Introdução à Econometria

Uma Abordagem Moderna

Jeffrey M. Wooldridge
Michigan State University

Tradução da quarta edição norte-americana

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

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

16

Modelos de Equações Simultâneas

No capítulo anterior, mostramos como o método das variáveis instrumentais pode solucionar dois tipos de problemas de endogeneidade: variáveis omitidas e erro de medida. Conceitualmente, esses problemas são claros. No caso de variáveis omitidas, existe uma variável (ou mais de uma) que gostaríamos de manter fixa quando estimamos o efeito *ceteris paribus* de uma ou mais das variáveis explicativas observadas. No caso do erro de medida, gostaríamos de estimar o efeito de certas variáveis explicativas sobre y, mas medimos incorretamente uma ou mais variáveis. Em ambos os casos, poderíamos estimar os parâmetros de interesse por MQO se pudéssemos coletar dados melhores.

Outra forma importante de endogeneidade de variáveis explicativas é a **simultaneidade**. Ela surge quando uma ou mais das variáveis explicativas são *determinadas conjuntamente* com a variável dependente, em geral por meio de um mecanismo de equilíbrio (como veremos mais tarde). Neste capítulo, estudamos métodos de estimar modelos de equações simultâneas (SEM) simples. Embora um tratamento completo de SEM esteja além do escopo desta obra, temos condições de abordar modelos que são amplamente usados.

O principal método para estimar modelos de equações simultâneas é o das variáveis instrumentais. Portanto, a solução dos problemas de simultaneidade é basicamente a mesma que a solução de VIs para os problemas de variáveis omitidas e erro de medida. Porém, elaborar e interpretar SEM é um trabalho desafiador. Dessa forma, iniciamos examinando a natureza e o escopo de modelos de equações simultâneas na Seção 16.1. Na Seção 16.2, confirmamos que o MQO aplicado a uma equação em um sistema simultâneo é geralmente viesado e inconsistente.

A Seção 16.3 fornece uma descrição geral sobre identificação e estimação em um sistema de duas equações, enquanto a Seção 16.4 trata resumidamente de modelos com mais de duas equações. Modelos de equações simultâneas são usados para modelar séries temporais agregadas, e na Seção 16.5 incluímos uma discussão sobre alguns problemas especiais que surgem em tais modelos. A Seção 16.6 refere-se a modelos de equações simultâneas com dados em painel.

16.1 A NATUREZA DOS MODELOS DE EQUAÇÕES SIMULTÂNEAS

O ponto mais importante a lembrar no uso de modelos de equações simultâneas é que cada equação no sistema deve ter uma interpretação causal, *ceteris paribus*. Como somente observamos os resultados em equilíbrio, precisamos usar raciocínio contrafactual na construção de equações de um modelo de equações simultâneas. Devemos pensar em termos de resultados potenciais assim como de resultados efetivos.

O exemplo clássico de SEM é uma equação de oferta e demanda de alguma mercadoria ou de algum insumo na produção (como a mão de obra). Concretamente, sejam h_s o total anual de horas cumpridas por trabalhadores na agricultura, medidas em nível municipal e w a média do salário por hora oferecida a tais trabalhadores. Uma função simples da oferta de mão de obra é

$$h_s = \alpha_1 w + \beta_1 z_1 + u_1, \tag{16.1}$$

em que z_1 é alguma variável observada afetando a oferta de mão de obra — digamos, a média dos salários da indústria no município. O termo de erro, u_1, contém outros fatores que afetam a oferta de mão de obra. [Muitos desses fatores são observados e poderiam ser incluídos na equação (16.1); para ilustrar os conceitos básicos, incluímos somente um de tais fatores, z_1.] A equação (16.1) é um exemplo de uma **equação estrutural**. Esse nome tem origem no fato de ser a função de oferta de mão de obra derivável da teoria econômica e tem uma interpretação causal. O coeficiente α_1 indica como a oferta de mão de obra muda quando o salário muda; se h_s e w estiverem na forma logarítmica, α_1 será a elasticidade da oferta de mão de obra. Em geral, esperamos que α_1 seja positiva (embora a teoria econômica não impeça $\alpha_1 \leq 0$). As elasticidades da oferta de mão de obra são importantes na determinação de como os trabalhadores alterarão o número de horas que desejam trabalhar quando os impostos sobre os salários se alteram. Se z_1 for o salário industrial, esperamos $\beta_1 \leq 0$: com outros fatores permanecendo iguais, se o salário industrial aumenta, mais trabalhadores irão para a indústria do que para a agricultura.

Quando fazemos o gráfico da oferta de mão de obra, descrevemos horas como uma função do salário, com z_1 e u_1 mantidos fixos. Uma alteração em z_1, assim como uma mudança em u_1, desloca a função de oferta de mão de obra. A diferença é que z_1 é observado, enquanto u_1 não é. Algumas vezes, z_1 é chamado de *deslocador observado da oferta*, e u_1 é chamado de *deslocador não observado da oferta*.

Como a equação (16.1) difere das que estudamos anteriormente? A diferença é sutil. Embora a equação (16.1) pretensamente deva ser válida para todos os valores possíveis de salários, não podemos, de forma geral, ver os salários variando exogenamente em um corte transversal de municípios. Se pudéssemos calcular um experimento no qual variássemos os níveis salariais industrial e agrícola por meio de amostra de municípios e pesquisar os trabalhadores para obtermos a oferta de mão de obra h_s, então, poderíamos estimar (16.1) por MQO. Infelizmente, esse não é um experimento exequível. Em vez disso, temos que coletar dados sobre salários médios nesses dois setores com informações sobre quantas horas-homem foram empregadas na produção agrícola. Ao decidir como analisar esses dados, devemos entender que eles são melhor descritos pela interação entre a oferta e a demanda de mão de obra. Sob a hipótese de que os mercados de mão de obra compensam-se mutuamente, de fato, observamos valores de *equilíbrio* de salários e horas trabalhadas.

Para descrever como os salários e horas de equilíbrio são determinados, necessitamos introduzir a demanda por mão de obra, que supomos ser dada por

$$h_d = \alpha_2 w + \beta_2 z_2 + u_2, \tag{16.2}$$

em que h_d representa horas demandadas. Como na função de oferta, escrevemos horas demandadas como uma função dos salários, w, mantendo z_2 e u_2 fixos. A variável z_2 — digamos área agrícola — é um *deslocador observável da demanda*, enquanto u_2 é um *deslocador não observável da demanda*.

Da mesma forma que na equação da oferta de mão de obra, a equação da demanda por mão de obra é uma equação estrutural: ela pode ser obtida a partir de considerações sobre a maximização de lucros dos fazendeiros. Se h_d e w estiverem em forma logarítmica, α_2 será a elasticidade da demanda por mão de obra. A teoria econômica nos diz que $\alpha_2 < 0$. Como mão de obra e terra são complementares na produção, esperamos $\beta_2 > 0$.

Observe como as equações (16.1) e (16.2) descrevem relações totalmente diferentes. A oferta de mão de obra é uma equação comportamental dos trabalhadores, e a demanda por mão de obra é uma relação comportamental dos fazendeiros. Cada equação tem uma interpretação *ceteris paribus* e é autossuficiente. Elas se tornam interligadas em uma análise econométrica somente porque salários e horas *observados* são determinados pela interseção da oferta e da demanda. Em outras palavras, em cada município i, as horas observadas h_i e os salários observados w_i são determinados pela condição de equilíbrio

$$h_{is} = h_{id}. \tag{16.3}$$

Como somente observamos horas de equilíbrio de cada município i, representamos horas observadas por h_i.

Quando combinamos a condição de equilíbrio em (16.3) com as equações de oferta e de demanda de mão de obra, obtemos

$$h_i = \alpha_1 w_i + \beta_1 z_{i1} + u_{i1} \tag{16.4}$$

e

$$h_i = \alpha_2 w_i + \beta_2 z_{i2} + u_{i2}, \tag{16.5}$$

em que explicitamente incluímos o subscrito i para enfatizar que h_i e w_i são os valores de equilíbrio observados de cada município i. Essas duas equações constituem um **modelo de equações simultâneas (SEM)**, que tem várias características importantes. Primeiro, dadas z_{i1}, z_{i2}, u_{i1} e u_{i2}, essas duas equações determinam h_i e w_i. (Na realidade, devemos presumir que $\alpha_1 \neq \alpha_2$, e significa que as inclinações das funções da oferta e da demanda diferem; veja o Problema 16.1 no final deste Capítulo.) Por essa razão, h_i e w_i são as **variáveis endógenas** nesse SEM. O que dizer de z_{i1} e de z_{i2}? Como elas são determinadas fora do modelo, as vemos como **variáveis exógenas**. Do ponto de vista estatístico, a hipótese fundamental concernente a z_{i1} e z_{i2} é que ambas são não correlacionadas com os erros da oferta e da demanda, u_{i1} e u_{i2}, respectivamente. Esses são exemplos de **erros estruturais** porque eles aparecem nas equações estruturais.

Um segundo ponto importante é que, sem a inclusão de z_1 e z_2 no modelo, não existe maneira de dizer qual das equações é a função de oferta e qual é a função de demanda. Quando z_1 representa salários industriais, o raciocínio econômico nos diz que ele é um fator na oferta de mão de obra agrícola, pois ele é uma indicação do custo da oportunidade de trabalhar na agricultura; quando z_2 representa a área agrícola, a teoria da produção sugere que ele apareça na função de demanda de mão de obra. Portanto, sabemos que (16.4) representa a oferta de mão de obra e (16.5) representa a demanda por mão de obra. Se z_1 e z_2 forem os mesmos — por exemplo, nível médio de educação dos adultos no município, que pode afetar tanto a oferta como a demanda —, as equações parecerão idênticas, e não há possibilidade de estimar qualquer uma delas. Resumidamente, isso ilustra o problema de identificação em modelos de equações simultâneas, que examinaremos de forma mais geral na Seção 16.3.

Os exemplos mais convincentes de SEM têm as mesmas formas dos exemplos de oferta e demanda. Cada equação deve ter uma interpretação comportamental própria, *ceteris paribus*. Como somente observamos resultados de equilíbrio, a especificação de um SEM exige que façamos perguntas contrafactuais como: quanta mão de obra os trabalhadores *ofereceriam* se os salários fossem diferentes de seus valores de equilíbrio? O Exemplo 16.1 oferece outra ilustração de um SEM na qual cada equação tem uma interpretação *ceteris paribus*.

EXEMPLO 16.1

(Taxa de Assassinatos e Tamanho da Força Policial)

As municipalidades frequentemente querem determinar em que proporção a imposição da lei diminuirá suas taxas de assassinatos. Um modelo simples de corte transversal para tratar dessa questão é

$$assaspc = \alpha_1 polpc + \beta_{10} + \beta_{11} rendapc + u_1, \qquad (16.6)$$

em que *assasspc* representa assassinatos *per capita*, *polpc* significa policiais *per capita* e *rendapc* é a renda *per capita*. (Deste ponto em diante, não incluiremos um subscrito *i*.) Consideramos a renda *per capita* como exógena nessa equação. Na prática, incluiríamos outros fatores, como as distribuições de idade e sexo, níveis de educação, talvez variáveis geográficas, e variáveis que indicassem a severidade da punição. Para organizar o raciocínio, consideramos a equação (16.6).

A questão que esperamos responder é: se uma cidade aumentar exogenamente sua força policial, esse aumento, em média, reduzirá a taxa de assassinatos? Se pudermos escolher exogenamente os tamanhos das forças policiais para uma amostra aleatória de cidades, poderíamos estimar (16.6) por MQO. Certamente, não podemos fazer tal experimento. Entretanto, podemos, de qualquer maneira, imaginar o tamanho da força policial como sendo exogenamente determinada? Provavelmente, não. O gasto de uma cidade com a imposição da lei é pelo menos parcialmente determinado pela taxa esperada de assassinatos. Para refletir isso, postulamos uma segunda relação:

$$polpc = \alpha_2 assaspc + \beta_{20} + outros\ fatores. \qquad (16.7)$$

Esperamos que $\alpha_2 > 0$: outros fatores sendo iguais, as cidades com taxas (esperadas) de homicídios mais elevadas terão mais policiais *per capita*. Assim que especificarmos os outros fatores em (16.7), teremos um modelo de equações simultâneas com duas equações. Na verdade, estamos interessados somente na equação (16.6), mas, como veremos na Seção 16.3, precisamos saber precisamente como a segunda equação é especificada para estimarmos a primeira.

Um ponto importante é que (16.7) descreve o comportamento dos policiais da cidade, enquanto (16.6) descreve as ações dos assassinos em potencial. Isso dá a cada equação uma clara interpretação *ceteris paribus*, o que faz das equações (16.6) e (16.7) um modelo de equações simultâneas apropriado.

A seguir damos um exemplo de uso inapropriado de SEM.

EXEMPLO 16.2

(Despesas e Poupança Familiares)

Suponha que, para uma família escolhida aleatoriamente na população, presumimos que os gastos e poupança familiares anuais sejam conjuntamente determinados por

$$gastofam = \alpha_1 poupfam + \beta_{10} + \beta_{11} renda + \beta_{12} educ + \beta_{13} idade + u_1 \qquad (16.8)$$

e

$$poupfam = \alpha_2 gastofam + \beta_{20} + \beta_{21} renda + \beta_{22} educ + \beta_{23} idade + u_2, \qquad (16.9)$$

em que *renda* é a renda anual e *educ* e *idade* são indicados em anos. Inicialmente, pode parecer que essas duas equações são uma maneira sensata de verificar como os gastos com habitação e poupança são determinados. Contudo, temos que perguntar: que valor teria uma dessas equações sem a outra? Nenhuma delas tem uma interpretação *ceteris paribus*, pois *gastofam* e *poupfam* são escolhidas pela mesma família. Por exemplo, não faz sentido fazer essa pergunta: se a renda anual crescer em US$ 10.000, como seriam alterados os gastos domésticos, *mantendo a poupança fixa*? Se a renda familiar aumentar, uma família geralmente alterará a composição ótima de gastos domésticos e poupança. Entretanto, a equação (16.8) faz parecer que queremos saber o efeito da alteração de *renda*, *educ* ou *idade*, mantendo a *poupança* fixa. Um experimento com esse enfoque não é interessante. Qualquer modelo baseado em princípios econômicos, particularmente a maximização da utilidade, teria a família feito a escolha ótima de *gastofam* e *poupfam* como funções da *renda* e dos preços relativos dos gastos domésticos e poupança. As variáveis *educ* e *idade* afetarão preferências de consumo, poupança e risco. Portanto, *gastofam* e *poupfam* serão cada uma função da renda, educação, idade e outras variáveis que afetem o problema da maximização da utilidade (tais como as diferentes taxas de retorno sobre gastos domiciliares e outras poupanças).

Mesmo que decidamos que os SEM em (16.8) e (16.9) têm lógica, não há maneira de estimarmos os parâmetros. (Discutiremos esse problema de forma mais geral na Seção 16.3.) As duas equações são indistintas, a menos que presumamos que renda, educação ou idade apareçam em uma equação, mas não na outra, o que não faria sentido.

Embora esse seja um exemplo pobre do SEM, podemos ter interesse em verificar se, com os outros fatores mantidos fixos, existe uma relação de substituição entre os gastos domésticos e a poupança. Contudo, nesse caso, estimaríamos somente, digamos (16.8) por MQO, a menos que haja um problema de variável omitida ou de erro de medida.

O Exemplo 16.2 tem as características de um grande número de aplicações SEM. O problema é que duas variáveis endógenas são selecionadas pelo mesmo agente econômico. Portanto, nenhuma das equações é autossuficiente. Outro exemplo de uso não apropriado de um SEM seria modelar horas semanais gastas estudando e horas semanais gastas trabalhando. Cada aluno selecionará essas variáveis simultaneamente — presumivelmente como uma função dos rendimentos que podem ser obtidos com o trabalho, talento como aluno, entusiasmo pela faculdade e assim por diante. Da mesma forma que no Exemplo 16.2, não faz sentido especificar duas equações em que cada uma é uma função da outra. A lição importante aqui: apenas o fato de duas variáveis serem determinadas simultaneamente *não* significa que um modelo de equações simultâneas seja adequado. Para que um SEM faça sentido, cada

equação deve ter uma interpretação *ceteris paribus* em separado da outra equação. Como discutido anteriormente, exemplos de demanda e oferta e o Exemplo 16.1 têm este componente. O raciocínio econômico básico, apoiado em alguns casos por modelos econômicos simples, pode nos ajudar a usar os SEMs de forma inteligente (e saber quando não usar o SEM).

QUESTÃO 16.1

Pindyck e Rubinfeld (1992, Seção 11.6) descrevem um modelo de publicidade no qual firmas monopolistas escolhem níveis de preços e gastos com propaganda que maximizam lucros. Isso significa que deveríamos usar um SEM para modelar essas variáveis no nível da firma?

16.2 VIÉS DE SIMULTANEIDADE NO MQO

É proveitoso ver, em um modelo simples, que uma variável explicativa que é determinada simultaneamente com a variável dependente geralmente é correlacionada com o termo de erro, o que conduz a viés e inconsistência no MQO. Consideremos o modelo estrutural de duas equações

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1 \qquad (16.10)$$

$$y_2 = \alpha_2 y_1 + \beta_2 z_2 + u_2 \qquad (16.11)$$

e nos concentremos em estimar a primeira equação. As variáveis z_1 e z_2 são exógenas, de forma que cada uma é não correlacionada com u_1 e u_2. Para simplificar, suprimimos o intercepto em cada equação.

Para mostrar que y_2 geralmente é correlacionada com u_1, solucionamos as duas equações para y_2 em termos das variáveis exógenas e do termo de erro. Se inserirmos o lado direito de (16.10) em y_1 para (16.11), obteremos

$$y_2 = \alpha_2(\alpha_1 y_2 + \beta_1 z_1 + u_1) + \beta_2 z_2 + u_2$$

ou

$$(1 - \alpha_2\alpha_1)y_2 = \alpha_2\beta_1 z_1 + \beta_2 z_2 + \alpha_2 u_1 + u_2. \qquad (16.12)$$

Agora, devemos fazer uma hipótese sobre os parâmetros para solucionar a equação para y_2:

$$\alpha_2\alpha_1 \neq 1. \qquad (16.13)$$

Com relação a essa hipótese ser restritiva, depende da aplicação. No Exemplo 16.1, entendemos que $\alpha_1 \leq 0$ e $\alpha_2 \geq 0$, o que implica $\alpha_1\alpha_2 \leq 0$; portanto (16.13) é bastante razoável para o Exemplo 16.1.

Desde que a condição (16.13) se mantenha, podemos dividir (16.12) por $(1 - \alpha_2\alpha_1)$ e escrever y_2 como

$$y_2 = \pi_{21} z_1 + \pi_{22} z_2 + v_2, \qquad (16.14)$$

em que $\pi_{21} = \alpha_2\beta_1/(1 - \alpha_2\alpha_1)$, $\pi_{22} = \beta_2/(1 - \alpha_2\alpha_1)$, e $v_2 = (\alpha_2 u_1 + u_2)/(1 - \alpha_2\alpha_1)$. A equação (16.14), que expressa y_2 em termos das variáveis exógenas e dos termos de erro, é a equação da **forma reduzida** de y_2, um conceito apresentado no Capítulo 15 no contexto da estimação de variáveis instrumentais. Os parâmetros π_{21} e π_{22} são chamados **parâmetros da forma reduzida**; observe como eles são funções não lineares dos **parâmetros estruturais**, que aparecem nas equações estruturais (16.10) e (16.11).

O **erro na forma reduzida**, v_2, é uma função linear dos termos de erro estruturais, u_1 e u_2. Como u_1 e u_2 são, individualmente, não correlacionados com z_1 e z_2, v_2 também é não correlacionado com z_1 e z_2. Portanto, podemos consistentemente estimar π_{21} e π_{22} por MQO, algo que é usado para a estimação por mínimos quadrados em dois estágios (ao qual retornaremos na próxima seção). Além disso, os parâmetros da forma reduzida são algumas vezes de interesse direto, embora estejamos, aqui, nos concentrando em estimar a equação (16.10).

Também existe uma forma reduzida de y_1 sob a hipótese (16.13); a álgebra é semelhante à usada para obter (16.14). Ela tem as mesmas propriedades da forma reduzida da equação de y_2.

Podemos usar a equação (16.14) para mostrar que, exceto sob hipóteses especiais, a estimação por MQO da equação (16.10) produzirá estimadores de α_1 e β_1 viesados e inconsistentes na equação (16.10). Como z_1 e u_1 pressupõe-se são não correlacionados, o problema está em saber se y_2 e u_1 são não correlacionados. A partir da forma reduzida em (16.14), vemos que y_2 e u_1 serão correlacionados se, e somente se, v_2 e u_1 forem correlacionados (pois z_1 e z_2 são considerados exógenos). Porém, v_2 é uma função linear de u_1 e u_2, de modo que ele geralmente é correlacionado com u_1. Na verdade, se considerarmos u_1 e u_2 não correlacionados, v_2 e u_1 *devem* ser correlacionados sempre que $\alpha_2 \neq 0$. Mesmo se α_2 for igual a zero — significando que y_1 não aparece na equação (16.11) —, v_2 e u_1 serão correlacionados se u_1 e u_2 forem correlacionados.

Quando $\alpha_2 = 0$ *e* u_1 e u_2 forem não correlacionados, y_2 e u_1 também serão não correlacionados. Esses são requisitos bastante fortes: se $\alpha_2 = 0$, y_2 não será simultaneamente determinado com y_1. Se adicionarmos correlação zero entre u_1 e u_2, isso eliminará variáveis omitidas ou erro de medida em u_1 que sejam correlacionados com y_2. Não devemos nos surpreender com o fato de que a estimação por MQO da equação (16.10) funciona nesse caso.

Quando y_2 for correlacionado com u_1 em razão da simultaneidade, dizemos que o MQO sofre de **viés de simultaneidade**. A obtenção da direção do viés nos coeficientes é geralmente complicada, como vimos com o viés de variáveis omitidas nos Capítulos 3 e 5. Contudo, em modelos sem complexidade, podemos determinar a direção do viés. Por exemplo, suponha que simplifiquemos a equação (16.10) retirando z_1 da equação e presumindo que u_1 e u_2 são não correlacionados. Então, a covariância entre y_2 e u_1 será

$$\begin{aligned}\text{Cov}(y_2,u_1) &= \text{Cov}(v_2,u_1) = [\alpha_2/(1 - \alpha_2\alpha_1)]\text{E}(u_1^2) \\ &= [\alpha_2/(1 - \alpha_2\alpha_1)]\sigma_1^2,\end{aligned}$$

em que $\sigma_1^2 = \text{Var}(u_1) > 0$. Portanto, o viés assimptótico (ou a inconsistência) no estimador MQO de α_1 terá o mesmo sinal de $\alpha_2/(1 - \alpha_2\alpha_1)$. Se $\alpha_2 > 0$ e $\alpha_2\alpha_1 < 1$, o viés assimptótico será positivo. (Infelizmente, como no caso de nosso cálculo do viés de variáveis omitidas da Seção 3.3 do Capítulo 3,

as conclusões não são transportadas para modelos mais gerais. Porém, elas servem como um guia útil.) Veja-se, no Exemplo 16.1, pensamos que $\alpha_2 > 0$ e $\alpha_2\alpha_1 \leq 0$, o que significa que o estimador MQO de α_1 teria um viés positivo. Se $\alpha_1 = 0$, o MQO estimará, em média, um impacto *positivo* de mais policiais sobre a taxa de assassinatos; geralmente, o estimador de α_1 é viesado para cima. Como esperamos um aumento no tamanho da força policial para reduzir as taxas de criminalidade (*ceteris paribus*), o viés para cima significa que o MQO subestimará a efetividade de uma força policial maior.

16.3 A IDENTIFICAÇÃO E A ESTIMAÇÃO DE UMA EQUAÇÃO ESTRUTURAL

Como vimos na Seção anterior, o MQO é viesado e inconsistente quando aplicado a uma equação estrutural em um sistema de equações simultâneas. No Capítulo 15, aprendemos que o método dos mínimos quadrados em dois estágios pode ser usado para solucionar o problema de variáveis explicativas endógenas. Agora mostramos como o MQ2E pode ser aplicado a SEM.

A mecânica do MQ2E é semelhante à do Capítulo 15. A diferença é que, como especificamos uma equação estrutural para cada variável endógena, podemos imediatamente verificar se existem VIs suficientes para estimar qualquer equação. Iniciamos discutindo o problema da identificação.

A Identificação em um Sistema de Duas Equações

Mencionamos a noção de identificação no Capítulo 15. Quando estimamos um modelo por MQO, a condição crucial de identificação é que cada variável explicativa seja não correlacionada com o termo de erro. Como demonstramos na Seção 16.2, de forma geral, essa condição fundamental não se mantém, para os SEMs. Porém, se tivermos algumas variáveis instrumentais, poderemos ainda identificar (ou estimar consistentemente) os parâmetros em uma equação SEM, da mesma forma com variáveis omitidas ou erro de medida.

Antes de considerarmos um SEM geral de duas equações, é útil adquirirmos conhecimento intuitivo ao considerarmos um exemplo simples de oferta e demanda. Escreva o sistema na forma de equilíbrio (isto é, impondo $q_s = q_d = q$) como

$$q = \alpha_1 p + \beta_1 z_1 + u_1 \qquad \textbf{(16.15)}$$

e

$$q = \alpha_2 p + u_2. \qquad \textbf{(16.16)}$$

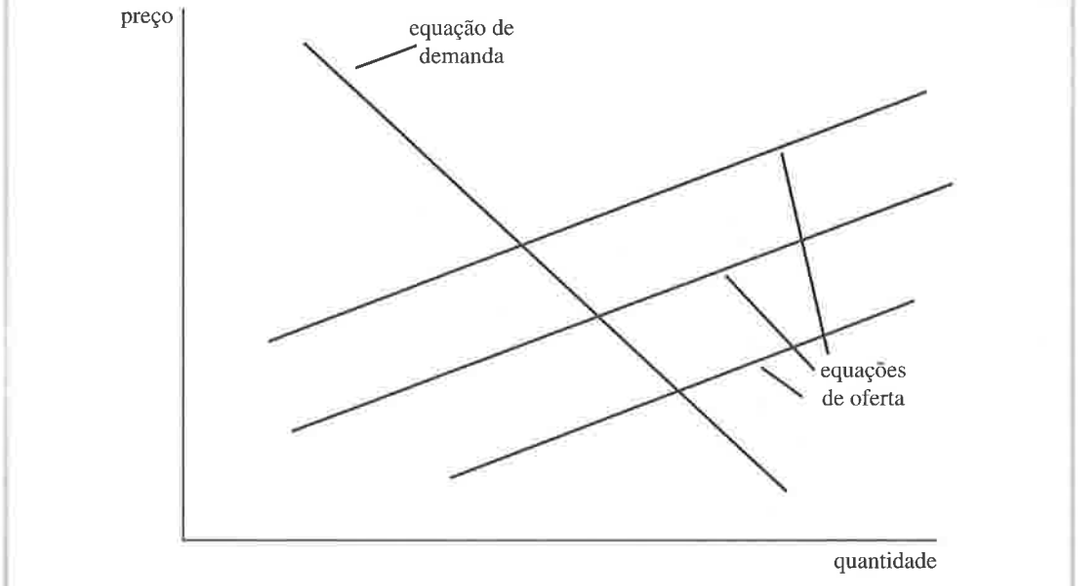
Concretamente, sejam q o consumo *per capita* de leite em nível municipal, p o preço médio por litro de leite no município e z_1 o preço da alimentação do gado, que consideremos ser exógeno nas equações de oferta e demanda de leite. Isso significa que (16.15) deve ser a função de oferta, já que o preço da alimentação do gado deslocará a oferta ($\beta_1 < 0$), mas não a demanda. A função de demanda não contém deslocadores observados da demanda.

Dada uma amostra aleatória de (q, p, z_1), qual dessas equações será estimada? Isto é, qual delas é uma **equação identificada**? É possível constatar que a equação de *demanda* (16.16) é identificada, mas a equação da oferta não é. Isso é fácil de verificar usando as regras de estimação de VI do Capítulo 15: podemos usar z_1 como uma VI do preço na equação (16.16). Porém, como z_1 aparece na equação (16.15), não temos uma VI do preço na equação de oferta.

Intuitivamente, o fato de a equação de demanda ser identificada é uma consequência de termos uma variável observada, z_1, que desloca a equação de oferta sem afetar a equação de demanda. Dada uma variação em z_1 e nenhum erro, podemos desenhar a curva de demanda, como mostrado na Figura 16.1. A presença do deslocador não observado da demanda u_2 faz com que estimemos a equação de demanda com erro, mas os estimadores serão consistentes, desde que z_1 seja não correlacionado com u_2.

Figura 16.1

O deslocamento nas equações de oferta permite desenhar a equação de demanda. Cada equação de oferta é traçada para um valor diferente da variável exógena, z_1.



A equação da oferta não pode ser desenhada porque não existem fatores exógenos não observados deslocando a curva de demanda. Não ajuda o fato de haver fatores não observados deslocando a função de demanda; necessitamos de algo observado. Se, como na função de demanda da mão de obra (16.2), tivéssemos um deslocador observado da demanda exógeno — como a renda na função de demanda do leite —, a função de oferta também poderia ser identificada.

Resumindo: *no sistema de (16.15) e (16.16), a presença de uma variável exógena na equação da oferta é que nos possibilita estimar a equação de demanda.*

Estender a discussão sobre a identificação a um modelo geral de duas equações não apresenta dificuldades. Escreva as duas equações como

$$y_1 = \beta_{10} + \alpha_1 y_2 + \mathbf{z}_1\boldsymbol{\beta}_1 + u_1 \qquad \textbf{(16.17)}$$

e

$$y_2 = \beta_{20} + \alpha_2 y_1 + \mathbf{z}_2\boldsymbol{\beta}_2 + u_2, \qquad \textbf{(16.18)}$$

em que y_1 e y_2 são as variáveis endógenas e u_1 e u_2 são os termos de erro estruturais. O intercepto na primeira equação é β_{10}, e o intercepto na segunda equação é β_{20}. A variável $\mathbf{z}_1$ representa um conjunto de k_1 variáveis exógenas aparecendo na primeira equação: $\mathbf{z}_1 = (z_{11}, z_{12}, \ldots, z_{1k_1})$. De forma semelhante, $\mathbf{z}_2$ é o conjunto de k_2 variáveis exógenas na segunda equação: $\mathbf{z}_2 = (z_{21}, z_{22}, \ldots, z_{2k_2})$. Em muitos casos, $\mathbf{z}_1$ e $\mathbf{z}_2$ se sobreporão. De forma abreviada, usamos a notação

$$\mathbf{z}_1\boldsymbol{\beta}_1 = \beta_{11}z_{11} + \beta_{12}z_{12} + \ldots + \beta_{1k_1}z_{1k_1}$$

ou

$$\mathbf{z}_2\boldsymbol{\beta}_2 = \beta_{21}z_{21} + \beta_{22}z_{22} + \ldots + \beta_{2k_2}z_{2k_2};$$

isto é, $\mathbf{z}_1\boldsymbol{\beta}_1$ representa todas as variáveis exógenas na primeira equação, cada uma multiplicada por um coeficiente, e semelhantemente $\mathbf{z}_2\boldsymbol{\beta}_2$. (Alguns autores usam a notação $\mathbf{z}_1'\boldsymbol{\beta}_1$ e $\mathbf{z}_2'\boldsymbol{\beta}_2$. Se você tiver interesse na abordagem de álgebra matricial em econometria, veja o Apêndice E, disponível no site da Cengage.)

O fato de que $\mathbf{z}_1$ e $\mathbf{z}_2$ geralmente contêm variáveis exógenas diferentes significa termos imposto **restrições de exclusão** no modelo. Em outras palavras, *presumimos* que certas variáveis exógenas não aparecem na primeira equação e outras estão ausentes da segunda equação. Como vimos nos exemplos anteriores de oferta e demanda, isso nos possibilita distinguir entre duas equações estruturais.

Quando podemos solucionar as equações (16.17) e (16.18) para y_1 e y_2 (como funções lineares de todas as variáveis exógenas e dos erros estruturais u_1 e u_2)? A condição é a mesma que em (16.13), ou seja, $\alpha_2\alpha_1 \neq 1$. A prova é praticamente idêntica à do modelo simples da Seção 16.2. Sob essa hipótese, existirão formas reduzidas para y_1 e y_2.

A pergunta principal é: sob quais hipóteses podemos estimar os parâmetros em, digamos, (16.17)? Esse é o problema da identificação. A **condição de classificação** para a identificação da equação (16.17) é fácil de estabelecer.

CONDIÇÃO DE CLASSIFICAÇÃO PARA A IDENTIFICAÇÃO DE UMA EQUAÇÃO ESTRUTURAL

A primeira equação em um modelo de equações simultâneas com duas equações será identificada se, e somente se, a *segunda* equação contiver ao menos uma variável exógena (com um coeficiente diferente de zero) que seja excluída da primeira equação.

Essa é a condição necessária e suficiente para que a equação (16.17) seja identificada. A **condição de ordem**, que discutimos no Capítulo 15, é necessária para a condição de classificação. A condição de ordem para a identificação da primeira equação estabelece que pelo menos uma variável exógena seja excluída dessa equação. A condição de ordem é fácil de ser verificada, uma vez que ambas as equações tenham sido especificadas. A condição de classificação exige mais: pelo menos uma das variáveis exógenas excluídas da primeira equação deve ter um coeficiente populacional diferente de zero na segunda equação. Isso garante que pelo menos uma das variáveis exógenas omitidas da primeira equação efetivamente apareça na forma reduzida de y_2, de maneira a podermos usar essas variáveis como instrumentais de y_2. Podemos verificar isso usando um teste t ou F, como no Capítulo 15; seguem-se alguns exemplos.

A identificação da segunda equação é, naturalmente, apenas a imagem espelhada da declaração para a primeira equação. Além disso, se escrevermos as equações como no exemplo da oferta e demanda de mão de obra da Seção 16.1 — de forma que y_1 apareça no lado esquerdo em *ambas* as equações, com y_2 no lado direito —, as condições de identificação serão idênticas.

EXEMPLO 16.3

(Oferta de Mão de Obra de Mulheres Casadas que Trabalham)

Para ilustrar o problema da identificação, considere a oferta de mão de obra de mulheres casadas que já estejam na força de trabalho. Em lugar da função de demanda, escrevemos a oferta de salários como uma função de horas e das variáveis de produtividade habituais. Com a condição de equilíbrio imposta, as duas equações estruturais serão

$$horas = \alpha_1 \log(salárioh) + \beta_{10} + \beta_{11}educ + \beta_{12}idade + \beta_{13}crianmed6 + \beta_{14}nesprend + u_1 \quad (16.19)$$

e

$$horas(salárioh) = \alpha_2 horas + \beta_{20} + \beta_{21}educ + \beta_{22}exper + \beta_{23}exper^2 + u_2 \quad (16.20)$$

A variável *idade* é a idade da mulher, em anos, *crianmed6* é o número de filhos menores de seis anos de idade, *nesprend* é a renda de outra pessoa da família que não a mulher (que inclui os ganhos do marido), e *educ* e *exper* são anos de educação e de experiência anterior, respectivamente. Todas as variáveis, com exceção de *horas* e log(*salárioh*) são consideradas exógenas. (Essa é uma hipótese fraca, já que *educ* pode ser correlacionado com a aptidão omitida em cada uma das equações. Mas com o propósito de ilustração, ignoramos o problema da aptidão omitida.) A forma funcional nesse sistema — no qual *horas* aparece na forma de nível, mas *salário*-hora está na forma logarítmica — é comum em economia do trabalho. Podemos escrever esse sistema como nas equações (16.17) e (16.18), definindo $y_1 = horas$ e $y_2 = \log(salárioh)$.

A primeira equação é a função de oferta. Ela satisfaz a condição de ordem porque duas variáveis exógenas, *exper* e *exper*^{2}, são omitidas da equação de oferta de mão de obra. Essas restrições de exclusão são hipóteses cruciais: presumimos que, uma vez que salário, educação, idade, número de filhos pequenos e outras rendas sejam controlados, a experiência passada não tem efeito na oferta corrente de mão de obra. Certamente poderia se questionar essa hipótese, mas nós a usamos a título de ilustração.

Dadas as equações (16.19) e (16.20), a condição de classificação para identificar a primeira equação é que pelo menos uma das variáveis *exper* ou *exper*^{2} tenha um coeficiente diferente de zero na equação (16.20). Se $\beta_{22} = 0$ e $\beta_{23} = 0$, não haverá variáveis exógenas aparecendo na segunda equação que também não apareçam na primeira equação (*educ* aparece em ambas). Podemos estabelecer a condição de classificação para a identificação de (16.19) em equivalência com os termos da forma reduzida de log(*salário*), que é

$$\log(salárioh) = \pi_{20} + \pi_{21}educ + \pi_{22}idade + \pi_{23}crianmed6 + \pi_{24}nesprend + \pi_{25}exper + \pi_{26}exper^2 + v_2. \quad (16.21)$$

EXEMPLO 16.3 (continuação)

Para a identificação, necessitamos que $\pi_{25} \neq 0$ ou $\pi_{26} \neq 0$, o que podemos testar usando uma estatística F padrão, como discutimos no Capítulo 15.

A equação de oferta de salário (16.20), será identificada se pelo menos uma das variáveis *idade*, *crianmed6* ou *nesprend* tiver um coeficiente diferente de zero na equação (16.19). Isso é o mesmo que supor que a forma reduzida de *horas* — tem a mesma forma do lado direito de (16.21) — depende de pelo menos uma das variáveis *idade*, *crianmed6* ou *nesprend*. Na especificação da equação de oferta de salário, estamos *presumindo* que *idade*, *crianmed6* ou *nesprend* não têm efeito sobre a oferta de salário, uma vez que horas, educação e experiência sejam levadas em conta. Essas serão hipóteses pobres se essas variáveis de alguma maneira tiverem efeitos diretos sobre a produtividade, ou se as mulheres forem discriminadas com base em sua idade ou número de filhos pequenos.

No Exemplo 16.3, consideramos a população de interesse como a de mulheres casadas que estejam na força de trabalho (de forma que horas em equilíbrio são positivas). Isso exclui o grupo de mulheres casadas que escolheram não trabalhar fora de casa. A inclusão de tais mulheres no modelo provocaria alguns problemas intrincados. Circunstancialmente, se uma mulher não trabalha, não poderemos observar sua oferta de salário. Abordaremos brevemente esses problemas no Capítulo 17; mas, por enquanto, temos que pensar nas equações (16.19) e (16.20) como válidas somente para mulheres que tenham $horas > 0$.

EXEMPLO 16.4

(Inflação e Abertura da Economia)

Romer (1993) propõe modelos teóricos de inflação que sugerem que países mais "abertos" devem ter taxas de inflação mais baixas. Sua análise empírica interpreta taxas médias anuais de inflação (desde 1973) em termos da participação média das importações no produto interno (ou nacional) bruto (PNB) desde 1973 — que é sua medida de abertura da economia. Além de estimar a equação-chave por MQO, ele usa variáveis instrumentais. Embora Romer não especifique ambas as equações em um sistema simultâneo, ele tem em mente um sistema de duas equações:

$$inf = \beta_{10} + \alpha_1 abertura + \beta_{11}\log(rendpc) + u_1 \qquad \textbf{(16.22)}$$

$$abertura = \beta_{20} + \alpha_2 inf + \beta_{21}\log(rendpc) + \beta_{22}\log(área) + u_2, \qquad \textbf{(16.23)}$$

em que *rendpc* é a renda *per capita* de1980, em dólares dos Estados Unidos (presumida como exógena), e *área* é a área do país, em milhas quadradas (também presumida como exógena). A equação (16.22) é a de interesse, com a hipótese de que $\alpha_1 < 0$. (Economias mais abertas têm menores taxas de inflação.) A segunda equação reflete o fato de que o grau da abertura pode depender da taxa de inflação, como também de outros fatores. A variável log(*rendpc*) aparece em ambas as equações, mas log(*área*) aparece *supostamente* somente na segunda equação. A ideia é que, *ceteris paribus*, um país menor provavelmente será mais aberto (portanto, $\beta_{22} < 0$).

Usando a regra de identificação que foi declarada anteriormente, a equação (16.22) será identificada, desde que $\beta_{22} \neq 0$. A equação (16.23) *não* é identificada porque contém ambas as variáveis exógenas. Mas estamos interessados em (16.22).

QUESTÃO 16.2

Se tivermos o crescimento da oferta de moeda desde 1973 de cada país, que presumimos ser exógeno, isso auxiliará a identificar a equação (16.23)?

Estimação por MQ2E

Uma vez que tenhamos determinado que uma equação é identificada, podemos estimá-la por mínimos quadrados em dois estágios. As variáveis instrumentais consistirão das variáveis exógenas que aparecem em cada equação.

EXEMPLO 16.5

(Oferta de Mão de Obra de Mulheres Casadas que Trabalham)

Utilizamos os dados sobre mulheres casadas que trabalham contidos no arquivo MROZ.RAW para estimar a equação da oferta de mão de obra (16.19) por MQ2E. No conjunto total de variáveis instrumentais estão incluídas *educ*, *idade*, *crianmed6*, *nesprend*, *exper* e *exper*^{2}. A curva da oferta de mão de obra é

$$\widehat{horas} = \underset{(574,56)}{2.225,66} + \underset{(470,58)}{1.639,56}\log(salárioh) - \underset{(59,10)}{183,75}\,educ$$
$$\underset{(9,38)}{-7,81}\,idade - \underset{(182,93)}{198,15}\,crianmed6 - \underset{(6,61)}{10,17}\,nesprend,\ n = 428, \qquad \textbf{(16.24)}$$

e mostra que a curva da oferta de mão de obra tem inclinação para cima. O coeficiente estimado de log(*salário*) tem a seguinte interpretação: mantendo fixos os outros fatores, $\Delta\widehat{horas} \approx 16,4(\%\Delta salárioh)$. Podemos calcular as elasticidades da oferta de mão de obra multiplicando ambos os lados dessa equação por 100/*horas*:

$$100 \cdot (\Delta\widehat{horas}/horas) \approx (1.640/horas)(\%\Delta salárioh)$$

ou

$$\%\Delta\widehat{horas} \approx (1.640/horas)(\%\Delta salárioh),$$

e implica que a elasticidade da oferta de mão de obra (com relação a salário) é simplesmente 1.640/*horas*. [A elasticidade não é constante nesse modelo porque *horas*, e não log(*horas*), é a variável dependente em (16.24).] Na média de horas trabalhadas, 1.303, a elasticidade estimada é 1.640/1.303 ≈ 1,26, implicando um aumento maior que 1% nas horas trabalhadas, dado um aumento de 1% no salário. Essa é uma grande elasticidade estimada. Com maior número de horas, a elasticidade será menor; com menor número de horas, como *horas* = 800, a elasticidade é maior que dois.

Comparativamente, quando (16.19) é estimada por MQO, o coeficiente de log(*salárioh*) é −2,05 (ep = 54,88), o que implica não haver nenhum efeito do salário sobre as horas trabalhadas. Para confirmar que log(*salárioh*) é de fato endógeno em (16.19), podemos aplicar o teste da Seção 15.5. Quando adicionamos os resíduos da forma reduzida $\hat{v}_2$ na equação e a estimamos por MQO, a estatística t de $\hat{v}_2$ é −6,61, que é muito significante e, portanto, log(*salárioh*) parece ser endógeno.

EXEMPLO 16.5 (continuação)

A equação da oferta de salário (16.20) também pode ser estimada por MQ2E. O resultado será

$$\widehat{\log(salário)} = -0{,}656 + 0{,}00013\ horas + 0{,}110\ educ \quad (16.25)$$
$$(0{,}338)\ (0{,}00025)\ \ (0{,}016)$$
$$+ 0{,}035\ exper - 0{,}00071\ exper^2,\ n = 428.$$
$$(0{,}019)\ \ (0{,}00045)$$

Essa equação difere das equações de salários anteriores pelo fato de que *horas* é incluída como uma variável explicativa e o MQ2E são usados para levar em conta a endogeneidade de *horas* (e presumimos que *educ* e *exper* são exógenos). O coeficiente de *horas* é estatisticamente não significante, e significa que não existe evidência de que a oferta de salário cresce com as horas trabalhadas. Os outros coeficientes são semelhantes aos que obteremos se eliminarmos *horas* e estimarmos a equação por MQO.

A estimativa do efeito da abertura sobre a inflação por variáveis instrumentais também é feita de forma direta.

EXEMPLO 16.6

(Inflação e Abertura da Economia)

Antes de estimar (16.22) usando os dados contidos no arquivo OPENNESS.RAW, fazemos uma verificação para ver se *abertura* tem correlação parcial suficiente com a VI proposta, log(*área*). A regressão da forma reduzida é

$$\widehat{abertura} = 117{,}08 + 0{,}546\log(rendpc) - 7{,}57\log(área)$$
$$(15{,}85)\ (1{,}493)\ \ (0{,}81)$$
$$n = 114,\ R^2 = 0{,}449.$$

A estatística *t* de log(*área*) é maior que nove, em valor absoluto, o que ratifica a assertiva de Romer, de que países menores são mais abertos economicamente. O fato de que log(*rendpc*) seja tão não significante nessa regressão é irrelevante.

A estimação de (16.22) usando log(*área*) como uma VI de *abertura* produz

$$\widehat{inf} = 26{,}90 - 0{,}337\ abertura + 0{,}376\log(rendpc),\ n = 114. \quad (16.26)$$
$$(15{,}40)\ (0{,}144)\ \ (2{,}015)$$

O coeficiente de *abertura* é estatisticamente significante no nível de aproximadamente 1% contra uma alternativa bilateral ($\alpha_1 < 0$). O efeito também é economicamente importante: para cada ponto percentual de aumento na participação das importações no PIB, a inflação anual será de um terço de ponto percentual mais baixa. A título de comparação, a estimativa por MQO é −0,215 (ep = 0,095).

QUESTÃO 16.3

Como é possível testar se a diferença entre as estimativas por MQO e por VI de *abertura* é estatisticamente diferente?

16.4 SISTEMAS COM MAIS DE DUAS EQUAÇÕES

Modelos de equações simultâneas podem consistir de mais de duas equações. O estudo da identificação geral desses modelos é difícil e requer álgebra matricial. Uma vez que uma equação em um sistema geral tenha sido identificada, ela poderá ser estimada por MQ2E.

Identificação em Sistemas com Três ou mais Equações

Usaremos um sistema de três equações para ilustrar os problemas que surgem na identificação de SEM complicados. Com interceptos suprimidos, escreva o modelo como

$$y_1 = \alpha_{12}y_2 + \alpha_{13}y_3 + \beta_{11}z_1 + u_1 \quad (16.27)$$

$$y_2 = \alpha_{21}y_1 + \beta_{21}z_1 + \beta_{22}z_2 + \beta_{23}z_3 + u_2 \quad (16.28)$$

$$y_3 = \alpha_{32}y_2 + \beta_{31}z_1 + \beta_{32}z_2 + \beta_{33}z_3 + \beta_{34}z_4 + u_3, \quad (16.29)$$

em que os y_g são as variáveis endógenas, e os z_j são exógenas. O primeiro subscrito nos parâmetros indica o número da equação e o segundo o número da variável; usamos α para os parâmetros das variáveis endógenas e β para os parâmetros das variáveis exógenas.

Quais dessas equações podem ser estimadas? Mostrar que uma equação é identificada em um SEM com mais de duas equações geralmente é difícil, mas é fácil verificar quando certas equações *não* são identificadas. No sistema (16.27) a (16.29), podemos verificar facilmente que (16.29) cai nessa categoria. Como cada variável exógena aparece nessa equação, não temos qualquer VI de y_2. Portanto, não podemos consistentemente estimar os parâmetros dessa equação. Pelas razões discutidas na Seção 16.2, a estimação por MQO normalmente não será consistente.

E quanto à equação (16.27)? Aqui o assunto parece promissor, pois as variáveis z_2, z_3 e z_4 foram todas excluídas da equação — esse é um outro exemplo de *restrições de exclusão*. Embora haja duas variáveis endógenas nessa equação, temos três VIs em potencial para y_2 e y_3. Portanto, a equação (16.27) passa na condição de ordem. Para finalizar, declaramos a condição de ordem geral dos SEMs.

Condição de Ordem para a Identificação

Uma equação em qualquer SEM satisfaz a condição de ordem para a identificação se o número de variáveis exógenas *excluídas* da equação for pelo menos tão grande quanto o número de variáveis endógenas existentes no lado direito da equação.

A segunda equação, (16.28), também passa na condição de ordem porque existe uma variável exógena excluída, z_4, e uma variável endógena, y_1, no lado direito.

Como discutimos no Capítulo 15 e na seção anterior, a condição de ordem é somente necessária, não suficiente, para a identificação. Por exemplo, se $\beta_{34} = 0$, z_4 não aparecerá em qualquer lugar no sistema, o que significa que ela não é correlacionada com y_1, y_2 ou y_3. Se $\beta_{34} = 0$, então, a segunda equação não será identificada, porque z_4 será inútil como uma VI de y_1. Isso novamente ilustra que a identificação de uma equação depende dos valores dos parâmetros (que nunca conhecemos com certeza) nas outras equações.

Existem muitas maneiras sutis de a identificação falhar em SEMs complicados. Para obter condições suficientes, necessitaremos estender a condição de classificação para a identificação em sistemas de duas equações. Isso é possível, mas requer álgebra matricial [veja, por exemplo, Wooldridge (2002, Capítulo 9)]. Em muitas aplicações, é possível presumir que, a menos que haja falha óbvia de identificação, uma equação que satisfaça a condição de ordem é identificada.

A nomenclatura sobre equações sobreidentificadas e exatamente identificadas do Capítulo 15 tem origem nos SEMs. Em termos da condição de ordem, (16.27) é uma **equação sobreidentificada** porque precisamos somente de duas VIs (para y_2 e y_3), mas temos três disponíveis (z_2, z_3 e z_4); existe uma restrição sobreidentificadora nessa equação. Em geral, o número de restrições sobreidentificadoras iguala o número total de variáveis exógenas no sistema, menos o número total de variáveis explicativas na equação. Isso pode ser verificado usando o teste de sobreidentificação da Seção 15.5. A equação (16.28) é uma **equação exatamente identificada**, e a terceira equação é uma **equação não identificada**.

Estimação

A despeito do número de equações em um SEM, cada equação identificada poderá ser estimada por MQ2E. Instrumentos de uma equação particular consistirão nas variáveis exógenas que aparecerem em qualquer lugar no sistema. Poderão ser obtidos testes de endogeneidade, heteroscedasticidade, correlação serial e de restrições sobreidentificadoras, da mesma forma como demonstrado no Capítulo 15.

Quando qualquer sistema com duas ou mais equações é corretamente especificado e certas hipóteses adicionais são válidas, os *métodos de estimação de sistemas* são geralmente mais eficientes do que estimar cada equação por MQ2E. O método de estimação de sistemas mais comum no contexto dos SEMs é o dos *mínimos quadrados em três estágios*. Esses métodos, com ou sem variáveis explicativas endógenas, estão além do escopo deste texto. [Veja, por exemplo, Wooldridge (2002, Capítulos 7 e 8).]

16.5 MODELOS DE EQUAÇÕES SIMULTÂNEAS COM SÉRIES TEMPORAIS

Entre as aplicações mais antigas dos SEMs estavam as estimações de grandes sistemas de equações simultâneas, que eram usados para descrever a economia de um país. Um modelo simples keynesiano de demanda agregada (que ignora exportações e importações) é

$$C_t = \beta_0 + \beta_1(Y_t - T_t) + \beta_2 r_t + u_{t1} \quad \textbf{(16.30)}$$

$$I_t = \gamma_0 + \gamma_1 r_t + u_{t2} \quad \textbf{(16.31)}$$

$$Y_t \equiv C_t + I_t + G_t, \quad \textbf{(16.32)}$$

em que C_t é o consumo, Y_t é a renda, T_t é a receita de impostos, r_t é a taxa de juros, I_t são os investimentos e G_t são os gastos governamentais. [Veja, por exemplo, Mankiw (1994, Capítulo 9).] Especificamente, presuma que t representa ano.

A primeira equação é uma função consumo agregado, na qual o consumo depende da renda disponível, da taxa de juros e do erro estrutural não observado u_{t1}. A segunda equação é uma função de investimento bastante simples. A equação (16.32) é uma *identidade* que é um resultado da contabilidade da renda nacional: ela se mantém, por definição, sem erro. Assim, não faz sentido estimar (16.32), mas precisamos dessa equação para completar o modelo.

Como existem três equações no sistema, também deve haver três variáveis endógenas. Dadas as primeiras duas equações, é claro que pretendemos C_t e I_t endógenas. Além disso, em razão da identidade contábil, Y_t é endógena. Presumiremos, pelo menos nesse modelo, que T_t, r_t e G_t são exógenas, de forma que elas são não correlacionadas com u_{t1} e u_{t2}. (Discutiremos sobre problemas com esse tipo de hipótese mais tarde.)

Se r_t for exógena, então, a estimação da equação (16.31) por MQO será natural. A função consumo, porém, depende da renda disponível, que é endógena porque Y_t é endógena. Temos duas variáveis instrumentais disponíveis sob as hipóteses de exogeneidade: T_t e G_t. Portanto, se seguirmos nossa receita de estimação de equações de corte transversal, estimaremos (16.30) por MQ2E usando as variáveis instrumentais (T_t, G_t, r_t).

Modelos como (16.30) a (16.32) hoje em dia são raramente estimados, por várias e boas razões. Primeiro, é muito difícil justificar, em um nível agregado, a hipótese de que impostos, taxas de juros e gastos governamentais sejam exógenos. Impostos dependem clara e diretamente da renda; por exemplo, com uma única alíquota de imposto de renda τ_t no ano t, $T_t = \tau_t Y_t$. Podemos com facilidade permitir isso substituindo $(Y_t - T_t)$ por $(1 - \tau_t)Y_t$ em (16.30) e ainda poderemos estimar a equação por MQ2E se presumirmos que o gasto governamental é exógeno. Poderemos também adicionar a alíquota do imposto na lista das variáveis instrumentais, se ela for exógena. Entretanto, serão os gastos governamentais e as alíquotas de impostos realmente exógenos? Certamente, em princípio, eles poderiam ser, se o governo definisse os gastos e as alíquotas de impostos independentemente do que estivesse acontecendo com a economia. Mas esse é um caso difícil de acontecer na realidade: gastos governamentais geralmente dependem do nível da renda, e com altos níveis de renda, a mesma receita de impostos é arrecadada com alíquotas de impostos menores. Além disso, presumir que as taxas de juros sejam exógenas é extremamente questionável. Poderíamos especificar um modelo mais realista que incluísse a demanda e a oferta de moeda e, então, as taxas de juros poderiam ser determinadas com C_t, I_t e Y_t. Contudo, nesse caso, encontrar suficientes variáveis exógenas para identificar as equações se torna bastante difícil (e os problemas seguintes com esses modelos ainda são pertinentes).

Algumas pessoas têm argumentado que certos componentes dos gastos governamentais, como os gastos com a defesa — veja, por exemplo, Hall (1988) e Ramey (1991) —, são exógenos em uma variedade de aplicações de equações simultâneas. Porém, não há unanimidade sobre isso e, de qualquer forma, os gastos com a defesa nem sempre são apropriadamente correlacionados com as variáveis explicativas endógenas [veja Shea (1993) para uma discussão sobre o assunto e o Exercício em Computador 16.6, disponível no site da Cengage, para um exemplo].

Um segundo problema com modelo como (16.30) a (16.32) é que ele se apresenta completamente estático. Especialmente com dados mensais ou trimestrais, mas até mesmo com dados anuais, frequentemente esperamos ajustes de defasagens. (Um argumento a favor dos modelos estáticos do tipo keynesiano é que eles pretendem descrever a dinâmica de longo prazo sem se preocupar com a dinâmica de curto prazo.) Permitir a dinâmica não é muito difícil. Por exemplo, poderíamos adicionar renda defasada à equação (16.31):

$$I_t = \gamma_0 + \gamma_1 r_t + \gamma_2 Y_{t-1} + u_{t2}. \qquad \textbf{(16.33)}$$

Em outras palavras, adicionamos uma **variável endógena defasada** (mas não I_{t-1}) à equação de investimentos. Podemos tratar Y_{t-1} como exógena nessa equação? Sob certas hipóteses sobre u_{t2}, a resposta é sim. Entretanto, em geral, chamamos uma variável endógena defasada em um SEM de **variável predeterminada**. Defasagens de variáveis exógenas também são predeterminadas. Se presumirmos que u_{t2} é não correlacionado com as variáveis exógenas correntes (o que é padrão) e com todas as variáveis endógenas e exógenas *passadas*, então, Y_{t-1} será não correlacionada com u_{t2}. Dada a exogeneidade de r_t, podemos estimar (16.33) por MQO.

Se adicionarmos consumo defasado à equação (16.30), poderemos tratar C_{t-1} como exógeno nessa equação sob as mesmas hipóteses em u_{t1} que fizemos para u_{t2} no parágrafo anterior. A renda disponível corrente ainda será endógena em

$$C_t = \beta_0 + \beta_1(Y_t - T_t) + \beta_2 r_t + \beta_3 C_{t-1} + u_{t1}, \qquad \textbf{(16.34)}$$

de modo que poderemos estimar essa equação por MQ2E usando as variáveis instrumentais (T_t, G_t, r_t, C_{t-1}); se o investimento for determinado por (16.33), Y_{t-1} deve ser incluído na lista de instrumentais. [Para verificar o motivo, use (16.32), (16.33) e (16.34) para encontrar a forma reduzida de Y_t em termos das variáveis exógenas e predeterminadas: T_t, r_t, G_t, C_{t-1} e Y_{t-1}. Como Y_{t-1} aparece nessa forma reduzida, ela deve ser usada como uma VI.]

A presença de dinâmica em SEMs agregados é, pelo menos para o propósito de previsões, uma clara melhoria sobre os SEMs estáticos. Contudo, ainda existem alguns problemas importantes na estimação de SEM usando dados agregados de séries temporais, alguns dos quais discutimos nos Capítulos 11 e 15. Lembre-se de que a validade dos procedimentos de inferência dos habituais MQO ou MQ2E em aplicações de séries temporais depende da noção de *dependência fraca*. Infelizmente, séries como as de consumo agregado, renda, investimentos e até mesmo de taxas de juros parecem violar os requisitos de dependência fraca. (Na terminologia do Capítulo 11, elas têm *raízes unitárias*.) Essas séries também estão propensas a ter tendências exponenciais, embora isso possa ser parcialmente compensado pelo uso da transformação logarítmica quando presumiu diferentes formas funcionais. Geralmente, mesmo as propriedades de amostras grandes, sem falar das propriedades pequenas, do MQO e do MQ2E são complicadas e dependentes de várias hipóteses quando aplicadas a equações com variáveis I(1). Abordaremos levemente esses problemas no Capítulo 18. Uma abordagem geral avançada é dada por Hamilton (1994).

A discussão anterior significa que os SEMs não são úteis quando aplicados a dados de séries temporais? Muito ao contrário. Os problemas com tendências e alta persistência podem ser evitados especificando sistemas em primeiras diferenças ou em taxas de crescimento. Contudo, devemos reconhecer que esse é um SEM diferente de outro especificado em nível. [Por exemplo, se especificarmos o crescimento do consumo como uma função do crescimento da renda disponível e das alterações das taxas de juros, isso será diferente de (16.30).] Além disso, como discutimos anteriormente, a incorporação de dinâmica não é especialmente difícil. Finalmente, o problema de encontrar variáveis verdadeiramente exógenas para serem incluídas nos SEMs geralmente é mais fácil com dados desagregados. Por exemplo, para indústrias transformadoras, Shea (1993) descreve como a produção (ou, mais precisamente, o crescimento da produção) em outros setores pode ser usada como uma variável instrumental na estimação de funções de oferta. Ramey (1991) também contém uma análise convincente da estimação de funções de custos industriais por variáveis instrumentais utilizando dados de séries temporais.

O próximo exemplo mostra como dados agregados podem ser usados para testar uma importante teoria econômica, a teoria do consumo da renda permanente, habitualmente chamada de *hipótese da renda permanente* (HRP). A abordagem usada nesse exemplo não é, a rigor, baseada em um modelo de equações simultâneas, mas podemos pensar no consumo e no crescimento da renda (como também nas taxas de juros) como sendo determinados conjuntamente.

EXEMPLO 16.7

(O Teste da Hipótese da Renda Permanente)

Campbell e Mankiw (1990) usaram métodos de variáveis instrumentais para testar várias versões da hipótese da renda permanente. Usaremos os dados anuais de 1959 a 1995 do arquivo CONSUMP.RAW para reproduzir uma de suas análises. Campbell e Mankiw usaram dados trimestrais até 1985.

Uma equação estimada por Campbell e Mankiw (usando nossa notação) foi

$$cc_t = \beta_0 + \beta_1 cy_t + \beta_2 r3_t + u_t, \qquad \textbf{(16.35)}$$

em que $cc_t = \Delta\log(c_t)$ é o crescimento anual do consumo *per capita* real (excluindo bens duráveis), cy_t é o crescimento da renda disponível real, e $r3_t$ é a (*ex post*) taxa de juros real, medida pelo rendimento da taxa das letras do Tesouro norte-americano de três meses: $r3_t = i3_t - inf_t$, em que a taxa de inflação é baseada no índice de preços ao consumidor. As taxas de crescimento do consumo e da renda disponível não apresentam tendência e são fracamente dependentes; presumiremos também ser esse o caso da $r3_t$ para podermos aplicar a teoria assimptótica padrão.

A principal característica da equação (16.35) é que a HRP indica ter termo de erro u_t uma média zero condicional em todas as informações observadas no momento $t-1$ ou anterior: $E(u_t|I_{t-1}) = 0$. Porém, u_t não é necessariamente não correlacionado com cy_t ou com $r3_t$; uma maneira tradicional de se pensar sobre isso é que essas variáveis são conjuntamente determinadas, mas não estamos escrevendo um sistema completo de três equações.

Como u_t é não correlacionado com todas as variáveis datadas em $t-1$ ou antes, as variáveis instrumentais válidas para estimar (16.35) são valores defasados de *cc*, *cy* e *r3* (e defasagens de outras variáveis observáveis, mas não as usaremos aqui). Quais são as hipóteses de interesse? A forma pura da HRP tem $\beta_1 = \beta_2 = 0$. Campbell e Mankiw argumentam que β_1 será positivo se alguma fração da população consome renda corrente, em vez de renda permanente. A HRP com uma taxa de juros real não constante implica que $\beta_2 > 0$.

Quando estimamos (16.35) por MQ2E, usando as instrumentais cc_{-1}, cy_{-1} e $r3_{-1}$, obtemos

$$\widehat{cc}_t = 0{,}0081 + 0{,}586\, cy_t - 0{,}00027\, r3_t$$
$$(0{,}0032)\quad (0{,}135)\qquad (0{,}00076) \qquad \textbf{(16.36)}$$
$$n = 35,\ R^2 = 0{,}678.$$

Portanto, a forma pura da HRP é fortemente rejeitada porque o coeficiente de *cy* é economicamente grande (um aumento de 1% na renda disponível aumenta em mais de 5% o consumo) e estatisticamente significante ($t = 4{,}34$). De outro lado, o coeficiente da taxa de juros real é bastante pequeno e estatisticamente não significante. Essas constatações são qualitativamente as mesmas de Campbell e Mankiw.

EXEMPLO 16.7 (continuação)

A HRP também implica que os erros $\{u_t\}$ são serialmente não correlacionados. Após a estimação por MQ2E, obtemos os resíduos $\hat{u}_t$ e incluímos $\hat{u}_{t-1}$ como uma variável explicativa adicional em (16.36); ainda usamos os instrumentos cc_{t-1}, cy_{t-1}, $r3_{t-1}$, e $\hat{u}_{t-1}$ age como sua própria variável instrumental (veja a Seção 15.7). O coeficiente de $\hat{u}_{t-1}$ é $\hat{\rho} = 0{,}187$ (ep = 0,133), de modo que existe alguma evidência de correlação serial positiva, embora não no nível de significância de 5%. Campbell e Mankiw (1990) discutem o motivo pelo qual, com os dados trimestrais disponíveis, pode ser encontrada correlação serial nos erros mesmo se a HRP se sustentar; algumas dessas preocupações se estendem aos dados anuais.

QUESTÃO 16.4

Suponha que para uma cidade específica você tenha dados mensais sobre o consumo *per capita* de peixe, renda *per capita*, preços de peixe e preços de frango e carne bovina; a renda e os preços de frango e carne são exógenos. Presuma que não há sazonalidade na função de demanda de peixe, mas que ela existe na função de oferta de peixe. Como você pode usar essa informação para estimar uma equação de demanda de peixe com elasticidade constante? Especifique uma equação e detalhe a identificação. (Sugestão: Você deve ter onze variáveis instrumentais do preço de peixe.)

O uso de taxas de variáveis de crescimento de tendências ou variáveis I(1) em SEM é bastante comum em aplicações de séries temporais. Por exemplo, Shea (1993) estima as curvas de ofertas industriais em termos de taxas de crescimento.

Se um modelo estrutural contiver uma tendência temporal — que pode capturar fatores exógenos de tendência que não sejam diretamente modelados —, então, a tendência agirá como sua própria VI.

16.6 MODELOS DE EQUAÇÕES SIMULTÂNEAS COM DADOS EM PAINEL

Modelos de equações simultâneas também surgem no contexto de dados em painel. Por exemplo, podemos imaginar a estimação de equações de oferta de mão de obra e de oferta de salários, como no Exemplo 16.3, de um grupo de pessoas que tenha trabalhado por certo período de tempo. Além de permitirmos a determinação simultânea das variáveis dentro de cada período de tempo, podemos admitir efeitos não observados em cada equação. Em uma função de oferta de mão de obra, seria útil possibilitar a preferência por lazer não observado que não se altere ao longo do tempo.

A abordagem básica para estimar SEM com dados em painel compreende duas etapas: (1) eliminar os efeitos não observados das equações de interesse usando a transformação de efeitos fixos ou a primeira diferença e (2) encontrar variáveis instrumentais das variáveis endógenas na equação transformada. Isso pode ser bastante desafiador, pois para uma análise convincente, precisaremos encontrar variáveis instrumentais que mudem ao longo do tempo. Para ver o motivo disso, escreva um SEM de dados em painel como

$$y_{it1} = \alpha_1 y_{it2} + \mathbf{z}_{it1}\boldsymbol{\beta}_1 + a_{i1} + u_{it1} \qquad \textbf{(16.37)}$$

$$y_{it2} = \alpha_2 y_{it1} + \mathbf{z}_{it2}\boldsymbol{\beta}_2 + a_{i2} + u_{it2}, \qquad \textbf{(16.38)}$$

em que i representa o corte transversal, t é o período de tempo, e $\mathbf{z}_{it1}\boldsymbol{\beta}_1$ ou $\mathbf{z}_{it2}\boldsymbol{\beta}_2$ são funções lineares de um conjunto de variáveis explicativas exógenas em cada equação. A análise mais geral permite que os efeitos não observados, a_{i1} e a_{i2}, sejam correlacionados com *todas* as variáveis explicativas, mesmo os elementos em $\mathbf{z}$. Porém, presumimos que os erros estruturais idiossincráticos, u_{it1} e u_{it2}, são não correlacionados com $\mathbf{z}$ em ambas as equações e ao longo de todos os períodos de tempo; é nesse sentido que $\mathbf{z}$ é exógeno. Exceto sob circunstâncias especiais, y_{it2} será correlacionado com u_{it1}, e y_{it1} será correlacionado com u_{it2}.

Suponha que estamos interessados na equação (16.37). Não podemos estimá-la por MQO, já que o erro composto $a_{i1} + u_{it1}$ é potencialmente correlacionado com todas as variáveis explicativas. Suponha que tiremos a diferença ao longo do tempo para remover o efeito não observado, a_{i1}:

$$\Delta y_{it1} = \alpha_1 \Delta y_{it2} + \Delta \mathbf{z}_{it1}\boldsymbol{\beta}_1 + \Delta u_{it1}. \qquad \textbf{(16.39)}$$

(Como é usual na diferença ou na centralização na média, podemos estimar os efeitos de variáveis que se alteram ao longo do tempo de pelo menos algumas unidades de corte transversal.) Agora, o termo de erro nessa equação é não correlacionado com $\Delta\mathbf{z}_{it1}$ por hipótese. Mas Δy_{it2} e Δu_{it1} possivelmente serão correlacionados. Portanto, precisamos de uma VI para Δy_{it2}.

Como no caso de dados de corte transversal puro ou de séries temporais puras, as VIs possíveis virão da *outra* equação: elementos de $\mathbf{z}_{it2}$ que não estejam também em $\mathbf{z}_{it1}$. Na prática, precisamos de elementos com *variação temporal* em $\mathbf{z}_{it2}$ que não estejam também em $\mathbf{z}_{it1}$. Isso porque necessitamos de uma variável instrumental de Δy_{it2}, e é pouco provável que uma mudança em variável de um período para o próximo seja altamente correlacionada com o *nível* das variáveis exógenas. Na verdade, se diferenciarmos (16.38), veremos que as VIs naturais de Δy_{it2} são os elementos $\Delta\mathbf{z}_{it2}$ que não estão também em $\Delta\mathbf{z}_{it1}$.

Como exemplo dos problemas que podem surgir, considere uma versão de dados em painel da função de oferta de mão de obra no Exemplo 16.3. Após fazer a diferenciação, suponha que temos a equação

$$\Delta horas_{it} = \beta_0 + \alpha_1 \Delta \log(salário\mathrm{h}_{it}) + \Delta(outros\ fatores_{it}),$$

e queremos usar $\Delta exper_{it}$ como uma variável instrumental de $\Delta\log(salário\mathrm{h}_{it})$. O problema é que, como estamos examinando pessoas que trabalham em todos os períodos de tempo, $\Delta exper_{it} = 1$ para todos os i e t. (Todas as pessoas adquirem mais um ano de experiência após a passagem de um ano.) Não podemos usar uma VI que tenha o mesmo valor para todos os i e t e, portanto, temos de continuar a procura.

Frequentemente, a participação em um programa experimental pode ser usada para obter VIs em contextos de dados em painel. No Exemplo 15.10, usamos o recebimento de subsídios de treinamento de pessoal como uma VI da mudança nas horas de treinamento na determinação dos efeitos do treinamento de pessoal sobre a produtividade do trabalhador. De fato, poderíamos ver que, em um contexto de SEM, o treinamento de pessoal e a produtividade do trabalhador seriam determinados conjuntamente, mas o recebimento de um subsídio de treinamento de pessoal seria exógeno na equação (15.57).

Podemos algumas vezes propor variáveis instrumentais engenhosas e convincentes em aplicações de dados em painel, como ilustra o exemplo seguinte.

EXEMPLO 16.8

(Efeito da População Prisional sobre as Taxas de Crimes Violentos)

Para estimar o efeito causal do aumento da população prisional sobre as taxas de criminalidade em nível estadual, Levitt (1996) usou exemplos de processos judiciais sobre superlotação prisional como variáveis instrumentais do crescimento da população prisional. A equação que Levitt estimou estava em primeiras diferenças; podemos escrever um modelo de efeitos fixos subjacentes como

$$\log(crime_{it}) = \theta_t + \alpha_1 \log(prisão_{it}) + \mathbf{z}_{it1}\boldsymbol{\beta}_1 + a_{i1} + u_{it1}, \qquad \textbf{(16.40)}$$

em que θ_t representa diferentes interceptos temporais, e *crime* e *prisão* são medidos por 100.000 habitantes. (A variável da população prisional é avaliada no último dia do ano anterior.) O vetor $\mathbf{z}_{it1}$ contém log de policiais *per capita*, log de renda *per capita*, taxa de desemprego, etnia e as proporções de distribuições metropolitanas e por idade.

Tirando a diferença de (16.40) produz a equação estimada por Levitt:

$$\Delta\log(crime_{it}) = \xi_t + \alpha_1\Delta\log(prisão_{it}) + \Delta\mathbf{z}_{it1}\boldsymbol{\beta}_1 + \Delta u_{it1}. \qquad \textbf{(16.41)}$$

A simultaneidade entre taxas criminais e população prisional, ou mais precisamente nas taxas de crescimento, faz com que a estimação por MQO de (16.41) seja geralmente inconsistente. Usando a taxa de crimes violentos e um subconjunto dos dados de Levitt (no arquivo PRISON.RAW, dos anos de 1980 a 1993, com um total de 51·14 = 714 observações), obtemos a estimativa por MQO agrupado de α_1, que é −0,181 (ep = 0,048). Também estimamos (16.41) por MQ2E agrupado, onde as variáveis instrumentais de $\Delta\log(prisão)$ são duas variáveis binárias, uma registrando se uma decisão final sobre o processo judicial da superlotação foi tomada no ano corrente e outra se nos dois anos anteriores. A estimativa por MQ2E agrupado de α_1 é −1,032 (ep = 0,370). Portanto, o efeito estimado por MQ2E é muito maior; não surpreende que ele seja também muito menos preciso. Levitt encontrou resultados semelhantes quando usou períodos de tempo mais longos (mas sem observações de períodos mais anteriores para alguns estados) e mais variáveis instrumentais.

O teste da existência de correlação serial AR(1) em $r_{it1} = \Delta u_{it1}$ é fácil. Após a estimação por MQ2E agrupado, obtenha os resíduos, $\hat{r}_{it1}$. Depois, inclua uma defasagem desses resíduos na equação original, e estime-a por MQ2E, em que $\hat{r}_{it1}$ age como sua própria variável instrumental. O primeiro ano será perdido em razão da defasagem. Então, a habitual estatística t do MQ2E do resíduo defasado será um teste válido para verificar a existência de correlação serial. No Exemplo 16.8, o coeficiente de $\hat{r}_{it1}$ está em torno de somente 0,076 com $t = 1,67$. Com um coeficiente tão pequeno e tão modesta estatística t, podemos, com segurança, presumir independência serial.

Uma abordagem alternativa de estimar SEMs com dados em painel é usar a transformação de efeitos fixos e depois aplicar uma técnica de VI como o MQ2E agrupado. Um procedimento simples é estimar a equação com centralização na média por MQ2E, que se pareceria com

$$\ddot{y}_{it1} = \alpha_1\ddot{y}_{t2} + \ddot{\mathbf{z}}_{it1}\boldsymbol{\beta}_1 + \ddot{u}_{it1}, t = 1, 2, \ldots, T, \qquad \textbf{(16.42)}$$

em que $\ddot{\mathbf{z}}_{it1}$ e $\ddot{\mathbf{z}}_{it2}$ são VIs. Isso é equivalente a usar o MQ2E na formulação de variáveis *dummy*, em que as variáveis *dummy* específicas-unitárias agem como suas próprias variáveis instrumentais. Ayres e Levitt (1998) aplicaram MQ2E em uma equação com centralização na média para estimar o efeito dos dispositivos eletrônicos Lojack contra roubos de automóveis nas cidades. Se (16.42) for estimada diretamente, então, os *gl* precisarão ser corrigidos para $N(T - 1) - k_1$, em que k_1 será o número total de elementos em α_1 e $\boldsymbol{\beta}_1$. A inclusão de variáveis *dummy* específicas-unitárias e a aplicação do MQ2E agrupado aos dados originais produzirão os *gl* corretos.

RESUMO

Modelos de equações simultâneas são apropriados quando cada equação no sistema tem uma interpretação *ceteris paribus*. Bons exemplos são quando equações separadas descrevem diferentes ângulos de um mercado ou as relações comportamentais de diferentes agentes econômicos. Exemplos de oferta e demanda são os principais casos, mas existem muitas outras aplicações dos SEMs em economia e nas ciências sociais.

Uma característica importante dos SEMs é que, pela completa especificação do sistema, fica claro quais variáveis são presumidas como exógenas e quais delas aparecem em cada equação. Dado um sistema completo, temos condições de determinar quais equações podem ser identificadas (isto é, podem ser estimadas). No importante caso de um sistema com duas equações, é fácil de identificar (declarar): pelo menos uma variável exógena deve ser excluída da primeira equação que apareça com um coeficiente diferente de zero na segunda equação.

Como vimos dos capítulos anteriores, a estimação por MQO de uma equação que contém uma variável explicativa endógena geralmente produz estimadores viesados e inconsistentes. Diferentemente, o MQ2E pode ser usado para estimar qualquer equação identificada em um sistema. Há métodos de sistemas mais avançados disponíveis, mas estão além do escopo deste livro.

A distinção entre variáveis omitidas e simultaneidade nas aplicações nem sempre é nítida. Ambos os problemas, sem falar no erro de medida, podem aparecer na mesma equação. Um bom exemplo é a oferta de mão de obra de mulheres casadas. Anos de escolaridade (*educ*) aparece tanto na função de oferta de mão de obra como na de oferta de salário [veja as equações (16.19) e (16.20)]. Se a aptidão omitida estiver no termo de erro da função de oferta de mão de obra, então, salário e educação serão ambos endógenos. O fator importante *i* é que uma equação estimada por MQ2E pode basear-se em si própria.

Os SEMs também podem ser aplicados a dados de séries temporais. Assim como na estimação por MQO, devemos estar atentos aos processos com tendências e integrados na aplicação do MQ2E. Problemas tais como a correlação serial podem ser tratados como na Seção 15.7. Também apresentamos um exemplo de como estimar um SEM usando dados em painel, onde tiramos a primeira diferença da equação para remover o efeito não observado. Depois, podemos estimar a equação diferenciada por MQ2E agrupado, como no Capítulo 15. Alternativamente, em alguns casos, podemos usar a centralização na média de todas as variáveis, inclusive as VIs e, então, aplicar o MQ2E agrupado; isso é o mesmo que incluir *dummies* de cada observação de corte transversal e usar o MQ2E, onde as *dummies* agem com suas próprias variáveis instrumentais. Aplicações de SEM com dados em painel são muito poderosas, já que nos possibilitam controlar a heterogeneidade não observada ao mesmo tempo em que estamos lidando com a simultaneidade. Elas estão se tornando cada vez mais comuns e não são especialmente difíceis de serem estimadas.

PROBLEMAS

16.1 Escreva um sistema de duas equações na forma de "oferta e demanda", isto é, com a mesma variável y_1 (em geral, "quantidade") aparecendo do lado esquerdo:

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1$$
$$y_1 = \alpha_2 y_2 + \beta_2 z_2 + u_2.$$

(i) Se $\alpha_1 = 0$ ou $\alpha_2 = 0$, explique por que existe uma forma reduzida de y_1. (Lembre-se, a forma reduzida expressa y_1 como uma função linear das variáveis exógenas e dos erros estruturais.) Se $\alpha_1 \neq 0$ e $\alpha_2 = 0$, encontre a forma reduzida de y_2.

(ii) Se $\alpha_1 \neq 0$, $\alpha_2 \neq 0$ e $\alpha_1 \neq \alpha_2$, encontre a forma reduzida de y_1. A variável y_2 tem uma forma reduzida nesse caso?

(iii) A condição $\alpha_1 \neq \alpha_2$ é possível de ser encontrada em exemplos de demanda? Explique.

16.2 Defina *milho* como o consumo *per capita* de milho em toneladas de grãos, no nível de município, *preço* como o preço por tonelada do milho, *renda* como a renda *per capita* do município, e defina *precpluv* como a precipitação pluviométrica em milímetros durante a última safra de plantio de milho. O seguinte modelo de equações simultâneas impõe a condição de equilíbrio em que a oferta se iguala à demanda:

$$milho = \alpha_1 preço + \beta_1 renda + u_1$$
$$milho = \alpha_2 preço + \beta_2 precpluv + \gamma_2 precpluv^2 + u_2.$$

Qual é a equação de oferta e qual é a de demanda? Explique.

16.3 No Problema 3.3 do Capítulo 3, estimamos uma equação para testar uma relação de substituição entre minutos por semana gastos dormindo (*dormir*) e minutos por semana gastos trabalhando (*trabtot*) de uma amostra aleatória de indivíduos. Também incluímos educação e idade na equação. Como *dormir* e *trabtot* são escolhidos conjuntamente por indivíduo, a relação de substituição entre dormir e trabalhar estimada está sujeita a uma crítica de "viés de simultaneidade"? Explique.

16.4 Suponha que os ganhos e o consumo de bebidas alcoólicas anuais sejam determinados pelo SEM

$$\log(ganhos) + \beta_0 + \beta_1 álcool + \beta_2 educ + u_1$$
$$álcool = \gamma_0 + \gamma_1 \log(ganhos) + \gamma_2 educ + \gamma_3 \log(preço) + u_2,$$

em que *preço* é o índice local de preços do álcool, que inclui impostos locais e estaduais. Suponha que *educ* e *preço* sejam exógenos. Se β_1, β_2, γ_1, γ_2 e γ_3 forem todas diferentes de zero, qual equação será identificada? Como você estimaria essa equação?

16.5 Um modelo simples para determinar a eficácia do uso da camisinha na redução das doenças sexualmente transmissíveis entre alunos do ensino médio sexualmente ativos é

$$taxainf = \beta_0 + \beta_1 usacamis + \beta_2 percmasc + \beta_3 rendfam + \beta_4 cidade + u_1,$$

em que *taxainf* é a porcentagem de alunos sexualmente ativos que tenham contraído doença venérea, *usacamis* é a porcentagem de rapazes que afirmam usar camisinha regularmente, *rendfam* é a renda familiar média e *cidade* é uma variável *dummy* indicando se a escola está em uma cidade; o modelo é construído no âmbito escolar.

(i) Interpretando a equação precedente de uma maneira causal, *ceteris paribus*, qual deverá ser o sinal de β_1?

(ii) Por que *taxainf* e *usacamis* podem ser conjuntamente determinadas?

(iii) Se o uso de camisinha aumentar com a taxa de doenças venéreas, de forma que $\gamma_1 > 0$ na equação
$usacamis = \gamma_0 + \gamma_1 taxainf + outros\ fatores$,
qual será o provável viés na estimativa de β_1 por MQO?

(iv) Defina *distcamis* como uma variável binária igual a um se uma escola tiver um programa de distribuição de camisinhas. Explique como isso pode ser usado para estimar β_1 (e os outros betas) por VIs. O que teremos de presumir sobre *distcamis* em cada equação?

16.6 Considere um modelo de probabilidade linear explicando se os empregadores oferecem um plano de pensão com base na porcentagem de trabalhadores que pertençam a um sindicato, bem como outros fatores:

$$pensão = \beta_0 + \beta_1 pertsind + \beta_2 idademed + \beta_3 educmed + \beta_4 percmasc + \beta_5 perccasad + u_1.$$

(i) Por que *pertsind* pode ser determinado conjuntamente com *pensão*?

(ii) Suponha que você possa pesquisar os trabalhadores nas firmas e colher informações sobre suas famílias. Você consegue pensar em uma informação que poderia ser usada para construir uma VI de *pertsind*?

(iii) Como você verificaria se sua variável é pelo menos uma candidata razoável a VI de *pertsind*?

16.7 Suponha que você seja solicitado a estimar a demanda por ingressos de jogos de basquete feminino de uma grande universidade. Você pode coletar dados de séries temporais de mais de dez temporadas, com um total de cerca de 150 observações. Um modelo possível seria

$$lPÚBLICO_t = \beta_0 + \beta_1 lPREÇO_t + \beta_2 PERCVIT_t + \beta_3 RIVAL_t + \beta_4 FIMSEMANA_t + \beta_5 t + u_1,$$

em que $PREÇO_t$ é o preço do ingresso, provavelmente indicado em termos reais — digamos, deflacionado por um índice local de preços ao consumidor x; $PERCVIT_t$ é a porcentagem de atuais vitórias da equipe x; $RIVAL_t$ é uma variável *dummy* indicando um jogo contra um rival x; e $FIMSEMANA_t$ é uma variável *dummy* indicando se o jogo é realizado durante o fim de semana.

O l representa o logaritmo natural, de forma que a função de demanda tem uma elasticidade-preço constante.

(i) Por que é uma boa ideia ter uma tendência temporal na equação?

(ii) A oferta de ingressos é fixada pela capacidade do estádio; suponha que ela não tenha mudado nos últimos dez anos. Isso significa que a quantidade oferecida não varia com o preço. Significa que o preço será necessariamente exógeno na equação de demanda? (*Sugestão*: A resposta é não.)

(iii) Suponha que o preço nominal do ingresso se altere lentamente — digamos, no início de cada temporada. O departamento esportivo determina os preços baseando-se parcialmente no público da temporada anterior, como também no sucesso obtido pela equipe na temporada anterior. Sob que hipóteses a porcentagem de vitórias na temporada anterior ($PERVITTEMP_{t-1}$) será uma variável instrumental válida de $lPREÇO_t$?

(iv) Parece razoável incluir o (log do) preço real dos ingressos dos jogos de basquetebol masculino na equação? Explique. Que indício a teoria econômica prevê para seu coeficiente? Você con-

segue pensar em outra variável relacionada ao basquetebol masculino que possa pertencer à equação do público nos jogos femininos?

(v) Se você está preocupado com que algumas das séries, particularmente *lPÚBLICO* e *lPREÇO*, tenham raízes unitárias, como você poderia alterar a equação estimada?

(vi) Se alguns jogos tiverem suas lotações esgotadas, que problemas isso causará para a estimativa da função de demanda? (*Sugestão*: Se um jogo tiver sua lotação esgotada, você necessariamente observará a demanda real?)

16.8 O quanto é grande o efeito dos gastos escolares por aluno sobre os preços de habitação local? Defina *PREÇOC* como sendo a mediana dos preços de habitação em um distrito escolar e defina *GASTO* como os gastos por aluno. Usando dados em painel dos anos de 1992, 1994 e 1996, postulamos o modelo

$$lPREÇOC_{it} = \theta_t + \beta_1 lGASTO_{it} + \beta_2 lPOLÍCIA_{it} + \beta_3 lMEDREND_{it} + \beta_4 lIMPPRO_{it} + a_{i1} + u_{it1},$$

em que $POLÍCIA_{it}$ são os gastos policiais *per capita*, $MEDREND_{it}$ é a mediana da renda e $IMPPRO_{it}$ é a alíquota do imposto sobre a propriedade; *l* denota logaritmo natural. Gastos e preços de habitação são determinados simultaneamente porque o valor dos imóveis afeta diretamente as receitas disponíveis para financiar as escolas.

Suponha que, em 1994, a maneira pela qual as escolas eram financiadas tenha mudado drasticamente: em vez de serem financiadas pelos impostos locais sobre a propriedade, os financiamentos das escolas tenham sido determinados basicamente em nível estadual. Defina $lALEST_{it}$ como o log da alocação da verba estadual ao distrito *i* no ano *t*, que é exógeno na equação precedente, uma vez que controlemos gastos e o efeito fixo de um distrito. Como você estimaria os β_j?