resíduos defasados poderão ser adicionados à equação para testar a existência de formas mais altas de correlação serial, usando um teste *F* conjunto.

O que acontece se detectarmos correlação serial? Alguns programas econométricos calculam erros-padrão robustos em relação a formas razoavelmente gerais de correlação serial e heteroscedasticidade. Esse é um bom e simples caminho a seguir se seu programa de econometria fizer isso. Os cálculos são muito semelhantes aos da Seção 12.5 do MQO. [Veja Wooldridge (1995) para fórmulas e outros métodos computacionais.]

Uma alternativa é usar o modelo AR(1) e corrigir a correlação serial. O procedimento é semelhante ao do MQO e coloca restrições adicionais sobre as variáveis instrumentais. A equação quase diferenciada é a mesma que a equação (12.32):

$$\tilde{y}_t = \beta_0(1 - \rho) + \beta_1 \tilde{x}_{t1} + \ldots + \beta_k \tilde{x}_{tk} + e_t, t \geq 2, \qquad \textbf{(15.56)}$$

em que $\tilde{x}_{tj} = x_{tj} - \rho x_{t-1,j}$. (Podemos usar a observação $t = 1$ como fizemos na Seção 12.3, mas, para facilitar, omitimos isso aqui.) A questão é: o que podemos usar como variáveis instrumentais? Parece natural usar instrumentos quase diferenciadas, $\tilde{z}_{tj} = z_{tj} - \rho z_{t-1,j}$. Isso somente funcionará, porém, se em (15.52) o erro u_1 original for não correlacionado com as instrumentos nos tempos *t*, $t - 1$ e $t + 1$. Isto é, as variáveis instrumentais devem ser estritamente exógenas em (15.52). Isso exclui variáveis dependentes defasadas como VIs, por exemplo. Também elimina casos em que movimentos futuros nas VIs reagirem a alterações correntes e passadas no erro, u_t.

MQ2E com Erros AR(1):

(i) Estime (15.52) por MQ2E e obtenha os resíduos MQ2E, $\hat{u}_t$, $t = 1, 2, \ldots, n$.
(ii) Obtenha $\hat{\rho}$ da regressão de $\hat{u}_t$ sobre $\hat{u}_{t-1}$, $t = 2, \ldots, n$ e construa as variáveis quase diferenciadas $\tilde{y}_t = y_t - \hat{\rho} y_{t-1}$, $\tilde{x}_{tj} = x_{tj} - \hat{\rho} x_{t-1,j}$ e $\tilde{z}_{tj} = z_{tj} - \hat{\rho} z_{t-1,j}$ para $t \geq 2$. (Lembre-se de que, na maioria dos casos, algumas das VIs também serão variáveis explicativas.)
(iii) Estime (15.56) (onde ρ é substituído por $\hat{\rho}$) por MQ2E, usando as $\tilde{z}_{tj}$ como variáveis instrumentais. Ao supor que (15.56) satisfaz as hipóteses do MQ2E do Apêndice deste capítulo, as estatísticas de testes habituais do MQ2E serão assimptoticamente válidas.

Podemos também usar o primeiro período de tempo como na estimação de Prais-Winsten do modelo com variáveis explicativas exógenas. As variáveis transformadas no primeiro período de tempo — a variável dependente, as variáveis explicativas e as variáveis instrumentais — são obtidas simplesmente pela multiplicação de todos os valores do primeiro período por $(1 - \hat{\rho})^{1/2}$. (Veja também a Seção 12.3 do Capítulo 12.)

15.8 A APLICAÇÃO DO MQ2E EM CORTES TRANSVERSAIS AGRUPADOS E EM DADOS EM PAINEL

A aplicação dos métodos de variáveis instrumentais em cortes transversais independentemente agrupados não apresenta novas dificuldades. Tal como acontece com os modelos estimados por MQO, deveremos frequentemente incluir variáveis *dummy* temporais para levar em conta os efeitos temporais agregados. Essas variáveis *dummy* são exógenas —, pois a passagem do tempo é exógena — e assim elas agem como suas próprias variáveis instrumentais.

EXEMPLO 15.9

(Efeito da Educação sobre a Fertilidade)

No Exemplo 13.1, usamos o corte transversal agrupado contido no arquivo FERTIL1.RAW para estimar o efeito de educação sobre a fertilidade das mulheres, controlando vários outros fatores. Como em Sander (1992), consideramos a possibilidade de que *educ* seja endógeno na equação. Como variáveis instrumentais de *educ*, usamos os níveis de educação da mãe e do pai (*educm*, *educp*). A estimativa de MQ2E de β_{educ} é $-0,153$ (ep $= 0,039$), comparada com a estimativa do MQO de $-0,128$ (ep $= 0,018$). A estimativa por MQ2E mostra um efeito de certa forma maior da educação sobre a fertilidade, mas o erro-padrão do MQ2E é mais de duas vezes maior que o do MQO. (Na verdade, o intervalo de confiança de 95% baseado no MQ2E facilmente contém a estimativa MQO.) As estimativas MQO e MQ2E de β_{educ} *não são estatisticamente* diferentes, o que pode ser visto testando a endogeneidade de *educ*, como na Seção 15.5: quando a forma reduzida residual, $\hat{v}_2$, é incluída com os outros regressores na Tabela 13.1 (inclusive *educ*), sua estatística t é 0,702, não significante em qualquer nível razoável. Portanto, nesse caso, concluímos que a diferença entre o MQ2E e o MQO é em razão do erro de amostragem.

A estimação de variáveis instrumentais pode ser combinada com métodos de dados em painel, particularmente a primeira diferença, para estimar consistentemente parâmetros na presença de efeitos não observados e de endogeneidade em uma ou mais variáveis explicativas com variação temporal. O exemplo a seguir ilustra essa combinação de métodos.

EXEMPLO 15.10

(Treinamento de Pessoal e Produtividade de Trabalhadores)

Suponha que queremos estimar o efeito de uma hora adicional de treinamento sobre a produtividade dos trabalhadores. Para os anos de 1987 e 1988, considere o modelo simples de dados em painel

$$\log(ref_{it}) = \beta_0 + \delta_0 d88_t + \beta_1 hrsemp_{it} + a_i + u_{it}, \; t = 1, 2,$$

em que ref_{it} é a taxa de refugo dos produtos da firma i no ano t, e $hrsemp_{it}$ representa horas de treinamento por empregado. Como sempre, permitimos interceptos diferentes para os anos e um efeito constante não observado da firma, a_i.

Pelas razões discutidas na Seção 13.2, podemos estar preocupados com a possibilidade de que $hrsemp_{it}$ seja correlacionado com a_i, que contém a aptidão não medida do trabalhador. Como antes, fazemos a diferenciação para remover a_i:

$$\Delta\log(ref_i) = \delta_0 + \beta_1 \Delta hrsemp_i + \Delta u_i. \qquad (15.57)$$

EXEMPLO 15.10 (continuação)

Normalmente, estimaríamos essa equação por MQO. Entretanto, e se Δu_i for correlacionado com $\Delta hrsemp_i$? Por exemplo, uma empresa pode empregar trabalhadores mais habilidosos e, simultaneamente, reduzir o nível de treinamento. Nesse caso, necessitamos de uma variável instrumental de $\Delta hrsemp_i$. Geralmente, seria difícil encontrar tal VI, mas podemos explorar o fato de que algumas empresas receberam subsídio de treinamento de pessoal em 1988. Se presumirmos que a destinação de subsídios é não correlacionada com Δu_i — possibilidade razoável, pois os subsídios foram concedidos no início de 1988 —, $\Delta subs$ será uma VI válida, desde que $\Delta hrsemp$ e $\Delta subs$ sejam correlacionados. Utilizando os dados contidos no arquivo JTRAIN.RAW, diferenciados entre 1987 e 1988, a regressão do primeiro estágio será

$$\widehat{\Delta hrsemp} = 0,51 + 27,88\,\Delta subs$$
$$(1,56) \quad (3,13)$$
$$n = 45, \; R^2 = 0,392.$$

Isso confirma que a alteração nas horas de treinamento por trabalhador é forte e positivamente relacionada com o recebimento de um subsídio de treinamento de pessoal em 1988. Na verdade, o recebimento de um subsídio de treinamento de pessoal aumenta o treinamento por empregado em quase 28 horas, e a destinação do subsídio foi responsável por quase 40% da variação em $\Delta hrsemp$. A estimação por mínimos quadrados de dois estágios de (15.57) produz

$$\widehat{\Delta\log(ref)} = -0,033 - 0,014\,\Delta hrsemp$$
$$(0,127) \quad (0,008)$$
$$n = 45, \; R^2 = 0,016.$$

E significa que dez horas a mais de treinamento por trabalhador reduziria a taxa de refugo em cerca de 14%. Nas empresas da amostra, a média de horas de treinamento em 1988 foi cerca de 17 horas por trabalhador, com um mínimo de zero e um máximo de 88.

A título de comparação, a estimação por MQO de (15.57) produz $\hat{\beta}_1 = -0,0076$ (ep $= 0,0045$), de modo que a estimativa de β_1 por MQ2E é quase duas vezes maior em magnitude e é levemente mais significante, estatisticamente.

Quando $T \geq 3$, a equação diferenciada pode conter correlação serial. O mesmo teste e a mesma correção da correlação serial AR(1) da Seção 15.7 podem ser usados, onde todas as regressões serão agrupadas ao longo de i e também de t. Como não queremos perder um período de tempo completo, a transformação de Prais-Winsten deverá ser usada para o período de tempo inicial.

Modelos de efeitos não observados que contenham variáveis dependentes defasadas também exigem métodos de VI para uma estimação consistente. A razão é que, após fazermos a diferenciação, $\Delta y_{i,t-1}$ será correlacionado com Δu_{it}, pois $y_{i,t-1}$ e $u_{i,t-1}$ são correlacionados. Podemos usar duas ou mais defasagens de y como VIs de $\Delta y_{i,t-1}$. [Veja Wooldridge (2002, Capítulo 11) para detalhes.]

As variáveis instrumentais após a diferenciação também podem ser usadas em amostras pareadas. Ashenfelter e Krueger (1994) diferenciaram a equação de salários-hora entre gêmeos para eliminar a aptidão não observada:

$$\log(salário\text{h}_2) - \log(salário\text{h}_1) = \delta_0 + \beta_1(educ_{2,2} - educ_{1,1}) + (u_2 - u_1),$$

em que $educ_{1,1}$ são os anos de escolaridade do primeiro gêmeo como por ele relatados, e $educ_{2,2}$ são os anos de escolaridade do segundo gêmeo, relatado por ele próprio. Para considerar a possibilidade de erro de medida nas indicações autoinformadas de escolaridade, Ashenfelter e Krueger usaram ($educ_{2,1}$ – $educ_{1,2}$) como uma VI de ($educ_{2,2}$ – $educ_{1,1}$), em que $educ_{2,1}$ representa os anos de escolaridade do segundo gêmeo como relatado pelo primeiro e $educ_{1,2}$ representa os anos de escolaridade do primeiro gêmeo como relatado pelo segundo gêmeo. A estimativa de VI de β_1 foi 0,167 (t = 3,88), comparada com a estimativa de MQO sobre as primeiras diferenças de 0,092 (t = 3,83) [veja Ashenfelter e Krueger (1994, Tabela 3)].

RESUMO

No Capítulo 15, apresentamos o método de variáveis instrumentais como uma maneira de estimar consistentemente os parâmetros em um modelo linear quando uma ou mais variáveis explicativas são endógenas. Uma variável instrumental deve ter duas propriedades: (1) ela deve ser exógena, isto é, não correlacionada com o termo de erro da equação estrutural; (2) ela deve ser parcialmente correlacionada com a variável explicativa endógena. Encontrar uma variável com essas duas propriedades é, normalmente, desafiador.

O método dos mínimos quadrados em dois estágios, que possibilita o uso de um maior número de variáveis instrumentais do que o de variáveis explicativas que temos, é usado rotineiramente em ciências sociais empíricas. Quando usado adequadamente, ele pode nos permitir estimar efeitos *ceteris paribus* na presença de variáveis explicativas endógenas. Isso é verdadeiro em aplicações de corte transversal, séries temporais e dados em painel. Mas quando as variáveis instrumentais são fracas — o que significa que elas são correlacionadas com o termo de erro, ou somente fracamente correlacionadas com a variável explicativa endógena, ou as duas coisas ao mesmo tempo —, então o MQ2E pode ser pior que o MQO.

Quando temos variáveis instrumentais válidas, podemos testar se uma variável explicativa é endógena, usando o teste da Seção 15.5. Além disso, embora nunca possamos verificar se todas as VIs são exógenas, podemos verificar se pelo menos algumas delas são — presumindo que temos mais variáveis instrumentais do que necessitamos para uma estimação consistente (isto é, o modelo é sobre-especificado). A heteroscedasticidade e a correlação serial podem ser testadas e tratadas usando métodos semelhantes ao caso de modelos com variáveis explicativas exógenas.

Neste capítulo, usamos variáveis omitidas e erro de medida para ilustrar o método das variáveis instrumentais. Métodos de VI também são indispensáveis nos modelos de equações simultâneas, os quais veremos no Capítulo 16.

PROBLEMAS

15.1 Considere um modelo simples para estimar o efeito da propriedade de um computador pessoal (PC) na nota média de graduação de formandos de uma grande universidade pública:

$$supGPA = \beta_0 + \beta_1 PC + u,$$

em que PC é uma variável binária que indica a propriedade de um PC.

(i) Por que a propriedade de um PC pode estar correlacionada com u?
(ii) Explique por que PC possivelmente está relacionado à renda anual dos pais. Isso significa que a renda dos pais será uma boa VI de PC? Por quê?
(iii) Suponha que, quatro anos atrás, a universidade tenha concedido subsídios para a compra de computadores a aproximadamente metade dos alunos novos, e que os alunos que receberam esses subsídios tenham sido escolhidos aleatoriamente. Explique cuidadosamente como você usaria essa informação para construir uma variável instrumental de PC.

15.2 Suponha que você queira estimar o efeito da frequência às aulas sobre o desempenho dos alunos, como no Exemplo 6.3. Um modelo básico é

$$respad = \beta_0 + \beta_1 taxafreq + \beta_2 prsGPA + \beta_3 ACT + u,$$

em que as variáveis foram definidas no Capítulo 6.

(i) Defina *dist* como a distância da residência do aluno até o local de estudos. Você considera que *dist* é não correlacionado com u?
(ii) Supondo que *dist* e u sejam não correlacionados, que outra hipótese *dist* terá que satisfazer para ser uma VI válida de *taxafreq*?
(iii) Suponha, como na equação (6.18), que adicionemos o termo de interação *prsGPA*·*taxafreq*:

$$respad = \beta_0 + \beta_1 taxafreq + \beta_2 prsGPA + \beta_3 ACT + \beta_4 prsGPA \cdot taxafreq + u.$$

Se *taxafreq* for correlacionado com u, então, em geral, $prsGPA \cdot taxafreq$ também será. O que poderia ser uma boa VI de $prsGPA \cdot taxafreq$? [*Sugestão*: Se E(u|*prsGPA, ACT, dist*) = 0, como acontece quando *prsGPA, ACT* e *dist* são todas variáveis exógenas, então, qualquer função de *prsGPA* e *dist* será não correlacionada com u.]

15.3 Considere o modelo de regressão simples

$$y = \beta_0 + \beta_1 x + u$$

e defina z como uma variável instrumental *binária* de x. Use (15.10) para mostrar que o estimador de VI $\hat{\beta}_1$ pode ser escrito como

$$\hat{\beta}_1 = (\bar{y}_1 - \bar{y}_0)/(\bar{x}_1 - \bar{x}_0),$$

em que $\bar{y}_0$ e $\bar{x}_0$ são as médias amostrais de y_i e x_i da parte da amostra com $z_i = 0$, e onde $\bar{y}_1$ e $\bar{x}_1$ são as médias amostrais de y_i e x_i da parte da amostra com $z_i = 1$. Esse estimador, conhecido como *estimador agrupado*, foi sugerido pela primeira vez por Wald (1940).

15.4 Suponha que você queira usar dados de séries temporais de determinado estado dos Estados Unidos, para estimar o efeito do salário mínimo em nível estadual sobre o emprego de pessoas entre 18 e 25 anos de idade (*EMP*). Um modelo simples é

$$cEMP_t = \beta_0 + \beta_1 cMIN_t + \beta_2 cPOP_t + \beta_3 cPEB_t + \beta_4 cPIB_t + u_t,$$

em que MIN_t é o salário mínimo, em dólares reais, POP_t é a população com idade entre 18 e 25 anos, PEB_t é o produto estadual bruto e PIB_t é o produto interno bruto norte-americano. O prefixo *c* indica a taxa de crescimento do ano $t - 1$ ao ano t, que em geral será aproximada pela diferença dos logs.

(i) Se estivermos preocupados que o estado escolha seu salário mínimo parcialmente baseado em fatores não observados (por nós) que afetem o emprego dos jovens, qual será o problema da estimação por MQO?
(ii) Defina $SMAm_t$ como o salário mínimo dos Estados Unidos, que também é indicado em termos de dólares reais. Você acha que $cSMAm_t$ é não correlacionado com u_t?
(iii) Por lei, qualquer salário mínimo estadual deve ser pelo menos igual ao salário mínimo nacional. Explique por que isso torna $cSMAm_t$ um candidato em potencial para ser uma VI de $cMIN_t$.

15.5 Retorne às equações (15.19) e (15.20). Suponha que $\sigma_u = \sigma_x$, de forma que a variação populacional no termo de erro seja a mesma contida em x. Suponha que a variável instrumental, z, seja levemente correlacionada com u: Corr(z,u) = 0,1. Suponha também que z e x tenham uma correlação um pouco maior: Corr(z,x) = 0,2.

(i) Qual será o viés assimptótico no estimador de VI?
(ii) Quanta correlação deverá existir entre x e u antes que o MQO tenha mais viés assimptótico que o MQ2E?

15.6 (i) No modelo com uma variável explicativa endógena, uma variável explicativa exógena e uma variável exógena extra, considere a forma reduzida de y_2, (15.26), inserindo-a na equação estrutural (15.22). Isso produzirá a forma reduzida de y_1:

$$y_1 = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + v_1.$$

Encontre os α_j em termos de β_j e π_j.

(ii) Encontre a forma reduzida do erro, v_1, em termos de u_1, v_2 e os parâmetros.
(iii) Como você estimaria consistentemente os α_j?

15.7 O que segue é um modelo simples para medir o efeito de um programa de escolha de escola sobre o desempenho em um teste padronizado [veja Rouse (1998), para entender a motivação do problema]:

$$nota = \beta_0 + \beta_1 escolha + \beta_2 rendfam + u_1,$$

em que *nota* é a nota em um teste de âmbito estadual, *escolha* é uma variável binária indicando se o aluno frequentou uma escola de sua escolha no último ano e *rendfam* é a renda familiar.
A VI de *escolha* é *conc*, o montante em dólares concedido aos alunos para ser usado como pagamento da anuidade da escola particular de sua escolha. O montante da concessão difere conforme o nível da renda familiar, razão pela qual controlamos *rendfam* na equação.

(i) Mesmo com *rendfam* na equação por que *escolha* pode ser correlacionada com u_1?

(ii) Se no interior de cada classe de rendimento os montantes de concessão fossem atribuídos aleatoriamente, *conc* seria não correlacionado com u_1?
(iii) Escreva a forma reduzida da equação de *escolha*. O que é necessário para *conc* ser parcialmente correlacionado com *escolha*?
(iv) Escreva a equação na forma reduzida de *nota*. Explique por que isso é importante. (*Sugestão*: Como você interpreta o coeficiente de *conc*?)

15.8 Suponha que você queira testar se meninas que frequentam uma escola de ensino médio só para meninas se saem melhor em matemática do que as que frequentam escolas mistas. Você tem uma amostra aleatória de meninas veteranas de escolas de ensino médio de um estado dos Estados Unidos, e *nota* é a nota de um teste padronizado de matemática. Defina *meninaem* como uma variável *dummy* indicando se uma aluna frequenta uma escola de ensino médio só para meninas.

(i) Que outros fatores você controlaria na equação? (Você deve ter condições razoáveis de coletar dados sobre esses fatores.)
(ii) Escreva uma equação que relaciona *nota* com *meninaem* e os outros fatores que você listou na parte (i).
(iii) Suponha que o suporte e o incentivo dos pais sejam fatores não indicados no termo de erro na parte (ii). É possível que eles sejam correlacionados com *meninaem*? Explique.
(iv) Discuta as hipóteses necessárias para que o número de escolas do ensino médio só para meninas situadas em um raio de 20 milhas (aproximadamente 32 km) da residência de uma menina seja uma VI válida de *meninaem*.

15.9 Suponha que na equação (15.8) você não tenha uma boa candidata a variável instrumental de *faltas*. Entretanto, você tem duas outras informações sobre os alunos: a nota média ponderada de matemática e habilidade verbal do estudante para ingresso em curso superior (*SAT*) e a nota média acumulada anterior ao semestre (*GPA*). O que você faria em vez da estimação de VI?

15.10 Em um artigo recente, Evans e Schwab (1995) estudaram os efeitos que frequentar uma escola católica do ensino médio teriam sobre a probabilidade de cursar uma faculdade. Concretamente, defina *faculdade* como uma variável binária igual a um se o aluno estiver na faculdade, e zero caso contrário. Defina *EMCat* como uma variável binária igual a um se o aluno frequenta uma escola católica do ensino médio. Um modelo de probabilidade linear é

$$faculdade = \beta_0 + \beta_1 EMCat + outros\ fatores + u,$$

em que, entre os outros fatores, estão sexo, raça, renda familiar e instrução dos pais.

(i) Por que *EMCat* pode ser correlacionado com u?
(ii) Evans e Schwab tinham dados sobre a nota de um teste padronizado feito quando cada estudante era aluno do 2º ano. O que pode ser feito com essa variável para melhorar a estimativa *ceteris paribus* de frequentar uma escola católica do ensino médio?
(iii) Defina *RelCat* como uma variável binária igual a um se o estudante for da religião católica. Detalhe os dois requisitos necessários para que essa seja uma VI válida de *EMCat* na equação precedente. Qual deles pode ser testado?
(iv) Não surpreendentemente, o fato de ser católico tem um efeito significante sobre frequentar uma escola católica do ensino médio. Você julga que *RelCat* é uma variável instrumental convincente de *EMCat*?

15.11 Considere um modelo simples de séries temporais no qual a variável explicativa tem erro clássico de medida:

$$y_t = \beta_0 + \beta_1 x_t^* + u_t$$
$$x_t = x_t^* + e_t, \qquad \textbf{(15.58)}$$

em que u_t tem média zero e é não correlacionado com x_t^* e e_t. Observamos somente y_t e x_t. Suponha que e_t tem média zero e é não correlacionado com x_t^* e que x_t^* também tem uma média zero (esta última hipótese é só para simplificar a álgebra).

(i) Escreva $x_t^* = x_t - e_t$ e insira essa expressão em (15.58). Mostre que o termo de erro na nova equação, digamos, v_t, é negativamente correlacionado com x_t se $\beta_1 > 0$. O que isso sugere sobre o estimador MQO de β_1 da regressão de y_t sobre x_t?

(ii) Além das hipóteses anteriores, presuma que u_t e e_t são não correlacionados com todos os valores passados de x_t^* e e_t; em particular com x_{t-1}^* e e_{t-1}. Mostre que $E(x_{t-1}v_t) = 0$, em que v_t é o termo de erro no modelo da parte (i).

(iii) É possível que x_t e x_{t-1} sejam correlacionados? Explique.

(iv) O que as partes (ii) e (iii) sugerem como uma estratégia vantajosa para estimarmos consistentemente β_0 e β_1?

APÊNDICE 15A

Hipóteses do Método de Mínimos Quadrados em Dois Estágios

Este apêndice abrange as hipóteses sob as quais o MQ2E tem propriedades desejáveis de amostra grande. Primeiro, declaramos as hipóteses para as aplicações de corte transversal sob amostragem aleatória. Depois, discutimos o que precisa ser adicionado para que elas se apliquem a séries temporais e dados em painel.

HIPÓTESE MQ2E.1 (LINEARIDADE NOS PARÂMETROS)
O modelo na população pode ser escrito como

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u,$$

em que $\beta_0, \beta_1, \ldots, \beta_k$ são os parâmetros desconhecidos (constantes) de interesse, e u é um erro aleatório não observável ou termo de perturbação aleatório. As variáveis instrumentais são representadas como z_j.

Vale a pena enfatizar que a Hipótese MQ2E.1 é praticamente idêntica à RLM.1 (com a pequena exceção que a MQ2E.1 menciona a notação das variáveis instrumentais, z_j).

Em outras palavras, o modelo pelo qual estamos interessados é o mesmo que o da estimação pelos MQO da β_j. Algumas vezes é fácil ignorarmos o fato de podermos aplicar métodos diferentes de *estimação* no *mesmo* modelo. Infelizmente, não é incomum ouvirmos pesquisadores dizendo "eu estimei um modelo MQO" ou "eu usei um modelo MQ2E". Tais declarações não têm sentido. MQO e MQ2E são métodos diferentes de *estimação* que são aplicados no *mesmo* modelo. É verdade que eles têm propriedades estatísticas desejáveis, sob diferentes conjuntos de hipóteses no modelo, mas o relacionamento que eles estão estimando é dado pela equação na MQ2E.1 (ou RLM.1). A questão

é semelhante à feita para o modelo de efeitos não observáveis em dados em painel discutidos nos Capítulos 13 e 14: MQO agrupados, primeira diferença, efeitos ajustados e efeitos aleatórios são métodos diferentes de estimação para o mesmo modelo.

HIPÓTESE MQ2E.2 (AMOSTRAGEM ALEATÓRIA)
Temos uma amostra aleatória de y, x_j e z_j.

HIPÓTESE MQ2E.3 (CONDIÇÃO DE CLASSIFICAÇÃO)
(i) Não há relações lineares perfeitas entre as variáveis instrumentais. (ii) A condição de classificação da identificação se mantém.

Com uma única variável explicativa endógena, como na equação (15.42), a condição de classificação é facilmente descrita. Sejam $z_1, \ldots, z_m$ as variáveis exógenas, em que $z_k, \ldots, z_m$ não aparecem no modelo estrutural (15.42). A forma reduzida de y_2 é

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \ldots + \pi_{k-1} z_{k-1} + \pi_k z_k + \ldots + \pi_m z_m + v_2.$$

Então, necessitamos de que pelo menos um dos $\pi_k, \ldots, \pi_m$ seja diferente de zero. Isso exige pelo menos uma variável exógena que não aparece em (15.42) (a condição de ordem). Declarar a condição de classificação com duas ou mais variáveis explicativas endógenas exige álgebra matricial. [Veja Wooldridge (2002, Capítulo 5).]

HIPÓTESE MQ2E.4 (VARIÁVEIS INSTRUMENTAIS EXÓGENAS)
O termo de erro u tem média zero, e cada VI é não correlacionada com u.

Lembre-se de que qualquer x_j que seja não correlacionado com u também age como uma VI.

TEOREMA 15A.1
Sob as hipóteses MQ2E.1 a MQ2E.4, o estimador MQ2E é consistente.

HIPÓTESE MQ2E.5 (HOMOSCEDASTICIDADE)
Seja $\mathbf{z}$ a coleção de todas as variáveis instrumentais. Então, $E(u^2|\mathbf{z}) = \sigma^2$.

TEOREMA 15A.2
Sob as Hipóteses MQ2E.1 a MQ2E.5, os estimadores MQ2E são assimptoticamente normalmente distribuídos. Estimadores consistentes da variância assimptótica são dados como na equação (15.43), em que σ^2 é substituída por $\hat{\sigma}^2 = (n-k-1)^2 \sum_{i=1}^{n} \hat{u}_i^2$ e os $\hat{u}_i$ são os resíduos MQ2E.

O estimador MQ2E também é o melhor estimador de VI sob as cinco hipóteses dadas. Definimos o resultado aqui. Uma prova pode ser encontrada em Wooldridge (2002, Capítulo 5).

TEOREMA 15A.3
Sob as Hipóteses MQ2E.1 a MQ2E.5, o estimador MQ2E é assimptoticamente eficiente na classe de estimadores de VI que usa combinações lineares das variáveis exógenas como variáveis instrumentais.

Se a hipótese de homoscedasticidade não se sustentar, os estimadores MQ2E ainda assim serão assimptoticamente normais, mas os erros-padrão (e as estatísticas t e F) precisarão ser ajustados; muitos programas econométricos fazem isso rotineiramente. Além disso, em geral, o estimador MQ2E não mais será o estimador de VI assimptoticamente eficiente. Não estudaremos aqui estimadores mais eficientes. [Veja Wooldridge (2002, Capítulo 8).]

Em aplicações de séries temporais, devemos adicionar algumas hipóteses. Primeiro, como no MQO, devemos presumir que todas as séries (inclusive as VIs) são fracamente dependentes: isso garante que a lei dos grandes números e o teorema do limite central sejam válidos. Para que os habituais erros-padrão e estatísticas de testes sejam válidos e também para a eficiência assimptótica, devemos adicionar uma hipótese de ausência de correlação serial.

HIPÓTESE MQ2E.6 (AUSÊNCIA DE CORRELAÇÃO SERIAL)
A equação (15.54) se mantém.

Uma hipótese semelhante de ausência de correlação serial é necessária em aplicações de dados em painel. Testes e correções de correlação serial foram examinados na Seção 15.7.