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

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

9

Problemas Adicionais de Especificação e de Dados

No Capítulo 8 estudamos uma violação das hipóteses de Gauss-Markov. Enquanto a heteroscedasticidade nos erros pode ser vista como uma má-especificação de modelo, ela é um problema relativamente de menor importância. A presença de heteroscedasticidade não causa viés ou inconsistência nos estimadores MQO. Além disso, é razoavelmente fácil ajustar intervalos de confiança e estatísticas t e F para obter inferência válida após a estimação MQO, ou mesmo para obter estimadores mais eficientes com o uso de mínimos quadrados ponderados.

Neste capítulo, retornamos a um problema muito mais sério da correlação entre o erro, u, e uma ou mais das variáveis explicativas. Lembre-se do Capítulo 3 em que, se u for, por qualquer razão, correlacionado com a variável explicativa x_j, então dizemos que x_j é uma **variável explicativa endógena**. Também fazemos uma discussão mais detalhada sobre três razões pelas quais uma variável explicativa pode ser endógena; em alguns casos discutimos possíveis correções.

Já vimos nos Capítulos 3 e 5 que a omissão de uma variável importante pode causar correlação entre o erro e algumas das variáveis explicativas, o que geralmente conduz a viés e inconsistência em *todos* os estimadores MQO. No caso especial em que a variável omitida é uma função de uma variável explicativa no modelo, este sofrerá de **má-especificação da forma funcional**.

Iniciamos a primeira seção discutindo as consequências da má-especificação da forma funcional e como testar sua existência. Na Seção 9.2, mostramos como o uso de variáveis *proxy* pode resolver, ou pelo menos aliviar, o viés de variáveis omitidas. Na Seção 9.3, derivamos e explicamos o viés no método MQO que pode aparecer sob certas formas de **erro de medida**. Problemas adicionais de dados são discutidos na Seção 9.4.

Todos os procedimentos descritos neste capítulo são baseados na estimação MQO. Como veremos, certos problemas que causam correlação entre o erro e algumas variáveis explicativas não podem ser resolvidos usando MQO em estudos de corte transversal. Postergamos uma abordagem sobre métodos de estimação alternativos para a Parte 3.

9.1 MÁ-ESPECIFICAÇÃO DA FORMA FUNCIONAL

Um modelo de regressão múltipla sofre de má-especificação da forma funcional quando não explica de maneira apropriada a relação entre as variáveis explicativas e a dependente observadas. Por exemplo, se o salário por hora é determinado por $\log(salário\text{h}) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u$, mas omitimos o termo de experiência elevado ao quadrado, $exper^2$, então estamos cometendo uma má-especificação da forma funcional. Já sabemos, como vimos no Capítulo 3, que isso geralmente conduz a estimadores viesados de β_0, β_1 e β_2. (Não estimamos β_3 porque o termo $exper^2$ foi excluído do modelo). Desse modo, especificando incorretamente como $exper$ afeta $\log(salário\text{h})$ geralmente resulta em um

estimador viesado do retorno da educação, β_1. A magnitude desse viés depende do tamanho de β_3 e da correlação entre $educ$, $exper$ e $exper^2$.

A situação é pior ao estimar o retorno da experiência: mesmo que pudéssemos conseguir um estimador não viesado de β_2, não seríamos capazes de estimar o retorno da experiência, pois ela é equivalente a $\beta_2 + 2\beta_3 exper$ (em forma decimal). Usar apenas o estimador viesado de β_2 pode ser enganoso, especialmente nos valores extremos de $exper$.

Como outro exemplo, suponha que a equação $\log(salário\text{h})$ seja

$$\log(salário\text{h}) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 feminino + \beta_5 feminino \cdot educ + u, \tag{9.1}$$

em que $feminino$ é uma variável binária. Se omitirmos o termo de interação, $feminino \cdot educ$, estaremos especificando mal a forma funcional. De maneira geral, não obteremos estimadores não viesados de nenhum dos outros parâmetros, e como o retorno da educação depende do gênero, não fica claro que tipo de retorno estaríamos estimando quando omitimos o termo de interação.

A omissão de funções de variáveis independentes não é a única maneira de um modelo sofrer o problema da má-especificação da forma funcional. Por exemplo, se (9.1) for o modelo verdadeiro para satisfazer as primeiras quatro hipóteses de Gauss-Markov, mas utilizarmos $salário\text{h}$, em lugar de $\log(salário\text{h})$, como variável dependente, não obteremos estimadores não viesados ou consistentes dos efeitos parciais. Os testes a seguir têm certa capacidade de detectar esse tipo de problema da forma funcional, mas existem testes melhores que serão mencionados nas subseções de testes contra alternativas não aninhadas.

A má-especificação da forma funcional de um modelo pode, certamente, trazer sérias consequências. No entanto, em um aspecto importante, o problema é secundário: por definição, temos dados de todas as variáveis necessárias para obter uma relação funcional que se ajuste bem aos dados. Isso pode ser contrastado com o problema tratado na próxima seção, na qual uma variável importante será omitida e sobre a qual não poderemos coletar dados.

Já temos uma ferramenta poderosa para detectar uma forma funcional mal-especificada: o teste F para restrições de exclusões conjuntas. Muitas vezes faz sentido adicionar termos quadráticos de quaisquer variáveis significantes a um modelo e executar um teste conjunto de significância. Se os termos quadráticos adicionados forem significantes, eles podem ser adicionados ao modelo (ao custo de complicar sua interpretação). Porém, termos quadráticos significantes podem ser sintomáticos de outros problemas de formas funcionais, como, por exemplo, usar uma variável em nível quando o logaritmo é mais apropriado, ou vice-versa. Pode ser difícil localizar a razão exata pela qual uma forma funcional está mal-especificada. Felizmente, em muitos casos, o uso de logaritmos de certas variáveis e a adição de termos quadráticos são suficientes para detectar muitas relações não lineares importantes em economia.

EXEMPLO 9.1

(Modelo Econômico do Crime)

A Tabela 9.1 contém estimativas MQO do modelo econômico do crime (veja Exemplo 8.3). Primeiro estimamos o modelo sem nenhum termo quadrático; os resultados estão na coluna (1).

EXEMPLO 9.1 (continuação)

Tabela 9.1
Variável dependente: *npre86*.

Variáveis independentes	(1)	(2)
pcond	−0,133 (0,040)	0,533 (0,154)
$pcond^2$	—	−0,730 (0,156)
sentmed	−0,011 (0,012)	−0,017 (0,012)
temptot	0,012 (0,009)	0,012 (0,009)
ptemp86	−0,041 (0,009)	0,287 (0,004)
$ptemp86^2$	—	−0,0296 (0,0039)
empr86	−0,051 (0,014)	−0,014 (0,017)
rend86	−0,0015 (0,0003)	−0,0034 (0,0008)
$rend86^2$	—	0,000007 (0,000003)
negro	0,327 (0,045)	0,292 (0,045)
hispan	0,194 (0,040)	0,164 (0,039)
intercepto	0,596 (0,036)	0,505 (0,037)
Observações *R*-quadrado	2.725 0,0723	2.725 0,1035

Na coluna (2) os quadrados de *pcond*, *ptemp86* e *rend86* foram adicionados; decidimos incluir os quadrados dessas variáveis porque cada termo em nível é significante na coluna (1). A variável *empr86* é uma variável discreta, ao presumirmos somente cinco valores, de modo que não incluímos seu quadrado na coluna (2).

EXEMPLO 9.1 (continuação)

Cada um dos termos quadráticos é significante e juntos são simultaneamente muito significantes ($F = 31,37$, com $gl = 3$ e 2.713; o *p*-valor é basicamente zero). Portanto, parece que o modelo inicial deixou de fora algumas não linearidades potencialmente importantes.

A presença dos termos quadráticos faz com que a interpretação do modelo seja um pouco difícil. Por exemplo, *pcond* não tem mais um efeito estritamente de dissuasão: a relação entre *npre86* e *pcond* é positiva até *pcond* = 0,365 e, a partir daí, a relação é negativa. Podemos concluir que existe pouco ou nenhum efeito de dissuasão em valores mais baixos de *pcond*; o efeito somente aparece com taxas de condenações anteriores mais altas. Teríamos que utilizar formas funcionais mais sofisticadas do que a quadrática para verificar essa conclusão. Pode ser que *pcond* não seja inteiramente exógena. Por exemplo, pessoas que não tenham sido condenadas no passado (de modo que *pcond* = 0) são, talvez, criminosos casuais, e, portanto, é menos provável que tenham sido presos em 1986. Isto poderia estar causando um viés nas estimativas.

Similarmente, a relação entre *npre86* e *ptemp86* é positiva até *ptemp86* = 4,85 (quase cinco meses na prisão), e a partir daí é negativa. A maioria das pessoas na amostra não passou tempo algum na prisão em 1986, de modo que, novamente, devemos ser cuidadosos ao interpretar os resultados.

A renda legal tem um efeito negativo sobre *npre86* até *rend86* = 242,85; como a renda é medida em centenas de dólares, isso representa uma renda anual de US$24.285. Somente 46 das pessoas na amostra têm rendimentos acima desse nível. Portanto, podemos concluir que *npre86* e *rend86* são negativamente relacionadas, com um efeito decrescente.

QUESTÃO 9.1

Por que não incluímos os quadrados de *negro* e *hispan* na coluna (2) da Tabela 9.1?

O exemplo 9.1 é um problema delicado de forma funcional, em razão da natureza da variável dependente. Outros modelos são, teoricamente, mais apropriados para se manipular variáveis dependentes considerando um número pequeno de valores inteiros. Discutiremos resumidamente esses modelos no Capítulo 17.

O Teste RESET como um Teste Geral da Má-Especificação da Forma Funcional

Alguns testes têm sido propostos para detectar a má-especificação da forma funcional. O **teste de erro de especificação da regressão** (**RESET**) de Ramsey (1969) tem se mostrado útil a esse respeito.

A ideia por trás do teste RESET é bastante simples. Se o modelo original

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u \qquad (9.2)$$

satisfizer RLM.4, nenhuma função não linear das variáveis independentes deve ser significante quando adicionada à equação (9.2). No Exemplo 9.1 adicionamos termos quadráticos às variáveis explicativas significantes. Embora isso muitas vezes detecte problemas de forma funcional, tem a desvantagem de gastar muitos graus de liberdade se houver muitas variáveis explicativas no modelo original (tanto quanto a forma direta do teste de White da heteroscedasticidade consome graus de liberdade). Além disso, certos tipos de não linearidades negligenciadas não serão detectadas pela adição de termos qua-

dráticos. O teste RESET adiciona polinômios aos valores estimados MQO na equação (9.2) para detectar tipos gerais de má-especificação de formas funcionais.

Para implementar o teste RESET, temos que decidir quantas funções dos valores estimados devem ser incluídas na regressão expandida. Não existe uma resposta certa para esta questão, mas os termos quadráticos e cúbicos têm demonstrado utilidade na maior parte das aplicações.

Sejam $\hat{y}$ os valores estimados MQO ao estimar (9.2). Considere a equação expandida

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + erro. \qquad (9.3)$$

Esta equação parece um tanto estranha, pois funções dos valores estimados na estimação inicial agora aparecem como variáveis explicativas. Na realidade, não estaremos interessados nos parâmetros estimados em (9.3); apenas usamos esta equação para testar se (9.2) tem não linearidades importantes ausentes. O que devemos lembrar é que $\hat{y}^2$ e $\hat{y}^3$ são apenas funções não lineares de x_j.

A hipótese nula é que (9.2) está corretamente especificada. Portanto, a estatística do teste RESET é a estatística F para testar H_0: $\delta_1 = 0, \delta_2 = 0$ no modelo expandido (9.3). Uma estatística F significante sugere algum tipo de problema na forma funcional. A distribuição da estatística F é, aproximadamente, $F_{2,n-k-3}$ em amostras grandes sob a hipótese nula (e sob as hipóteses de Gauss-Markov). Os *gl* na equação expandida (9.3) são $n - k - 1 - 2 = n - k - 3$. Uma versão *LM* também está disponível (e a distribuição qui-quadrado terá dois *gl*). Além disso, o teste pode ser transformado em robusto em relação à heteroscedasticidade utilizando os métodos discutidos na Seção 8.2.

EXEMPLO 9.2

(Equação dos Preços de Imóveis)

Estimamos dois modelos para os preços de imóveis. O primeiro tem todas as variáveis em forma de nível:

$$preço = \beta_0 + \beta_1 tamterr + \beta_2 arquad + \beta_3 qtdorm + u. \qquad (9.4)$$

O segundo utiliza os logaritmos de todas as variáveis, exceto *qtdorm*:

$$lpreço = \beta_0 + \beta_1 ltamterr + \beta_2 larquad + \beta_3 qtdorm + u. \qquad (9.5)$$

Usando $n = 88$ imóveis do arquivo HPRICE.RAW, constata-se que a estatística do teste RESET da equação (9.4) é 4,67; este é o valor de uma variável aleatória $F_{2,82}$ ($n = 88$, $k = 3$), e o p-valor associado é 0,012. Isto é uma evidência de má-especificação da forma funcional em (9.4).

A estatística do teste RESET em (9.5) é 2,56, com p-valor = 0,084. Portanto, não rejeitamos (9.5) no nível de significância de 5% (embora o faríamos ao nível de 10%). Com base no teste RESET, o modelo log-log em (9.5) é preferido.

No exemplo anterior, tentamos dois modelos para explicar os preços de imóveis. Um foi rejeitado pelo teste RESET, ao passo que o outro não o foi (pelo menos ao nível de 5%). Muitas vezes, as coisas

não são tão simples. Uma desvantagem do teste RESET é que ele não fornece uma orientação prática de como proceder se o modelo for rejeitado. A rejeição de (9.4), pelo uso do teste RESET não sugere diretamente que (9.5) seja o passo seguinte. A equação (9.5) foi estimada porque modelos de elasticidade constante são fáceis de serem interpretados e podem apresentar boas propriedades estatísticas. Neste exemplo, o modelo também passa no teste da forma funcional.

Algumas pessoas argumentaram que o teste RESET é demasiadamente generalizado da má-especificação de modelos, incluindo variáveis omitidas não observadas e heteroscedasticidade. Infelizmente, tal uso do teste é bastante equivocado. Pode ser demonstrado que o teste RESET não tem poder para detectar variáveis omitidas sempre que houver expectativa de que elas sejam lineares nas variáveis independentes incluídas no modelo [veja Wooldridge (1995) para um enunciado preciso]. Além disso, se a forma funcional for apropriadamente especificada, o teste RESET não tem poder para detectar heteroscedasticidade. O ponto principal é que o teste RESET é um teste da forma funcional, e nada mais que isso.

Testes contra Alternativas não Aninhadas

Obter testes para outros tipos de má-especificação da forma funcional — por exemplo, tentar decidir se uma variável independente deveria aparecer em nível ou em forma logarítmica — nos leva para fora do âmbito dos testes de hipótese clássicos. É possível testar o modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \qquad (9.6)$$

contra o modelo

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 log(x_2) + u, \qquad (9.7)$$

e vice-versa. Porém, esses são **modelos não aninhados** (veja Capítulo 6), e portanto não podemos simplesmente usar um teste F padrão. Dois métodos diferentes podem ser sugeridos. O primeiro é construir um modelo abrangente que contenha cada modelo como um caso especial e, em seguida, testar as restrições que conduziram a cada um dos modelos. No exemplo, o modelo abrangente é

$$y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 \log(x_1) + \gamma_4 log(x_2) + u. \qquad (9.8)$$

Podemos primeiro testar H_0: $\gamma_3 = 0, \gamma_4 = 0$, como um teste de (9.6). Podemos também testar H_0: $\gamma_1 = 0, \gamma_2 = 0$ como um teste de (9.7). Esta abordagem foi sugerida por Mizon e Richard (1986).

Outro método foi sugerido por Davidson e MacKinnon (1981). Eles salientam que, se (9.6) for verdadeira, então os valores estimados do modelo (9.7) deveriam ser não significantes em (9.6). Assim, para testar (9.6), primeiro estimamos o modelo (9.7) por MQO para obtermos os valores estimados. Chamemos esses valores de $\hat{\hat{y}}$. Então, o **teste de Davidson-MacKinnon** baseia-se na estatística t sobre $\hat{\hat{y}}$ na equação

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \theta_1 \hat{\hat{y}} + erro.$$

Uma estatística t significante (contra uma alternativa bicaudal) é uma rejeição de (9.6).

Similarmente, se $\hat{y}$ representar os valores estimados da estimação de (9.6), o teste de (9.7) é a estatística t sobre $\hat{y}$ no modelo

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + \theta_1 \hat{y} + erro;$$

uma estatística t significante é evidência contra (9.7). Os mesmos dois testes podem ser usados para testar quaisquer dois modelos não aninhados com a mesma variável dependente.

Existem alguns problemas com testes não aninhados. Primeiro, não necessariamente, um dos modelos será claramente o escolhido. Ambos os modelos, ou nenhum deles, podem ser rejeitados. Se nenhum deles for rejeitado, podemos usar o R-quadrado ajustado para selecionar um deles. Se ambos os modelos forem rejeitados, teremos mais trabalho. Porém, é importante sabermos das consequências práticas ao utilizarmos cada um deles: se os efeitos de importantes variáveis independentes sobre y não forem muito diferentes, então não importa qual dos modelos será usado.

Um segundo problema é que a rejeição de (9.6) pela utilização, digamos, do teste de Davidson-MacKinnon, não significa que (9.7) seja o modelo correto. Esse modelo (9.6) pode ser rejeitado por uma diversidade de más-especificações da forma funcional.

Um problema ainda mais difícil é obter testes não aninhados quando os modelos concorrentes têm variáveis dependentes diferentes. O caso principal é y *versus* $\log(y)$. Vimos no Capítulo 6 que apenas a obtenção de medidas de qualidades de ajustes que possam ser comparadas necessita de algum cuidado. Alguns testes foram propostos para resolver este problema, mas estão além do escopo deste texto. [Veja Wooldridge (1994a) para um teste que tem uma interpretação simples e é fácil de ser implementado.]

9.2 UTILIZANDO VARIÁVEIS *PROXY* PARA VARIÁVEIS EXPLICATIVAS NÃO OBSERVADAS

Um problema mais difícil surge quando um modelo exclui uma variável importante, normalmente em razão da não disponibilidade de dados. Considere uma equação de salários que explicitamente reconheça que a aptidão (*aptid*) afeta log(*salário*h):

$$log(salário) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 aptid + u. \qquad (9.9)$$

Este modelo mostra explicitamente que queremos manter fixa a aptidão quando medimos os retornos de *educ* e de *exper*. Se, digamos, *educ* for correlacionada com *aptid*, colocar *aptid* no termo de erro fará com que o estimador MQO de β_1 (e β_2) sejam viesados, um tema que tem aparecido repetidamente.

Nosso interesse primordial na equação (9.9) está nos parâmetros de inclinação β_1 e β_2. Realmente, não nos interessa se obteremos um estimador não viesado ou consistente do intercepto β_0; como veremos em breve, normalmente isso não é possível. Tampouco podemos esperar estimar β_3, pois *aptid* não é observada; na verdade, de qualquer forma, não saberíamos como interpretar β_3, pois a aptidão é, na melhor das hipóteses, um conceito vago.

Como podemos resolver, ou pelo menos aliviar, o problema do viés de variáveis omitidas em uma equação como (9.9)? Uma possibilidade é obter uma **variável *proxy*** da variável omitida. *Genericamente falando*, uma variável *proxy* é algo que está relacionado com a variável não observada que gostaríamos de controlar em nossa análise. Na equação do salário, uma possibilidade é usar o quociente de inteligência, ou QI, como uma *proxy* da aptidão. Isso *não* exige que QI seja a mesma coisa que aptidão; o que precisamos é que QI seja correlacionado com aptidão, o que esclareceremos na discussão a seguir.

Todas essas ideias podem ser ilustradas em um modelo com três variáveis independentes, duas das quais são observadas:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u. \qquad (9.10)$$

Presumimos que os dados estão disponíveis para y, x_1 e x_2 — no exemplo do salário, eles são log(*salário*h), *educ* e *exper*, respectivamente. A variável explicativa x_3^* é não observada, mas temos uma variável *proxy* de x_3^*. Chamemos essa variável *proxy* de x_3.

O que necessitamos de x_3? No mínimo, ela deve ter alguma relação com x_3^*. Isso é capturado pela equação de regressão simples

$$x_3^* = \delta_0 + \delta_3 x_3 + v_3, \qquad (9.11)$$

em que v_3 é um erro pelo fato de que x_3^* e x_3 não são exatamente relacionadas. O parâmetro δ_3 mede a relação entre x_3^* e x_3; em geral, pensamos em x_3^* e x_3 como positivamente relacionadas, de forma que $\delta_3 > 0$. Se $\delta_3 = 0$, então x_3 não é uma *proxy* adequada de x_3^*. O intercepto δ_0 em (9.11), que pode ser positivo ou negativo, simplesmente permite que x_3^* e x_3 sejam medidas em diferentes escalas. (Por exemplo, é evidente não ser necessário que a aptidão não observada tenha o mesmo valor médio do QI na população dos Estados Unidos.)

Como podemos usar x_3 para obtermos estimadores não viesados (ou pelo menos consistentes) de β_1 e β_2? A proposta é simular que x_3 e x_3^* sejam as mesmas, de forma que possamos calcular a regressão de

$$y \text{ sobre } x_1, x_2, x_3. \qquad (9.12)$$

Chamamos a isso de **solução plugada do problema de variáveis omitidas** porque a variável x_3 está "plugada" em x_3^* antes de executarmos o MQO. Se x_3 for verdadeiramente relacionada com x_3^*, isso parece sensato. Porém, como x_3 e x_3^* não são as mesmas, devemos determinar quando esse procedimento produzirá, de fato, estimadores consistentes de β_1 e β_2.

As hipóteses necessárias para que a solução plugada forneça estimadores consistentes de β_1 e β_2 podem ser decompostas em hipóteses sobre u e v_3:

1) O erro u é não correlacionado com x_1, x_2 e x_3^*, que justamente é a hipótese-padrão no modelo (9.10). Além disso, u é não correlacionado com x_3. Esta última hipótese significa exatamente que x_3 é irrelevante no modelo populacional, já que as variáveis x_1, x_2 e x_3^* foram incluídas. Isso é basicamente verdadeiro por definição, visto que x_3 é uma variável *proxy* de x_3^*: é x_3^* que diretamente afeta y, não x_3. Assim, a hipótese de que u é não correlacionada com x_1, x_2, x_3^* e x_3 não é muito controversa. (Outra maneira de explicar essa hipótese é que o valor esperado de u, dadas todas essas variáveis, é zero.)
2) O erro v_3 é não correlacionado com x_1, x_2 e x_3. Supor que v_3 é não correlacionado com x_1 e x_2 exige que x_3 seja uma "boa" *proxy* de x_3^*. Pode-se ver isso de maneira mais fácil, escrevendo-se a expressão análoga dessas hipóteses em termos de expectativas condicionais:

$$E(x_3^* | x_1, x_2, x_3) = E(x_3^* | x_3) = \delta_0 + \delta_3 x_3. \qquad (9.13)$$

A primeira igualdade, que é a mais importante, diz que, como x_3 é controlada, o valor esperado de x_3^* não depende de x_1 ou de x_2. Alternativamente, x_3^* tem correlação zero com x_1 e com x_2, dado que x_3 é parcializada.

Na equação de salários-hora (9.9), em que QI é a *proxy* da aptidão, a condição (9.13) torna-se

$$\mathrm{E}(aptid|educ,exper,QI) = \mathrm{E}(aptid|QI) = \delta_0 + \delta_3 QI.$$

Assim, a média de aptidão somente muda com *QI*, não com *educ* e *exper*. Isso é razoável? Talvez não seja exatamente verdade, mas está perto de ser. Certamente, vale a pena incluir *QI* na equação de salários para vermos o que acontece com o retorno estimado da variável educação.

Podemos facilmente verificar a razão de as hipóteses anteriores serem suficientes para que a solução plugada funcione. Se integrarmos a equação (9.11) na equação (9.10) e aplicarmos álgebra simples teremos

$$y = (\beta_0 + \beta_3\delta_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3\delta_3 x_3 + u + \beta_3 v_3.$$

Chamemos o erro composto nesta equação de $e = u + \beta_3 v_3$; ele depende do erro no modelo de interesse (9.10), e do erro na equação da variável *proxy*, v_3. Como tanto u quanto v_3 têm média zero e ambos são não correlacionados com x_1, x_2 e x_3, o erro e também tem média zero e é não correlacionado com x_1, x_2 e x_3. Escreva esta equação como

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha_3 x_3 + e,$$

em que $\alpha_0 = (\beta_0 + \beta_3\delta_0)$ é o novo intercepto e $\alpha_3 = \beta_3\delta_3$ é o parâmetro de inclinação da variável *proxy* x_3. Como mencionamos anteriormente, quando calculamos a regressão em (9.12), não obteremos estimadores não viesados de β_0 e β_3; em vez disso, obteremos estimadores não viesados (ou pelo menos consistentes) de α_0, β_1, β_2 e α_3. O importante é que obtenhamos boas estimativas dos parâmetros β_1 e β_2.

Em muitos casos, a estimativa de α_3 é efetivamente mais interessante do que uma estimativa de β_3. Por exemplo, na equação de salários, α_3 mede o retorno do salário se um ou mais pontos forem atribuídos à pontuação do QI.

EXEMPLO 9.3

(QI como *Proxy* de Aptidão)

O arquivo WAGE2.RAW, de Blackburn e Neumark (1992), contém informações sobre renda mensal, educação, diversas variáveis demográficas e pontuação de QI para 935 homens, em 1980. Como um método para explicar o viés da variável omitida aptidão, adicionamos *QI* a uma equação de log salário padrão. Os resultados estão mostrados na Tabela 9.2.

Nosso principal interesse está no que acontece com o retorno estimado da educação. A coluna (1) contém as estimativas sem a utilização de *QI* como uma variável *proxy*. O retorno estimado da educação é de 6,5%. Se imaginarmos que a aptidão omitida é positivamente correlacionada com *educ*, admitimos que esta estimativa é alta demais. (Mais precisamente, as estimativas médias de todas as amostras aleatórias seriam altas demais.) Quando *QI* é adicionada à equação, o retorno da educação cai para 5,4%, o que corresponde a nossa opinião anterior sobre o viés de omitir a variável aptidão.

EXEMPLO 9.3 (continuação)

Tabela 9.2

Variável dependente: log(*salário*h).

Variáveis independentes	(1)	(2)	(3)
educ	0,065 (0,006)	0,054 (0,007)	0,018 (0,041)
exper	0,014 (0,003)	0,014 (0,003)	0,014 (0,003)
perm	0,012 (0,002)	0,011 (0,002)	0,011 (0,002)
casado	0,199 (0,039)	0,200 (0,039)	0,201 (0,039)
sulista	−0,091 (0,026)	−0,080 (0,026)	−0,080 (0,026)
urbano	0,184 (0,027)	0,182 (0,027)	0,184 (0,027)
negro	−0,188 (0,038)	−0,143 (0,039)	−0,147 (0,040)
QI	—	−0,0036 (0,0010)	−0,0009 (0,0052)
educ · *QI*	—	—	0,00034 (0,00038)
intercepto	5,395 (0,113)	5,176 (0,128)	5,648 (0,546)
Observações *R*-quadrado	935 0,253	935 0,263	935 0,263

O efeito do QI nos resultados socioeconômicos foi recentemente documentado no controverso livro *The Bell Curve* (*A curva normal*), de Herrnstein e Murray (1994). A coluna (2) mostra que a variável QI tem um efeito estatisticamente significante e positivo sobre a renda, após vários outros fatores terem sido controlados. Todos os outros fatores permanecendo inalterados, um aumento de dez pontos no QI aumenta a renda em 3,6%. O desvio-padrão do QI na população dos Estados Unidos é 15, de modo que um aumento de desvio-padrão no QI está associado a uma elevação na renda de 5,4%. Essa elevação é idêntica ao aumento previsto em salário motivado por um ano a mais de educação. Fica claro, a partir da coluna (2),

EXEMPLO 9.3 (continuação)

que a educação ainda tem um papel importante no aumento da renda, embora o efeito não seja tão grande quanto inicialmente se estimava.

Outras observações interessantes surgem do exame das colunas (1) e (2). Adicionar a variável QI à equação somente aumenta o *R*-quadrado de 0,253 para 0,263. A maior parte da variação em log(*salário*h) não é explicada pelos fatores da coluna (2). Além disso, a adição de QI à equação não elimina a diferença da renda estimada entre negros e brancos: estima-se que um negro, com o mesmo QI, educação, experiência etc. de um branco ganhe cerca de 14,3% a menos, e essa diferença é estatisticamente bastante significante.

A coluna (3) na Tabela 9.2 inclui o termo de interação *educ*· *QI*. Ele possibilita que *educ* e *aptid* interajam na determinação de log(*salário*h). Podemos pensar que o retorno da educação seja mais alto para pessoas com mais aptidão, mas este acaba não sendo o caso: o termo de interação não é significante e sua adição torna *educ* e *QI* individualmente não significantes, além de complicar o modelo. Portanto, as estimativas da coluna (2) são preferidas.

Não existe razão para usarmos somente uma variável *proxy* da aptidão neste exemplo. O conjunto de dados do arquivo WAGE2.RAW também contém registros da pontuação de cada pessoa no teste KWW — *Knowledge of the World of Work* (Conhecimento do Mundo do Trabalho)*. Essa pontuação produz uma medida diferente da aptidão, que possa ser usada isoladamente ou em conjunto com o QI para estimar o retorno da educação (veja o Exercício em Computador 9.2, no site da Cengage).

* Pesquisa patrocinada e dirigida pelo *Bureau of Labor Statistics* (Agência de Estatísticas do Trabalho) do *U.S. Department of Labor* (Departamento de Trabalho dos Estados Unidos) (NRT).

QUESTÃO 9.2

O que é possível concluir do pequeno e estatisticamente não significante coeficiente da variável *educ* na coluna (3) da Tabela 9.2? (*Sugestão*: Quando *educ*· *QI* está na equação, qual é a interpretação do coeficiente de *educ*?)

É fácil observar como o uso de uma variável *proxy* ainda pode conduzir a viés, se ela não satisfizer as hipóteses precedentes. Suponha que, em lugar de (9.11), a variável não observada, x_3^*, seja relacionada com todas as variáveis observadas por

$$x_3^* = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + v_3, \qquad (9.14)$$

em que v_3 tem média zero e é não correlacionada com x_1, x_2 e x_3. A equação (9.11) presume que δ_1 e δ_2 são ambos zero. Plugando a equação (9.14) à (9.10), obtemos

$$y = (\beta_0 + \beta_3\delta_0) + (\beta_1 + \beta_3\delta_1)x_1 + (\beta_2 + \beta_3\delta_2)x_2 + \beta_3\delta_3 x_3 + u + \beta_3 v_3, \qquad (9.15)$$

da qual segue que $\text{plim}(\hat{\beta}_1) = \beta_1 + \beta_3\delta_1$ e $\text{plim}(\hat{\beta}_2) = \beta_2 + \beta_3\delta_2$. [Isso acontece porque o erro em (9.15), $u + \beta_3 v_3$, tem média zero e é não correlacionado com x_1, x_2 e x_3.] No exemplo anterior, em que $x_1 = educ$ e $x_3^* = aptid$, $\beta_3 > 0$, de modo que existe um viés positivo (inconsistência), se *aptid* tem

uma correlação parcial com *educ* ($\delta_1 > 0$). Desse modo, ainda poderíamos continuar obtendo um viés para cima no retorno da educação, utilizando *QI* como *proxy* de *aptid*, se *QI* não for uma boa *proxy*. Porém, podemos ter alguma esperança de que esse viés será menor do que se ignorarmos totalmente o problema da aptidão omitida.

Variáveis *proxy* também podem aparecer na forma de informação binária. No exemplo 7.9 [veja equação (7.15)], discutimos as estimativas de Krueger (1993) do retorno do uso de computador no trabalho. Krueger também incluiu uma variável binária indicando se o trabalhador utiliza um computador em casa (como também um termo de interação entre a utilização de computador no trabalho e em casa). Sua principal razão para incluir a utilização de computador em casa na equação foi a de substituir "aptidão técnica" não observada que pudesse afetar diretamente o salário e estar relacionada com a utilização de computador no trabalho.

O Uso de Variáveis Dependentes Defasadas como Variáveis *Proxy*

Em algumas aplicações, como no exemplo anterior dos salários-hora, temos pelo menos uma vaga ideia de qual fator não observado gostaríamos de controlar. Isto facilita a escolha das variáveis *proxy*. Em outras aplicações, suspeitamos que uma ou mais variáveis independentes seja correlacionada com variável omitida, mas não temos a menor ideia de como obter uma *proxy* para a variável omitida. Nesses casos, podemos incluir, como um controle, o valor da variável dependente de um período anterior. Isso é especialmente útil para a análise de políticas públicas.

O uso de uma **variável dependente defasada** em equação de corte transversal aumenta os requisitos de dados, mas também fornece uma maneira simples de explicar fatores históricos que causam diferenças *correntes* na variável dependente que são difíceis de explicar de outras maneiras. Por exemplo, algumas cidades apresentaram altas taxas de criminalidade no passado. Muitos dos mesmos fatores não observados contribuem para as taxas de criminalidade tanto atuais como passadas. Da mesma forma, algumas universidades são, de modo tradicional, academicamente melhores que outras. Efeitos inerciais também são capturados trabalhando-se com defasagens de *y*.

Considere uma equação simples para explicar as taxas de criminalidade de uma cidade:

$$crime = \beta_0 + \beta_1 desemp + \beta_2 gasto + \beta_3 crime_{-1} + u, \qquad (9.16)$$

em que *crime* é uma medida do *crime per capita*, *desemp* é a taxa de desemprego da cidade, *gasto* é o dispêndio *per capita* para a imposição da lei e $crime_{-1}$ indica a taxa de criminalidade medida em algum ano anterior (que poderá ser o ano anterior ou vários anos atrás). Estamos interessados nos efeitos de *desemp* sobre *crime*, como também nos efeitos dos dispêndios com a imposição da lei sobre a variável crime.

Qual o propósito de se incluir $crime_{-1}$ na equação? Certamente, esperamos que $\beta_3 > 0$, já que crime possui inércia. Entretanto, a principal razão para colocá-la na equação é o fato de que cidades com taxas históricas elevadas de criminalidade devem gastar mais com a prevenção do crime. Assim, fatores não observados para nós (econometristas) que afetem *crime* são propensos a estarem correlacionados com *gasto* (e *desemp*). Se usarmos uma análise de corte transversal pura é improvável que obtenhamos um estimador não viesado do efeito causal dos dispêndios com a imposição da lei sobre o crime. Porém, ao incluirmos $crime_{-1}$ na equação, podemos, no mínimo, fazer a seguinte experiência: se duas cidades têm as mesmas taxas, anterior de criminalidade e atual de desemprego, então β_2 mede o efeito do gasto de mais um dólar com a imposição da lei sobre o crime.

EXEMPLO 9.4

(Taxas de Criminalidade em Cidades)

Estimamos uma versão de elasticidade constante do modelo do crime na equação (9.16) (*desemp*, por ser uma percentagem, é deixada em forma de nível). Os dados do arquivo CRIME2.RAW são de 46 cidades para o ano de 1987. A taxa de criminalidade também está disponível para 1982, e a utilizamos como variável independente adicional na tentativa de controlar as variáveis não observáveis das cidades que afetem o crime e possam estar correlacionadas com os dispêndios atuais com a imposição da lei. A Tabela 9.3 contém os resultados.

Tabela 9.3

Variável dependente: $\log(txcrim_{87})$.

Variáveis independentes	(1)	(2)
$desemp_{87}$	−0,029 (0,032)	0,009 (0,020)
$\log(disppclei_{87})$	0,203 (0,173)	−0,140 (0,109)
$\log(txcrim_{82})$	—	1,194 (0,132)
intercepto	3,34 (1,25)	0,076 (0,821)
Observações *R*-quadrado	46 0,057	46 0,680

Sem a taxa de criminalidade defasada na equação, os efeitos da taxa de desemprego e dos dispêndios com a imposição da lei não são intuitivos; nenhum dos coeficientes é estatisticamente significante, embora a estatística *t* da variável $\log(disppclei_{87})$ seja 1,17. É possível que o aumento dos dispêndios com a imposição da lei melhore a burocracia dos registros, e assim mais crimes serão *informados*. Entretanto, também é provável que cidades com taxas elevadas de criminalidade recentes gastem mais com a imposição da lei.

A adição do log da taxa de criminalidade de cinco anos atrás produz um grande efeito no coeficiente dos dispêndios. A elasticidade da taxa de criminalidade em relação aos dispêndios passa a ser de −0,14 com $t = -1,28$. Isso não é muito significante, mas sugere que um modelo mais sofisticado com mais cidades na amostra poderia produzir resultados melhores.

Não surpreende que as taxas atuais de criminalidade sejam fortemente relacionadas às taxas passadas. A estimativa indica que se a taxa de criminalidade em 1982 fosse 1% mais alta, então a taxa de criminalidade prevista de 1987 seria cerca de 1,19% mais alta. Não podemos rejeitar a hipótese de que a elasticidade da criminalidade corrente em relação à criminalidade passada seja unitária [$t = (1,194 - 1)/0,132 \approx 1,47$]. A adição da taxa de criminalidade passada aumenta o poder explicativo da regressão de maneira marcante, mas isso não surpreende. A razão principal para incluir a taxa de criminalidade defasada é a obtenção de uma melhor estimativa do efeito *ceteris paribus* de $\log(disppclei_{87})$ sobre $\log(txcrim_{87})$.

A prática de usar uma variável *y* defasada como um método geral para controlar variáveis não observadas está longe de ser perfeita. Porém, ela pode auxiliar na obtenção de uma melhor estimativa dos efeitos das variáveis de políticas de governo em diversos resultados.

A adição de um valor defasado de *y* não é a única maneira de utilizarmos dois anos de dados para controlar fatores omitidos. Quando discutirmos sobre métodos de dados em painel nos Capítulos 13 e 14, abordaremos outras maneiras sobre o uso de dados repetidos nas mesmas unidades de corte transversal em diferentes pontos no tempo.

Uma Inclinação Diferente na Regressão Múltipla

A discussão sobre variáveis identificadoras nesta seção sugere uma maneira alternativa de interpretar uma análise de regressão múltipla quando não observamos, necessariamente, todas as variáveis explicativas relevantes. Até agora, temos especificado o modelo populacional de interesse, com um erro suplementar, como na equação (9.9). Nossa discussão daquele exemplo dependia se tínhamos uma variável identificadora adequada (pontuação do QI, nesse caso, outros testes de pontuação de forma mais geral) das variáveis explicativas não observadas, que nós chamamos de "destreza".

Um método mais geral, menos estruturado, para multiplicar regressão é abrir mão da especificação de modelos com não observáveis. Em vez disso, começamos com a premissa que temos acesso a um conjunto de variáveis explicativas observáveis – que inclui as variáveis de interesse primordial, tais como anos de escolaridade, e de controles, tais como testes de pontuação observáveis. Então, modelamos a média de *y* condicional nas variáveis explicativas observáveis. Por exemplo, no exemplo salarial com *lsalário* denotando log(*salário*), podemos estimar E(*lsalárioleduc,exper,perm,sul,urban,negro,QI*) – exatamente o que está relatado na Tabela 9.2. A diferença é que agora definimos nossas metas mais modestamente, ou seja, em lugar de introduzirmos o conceito nebuloso de "destreza" na equação (9.9), estabelecemos desde o início que vamos estimar o efeito *ceteris paribus* da educação mantendo fixa a *QI* (e os outros fatores observados). Não há necessidade de argumentarmos se *QI* é um indicador adequado da destreza. Consequentemente, embora possamos não estar respondendo à pergunta subjacente da equação (9.9), estaremos respondendo uma pergunta de interesse: se duas pessoas tiverem os mesmos níveis de *QI* (e mesmos valores de experiência, permanência, e assim por diante), mas diferirem em níveis de escolaridade em um ano, qual é a diferença esperada em seus log salários?

Como outro exemplo, se incluirmos como uma variável explicativa a taxa de pobreza numa regressão de nível de escolaridade para avaliarmos o efeito dos gastos com pontuação do teste padronizado, devemos reconhecer que a taxa de pobreza somente captura de forma grosseira as diferenças relevantes nas crianças e pais entre escolas. Mas frequentemente é tudo o que temos e é melhor controlarmos a taxa de pobreza do que nada fazermos, pois não podemos encontrar indicadores adequados da "destreza" do estudante, o "envolvimento" dos pais e assim por diante. Quase certamente, o controle da taxa de pobreza nos levará para mais perto dos efeitos *ceteris paribus* dos gastos do que se deixarmos a taxa de pobreza de fora da análise.

Em algumas aplicações de análise de regressão, estamos interessados em predizer o efeito, *y*, em face de um conjunto de variáveis explicativas $(x_1, \ldots, x_k)$. Em tais casos, não faz muito sentido pensar em termos de "viés" nos coeficientes estimados em razão de variáveis omitidas. Em lugar disso, nos ateríamos na obtenção de um modelo que faça a predição tão bem quanto possível, e nos certificaremos de não incluirmos como regressores variáveis que não possam ser observadas no momento da predição. Por exemplo, um responsável pela admissão numa faculdade ou universidade pode ter interesse em predizer o êxito do aluno na faculdade, avaliado pela nota média de graduação, em termos de variáveis que possam ser avaliadas no momento da matrícula. Entre essas variáveis estariam incluídas o desempenho no ensino médio (talvez somente a nota média de graduação, mas possivelmente o desempenho em tipos específicos de cursos), escores de testes padronizados, a participação em diversas atividades

(como debates ou clube de matemática), e até mesmo variáveis sobre os antecedentes familiares. Não incluiríamos uma variável que avalia a frequência à aulas da faculdade porque não observaríamos a frequência no momento da matrícula. Tampouco nos preocuparíamos com potenciais "vieses" causados pela omissão de uma variável de frequência: não temos interesse em, digamos, medir o efeito da nota de conclusão do ensino médio mantendo fixa a frequência na faculdade. Do mesmo modo, não nos preocuparíamos sobre vieses nos coeficientes, pois não podemos observar fatores tais como motivação. Naturalmente, para propósitos preditivos, seria de grande ajuda se tivéssemos um indicador de motivação, mas na sua ausência nos adaptamos ao melhor modelo que pudermos com variáveis explicativas observadas.

9.3 MODELOS COM INCLINAÇÕES ALEATÓRIAS

Em nossa abordagem sobre regressão, até agora, temos presumido que os coeficientes de inclinação são os mesmos em todos os indivíduos na população, ou que, se as inclinações diferem, elas o fazem por características mensuráveis, caso em que somos levados a modelos de regressão que contêm termos de interação. Por exemplo, como vimos na Seção 7.4, podemos permitir que o retorno da educação difira entre homens e mulheres pela interação da educação com uma variável simulada do gênero numa equação log salário.

Aqui estamos interessados numa questão relacionada, mas diferente: e se o efeito parcial de uma variável depender de fatores não observados que variam por unidade populacional? Se tivermos somente uma variável explicativa, x, poderemos escrever um modelo geral (de uma extração aleatória, i, da população, para dar ênfase) como

$$y_i = a_i + b_i x_i \qquad \textbf{9.17}$$

em que a_i é o intercepto da unidade i e b_i é a inclinação. No modelo de regressão simples do Capítulo 2 consideramos $b_i = \beta$ e classificamos a_i como o erro u_i. O modelo em (9.17) é algumas vezes chamado de um **modelo de coeficiente aleatório** ou **modelo de inclinação aleatória** porque o coeficiente de inclinação não observado, b_i é visto como uma extração aleatória da população juntamente com os dados observados (x_i, y_i), e o intercepto não observado, a_i. Como um exemplo, se $y = \log(salário_i)$ e $x_i = educ_i$, então (9.17) permite que o retorno da educação, b_i, varie por pessoa. Se, digamos, b_i contiver habilidade não medida (da mesma forma que a_i conteria), o efeito parcial de um ano adicional de educação formal pode depender da habilidade.

Com uma amostra aleatória de tamanho n, nós (implicitamente) extraímos n valores de b_i juntamente com n valores de a_i (e os dados observados em x e em y). Obviamente, não podemos estimar uma inclinação — ou, de fato também, um intercepto — de cada i. Mas podemos esperar estimar a média de inclinação (e intercepto médio), em que a média está em toda a população. Portanto, defina $\alpha = \mathrm{E}(a_i)$ e $\beta = \mathrm{E}(b_i)$. Então β será a média do efeito parcial de x sobre y, e assim chamamos β de **efeito parcial médio (APE)**, ou *efeito marginal médio* (*AME*). No contexto de uma equação log salário-hora, β é o retorno médio de um ano adicional de educação formal na população.

Se escrevermos $a_i = \alpha + c_i$ e $b_i = \beta + d_i$, então d_i será o desvio específico do indivíduo da APE. Por construção, $\mathrm{E}(c_i) = 0$ e $\mathrm{E}(d_i) = 0$. Fazendo a substituição na (9.17) teremos

$$y_i = \alpha + \beta x_i + c_i + d_i x_i \equiv \alpha + \beta x_i + u_i \qquad \textbf{9.18}$$

em que $u_i = c_i + d_i x_i$. (Para tornar mais fácil seguir a notação, agora usaremos α, o valor médio de a_i, como o intercepto, e β, a média de b_i, como a inclinação.) Em outras palavras, podemos escrever o coeficiente aleatório como um modelo de coeficiente constante, mas onde o termo de erro contém uma interação entre uma não observável, d_i, e a variável explicativa observada, x_i.

E quando uma regressão simples de y_i sobre x_i produz uma estimativa não viesada de β (e de α)? Podemos aplicar o efeito de não viesamento do Capítulo 2. Se $\mathrm{E}(u_i \mid x_i) = 0$, então os MQO serão de forma geral não viesados. Quando $u_i = c_i + d_i x_i$, será suficiente $\mathrm{E}(c_i \mid x_i) = \mathrm{E}(c_i) = 0$ e $\mathrm{E}(d_i \mid x_i) = \mathrm{E}(d_i) = 0$. Podemos escrevê-las em termos do intercepto e da inclinação específicos da unidade, desta forma

$$\mathrm{E}(a_i \mid x_i) = \mathrm{E}(a_i) \quad \text{e} \quad \mathrm{E}(b_i \mid x_i) = \mathrm{E}(b_i); \qquad \textbf{9.19}$$

isto é, a_i e b_i são ambas independentes da média de x_i. Esta é uma conclusão útil se permitir inclinações específicas para a unidade, pois MQO estima consistentemente as médias dessas inclinações quando são independentes da média da variável explicativa. (Veja o Problema 9.6 sobre um conjunto mais fraco de condições que implica consistência dos MQO.)

O termo de erro na (9.18) quase com certeza contém heteroscedasticidade. Aliás, se $\mathrm{Var}(c_i \mid x_i) = \sigma_c^2$, $\mathrm{Var}(d_i \mid x_i) = \sigma_d^2$, e $\mathrm{Cov}(c_i, d_i \mid x_i) = 0$, então

$$\mathrm{Var}(u_i \mid x_i) = \sigma_c^2 + \sigma_d^2 x_i^2 \qquad \textbf{9.20}$$

e portanto deve haver heteroscedasticidade em u_i a menos que $\sigma_d^2 = 0$, o que significa $b_i = \beta$ de todos i. Sabemos como avaliar a heteroscedasticidade desse tipo. Podemos usar os MQO e calcular os erros-padrão de heteroscedasticidade robusta e dos testes estatísticos, ou podemos estimar a função de variância na (9.20) e aplicar os mínimos quadrados ponderados. Claro, esta última estratégia impõe homoscedasticidade nos intercepto e inclinação aleatórios, e assim faríamos uma análise dos MQP plenamente robustos das violações da (9.20).

Em razão da equação (9.20), alguns autores gostam de ver a heteroscedasticidade em modelos de regressão em geral como surgindo dos coeficientes das inclinações aleatórias. Mas devemos lembrar que a forma de (9.20) é especial, e ela não permite heteroscedasticidade em a_i ou em b_i. Não podemos, de maneira convincente, fazer a distinção entre um modelo de inclinação aleatória, em que o intercepto e a inclinação são independentes de x_i, e um modelo de inclinação constante com heteroscedasticidade na a_i.

O tratamento da regressão múltipla é semelhante. De forma geral, escrevemos

$$y_i = a_i + b_{i1} x_{i1} + b_{i2} x_{i2} + \ldots + b_{ik} x_{ik} \,. \qquad \textbf{9.21}$$

Então, escrevendo $a_i = \alpha + c_i$ e $b_{ij} = \beta_j + d_{ij}$, teremos

$$y_i = \alpha + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i \,, \qquad \textbf{9.22}$$

em que $u_i = c_i + d_{i1} x_{i1} + \ldots + d_{ik} x_{ik}$. Se mantivermos as hipóteses da independência da média $\mathrm{E}(a_i \mid \mathbf{x}_i) = \mathrm{E}(a_i)$ e $\mathrm{E}(b_{ij} \mid \mathbf{x}_i) = \mathrm{E}(b_{ij})$, $j = 1, \ldots, k$, então $\mathrm{E}(y_i \mid \mathbf{x}_i) = \alpha + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$, e assim os MQO usando

uma amostra aleatória produz estimadores não viesados de α e de β_j. Como no caso da regressão simples, Var $(u_i|\mathbf{x}_i)$ quase certamente será heteroscedástica.

Podemos permitir que b_{ij} dependa das variáveis explicativas observáveis como também outros observáveis. Por exemplo, suponha que, com $k = 2$ o efeito de x_{i2} dependa de x_{i1}, e escrevemos $b_{i2} = \beta_2 + \delta_1(x_{i1} - \mu_1) + d_{i2}$, em que $\mu_1 = \mathrm{E}(x_{i1})$. Se considerarmos $\mathrm{E}(d_{i2}|\mathbf{x}_i) = 0$ (e de forma semelhante para c_i e d_{i1}), então $\mathrm{E}(y_i|x_{i1}, x_{i2}) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \delta_1 (x_{i1} - \mu_1)x_{i2}$, o que significa que temos uma interação entre x_{i1} e x_{i2}. Como subtraímos a média μ_1 de x_{i1}, β_2 será o efeito parcial médio de x_{i2}.

O ponto principal desta seção é que a permissão de inclinações aleatórias é razoavelmente simples se as inclinações forem independentes, ou pelo menos independentes da média, das variáveis explicativas. Além disso, podemos facilmente modelar as inclinações como funções das variáveis exógenas, o que leva a modelos com quadrados e interações. Claro, no Capítulo 6 discutimos como tais modelos podem ser úteis sem jamais introduzirmos a noção de uma inclinação aleatória. A especificação das inclinações aleatórias fornece uma justificativa em separado de tais modelos. A estimação torna-se consideravelmente mais difícil se o intercepto aleatório, como também algumas inclinações são correlacionadas com alguns dos regressores. Tratamos do problema das variáveis explicativas endógenas no Capítulo 15.

9.4 PROPRIEDADES DO MÉTODO MQO QUANDO HÁ ERROS DE MEDIDA

Algumas vezes, em aplicações econômicas, não podemos coletar dados da variável que verdadeiramente afetam o comportamento econômico. Um bom exemplo é o ganho marginal do imposto de renda com que se defronta uma família que esteja tentando determinar quanto contribuir para instituições de caridade em determinado ano. O ganho marginal pode ser difícil de ser obtido ou resumido como um número único para todos os níveis de renda. Em vez disso, podemos calcular o ganho médio baseado na renda total e no pagamento do imposto.

Quando utilizamos uma medida imprecisa de uma variável econômica em um modelo de regressão, nosso modelo conterá um erro de medida. Nesta seção derivamos as consequências do erro de medida para a estimação dos mínimos quadrados ordinários. O método MQO será coerente sob certas hipóteses, mas existem outras sob as quais ele será inconsistente. Em alguns desses casos, podemos inferir o tamanho do viés assimptótico.

Como veremos, o problema do erro de medida tem uma estrutura estatística similar ao problema variável — variável *proxy* omitida — discutido na seção anterior, mas eles são conceitualmente diferentes. No caso da variável *proxy*, estamos procurando uma variável que, de certo modo, é associada à variável não observada. No caso do erro de medida, a variável que não observamos tem um significado quantitativo bem definido (como o ganho marginal do imposto ou a renda anual), mas as medidas sobre elas registradas por nós podem conter erros. Por exemplo, a renda anual registrada é uma medida da renda anual efetiva, ao passo que a pontuação de QI é uma *proxy* da aptidão.

Outra diferença importante entre os problemas da variável *proxy* e do erro de medida, é que, no último caso, muitas vezes a variável independente mal medida é a de maior interesse. No caso da variável *proxy*, o efeito parcial da variável omitida raramente é de interesse central: normalmente estamos preocupados com os efeitos das outras variáveis independentes.

Antes de considerarmos os detalhes, devemos nos lembrar que o erro de medida é um problema somente quando as variáveis cujos dados o econometrista pode coletar diferem das variáveis que influenciam as decisões de indivíduos, famílias, firmas etc.

Erro de Medida na Variável Dependente

Começaremos com o caso no qual somente a variável dependente é medida com erro. Vamos chamar de y^* a variável (na população, como sempre) que queremos explicar. Por exemplo, y^* poderia ser a poupança familiar anual. O modelo de regressão tem a forma usual

$$y^* = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u, \qquad (9.23)$$

e supomos que satisfaz as hipóteses de Gauss-Markov. Seja y a medida observável de y^*. No caso da poupança, y é a poupança anual registrada. Infelizmente, as famílias não declaram com perfeição suas poupanças anuais; é fácil deixar categorias de fora ou superestimar o montante contribuído para determinado fundo. Geralmente, podemos esperar que y e y^* sejam diferentes, pelo menos em alguns subconjuntos de famílias na população.

O erro de medida (na população) é definido como a diferença entre o valor observado e o valor real:

$$e_0 = y - y^*. \qquad (9.24)$$

Para uma extração aleatória i na população, podemos escrever $e_{i0} = y_i - y_i^*$, mas o importante é como o erro de medida na população está relacionado a outros fatores. Para obter um modelo que pode ser estimado, escrevemos $y^* = y - e_0$, inserimos essa expressão na equação (9.23) e reorganizamos esta última:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u + e_0. \qquad (9.25)$$

O termo de erro na equação (9.25) é $u + e_0$. Como $y, x_1, x_2, \ldots, x_k$ são observados, podemos estimar este modelo por MQO. Na verdade, simplesmente ignoramos o fato de que y é uma medida imperfeita de y^* e prosseguimos da maneira habitual.

Quando o método MQO com y em lugar de y^* produz estimadores consistentes de β_j? Como o modelo original (9.23) satisfaz as hipóteses de Gauss-Markov, u tem média zero e é não correlacionado com cada x_j. É natural supor que o erro de medida tem média zero; se não for assim, simplesmente obteremos um estimador viesado do intercepto, β_0, o que raramente é motivo de preocupação. Muito mais importante é nossa suposição sobre a relação entre o erro de medida, e_0, e as variáveis explicativas, x_j. A suposição habitual é que o erro de medida em y é estatisticamente independente de cada variável explicativa. Se isso for verdade, então os estimadores MQO de (9.25) são não viesados e consistentes. Além disso, os procedimentos de inferência do método MQO (estatísticas t, F e LM) são válidos.

Se e_0 e u forem não correlacionados, como normalmente se supõe, então $\mathrm{Var}(u + e_0) = \sigma_u^2 + \sigma_0^2 > \sigma_u^2$. Isso significa que o erro de medida na variável dependente resulta em uma variância de erro maior do que quando não ocorre erro algum; isso produz, evidentemente, variâncias maiores dos estimadores MQO. Esses resultados devem ser esperados, e não há nada que possamos fazer (exceto coletar dados melhores). O resultado é que se o erro de medida for não correlacionado com as variáveis independentes, a estimação MQO possuirá boas propriedades.

EXEMPLO 9.5

(Função de Poupança com Erro de Medida)

Considere uma função de poupança

$$poup^* = \beta_0 + \beta_1 rend + \beta_2 tam + \beta_3 educ + \beta_4 idade + u,$$

EXEMPLO 9.5 (continuação)

mas na qual a poupança real (*poup**) possa desviar-se da poupança registrada (*poup*). A questão é saber se o tamanho do erro de medida em *poup* está ou não sistematicamente relacionado com as outras variáveis. Pode ser razoável presumir que o erro de medida não esteja correlacionado com *rend, tam, educ* e *idade*. De outro lado, podemos pensar que famílias com rendas mais elevadas, ou mais educação, declarem suas poupanças com mais precisão. Não podemos ter certeza se o erro de medida está correlacionado com *rend* ou *educ*, a menos que possamos coletar dados de *poup**; então o erro de medida poderá ser calculado para cada observação como $e_{i0} = poup_i - poup_i^*$.

Quando a variável dependente está na forma logarítmica, para a qual $\log(y^*)$ é a variável dependente, é natural que a equação do erro de medida seja da forma

$$\log(y) = \log(y^*) + e_0. \tag{9.26}$$

Isso é proveniente de um **erro de medida multiplicativo** de y: $y = y^*a_0$, em que $a_0 > 0$ e $e_0 = \log(a_0)$.

EXEMPLO 9.6

(Erro de Medida nas Taxas de Rejeição de Produtos Industriais)

Na Seção 7.6 discutimos um exemplo no qual queríamos determinar se a concessão de subsídios para treinamento de pessoal reduzia a taxa de rejeição de produtos das indústrias. Podemos certamente imaginar que a taxa de rejeição registrada pela empresa seja medida com erro. (De fato, a maioria das empresas da amostra sequer computa taxas de rejeição de seus produtos.) Em uma estrutura de regressão simples isso é capturado pela equação

$$\log(rejei^*) = \beta_0 + \beta_1 subs + u,$$

em que *rejei** é a rejeição verdadeira e *subs* é a variável *dummy* indicando se uma empresa recebeu subsídios. A equação do erro de medida é

$$\log(rejei) = \log(rejei^*) + e_0.$$

O erro de medida e_0 é independente de a empresa ter, ou não, recebido subsídios? De um ponto de vista crítico, é possível pensar que uma empresa que tenha recebido subsídios está mais propensa a esconder sua taxa de rejeição para fazer com que os subsídios pareçam efetivos. Se isso acontecer, então, na equação estimada

$$\log(rejei) = \beta_0 + \beta_1 subs + u + e_0,$$

o erro $u + e_0$ é negativamente correlacionado com *subs*. Isso produziria um viés para baixo em β_1, o que tenderia a fazer com que o programa de treinamento parecesse mais efetivo do que na realidade foi. (Lembre-se: um β_1 mais negativo significa que o programa foi mais efetivo, pois uma melhor produtividade do trabalhador está associada a uma taxa de rejeição mais baixa.)

O fator preponderante desta subseção é que o erro de medida na variável dependente pode causar vieses no método MQO se ele for sistematicamente relacionado com uma ou mais das variáveis explicativas. Se o erro de medida for apenas um erro de informação aleatório que seja independente das variáveis explicativas, como muitas vezes é presumido, o método MQO é perfeitamente apropriado.

Erro de Medida em uma Variável Explicativa

Tradicionalmente, o erro de medida em uma variável explicativa tem sido considerado um problema muito mais importante do que o erro de medida em uma variável dependente. Nesta subseção veremos a razão de isso ser assim.

Comecemos com o modelo de regressão simples

$$y = \beta_0 + \beta_1 x_1^* + u, \tag{9.27}$$

supondo que ele satisfaz pelo menos as primeiras quatro hipóteses de Gauss-Markov. Isso significa que a estimação de (9.27) por MQO produziria estimadores de β_0 e β_1 não viesados e consistentes. O problema é que x_1^* não é observado. Em vez disso, temos uma medida de x_1^*, que pode ser chamada de x_1. Por exemplo, x_1^* poderia ser a verdadeira renda e x_1 poderia ser a renda registrada.

O erro de medida na população é simplesmente

$$e_1 = x_1 - x_1^*, \tag{9.28}$$

e pode ser positivo, negativo ou zero. Presumimos que o erro de medida *médio* na população é zero: $E(e_1) = 0$. Isso é natural e, de qualquer forma, não afeta a importante conclusão a seguir. Uma suposição sustentada no que segue é que u é não correlacionado com x_1^* e x_1. Em termos de expectativa condicional, podemos escrevê-la como $E(y|x_1^*, x_1) = E(y|x_1^*)$, que apenas diz que x_1 não afeta y após ter-se controlado x_1^*. Usamos a mesma suposição no caso da variável *proxy* e isso não é controverso; ela se mantém quase que por definição.

Queremos saber as propriedades de MQO se simplesmente substituirmos x_1^* por x_1 e executarmos a regressão de y sobre x_1. Elas dependerão crucialmente das suposições que fizermos sobre o erro de medida. Duas hipóteses têm sido enfatizadas na literatura econométrica, e ambas representam extremos opostos. A primeira hipótese é que e_1 é não correlacionado com a medida *observada*, x_1:

$$\text{Cov}(x_1, e_1) = 0. \tag{9.29}$$

Da relação em (9.28), se a hipótese (9.29) for verdadeira, então e_1 deve ser correlacionado com a variável não observada x_1^*. Para determinar as propriedades de MQO neste caso, escrevemos $x_1^* = x_1 - e_1$ e inserimos esta expressão na equação (9.27):

$$y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1). \tag{9.30}$$

Como presumimos que tanto u quanto e_1 têm média zero e são não correlacionados com x_1, $u - \beta_1 e_1$ tem média zero e é não correlacionado com x_1. Em consequência, a estimação de MQO com x_1 em lugar

de x_1^* produz um estimador consistente de β_1 (e também de β_0). Como u é não correlacionado com e_1, a variância do erro em (9.30) é $\text{Var}(u - \beta_1 e_1) = \sigma_u^2 + \beta_1^2\sigma_{e_1}^2$. Assim, exceto quando $\beta_1 = 0$, o erro de medida aumenta a variância do erro. Porém, isso não afeta quaisquer das propriedades de MQO (exceto pelo fato de que as variâncias de $\hat{\beta}_j$ serão maiores do que se observarmos x_1^* diretamente).

A hipótese de que e_1 é não correlacionada com x_1 é análoga à hipótese da variável *proxy* que fizemos na Seção 9.2. Como esta hipótese significa que o método MQO tem todas as suas propriedades perfeitas, não é isso o que os econometristas têm em mente quando se referem ao erro de medida em uma variável explicativa. A suposição de **erro clássico nas variáveis (CEV)** é que o erro de medida é não correlacionado com a variável explicativa *não observada*:

$$\text{Cov}(x_1^*, e_1) = 0. \tag{9.31}$$

Esta hipótese provém de ter-se escrito a medida observada como a soma da variável explicativa verdadeira com o erro de medida,

$$x_1 = x_1^* + e_1,$$

e em seguida presumindo que os dois componentes de x_1 são não correlacionados. (Isso não tem nada a ver com as hipóteses sobre u; sempre supomos que u é não correlacionado com x_1^* e x_1, e, portanto, com e_1.)

Se a hipótese (9.31) for válida, então x_1 e e_1 *devem* ser correlacionadas:

$$\text{Cov}(x_1, e_1) = \text{E}(x_1 e_1) = \text{E}(x_1^* e_1) + \text{E}(e_1^2) = 0 + \sigma_{e_1}^2 = \sigma_{e_1}^2. \tag{9.32}$$

Assim, a covariância entre x_1 e e_1 é igual à variância do erro de medida sob a hipótese CEV.

Com referência à equação (9.30), podemos ver que a correlação entre x_1 e e_1 causará problemas. Como u e x_1 são não correlacionados, a covariância entre x_1 e o erro composto $u - \beta_1 e_1$ é

$$\text{Cov}(x_1, u - \beta_1 e_1) = -\beta_1 \text{Cov}(x_1, e_1) = -\beta_1 \sigma_{e_1}^2.$$

Assim, no caso CEV, a regressão de MQO de y sobre x_1 produz um estimador viesado e inconsistente.

Utilizando os resultados assimptóticos do Capítulo 5, podemos determinar o montante de inconsistência no método MQO. O limite de probabilidade de $\hat{\beta}_1$ é β_1 mais a razão da covariância entre x_1 e $u - \beta_1 e_1$ e a variância de x_1:

$$\begin{aligned} \text{plim}(\hat{\beta}_1) &= \beta_1 + \frac{\text{Cov}(x_1, u - \beta_1 e_1)}{\text{Var}(x_1)} \\ &= \beta_1 - \frac{\beta_1 \sigma_{e_1}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2} \end{aligned} \tag{9.33}$$

$$\begin{aligned} &= \beta_1\left(1 - \frac{\sigma_{e_1}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2}\right) \\ &= \beta_1\left(\frac{\sigma_{x_1^*}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2}\right), \end{aligned} \tag{9.33}$$

em que usamos o fato de que $\text{Var}(x_1) = \text{Var}(x_1^*) + \text{Var}(e_1)$.

A equação (9.33) é bastante interessante. O termo que multiplica β_1, que é a razão $\text{Var}(x_1^*)/\text{Var}(x_1)$, é sempre menor que um [uma implicação da hipótese CEV (9.31)]. Assim, $\text{plim}(\hat{\beta}_1)$ está sempre mais perto de zero que β_1. Isso é chamado de **viés de atenuação** no MQO em razão do erro clássico nas variáveis: em médias (ou grandes) amostras, o efeito estimado de MQO será *atenuado*. Em particular, se β_1 for positivo, $\hat{\beta}_1$ tenderá a subestimar β_1. Esta é uma conclusão importante, porém ela depende da configuração CEV.

Se a variância de x_1^* for grande, em relação à variância no erro de medida, então a inconsistência no MQO será pequena. Isso é em razão do fato de $\text{Var}(x_1^*)/\text{Var}(x_1)$ ficar próximo da unidade, quando $\sigma_{x_1^*}^2/\sigma_{e_1}^2$ for grande. Portanto, dependendo do volume de variação em x_1^*, com relação a e_1, o erro de medida não causará, necessariamente, grandes vieses.

As coisas se complicam quando adicionamos mais variáveis explicativas. Como ilustração, considere o modelo

$$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + \beta_3 x_3 + u, \tag{9.34}$$

em que a primeira das três variáveis explicativas é medida com erro. Naturalmente, supomos que u é não correlacionado com x_1^*, x_2, x_3 e x_1. Novamente, a hipótese crucial refere-se ao erro de medida e_1. Em quase todos os casos, presume-se que e_1 é não correlacionado com x_2 e x_3 — as variáveis explicativas não medidas com erro. O grande problema é se e_1 é não correlacionado com x_1. Se for, então a regressão MQO de y sobre x_1, x_2 e x_3 produzirá estimadores consistentes. Pode-se ver isso facilmente escrevendo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u - \beta_1 e_1, \tag{9.35}$$

em que u e e_1 são ambos não correlacionados com todas as variáveis explicativas.

Sob a hipótese CEV em (9.31), o MQO será viesado e inconsistente, pois e_1 é correlacionado com x_1 na equação (9.35). Lembre-se de que isso significa que, em geral, *todos* os estimadores MQO serão viesados, e não somente $\hat{\beta}_1$. E quanto ao viés de atenuação derivado na equação (9.33)? Ainda existe um viés de atenuação ao se estimar β_1: pode ser demonstrado que

$$\text{plim}(\hat{\beta}_1) = \beta_1\left(\frac{\sigma_{r_1^*}^2}{\sigma_{r_1^*}^2 + \sigma_{e_1}^2}\right), \tag{9.36}$$

em que r_1^* é o erro populacional na equação $x_1^* = \alpha_0 + \alpha_1 x_2 + \alpha_2 x_3 + r_1^*$. A fórmula (9.36) também funciona no caso da variável geral k, quando x_1 for a única variável medida erroneamente.

As coisas são menos nítidas ao se estimar β_j nas variáveis não medidas com erro. No caso especial em que x_1^* é não correlacionado com x_2 e x_3, $\hat{\beta}_2$ e $\hat{\beta}_3$ são consistentes. Entretanto, na prática isso é raro. Geralmente, o erro de medida em uma única variável provoca inconsistência em todos os estimadores. Infelizmente, os tamanhos, e até mesmo as direções dos vieses, não são facilmente derivados.

EXEMPLO 9.7

(Equação da Nota Média com Erro de Medida)

Considere o problema de estimar o efeito da renda familiar na nota média da graduação, após ter-se controlado *emGPA* (nota média do ensino médio) e *SAT* (teste de aptidão acadêmica). Pode ser que, embora a renda familiar seja importante para o desempenho escolar antes da faculdade, ela não tenha efeito direto no desempenho na faculdade. Para testarmos isso, podemos postular o modelo

$$supGPA = \beta_0 + \beta_1 rendfam^* + \beta_2 emGPA + \beta_3 SAT + u,$$

em que *rendfam** é a renda anual familiar efetiva. (Ela pode aparecer na forma logarítmica, mas, para fins de ilustração, deixaremos na forma em nível.) Dados precisos sobre *supGPA*, *emGPA* e *SAT* são relativamente fáceis de ser obtidos. Porém, a renda familiar, especialmente as informadas pelos estudantes, podem facilmente ser incorretamente medidas. Se $rendfam = rendfam^* + e_1$, e a hipótese CEV for válida, então o uso da renda familiar informada em lugar da renda familiar efetiva viesará o estimador MQO de β_1 em direção a zero. Uma consequência disso é que um teste de H_0: $\beta_1 = 0$ terá menos possibilidade de detectar $\beta_1 > 0$.

Evidentemente, o erro de medida pode estar presente em mais de uma variável explicativa, ou em algumas das variáveis explicativas e na variável dependente. Como discutido anteriormente, qualquer erro de medida na variável dependente é usualmente presumido como não correlacionado com todas as variáveis explicativas, seja ele observado ou não. Derivar o viés nos estimadores MQO sob extensões das hipóteses CEV é complicado e não leva a resultados claros.

Em alguns casos, fica evidente que a hipótese CEV em (9.31) não pode ser verdadeira. Considere uma variante do Exemplo 9.7:

$$supGPA = \beta_0 + \beta_1 fumou^* + \beta_2 emGPA + \beta_3 SAT + u,$$

em que *fumou** é o número efetivo de vezes que um estudante fumou maconha nos últimos 30 dias. A variável *fumou* é a resposta à questão: em quantas ocasiões distintas um estudante fumou maconha nos últimos 30 dias? Suponha que postulemos o modelo-padrão de erro de medida

$$fumou = fumou^* + e_1,$$

Mesmo que admitamos que os estudantes tentem informar a verdade, é pouco provável que a hipótese CEV se mantenha. As pessoas que nunca fumam maconha sobretudo — de forma que $fumou^* = 0$ — provavelmente responderão $fumou = 0$, de modo que o erro de medida será, provavelmente, zero para os estudantes que nunca fumaram maconha. Quando $fumou^* > 0$ é muito mais provável que o estudante tenha errado na contagem de quantas vezes fumou maconha nos últimos 30 dias. Isso significa que o erro de medida e_1 e o número *efetivo* de vezes em que fumou, *fumou**, são correlacionados, o que infringe a hipótese CEV em (9.31). Infelizmente, derivar as implicações do erro de medida que não satisfaçam (9.29) ou (9.31) é difícil e está além do escopo deste livro.

QUESTÃO 9.3

Seja *educ** o grau efetivo de escolaridade, medido em anos (que pode ser um número não inteiro), e seja *educ* o número de anos mais elevado de educação formal. Você acha que *educ* e *educ** são relacionados pelo modelo clássico de erro nas variáveis?

Antes de encerrarmos esta seção, enfatizamos que a hipótese CEV (9.31), embora mais verossímil que a hipótese (9.29), ainda é uma hipótese forte. A verdade está, provavelmente, em algum ponto entre as duas, e se e_1 for correlacionado com x_1^* e x_1, MQO é inconsistente. Isto levanta uma questão importante: temos que conviver com estimadores inconsistentes sob o modelo clássico de erro nas variáveis, ou com outros tipos de erros de medida que são correlacionados com x_1? Felizmente, a resposta é não. O Capítulo 15 mostra como, sob certas hipóteses, os parâmetros podem ser consistentemente estimados na presença de erros gerais de medida. Adiamos a discussão para mais tarde, pois ela exige que abandonemos o âmbito da estimação MQO. (Veja o Problema 9.7, no final deste capítulo sobre como múltiplos indicadores podem ser usados para reduzir o viés de atenuação).

9.5 AUSÊNCIA DE DADOS, AMOSTRAS NÃO ALEATÓRIAS E OBSERVAÇÕES EXTREMAS

O problema do erro de medida discutido na seção anterior pode ser visto como um problema de dados: não podemos obter dados sobre as variáveis de interesse. Além disso, sob o modelo clássico de erro nas variáveis, o termo de erro composto é correlacionado com a variável independente incorretamente medida, violando as hipóteses de Gauss-Markov.

Outro problema de dados que discutimos várias vezes em capítulos anteriores é a multicolinearidade entre as variáveis explicativas. Lembremo-nos de que a correlação entre variáveis explicativas não infringe hipótese alguma. Quando duas variáveis independentes são altamente correlacionadas, pode ser difícil estimar o efeito parcial de cada uma delas. Entretanto, isto é adequadamente refletido nas estatísticas de MQO usuais.

Nesta seção apresentamos uma introdução aos problemas de dados que podem violar a hipótese de amostragem aleatória, RLM.2. Podemos isolar casos nos quais a amostragem não aleatória não tem efeito prático sobre o método MQO. Em outros casos, a amostragem não aleatória faz com que os estimadores MQO sejam viesados e inconsistentes. Um tratamento mais completo que comprova várias das afirmações feitas aqui é apresentado no Capítulo 17.

Ausência de Dados

O problema de **ausência de dados** pode surgir de várias formas. Muitas vezes coletamos uma amostra aleatória de pessoas, escolas, cidades etc. e mais tarde descobrimos que estão faltando informações de algumas variáveis importantes para diversas unidades na amostra. Por exemplo, no conjunto de dados do arquivo BWGHT.RAW, 197 das 1.388 observações não têm informação alguma sobre a educação do pai, da mãe ou de ambos. No conjunto de dados sobre salários medianos iniciais dos recém-formados em faculdades de direito, no arquivo LAWSCH85.RAW, seis das 156 faculdades não têm informações sobre as medianas da pontuação LSAT dos alunos, relativas a classes novas; também faltam outras variáveis de algumas faculdades de direito.

Quando estão faltando dados de uma observação na variável dependente ou em uma das variáveis independentes, a observação não pode ser usada em uma análise de regressão múltipla padrão. Aliás,

desde que os dados ausentes tenham sido adequadamente indicados, todos os modernos programas de regressão rastreiam os dados e simplesmente ignoram as observações ao calcularem uma regressão. Vimos isso claramente na situação do peso de nascimento no Exemplo 4.9, quando 197 observações foram eliminadas em razão da não existência de informações sobre o nível de educação dos pais.

Além de reduzir o tamanho da amostra disponível para uma regressão, há consequências *estatísticas* provocadas pela ausência de dados? Depende do motivo da ausência dos dados. Se estes estiverem faltando aleatoriamente, então o tamanho da amostra aleatória disponível da população será simplesmente reduzido. Embora isso torne os estimadores menos precisos, não produz viés algum: a hipótese de amostragem aleatória, RLM.2, ainda é válida. Existem maneiras de usar as informações das observações nas quais somente algumas variáveis estão faltando, mas, na prática, não se faz isso com frequência. A melhoria nos estimadores normalmente é pequena, embora o método seja um pouco complicado. Na maioria dos casos, simplesmente ignoramos as observações que representam falta de informação.

Amostras Não Aleatórias

A ausência de dados é mais problemática quando resulta de uma **amostra não aleatória** da população. Por exemplo, no conjunto de dados sobre pesos de nascimento, o que acontecerá se a probabilidade de que estejam faltando informações sobre o nível de educação for mais alta para as pessoas cujo nível de educação seja mais baixo que a média? Ou, na Seção 9.2, tenhamos usado um conjunto de dados sobre salários-hora que tenha incluído pontuações de QI. Esse conjunto de dados foi construído com a omissão de várias pessoas da amostra para as quais não havia informações sobre a pontuação do QI. Se a obtenção de escores de QI é mais fácil para as pessoas com QI mais elevado, a amostra não será representativa da população. A hipótese de amostragem aleatória RLM.2 está sendo violada e devemos nos preocupar com suas consequências durante a estimação MQO.

Felizmente, certos tipos de amostragens não aleatórias *não* causam viés ou inconsistência no MQO. Sob as hipóteses de Gauss-Markov (mas sem a RLM.2), a amostra pode ser escolhida com base nas variáveis *independentes* sem causar nenhum problema estatístico. Isso é chamado de *seleção da amostra com base nas variáveis independentes*, e é um exemplo de **seleção amostral exógena**. Para ilustrar, suponha que estejamos estimando uma função de poupança, na qual a poupança anual depende da renda, idade, tamanho da família e talvez de alguns outros fatores. Um modelo simples é

$$poupança = \beta_0 + \beta_1 renda + \beta_2 idade + \beta_3 tamanho + u. \qquad (9.37)$$

Suponha que nosso conjunto de dados foi montado com base em pesquisa feita com pessoas com mais de 35 anos de idade, com isso deixando-nos com uma amostra não aleatória de todos os adultos. Embora isso não seja o ideal, ainda podemos obter estimadores não viesados e consistentes dos parâmetros no modelo populacional (9.37), utilizando a amostra não aleatória. Não demonstraremos isso formalmente aqui, mas a razão pela qual o MQO na amostra não aleatória é não viesado é o fato de a função de regressão $E(poupança|renda,idade,tamanho)$ ser a mesma para qualquer subconjunto da população descrito por *renda, idade* ou *tamanho*. Desde que haja variação suficiente nas variáveis independentes na subpopulação, a seleção com base nas variáveis independentes não será um problema sério, exceto pelo fato de resultar em estimadores ineficientes.

No exemplo mencionado da pontuação do QI, as coisas não são tão nítidas, porque nenhuma regra fixa baseada no QI foi utilizada para incluir algo na amostra. Ao contrário, a *probabilidade* de estar na amostra aumenta com o QI. Se os outros fatores que determinam a seleção na amostra forem independentes do termo de erro na equação de salários, teremos outro caso de seleção amostral exógena, e o MQO usando a amostra selecionada terá todas as propriedades desejáveis sob as outras hipóteses de Gauss-Markov.

A situação é muito diferente quando a seleção é baseada na variável dependente, *y*, que é chamada de *seleção de amostra com base na variável dependente* e é um exemplo de **seleção amostral endógena**. Se a amostra tiver como base o fato de a variável dependente estar acima ou abaixo de determinado valor, sempre ocorrerá viés no MQO ao estimarmos o modelo populacional. Por exemplo, suponha que queiramos estimar a relação entre a riqueza individual e vários outros fatores na população adulta:

$$riqueza = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 idade + u. \qquad (9.38)$$

Suponha que somente pessoas com riqueza abaixo de US$ 250.000 sejam incluídas na amostra. Essa é uma amostra não aleatória da população de interesse, e se baseia no valor da variável dependente. A utilização de uma amostra de pessoas com riqueza abaixo de US$ 250.000 resultará em estimadores viesados e inconsistentes dos parâmetros em (9.32). Resumidamente, a razão é que a regressão populacional $E(riqueza|educ,exper,idade)$ não é a mesma que o valor esperado condicional da *riqueza* ser menor que US$ 250.000.

Outros esquemas de amostragem levam a amostras não aleatórias da população, em geral intencionalmente. Um método comum de coleta de dados é a **amostragem estratificada**, na qual a população é dividida em grupos ou estratos não sobrepostos. Então, alguns grupos aparecem com mais frequência do que a determinada por sua representação populacional e outros aparecem com menor frequência. Por exemplo, algumas pesquisas propositalmente superdimensionam grupos minoritários ou grupos de baixa renda. Quando forem necessários métodos especiais, isso dependerá outra vez de ser a estratificação exógena (baseada em variáveis explicativas exógenas) ou endógena (baseada na variável dependente). Suponha que uma pesquisa sobre o contingente militar superdimensionou mulheres porque o interesse inicial era estudar os fatores que determinam o pagamento às mulheres no serviço militar. (Superdimensionar um grupo que seja relativamente pequeno na população é comum na coleta de amostras estratificadas.) Desde que os homens também tenham sido representados na amostra, podemos usar o MQO na amostra estratificada para estimar qualquer diferencial de gênero, juntamente com os retornos da educação e da experiência de todo o contingente militar. (Podemos querer pressupor que retorno da educação e da experiência não sejam específicos quanto ao gênero.) A razão pela qual o MQO é não viesado e consistente está ligado ao fato de a estratificação ser feita com relação a uma variável explicativa, ou seja, o gênero.

Se, em vez disso, a pesquisa superdimensionou o contingente militar de salários mais baixos, então o MQO que utiliza a amostra estratificada não estima consistentemente os parâmetros da equação de salários dos militares, porque a estratificação é endógena. Em tais casos, são necessários métodos econométricos especiais [veja Wooldridge (2002, Capítulo 17)].

A amostragem estratificada é uma forma bastante óbvia de amostragem não aleatória. Outros problemas de seleção de amostras são mais sutis. Por exemplo, em vários dos exemplos anteriores estimamos os efeitos de diversas variáveis, particularmente educação e experiência, sobre os salários por hora. O conjunto de dados do arquivo WAGE1.RAW, que utilizamos amplamente, é, essencialmente, uma amostra aleatória de indivíduos *trabalhadores*. Economistas especializados na área trabalhista com frequência estão interessados em estimar os efeitos da, digamos, educação sobre os salários-hora *oferecidos*. A ideia é a seguinte: toda pessoa em idade de trabalhar se defronta com uma oferta de salários por hora e pode trabalhar, ou não, por aquele salário. Para alguém que trabalhe, o salário ofertado é o salário ganho. Para pessoas que não trabalhem não podemos, normalmente, observar o salário-hora oferecido. Agora, como a equação da oferta salarial

$$\log(salário\hat{h}^0) = \beta_0 + \beta_1 educ + \beta_2 exper + u \qquad (9.39)$$

representa a população de trabalhadores de todas as idades, não podemos estimá-la utilizando uma amostra aleatória desta população; em vez disso, temos informações sobre a oferta salarial somente para pessoas que trabalham (embora possamos obter dados sobre *educ* e *exper* de pessoas que não trabalham). Se utilizarmos uma amostra aleatória de pessoas que trabalham para estimar (9.39), obteremos estimadores não viesados? Este caso não é muito claro. Como a amostra é selecionada com base na decisão de alguém de trabalhar (em oposição ao tamanho do salário ofertado), este não é como o caso anterior. Porém, como a decisão de trabalhar pode estar relacionada com fatores não observados que afetem a oferta salarial, a seleção pode ser endógena, e isso pode resultar em um viés de seleção de amostra nos estimadores MQO. Trataremos os métodos que podem ser usados para testar e corrigir vieses de seleção de amostras no Capítulo 17.

QUESTÃO 9.4

Suponha que estejamos interessados nos efeitos dos gastos com campanha eleitoral feitos por candidatos, sobre o apoio dos eleitores. Alguns candidatos decidem não apresentar sua candidatura à reeleição. Se somente pudermos coletar os resultados da votação e dos gastos dos candidatos que efetivamente participaram da eleição, existe a possibilidade de ocorrer uma seleção endógena da amostra?

Observações Extremas (*outliers*) e Influentes

Em algumas aplicações, especialmente, mas não apenas nelas, com conjuntos de dados pequenos, as estimativas MQO são sensíveis à inclusão de uma ou várias observações. Uma abordagem completa das **observações extremas** e das **observações influenciadoras** está além do objetivo deste livro, pois um desenvolvimento formal exige álgebra matricial. De maneira vaga, uma observação é uma observação influenciadora se sua eliminação da análise muda as estimativas principais dos MQO em montante praticamente "grande". A noção de uma observação extrema também é um pouco vaga, pois ela exige que se compare os valores das variáveis de uma observação com aqueles do restante da amostra. Apesar disso, é desejável ficar vigilante quanto a observações "incomuns" porque elas podem afetar grandemente as estimativas MQO.

O MQO é suscetível a observações extremas porque ele minimiza a soma dos quadrados dos resíduos: grandes resíduos (positivos ou negativos) recebem muita carga no problema de minimização de mínimos quadrados. Se as estimativas mudarem em quantidade significativa quando modificamos ligeiramente nossa amostra, devemos nos preocupar.

Quando estatísticos e econometristas estudam teoricamente o problema dessas observações extremas, algumas vezes os dados são vistos como provenientes de uma amostra aleatória de determinada população — embora com uma distribuição pouco comum que pode resultar em valores extremos — e às vezes se presume que as observações extremas provêm de uma população diferente. De uma perspectiva prática, tais observações podem ocorrer por duas razões. O caso mais fácil de tratar é quando um engano foi cometido na entrada dos dados. A adição de zeros extras a um número ou a má colocação do ponto decimal podem mascarar as estimativas MQO, especialmente em amostras de pequeno tamanho. É sempre uma boa ideia calcular as estatísticas sumárias, especialmente os mínimos e os máximos, para detectar enganos da entrada dos dados. Infelizmente, entradas incorretas nem sempre são óbvias.

Os valores extremos também podem surgir quando se faz a amostragem de uma pequena população, se um ou vários membros da população forem muito diferentes em alguns aspectos relevantes do resto da população. Pode ser difícil tomar a decisão de manter ou desprezar tais observações em uma

análise de regressão, e as propriedades estatísticas dos estimadores resultantes são complicadas. Observações extremas podem fornecer informações importantes aumentando-se a variação das variáveis explicativas (o que reduz os erros-padrão). Porém, os resultados de MQO deverão ser descritos com e sem as observações extremas, nos casos em que um ou vários pontos dos dados alteram substancialmente os resultados.

EXEMPLO 9.8

(Intensidade de Pesquisa e Desenvolvimento (P&D) e Tamanho das Empresas)

Suponha que os gastos com pesquisa e desenvolvimento como uma percentagem das vendas (*pdintens*) sejam relacionados a *vendas* (em milhões) e a lucros como uma percentagem das vendas (*lucrmarg*):

$$pdintens = \beta_0 + \beta_1 vendas + \beta_2 lucrmarg + u. \qquad (9.40)$$

A equação MQO utilizando os dados de 32 indústrias químicas, contidos no arquivo RDCHEM.RAW é

$$\widehat{pdintens} = 2{,}625 + 0{,}000053\ vendas + 0{,}0446\ lucrmarg$$
$$(0{,}586)\ \ (0{,}000044)\qquad\quad (0{,}0462)$$
$$n = 32,\ R^2 = 0{,}0761,\ \bar{R}^2 = 0{,}0124.$$

Nem *vendas* nem *lucrmarg* são estatisticamente significantes, mesmo no nível de 10% nesta regressão.

Das 32 empresas, 31 têm vendas anuais abaixo de US$ 20 bilhões. Uma empresa tem vendas anuais de quase US$ 40 bilhões. A Figura 9.1 mostra a distância desta empresa do resto da amostra. Em termos de vendas, esta empresa é quase duas vezes maior do que qualquer outra, de modo que deve ser uma boa ideia estimar o modelo sem ela. Quando fazemos isso, obtemos

$$\widehat{pdintens} = 2{,}297 + 0{,}000186\ vendas + 0{,}0478\ lucrmarg$$
$$(0{,}592)\ \ (0{,}000084)\qquad\quad (0{,}0445)$$
$$n = 31,\ R^2 = 0{,}1728,\ \bar{R}^2 = 0{,}1137.$$

Se a maior empresa for eliminada da regressão, o coeficiente de *vendas* mais do que triplica, e agora ela tem uma estatística *t* acima de dois. Usando a amostra de empresas menores, concluiríamos que existe um efeito positivo estatisticamente significante entre a intensidade de P&D e o tamanho das empresas. A margem de lucro ainda não é significante, e seu coeficiente não mudou muito.

Algumas vezes, as observações extremas são definidas pelo tamanho do resíduo em uma regressão MQO, na qual todas as observações são usadas. Geralmente, essa não é uma boa ideia porque as estimativas MQO se ajustam para tornarem as somas dos quadrados dos resíduos tão pequenas quanto possível. No exemplo anterior, a inclusão da maior firma achatou consideravelmente as linhas de regressão do MQO, o que não tornou especialmente grande o resíduo daquela estimação. Aliás, o resíduo da maior firma é −1,62 quando todas 32 observações são usadas. Este valor do resíduo não é nem mesmo um desvio-padrão estimado, $\hat{\sigma} = 1{,}82$ da média dos resíduos, que é zero pela construção.

Resíduos estudentizados são obtidos dos resíduos originais dos MQO pela divisão deles por uma estimativa de seus desvios-padrão (condicionais às variáveis explicativas na amostra). A fórmula dos resíduos estudentizados vale-se de álgebra matricial, mas acontece de existir um artifício simples para calcular um resíduo estudentizado de qualquer observação. Ou seja, defina uma variável *dummy* igual

a um para aquela observação — digamos, observação h — e então inclua-a na regressão (usando todas as observações) juntamente com as outras variáveis explicativas. O coeficiente na variável *dummy* tem uma interpretação útil: ele é o resíduo da observação h calculada da linha de regressão usando somente as *outras* observações. Portanto, o coeficiente *dummy* pode ser usado para verificarmos quão longe a observação está da linha de regressão obtida sem o uso daquela observação. Melhor ainda, a estatística t na variável *dummy* é igual ao resíduo estudentizado da observação h. Sob as hipóteses do modelo linear clássico esta estatística t tem uma distribuição t_{n-k-1}. Portanto, um valor grande da estatística t (em valor absoluto) infere um resíduo grande relativo ao desvio-padrão estimado.

Para o Exemplo 9.8, se definirmos uma variável *dummy* para a maior firma (observação 10 no arquivo de dados), e a incluirmos como um regressor adicional, seu coeficiente será −6,57, verificando que a observação da maior firma está muito longe da linha de regressão obtida usando as outras observações. Porém, quando estudentizado, o resíduo é de apenas −1,82. Embora isto seja uma estatística t marginalmente significante (p-valor bilateral = 0,08), ela não está nem perto de ser o maior resíduo estudentizado na amostra. Se usarmos o mesmo método para a observação com o mais alto valor de *pdintens* — a primeira observação, com *pdintens* ≈ 9,42 — o coeficiente na variável *dummy* será 6,72 com uma estatística t de 4,56. Portanto, por este indicador, a primeira observação está mais para extremos do que a décima. Todavia a eliminação da primeira observação altera o coeficiente em *vendas* em apenas um pequeno montante (para aproximadamente 0,000051 de 0,000053), embora o coeficiente na *lucrmarg* se torne maior e estatisticamente significante. Então, a primeira observação é "extrema" também? Esses cálculos mostram o enigma em que podemos nos meter quando tentamos determinar observações que deveriam ser excluídas de uma análise de regressão, mesmo quando o conjunto de dados é pequeno. Infelizmente, o tamanho do resíduo estudentizado não necessita corresponder a quão influente uma observação é para as estimativas de inclinação pelos MQO, e certamente não para todas elas simultaneamente.

Um problema geral com o uso do resíduo estudentizado é que, na realidade, todas as outras observações são usadas para estimar a linha de regressão para calcular o resíduo de uma determinada observação. Em outras palavras, quando o resíduo estudentizado é obtido da primeira observação, a décima observação foi usada na estimativa do intercepto e da inclinação. Conhecido quão plana é a linha de regressão com a maior firma (décima observação) incluída, não é tão surpreendente que a primeira observação, com seu alto valor de *pdintens*, esteja tão afastada da linha de regressão.

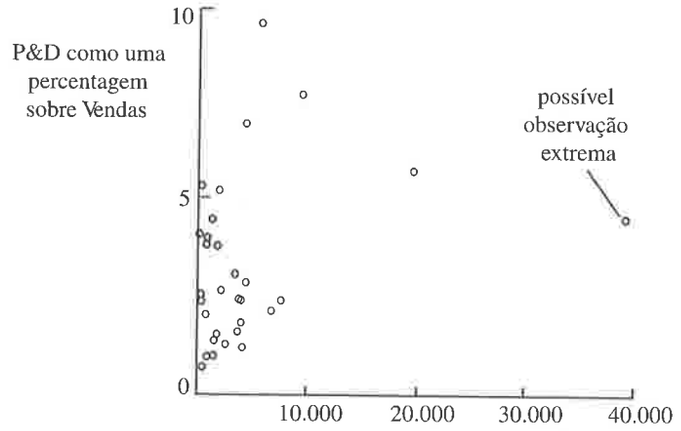
Claro, podemos adicionar duas variáveis *dummy* ao mesmo tempo — uma para a primeira observação e outra para a décima – que tem o efeito de usar somente as restantes 30 observações para estimar a linha de regressão. Se estimarmos a equação sem a primeira e a décima observações, os resultados serão

$$\widehat{pdintens} = 1{,}939 + 0{,}000160\ vendas + 0{,}0701\ lucrmarg$$
$$(0{,}459)\quad (0{,}00065)\qquad\quad (0{,}0343)$$
$$n = 30,\ R^2 = 0{,}2711,\ \bar{R}^2 = 0{,}2171$$

O coeficiente da *dummy* da primeira observação é 6,47 (t = 4,58), e da décima observação é −5,41 (t = −1,95). Observe que os coeficientes em *vendas* e em *lucrmarg* são ambos estatisticamente significantes, a última quase o nível de 5% contra uma alternativa bilateral (p-valor = 0,051). Mesmo nesta regressão ainda existem duas observações com resíduos estudentizados maiores que dois (correspondendo às duas observações restantes com intensidade de P&D acima de seis).

Algumas formas funcionais são menos sensíveis a observações extremas. Na Seção 6.2, mencionamos que, para a maioria das variáveis econômicas, a transformação logarítmica estreita significantemente a amplitude dos dados e também produz formas funcionais — tais como os modelos de elasticidade constante — que podem explicar uma gama mais ampla de dados.

Figura 9.1
Diagrama de dispersão da intensidade de pesquisa e desenvolvimento em relação às vendas da empresa.



EXEMPLO 9.9
(Intensidade de P&D)

Podemos testar se a intensidade de P&D aumenta com o tamanho das empresas, começando com o modelo

$$pd = vendas^{\beta_1} exp(\beta_0 + \beta_2 lucrmarg + u). \qquad (9.41)$$

Assim, mantendo fixos os outros fatores, a intensidade de P&D aumenta com *vendas* se, e somente se, $\beta_1 > 1$. Considerando o log de (9.41) teremos

$$\log(pd) = \beta_0 + \beta_1 \log(vendas) + \beta_2 lucrmarg + u. \qquad (9.42)$$

Quando usamos todas as 32 empresas, a equação de regressão é

$$\widehat{\log(pd)} = -4{,}378 + 1{,}084\ \log(vendas) + 0{,}0217\ lucrmarg,$$
$$(0{,}468)\quad (0{,}062)\qquad\quad (0{,}0128)$$
$$n = 32,\ R^2 = 0{,}9180, \bar{R}^2 = 0{,}9123,$$

ao passo que ao se retirar a maior empresa resulta na equação

$$\widehat{\log(pd)} = -4{,}404 + 1{,}088\ \log(vendas) + 0{,}0218\ lucrmarg,$$
$$(0{,}511)\quad (0{,}067)\qquad\quad (0{,}0130)$$
$$n = 31,\ R^2 = 0{,}9037, \bar{R}^2 = 0{,}8968.$$

EXEMPLO 9.9 (continuação)

Esses resultados são praticamente os mesmos. Em nenhum dos casos rejeitamos a hipótese nula $H_0: \beta_1 = 1$ contra $H_1: \beta_1 > 1$ (Por quê?).

Em alguns casos, certas observações são, desde o princípio, suspeitas de serem fundamentalmente diferentes do restante da amostra. Isso acontece com frequência quando utilizamos dados em níveis muito agregados, como os níveis de cidades, municípios ou estados. O que segue é um exemplo disso.

EXEMPLO 9.10

(Taxas Estaduais de Mortalidade Infantil)

Informações sobre mortalidade infantil, renda *per capita* e saúde podem ser obtidas, em nível de estados, no *Statistical Abstracts of the United States* (Resumo Estatístico dos Estados Unidos). Faremos aqui uma análise simples apenas para ilustrar o efeito das observações extremas. Os dados são para o ano de 1990, e temos todos os 50 estados dos Estados Unidos, mais a capital, o Distrito de Colúmbia (D.C.). A variável *mortinf* é o número de mortes no primeiro ano de vida por 1.000 nascimentos, *rendpc* é a renda *per capita*, *medic* representa médicos por 100.000 habitantes, e *popul* é a população (em milhares). Os dados estão contidos no arquivo INFMRT.RAW. Incluímos todas as variáveis independentes na forma logarítmica:

$$\widehat{mortinf} = \underset{(20,43)}{33,86} - \underset{(2,60)}{4,68}\log(rendpc) + \underset{(1,51)}{4,15}\log(medic) - \underset{(0,287)}{0,088}\log(popul) \quad (9.43)$$
$$n = 51, R^2 = 0,139, \bar{R}^2 = 0,084.$$

A renda *per capita* mais alta tem uma relação estimada inversa em relação à mortalidade infantil, um resultado esperado. Porém, a variável médicos *per capita* está associada com taxas maiores de mortalidade infantil, o que é contrário à intuição. As taxas de mortalidade infantil parecem não estar relacionadas ao tamanho da população.

O Distrito de Colúmbia é incomum por ter bolsões de extrema pobreza e de grande riqueza em uma área tão pequena. Aliás, a taxa de mortalidade infantil do D.C. em 1990 foi de 20,7, comparado com 12,4 do estado seguinte com maior taxa. Ele também tem 615 médicos por 100.000 habitantes, comparados com 337 do segundo estado. A grande quantidade de médicos complementada pela elevada taxa de mortalidade infantil no D.C. poderia certamente influenciar os resultados. Se retirarmos o D.C. da regressão, teremos

$$\widehat{mortinf} = \underset{(12,42)}{23,95} - \underset{(1,64)}{0,57}\log(rendpc) + \underset{(1,19)}{2,74}\log(medic) + \underset{(0,191)}{0,629}\log(popul) \quad (9.44)$$
$$n = 50, R^2 = 0,273, \bar{R}^2 = 0,226.$$

EXEMPLO 9.10 (continuação)

Agora verificamos que o número maior de médicos *per capita* reduz a mortalidade infantil, e a estimativa é estatisticamente diferente de zero no nível de 5%. O efeito da renda *per capita* caiu drasticamente e não mais é estatisticamente significante. Na equação (9.44), as taxas de mortalidade infantil são mais elevadas nos estados mais populosos, e a relação é estatisticamente bastante significante. Além disso, muito mais variação em *mortinf* é explicada quando o D.C. é retirado da regressão. Claramente, o D.C. tem uma influência considerável nas estimativas iniciais, e provavelmente o deixaríamos de fora de qualquer análise futura.

Como o Exemplo 9.8 demonstra, inspecionar as observações na tentativa de determinar quais são extremas, e até quais têm influência substancial nas estimativas MQO, é um empreendimento difícil. Tratamentos mais avançados permitem abordagens mais formais para se determinar quais observações são mais prováveis de serem observações influenciadoras. Usando álgebra matricial, Belsley, Kuh, e Welsh (1980) definiram a *alavancagem* de uma observação, que formaliza a noção que uma observação tem uma grande ou pequena influência nas estimativas MQO. Esses autores também fornecem uma discussão mais minuciosa dos resíduos padronizados e estudentizados.

9.6 ESTIMAÇÃO DOS MÍNIMOS DESVIOS ABSOLUTOS

Em vez de tentar determinar quais observações, se alguma, têm influência indevida nas estimativas MQO, um método diferente para se defender contra observações extremas é usar um método de estimação que seja menos sensível às observações extremas que os MQO. Um método desse tipo, que se tornou popular entre os econometricistas dedicados é chamado **mínimos desvios absolutos (MDA)**. Os estimadores MDA de β_j num modelo linear minimiza a soma dos valores absolutos dos resíduos,

$$\min_{b_0, b_1, \ldots, b_k} \sum_{i=1}^{n} |y_i - b_0 - b_1 x_{i1} - \ldots - b_k x_{ik}|. \quad (9.45)$$

Diferente dos MQO, que minimizam a soma dos quadrados dos resíduos, as estimativas MDA não estão disponíveis em forma fechada — isto é, não podemos escrever as fórmulas para elas. Aliás, historicamente, a solução do problema na equação (9.45) foi computacionalmente difícil, especialmente com amostras de tamanhos grandes e muitas variáveis explicativas. Mas com as grandes melhorias na velocidade de computação nas últimas duas décadas, as estimativas MDA são razoavelmente fáceis de serem obtidas mesmo de grandes conjuntos de dados.

Como os MDA não produzem ponderação crescente de resíduos maiores, eles são muito menos sensíveis a mudanças nos valores extremos dos dados que os MQO. Aliás, é sabido que os MDA são projetados para estimar os parâmetros da **mediana condicional** de y dado $x_1, x_2, \ldots, x_k$ em vez da média condicional. Como a mediana não é afetada por grandes mudanças nas observações extremas, deduz-se que as estimativas do parâmetro MDA são mais resistentes às observações extremas. (Veja a Seção A.1 para uma breve discussão sobre a mediana da amostra). Ao escolher as estimativas, os MQO elevam ao quadrado cada resíduo, e assim, as estimativas MQO podem ser bastante sensíveis a observações atípicas, como vimos nos Exemplos 9.8 e 9.10.

Além de os MDA serem mais exigentes computacionalmente do que os MQO, uma segunda desvantagem dos MDA é que toda inferência estatística que envolve os estimadores MDA será justificável

somente na medida em que o tamanho da amostra cresça. [As fórmulas são complicadas e exigem álgebra matricial, e nós não precisamos delas aqui. Koenker (2005) fornece um tratamento abrangente.] Lembre-se que, no âmbito do modelo clássico de hipóteses lineares, as estatísticas t dos MQO têm exatas t distribuições, e as estatísticas F têm exatas F distribuições. Enquanto as versões assimptóticas dessas estatísticas estão disponíveis para MDA — e relatadas rotineiramente por pacotes de softwares que calculam as estimativas MDA — esses só se justificam em grandes amostras. Assim como o ônus adicional envolvido no cálculo das estimativas MDA, a falta de inferência exata dos MDA é uma preocupação marginal, pois a maioria das aplicações dos MDA envolve várias centenas, se não vários milhares, de observações. Claro, podemos estar abusando se aplicarmos aproximações de amostra grande em um exemplo como o Exemplo 9.8, com $n = 32$. De certa forma, isto não é muito diferente dos MQO porque, na maioria dos casos, precisamos recorrer a aproximações de amostra grande para justificar a inferência MQO sempre que qualquer uma das hipóteses do MLC falha.

Outra inconveniência sutil, mas importante, do MDA é o fato de que nem sempre ele estima consistentemente os parâmetros que aparecem na função de média condicional, $E(y|x_1, \ldots, x_k)$. Como mencionado anteriormente, o MDA foi construído para estimar os efeitos sobre a mediana condicional. Geralmente, a média e a mediana são as mesmas somente quando a distribuição de y, dadas as covariadas $x_1, \ldots, x_k$, é simétrica em relação a $\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$. (Equivalentemente, o termo de erro da população, u, é simétrico em relação a zero.) Lembre-se de que o MQO produz estimadores não viesados e consistentes dos parâmetros na média condicional, seja ou não simétrica a distribuição do erro; a simetria não aparece entre as hipóteses de Gauss-Markov. Quando MDA e MQO são aplicados em casos com distribuições assimétricas, o efeito parcial estimado de, digamos, x_1, obtido com o MDA pode ser muito diferente do efeito parcial obtido com o MQO. Porém, tal diferença pode apenas refletir a diferença entre a mediana e a média e pode não ter nada a ver com as observações extremas.

Se presumirmos que o erro populacional u no modelo (9.2) é *independente* de $(x_1, \ldots, x_k)$, as estimativas de inclinação MQO e MDA devem diferir apenas por erro de amostragem, seja ou não simétrica a distribuição de u. As estimativas dos interceptos geralmente serão diferentes, refletindo o fato de que, se a média de u for zero, sua mediana é diferente de zero sob assimetria. Infelizmente, a independência entre o erro e as variáveis explicativas é, muitas vezes, inacreditavelmente forte quando o MDA é aplicado. Em particular, a independência impede a heteroscedasticidade, um problema que muitas vezes surge em aplicações com distribuições assimétricas.

Mínimos desvios absolutos são um caso especial do que muitas vezes é chamado de regressão *robusta*. Infelizmente, a maneira como "robusta" está sendo usada aqui pode criar confusão. Na literatura estatística, um estimador de regressão robusta é relativamente insensível a observações extremas. Efetivamente, observações com grandes resíduos recebem menos peso do que nos mínimos quadrados. [Berk (1990) contém um tratamento introdutório de estimadores que são robustos em relação a observações extremas.] Com base em nossa discussão anterior, em linguagem econométrica, o MDA não é um estimador robusto da média condicional, pois ele exige hipóteses extras para estimar consistentemente os parâmetros da média condicional. Na equação (9.2), a distribuição de u dado $(x_1, \ldots, x_k)$ tem que ser simétrica em relação a zero, ou u deve ser independente de $(x_1, \ldots, x_k)$. Nada disso é exigido pelo MQO.

RESUMO

Investigamos mais detalhadamente alguns problemas importantes de especificação e de dados que muitas vezes surgem na análise de corte transversal empírica. Formas funcionais mal especificadas fazem com que a equação estimada seja difícil de ser interpretada. No entanto, a forma funcional incorreta pode ser detectada pela adição de termos quadráticos, pelo cálculo do teste RESET, ou fazendo-se o teste contra um modelo alternativo não aninhado, utilizando o teste de Davidson-MacKinnon. Não é necessária nenhuma coleta adicional de dados.

Resolver o problema das variáveis omitidas é mais difícil. Na Seção 9.2 discutimos uma possível solução com base no uso de uma variável *proxy* substituindo a variável omitida. Sob hipóteses razoáveis, a inclusão da variável *proxy* em uma regressão MQO elimina ou pelo menos reduz o viés. A dificuldade em aplicar este método é que variáveis *proxy* podem ser difíceis de encontrar. Uma possibilidade geral é usar dados de uma variável dependente de um ano anterior.

Os economistas que trabalham em áreas aplicadas estão frequentemente preocupados com os erros de medida. Sob as hipóteses do erro clássico nas variáveis (CEV), o erro de medida na variável dependente não tem efeito nas propriedades estatísticas do MQO. Por outro lado, sob as hipóteses CEV para uma variável independente, o estimador MQO do coeficiente na variável incorretamente medida é viesado em direção a zero. O viés nos coeficientes das outras variáveis pode ser para qualquer lado e é difícil de ser determinado.

Amostras não aleatórias de uma população subjacente podem levar a vieses no MQO. Quando a seleção da amostra está correlacionada com o termo de erro u, o MQO é geralmente viesado e inconsistente. Por outro lado, a seleção amostral exógena — que se baseia nas variáveis explicativas ou, ao contrário, é independente de u — não causa problemas para o MQO. Observações extremas em conjuntos de dados podem produzir grandes impactos nas estimativas MQO, especialmente em amostras pequenas. É importante, pelo menos informalmente, identificar as observações extremas e reestimar os modelos com as observações extremas suspeitas excluídas.

A estimação dos mínimos desvios absolutos é uma alternativa aos MQO que é menos sensível às observações extremas e que produz estimativas consistentes dos parâmetros das medianas condicionais.

PROBLEMAS

9.1 No Exercício 4.11 o R-quadrado da estimativa do modelo

$$\log(salário) = \beta_0 + \beta_1 \log(vendas) + \beta_2 \log(valmerc) + \beta_3 lucrmarg + \beta_4 perceo + \beta_5 percomp + u,$$

usando os dados contidos no arquivo CEOSAL2.RAW, era $R^2 = 0{,}353$ ($n = 177$). Quando $perceo^2$ e $percomp^2$ são adicionados, $R^2 = 0{,}375$. Existe evidência de má-especificação da forma funcional neste modelo?

9.2 Modifiquemos o Exercício em Computador 8.4, disponível no site da Cengage, utilizando os resultados das eleições de 1990 dos candidatos que foram eleitos em 1988. O candidato A foi eleito em 1988 e buscava a reeleição em 1990; *votoA90* é a percentagem de votos do Candidato A na eleição de 1990. A percentagem de votos do Candidato A na eleição de 1988 é usada como uma variável *proxy* da qualidade do candidato. Todas as demais variáveis são das eleições de 1990. As seguintes equações foram estimadas, utilizando os dados contidos no arquivo VOTE2.RAW:

$$\widehat{votoA90} = \underset{(9{,}25)}{75{,}71} + \underset{(0{,}046)}{0{,}312}\, forpartA + \underset{(1{,}01)}{4{,}93}\, democA - \underset{(0{,}684)}{0{,}929} \log(gastoA) - \underset{(0{,}281)}{1{,}950} \log(gastoB)$$
$$n = 186,\ R^2 = 0{,}495,\ \bar{R}^2 = 0{,}483,$$

e

$$\widehat{votoA90} = 70,81 + 0,282\,forpartA + 4,52\,democA$$
$$(10,01)\ \ (0,052)\qquad\quad (1,06)$$
$$-\,0,839\log(gastoA) - 1,846\log(gastoB) + 0,067\,votoA88$$
$$(0,687)\qquad\quad (0,292)\qquad\quad (0,053)$$
$$n = 186,\ R^2 = 0,499,\ \overline{R^2} = 0,485.$$

(i) Interprete o coeficiente de *votoA88* e comente sobre sua significância estatística.

(ii) A adição de *votoA88* tem muito efeito sobre os outros coeficientes?

9.3 Seja *mate10* a percentagem de aprovação em um teste-padrão de matemática de estudantes de uma escola secundária de Michigan (veja também o Exemplo 4.2). Estamos interessados em estimar o efeito do gasto por estudante no desempenho em matemática. Um modelo simples é

$$mate10 = \beta_0 + \beta_1\log(gasto) + \beta_2\log(matricl) + \beta_3 pobreza + u,$$

em que *pobreza* é a percentagem de estudantes vivendo em condições de pobreza.

(i) A variável *prgalm* é a percentagem de estudantes qualificados para o programa de merenda escolar financiado pelo governo federal. Por que ela é uma variável *proxy* razoável de *pobreza*?

(ii) A tabela seguinte contém estimativas MQO, com e sem *prgalm* como uma variável explicativa.

Variável dependente: *mate10*

Variáveis independentes	(1)	(2)
log(*gasto*)	11,13 (3,30)	7,75 (3,04)
log(*matricl*)	0,022 (0,615)	−1,26 (0,58)
prgalm	—	−0,324 (0,036)
intercepto	−69,24 (26,72)	−23,14 (24,99)
Observações *R*-quadrado	428 0,0297	428 0,1893

Explique por que o efeito dos gastos sobre *mate10* é menor na coluna (2) do que na coluna (1). O efeito na coluna (2) ainda é estatisticamente maior que zero?

(iii) Parece que as taxas de aprovação são menores em escolas maiores, com os outros fatores sendo iguais? Explique.

(iv) Interprete o coeficiente de *prgalm* na coluna (2).

(v) O que você deduz do substancial aumento de R^2 da coluna (1) para a coluna (2)?

9.4 A equação seguinte explica o número de horas por semana que uma criança passa assistindo televisão, em termos da idade da criança, educação da mãe, educação do pai e número de irmãos:

$$tvhoras^* = \beta_0 + \beta_1 idade + \beta_2 idade^2 + \beta_3 educm + \beta_4 educp + \beta_5 irms + u.$$

Estamos preocupados com a possibilidade de que *tvhoras** tenha sido medida com erro em nossa pesquisa. Seja *tvhoras* o número de horas por semana que se gasta assistindo televisão.

(i) O que as hipóteses do erro clássico nas variáveis (CEV) requerem nesta aplicação?

(ii) Você acha que as hipóteses CEV têm possibilidades de se manter? Explique.

9.5 No Exemplo 4.4 estimamos um modelo relacionando número de crimes no *campus* às matrículas de estudantes em um grupo de faculdades. A amostra que usamos não era uma amostra aleatória de faculdades nos Estados Unidos, pois muitas escolas em 1992 não registraram crimes no *campus*. Você acha que a falha das faculdades em informar os crimes pode ser vista como uma seleção amostral exógena? Explique.

9.6 No modelo (9.17), prove que os MQO estimam consistentemente α e β se a_i for não correlacionada com x_i e b_i for não correlacionada com x_i e x_i^2, que são hipóteses mais fracas que em (9.19). [*Sugestão:* escreva a equação como em (9.18) e recorde do Capítulo 5 que a suficiência para a consistência dos MQO do intercepto e da inclinação são $E(u_i) = 0$ e $Cov(x_i, u_i) = 0$.]

9.7 Considere o modelo de regressão simples com erro de medição clássico, $y = \beta_0 + \beta_1 x^* + u$ em que temos m medidas na x^*. Escreva como $z_h = x^* + e_h, h = 1,...,m$. Suponha que x^* é não correlacionada com $u, e_1, ..., e_m$, que os erros de medição são não correlacionados em pares, e têm a mesma variância, σ_e^2. Que $w = (z_1 + ... + z_m)/m$ seja a média das medidas na x^*, de forma que, de cada observação i, $w_i = (z_{i1} + ... + z_{im})/m$ será a média das medidas m. Que $\overline{\beta}_1$ seja o estimador MQO da regressão simples y_i sobre $1, w_i, i = 1,...,n$, usando uma amostra aleatória de dados.

(i) Prove que

$$\text{plim}(\overline{\beta}_1) = \beta_1\left\{\frac{\sigma_x^{2}*}{[\sigma_x^{2}* + (\sigma_e^2/m)]}\right\}.$$

[*Sugestão:* o plim de $\overline{\beta}_1$ é Cov(w, y)/Var(w).]

(ii) Como a inconsistência na $\overline{\beta}_1$ se compara com isso quando somente um único indicador está disponível (isto é, $m = 1$)? O que acontece à medida que m cresce? Comente.