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

17

Modelos com Variáveis Dependentes Limitadas e Correções da Seleção Amostral

No Capítulo 7, estudamos o modelo de probabilidade linear, que simplesmente é uma aplicação do modelo de regressão múltipla a uma variável dependente binária. Uma variável dependente binária é um exemplo de uma **variável dependente limitada (VDL)**. Uma VDL é definida, de modo geral, como uma variável dependente cujo intervalo de valores é substancialmente restrito. Uma variável binária assume somente dois valores, zero e um. Já vimos vários outros exemplos de variáveis dependentes limitadas: a porcentagem de participação em um plano de pensão deve estar entre zero e 100, o número de vezes em que uma pessoa é presa em determinado ano é um inteiro não negativo, e a nota média da graduação está entre zero e 4,0 em quase todas faculdades dos Estados Unidos.

A maioria das variáveis econômicas que gostaríamos de explicar é de alguma forma limitada, muitas vezes porque ela deve ser positiva. Por exemplo, o salário por hora, os preços de imóveis e as taxas nominais de juros devem ser maiores que zero. Mas nem todas as variáveis desse tipo precisam de tratamento especial. Se uma variável estritamente positiva assumir vários valores diferentes, raramente será necessário um modelo econométrico especial. Quando y for discreta e assumir um pequeno número de valores, não fará sentido tratá-la como uma variável aproximadamente contínua. A descontinuidade de y não significa, por si só, que os modelos lineares sejam inadequados. Porém, como vimos no Capítulo 7 sobre a resposta binária, o modelo de probabilidade linear tem certas desvantagens. Na Seção 17.1, discutiremos os modelos logit e probit, que compensam as desvantagens do MPL; a desvantagem é que eles são mais difíceis de ser interpretados.

Outros tipos de variáveis dependentes limitadas surgem na análise econométrica, especialmente quando estamos modelando o comportamento de indivíduos, famílias ou firmas. A otimização de comportamentos frequentemente leva a uma **resposta de solução de canto** para alguma fração relevante da população. Ou seja, uma quantidade ou valor em dólar zero, por exemplo, é uma escolha ótima. Durante qualquer determinado ano, um número significativo de famílias fará zero contribuições de caridade. Portanto, as contribuições de caridade familiares anuais têm uma distribuição populacional espalhada em uma ampla gama de valores positivos, mas com um acúmulo no valor zero. Embora um modelo linear possa ser apropriado para capturar os valores esperados de contribuições de caridade, muito provavelmente levará a previsões negativas para algumas famílias. Não será possível usar o log natural, pois muitas observações serão zero. O modelo Tobit, que abordaremos na Seção 17.2, é especificamente projetado para modelar variáveis dependentes que tenham soluções de canto.

Outro importante tipo de VDL é uma variável de contagem, que assume valores inteiros não negativos. A Seção 17.3 ilustra como os modelos de regressão de Poisson são bem apropriados para modelar variáveis de contagem.

Em alguns casos, observamos variáveis dependentes limitadas em razão da censura dos dados, um tópico que trataremos na Seção 17.4. O problema geral da seleção amostral, no qual observamos uma amostra não aleatória da população subjacente, é tratado na Seção 17.5.

Modelos de variáveis dependentes limitadas podem ser usados para séries temporais e dados em painel, mas são aplicados com mais frequência a dados em corte transversal. Problemas de seleção amostral geralmente estão restritos a dados em corte transversal ou em painel. Neste capítulo, concentramo-nos em aplicações de corte transversal. Wooldridge (2002) apresenta esses problemas no contexto de modelos de dados em painel e fornece muitos outros detalhes sobre aplicações de corte transversal e dados em painel.

17.1 MODELOS LOGIT E PROBIT DE RESPOSTA BINÁRIA

O modelo de probabilidade linear é fácil de ser estimado e usado, mas tem algumas desvantagens que discutimos na Seção 7.5 do Capítulo 7. As duas desvantagens mais relevantes são que as probabilidades ajustadas podem ser menores que zero ou maiores que um e o efeito parcial de qualquer variável explicativa (aparecendo na forma de nível) é constante. Essas limitações do MPL podem ser compensadas pelo uso de **modelos de resposta binária**.

Em um modelo de resposta binária, o interesse reside, principalmente, na **probabilidade de resposta**

$$P(y = 1|\mathbf{x}) = P(y = 1|x_1, x_2, \ldots, x_k), \tag{17.1}$$

em que usamos $\mathbf{x}$ para representar o conjunto completo de variáveis explicativas. Por exemplo, quando y for um indicador de emprego, $\mathbf{x}$ poderá conter várias características individuais, como educação, idade, estado civil e outros fatores que afetem a situação de emprego, inclusive uma variável binária indicadora da participação em um recente programa de treinamento de pessoal.

A Especificação de Modelos Logit e Probit

No MPL, assumimos que a probabilidade de resposta é linear em um conjunto de parâmetros, β_j; [veja a equação (7.27)]. Para evitar as limitações do MPL, considere uma classe de modelos de resposta binária da forma

$$P(y = 1|\mathbf{x}) = G(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) = G(\beta_0 + \mathbf{x}\boldsymbol{\beta}), \tag{17.2}$$

em que G é uma função assumindo valores estritamente entre zero e um: $0 < G(z) < 1$, para todos os números z reais. Isso garante que as probabilidades estimadas de resposta estejam estritamente entre zero e um. Como nos capítulos anteriores, escrevemos $\mathbf{x}\boldsymbol{\beta} = \beta_1 x_1 + \ldots + \beta_k x_k$.

Várias funções não lineares têm sido sugeridas para a função G para garantir que as probabilidades estejam entre zero e um. As duas que trataremos aqui são usadas na grande maioria das aplicações (juntamente com o MPL). No **modelo logit**, G é a função logística:

$$G(z) = \exp(z)/[1 + \exp(z)] = \Lambda(z), \tag{17.3}$$

que está entre zero e um para todos os números z reais. Essa é a função de distribuição cumulativa de uma variável aleatória logística padrão. No **modelo probit**, G é a função de distribuição cumulativa (fdc) normal padrão, que é expressa como uma integral:

$$G(z) = \Phi(z) \equiv \int_{-\infty}^{z} \phi(v)dv, \tag{17.4}$$

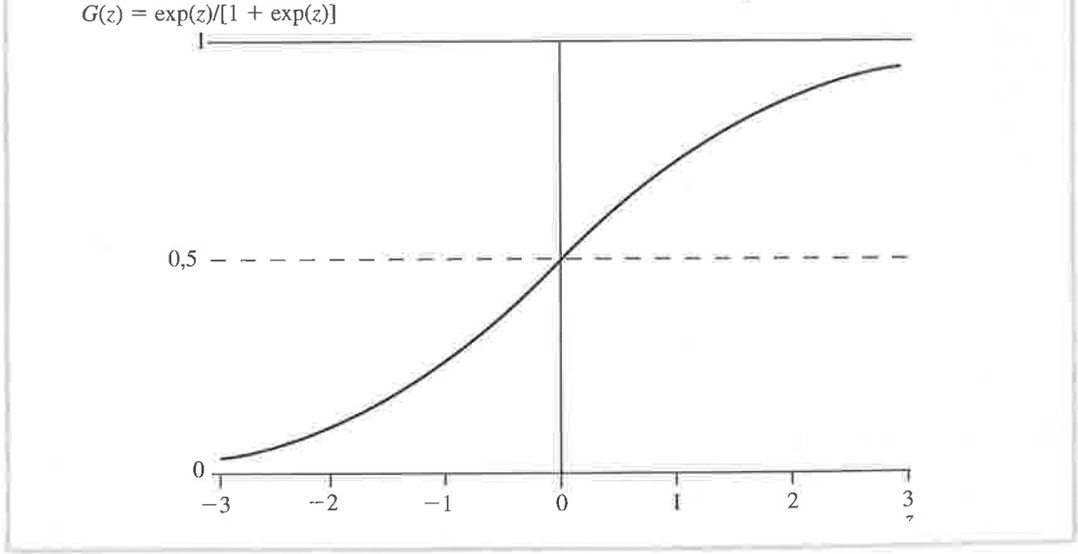
em que $\phi(z)$ é a densidade normal padrão

$$\phi(z) = (2\pi)^{-1/2}\exp(-z^2/2). \tag{17.5}$$

A escolha de G mais uma vez assegura que (17.2) esteja estritamente entre zero e um para todos os valores dos parâmetros e para x_j.

As funções G em (17.3) e (17.4) são ambas funções crescentes. Cada uma delas cresce mais rapidamente com $z = 0$, $G(z) \to 0$ quando $z \to -\infty$ e $G(z) \to 1$ quando $z \to \infty$. A função logística está representada na Figura 17.1. A fdc normal padrão tem uma forma muito semelhante à da fdc logística.

Figura 17.1
Gráfico da função logística $G(z) = \exp(z)/[1 + \exp(z)]$.



Os modelos logit e probit podem ser derivados de um **modelo de variável latente** subjacente. Seja y^* uma variável não observada, ou *latente*, determinada por

$$y^* = \beta_0 + \mathbf{x}\boldsymbol{\beta} + e, \; y = 1[y^* > 0], \tag{17.6}$$

em que introduzimos a notação $1[\cdot]$ para definir um resultado binário. A função $1[\cdot]$ é chamada de *função indicadora*, que assume o valor um se o evento entre colchetes for verdadeiro e zero, caso contrário. Portanto, y será um se $y^* > 0$, e zero se $y^* \leq 0$. Assumimos que e é independente de $\mathbf{x}$ e que e tem a distribuição logística padrão ou a distribuição normal padrão. Em qualquer caso, e será simetricamente distribuída ao redor de zero, o que significa que $1 - G(-z) = G(z)$ para todos os números z reais. Os economistas tendem a preferir a hipótese de normalidade de e, razão pela qual o modelo probit é mais popular que o logit em econometria. Além disso, vários problemas de especificação, sobre os quais comentaremos mais tarde, são muito mais facilmente analisados com o uso do probit em razão das propriedades da distribuição normal.

Com base em (17.6) e nas hipóteses dadas, podemos derivar a probabilidade de resposta de y:

$$\begin{aligned} \mathrm{P}(y = 1|\mathbf{x}) &= \mathrm{P}(y^* > 0|\mathbf{x}) = \mathrm{P}[e > -(\beta_0 + \mathbf{x}\boldsymbol{\beta})|\mathbf{x}] \\ &= 1 - G[-(\beta_0 + \mathbf{x}\boldsymbol{\beta})] = G(\beta_0 + \mathbf{x}\boldsymbol{\beta}), \end{aligned}$$

que é exatamente igual a (17.2).

Na maioria das aplicações de modelos de resposta binária, a meta principal é explicar os efeitos de x_j sobre a probabilidade de resposta $\mathrm{P}(y = 1|\mathbf{x})$. A formulação da variável latente tende a dar a impressão de que estamos interessados primeiramente nos efeitos de cada x_j sobre y^*. Como veremos, no logit e no probit, a *direção* do efeito de x_j sobre $\mathrm{E}(y^*|\mathbf{x}) = \beta_0 + \mathbf{x}\boldsymbol{\beta}$ e $\mathrm{E}(y|\mathbf{x}) = \mathrm{P}(y = 1|\mathbf{x}) = G(\beta_0 + \mathbf{x}\boldsymbol{\beta})$ é sempre a mesma. Contudo, a variável latente y^* raramente tem uma unidade de medida bem definida. (Por exemplo, y^* pode ser a diferença, em níveis de utilidade, de duas ações diferentes.) Assim, as magnitudes de cada β_j não são, em si mesmas, de grande valia (ao contrário do que ocorre no modelo de probabilidade linear). Para muitos propósitos, queremos estimar o efeito de x_j sobre a probabilidade de êxito $\mathrm{P}(y = 1|\mathbf{x})$, mas isso é complicado em razão da natureza não linear de $G(\cdot)$.

Para encontrarmos o efeito parcial de variáveis, aproximadamente contínuas, temos que confiar no cálculo. Se x_j for variável aproximadamente contínua, seu efeito parcial sobre $p(\mathbf{x}) = \mathrm{P}(y = 1|\mathbf{x})$ será obtido da derivada parcial:

$$\frac{\partial p(\mathbf{x})}{\partial x_j} = g(\beta_0 + \mathbf{x}\boldsymbol{\beta})\beta_j, \text{ em que } g(z) \equiv \frac{dG}{dz}(z). \qquad \textbf{(17.7)}$$

Como G é a fdc de uma variável aleatória contínua, g é uma função de densidade de probabilidade. Nos casos logit e probit, $G(\cdot)$ será uma fdc estritamente crescente, e assim $g(z) > 0$ para todo z. Portanto, o efeito parcial de x_j sobre $p(\mathbf{x})$ depende de $\mathbf{x}$ em razão da quantidade positiva $g(\beta_0 + \mathbf{x}\boldsymbol{\beta})$, e significa que o efeito parcial sempre terá o mesmo sinal de β_j.

A equação (17.7) mostra que os efeitos *relativos* de duas variáveis explicativas contínuas quaisquer não dependem de $\mathbf{x}$: a razão dos efeitos parciais de x_j e x_h é β_j/β_h. No caso típico em que g é uma densidade simétrica ao redor de zero, com uma única moda em zero, o maior efeito ocorre quando $\beta_0 + \mathbf{x}\boldsymbol{\beta} = 0$. Por exemplo, no caso probit com $g(z) = \phi(z)$, $g(0) = \phi(0) = 1/\sqrt{2\pi} \approx 0{,}40$. No caso logit, $g(z) = \exp(z)/[1 + \exp(z)]^2$ e, portanto, $g(0) = 0{,}25$.

Se, digamos, x_1 for uma variável explicativa binária, o efeito parcial de alterar x_1 de zero para um, mantendo-se todas as outras variáveis fixas, será simplesmente

$$G(\beta_0 + \beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k) - G(\beta_0 + \beta_2 x_2 + \ldots + \beta_k x_k). \qquad \textbf{(17.8)}$$

Mais uma vez, isso depende de todos os valores dos outros x_j. Por exemplo, se y for um indicador de emprego e x_1 for uma variável *dummy* indicando a participação em um programa de treinamento de pessoal, então (17.8) será a mudança na probabilidade do emprego em razão do programa de treinamento de pessoal; isso dependerá de outras características que afetem a empregabilidade, tais como a educação e a experiência. Observe que o conhecimento do sinal de β_1 será suficiente para determinar se o programa teve um efeito positivo ou negativo. Entretanto, para encontrar a *magnitude* do efeito, teremos que estimar a quantidade em (17.8).

Também podemos usar a diferença em (17.8) para outros tipos de variáveis discretas (como o número de filhos). Se x_k representar essa variável, o efeito sobre a probabilidade de x_k ir de c_k para $c_k + 1$ será simplesmente

$$\begin{aligned} &G[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k (c_k + 1)] \\ &- G(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k c_k). \end{aligned} \qquad \textbf{(17.9)}$$

A inclusão de formas funcionais padrão entre as variáveis explicativas é feita de forma direta. Por exemplo, no modelo

$$\mathrm{P}(y = 1|\mathbf{z}) = G(\beta_0 + \beta_1 z_1 + \beta_2 z_1^2 + \beta_3 \log(z_2) + \beta_4 z_3),$$

o efeito parcial de z_1 sobre $\mathrm{P}(y = 1|\mathbf{z})$ será $\partial \mathrm{P}(y = 1|\mathbf{z})/\partial z_1 = g(\beta_0 + \mathbf{x}\boldsymbol{\beta})(\beta_1 + 2\beta_2 z_1)$, e o efeito parcial de z_2 sobre a probabilidade de resposta será $\partial \mathrm{P}(y = 1|\mathbf{z})/\partial z_2 = g(\beta_0 + \mathbf{x}\boldsymbol{\beta})(\beta_3/z_2)$, em que $\mathbf{x}\boldsymbol{\beta} = \beta_1 z_1 + \beta_2 z_1^2 + \beta_3 \log(z_2) + \beta_4 z_3$. Portanto, $g(\beta_0 + \mathbf{x}\boldsymbol{\beta})(\beta_3/100)$ será a mudança aproximada na probabilidade de resposta quando z_2 é aumentado em 1%.

Algumas vezes queremos calcular a elasticidade da probabilidade de resposta com respeito a uma variável explicativa, embora devamos ser cuidadosos na interpretação das mudanças percentuais nas probabilidades. Por exemplo, uma alteração em uma probabilidade de 0,04 para 0,06 representa um aumento de 2 *pontos* percentuais na probabilidade, mas um aumento de 50% em relação ao valor inicial. Usando o cálculo infinitesimal, no modelo anterior a elasticidade de $\mathrm{P}(y = 1|\mathbf{z})$ com respeito à z_2 pode ser demonstrado como sendo $\beta_3 [g(\beta_0 + \mathbf{x}\boldsymbol{\beta})/G(\beta_0 + \mathbf{x}\boldsymbol{\beta})]$. A elasticidade com respeito à z_3 é $(\beta_4 z_3)[g(\beta_0 + \mathbf{x}\boldsymbol{\beta})/G(\beta_0 + \mathbf{x}\boldsymbol{\beta})]$. No primeiro caso, a elasticidade terá sempre o mesmo sinal da β_2, mas ela geralmente depende de todos os parâmetros e de todos os valores das variáveis explicativas. Se $z_3 > 0$, a segunda elasticidade sempre terá o mesmo sinal do parâmetro β_4.

Modelos com interações entre as variáveis explicativas podem ser um pouco ardilosos, mas deve-se calcular as derivadas parciais e então avaliar os efeitos parciais resultantes e valores de interesse. Quando estivermos medindo os efeitos de variáveis discretas — independentemente de quão complicado seja o modelo — devemos usar a (17.9). Discutiremos isto com mais detalhes mais adiante na subseção sobre interpretação de estimativas.

Estimação de Máxima Verossimilhança de Modelos Logit e Probit

Como devemos estimar modelos de resposta binária não linear? Para estimar o MPL, podemos usar mínimos quadrados ordinários (veja a Seção 7.5) ou, em alguns casos, mínimos quadrados ponderados (veja a Seção 8.5). Em razão da natureza não linear de $\mathrm{E}(y|\mathbf{x})$, MQO e MQP não são aplicáveis. Poderíamos usar versões não lineares desses métodos, mas o uso da **estimação de máxima verossimilhança (EMV)** não é mais complicada (veja uma discussão resumida no Apêndice 17A no final deste Capítulo). Até agora, precisamos muito pouco da EMV, embora tenhamos notado que, sob as hipóteses do modelo linear clássico, o estimador MQO é o estimador de máxima verossimilhança (condicional nas variáveis explicativas). Para estimar modelos de variáveis dependentes limitadas, os métodos de

máxima verossimilhança são indispensáveis. Como a estimação de máxima verossimilhança é baseada na distribuição de y dado $\mathbf{x}$, a heteroscedasticidade em $\text{Var}(y|\mathbf{x})$ é automaticamente considerada.

Assuma que temos uma amostra aleatória de tamanho n. Para obter o estimador de máxima verossimilhança, condicional nas variáveis explicativas, precisamos da densidade y_i dado $\mathbf{x}_i$. Podemos escrever isso como

$$f(y|\mathbf{x}_i;\boldsymbol{\beta}) = [G(\mathbf{x}_i\boldsymbol{\beta})]^y[1 - G(\mathbf{x}_i\boldsymbol{\beta}]^{1-y}, y = 0, 1, \quad \textbf{(17.10)}$$

em que, para simplificar, absorvemos o intercepto no vetor $\mathbf{x}_i$. Podemos facilmente ver que quando $y = 1$, obtemos $G(\mathbf{x}_i\boldsymbol{\beta})$ e quando $y = 0$, obtemos $1 - G(\mathbf{x}_i\boldsymbol{\beta})$. A **função log-verossimilhança** da observação i é uma função dos parâmetros e dos dados $(\mathbf{x}_i,y_i)$ e é obtida tomando o log de (17.10):

$$\ell_i(\boldsymbol{\beta}) = y_i\log[G(\mathbf{x}_i\boldsymbol{\beta})] + (1 - y_i)\log[1 - G(\mathbf{x}_i\boldsymbol{\beta})]. \quad \textbf{(17.11)}$$

Como $G(\cdot)$ está estritamente entre zero e um no logit e no probit, $\ell_i(\boldsymbol{\beta})$ será bem definido para todos os valores de $\boldsymbol{\beta}$.

A log-verossimilhança de uma amostra de tamanho n é obtida pela soma de (17.11) para todas as observações: $\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\ell_i(\boldsymbol{\beta})$. A EMV de $\boldsymbol{\beta}$, representada por $\hat{\boldsymbol{\beta}}$, maximiza essa log-verossimilhança. Se $G(\cdot)$ for a fdc logit padrão, $\hat{\boldsymbol{\beta}}$ será o *estimador logit*; se $G(\cdot)$ for a fdc normal padrão, então, $\hat{\boldsymbol{\beta}}$ será o *estimador probit*.

Devido à natureza não linear do problema de maximização, não podemos escrever fórmulas para as estimativas de máxima verossimilhança logit ou probit. Além de levantar questões computacionais, isso torna a teoria estatística do logit e do probit muito mais difícil do que o MQO ou mesmo o MQ2E. No entanto, a teoria geral da EMV de amostras aleatórias implica, sob condições bastante gerais, a EMV consistente, assimptoticamente normal e assimptoticamente eficiente. [Veja uma discussão geral sobre este tópico em Wooldridge (2002, Capítulo 13).] Aqui, usaremos apenas os resultados: a aplicação de modelos logit e probit é razoavelmente fácil, desde que entendamos o significado das estatísticas.

Cada $\hat{\beta}_j$ vem com um erro-padrão (assimptótico), cuja fórmula é complicada e é apresentada no apêndice deste capítulo. Uma vez que tenhamos os erros-padrão — e eles são descritos com as estimativas dos coeficientes por qualquer programa econométrico que suporte probit e logit —, poderemos construir testes t e intervalos de confiança (assimptóticos), exatamente como nos métodos MQO e MQ2E, e como para os outros estimadores que tenhamos encontrado. Particularmente, para testar H_0: $\beta_j = 0$, formamos a estatística $t = \hat{\beta}_j/\text{ep}(\hat{\beta}_j)$ e conduzimos o teste da maneira habitual, logo que tenhamos decidido sobre uma alternativa unilateral ou bilateral.

Testes de Hipóteses Múltiplas

Também podemos testar restrições múltiplas em modelos logit e probit. Na maioria dos casos, esses serão testes de múltiplas restrições de exclusões, como na Seção 4.5 do Capítulo 4. Aqui nos concentraremos nas restrições de exclusão.

Existem três maneiras de testarmos restrições de exclusão nos modelos logit e probit. O teste do multiplicador de Lagrange ou de estatística escore exige que apenas se estime o modelo sob a hipótese nula, da mesma forma como no caso linear na Seção 5.2; não trataremos aqui da estatística

escore, já que raramente é necessário testar as restrições de exclusão. [Veja Wooldridge (2002, Capítulo 15) sobre outros usos da estatística escore em modelos de resposta binária.]

O teste de Wald exige a estimação somente do modelo irrestrito. No caso do modelo linear, a **estatística de Wald**, após uma transformação simples, é essencialmente a estatística F, de modo que não há a necessidade de estudar a estatística de Wald separadamente. A fórmula da estatística de Wald é dada em Wooldridge (2002, Capítulo 15). Essa estatística é calculada por programas econométricos que permitem testar as restrições de exclusão após o modelo irrestrito ter sido estimado. Ela tem uma distribuição qui-quadrada assimptótica, com gl iguais ao número de restrições sendo testadas.

Se tanto o modelo restrito como o irrestrito forem fáceis de se estimar — como normalmente é o caso com restrições de exclusão —, então, o *teste da razão de verossimilhança* (*RV*) se torna bastante atraente. O teste *RV* baseia-se no mesmo conceito de o teste F em um modelo linear. O teste F mede o acréscimo na soma dos quadrados dos resíduos quando variáveis são retiradas do modelo. O teste RV baseia-se na diferença das funções log-verossimilhança dos modelos irrestrito e restrito. A ideia é a seguinte: como a EMV maximiza a função log-verossimilhança, a eliminação de variáveis geralmente conduz a uma log-verossimilhança *menor* — ou, pelo menos, não maior. (Isso é semelhante ao fato de o R-quadrado nunca aumentar quando variáveis são eliminadas de uma regressão.) A questão é se a queda na log-verossimilhança será suficientemente grande para concluirmos que as variáveis eliminadas são importantes. Poderemos tomar essa decisão logo que tenhamos uma estatística de teste e um conjunto de valores críticos.

A **estatística da razão de verossimilhança** é o *dobro* da diferença nas log-verossimilhanças:

$$RV = 2(\mathcal{L}_{ur} - \mathcal{L}_r), \quad \textbf{(17.12)}$$

em que $\mathcal{L}_{ur}$ é o valor de log-verossimilhança do modelo irrestrito e $\mathcal{L}_r$ é o valor de log-verossimilhança do modelo restrito. Como $\mathcal{L}_{ur} \geq \mathcal{L}_r$, *RV* será não negativa e usualmente estritamente positiva. Ao calcular a estatística *RV* de modelos de resposta binária, é importante saber que a função log-verossimilhança será sempre um número negativo. Esse fato advém da equação (17.11), porque y_i é ou zero ou um, e ambas as variáveis no interior da função log estão estritamente entre zero e um, o que significa que seus logs naturais são negativos. O fato de as funções log-verossimilhança serem ambas negativas não altera a maneira de calcularmos a estatística *RV*; simplesmente preservamos os sinais negativos na equação (17.12).

A multiplicação por dois em (17.12) é necessária para que a *RV* tenha uma distribuição qui-quadrada aproximada sob H_0. Se estivermos testando q restrições de exclusão, $RV \overset{a}{\sim} X_q^2$. Isso significa que, para testar H_0 ao nível de 5%, usamos como nosso valor crítico o 95º percentil na distribuição X_q^2. Calcular os p-valores é fácil na maioria dos programas econométricos.

QUESTÃO 17.1

Um modelo probit para explicar se uma firma será adquirida por outra durante determinado ano é

$$\begin{aligned}\text{P}(aquisição = 1|\mathbf{x}) = \Phi(&\beta_0 + \beta_1 lucrmed + \beta_2 valmerc \\ &+ \beta_3 dividareceita + \beta_4 perceo \\ &+ \beta_5 salceo + \beta_6 idadeceo),\end{aligned}$$

QUESTÃO 17.1 (continuação)

em que *aquisição* é uma variável de resposta binária, *lucrmed* é a média da margem de lucro da firma de vários anos anteriores, *valmerc* é o valor de mercado da firma, *dividareceita* é a razão dívida/receitas, e *perceo*, *salceo* e *idadeceo* são permanência, salário anual e idade do diretor-executivo (CEO), respectivamente. Estabeleça a hipótese nula que, outros fatores permanecendo iguais, as variáveis relacionadas com o CEO não têm efeito sobre a probabilidade de a firma ser adquirida por terceiros. Quantos *gl* existem na distribuição qui-quadrada do teste *RV* ou de Wald?

A Interpretação das Estimativas Logit e Probit

Considerando os modernos computadores, de uma perspectiva prática o aspecto mais difícil dos modelos logit e probit é apresentar e interpretar os resultados. As estimativas dos coeficientes, seus erros-padrão e o valor da função log-verossimilhança são descritos por todos os programas que executam logit e probit, e essas informações devem ser descritas em qualquer aplicação. Os coeficientes dão os sinais dos efeitos parciais de cada x_j sobre a probabilidade de resposta, e a significância estatística de x_j é determinada pela condição de podermos rejeitar H_0: $\beta_j = 0$ a um nível de significância suficientemente pequeno.

Como discutimos brevemente na Seção 7.5, para o modelo de probabilidade linear podemos calcular um indicador de qualidade de ajuste chamado **percentagem corretamente predita**. Como antes, definimos um preditor binário das y_i para ser um, se a probabilidade predita for pelo menos 0,5, e zero caso contrário. Matematicamente, $\tilde{y}_i = 1$ se $G(\hat{\beta}_0 + x_i\hat{\beta}) \geq 0{,}5$ e $\tilde{y}_i = 0$ se $G(\hat{\beta}_0 + x_i\hat{\beta}) < 0{,}5$. Dado $\{\tilde{y}_i: i = 1, 2, ..., n\}$, podemos ver o grau de perfeição com que $\tilde{y}_i$ prediz y_i ao longo de todas as observações. Existem quatro resultados possíveis em cada par $(y_i, \tilde{y}_i)$; quando ambas são zero ou ambas são um, fazemos a predição correta. Nos dois casos em que um dos pares é zero e o outro é um, fazemos a predição incorreta. A percentagem corretamente predita é a percentagem de vezes em que $\tilde{y}_i = y_i$.

Embora a percentagem corretamente predita seja útil como um indicador de qualidade de ajuste, ela pode ser enganosa. Em particular, será possível obter percentagens corretamente preditas bastante altas mesmo quando o resultado menos provável é muito pobremente predito. Por exemplo, suponha que $n = 200$, 160 observações têm $y_i = 0$, e, dessas 160 observações, 140 das $\tilde{y}_i$ também sejam zero (assim, predizemos corretamente 87,5% dos resultados zero). Mesmo se *nenhuma* das predições for correta quando $y_i = 1$, ainda assim prediremos corretamente 70% de todos os resultados (140/200 = 0,70). Muitas vezes esperamos ter alguma habilidade para predizer o resultado menos provável (como, por exemplo, se alguma pessoa é presa por cometer um crime) e assim devemos ser sinceros quanto à eficiência para predizer cada resultado. Portanto, faz sentido também calcular a percentagem corretamente predita de cada um dos resultados. O Problema 17.1 pede que você demonstre que a percentagem corretamente predita global é uma média ponderada de $\hat{q}_0$ (a percentagem corretamente predita de $y_i = 0$) e $\hat{q}_1$ (a percentagem corretamente predita de $y_i = 1$) em que os pesos são as frações de zeros e uns na amostra, respectivamente.

Algumas pessoas têm criticado a regra de predição que acabamos de descrever, pelo uso de um limite de 0,5, especialmente quando um dos resultados é improvável. Por exemplo, se $\bar{y}_i = 0{,}08$ (somente 8% de "êxitos" na amostra), pode ser que *nunca* seja predito $y_i = 1$, pois a probabilidade de êxito estimada nunca será maior que 0,5. Uma alternativa é usar a fração de êxitos na amostra como o limite — 0,08 no exemplo anterior. Em outras palavras, definimos $\tilde{y}_i = 1$ quando $G(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}) \geq 0{,}08$ e zero, caso contrário. O uso dessa regra certamente aumentará o número de êxitos preditos, mas não sem um custo: cometeremos necessariamente mais enganos — talvez muitos mais — na predição de zeros ("fracassos"). Em termos da percentagem corretamente predita global, poderemos ter um pior resultado do que se usarmos o limite de 0,5.

Uma terceira possibilidade é selecionar o limite de tal modo que a fração de $\tilde{y}_i = 1$ na amostra seja a mesma (ou muito próxima de) $\bar{y}$. Em outras palavras, procuramos por valores de limites τ, $0 < \tau < 1$, de tal forma que se definirmos $\tilde{y}_i = 1$ quando $G(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}) \geq \tau$, então $\sum_{i=1}^{n} \tilde{y}_i \approx \sum_{i=1}^{n} y_i$. (O processo de tentativa e erro necessário para encontrarmos o valor desejado de τ pode ser enfadonho, mas é viável. Em alguns casos, não será possível fazer que o número de êxitos preditos seja exatamente o mesmo que o número de êxitos na amostra). Agora, dado este conjunto de $\tilde{y}_i$, poderemos calcular a percentagem corretamente predita de cada um dos dois resultados como também a percentagem corretamente predita global.

Existem também vários indicadores de **pseudo *R*-quadrados** de resposta binária. McFadden (1974) sugere o indicador $1 - \mathcal{L}_{ur}/\mathcal{L}_o$, em que $\mathcal{L}_{ur}$ é a função log-verossimilhança do modelo estimado e $\mathcal{L}_o$ é a função log-verossimilhança no modelo com somente um intercepto. Por que esse indicador faz sentido? Lembre-se de que as funções log-verossimilhança são negativas e, portanto, $\mathcal{L}_{ur}/\mathcal{L}_o = |\mathcal{L}_{ur}|/|\mathcal{L}_o|$. Além disso, $|\mathcal{L}_{ur}| \leq |\mathcal{L}_o|$. Se as covariadas não tiverem poder explicativo, então, $\mathcal{L}_{ur}/\mathcal{L}_o = 1$, e o pseudo *R*-quadrado será zero, da mesma forma que o *R*-quadrado normal é zero em uma regressão linear quando as covariadas não têm poder explicativo. Em geral, $|\mathcal{L}_{ur}| \leq |\mathcal{L}_o|$, caso em que $1 - \mathcal{L}_{ur}/\mathcal{L}_o > 0$. Se $\mathcal{L}_{ur}$ fosse zero, o pseudo *R*-quadrado seria igual à unidade. Aliás, $\mathcal{L}_{ur}$ não pode atingir zero em um modelo probit ou logit, já que isso exigiria que as probabilidades estimadas quando $y_i = 1$ fossem todas a unidade e as probabilidades estimadas quando $y_i = 0$ fossem todas zero.

R-quadrados alternativos do probit e logit estão mais diretamente relacionados com o *R*-quadrado habitual da estimação por MQO de um modelo de probabilidade linear. Para o probit ou para o logit, defina $\hat{y}_i = G(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})$ serem as probabilidades estimadas. Como essas probabilidades são, também, estimativas de $E(y_i|\mathbf{x}_i)$, podemos basear um *R*-quadrado na proximidade de $\hat{y}_i$ com y_i. Uma possibilidade que é sugerida por uma análise padrão de regressão é calcular a correlação quadrada entre y_i e $\hat{y}_i$. Lembre-se de que, em uma estrutura de regressão linear, essa é uma maneira algebricamente equivalente de obter o *R*-quadrado habitual; veja a equação (3.29) do Capítulo 3. Portanto, podemos calcular um pseudo *R*-quadrado do probit e logit que seja diretamente comparável ao habitual *R*-quadrado da estimação de um modelo de probabilidade linear. Em qualquer caso, a qualidade de ajuste é menos importante que tentar obter estimativas convincentes dos efeitos *ceteris paribus* das variáveis explicativas.

Frequentemente, queremos estimar os efeitos das x_j sobre as probabilidades de resposta, $P(y = 1|\mathbf{x})$. Se x_j for (em linhas gerais) contínuo, então,

$$\Delta\hat{P}(y = 1|\mathbf{x}) \approx [g(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})\hat{\beta}_j]\Delta x_j, \qquad \textbf{(17.13)}$$

para "pequenas" alterações nas x_j. Assim, para $\Delta x_j = 1$, a alteração na probabilidade de êxito estimada será, aproximadamente, $g(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})\hat{\beta}_j$. Comparado com o modelo de probabilidade linear, o custo por se usar modelos probit e logit é que os efeitos parciais na equação (17.13) serão mais difíceis de resumir devido ao fato de que o fator de escala, $g(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})$ é dependente da $\mathbf{x}$ (isto é, de todas as variáveis explicativas). Uma possibilidade é integrar valores interessantes para as x_j –tais como médias, medianas, mínimos, máximos e quartis inferiores e superiores – e depois vermos como $g(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})$ se altera. Embora seja atraente, isto pode ser entediante e resultar em demasiada informação, mesmo se o número de variáveis explicativas for moderado.

Como um breve resumo para obter as magnitudes dos efeitos parciais, é conveniente ter-se um único fator de escala que possa ser usado para multiplicar cada $\hat{\beta}_j$ (ou pelo menos os coeficientes nas variáveis mais ou menos contínuas). Um método, usado comumente em pacotes econométricos que rotineiramente estima modelos probit e logit é substituir cada variável explicativa por suas médias amostrais. Em outras palavras, o fator de ajuste será

$$g(\hat{\beta}_0 + \bar{\mathbf{x}}\hat{\boldsymbol{\beta}}) = g(\hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \hat{\beta}_2\bar{x}_2 + \ldots \hat{\beta}_k\bar{x}_k), \qquad \textbf{(17.14)}$$

em que $g(\cdot)$ é a densidade normal padrão no caso probit e $g(z) = \exp(z)/[1 + \exp(z)]^2$ no caso logit. A ideia por traz da (17.14) é que, quando ela é multiplicada pelas , obtemos o efeito parcial das x_j da pessoa "média" na amostra. Assim, se multiplicarmos um coeficiente pela (17.14), geralmente obtemos o **efeito parcial na média (PEA)**.

Existem pelo menos dois problemas potenciais com o uso dos PEAs para sintetizarmos os efeitos parciais das variáveis explicativas. Primeiro, se algumas das variáveis explicativas forem discretas, suas médias não representarão ninguém na amostra (ou população, por sinal). Por exemplo, se x_1 = *feminino* e 47,5% da amostra for do sexo feminino, que sentido fará agregarmos na $\bar{x}_1 = 0{,}475$ para representar a pessoa "média"? Segundo, se uma variável explicativa contínua aparecer como uma função não linear — digamos, como um log natural ou em um quadrático — não ficará claro se queremos calcular a média da função não linear ou agregarmos a média na função não linear. Por exemplo, devemos usar $\overline{\log(vendas)}$ ou $\log(\overline{vendas})$ para representar o tamanho médio da empresa? Pacotes econométricos que calculam o fator de escala na (17.14) como padrão usam a primeira: o programa foi escrito para computar as médias dos regressores incluídos na estimação probit ou tobit.

Um método diferente de calcular um fator de escala contorna o problema de quais valores se devem agregar para as variáveis explicativas. Em vez disso, o segundo fator de escala resulta do cálculo da média dos efeitos parciais individuais ao longo da amostra, levando ao que algumas vezes é chamado de **efeito parcial médio (APE)**. Para uma variável explicativa contínua x_j, o efeito parcial médio será $n^{-1}\sum_{i=1}^{n}[g(\hat{\beta}_0 + x_i\hat{\beta})\hat{\boldsymbol{\beta}}_j] = \left[n^{-1}\sum_{i=1}^{n} g(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})\right]\hat{\beta}_j$. O termo multiplicador $\hat{\beta}_j$ age como um fator de escala:

$$n^{-1}\sum_{i=1}^{n} g(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}) \qquad \textbf{(17.15)}$$

A equação (17.15) é facilmente calculada após as estimações probit e logit, em que $g(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}) = \phi(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})$ no caso probit e $g(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}) = \exp(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})/[1 + \exp((\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})]^2$ no caso logit. Os dois fatores de escala são diferentes – e possivelmente muito diferentes – pois na (17.15) estamos usando a média da função não linear em lugar da função não linear da média [como na (17.14)].

Como ambos os fatores de escala que acabamos de descrever dependem da aproximação do cálculo integral na (17.13), nenhum deles faz sentido para variáveis explicativas discretas. Em vez disso é melhor usarmos a equação (17.9) para estimarmos diretamente a alteração na probabilidade. Para uma alteração na x_k de c_k para $c_k + 1$, o equivalente discreto do efeito parcial baseado na (17.14) será

$$G[\hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \ldots + \hat{\beta}_{k-1}\bar{x}_{k-1} + \hat{\beta}_k(c_k + 1)] - G(\hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \ldots + \hat{\beta}_{k-1}\bar{x}_{k-1} + \hat{\beta}_k c_k). \qquad \textbf{(17.16)}$$

em que G é a fdc normal padrão no caso probit e $G(z) = \exp(z)/[1 + \exp(z)]$ no caso logit. [Quanto à x_k binária, (17.16) é calculada rotineiramente por certos pacotes de econometria, como o Stata®.] O efeito parcial médio, que normalmente é mais comparável com as estimativas do MPL, é

$$n^{-1}\sum_{i=1}^{n}\{G[\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots \hat{\beta}_{l-1}x_{ik-1} + \hat{\beta}_k(c_k + 1)] - G(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_{k-1}x_{ik-1} + \hat{\beta}_k c_k)\}. \qquad \textbf{(17.17)}$$

A obtenção da expressão (17.17) tanto para o probit quanto para o logit é na realidade bastante simples. Primeiro, de cada observação estimamos a probabilidade de êxito dos dois valores selecionados da x_k, integrando os resultados reais das variáveis explicativas. (Assim, teremos n diferenças estimadas.) Depois, calculamos a média das diferenças nas probabilidades estimadas em todas as observações.

A expressão na (17.17) tem interpretação particularmente útil quando x_k é uma variável binária. A cada unidade i, estimamos a diferença predita na probabilidade que $y_i = 1$ quando $x_k = 1$ e $x_k = 0$, ou seja,

$$G(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_{k-1}x_{i,k-1} + \hat{\beta}_k) - G(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_{k-1}x_{i,k-1}).$$

A cada i, esta diferença é o efeito estimado da mudança na x_k de zero para um, quer a unidade i tivesse $x_{ik} = 1$ ou $x_{ik} = 0$. Por exemplo, se y for um indicador de desemprego (igual a um se a pessoa estiver empregada) após a participação em um programa de treinamento de pessoal, indicado por x_k então podemos estimar a diferença nas probabilidades de emprego de cada pessoa em ambos os estados do universo. Este *raciocínio contrafactual* é semelhante ao do Capítulo 16, que usamos para induzir modelos de equações simultâneas. O efeito estimado do programa de treinamento de pessoal na probabilidade de emprego é a média das diferenças em probabilidades estimadas. Como outro exemplo, suponha que y indica se uma família foi aprovada para fazer uma hipoteca, e x_k é um indicador binário de etnia (digamos, igual a um para os não brancos). Então, para cada família podemos estimar a diferença predita de ter aprovação para fazer uma hipoteca como uma função da renda, bens, risco de crédito, e assim por diante — que seriam elementos de $(x_{i1}, x_{i2}, \ldots, x_{i,k-1})$ — sob os dois cenários de que o chefe da família seja não branco *versus* branco. Esperançosamente, controlamos fatores suficientes de forma que fazer o nivelamento pela média das diferenças em probabilidades resulte numa estimativa convincente do efeito da etnia.

Em aplicações onde são empregados probit, logit e o MPL, faz sentido calcular os fatores de escala descritos acima para os probit e logit quando fazemos comparações dos efeitos parciais. Ainda assim, busca-se uma maneira mais rápida para se comparar as magnitudes das diferentes estimativas. Como mencionado anteriormente, para o probit $g(0) \approx 0{,}4$ e para o logit $g(0) = 0{,}25$. Assim, para tornarmos as magnitudes do probit e do logit mais ou menos comparáveis podemos multiplicar os coeficientes probit por $0{,}4/0{,}25 = 1{,}6$, ou podemos multiplicar as estimativas logit por 0,625. No MPL, $g(0)$ é efetivamente um, portanto as estimativas de inclinação logit podem ser divididas por 4 para torná-las comparáveis às estimativas MPL; as estimativas de inclinação probit podem ser divididas por 2,5 para torná-las comparáveis às estimativas MPL. Ainda assim, na maioria dos casos, queremos as comparações mais exatas obtidas com o uso dos fatores de escala na (17.15) do logit e probit

EXEMPLO 17.1

(Participação das Mulheres Casadas na Força de Trabalho)

Agora usamos os dados do arquivo MROZ.RAW para estimar o modelo de participação na força de trabalho do Exemplo 8.8 — veja também a Seção 7.5 — por logit e probit. Também descrevemos as estimativas do modelo de probabilidade linear do Exemplo 8.8, usando os erros-padrão robustos em relação à heteroscedasticidade. Os resultados, com os erros-padrão entre parênteses, são apresentados na Tabela 17.1.

EXEMPLO 17.1 (continuação)

Tabela 17.1

Estimativas MPL, Logit e Probit da participação na força de trabalho.

Variável dependente: *naft*			
Variáveis independentes	**MPL(MQO)**	**Logit(EMV)**	**Probit(EMV)**
nesprend	−0,0034 (0,0015)	−0,021 (0,008)	−0,012 (0,005)
educ	0,038 (0,007)	0,221 (0,043)	0,131 (0,025)
exper	0,039 (0,006)	0,206 (0,032)	0,123 (0,019)
$exper^2$	−0,0060 (0,00018)	−0,0032 (0,0010)	−0,0019 (0,0006)
idade	−0,016 (0,002)	−0,088 (0,015)	−0,053 (0,008)
crianmed6	−0,262 (0,032)	−1,443 (0,204)	−0,868 (0,119)
crianma6	0,013 (0,013)	0,060 (0,075)	0,036 (0,043)
constante	0,586 (0,151)	0,425 (0,860)	0,270 (0,509)
Percentagem Corretamente Prevista	73,4	73,6	73,4
Valor de Log-Verossimilhança	—	−401,77	−401,30
Pseudo *R*-Quadrado	0,264	0,220	0,221

As estimativas dos três modelos contam uma história consistente. Os sinais dos coeficientes são os mesmos em todos os modelos, e as mesmas variáveis são estatisticamente significantes em cada modelo. O pseudo *R*-quadrado do MPL é o mesmo *R*-quadrado usual descrito pelo MQO; no logit e probit, o pseudo *R*-quadrado é o indicador baseado nas log-verossimilhanças descritas anteriormente.

Como já enfatizamos anteriormente, as *magnitudes* das estimativas de coeficiente entre modelos não são diretamente comparáveis. Em vez disso, calculamos os fatores de escala nas equações (17.14) e (17.15). Se avaliarmos a função de densidade de probabilidade normal padrão $\theta(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \ldots + \hat{\beta}_k x_k)$ nas médias amostrais das variáveis explicativas (incluindo a média de $exper^2$, *crianmed6* e *crianma6*), o resultado será, aproximadamente, 0,391. Quando calculamos a (17.14) para o caso logit, obtemos, aproximadamente, 0,243. A razão destas, 0,391/0,243 ≈ 1,61, é muito próxima da simples regra prática para escalonar as estimativas probit para torná-las comparáveis às estimativas logit: multiplique as estimativas probit por 1,6.

EXEMPLO 17.1 (continuação)

Todavia, para comparar probit e logit com as estimativas MPL, é melhor usarmos a (17.15). Esses fatores de escala são em torno de 0,301 (probit) e 0,179 (logit). Por exemplo, o coeficiente do logit escalonado na *educ* está em torno de 0,179 (0,221) ≈ 0,040, e o coeficiente do probit escalonado na *educ* está em torno de 0,301(0,131) ≈ 0,039; ambos são notadamente próximos da estimativa MPL de 0,038. Mesmo na discreta variável *crianmed6*, os coeficientes probit e logit escalonados são semelhantes ao coeficiente MPL de −0,262. Eles são 0,179 (−1,443) ≈ −0,258 (logit) e 0,301(−0,868) ≈ −0,261 (probit).

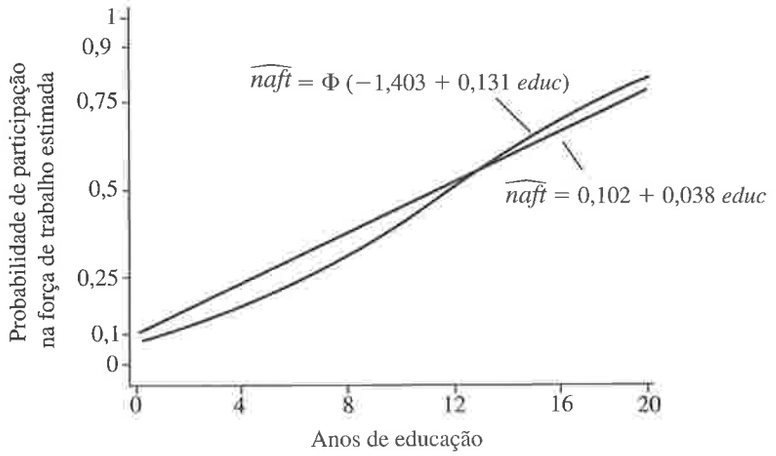
A maior diferença entre o MPL e os modelos logit e probit é que o MPL assume efeitos marginais constantes para *educ, crianmed6*, e assim por diante, enquanto os modelos logit e probit implicam magnitudes decrescentes dos efeitos parciais. No MPL, estima-se que uma criança a mais reduz a probabilidade de participação na força de trabalho em aproximadamente 0,262, independentemente de quantos filhos pequenos a mulher já tenha (e independentemente dos níveis das outras variáveis explicativas). Podemos contrastar isso com o efeito marginal estimado pelo probit. Em termos concretos, consideremos uma mulher com *nesprend* = 20,13, *educ* = 12,3, *exper* = 10,6 e *idade* = 42,5 — que são, *aproximadamente*, as médias da amostra — e *crianmed6* = 1. Qual será a redução estimada na probabilidade de trabalhar quando o número de crianças pequenas aumenta de zero para um? Avaliamos a fdc normal padrão, $\Phi(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k)$, com *crianmed6* = 1 e *crianmed6* = 0, e com as outras variáveis independentes definidas com os valores precedentes. Obtemos aproximadamente 0,373 − 0,707 = −0,334, o que significa que a probabilidade de participação na força de trabalho será cerca de 0,334 menor quando uma mulher tem um filho pequeno. Se uma mulher passa de uma para duas crianças pequenas, a probabilidade cai ainda mais, mas o efeito marginal não é tão grande: 0,117 − 0,373 = −0,256. Curiosamente, a estimativa do modelo de probabilidade linear, que pretensamente estima o efeito próximo da média, está, na realidade, entre essas duas estimativas.

QUESTÃO 17.2

Usando as estimativas probit e a aproximação por cálculo infinitesimal, qual será a alteração aproximada na probabilidade de resposta quando *exper* aumenta de 10 para 11?

A Figura 17.2 ilustra como as probabilidades de resposta estimadas a partir de modelos não lineares de respostas binárias podem diferir daquelas do modelo de probabilidade linear. A probabilidade estimada da participação na força de trabalho é traçada em relação aos anos de educação para o modelo de probabilidade linear e o modelo probit. (O gráfico do modelo logit é muito semelhante ao do modelo probit.) Em ambos os casos, as variáveis explicativas, outras que não *educ*, são definidas com o valor de suas médias amostrais. Especificamente, as duas equações traçadas são $\widehat{naft}$ = 0,102 + 0,038 *educ* para o modelo linear e $\widehat{naft} = \Phi(-1{,}403 + 0{,}131\ educ)$. Em níveis mais baixos de educação, o modelo de probabilidade linear estima probabilidades de participação na força de trabalho mais altas do que o modelo probit. Por exemplo, com oito anos de educação, o modelo de probabilidade linear estima uma probabilidade de participação na força de trabalho de 0,406, enquanto a estimativa do modelo probit é de aproximadamente 0,361. As estimativas são as mesmas ao redor de $11^1/_3$ anos de educação. Em níveis mais altos de educação, o modelo probit prevê probabilidades de participação na força de trabalho mais altas. Na amostra, o menor nível de educação é de cinco anos e o maior é de 17 anos, de modo que não devemos fazer comparações fora dessa faixa.

Figura 17.2
Probabilidade de resposta estimada em relação à educação para os modelos de probabilidade linear e probit.



Os mesmos problemas relativos às variáveis explicativas endógenas em modelos lineares também aparecem nos modelos logit e probit. Não temos espaço para tratar deles, mas é possível testar e corrigir variáveis explicativas endógenas com o uso de métodos relacionados aos mínimos quadrados em dois estágios. Evans e Schwab (1995) estimaram um modelo probit analisando se um aluno fazia curso superior, onde a variável explicativa principal era uma variável *dummy* que indicava se o aluno frequentava uma escola católica. Evans e Schwab estimaram um modelo por máxima verossimilhança que permitia que o fato de o aluno frequentar uma escola católica fosse considerado endógeno. [Veja Wooldridge (2002, Capítulo 15) para uma explicação desses métodos.]

Dois outros problemas receberam atenção no contexto dos modelos probit. O primeiro é a não normalidade de e no modelo de variável latente (17.6). Naturalmente, se e não tiver uma distribuição normal padrão, a probabilidade de resposta não terá a forma probit. Alguns autores tendem a enfatizar a inconsistência de estimar β_j, mas esse é o foco errado, a menos que estejamos interessados somente na direção dos efeitos. Como a probabilidade de resposta é desconhecida, não podemos estimar a magnitude dos efeitos parciais mesmo que tenhamos estimativas consistentes de β_j.

Um segundo problema de especificação, também definido em termos do modelo de variável latente, é a heteroscedasticidade em e. Se $\text{Var}(e|\mathbf{x})$ depender de $\mathbf{x}$, a probabilidade de resposta não mais terá a forma $G(\beta_0 + \mathbf{x}\boldsymbol{\beta})$; ao contrário, ela dependerá da forma da variância e exigirá estimação mais genérica. Tais modelos, na prática, não são usados com muita frequência, já que o logit e o probit com formas funcionais flexíveis nas variáveis independentes tendem a funcionar bem.

Modelos de resposta binária aplicam-se com pequenas modificações a cortes transversais agrupados independentemente ou a outros conjuntos de dados nos quais as observações são independentes, mas não necessariamente identicamente distribuídas. Muitas vezes, variáveis *dummy* anuais ou de outro período de tempo são incluídas para avaliar efeitos temporais agregados. Da mesma forma que com os modelos lineares, o logit e o probit podem ser usados para avaliar o impacto de certas decisões políticas no contexto de uma experimentação natural.

O modelo de probabilidade linear pode ser aplicado com dados em painel; em geral, ele será estimado por efeitos fixos (veja o Capítulo 14). Recentemente, modelos logit e probit com efeitos

não observados estão se tornando populares. Esses modelos são complicados pela natureza não linear das probabilidades de resposta, e são difíceis de estimar e interpretar. [Veja Wooldridge (2002, Capítulo 15).]

17.2 O MODELO TOBIT PARA RESPOSTA DE SOLUÇÃO DE CANTO

Conforme mencionado no capítulo introdutório, outro tipo importante de variável dependente limitada é uma resposta de solução de canto. Esse tipo de variável é zero para uma fração não desprezível da população, mas é aproximadamente distribuída de forma contínua ao longo de valores positivos. Um exemplo é o valor gasto por um indivíduo com bebida alcoólica em determinado mês. Na população de pessoas com mais de 21 anos dos Estados Unidos, essa variável assume uma ampla gama de valores. Para certa fração significante, o montante gasto com álcool é zero. A seguinte abordagem omite a verificação de alguns detalhes relacionados ao modelo Tobit. [Eles são dados em Wooldridge (2002, Capítulo 16).]

Seja y uma variável essencialmente contínua ao longo de valores estritamente positivos, mas que assuma zero com probabilidade positiva. Nada nos impede de usar um modelo linear para y. Na verdade, um modelo linear pode ser uma boa aproximação de $\text{E}(y|x_1, x_2, \ldots, x_k)$, especialmente para x_j próximos dos valores médios. Mas provavelmente obteremos valores estimados negativos, o que conduz a previsões negativas de y; isso é parecido com os problemas do MPL de resultados binários. Além disso, a hipótese de que uma variável explicativa que apareça em forma de nível tenha um efeito parcial constante sobre $\text{E}(y|\mathbf{x})$ pode ser enganosa. Provavelmente, $\text{Var}(y|\mathbf{x})$ será heteroscedástica, embora possamos facilmente lidar com a heteroscedasticidade generalizada calculando erros-padrão e estatísticas de testes robustos. Como a distribuição de y se acumula em zero, claramente y não pode ter uma distribuição normal condicional. Portanto, toda a inferência terá somente justificativa assimptótica, como acontece com o modelo de probabilidade linear.

Em alguns casos, é importante ter um modelo que implique valores previstos não negativos de y, e que tenha efeitos parciais sensíveis sobre uma ampla faixa das variáveis explicativas. Mais ainda, algumas vezes queremos estimar características da distribuição de y, dados outros $x_1, \ldots, x_k$ além do valor esperado condicional. O **modelo Tobit** é bastante conveniente para esses propósitos. Em geral, o modelo Tobit expressa a resposta observada, y, em termos de uma variável latente subjacente:

$$y^* = \beta_0 + \mathbf{x}\boldsymbol{\beta} + u,\ u|\mathbf{x} \sim \text{Normal}(0, \sigma^2) \qquad \textbf{(17.18)}$$

$$y = \max(0, y^*). \qquad \textbf{(17.19)}$$

A variável latente y^* satisfaz as hipóteses do modelo linear clássico; em particular, ela tem uma distribuição normal, homoscedástica, com uma média condicional linear. A equação (17.19) indica que a variável observada, y, será igual a y^* quando $y^* \geq 0$, mas $y = 0$ quando $y^* < 0$. Como y^* é normalmente distribuída, y terá uma distribuição contínua sobre valores estritamente positivos. Em particular, a densidade de y, dado $\mathbf{x}$ será igual à densidade de y^* dado $\mathbf{x}$ para valores positivos. Além disso,

$$\text{P}(y = 0|\mathbf{x}) = \text{P}(y^* < 0|\mathbf{x}) = \text{P}(u < -\mathbf{x}\boldsymbol{\beta}|\mathbf{x})$$

$$= \text{P}(u/\sigma < -\mathbf{x}\boldsymbol{\beta}/\sigma|\mathbf{x}) = \Phi(-\mathbf{x}\boldsymbol{\beta}/\sigma) = 1 - \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma),$$

em razão do fato de u/σ ter distribuição normal padrão e ser independente de $\mathbf{x}$; absorvemos o intercepto em $\mathbf{x}$ por simplicidade de notação. Portanto, se $(\mathbf{x}_i, y_i)$ for retirada aleatoriamente da população, a densidade de y_i, dado $\mathbf{x}_i$ será

$$(2\pi\sigma^2)^{-1/2}\exp[-(y - \mathbf{x}_i\boldsymbol{\beta})^2/(2\sigma^2)] = (1/\sigma)\phi[(y - \mathbf{x}_i\boldsymbol{\beta})/\sigma],\ y > 0 \qquad \textbf{(17.20)}$$

$$P(y_i = 0|\mathbf{x}_i) = 1 - \Phi(\mathbf{x}_i\boldsymbol{\beta}/\sigma), \qquad \textbf{(17.21)}$$

em que ϕ é a função densidade normal padrão.

De (17.20) e (17.21), podemos obter a função log-verossimilhança de cada observação i:

$$\ell_i(\boldsymbol{\beta},\sigma) = 1(y_i = 0)\log[1 - \Phi(\mathbf{x}_i\boldsymbol{\beta}/\sigma)] + 1(y_i > 0)\log\{(1/\sigma)\phi[(y_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\}; \qquad \textbf{(17.22)}$$

observe como essa função depende de σ, o desvio-padrão de u, como também de β_j. A log-verossimilhança de uma amostra aleatória de tamanho n é obtida somando (17.22) ao longo de todas as observações i. As estimativas de máxima verossimilhança de $\boldsymbol{\beta}$ e σ são obtidas pela maximização da log-verossimilhança; exigirá métodos numéricos, embora na maioria dos casos isso seja facilmente feito usando rotina de um programa econométrico.

QUESTÃO 17.3

Seja y o número de casos extraconjugais de uma mulher casada da população dos Estados Unidos; gostaríamos de explicar essa variável em termos de outras características da mulher — em especial se ela trabalha fora de casa — assim como de seu marido e sua família. Esse exemplo seria um bom candidato para um modelo Tobit?

Como nos casos logit e probit, cada estimativa Tobit é acompanhada de um erro-padrão, e isso pode ser usado para construir estatísticas t de cada $\hat{\beta}_j$; a forma matricial usada para encontrar os erros-padrão é complicada e não será apresentada aqui. [Veja, por exemplo, Wooldridge (2002, Capítulo 16).]

O teste de múltiplas restrições de exclusão é feito facilmente com o uso do teste de Wald ou do teste da razão de verossimilhança. O teste de Wald tem uma forma semelhante às dos casos logit e probit; o teste RV é constante em (17.12), no qual, é claro, usamos as funções log-verossimilhança Tobit dos modelos com e sem restrições.

A Interpretação das Estimativas Tobit

Com o uso de computadores modernos, as estimativas de máxima verossimilhança dos modelos Tobit usualmente não são muito mais difíceis de serem obtidas do que as estimativas MQO de um modelo linear. Além disso, os resultados do Tobit e do MQO são, muitas vezes, semelhantes. Isso torna tentador interpretar os $\hat{\beta}_j$ do Tobit como se fossem estimativas de uma regressão linear. Infelizmente, as coisas não são tão fáceis.

A partir da equação (17.18), vemos que os $\hat{\beta}_j$ medem os efeitos parciais dos x_j sobre $E(y^*|\mathbf{x})$, em que y^* é a variável latente. Algumas vezes, y^* tem um significado econômico interessante, mas, na

maioria das vezes, não. A variável que queremos explicar é y, já que ela é o resultado observado (tal como horas trabalhadas ou montante de contribuições de caridade). Por exemplo, em uma questão de critério de decisão, estamos interessados na sensibilidade das horas trabalhadas quanto a mudanças na alíquota de um imposto.

Podemos estimar $P(y = 0|\mathbf{x})$ a partir de (17.21), a qual, naturalmente, permite estimar $P(y > 0|\mathbf{x})$. O que acontecerá se quisermos estimar o valor esperado de y como uma função de $\mathbf{x}$? Em modelos Tobit, dois valores esperados são de especial interesse: $E(y|y > 0,\mathbf{x})$, que algumas vezes é chamada de "esperança condicional" por ser condicional a $y > 0$, e $E(y|\mathbf{x})$, que, infelizmente, é chamado de "esperança incondicional". (Ambos os valores esperados são condicionais às variáveis explicativas.) A expectativa $E(y|y > 0,\mathbf{x})$ nos informa, para determinados valores de $\mathbf{x}$, o valor esperado de y da subpopulação em que y é positivo. Dado $E(y|y > 0,\mathbf{x})$, podemos facilmente encontrar $E(y|\mathbf{x})$:

$$E(y|\mathbf{x}) = P(y > 0|\mathbf{x})\cdot E(y|y > 0,\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)\cdot E(y|y > 0,\mathbf{x}). \qquad \textbf{(17.23)}$$

Para obter $E(y|y > 0,\mathbf{x})$, usamos um resultado das variáveis aleatórias normalmente distribuídas: se $z \sim$ Normal(0,1), então, $E(z|z > c) = \phi(c)/[1 - \Phi(c)]$ para qualquer constante c. Mas $E(y|y > 0,\mathbf{x}) = \mathbf{x}\boldsymbol{\beta} + E(u|u > -\mathbf{x}\boldsymbol{\beta}) = \mathbf{x}\boldsymbol{\beta} + \sigma E[(u/\sigma)|(u/\sigma) > -\mathbf{x}\boldsymbol{\beta}/\sigma] = \mathbf{x}\boldsymbol{\beta} + \sigma\,\phi(\mathbf{x}\boldsymbol{\beta}/\sigma)/\Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)$, porque $\phi(-c) = \phi(c)$, $1 - \Phi(-c) = \Phi(c)$ e u/σ tem uma distribuição normal padrão independente de $\mathbf{x}$.

Podemos resumir isso como

$$E(y|y > 0,\mathbf{x}) = \mathbf{x}\boldsymbol{\beta} + \sigma\lambda(\mathbf{x}\boldsymbol{\beta}/\sigma), \qquad \textbf{(17.24)}$$

em que $\lambda(c) = \phi(c)/\Phi(c)$ é chamado de **razão inversa de Mills**; essa é a razão entre a fdp normal padrão e a fdc normal padrão, cada uma avaliada em c.

A equação (17.24) é importante. Ela mostra que o valor esperado de y condicional a $y > 0$ é igual a $\mathbf{x}\boldsymbol{\beta}$, mais um termo estritamente positivo, que é σ vezes a razão de Mills inversa avaliada em $\mathbf{x}\boldsymbol{\beta}/\sigma$. Essa equação também mostra porque o uso do MQO somente para observações nas quais $y_i > 0$ nem sempre estimará $\boldsymbol{\beta}$ consistentemente; essencialmente, a razão de Mills inversa é uma variável omitida, e geralmente ela é correlacionada com os elementos de $\mathbf{x}$.

A combinação da (17.23) com (17.24) produz

$$E(y|\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)\{\mathbf{x}\boldsymbol{\beta} + \sigma\lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)] = \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)\mathbf{x}\boldsymbol{\beta} + \sigma\phi(\mathbf{x}\boldsymbol{\beta}/\sigma), \qquad \textbf{(17.25)}$$

em que a segunda igualdade decorre porque $\Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)\lambda(\mathbf{x}\boldsymbol{\beta}/\sigma) = \phi(\mathbf{x}\boldsymbol{\beta}/\sigma)$. Essa equação mostra que, quando y segue um modelo Tobit, $E(y|\mathbf{x})$ será uma função não linear de $\mathbf{x}$ *e* $\boldsymbol{\beta}$. Embora não seja óbvio, pode ser mostrado que o lado direito da equação (17.25) será positivo para quaisquer valores de $\mathbf{x}$ e de $\boldsymbol{\beta}$. Portanto, logo que tivermos as estimativas de $\boldsymbol{\beta}$, podemos ter certeza de que os valores previstos de y — isto é, estimativas de $E(y|\mathbf{x})$ — são positivos. O custo de garantir previsões positivas de y é que a equação (17.25) é mais complicada que um modelo linear de $E(y|\mathbf{x})$. Mais importante ainda, os efeitos parciais de (17.25) são mais complicados do que os de um modelo linear. Como veremos, os efeitos parciais de x_j sobre $E(y|y > 0,\mathbf{x})$ e sobre $E(y|\mathbf{x})$ têm o mesmo sinal do coeficiente, β_j, mas a magnitude dos efeitos depende de *todos* os valores das variáveis explicativas e dos parâmetros. Como σ aparece em (17.25), não é de surpreender que os efeitos parciais também dependam de σ.

Se x_j for uma variável contínua, poderemos encontrar os efeitos parciais usando cálculo infinitesimal. Primeiro,

$$\partial \mathrm{E}(y|y > 0, \mathbf{x})/\partial x_j = \beta_j + \beta_j \cdot \frac{d\lambda}{dc}(\mathbf{x}\boldsymbol{\beta}/\sigma),$$

assumindo que x_j não seja funcionalmente relacionado a outros regressores. Tirando a diferença de $\lambda(c) = \phi(c)/\Phi(c)$ e usando $d\Phi/dc = \phi(c)$ e $d\phi/dc = -c\phi(c)$, pode ser demonstrado que $d\lambda/dc = -\lambda(c)[c + \lambda(c)]$. Portanto,

$$\partial \mathrm{E}(y|y > 0, \mathbf{x})/\partial x_j = \beta_j\{1 - \lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)[\mathbf{x}\boldsymbol{\beta}/\sigma + \lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)]\}. \qquad \textbf{(17.26)}$$

Isso mostra que o efeito parcial de x_j sobre $\mathrm{E}(y|y > 0,\mathbf{x})$ não é determinado apenas por β_j. O fator de ajuste é dado pelo termo entre chaves $\{\cdot\}$, e depende de uma função linear de $\mathbf{x}$, $\mathbf{x}\boldsymbol{\beta}/\sigma = (\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)/\sigma$. Pode ser mostrado que o fator de ajuste está estritamente entre zero e um. Na prática, podemos estimar (17.26) inserindo as EMVs de β_j e σ. Como com os modelos logit e probit, devemos inserir valores de x_j, usualmente os valores médios ou outros valores interessantes. A equação (17.26) revela um ponto sutil que algumas vezes é perdido na aplicação do modelo Tobit em respostas de solução de canto: o parâmetro σ aparece diretamente nos efeitos parciais, e assim ter uma estimativa de σ é crucial para estimar os efeitos parciais. Algumas vezes, σ é chamado de um parâmetro "ancilar" (e significa que ele é auxiliar, ou sem importância). Embora seja verdade que o valor de σ não afeta o sinal dos efeitos parciais, ele afeta as magnitudes, e frequentemente estamos interessados na importância econômica das variáveis explicativas. Portanto, caracterizar σ como ancilar é equivocado e advém de uma confusão entre o modelo Tobit de aplicações de solução de canto e aplicações de censura de dados. (Veja a Seção 17.4.)

Todas as quantidades econômicas habituais, como as elasticidades, podem ser calculadas. Por exemplo, a elasticidade de y em relação a x_1, condicional a $y > 0$, é

$$\frac{\partial \mathrm{E}(y|y > 0, \mathbf{x})}{\partial x_1} \cdot \frac{x_1}{\mathrm{E}(y|y > 0, \mathbf{x})}. \qquad \textbf{(17.27)}$$

Essa equação pode ser calculada quando x_1 aparece em várias formas funcionais, inclusive nas formas em nível, logarítmica e quadrática.

Se x_1 for uma variável binária, o efeito de interesse é obtido como a diferença entre $\mathrm{E}(y|y > 0,\mathbf{x})$, com $x_1 = 1$ e $x_1 = 0$. Efeitos parciais que envolvam outras variáveis discretas (como o número de filhos) podem ser tratados de maneira semelhante.

Podemos usar (17.25) para encontrar a derivada parcial de $\mathrm{E}(y|\mathbf{x})$ em relação a x_j contínua. Essa derivada leva em conta o fato de que as pessoas que iniciam em $y = 0$ podem escolher $y > 0$ quando x_j muda:

$$\frac{\partial \mathrm{E}(y|\mathbf{x})}{\partial x_j} = \frac{\partial \mathrm{P}(y > 0|\mathbf{x})}{\partial x_j} \cdot \mathrm{E}(y|y > 0, \mathbf{x}) + \mathrm{P}(y > 0|\mathbf{x}) \cdot \frac{\partial \mathrm{E}(y|y > 0, \mathbf{x})}{\partial x_j}. \qquad \textbf{(17.28)}$$

Como $\mathrm{P}(y > 0|\boldsymbol{x}) = \Phi(\boldsymbol{x}\boldsymbol{\beta}/\sigma)$,

$$\frac{\partial P(y > 0|\mathbf{x})}{\partial x_j} = (\beta_j/\sigma)\phi(\mathbf{x}\,\beta/\sigma), \qquad \textbf{(17.29)}$$

e, dessa forma, podemos estimar cada termo em (17.28), assim que inserirmos as EMVs de β_j e σ e valores particulares de x_j.

É importante observar que ao inserir (17.26) e (17.29) em (17.28) e usar o fato de que $\Phi(c)\lambda(c) = \phi(c)$ para qualquer c, obtém- se

$$\frac{\partial \mathrm{E}(y|\mathbf{x})}{\partial x_j} = \beta_j \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma). \qquad \textbf{(17.30)}$$

A equação (17.30) permite fazer comparações aproximadas entre as estimativas MQO e Tobit. [A equação (17.30) também pode ser derivada diretamente da equação (17.25) usando o fato de que $d\phi(z)/dz = -z\phi(z)$.] Os coeficientes de inclinação MQO, digamos $\hat{\gamma}_j$, da regressão de y_i sobre $x_{i1}, x_{i2}, \ldots, x_{ik}$, $i = 1, \ldots, n$ — isto é, usando todos os dados — são estimativas diretas de $\partial \mathrm{E}(y|\mathbf{x})/\partial x_j$. Para tornar o coeficiente Tobit, $\hat{\beta}_j$, comparável com $\hat{\gamma}_j$, devemos multiplicar $\hat{\beta}_j$ por um fator de ajuste.

Como nos casos probit e logit, existem dois métodos para calcular-se um fator de ajuste para a obtenção de efeitos parciais — pelo menos de variáveis explicativas contínuas. Ambos são na equação (17.30). Primeiro, o efeito parcial na média, PEA é obtido pela avaliação de $\Phi(\mathbf{x}\hat{\boldsymbol{\beta}}/\hat{\sigma})$, que denotamos $\Phi(\bar{\mathbf{x}}\hat{\boldsymbol{\beta}}/\hat{\sigma})$. Podemos usar este único fator para multiplicarmos os coeficientes nas variáveis explicativas contínuas. O PEA tem as mesmas inconveniências neste caso como nos casos logit e probit: não podemos ter interesse no efeito parcial da "média", pois a média ou é desinteressante ou é insignificante. E mais, temos de decidir se usaremos médias de funções não lineares ou se agregaremos as médias às funções não lineares.

O APE é preferido na maioria dos casos. Aqui, calculamos o fator de escalonamento como $n^{-1}\sum_{i=1}^{n} \Phi(\mathbf{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma})$. Diferentemente do PAE, o APE não exige que agreguemos uma unidade fictícia ou inexistente da população, e não há decisões a serem feitas sobre a agregação de médias nas funções não lineares. Como no PAE, o fator de escalonamento do APE estará sempre entre zero e um, pois $0 < \Phi(\mathbf{x}\hat{\boldsymbol{\beta}}/\hat{\sigma}) < 1$, para quaisquer valores das variáveis explicativas. Na verdade, $\hat{\mathrm{P}}(y_i > 0|x_i) = \Phi(\mathbf{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma}) < 1$ e assim o fator de escalonamento do APE e o do PAE tendem a ser mais próximos de um quando existem poucas observações com $y_i = 0$. No caso em que $y_i > 0$ de todas as i, as estimativas dos parâmetros pelo Tobit e pelos MQO são idênticas. [É claro, se $y_i > 0$ de todas as i, não podemos, de qualquer forma, justificar o uso de um modelo Tobit. O uso de $\log(y_i)$ num modelo de regressão linear faz muito mais sentido.]

Infelizmente, para variáveis explicativas discretas, comparar as estimativas Tobit e MQO não é tão fácil (embora o uso do fator de escalonamento de variáveis explicativas contínuas seja sempre uma aproximação útil.). Do Tobit, o efeito parcial das variáveis explicativas contínuas, por exemplo, uma variável binária, deve realmente ser obtida estimando-se $\mathrm{E}(y|\mathbf{x})$ da equação (17.25). Por exemplo, se x_1 for uma variável binária, devemos primeiro agregar $x_1 = 1$ e depois, $x_1 = 0$. Se especificarmos as outras variáveis explicativas em suas médias amostrais, obteremos um indicador análogo ao da (17.16) para os casos probit e logit. Se calcularmos a diferença nos valores esperados de cada indivíduo, e depois nivelarmos a diferença, teremos um APE análogo ao da (17.17).

EXEMPLO 17.2

(Oferta de Mão de Obra Anual de Mulheres Casadas)

O arquivo MROZ.RAW inclui dados sobre horas trabalhadas de 753 mulheres casadas, 428 das quais trabalharam fora de casa por um salário durante o ano; 325 mulheres trabalharam zero horas. Para as mulheres que trabalharam horas positivas, a faixa é bastante ampla, de 12 a 4.950. Assim, horas anuais trabalhadas é uma boa candidata a modelo Tobit. Também estimamos um modelo linear (usando todas as 753 observações) por MQO. Os resultados estão na Tabela 17.2

Tabela 17.2

Estimação MQO e Tobit de horas anuais trabalhadas.

Variável dependente: *horas*		
Variáveis independentes	**Linear (MQO)**	**Tobit (EMV)**
nesprend	−3,45 (2,54)	−8,81 (4,46)
educ	28,76 (12,95)	80,65 (21,58)
exper	65,67 (9,96)	131,56 (17,28)
$exper^2$	−0,700 (0,325)	−1,86 (0,54)
idade	−30,51 (4,36)	−54,41 (7,42)
crianmed6	−442,09 (58,85)	−894,02 (111,88)
crianma6	32,78 (23,18)	−16,22 (38,64)
constante	1.330,48 (270,78)	965,31 (446,44)
Valor de Log-Verossimilhança *R*-Quadrado $\hat{\sigma}$	——— 0,266 750,18	−3.819,09 0,274 1.122,02

Essa tabela tem várias características às quais devemos atentar. Primeiro, as estimativas dos coeficientes Tobit têm o mesmo sinal das correspondentes estimativas MQO, e a significância estatística das estimativas são semelhantes. (Possíveis exceções são os coeficientes de *nesprend* e *crianma6*, mas as estatísticas *t* têm magnitudes semelhantes.) Segundo, embora seja tentador comparar as magnitudes das estimativas

EXEMPLO 17.2 (continuação)

MQO e Tobit, isso não será muito informativo. Devemos ter cuidado para não pensar que, como o coeficiente Tobit de *crianmed6* é aproximadamente o dobro do coeficiente MQO, o modelo Tobit indica uma resposta muito maior de horas trabalhadas em relação a crianças pequenas.

Podemos multiplicar as estimativas Tobit por fatores de ajuste apropriados para torná-las, grosso modo, comparáveis com as estimativas MQO. O fator de escalonamento APE $n^{-1}\sum_{i=1}^{n}\Phi(\mathbf{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma})$ acaba sendo em torno de 0,589, que podemos usar para obter os efeitos parciais médios da estimação Tobit. Se, por exemplo, multiplicarmos o coeficiente da educ por 0,589 obteremos 0,589(80,65) ≈ 47,50 (isto é 47,5 horas a mais), que é bastante maior que o efeito parcial do MQO, cerca de 28,8 horas. Assim mesmo estimando um efeito médio, as estimativas Tobit são notavelmente maiores em magnitude que a estimativa MQO correspondente. Se, ao contrário, quisermos estimar o efeito de mais um ano de escolaridade começando nos valores médios de todas as variáveis explicativas, então calculamos o fator de escalonamento PEA $\Phi(\bar{\mathbf{x}}\hat{\boldsymbol{\beta}}/\hat{\sigma})$. Isto acaba sendo em torno de 0,645 [quando usamos a média quadrada de experiência, $(exper)^2$, em lugar da média de $exper^2$]. Este efeito parcial, que está em torno de 52 horas é quase duas vezes maior que a estimativa MQO. Com exceção da *crianma6*, os coeficientes da inclinação do escalonado todos serão maiores em magnitude do que o coeficiente MQO correspondente.

Descrevemos um *R*-quadrado tanto para o modelo de regressão linear como para o modelo Tobit. O *R*-quadrado do MQO é o habitual. Para o Tobit, o *R*-quadrado é o quadrado do coeficiente de correlação entre y_i e $\hat{y}_i$, no qual $\hat{y}_i = \Phi(\mathbf{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma})\,\mathbf{x}_i\hat{\boldsymbol{\beta}} + \hat{\sigma}\,\phi(\mathbf{x}_i\hat{\boldsymbol{\beta}}/\hat{\sigma})$ é a estimativa de $E(y|\mathbf{x} = \mathbf{x}_i)$. Isso é motivado pelo fato de o *R*-quadrado habitual do MQO ser igual à correlação elevada ao quadrado entre y_i e os valores estimados [veja a equação (3.29)]. Em modelos não lineares como o modelo Tobit, o coeficiente de correlação elevado ao quadrado não é idêntico a um *R*-quadrado baseado na soma dos quadrados dos resíduos como em (3.28). Isso ocorre porque os valores estimados, como definidos anteriormente, e os resíduos $y_i - \hat{y}_i$ não são não correlacionados na amostra. Um *R*-quadrado definido como o coeficiente de correlação elevado ao quadrado entre y_i e $\hat{y}_i$ tem a vantagem de sempre estar entre zero e um; um *R*-quadrado baseado na soma dos quadrados dos resíduos não precisa ter essa característica.

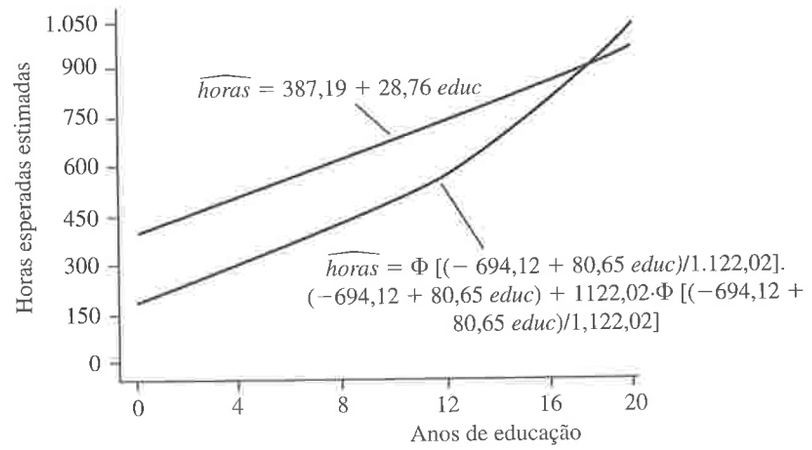
Podemos ver que, com base nas medidas do *R*-quadrado, a função Tobit da média condicional ajusta um pouco os dados sobre horas, mas não substancialmente melhor. Porém, devemos nos lembrar que as estimativas Tobit não são escolhidas para maximizar um *R*-quadrado — elas maximizam a função log-verossimilhança —, enquanto as estimativas MQO são os valores que efetivamente produzem o mais alto *R*-quadrado, dada a forma funcional linear.

Por construção, todos os valores estimados Tobit de *horas* são positivos. Em contraposição, 39 dos valores estimados MQO são negativos. Embora previsões negativas gerem alguma preocupação, 39 em 753 é apenas um pouco mais de cinco por cento das observações. Não fica totalmente claro como valores estimados negativos do MQO se traduzem em diferenças nos efeitos parciais estimados. A Figura 17.3 traça estimativas de $E(horas|\mathbf{x})$ como uma função da educação; no modelo Tobit, as outras variáveis explicativas são definidas em seus valores médios. No modelo linear, a equação traçada é $\widehat{horas} = 387{,}19 + 28{,}76\,educ$. No modelo Tobit, a equação traçada é $\widehat{horas} = \Phi[(-694{,}12 + 80{,}65\,educ)/1.122{,}02] \cdot (-694{,}12 + 80{,}65\,educ) + 1.122{,}02 \cdot \phi[(-694{,}12 + 80{,}65\,educ)/1.122{,}02]$. Como pode ser visto na figura, o modelo linear produz estimativas notavelmente mais altas das horas trabalhadas esperadas mesmo com altos níveis de educação. Por exemplo, com oito anos de educação, o valor previsto de horas pelo MQO é de aproximadamente 617,5, enquanto a estimativa Tobit está em torno de 423,9. Com 12 anos de educação, o valor previsto de *horas* é de, aproximadamente, 732,7 e 598,3, respectivamente. As duas linhas de previsão se cruzam após 17 anos de educação, mas nenhuma mulher na amostra tem mais de 17 anos de educação. A inclinação crescente da linha Tobit indica claramente o efeito marginal crescente da educação sobre as horas trabalhadas esperadas.

EXEMPLO 17.2 (continuação)

Figura 17.3

Valores esperados estimados de horas em relação à educação para os modelos Linear e Tobit.



Problemas de Especificação nos Modelos Tobit

O modelo Tobit, e em particular as fórmulas dos valores esperados em (17.24) e (17.25), dependem essencialmente da normalidade e da homoscedasticidade no modelo subjacente da variável latente. Quando $E(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$, sabemos, do Capítulo 5, que a normalidade condicional de y não desempenha nenhum papel na inexistência de viés, na consistência ou na inferência de amostras grandes. A heteroscedasticidade não afeta a inexistência de viés ou a consistência do MQO, embora devamos calcular erros-padrão e estatísticas de testes robustos para efetuarmos inferência aproximada. Em um modelo Tobit, se qualquer das hipóteses em (17.18) falhar, será difícil saber o que a EMV Tobit estará estimando. No entanto, para desvios moderados das hipóteses, o modelo Tobit pode produzir boas estimativas dos efeitos parciais sobre as médias condicionais. É possível levar em conta mais hipóteses gerais em (17.18), mas tais modelos são muito mais complicados de se estimar e interpretar.

Uma limitação potencialmente importante do modelo Tobit, pelo menos em certas aplicações, é que o valor esperado condicional em $y > 0$ está muito estreitamente ligado com a probabilidade de que $y > 0$. Isso está claro nas equações (17.26) e (17.29). Em particular, o efeito de x_j sobre $P(y > 0|\mathbf{x})$ é proporcional a β_j, como também o é o efeito sobre $E(y|y > 0,\mathbf{x})$, no qual ambas as funções que multiplicam β_j são positivas e dependem de $\mathbf{x}$ somente por meio de $\mathbf{x}\boldsymbol{\beta}/\sigma$. Isso elimina algumas possibilidades interessantes. Por exemplo, considere a relação entre o valor de cobertura de um seguro de vida e a idade da pessoa. Pessoas jovens podem ser menos propensas a fazer seguro de vida, de modo que a probabilidade de $y > 0$ cresce com a idade (pelo menos até determinado ponto). Condicional a ter um seguro de vida, o valor das apólices pode decrescer com a idade, já que seguros de vida se tornam menos importantes à medida que as pessoas se aproximam do fim de suas vidas. Essa possibilidade não é considerada no modelo Tobit.

Uma maneira de avaliar informalmente se o modelo Tobit é apropriado é estimar um modelo probit no qual o resultado binário, digamos w, será igual a um se $y > 0$, e $w = 0$ se $y = 0$. Assim, de (17.21), w segue um modelo probit, em que o coeficiente de x_j será $\gamma_j = \beta_j/\sigma$. E significa que podemos estimar a razão de β_j com σ pelo probit, para cada j. Se o modelo Tobit for válido, a estimativa probit, $\hat{\gamma}_j$, deve ficar "próxima" de $\beta_j/\hat{\sigma}$, em que β_j e $\hat{\sigma}$ são as estimativas Tobit. Elas nunca serão idênticas devido ao erro de amostragem. Mas podemos procurar por certos sinais problemáticos. Por exemplo, se $\hat{\gamma}_j$ for significante e negativo, mas β_j for positivo, o modelo Tobit poderá não ser apropriado. Ou, se $\hat{\gamma}_j$ e β_j tiverem o mesmo sinal, mas $|\beta_j/\hat{\sigma}|$ for muito maior, ou menor, que $|\hat{\gamma}_j|$, isso também pode indicar problemas. Não devemos nos preocupar muito com as mudanças de sinais ou diferenças em magnitudes nas variáveis explicativas que sejam não significantes em ambos os modelos.

No exemplo de horas anuais trabalhadas, $\hat{\sigma} = 1.122{,}02$. Quando dividimos o coeficiente Tobit de *nesprend* por $\hat{\sigma}$, obtemos $-8{,}81/1.122{,}02 \approx -0{,}0079$; o coeficiente probit de *nesprend* está em torno de $-0{,}012$, que é diferente, mas não de forma substancial. Para *crianmed6*, a estimativa do coeficiente de $\hat{\sigma}$ está em torno de $-0{,}797$, comparada com a estimativa probit de $-0{,}868$. Novamente, essa não é uma diferença muito grande, mas indica que o fato de ter filhos pequenos tem efeito maior sobre a decisão inicial de participar da força de trabalho do que uma mulher decidir quantas horas trabalhar, uma vez que ela faça parte da força de trabalho. (O Tobit efetivamente leva em conta a média desses dois efeitos simultaneamente.) Não sabemos se os efeitos são estatisticamente diferentes, mas eles são da mesma ordem de magnitude.

O que acontecerá se concluirmos que o modelo Tobit não é apropriado? Existem modelos, normalmente chamados de modelos de *saltos* ou de *duas partes*, que podem ser usados quando o Tobit parecer inadequado. Todos eles têm a propriedade de que $P(y > 0|\mathbf{x})$ e $E(y|y > 0,\mathbf{x})$ dependem de parâmetros diferentes, de modo que x_j pode ter efeitos diferentes sobre essas duas funções. [Veja Wooldridge (2002, Capítulo 16) para uma descrição desses modelos.]

17.3 O MODELO DE REGRESSÃO DE POISSON

Outro tipo de variável dependente não negativa é uma **variável de contagem**, que pode assumir valores inteiros não negativos: {0, 1, 2,...}. Estamos especialmente interessados em casos nos quais y assume um número relativamente pequeno de valores, inclusive zero. Os exemplos incluem número de filhos de uma mulher, o número de vezes em que alguém foi preso durante o ano, ou o número de patentes solicitadas por uma firma durante um ano. Pelas mesmas razões discutidas quanto às respostas binárias e Tobit, um modelo linear para $E(y|x_1, \ldots, x_k)$ pode não fornecer o melhor ajuste para todas as variáveis explicativas. (Mesmo assim, sempre é interessante iniciar com um modelo linear, como fizemos no Exemplo 3.5 do Capítulo 3).

Assim como um resultado Tobit, não podemos tomar o logaritmo de uma variável de contagem porque ela assume o valor zero. Um método eficaz é modelar o valor esperado como uma função exponencial:

$$E(y|x_1,x_2, \ldots, x_k) = \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k). \qquad \textbf{(17.31)}$$

Como $\exp(\cdot)$ é sempre positivo, (17.31) garante que os valores previstos de y também serão positivos. A função exponencial está traçada na Figura A.5 do Apêndice A, disponível no site da Cengage.

Embora (17.31) seja mais complicada que um modelo linear, basicamente já sabemos como interpretar os coeficientes. Tomando o log da equação (17.31):

$$\log[\mathrm{E}(y|x_1, x_2, \ldots, x_k)] = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \tag{17.32}$$

de forma que o log do valor esperado é linear. Portanto, usando as propriedades de aproximação da função log que temos usado nos capítulos anteriores,

$$\%\Delta \mathrm{E}(y|\mathbf{x}) \approx (100\beta_j)\Delta x_j.$$

Em outras palavras, $100\beta_j$ é aproximadamente a porcentagem de mudança em $\mathrm{E}(y|\mathbf{x})$, dado um aumento de uma unidade em x_j. Algumas vezes, necessitamos de uma estimativa mais precisa, e podemos encontrar uma, facilmente, verificando as mudanças discretas no valor esperado. Mantenha fixas todas as variáveis explicativas, exceto x_k, e defina x_k^0 como o valor inicial e x_k^1 como o valor subsequente. Então, a mudança proporcional no valor esperado será

$$[\exp(\beta_0 + \mathbf{x}_{k-1}\boldsymbol{\beta}_{k-1} + \beta_k x_k^1)/\exp(\beta_0 + \mathbf{x}_{k-1}\boldsymbol{\beta}_{k-1} + \beta_k x_k^0)] - 1 = \exp(\beta_k \Delta x_k) - 1,$$

em que $\mathbf{x}_{k-1}\boldsymbol{\beta}_{k-1}$ é a forma abreviada de $\beta_1 x_1 + \ldots + \beta_{k-1}x_{k-1}$, e $\Delta x_k = x_k^1 - x_k^0$. Quando $\Delta x_k = 1$ — por exemplo, se x_k for uma variável *dummy* que alteramos de zero para um —, então, a mudança será $\exp(\beta_k) - 1$. Dado $\hat{\beta}_k$, obtemos $\exp(\hat{\beta}_k) - 1$ e o multiplicamos por 100 para transformar a mudança proporcional em uma mudança percentual.

Se, digamos, $x_j = \log(z_j)$ de algumas variáveis $z_j > 0$, então seu coeficiente, β_j, é interpretado como uma elasticidade com relação à z_j. Tecnicamente, ela é uma elasticidade do *valor esperado* de y com relação à z_j, pois não podemos calcular a percentagem de alteração em casos onde $y = 0$. Isso, para nossa finalidade, não é importante.

O ponto principal é que, para propósitos práticos, podemos interpretar os coeficientes na equação (17.31) como se tivéssemos um modelo linear, com $\log(y)$ como a variável dependente. Existem algumas diferenças sutis que não precisamos estudar aqui.

Como (17.31) é não linear em seus parâmetros — lembre-se, $\exp(\cdot)$ é uma função não linear —, não podemos usar métodos de regressão linear. Poderíamos usar *mínimos quadrados não lineares*, os quais, como acontece com o MQO, minimizam a soma dos quadrados dos resíduos. Acontece, porém, que todas as distribuições de dados de contagem padrão exibem heteroscedasticidade, e os mínimos quadrados não lineares não exploram isso [veja Wooldridge (2002, Capítulo 12)]. Em vez disso, vamos nos valer da máxima verossimilhança e do importante método relacionado da *estimação de quase-máxima verossimilhança*.

No Capítulo 4, introduzimos a normalidade como a hipótese de distribuição padrão da regressão linear. A hipótese de normalidade é razoável para (em linhas gerais) variáveis dependentes contínuas que podem assumir um grande intervalo de valores. Uma variável de contagem não pode ter uma distribuição normal (pois a distribuição normal é de variáveis contínuas que podem assumir todos os valores), e se ela assumir muito poucos valores, a distribuição pode ser muito diferente da normal. Em vez desta, a distribuição normal de dados de contagem é a **distribuição de Poisson**.

Como estamos interessados no efeito das variáveis explicativas sobre y, devemos olhar a distribuição de Poisson condicional em $\mathbf{x}$. A distribuição de Poisson é inteiramente determinada por sua média, de modo que só precisamos especificar $\mathrm{E}(y|\mathbf{x})$. Assumimos que ela tem a mesma forma de (17.31), que escrevemos de forma abreviada como $\exp(\mathbf{x}\boldsymbol{\beta})$. Assim, a probabilidade de que y seja igual ao valor h, condicional em $\mathbf{x}$, será

$$\mathrm{P}(y = h|\mathbf{x}) = \exp[-\exp(\mathbf{x}\boldsymbol{\beta})][\exp(\mathbf{x}\boldsymbol{\beta})]^h/h!, \; h = 0, 1, \ldots,$$

em que $h!$ significa o fatorial (veja o Apêndice B no site da Cengage). Essa distribuição, que é a base do **modelo de regressão de Poisson**, permite encontrar as probabilidades condicionais de qualquer valor das variáveis explicativas. Por exemplo, $\mathrm{P}(y = 0|\mathbf{x}) = \exp[-\exp(\mathbf{x}\boldsymbol{\beta})]$. Logo que tenhamos as estimativas dos β_j, poderemos inseri-las nas probabilidades de vários valores de $\mathbf{x}$.

Dada uma amostra aleatória $\{(\mathbf{x}_i, y_i): i = 1, 2, \ldots, n\}$, podemos construir a função log-verossimilhança:

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \mathbf{x}_i\boldsymbol{\beta} - \exp(\mathbf{x}_i\boldsymbol{\beta})\}, \tag{17.33}$$

onde eliminamos o termo $-\log(y_i!)$ porque ele não depende de $\boldsymbol{\beta}$. Essa função log-verossimilhança é simples de maximizar, embora as EMVs de Poisson não sejam obtidas em forma fechada.

Os erros-padrão das estimativas de Poisson $\hat{\beta}_j$ são fáceis de ser obtidos depois de a função log-verossimilhança ter sido maximizada; a fórmula se encontra no apêndice deste capítulo. Elas são descritas com os $\hat{\beta}_j$ por qualquer programa econométrico.

Assim como nos modelos probit, logit e Tobit, não podemos comparar diretamente as magnitudes das estimativas de Poisson de uma função exponencial com as estimativas MQO de uma função linear. No entanto, é possível fazer uma comparação aproximada, pelo menos para as variáveis explicativas contínuas. Se (17.31) for válida, o efeito parcial de x_j em relação a $\mathrm{E}(y|x_1, x_2, \ldots, x_k)$ será $\partial \mathrm{E}(y|x_1, x_2, \ldots, x_k)/x_j = \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) \cdot \beta_j$. Esta expressão é derivada da regra da cadeia em cálculo infinitesimal, pois as derivativas da função exponencial é simplesmente a função exponencial. Se especificarmos que $\hat{\gamma}_j$ denote um coeficiente de inclinação dos MQO da regressão de y sobre $x_1, x_2, \ldots, x_k$, então poderemos, em termos gerais, comparar a magnitude de $\hat{\gamma}_j$ e o efeito parcial médio de uma função de regressão exponencial. Curiosamente, o fator de escalonamento do APE, neste caso, $n^{-1}\sum_{i=1}^{n} \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik}) = n^{-1}\sum_{i=1}^{n} \hat{y}_i$, é simplesmente a média amostral $\bar{y}$ da y_i, em que definimos os valores ajustados $\hat{y}_i = \exp(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})$. Em outras palavras, para uma regressão de Poisson com uma função de média exponencial, a média dos valores ajustados é a mesma que a dos resultados originais na y_i — exatamente como no caso de regressão linear. Isso torna simples escalonar as estimativas de Poisson, $\hat{\beta}_j$, para torná-las comparáveis com as estimativas pelos MQO correspondentes, $\hat{\gamma}_j$: de uma variável explicativa contínua, podemos comparar $\hat{\gamma}_j$ com $\bar{y} \cdot \hat{\beta}_j$.

Embora a análise da EMV de Poisson seja o primeiro passo natural para dados de contagem, frequentemente ela é restritiva demais. Todas as probabilidades e os momentos de ordem mais alta da distribuição de Poisson são inteiramente determinados pela média. Em particular, a variância é igual à média:

$$\mathrm{Var}(y|\mathbf{x}) = \mathrm{E}(y|\mathbf{x}). \tag{17.34}$$

Isso é restritivo e já foi mostrado que é violado em muitas aplicações. Felizmente, a distribuição de Poisson tem uma propriedade de robustez bastante precisa: independentemente de a distribuição de Poisson ser válida, ainda assim obtemos estimadores dos β_j consistentes e assimptoticamente normais. [Veja Wooldridge (2002, Capítulo 19), para detalhes.] Isso é análogo ao estimador MQO, que é consistente e assimptoticamente normal, independentemente de a hipótese de normalidade ser válida; contudo, o MQO é o EMV sob normalidade.

Quando usamos a EMV de Poisson, mas não assumimos que a distribuição de Poisson seja inteiramente correta, chamamos a análise de **estimação de quase-máxima verossimilhança (EQMV)**. A EQMV de Poisson é bastante prática, e está incluída em vários programas econométricos. Porém, a menos que a hipótese de variância de Poisson (17.34) se mantenha, os erros-padrão terão que ser ajustados.

Um ajuste simples dos erros-padrão está disponível quando assumimos que a variância é proporcional à média:

$$\text{Var}(y|\mathbf{x}) = \sigma^2 \text{E}(y|\mathbf{x}), \tag{17.35}$$

em que $\sigma^2 > 0$ é um parâmetro desconhecido. Quando $\sigma^2 = 1$, obtemos a hipótese de variância de Poisson. Quando $\sigma^2 > 1$, a variância será maior que a média para todos os $\mathbf{x}$; isso é chamado de **superdispersão**, porque a variância é maior do que no caso Poisson, e é observado em muitas aplicações de regressões de contagem. O caso $\sigma^2 < 1$, chamado de *subdispersão*, é menos comum, mas é permitido em (17.35).

Sob (17.35), é fácil ajustar os erros-padrão habituais da EMV de Poisson. Seja $\hat{\beta}_j$ a EQMV de Poisson e defina os residuais como $\hat{u}_i = y_i - \hat{y}_i$, em que $\hat{y}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik})$ é o valor ajustado. Como sempre, o resíduo da observação i é a diferença entre y_i e seu valor ajustado. Um estimador consistente de σ^2 é $(n - k - 1)^{-1}\sum_{i=1}^{n} \hat{u}_i^2/\hat{y}_i$, em que a divisão por $\hat{y}_i$ é o ajuste adequado da heteroscedasticidade, e $n - k - 1$ representa os gl, dadas n observações e $k + 1$ estimativas $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. Definindo $\hat{\sigma}$ como a raiz quadrada positiva de $\hat{\sigma}^2$, multiplicamos os erros-padrão habituais de Poisson por $\hat{\sigma}$. Se $\hat{\sigma}$ for notavelmente maior que 1, os erros-padrão corrigidos podem ser muito maiores que os erros-padrão nominais da EMV de Poisson, geralmente incorretos.

Mesmo (17.35) não é inteiramente geral. Como no modelo linear, podemos também obter erros-padrão da EQMV de Poisson que não restrinjam a variância. [Veja Wooldridge (2002, Capítulo 19), para explicações adicionais.]

QUESTÃO 17.4

Suponha que obtemos $\hat{\sigma}^2 = 2$. Como os erros-padrão ajustados podem ser comparados com os erros-padrão habituais da EMV de Poisson? Como a estatística quase-*RV* pode ser comparada com a estatística *RV* habitual?

Sob a hipótese de distribuição de Poisson, podemos usar a estatística razão de verossimilhança para testar restrições de exclusão, as quais, como sempre, têm a forma de (17.12). Se tivermos q restrições de exclusão, a estatística será aproximadamente distribuída como χ_q^2 sob a hipótese nula. Sob a hipótese menos restritiva (17.35), há um ajuste simples (e então chamamos a estatística de **estatística quase-razão de verossimilhança**): dividimos (17.12) por $\hat{\sigma}^2$, na qual $\hat{\sigma}^2$ é obtida do modelo sem restrições.

EXEMPLO 17.3

(Regressão de Poisson do Número de Prisões)

Agora aplicamos o modelo de regressão de Poisson aos dados de prisões em CRIME1.RAW usados, entre outros locais, no Exemplo 9.1. A variável dependente, *npre86*, é o número de vezes que um homem foi preso em 1986. Essa variável é zero para 1.970 de 2.725 homens na amostra, e somente oito valores de *npre86* são maiores que cinco. Assim, um modelo de regressão de Poisson é mais apropriado que um modelo de regressão linear. A Tabela 17.3 também apresenta os resultados da estimação por MQO de um modelo de regressão linear.

EXEMPLO 17.3 (continuação)

Tabela 17.3
Determinantes do número de prisões de homens jovens.

Variável dependente: *npre86*		
Variáveis independentes	**Linear (MQO)**	**Exponencial (EQMV de Poisson)**
pcond	−0,132 (0,040)	−0,402 (0,085)
sentmed	−0,011 (0,012)	−0,024 (0,020)
temptot	0,012 (0,009)	0,024 (0,015)
ptemp86	−0,041 (0,009)	−0,099 (0,021)
empr86	−0,051 (0,014)	−0,038 (0,029)
rend86	−0,0015 (0,0003)	−0,0081 (0,0010)
negro	0,327 (0,045)	0,661 (0,074)
hispan	0,194 (0,040)	0,500 (0,074)
nasc60	−0,022 (0,033)	−0,051 (0,064)
constante	0,577 (0,038)	−0,600 (0,067)
Valor de Log-Verossimilhança *R*-Quadrado $\hat{\sigma}$	——— 0,073 0,829	−2.248,76 0,077 1,232

Os erros-padrão do MQO são os habituais; com certeza, poderíamos tê-los tornado robustos quanto à heteroscedasticidade. Os erros-padrão da regressão de Poisson são os erros-padrão usuais de máxima verossimilhança. Como $\hat{\sigma} = 1{,}232$, os erros-padrão da regressão de Poisson devem ser corrigidos por esse fator (de forma que cada erro-padrão retificado seja aproximadamente 23% maior.) Por exemplo, um erro-padrão

EXEMPLO 17.3 (continuação)

mais confiável de *temptot* seria 1,23(0,015) ≈ 0,0185, o que produz uma estatística *t* de aproximadamente 1,3. O ajustamento dos erros-padrão reduz a significância de todas as variáveis, mas várias delas ainda serão estatisticamente bastante significantes.

Os coeficientes do MQO e de Poisson não são diretamente comparáveis e possuem significados bastante diferentes. Por exemplo, o coeficiente de *pcond* indica que, se $\Delta pcond = 0,10$, o número esperado de prisões cai em 0,013 (*pcond* é a proporção de prisões anteriores que levaram a uma condenação). O coeficiente de Poisson indica que $\Delta pcond = 0,10$ reduz as prisões esperadas em cerca de 4% [0,402(0,10) = 0,0402, e multiplicamos esse resultado por 100 para obtermos a porcentagem do efeito]. Como regra, isso sugere que podemos reduzir o número total de prisões em cerca de 4% se pudermos aumentar a probabilidade de condenação em 0,1.

O coeficiente de Poisson de *negro* indica que, outros fatores mantidos iguais, o número esperado de prisões de homens negros é estimado em cerca de 100 · [exp(0,661) − 1] ≈ 93,7% mais alto que o de um homem branco com os mesmos valores das outras variáveis explicativas..

Como na aplicação Tobit na Tabela 17.2, apresentamos um *R*-quadrado da regressão de Poisson. Esse é o quadrado do coeficiente de correlação entre y_i e $\hat{y}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik})$. A motivação para essa medida da qualidade de ajuste é a mesma do modelo Tobit. Observamos que o modelo de regressão exponencial, estimado pela EQMV de Poisson, se ajusta um pouco melhor. Lembre-se de que as estimativas MQO são escolhidas para maximizar o *R*-quadrado, mas as estimativas de Poisson não têm essa finalidade. (Elas são escolhidas para maximizar a função log-verossimilhança.)

Outros modelos de regressão de dados de contagem têm sido propostos e usados em aplicações, que generalizam a distribuição de Poisson de várias maneiras. Se estivermos interessados nos efeitos dos x_j sobre a resposta média, há poucas razões para irmos além da regressão de Poisson: ela é simples, frequentemente produz bons resultados, e tem a propriedade de robustez discutida anteriormente. Na verdade, podemos aplicar a regressão de Poisson a uma variável *y* que é um resultado do tipo Tobit, desde que (17.31) se mantenha. Isso pode produzir boas estimativas dos efeitos médios. Extensões da regressão de Poisson são mais úteis quando estamos interessados em estimar probabilidades, tal como em $P(y > 1|\mathbf{x})$. [Veja, por exemplo, Cameron e Trivedi (1998).]

17.4 MODELOS DE REGRESSÃO CENSURADA E TRUNCADA

Os modelos nas Seções 17.1, 17.2 e 17.3 aplicam-se a vários tipos de variáveis dependentes limitadas que frequentemente surgem aplicados em trabalhos de econometria. Ao usarmos esses métodos, é importante lembrar que a razão de usarmos um modelo probit ou logit para uma resposta binária, um modelo Tobit para um resultado de solução de canto, ou um modelo de regressão de Poisson para uma resposta de contagem é porque queremos modelos que avaliem características importantes da distribuição de *y*. Não há nenhuma questão de observação dos dados. Por exemplo, na aplicação Tobit da participação das mulheres na força de trabalho, no Exemplo 17.2, não existe nenhum problema para se observar horas trabalhadas: é simplesmente o fato de que uma fração não desprezível das mulheres casadas na população escolhem não trabalhar em troca de uma remuneração. Na aplicação da regressão de Poisson nas prisões anuais, observamos a variável dependente para cada homem jovem em uma amostra aleatória da população, mas a variável dependente pode ter valor zero como também outros valores inteiros pequenos.

Infelizmente, a distinção entre aglomeração numa variável resultante (tal como assumir o valor zero de uma fração não desprezível da população) e problemas de censura de dados pode ser confusa. Isto é especialmente verdadeiro quando se aplica o modelo Tobit. Neste livro, o modelo Tobit padrão descrito na Seção 17.2 é somente para resultados de solução de canto. Mas a literatura sobre modelos Tobit trata outra situação dentro da mesma estrutura: a variável de resposta foi censurada acima ou abaixo de algum limite. Tipicamente, a censura é devida a uma concepção de pesquisa e, em alguns casos, a restrições institucionais. Em vez de tratarmos censura de dados juntamente com resultados de solução de canto, resolvemos o problema da censura de dados aplicando o **modelo de regressão censurada**. Essencialmente, o problema resolvido por um modelo de regressão censurada é o de falta de dados na variável de resposta, *y*. Embora tenhamos condições de extrair unidades da população e obter informações nas variáveis explicativas de todas as unidades, o resultado na y_i está faltando para algum *i*. Mesmo assim, saberemos se os valores faltantes estão acima ou abaixo de determinado limite, e este conhecimento fornece informação útil para estimarmos os parâmetros.

Um **modelo de regressão truncada** surge quando excluímos, na base de *y*, um subconjunto da população em nosso esquema de amostragem. Em outras palavras, não temos uma amostra aleatória da população subjacente, mas conhecemos a regra que foi usada para incluir unidades na amostra. Essa regra é determinada pelo critério de *y* estar acima ou abaixo de certo valor limite. Mais adiante, explicaremos, de maneira mais completa a diferença entre modelos de regressão censurada e truncada.

Modelos de Regressão Censurada

Embora os modelos de regressão censurada possam ser definidos sem hipóteses sobre distribuições, nesta subseção estudaremos o **modelo de regressão normal censurada**. A variável que gostaríamos de explicar, *y*, segue o modelo linear clássico. Para enfatizar, colocamos um subscrito *i* em uma extração aleatória da população:

$$y_i = \beta_0 + \mathbf{x}_i\boldsymbol{\beta} + u_i,\ u_i|\mathbf{x}_i,\ c_i \sim \text{Normal}(0, \sigma^2) \qquad \textbf{(17.36)}$$

$$w_i = \min(y_i, c_i). \qquad \textbf{(17.37)}$$

Em vez de observar y_i, somente a observaremos se ela for inferior a um valor de censura, c_i. Observe que (17.36) inclui a hipótese de ser u_i independente de c_i. (Concretamente, consideramos explicitamente a censura acima, ou *censura à direita*; o problema de fazer a censura abaixo, ou *censura à esquerda*, é tratado de forma semelhante).

QUESTÃO 17.5

Defina vpm_i como o valor do produto marginal do trabalhador *i*; esse é o preço do produto de uma firma multiplicado pelo produto marginal do trabalhador. Assuma que vpm_i é uma função linear de variáveis exógenas, tais como educação, experiência e assim por diante, e um erro não observável. Sob concorrência perfeita e sem restrições institucionais, cada trabalhador recebe seu valor do produto marginal. Defina $salmín_i$ como o salário mínimo do trabalhador *i*, que varia por estado. Observamos $salário_i$, que é o maior dos vpm_i e $salmín_i$. Escreva o modelo apropriado para o salário observado.

Um exemplo da censura à direita de dados é a **codificação superior**. Quando uma variável tem codificação superior, conhecemos seu valor somente até certo limite. Para respostas maiores que o limite, somente sabemos que a variável é pelo menos tão grande quanto o limite. Por exemplo, em algumas pesquisas, a riqueza familiar tem codificação superior. Suponha que os entrevistados sejam questionadas sobre sua riqueza, mas as pessoas poderão responder "mais de 500.000 dólares". Assim, observamos a verdadeira riqueza dos entrevistados cujo valor dos bens seja inferior a 500.000 dólares, mas não daqueles cujo valor dos bens seja superior a 500.000 dólares. Nesse caso, o limite da censura, c_i, é o mesmo para todo i. Em muitas situações, o limite da censura muda com as características individuais ou familiares.

Se observarmos uma amostra aleatória de (**x**, y), simplesmente estimaremos $\boldsymbol{\beta}$ por MQO, e a inferência estatística será padrão. (Novamente absorvemos o intercepto em **x** para simplificar.) A censura causa problemas. Usando argumentos semelhantes aos de um modelo Tobit, uma regressão MQO que use somente observações não censuradas — isto é, aquelas com $y_i < c_i$ — produz estimadores inconsistentes dos β_j. Uma regressão MQO de w_i sobre $\mathbf{x}_i$, usando todas as observações, não estima consistentemente os β_j, a menos que não haja censura. Isso é semelhante ao caso Tobit, mas o problema é muito diferente. No modelo Tobit, estamos modelando comportamento econômico, que muitas vezes produz resultados iguais a zero; supõe-se que o modelo Tobit reflita isso. Com a regressão censurada, temos um problema de coleta de dados porque, por alguma razão, os dados são censurados.

Sob as hipóteses em (17.36) e (17.37), podemos estimar $\boldsymbol{\beta}$ (e σ^2) por máxima verossimilhança, dada uma amostra aleatória de ($\mathbf{x}_i$, w_i). Para isso, precisamos da densidade de w_i, dado ($\mathbf{x}_i$, c_i). Para observações não censuradas, $w_i = y_i$, e a densidade de w_i será a mesma da y_i: Normal($\mathbf{x}_i\boldsymbol{\beta}$,$\sigma^2$). Para observações censuradas, precisamos da probabilidade de que w_i seja igual ao valor de censura, c_i, dado $\mathbf{x}_i$:

$$P(w_i = c_i|\mathbf{x}_i) = P(y_i \geq c_i|\mathbf{x}_i) = P(u_i \geq c_i - \mathbf{x}_i\boldsymbol{\beta}) = 1 - \Phi[(c_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma].$$

Podemos combinar essas duas partes para obter a densidade de w_i, dados $\mathbf{x}_i$ e c_i:

$$f(w|\mathbf{x}_i,c_i) = 1 - \Phi[(c_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma],\ w = c_i, \qquad \textbf{(17.38)}$$

$$= (1/\sigma)\phi\,[(w - \mathbf{x}_i\boldsymbol{\beta})/\sigma],\ w < c_i. \qquad \textbf{(17.39)}$$

O log-verossimilhança da observação i é obtido tomando o log natural da densidade de cada i. Podemos maximizar a soma deles ao longo de i, com relação a β_j e σ para obtermos as EMVs.

É bom saber que podemos interpretar os β_j do mesmo jeito que em um modelo de regressão linear sob amostragem aleatória. Isso é muito diferente das aplicações Tobit nas respostas de solução de canto, nas quais os valores esperados de interesse são funções não lineares dos β_j.

Uma aplicação importante dos modelos de regressão censurada é a **análise de duração**. Uma *duração* é uma variável que registra o tempo antes da ocorrência de certo evento. Por exemplo, podemos explicar o número de dias antes de um criminoso solto da prisão ser preso novamente. Para alguns criminosos, isso pode nunca acontecer, ou talvez aconteça após um período tão longo que precisaremos censurar a duração para podermos analisar os dados.

Em aplicações de duração de regressão normal censurada, como também na codificação superior, usamos com frequência o log natural como a variável dependente, significando que também tomamos o log do valor de limite censura em (17.37). Como temos visto ao longo de todo este texto, o uso da transformação logarítmica da variável dependente faz com que os parâmetros sejam interpretados como mudanças percentuais. Além disso, como acontece com muitas variáveis positivas, o log de uma duração em geral tem uma distribuição mais próxima da (condicional) normal do que a própria duração.

EXEMPLO 17.4

(Intervalo de Reincidência)

O arquivo RECID.RAW contém dados sobre o tempo em meses até que um ex-recluso de uma prisão da Carolina do Norte foi preso após ter sido solto; vamos chamar de *durat*. Alguns dos presidiários participaram de um programa de trabalho durante o tempo em que estiveram na prisão. Também controlamos diversas variáveis demográficas, bem como medidas de prisões e históricos criminais.

De 1.445 reclusos, 893 não foram presos durante o tempo em que foram vigiados; portanto, essas observações foram censuradas. O tempo censurado diferiu entre os reclusos, variando de 70 a 81 meses.

A Tabela 17.4 mostra os resultados da regressão normal censurada de log(*durat*). Cada um dos coeficientes, quando multiplicado por 100, informa a mudança percentual estimada na duração esperada, dado um aumento *ceteris paribus* de uma unidade na variável explicativa correspondente.

Vários dos coeficientes na Tabela 17.4 são interessantes. As variáveis *anteriores* (número de condenações anteriores) e *totpris* (total de meses passados na prisão) têm efeitos negativos sobre o tempo até que ocorra a nova prisão. Isso sugere que essas variáveis medem a tendência da atividade criminal, em vez de representar um efeito dissuasor. Por exemplo, um recluso com uma condenação anterior a mais tem uma intervalo até a próxima prisão que é quase 14% menor. Um ano de reclusão reduz o intervalo em cerca de $100 \cdot 12(0{,}019) = 22{,}8\%$. Uma constatação surpreendente é que um homem que esteja cumprindo pena por delito grave tem uma duração esperada estimada quase 56% [exp(0,444) − 1 ≈ 0,56] *mais longa* que alguém que esteja cumprindo pena por um crime menos grave.

Tabela 17.4

Estimação da regressão censurada de reincidência criminal.

Variável dependente: log(*durat*)	
Variáveis independentes	**Coeficientes (erro-padrão)**
protrab	−0,063 (0,120)
anteriores	−0,137 (0,021)
totpris	−0,019 (0,003)
criminoso	0,444 (0,145)
álcool	−0,635 (0,144)
drogas	−0,298 (0,133)
negro	−0,543 (0,117)

(*cont.*)

EXEMPLO 17.4 (continuação)

(*cont.*)

Variável dependente: log(*durat*)	
Variáveis independentes	**Coeficientes (erro-padrão)**
casado	0,341 (0,140)
educ	0,023 (0,025)
idade	0,0039 (0,0006)
constante	4,099 (0,348)
Valor de Log-Verossimilhança	−1.597,06
$\hat{\sigma}$	1,810

Os que têm um histórico de abuso de drogas ou álcool têm intervalo esperado substancialmente mais curto até a próxima prisão. (As variáveis *álcool* e *drogas* são binárias.) Estima-se que para homens mais velhos e homens que eram casados quando do encarceramento os intervalos sejam significantemente mais longos até suas próximas prisões. Os negros têm intervalos substancialmente mais curtos, da ordem de 42% [exp(−0,543) − 1 ≈ −0,42].

A variável de decisão crucial, *protrab*, não tem o efeito desejado. A estimativa por ponto é que, outros fatores permanecendo inalterados, homens que tenham participado do programa de trabalho têm intervalos de reincidência estimadas cerca de 6,3% mais curtas se comparadas aos que não participaram do programa. O coeficiente tem uma estatística *t* pequena, de modo que provavelmente concluiríamos que o programa de trabalho não tem efeito nenhum. Isso pode ser motivado por um problema de autosseleção, ou pode ser produto da maneira pela qual os homens são alocados para o programa. Naturalmente, é possível que o programa tenha sido ineficiente.

Nesse exemplo, é essencial explicar a censura, especialmente porque quase 62% das durações são censuradas. Se aplicarmos o MQO puro à totalidade da amostra e tratarmos as durações censuradas como se não fossem censuradas, as estimativas dos coeficientes serão notavelmente diferentes. Na verdade, todas elas se reduzem em direção a zero. Por exemplo, o coeficiente em *anteriores* se tornará −0,059 (ep = 0,009), e o de *álcool* se tornará −0,262 (ep = 0,060). Embora as direções dos efeitos sejam as mesmas, a importância dessas variáveis é bastante reduzida. As estimativas da regressão censurada são muito mais confiáveis.

Existem outras maneiras de medir os efeitos de cada uma das variáveis explicativas da Tabela 17.4 sobre a duração, em vez de nos concentrarmos apenas na duração esperada. Uma abordagem sobre a análise de duração moderna está além do escopo deste livro. [Para uma introdução ao tema, veja Wooldridge (2002, Capítulo 20).]

Se qualquer das hipóteses da regressão normal censurada for violada — particularmente se houver heteroscedasticidade ou não normalidade em u_i — , as EMVs geralmente serão inconsistentes. Isso mostra que a censura é potencialmente muito onerosa, já que o MQO usando uma amostra não censurada não exige normalidade ou heteroscedasticidade para a consistência. Há métodos que não exigem que assumamos uma distribuição, mas eles são mais avançados. [Veja Wooldridge (2002, Capítulo 16).]

Modelos de Regressão Truncada

O modelo de regressão truncada difere num aspecto importante do modelo de regressão censurada. No caso de censura de dados nós extraímos, aleatoriamente, unidades amostrais da população. O problema de censurar é que, embora sempre observemos as variáveis explicativas de cada unidade aleatoriamente extraída, observamos o resultado na y somente quando ela for não censurada acima ou abaixo de determinado limite. Com o truncamento de dados, restringimos a atenção em um subconjunto da população antes da amostragem; assim, há uma parte da população da qual não observamos qualquer informação. Particularmente, não teremos qualquer informação das variáveis explicativas. O cenário da amostragem truncada caracteristicamente surge quando uma pesquisa objetiva um subconjunto particular da população e, talvez devido a considerações de custo, ignora totalmente a outra parte da população. Posteriormente, os pesquisadores podem usar a amostra truncada para responder questões sobre a população inteira, mas deve-se reconhecer que o esquema de amostragem não gerou uma amostra aleatória de toda a população.

Por exemplo, Hausman e Wise (1977) utilizaram dados de um experimento de imposto de renda negativo para estudar vários determinantes de receitas. Para ser incluída no estudo, uma família tinha que ter renda inferior a 1,5 vezes a linha de pobreza de 1967, em que a linha de pobreza dependia do tamanho da família. Hausman e Wise queriam usar os dados para estimarem uma equação de ganhos da totalidade da população.

O **modelo de regressão truncada normal** começa com um modelo de população básica que satisfaça as hipóteses do modelo linear clássico:

$$y = \beta_0 + \mathbf{x}\boldsymbol{\beta} + u,\ u|\mathbf{x} \sim \text{Normal}(0,\sigma^2). \qquad \textbf{(17.40)}$$

Lembre-se de que esse é um forte conjunto de hipóteses, pois u deve não só ser independente de $\mathbf{x}$, mas também normalmente distribuído. Vamos nos concentrar nesse modelo, pois é difícil relaxar as hipóteses.

Sob (17.40) sabemos que, dada uma amostra aleatória da população, o MQO é o procedimento mais eficiente de estimação. O problema surge porque não observamos uma amostra aleatória da população: a Hipótese RLM.2 é violada. Em particular, uma extração aleatória $(\mathbf{x}_i,y_i)$ é observada somente se $y_i \leq c_i$, em que c_i é o valor limite do truncamento que pode depender de variáveis exógenas — particularmente de $\mathbf{x}_i$. (No exemplo de Hausman e Wise, c_i depende do tamanho da família). Isso significa que, se $\{(\mathbf{x}_i, y_i): i = 1, \ldots, n\}$ for nossa amostra *observada*, y_i será necessariamente menor ou igual a c_i. Isso difere do modelo de regressão censurada; em um modelo de regressão censurada, observamos $\mathbf{x}_i$ para qualquer observação extraída aleatoriamente da população; no modelo truncado, somente observamos $\mathbf{x}_i$ se $y_i \leq c_i$.

Para estimar os β_j (com σ), necessitamos da distribuição de y_i, dados $y_i \leq c_i$ e $\mathbf{x}_i$. Isso é escrito da seguinte maneira

$$g(y|\mathbf{x}_i,c_i) = \frac{f(y|\mathbf{x}_i\boldsymbol{\beta},\sigma^2)}{F(c_i|\mathbf{x}_i\boldsymbol{\beta},\sigma^2)},\ y \leq c_i, \qquad \textbf{(17.41)}$$

em que $f(y|\mathbf{x}_i\boldsymbol{\beta},\sigma^2)$ representa a densidade normal com média $\beta_0 + \mathbf{x}_i\boldsymbol{\beta}$ e variância σ^2, e $F(c_i|\mathbf{x}_i\boldsymbol{\beta},\sigma^2)$ é a fdc normal com as mesmas média e variância, avaliadas em c_i. Essa expressão para a densidade,

condicional a $y_i \leq c_i$, tem sentido intuitivo: é a densidade da população para y, dado $\mathbf{x}$, dividida pela probabilidade de que y_i seja menor ou igual a c_i (dado $\mathbf{x}_i$), $P(y_i \leq c_i|\mathbf{x}_i)$. Na realidade, normalizamos outra vez a densidade dividindo-a pela área sob $f(\cdot|\mathbf{x}_i\boldsymbol{\beta},\sigma^2)$ que está à esquerda de c_i.

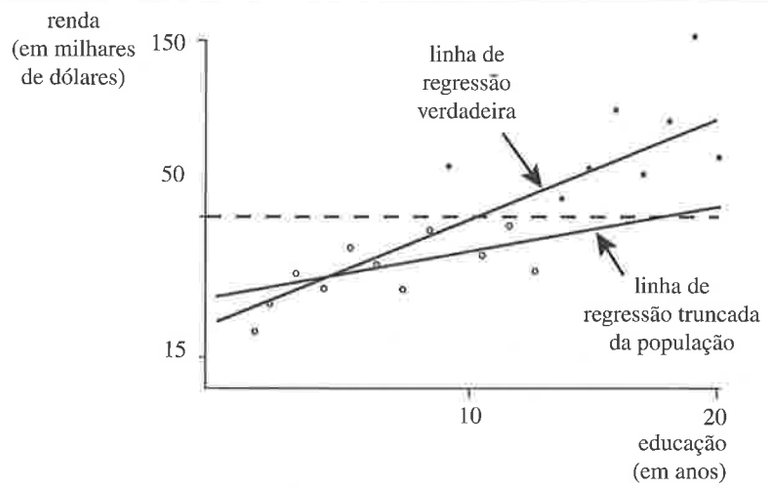
Se tomarmos o log de (17.41), somarmos ao longo de todos os i, e maximizarmos o resultado em relação a β_j e σ^2, obteremos os estimadores de máxima verossimilhança. Isso conduz a estimadores consistentes e aproximadamente normais. A inferência, inclusive os erros-padrão e estatísticas de log verossimilhança, é padrão.

Poderíamos analisar os dados do Exemplo 17.4 como uma amostra truncada se eliminássemos todos os dados de uma observação sempre que ela tivesse sido censurada. Isso nos daria 552 observações de uma distribuição normal truncada, na qual o ponto de truncamento difere ao longo das observações i. Porém, nunca analisaríamos dados de duração (ou de codificação superior) dessa maneira, pois ela elimina informações úteis. O fato de conhecermos um limite inferior de 893 durações, com as variáveis explicativas, é informação útil; a regressão censurada usa essas informações, o que a regressão truncada não faz.

Um exemplo melhor é dado por Hausman e Wise (1977), no qual eles enfatizam que o MQO aplicado a uma amostra truncada acima geralmente produz estimadores viesados para zero. Intuitivamente, isso faz sentido. Suponha que a relação de interesse seja entre os níveis de renda e educação. Se apenas observarmos pessoas cuja renda esteja abaixo de certo valor, estaremos eliminando a parte superior. Isso tende a nivelar a linha estimada, em relação à verdadeira linha da regressão, na totalidade da população. A Figura 17.4 ilustra o problema quando a renda é truncada acima de 50.000 dólares. Embora observemos os pontos dos dados representados pelos círculos abertos, não observamos os conjuntos de dados representados pelos círculos escuros. Uma análise de regressão usando amostra truncada não conduz a estimadores consistentes. A propósito, se a amostra da Figura 17.4 tivesse sido censurada em vez de truncada — isto é, tivéssemos dados com codificação superior —, observaríamos níveis de educação para todos os pontos na Figura 17.4, mas para indivíduos com renda acima de 50.000 dólares não saberíamos o montante exato da renda. Apenas saberíamos que a renda seria de pelo menos 50.000 dólares. Na verdade, todas as observações representadas pelos círculos escuros seriam levadas para baixo na linha horizontal de *renda* = 50.

Figura 17.4

Uma linha de regressão verdadeira, ou populacional, e a linha de regressão incorreta da população truncada com renda abaixo de US$ 50.000.



Assim como na regressão censurada, se a hipótese normal homoscedástica subjacente em (17.40) for violada, a EMV normal truncada será viesada e inconsistente. Há métodos que não necessitam dessas hipóteses; veja Wooldridge (2002, Capítulo 17), para discussão e referências.

17.5 CORREÇÕES DA SELEÇÃO AMOSTRAL

A regressão truncada é um caso especial de um problema genérico conhecido como **seleção de amostra não aleatória**. Entretanto, o projeto da pesquisa não é a única causa da seleção de amostra não aleatória. Com frequência, entrevistados não respondem a certas perguntas, o que leva a dados ausentes das variáveis dependentes ou independentes. Como não podemos usar essas observações em nossa estimação, devemos imaginar se suas eliminações conduzirão a viés em nossos estimadores.

Outro exemplo genérico é habitualmente chamado de **truncamento ocasional**. Nesse caso, não observamos y em razão do resultado de outra variável. O principal exemplo é estimar a chamada *função de oferta de salário* na área da economia do trabalho. O interesse reside em vários fatores, como a educação, afetam o salário que um indivíduo poderia ganhar na força de trabalho. Para as pessoas que estão na força de trabalho, observamos a oferta de salário como o salário corrente. Contudo, para aqueles que estejam desempregados, não observamos a oferta de salário. Como trabalhar pode estar sistematicamente correlacionado a fatores não observáveis que afetam a oferta de salário, usar somente pessoas que estejam trabalhando — o que temos feito em todos os exemplos de salários até agora — pode produzir estimadores viesados dos parâmetros na equação de oferta de salário.

A seleção de amostra não aleatória também pode surgir quando temos dados em painel. No caso mais simples, teremos dois anos de dados, mas, em razão das demissões, algumas pessoas saem da amostra. Isso é particularmente um problema na análise de políticas empresariais, nas quais as demissões podem estar relacionadas à eficácia de um programa administrativo.

Quando o MQO é Consistente na Amostra Selecionada?

Na Seção 9.4, apresentamos uma breve explicação dos tipos de seleções amostrais que podem ser ignorados. A distinção crucial é entre seleções amostrais *exógenas* e *endógenas*. No caso Tobit truncado, claramente temos seleção amostral endógena e o MQO é viesado e não consistente. De outro lado, se nossa amostra for determinada somente por uma variável explicativa exógena, teremos seleção amostral exógena. Casos entre esses dois extremos são menos claros, e agora apresentamos cuidadosas definições e hipóteses para eles. O modelo populacional é

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u,\ \mathrm{E}(u|x_1, x_2, \ldots, x_k) = 0. \qquad \textbf{(17.42)}$$

É útil escrever o modelo populacional de uma extração *aleatória* como

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + u_i, \qquad \textbf{(17.43)}$$

em que usamos $\mathbf{x}_i\boldsymbol{\beta}$ como uma forma abreviada de $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$. Agora, seja n o tamanho de uma *amostra aleatória* da população. Se pudéssemos observar y_i e cada x_{ij} de todas as observações i, simplesmente usaríamos o MQO. Assuma que, por alguma razão, y_i ou algumas das variáveis independentes não sejam observadas para determinado i. Para ao menos algumas observações,

encontramos o conjunto completo de variáveis. Defina um *indicador de seleção* s_i de cada i como $s_i = 1$ se observarmos todos os $(y_i, \mathbf{x}_i)$, e $s_i = 0$ caso contrário. Assim, $s_i = 1$ indica que usaremos a observação em nossa análise; $s_i = 0$ significa que a observação não será usada. Estamos interessados nas propriedades estatísticas dos estimadores MQO usando a **amostra selecionada**, isto é, usando as observações de cada $s_i = 1$. Portanto, aproveitamos menos que n observações, digamos n_1.

Ocorre que é fácil obter condições sob as quais o MQO será consistente (e mesmo não viesado). Efetivamente, em vez de estimar (17.43), podemos somente estimar a equação

$$s_i y_i = s_i \mathbf{x}_i \boldsymbol{\beta} + s_i u_i. \qquad \textbf{(17.44)}$$

Quando $s_i = 1$, simplesmente teremos (17.43); quando $s_i = 0$, simplesmente teremos $0 = 0 + 0$, o que obviamente não nos diz nada a respeito de $\boldsymbol{\beta}$. Fazer a regressão de $s_i y_i$ sobre $s_i \mathbf{x}_i$ com $i = 1, 2, \ldots, n$ é o mesmo que fazer a regressão de y_i sobre $\mathbf{x}_i$ usando as observações para as quais $s_i = 1$. Assim, podemos verificar a consistência de $\hat{\beta}_j$ estudando (17.44) em uma amostra aleatória.

Conforme nossa análise no Capítulo 5, os estimadores MQO de (17.44) serão consistentes se o termo de erro tiver média zero e for não correlacionado com cada variável explicativa. Na população, a hipótese de média zero é $E(su) = 0$, e a hipótese de correlação zero pode ser estabelecida como

$$E[(sx_j)(su)] = E(sx_j u) = 0, \qquad \textbf{(17.45)}$$

em que s, x_j e u são variáveis aleatórias representando a população; usamos o fato de $s^2 = s$ porque s é uma variável binária. A condição (17.45) é diferente do que necessitamos se observarmos todas as variáveis de uma amostra aleatória: $E(x_j u) = 0$. Portanto, na população, precisamos que u seja não correlacionado com sx_j.

A condição mais importante para a inexistência de viés é $E(su|sx_1, \ldots, sx_k) = 0$. Como sempre, essa é uma hipótese mais forte do que a necessária para consistência.

Se s for uma função somente das variáveis explicativas, então, sx_j será apenas uma função de x_1, $x_2, \ldots, x_k$; pela hipótese da média condicional em (17.42), sx_j também será não correlacionada com u. Na verdade, $E(su|sx_1, \ldots, sx_k) = sE(u|sx_1, \ldots, sx_k) = 0$, pois $E(u|x_1, \ldots, x_k) = 0$. Esse é o caso da **seleção amostral exógena**, na qual $s_i = 1$ é determinado inteiramente por $x_{i1}, \ldots, x_{ik}$. Como um exemplo, ao estimarmos uma equação de salários na qual as variáveis explicativas sejam educação, experiência, permanência, gênero, estado civil, e assim por diante — que são assumidas como exógenas —, poderemos selecionar a amostra com base em qualquer ou todas as variáveis explicativas.

Se a seleção da amostra for inteiramente aleatória no sentido de que s_i é *independente* de $(\mathbf{x}_i, u_i)$, então $E(sx_j u) = E(s)E(x_j u) = 0$, pois $E(x_j u) = 0$ sob (17.42). Portanto, se começarmos com uma amostra aleatória e aleatoriamente eliminarmos observações, o MQO ainda será consistente. De fato, o MQO novamente será não viesado nesse caso, desde que não haja multicolinearidade perfeita na amostra selecionada.

Se s depender das variáveis explicativas e de termos aleatórios adicionais que sejam independentes de $\mathbf{x}$ e u, o MQO também será consistente e não viesado. Por exemplo, suponha que a pontuação do QI seja uma variável explicativa em uma equação de salários, mas que não esteja presente para algumas pessoas. Suponha que pensemos poder ser a seleção descrita por $s = 1$ se $QI \geq v$, e $s = 0$ se $QI < v$, em que v é uma variável aleatória não observada independente de QI, de u e das outras variáveis explicativas. Isso significa que é mais provável observarmos um QI que seja alto, mas sempre existe alguma probabilidade de não observarmos nenhum QI. Condicional às variáveis explicativas, s será independente de u, o que significa que $E(u|x_1, \ldots, x_k, s) = E(u|x_1, \ldots, x_k)$, e o último valor esperado será zero

por hipótese no modelo populacional. Se adicionarmos a hipótese de homoscedasticidade $E(u^2|\mathbf{x},s) = E(u^2) = \sigma^2$, os habituais erros-padrão e as estatísticas de testes do MQO serão válidos.

Até agora, mostramos várias situações nas quais o MQO na amostra selecionada é não viesado, ou pelo menos consistente. Quando o MQO na amostra selecionada será inconsistente? Já vimos um exemplo: a regressão usando uma amostra truncada. Quando o truncamento é acima, $s_i = 1$ se $y_i \leq c_i$, no qual c_i é o valor limite do truncamento. De forma equivalente, $s_i = 1$ se $u_i \leq c_i - \mathbf{x}_i\boldsymbol{\beta}$. Como s_i depende diretamente de u_i, s_i e u_i não serão não correlacionados, mesmo condicionais a $\mathbf{x}_i$. Essa é a razão pela qual o MQO na amostra selecionada não estima com consistência os β_j. Existem meios menos óbvios de s e u serem correlacionados; consideraremos isso na próxima subseção.

Os resultados sobre a consistência do MQO se estendem para a estimação de variáveis instrumentais. Se as VIs forem chamadas de z_h na população, a condição crucial para a consistência do MQ2E será $E(sz_h u) = 0$, que será válido se $E(u|\mathbf{z},s) = 0$. Portanto, se a seleção for determinada inteiramente pelas variáveis exógenas $\mathbf{z}$, ou se s depender de outros fatores que sejam independentes de u e de $\mathbf{z}$, então, o MQ2E na amostra selecionada geralmente será consistente. Temos que assumir que as variáveis explicativas e instrumentais são apropriadamente correlacionadas na parte selecionada da população. Wooldridge (2002, Capítulo 17) contém definições precisas dessas hipóteses.

Também pode ser mostrado que, quando a seleção é inteiramente uma função das variáveis exógenas, a estimação de máxima verossimilhança de um modelo não linear — tal como um modelo logit ou probit — produz estimadores consistentes e assimptoticamente normais, e os habituais erros-padrão e estatísticas de testes são válidos. [Novamente, veja Wooldridge (2002, Capítulo 17)].

Truncamento Ocasional

Como mencionamos anteriormente, uma forma comum de seleção amostral é chamada de truncamento ocasional. Começamos novamente com o modelo populacional em (17.42). Porém, assumimos que sempre observaremos as variáveis explicativas x_j. O problema é que somente observamos y para um subconjunto da população. A regra que determina se observamos y *não* depende diretamente do resultado de y. Um exemplo importante é quando $y = \log(salário°)$, no qual *salário*° é a *oferta de salário*, ou a remuneração por hora que um indivíduo poderia receber no mercado de trabalho. Se a pessoa estiver trabalhando no momento da pesquisa, observaremos a oferta de salário, porque assumimos que ela é o salário observado. Porém, para as pessoas fora da força de trabalho, não podemos observar o *salário*°. Portanto, o truncamento da oferta salarial é *ocasional*, pois ele depende de outra variável, ou seja, a participação na força de trabalho. É importante mencionar que geralmente observaremos todas as outras informações sobre um indivíduo, tais como educação, experiência anterior, gênero, estado civil etc.

A abordagem habitual para o truncamento ocasional é adicionar uma equação de seleção explícita ao modelo populacional de interesse:

$$y = \mathbf{x}\boldsymbol{\beta} + u, \; E(u|\mathbf{x}) = 0 \qquad \textbf{(17.46)}$$

$$s = 1[\mathbf{z}\boldsymbol{\gamma} + v \geq 0], \qquad \textbf{(17.47)}$$

em que $s = 1$ se observarmos y, e zero, caso contrário. Assumimos que elementos de $\mathbf{x}$ e $\mathbf{z}$ são sempre observados, e escrevemos $\mathbf{x}\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ e $\mathbf{z}\boldsymbol{\gamma} = \gamma_0 + \gamma_1 z_1 + \ldots + \gamma_m z_m$.

A equação de maior interesse é (17.46), e é possível estimar $\boldsymbol{\beta}$ por MQO, dada uma amostra aleatória. A equação de seleção (17.47), depende das variáveis observadas, z_h, e de um erro não observado, v. Uma hipótese padrão que faremos é que $\mathbf{z}$ é exógeno em (17.46):

$$E(u|\mathbf{x},\mathbf{z}) = 0$$

Na verdade, para que os métodos seguintes propostos funcionem bem, necessitaremos que $\mathbf{x}$ seja um subconjunto estrito de $\mathbf{z}$: qualquer x_j também é um elemento de $\mathbf{z}$, e temos alguns elementos de $\mathbf{z}$ que não estão, também, em $\mathbf{x}$. Veremos mais tarde por que isso é essencial.

Assume-se que o termo de erro v na equação de seleção amostral é independente de $\mathbf{z}$ (e portanto de $\mathbf{x}$). Também assumimos que v tem uma distribuição normal padrão. Podemos facilmente ver que a correlação entre u e v geralmente causa um problema de seleção amostral. Para ver o motivo, assuma que (u, v) seja independente de $\mathbf{z}$. Então, considerando o valor esperado de (17.46), condicional a $\mathbf{z}$ e v, e usando o fato de que $\mathbf{x}$ é um subconjunto de $\mathbf{z}$ produz

$$E(y|\mathbf{z},v) = \mathbf{x}\boldsymbol{\beta} + E(u|\mathbf{z},v) = \mathbf{x}\boldsymbol{\beta} + E(u|v),$$

em que $E(u|\mathbf{z},v) = E(u|v)$ porque (u, v) é independente de $\mathbf{z}$. Agora, se u e v forem conjuntamente normais (com média zero), $E(u|v) = \rho v$ para algum parâmetro ρ. Portanto,

$$E(y|\mathbf{z},v) = \mathbf{x}\boldsymbol{\beta} + \rho v.$$

Não observamos v, mas podemos usar essa equação para computar $E(y|\mathbf{z},s)$ e depois a limitarmos em $s = 1$. Agora temos:

$$E(y|\mathbf{z},s) = \mathbf{x}\boldsymbol{\beta} + \rho E(v|\mathbf{z},s).$$

Como s e v são relacionados por (17.47), e v tem uma distribuição normal padrão, podemos mostrar que $E(v|\mathbf{z},s)$ é simplesmente o inverso da razão de Mills, $\lambda(\mathbf{z}\boldsymbol{\gamma})$, quando $s = 1$. Isso leva à importante equação

$$E(y|\mathbf{z},s = 1) = \mathbf{x}\beta + \rho\lambda(\mathbf{z}\boldsymbol{\gamma}). \qquad \textbf{(17.48)}$$

A equação (17.48) mostra que o valor esperado de y, dados $\mathbf{z}$ *e* a observabilidade de y, é igual a $\mathbf{x}\boldsymbol{\beta}$, mais um termo adicional que depende do inverso da razão de Mills avaliado em $\mathbf{z}\boldsymbol{\gamma}$. Lembre-se de que esperamos estimar $\boldsymbol{\beta}$. Essa equação mostra que isso é possível usando somente a amostra selecionada, desde que incluamos o termo $\lambda(\mathbf{z}\boldsymbol{\gamma})$ como um regressor adicional.

Se $\rho = 0$, $\lambda(\mathbf{z}\boldsymbol{\gamma})$ não aparecerá, e o MQO de y sobre $\mathbf{x}$ usando a amostra selecionada estima consistentemente $\boldsymbol{\beta}$. Fora isso, omitimos efetivamente uma variável, $\lambda(\mathbf{z}\boldsymbol{\gamma})$, que geralmente está correlacionada com $\mathbf{x}$. Em que situação $\rho = 0$? A resposta é: quando u e v forem não correlacionados.

Como $\boldsymbol{\gamma}$ é desconhecido, não podemos avaliar $\lambda(\mathbf{z}_i\boldsymbol{\gamma})$ para cada i. Porém, com base nas hipóteses que fizemos, s dado z segue um modelo probit:

$$P(s = 1|\mathbf{z}) = \Phi(\mathbf{z}\boldsymbol{\gamma}). \qquad \textbf{(17.49)}$$

Portanto, podemos estimar $\boldsymbol{\gamma}$ pelo probit de s_i sobre $\mathbf{z}_i$, usando a amostra *inteira*. Em uma segunda etapa, poderemos estimar $\boldsymbol{\beta}$. Resumimos o procedimento, que recentemente foi batizado de **método Heckit** na literatura econométrica depois do trabalho de Heckman (1976).

Correção da Seleção Amostral

(i) Usando todas as n observações, estime um modelo probit de s_i sobre $\mathbf{z}_i$ e obtenha as estimativas $\hat{\gamma}_h$. Calcule o inverso da razão de Mills, $\hat{\lambda}_i = \lambda(\mathbf{z}_i\hat{\boldsymbol{\gamma}})$ para cada i. (Na realidade, somente necessitaremos desse cálculo para a observação i com $s_i = 1$).

(ii) Usando a amostra selecionada, ou seja, as observações nas quais $s_i = 1$ (digamos, n_1 delas), calcule a regressão de

$$y_i \text{ sobre } \mathbf{x}_i, \hat{\lambda}_i. \qquad \textbf{(17.50)}$$

Os $\hat{\beta}_j$ são consistentes e aproximadamente normalmente distribuídos.

Um teste simples do viés de seleção está disponível a partir da regressão (17.50). Em outras palavras, podemos usar a estatística t habitual de $\hat{\lambda}_i$ como um teste de H_0: $\rho = 0$. Sob H_0, não há problema de seleção de amostra.

Quando $\rho \neq 0$, os erros-padrão habituais do MQO descritos em (17.50) não serão exatamente corretos. Isso porque eles não explicam a estimação de $\boldsymbol{\gamma}$, que utiliza as mesmas observações na regressão (17.50), e por outros motivos mais. Alguns programas econométricos calculam corretamente os erros-padrão. [Infelizmente, isso não é tão simples quanto um ajuste de heteroscedasticidade. Veja Wooldridge (2002, Capítulo 6), para discussões adicionais.] Em muitos casos, os ajustes não levam a diferenças importantes, mas é difícil saber isso antecipadamente (a menos que $\hat{\rho}$ seja pequeno e não significante).

Há pouco mencionamos que $\mathbf{x}$ deveria ser um subconjunto estrito de $\mathbf{z}$. Isso traz duas implicações. Primeira, qualquer elemento que apareça como uma variável explicativa em (17.46) também deve ser uma variável explicativa na equação de seleção. Embora em raras ocasiões faça sentido excluir elementos da equação de seleção, não custa muito incluir todos os elementos de $\mathbf{x}$ em $\mathbf{z}$; a exclusão deles pode levar a inconsistências se eles forem excluídos incorretamente.

Uma segunda implicação importante é que temos pelo menos um elemento de $\mathbf{z}$ que não está também em $\mathbf{x}$. Isso significa que necessitamos de uma variável que afete a seleção, mas que *não* tenha um efeito parcial sobre y. Isso não é absolutamente necessário para aplicar o procedimento — de fato, podemos conduzir mecanicamente as duas etapas quando $\mathbf{z} = \mathbf{x}$ —, mas os resultados em geral serão menos que convincentes, a não ser que tenhamos uma *restrição de exclusão* em (17.46). A razão para isso é que, embora o inverso da razão de Mills seja uma função não linear de $\mathbf{z}$, ele frequentemente é bem aproximado por uma função linear. Se $\mathbf{z} = \mathbf{x}$, $\hat{\lambda}_i$ pode ser altamente correlacionado com os elementos de $\mathbf{x}_i$. Como sabemos, tal multicolinearidade pode conduzir a erros-padrão muito altos dos $\hat{\beta}_j$. Intuitivamente, se não tivermos uma variável que afete a seleção, exceto y, será extremamente difícil, se não impossível, distinguir seleção amostral de uma forma funcional mal-especificada em (17.46).

EXEMPLO 17.5

(Equação da Oferta de Salário para Mulheres Casadas)

Aplicamos a correção da seleção amostral aos dados sobre mulheres casadas contidos no arquivo MROZ.RAW. Lembre-se de que das 753 mulheres na amostra, 428 trabalharam por salário durante o ano. A equação da oferta de salário é padrão, com log(*salário*) como a variável dependente e *educ*, *exper* e $exper^2$ como as variáveis explicativas. Para testarmos e corrigirmos o viés da seleção amostral — em razão da impossibilidade de observar a oferta de salário para as mulheres que não trabalham —, precisamos estimar um modelo probit da participação na força de trabalho. Adicionalmente às variáveis educação e experiência, incluímos os fatores descritos na Tabela 17.1: outra renda, idade, número de filhos pequenos e número de filhos mais

EXEMPLO 17.5 (continuação)

velhos. O fato de essas quatro variáveis serem excluídas da equação de oferta de salário é uma *hipótese*: assumimos que, dados os fatores de produtividade, *nesprend*, *idade*, *crianmed6* e *crianma6* não têm efeito sobre a oferta de salário. É evidente, pelos resultados probit na Tabela 17.1, que pelo menos *idade* e *crianmed6* têm um forte efeito sobre a participação na força de trabalho.

A Tabela 17.5 contém os resultados do MQO e de Heckit. [Os erros-padrão dos resultados de Heckit são os mesmos erros-padrão habituais do MQO da regressão (17.50).] Não existe evidência de um problema de seleção amostral na estimativa da equação de oferta de salário. O coeficiente de $\hat{\lambda}$ tem uma estatística t bastante pequena (0,239), e assim não podemos rejeitar $H_0: \rho = 0$. De mesma importância, é o fato de não haver grandes diferenças práticas nos coeficientes de inclinação estimados na Tabela 17.5. Os retornos da educação estimados diferem somente em um décimo de ponto percentual.

Tabela 17.5

Equação da oferta de salário para mulheres casadas.

Variável dependente: log(*salário*)		
Variáveis independentes	**MQO**	**Heckit**
educ	0,108 (0,014)	0,109 (0,016)
exper	0,042 (0,012)	0,044 (0,016)
$exper^2$	−0,00081 (0,00039)	−0,00086 (0,00044)
constante	−0,522 (0,199)	−0,578 (0,307)
$\hat{\lambda}$	———	0,032 (0,134)
Tamanho da Amostra *R*-quadrado	428 0,157	428 0,157

Uma alternativa ao método precedente de estimação em duas etapas é a estimação de máxima verossimilhança completa. Ela é mais complicada, já que requer que obtenhamos a distribuição conjunta de y e s. Muitas vezes, faz sentido testar a seleção amostral usando o procedimento anterior; se não houver evidência de seleção amostral, não haverá razão para continuar. Se detectarmos viés de seleção amostral, poderemos tanto usar a estimativa em duas etapas como estimarmos conjuntamente as equações de regressão e seleção por EMV. [Veja Wooldridge (2002, Capítulo 17).]

No Exemplo 17.5, conhecemos mais do que apenas se uma mulher trabalhou durante o ano: sabemos quantas horas cada mulher trabalhou. Acontece que podemos usar essa informação em um procedimento alternativo de seleção amostral. Em lugar do inverso da razão de Mills $\hat{\lambda}_i$, usamos os resíduos

Tobit, digamos, $\hat{v}_i$, que é calculado como $\hat{v}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}$ sempre que $y_i > 0$. Pode ser mostrado que a regressão em (17.50) com $\hat{v}_i$ no lugar de $\hat{\lambda}_i$ também produz estimativas consistentes dos β_j, e que a estatística t padrão de $\hat{v}_i$ é um teste válido para o viés de seleção amostral. Essa abordagem tem a vantagem de usar mais informações, mas é menos aplicada. [Veja Wooldridge (2002, Capítulo 17).]

Há mais tópicos envolvendo a questão da seleção amostral. Um digno de ser mencionado é o de modelos com variáveis explicativas endógenas, *em adição ao* possível viés de seleção amostral. Escreva um modelo com uma única variável explicativa endógena como

$$y_1 = \alpha_1 y_2 + \mathbf{z}_1\boldsymbol{\beta}_1 + u_1, \qquad \textbf{(17.51)}$$

em que y_1 somente será observado quando $s = 1$, e y_2 poderá ser observado com y_1. Um exemplo é quando y_1 é a porcentagem de votos recebidos por um candidato, e y_2 é a porcentagem do total de gastos de campanha registrado pelo candidato. Para os candidatos que não concorrem, não poderemos observar y_1 ou y_2. Se tivermos fatores exógenos que afetem a decisão de concorrer e que estejam correlacionados com os gastos de campanha, poderemos estimar consistentemente α_1 e os elementos de $\boldsymbol{\beta}_1$ por variáveis instrumentais. Para ser convincente, precisamos de *duas* variáveis exógenas que não apareçam em (17.51). Efetivamente, uma deve afetar a decisão de seleção, e uma deve ser correlacionada com y_2 [a exigência normal para estimar (17.51) por MQ2E]. Resumidamente, o método é estimar a equação de seleção por probit, no qual *todas* as variáveis exógenas aparecem na equação probit. Depois, adicionamos o inverso da razão de Mills a (17.51) e estimamos a equação por MQ2E. O inverso da razão de Mills age como sua própria instrumental, já que depende somente de variáveis exógenas. Usamos todas as variáveis exógenas igual às outras instrumentais. Como antes, podemos usar a estatística t de $\hat{\lambda}_i$ na configuração de teste para o viés de seleção. [Veja Wooldridge (2002, Capítulo 17), para informações adicionais.]

RESUMO

Neste capítulo, estudamos vários métodos avançados que são frequentemente usados em aplicações, especialmente em microeconomia. Os modelos logit e probit são usados para variáveis de resposta binária. Esses modelos oferecem algumas vantagens sobre o modelo de probabilidade linear: as probabilidades estimadas estão entre zero e um, e os efeitos parciais decrescem. O principal custo do logit e do probit é que eles são mais difíceis de interpretar.

O modelo Tobit é aplicável a resultados não negativos que se acumulam em zero, mas que também assumem uma ampla gama de valores positivos. Muitas variáveis de escolha individual, tais como a oferta de mão de obra, o valor do seguro de vida, e o montante do fundo de pensão investido em ações, possuem essa característica. Assim como no logit e no probit, os valores esperados de y dado $\mathbf{x}$ — sejam condicionais a $y > 0$ ou incondicionais — dependem de $\mathbf{x}$ e de $\boldsymbol{\beta}$ de maneiras não lineares. Demos as expressões desses valores esperados, como também as fórmulas dos efeitos parciais de cada x_j sobre as expectativas. Elas poderão ser estimadas após o modelo Tobit ter sido estimado por máxima verossimilhança.

Quando a variável dependente é uma variável de contagem — isto é, ela assume valores inteiros não negativos —, um modelo de regressão de Poisson será apropriado. O valor esperado de y, dados os x_j, tem uma forma exponencial. Isso dá aos parâmetros interpretações como semielasticidades ou elasticidades, dependendo se os x_j estão em nível ou na forma logarítmica. Em resumo, podemos interpretar os parâmetros *como se* eles estivessem em um modelo linear com log(y) como a variável dependente.

Os parâmetros podem ser estimados por EMV. Porém, como a distribuição de Poisson impõe igualdade entre a variância e a média, frequentemente é necessário calcular erros-padrão e estatísticas de testes que admitam superdispersão ou subdispersão. Trata-se de simples ajustes dos habituais erros padrão e estatísticas da EMV.

Modelos de regressões censurada e truncada resolvem tipos específicos de problemas de ausência de dados. Na regressão censurada, a variável dependente é censurada acima ou abaixo de um valor limite. Podemos usar as informações sobre os resultados censurados porque sempre observamos as variáveis explicativas, como em aplicações de duração ou codificação superior de observações. Um modelo de regressão truncada surge quando uma parte da população é inteiramente excluída: não observamos qualquer informação em unidades que não estejam cobertas pelo esquema de amostragem. Este é um caso especial de problema de seleção amostral.

A Seção 17.5 oferece um tratamento sistemático da seleção amostral não aleatória. Mostramos que a seleção amostral exógena não afeta a consistência do MQO quando aplicada na subamostra, mas a seleção amostral endógena afeta. Mostramos como testar e corrigir o viés de seleção amostral para o problema geral do truncamento ocasional, no qual observações estão faltando em y em razão do resultado de outra variável (como a participação na força de trabalho). O método de Heckman é relativamente fácil de ser implementado nessas situações.

PROBLEMAS

17.1 (i) Para uma resposta binária y, seja $\bar{y}$ a proporção de uns na amostra (que é igual à média amostral de y_i). Sejam $\hat{q}_0$ a porcentagem corretamente prevista do resultado $y = 0$ e $\hat{q}_1$ a porcentagem corretamente prevista do resultado $y = 1$. Se $\hat{\rho}$ é a porcentagem global corretamente prevista, mostre que $\hat{\rho}$ é uma média ponderada de $\hat{q}_0$ e $\hat{q}_1$:

$$\hat{\rho} = (1 - \bar{y})\hat{q}_0 + \bar{y}\hat{q}_1.$$

(ii) Em uma amostra de 300 observações, suponha que $\bar{y} = 0{,}70$, de modo que existem 210 resultados com $y_i = 1$ e 90 com $y_i = 0$. Suponha que a porcentagem corretamente prevista quando $y = 0$ seja 80, e quando $y = 1$ seja 40. Encontre a porcentagem global corretamente prevista.

17.2 Defina *grad* como uma variável *dummy* informando se um estudante-atleta de uma grande universidade se formará em cinco anos. Sejam *GPA* e *SAT* a nota média do ensino médio e a nota do exame SAT, respectivamente. Defina *estudo* como o número de horas gastas por semana em uma sala de estudo organizada. Suponha que, usando os dados de 420 estudantes-atletas, obtenha-se o seguinte modelo logit:

$$\hat{\mathrm{P}}(grad = 1|GPA,SAT,estudo) = \Lambda(-1{,}17 + 0{,}24\ GPA + 0{,}00058\ SAT + 0{,}073\ estudo),$$

em que $\Lambda(z) = \exp(z)/[1 + \exp(z)]$ é a função logit. Mantendo fixos *GPA* em 3,0 e *SAT* em 1.200, calcule a diferença estimada na probabilidade de formatura de alguém que passou dez horas por semana em uma sala de estudo e de alguém que passou cinco horas por semana.

17.3 (Exige cálculo infinitesimal.)

(i) Suponha no modelo Tobit que $x_1 = \log(z_1)$, e que esse é o único lugar em que z_1 aparece em $\mathbf{x}$. Mostre que

$$\frac{\partial \mathrm{E}(y|y > 0,\mathbf{x})}{\partial z_1} = (\beta_1/z_1)\{1 - \lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)[\mathbf{x}\boldsymbol{\beta}/\sigma + \lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)]\}, \qquad \textbf{(17.52)}$$

em que β_1 é o coeficiente de $\log(z_1)$.

(ii) Se $x_1 = z_1$, e $x_2 = z_1^2$, mostre que

$$\frac{\partial \mathrm{E}(y|y > 0,\mathbf{x})}{\partial z_1} = (\beta_1 + 2\beta_2 z_1)\{1 - \lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)[\mathbf{x}\boldsymbol{\beta}/\sigma + \lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)]\},$$

em que β_1 é o coeficiente de z_1 e β_2 é o coeficiente de z_1^2.

17.4 Defina vpm_i como o valor do produto marginal do trabalhador i, que é o preço do bem de uma firma multiplicado pelo produto marginal do trabalhador. Assuma que

$$\log(vpm_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i$$
$$salário_i = \max(vpm_i, salmín_i),$$

em que estão incluídas como variáveis explicativas educação, experiência etc., e $salmín_i$ é o salário mínimo relevante para o indivíduo i. Escreva $\log(salário_i)$ em termos de $\log(vpm_i)$ e $\log(salmín_i)$.

17.5 (Exige cálculo infinitesimal.)
Defina *patentes* como o número de patentes requeridas por uma firma durante determinado ano. Assuma que o valor esperado condicional de *patentes*, dados *vendas* e *PD* é

$$\mathrm{E}(patentes|vendas,PD) = \exp[\beta_0 + \beta_1 \log(vendas) + \beta_2 PD + \beta_3 PD^2],$$

em que *vendas* representa as vendas anuais da firma e *PD* é o total de gastos com pesquisa e desenvolvimento nos últimos 10 anos.

(i) Como você estimaria os β_j? Justifique sua resposta detalhando a natureza de *patentes*.

(ii) Como você interpreta β_1?

(iii) Encontre o efeito parcial de *PD* sobre $\mathrm{E}(patentes|vendas,PD)$.

17.6 Considere uma função de poupança familiar para a população de todas as famílias dos Estados Unidos:

$$poup = \beta_0 + \beta_1 renda + \beta_2 tamfam + \beta_3 educ + \beta_4 idade + u,$$

em que *tamfam* é o tamanho da família, *educ* são anos de escolaridade do chefe da família e *idade* é a idade do chefe da família. Assuma que $\mathrm{E}(u|renda,tamfam,educ,idade) = 0$.

(i) Suponha que a amostra inclua apenas famílias cuja idade de seu chefe é superior a 25 anos. Se usarmos o MQO em tal amostra, obteremos estimadores não viesados dos β_j? Explique.

(ii) Agora, suponha que nossa amostra inclua somente casais sem filhos. Podemos estimar todos os parâmetros na equação de poupança? Quais podemos estimar?

(iii) Suponha que excluamos de nossa amostra as famílias que poupam mais de 25.000 dólares por ano. O MQO produzirá estimadores consistentes dos β_j?

17.7 Suponha que você seja contratado por uma universidade para estudar os fatores que determinam se os alunos admitidos na universidade matricularam-se efetivamente na universidade. Você recebe uma grande amostra aleatória dos alunos que foram admitidos no ano anterior. Também são disponibilizadas informações sobre se cada aluno decidiu matricular-se, o desempenho no ensino médio, a renda familiar, o auxílio financeiro oferecido, etnia e variáveis geográficas. Alguém lhe diz, "Qualquer análise desses dados conduzirá a resultados viesados, pois não se trata de uma amostra aleatória de todos os candidatos às universidades, mas somente daqueles que se candidataram nesta universidade". Qual sua opinião sobre essa crítica?

APÊNDICE 17A

Estimação De Máxima Verossimilhança Com Variáveis Explicativas

O Apêndice C (disponível no site da Cengage) fornece uma análise crítica da estimação da máxima verossimilhança (EMV) no caso mais simples de se estimar os parâmetros de uma distribuição incondicional. Mas a maioria dos modelos em econometria possuem variáveis explicativas, quer estimemos esses modelos pelos MQO, quer pela EMV. A última é indispensável para modelos não lineares, e vamos fornecer aqui uma descrição muito breve da abordagem geral.

Todos os modelos cobertos neste capítulo podem ser postos na seguinte forma. Considere que $f(y|\mathbf{x}, \boldsymbol{\beta})$ denote a função de densidade de uma extração aleatória y_i da população, condicional em $\mathbf{x}_i = \mathbf{x}$. O estimador da máxima verossimilhança (EMV) da $\boldsymbol{\beta}$ maximiza a função log-verossimilhança,

$$\max_{b} \sum_{i=1}^{n} \log f(y_i|\mathbf{x}_i, \mathbf{b}), \tag{17.53}$$

em que o vetor **b** é o argumento simulado no problema de maximização. Na maioria dos casos, o EMV, que escrevemos como $\hat{\boldsymbol{\beta}}$, é consistente e tem uma distribuição normal aproximada em amostras grandes. Isso é verdadeiro embora não possamos escrever uma fórmula para a $\hat{\boldsymbol{\beta}}$ exceto em circunstâncias muito especiais.

Em relação ao caso de resposta binária (logit e probit), a densidade condicional é determinada por dois valores, $f(1|\mathbf{x},\boldsymbol{\beta}) = \mathrm{P}(y_i = 1|\mathbf{x}_i) = G(\mathbf{x}_i\boldsymbol{\beta})$ e $f(0|\mathbf{x},\boldsymbol{\beta}) = \mathrm{P}(y_i = 0|\mathbf{x}_i) = 1 - \mathrm{G}(\mathbf{x}_i\boldsymbol{\beta})$. Aliás, uma maneira sucinta de se escrever a densidade é $f(y|\mathbf{x},\boldsymbol{\beta}) = [1 - G(\mathbf{x}\,\boldsymbol{\beta})]^{(1-y)}[G(\mathbf{x}\,\boldsymbol{\beta})]^{y}$ para $y = 0, 1$. Assim, podemos escrever a (17.53) da seguinte maneira

$$\max_{\mathrm{b}} \sum_{i=1}^{n} \{(1 - y_i)\log[1 - G(\mathbf{x}_i\mathbf{b}) + y_i\log[G(\mathbf{x}_i\mathbf{b})]\}. \tag{17.54}$$

De modo geral, as soluções para a (17.54) são rapidamente encontradas por computadores modernos utilizando-se métodos iterativos para se maximizar uma função. O tempo total de computação mesmo para conjuntos de dados razoavelmente grandes é caracteristicamente bastante rápido.

A função log-verossimilhança do modelo tobit e de regressões censuradas e truncadas são apenas um pouco mais complicadas, dependendo de um parâmetro de variação adicional em adição à $\boldsymbol{\beta}$. Elas são facilmente derivadas das densidades obtidas no texto. Para detalhes, veja Wooldridge (2002).

APÊNDICE 17B

Erros-Padrão Assimptóticos em Modelos de Variável Dependente Limitada

As derivações dos erros-padrão assimptóticos dos modelos e métodos apresentados neste capítulo estão bem além do escopo deste texto. Não apenas as deduções exigem álgebra matricial como também exigem teoria assimptótica avançada de estimação não linear. Os fundamentos necessários para uma análise cuidadosa desses métodos e das várias derivações são fornecidos em Wooldridge (2002).

É instrutivo ver as fórmulas de obtenção de erros-padrão assimptóticos de pelo menos alguns dos métodos. Dado um modelo de resposta binária $\mathrm{P}(y = 1|\mathbf{x}) = G(\mathbf{x}\boldsymbol{\beta})$, em que $G(\cdot)$ é a função logit ou probit, e $\boldsymbol{\beta}$ é o vetor de parâmetros $k \times 1$, a matriz de variância assimptótica de $\hat{\boldsymbol{\beta}}$ é estimada como

$$\widehat{\mathrm{Avar}}(\hat{\boldsymbol{\beta}}) \equiv \left(\sum_{i=1}^{n} \frac{[g(\mathbf{x}_i\hat{\boldsymbol{\beta}})]^2\mathbf{x}_i'\,\mathbf{x}_i}{G(\mathbf{x}_i\hat{\boldsymbol{\beta}})[1 - G(\mathbf{x}_i\hat{\boldsymbol{\beta}})]}\right)^{-1} \tag{17.55}$$

que é uma matriz $k \times k$. (Veja o Apêndice D, no site da Cengage, para um resumo de álgebra matricial.) Sem os termos que envolvem $g(\cdot)$ e $G(\cdot)$, essa fórmula se parece muito com a matriz de variância estimada do estimador MQO, exceto pelo termo $\hat{\sigma}^2$. A expressão em (17.55) leva em consideração a natureza não linear da probabilidade de resposta — isto é, a natureza não linear de $G(\cdot)$ — como também a forma particular de heteroscedasticidade em um modelo de resposta binária: $\mathrm{Var}(y|\mathbf{x}) = G(\mathbf{x}\boldsymbol{\beta})[1 - G(\mathbf{x}\boldsymbol{\beta})]$.

As raízes quadradas dos elementos diagonais de (17.55) são os erros-padrão assimptóticos dos $\hat{\beta}_j$, e eles são rotineiramente descritos por programas econométricos que suportam análises logit e probit. Dadas essas informações, estatísticas t (assimptóticas) e intervalos de confiança serão obtidos das maneiras habituais.

A matriz em (17.55) também é a base para os testes de Wald de restrições múltiplas de $\boldsymbol{\beta}$. [Veja Wooldridge (2002, Capítulo 15).]

A matriz de variância assimptótica do Tobit é mais complicada, mas tem uma estrutura semelhante. Observe que também podemos obter um erro-padrão para $\hat{\sigma}$. A variância assimptótica da regressão de Poisson, considerando $\sigma^2 \neq 1$ em (17.35), tem uma forma muito parecida com (17.55):

$$\widehat{\mathrm{Avar}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2\left(\sum_{i=1}^{n} \exp(\mathbf{x}_i\hat{\boldsymbol{\beta}})\mathbf{x}_i'\,\mathbf{x}_i\right)^{-1}. \tag{17.56}$$

As raízes quadradas dos elementos diagonais dessa matriz são os erros-padrão assimptóticos. Se a hipótese de Poisson se mantiver, podemos eliminar $\hat{\sigma}^2$ da fórmula (pois $\sigma^2 = 1$).

Os erros-padrão assimptóticos das regressões censurada e truncada, e da correção da seleção amostral de Heckit são mais complicados, embora compartilhem partes essenciais com as fórmulas anteriores. [Veja Wooldridge (2002) para mais detalhes.]