

APÊNDICE A

Ferramentas matemáticas básicas

Este apêndice cobre a matemática básica que é usada na análise econométrica. Resumimos várias propriedades do operador de soma, estudamos propriedades de equações lineares e determinadas equações não lineares e revisamos proporções e porcentagens. Também apresentamos algumas funções especiais que geralmente surgem na econometria aplicada, incluindo funções quadráticas e o logaritmo natural. As primeiras quatro seções exigem apenas habilidades de álgebra básicas. A Seção A.5 contém uma breve revisão do cálculo diferencial; embora não seja necessário um conhecimento de cálculo para entender a maior parte do texto, ele é usado em alguns apêndices do fim dos capítulos e em vários dos capítulos mais avançados da Parte 3.

A.1 Operador de soma e estatísticas descritivas

O **operador de soma** é uma abreviatura útil para manipular expressões que envolvem as somas de muitos números e tem papel fundamental nas análises estatísticas e econométricas. Se $\{x_i: i = 1, \dots, n\}$ denota uma sequência de n números, então, escrevemos a soma destes números como

$$\sum_{i=1}^n x_i \equiv x_1 + x_2 + \dots + x_n. \quad (\text{A.1})$$

Com esta definição, pode ser facilmente mostrado que o operador de soma tem as seguintes propriedades:

Propriedade de soma 1: Para qualquer c constante,

$$\sum_{i=1}^n c = nc. \quad (\text{A.2})$$

Propriedade de soma 2: Para qualquer c constante,

$$\sum_{i=1}^n cx_i = c \sum_{i=1}^n x_i. \quad (\text{A.3})$$

Propriedade de soma 3: Se $\{(x_i, y_i): i = 1, 2, \dots, n\}$ é um conjunto de n pares de números, e a e b são constantes, então

$$\sum_{i=1}^n (ax_i + by_i) = a \sum_{i=1}^n x_i + b \sum_{i=1}^n y_i. \quad (\text{A.4})$$

Também é importante estar ciente de algumas coisas que *não podem* ser feitas com o operador de soma. Defina $\{(x_i, y_i): i = 1, 2, \dots, n\}$ novamente como um conjunto de n pares de números, com $y_i \neq 0$ para cada i . Então,

$$\sum_{i=1}^n (x_i/y_i) \neq \left(\sum_{i=1}^n x_i \right) / \left(\sum_{i=1}^n y_i \right).$$

Em outras palavras, a soma das relações não é a relação das somas. No caso de $n = 2$, a aplicação de álgebra básica conhecida também revela essa ausência de igualdade: $x_1/y_1 + x_2/y_2 \neq (x_1 + x_2)/(y_1 + y_2)$. De forma similar, a soma dos quadrados não é o quadrado da soma: $\sum_{i=1}^n x_i^2 \neq (\sum_{i=1}^n x_i)^2$ exceto em casos especiais. Perceber que essas duas quantidades geralmente não são iguais é mais fácil quando $n = 2$: $x_1^2 + x_2^2 \neq (x_1 + x_2)^2 = x_1^2 + 2x_1x_2 + x_2^2$.

Dados n números $\{x_i: i = 1, \dots, n\}$, calculamos sua **média** somando-os e dividindo por n :

$$\bar{x} = (1/n) \sum_{i=1}^n x_i. \quad (\text{A.5})$$

Quando os x_i são uma amostra de dados sobre determinada variável (como anos de escolaridade), geralmente a chamamos de *média amostral* para enfatizar que foi calculada a partir de determinado conjunto de dados. A média amostral é um exemplo de **estatística descritiva**; neste caso, a estatística descreve a tendência central do conjunto de pontos x_i .

Existem algumas propriedades básicas relativas às médias que são importantes para se entender. Primeiro, suponha que selecionemos cada observação de x e subtraímos a média: $d_i \equiv x_i - \bar{x}$ (o “ d ” significa *desvio* em relação à média). Assim, a soma desses desvios é sempre zero:

$$\sum_{i=1}^n d_i = \sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x} = \sum_{i=1}^n x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0.$$

Resumimos como

$$\sum_{i=1}^n (x_i - \bar{x}) = 0. \quad (\text{A.6})$$

Um simples exemplo numérico mostra como isto funciona. Suponha que $n = 5$ e $x_1 = 6$, $x_2 = 1$, $x_3 = -2$, $x_4 = 0$ e $x_5 = 5$. Então, $\bar{x} = 2$ e a amostra sem a média é $\{4, -1, -4, -2, 3\}$. A soma dos valores dá zero, o que é exatamente o que a equação (A.6) mostra.

Em nossa abordagem sobre análises de regressão, no Capítulo 2, precisamos conhecer alguns fatos algébricos adicionais que envolvem desvios das médias amostrais.

Um aspecto importante é que a soma dos quadrados dos desvios é a soma de x_i ao quadrado diminuindo de n vezes o quadrado de $\bar{x}$:

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n(\bar{x})^2. \quad (\text{A.7})$$

Isso pode ser demonstrado usando propriedades básicas do operador de soma:

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n(\bar{x})^2 \\ &= \sum_{i=1}^n x_i^2 - 2n(\bar{x})^2 + n(\bar{x})^2 = \sum_{i=1}^n x_i^2 - n(\bar{x})^2. \end{aligned}$$

Dado um conjunto de dados sobre duas variáveis, $\{(x_i, y_i): i = 1, 2, \dots, n\}$, também se pode mostrar que

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^n x_i(y_i - \bar{y}) \\ &= \sum_{i=1}^n (x_i - \bar{x})y_i = \sum_{i=1}^n x_i y_i - n(\bar{x}\bar{y}); \end{aligned} \quad (\text{A.8})$$

esta é uma generalização da equação (A.7). (Lá, $y_i = x_i$ para todo i .)

A média é a medida da tendência central em que nos focaremos na maior parte do texto. No entanto, às vezes é informativo usar a **mediana** (ou *mediana amostral*) para descrever o valor central. Para obter a mediana dos n números $\{x_1, \dots, x_n\}$, primeiro ordenamos os valores de x_i do menor para o maior. Assim, se n for ímpar, a mediana amostral é o número do meio das observações ordenadas. Por exemplo, dados os números $\{-4, 8, 2, 0, 21, -10, 18\}$, o valor mediano é 2 (porque a sequência ordenada é $\{-10, -4, 0, 2, 8, 18, 21\}$). Se mudarmos o número maior desta lista, 21, para o dobro de seu valor, 42, a mediana ainda é 2. Em contrapartida, a média amostral aumentaria de 5 para 8, uma mudança considerável. Geralmente a mediana é menos sensível do que a média a alterações nos valores extremos (grandes ou pequenas) em uma lista de números. É por isso que “mediana de rendas” ou “valores medianos de residências” são reportados com frequência, em vez das médias, quando resumimos valores de renda ou residências em uma cidade ou estado.

Se n for par, não há uma forma única de definir a mediana porque existem dois números no centro. Geralmente a mediana é definida como a média dos dois valores centrais (novamente, depois de ordenar os números do menor para o maior). Usando esta regra, a mediana para o conjunto de números $\{4, 12, 2, 6\}$ seria $(4 + 6)/2 = 5$.

A.2 Propriedades das funções lineares

Funções lineares desempenham papel importante na econometria porque são simples de interpretar e manipular. Se x e y são duas variáveis relacionadas por

$$y = \beta_0 + \beta_1 x, \quad (\text{A.9})$$

então, podemos dizer que y é uma **função linear** de x , e β_0 e β_1 são dois parâmetros (números) que descrevem essa relação. O **intercepto** é β_0 e a **inclinação** é β_1 .

A característica definidora de uma função linear é que a variação de y é sempre β_1 vezes a variação em x :

$$\Delta y = \beta_1 \Delta x, \quad (\text{A.10})$$

onde Δ significa “variação”. Em outras palavras, o **efeito marginal** de x sobre y é constante e igual a β_1 .

EXEMPLO A.1 Função linear de gastos com habitação

Suponha que a relação entre os gastos mensais com habitação e a renda mensal seja

$$\text{habitação} = 164 + 0,27 \text{ renda}. \quad (\text{A.11})$$

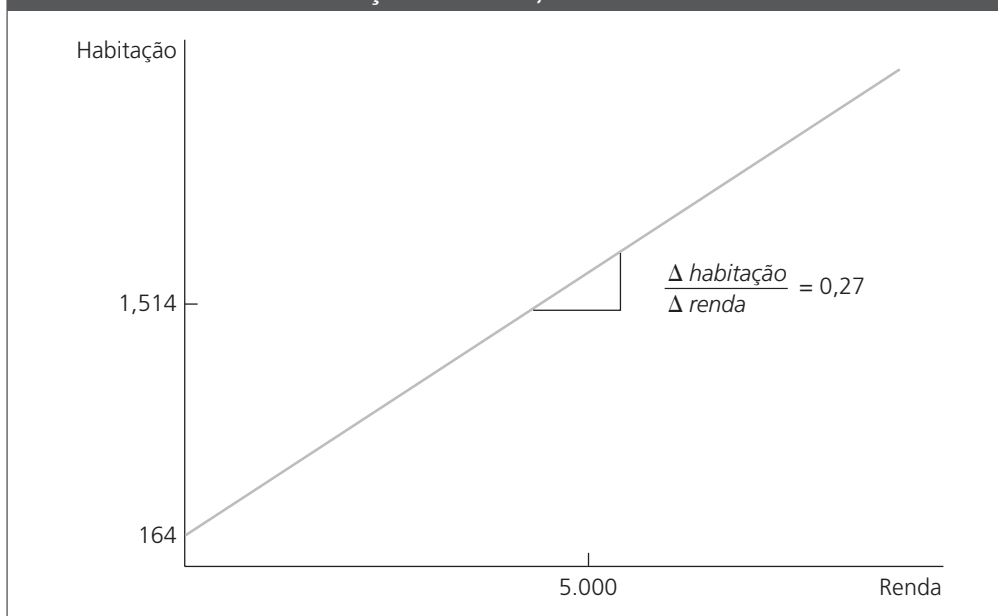
Assim, para cada dólar adicional de renda, 27 centavos são gastos com habitação. Se a renda familiar aumentar US\$ 200, os gastos com habitação aumentam $(0,27)200 = \text{US\$ } 54$. Essa função é representada graficamente na Figura A.1.

De acordo com a equação (A.11), uma família sem renda gasta US\$ 164 com habitação, o que, é claro, não pode ser literalmente verdade. Para níveis baixos de renda, essa função linear não descreveria muito bem a relação entre *habitação* e *renda*, e esse é o motivo para que tenhamos de usar outros tipos de funções para descrever essas relações.

Na equação (A.11), a *propensão marginal a consumir* (PMgC) de habitação pela renda é 0,27. Isso difere da *propensão média a consumir* (PmeC), que é

$$\frac{\text{habitação}}{\text{renda}} = 164/\text{renda} + 0,27.$$

FIGURA A.1 Gráfico de $\text{habitação} = 164 + 0,27 \text{ renda}$.



A PmeC não é constante; ela é sempre maior do que a PMgC e fica mais próxima dela à medida que a renda aumenta.

Funções lineares são facilmente definidas para mais de duas variáveis. Suponha que y esteja relacionado com duas variáveis, x_1 e x_2 , na forma geral

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2. \quad (\text{A.12})$$

É bem difícil visualizar essa função porque seu gráfico é tridimensional. No entanto, β_0 ainda é o intercepto (o valor de y quando $x_1 = 0$ e $x_2 = 0$), e β_1 e β_2 medem inclinações particulares. A partir da equação (A.12), a variação em y , para determinadas alterações em x_1 e x_2 , é

$$\Delta y = \beta_1 \Delta x_1 + \beta_2 \Delta x_2. \quad (\text{A.13})$$

Se x_2 não variar, isto é, $\Delta x_2 = 0$, então temos

$$\Delta y = \beta_1 \Delta x_1 \text{ se } \Delta x_2 = 0,$$

assim, β_1 é a inclinação da relação na direção de x_1 :

$$\beta_1 = \frac{\Delta y}{\Delta x_1} \text{ se } \Delta x_2 = 0.$$

Visto que ele mede como y muda com x_1 , mantendo x_2 fixo, β_1 é geralmente chamado de **efeito parcial** de x_1 sobre y . Como o efeito parcial envolve manter outros fatores fixos, ele é fortemente ligado à noção de *ceteris paribus*. O parâmetro β_2 tem uma interpretação similar: $\beta_2 = \Delta y / \Delta x_2$ se $\Delta x_1 = 0$, assim, β_2 é o efeito parcial de x_2 sobre y .

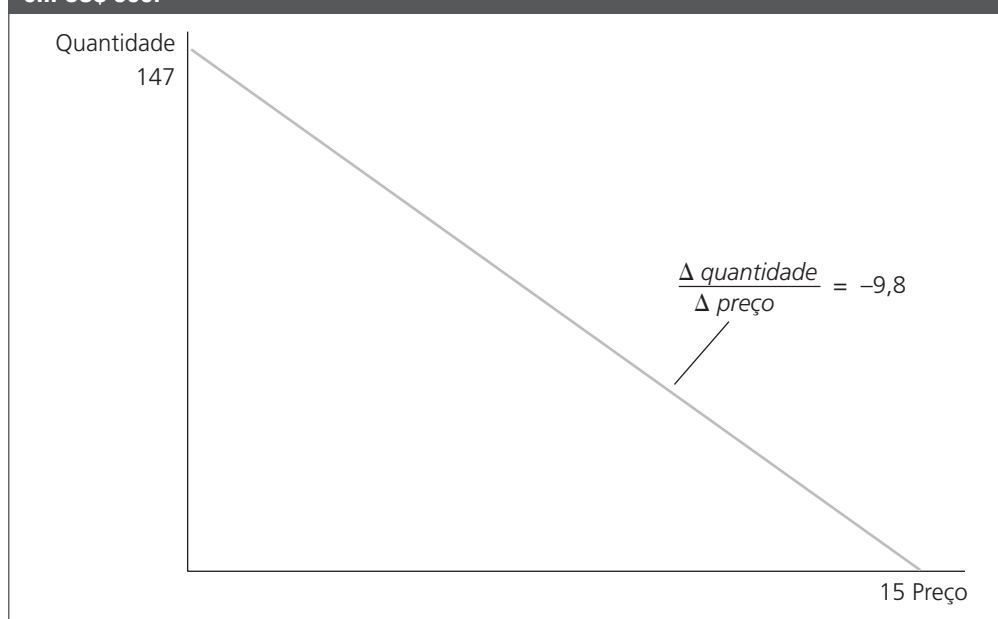
EXEMPLO A.2

Demanda por CDs

Para estudantes universitários, suponha que a quantia mensal demandada de CDs está relacionada ao preço dos CDs e à renda discricionária mensal por

$$\text{quantidade} = 120 - 9,8 \text{ preço} + 0,03 \text{ renda},$$

em que *preço* é o valor por disco e *renda* é medida em dólares. A *curva de demanda* é a relação entre *quantidade* e *preço*, mantendo *renda* (e outros fatores) fixa. Isso foi representado graficamente em duas dimensões na Figura A.2, com um nível de renda de US\$ 900. A inclinação da curva de demanda, $-9,8$, é o *efeito parcial* do preço sobre a quantidade: mantendo a renda fixa, se o preço dos CDs aumentar um dólar, a quantia demandada cai 9,8. (Não levamos em conta o fato de que CDs só podem ser comprados em unidades separadas.) Um aumento na renda simplesmente desloca a curva de demanda para cima (muda o intercepto), mas a inclinação continua a mesma.

FIGURA A.2 Gráfico de $\text{quantidade} = 120 - 9,8 \text{ preço} + 0,03 \text{ renda}$, com renda fixa em US\$ 900.

A.3 Proporções e porcentagens

Proporções e porcentagens têm papel tão importante na economia aplicada de tal forma, que é preciso estar muito seguro para trabalhar com elas. Muitas quantidades relatadas na imprensa popular estão na forma de porcentagem; alguns exemplos são taxas de juros, taxas de desemprego e índices graduados no ensino médio.

Uma habilidade importante é ser capaz de converter proporções em porcentagens e vice-versa. Uma porcentagem é facilmente obtida multiplicando-se uma proporção por 100. Por exemplo, se a proporção de adultos com ensino médio completo em um condado é de 0,82, podemos dizer que 82% (82 por cento) dos adultos têm diploma do ensino médio. Outra forma de pensar em porcentagens e proporções é que uma proporção é a forma decimal de uma porcentagem. Por exemplo, se a alíquota de imposto marginal de uma família que ganha US\$ 30.000 por ano é registrada como 28%, então a proporção para o próximo dólar de renda que é pago na forma de imposto sobre rendimentos é de 0,28 (ou 28¢).

Quando usamos porcentagens, geralmente precisamos convertê-las para a forma decimal. Por exemplo, se o imposto sobre vendas de um estado é 6% e forem gastos US\$ 200 em um item tributável, então o imposto pago sobre a venda é de $200(0,06) = \text{US\$ } 12$. Se o retorno anual sobre um certificado de depósito (CD) é de 7,6% e investimos US\$ 3.000 em um CD no início do ano, nosso rendimento de juros será $3.000(0,076) = \text{US\$ } 228$. Ainda que quiséssemos, o rendimento de juros não é obtido multiplicando 3.000 por 7,6.

Precisamos tomar cuidado com proporções que às vezes são reportadas incorretamente como porcentagens na mídia popular. Se lermos: “A porcentagem de estudantes do ensino médio que consome bebidas alcoólicas é 0,57”, sabemos que isso quer dizer

57% (e não pouco mais de meio por cento, como a afirmação indica literalmente). Fãs de voleibol universitário estão provavelmente familiarizados com cliques da imprensa que contêm afirmações como “Sua porcentagem de acertos foi 0,372”. Isso quer dizer, na verdade, que a porcentagem de acertos foi 37,2%.

Em econometria, frequentemente estamos interessados em medir as *variações* em diversas grandezas. Seja x uma variável que indique algo, como a renda de um indivíduo, o número de crimes cometidos em uma comunidade ou os lucros de uma empresa. Seja x_0 e x_1 dois valores designados para x : x_0 é o valor inicial e x_1 é o valor subsequente. Por exemplo, x_0 pode ser a renda anual de um indivíduo em 1994 e x_1 a renda do mesmo indivíduo em 1995. A **variação proporcional** de x ao se mover de x_0 a x_1 , às vezes chamada de **variação relativa**, é simplesmente

$$(x_1 - x_0)/x_0 = \Delta x/x_0, \quad (\text{A.14})$$

supondo, é claro, que $x_0 \neq 0$. Em outras palavras, para obter a variação proporcional, simplesmente dividimos a variação em x por seu valor inicial. Essa é uma forma de padronizar a mudança para que fique livre de unidades de medida. Por exemplo, se a renda de um indivíduo vai de US\$ 30.000 para US\$ 36.000 por ano, a variação proporcional é $6.000/30.000 = 0,20$.

É mais comum declarar alterações em termos de porcentagens. A **variação percentual** de x ao ir de x_0 para x_1 é simplesmente 100 vezes a variação proporcional:

$$\% \Delta x = 100(\Delta x/x_0); \quad (\text{A.15})$$

a notação “ $\% \Delta x$ ” é lida como “a variação percentual de x ”. Por exemplo, quando a renda vai de US\$ 30.000 a US\$ 33.750, ela aumentou 12,5%; para obter este valor, simplesmente multiplicamos a variação proporcional, 0,125, por 100.

Novamente, precisamos ficar atentos a variações proporcionais que são relatadas como variações percentuais. No exemplo anterior, relatar a variação percentual da renda como 0,125 é incorreto e pode levar a confusões.

Quando analisamos mudanças em coisas como montantes de dólar ou população, por exemplo, não há ambiguidade sobre o que significa alteração percentual. Em contrapartida, interpretar cálculos de variação percentual pode ser difícil quando a variável de interesse já é uma porcentagem, algo que ocorre frequentemente na economia e em outras ciências sociais. Para ilustrar, digamos que x denote a porcentagem de adultos que tenham educação superior em determinada cidade. Suponha que o valor inicial seja $x_0 = 24$ (24% tem curso superior), e o novo valor seja $x_1 = 30$. Podemos calcular duas medidas para descrever como a porcentagem de pessoas com curso superior mudou. A primeira é a mudança em x , Δx . Neste caso, $\Delta x = x_1 - x_0 = 6$: a porcentagem de pessoas com educação superior aumentou seis *pontos percentuais*. Por outro lado, podemos calcular a variação percentual de x usando a equação (A.15): $\% \Delta x = 100[(30 - 24)/24] = 25$.

Neste exemplo, a variação de pontos percentuais e a variação percentual são muito diferentes. A **variação de ponto percentual** é apenas a diferença entre as porcentagens. A variação percentual é a mudança em relação ao valor inicial. Normalmente devemos prestar muita atenção em qual número está sendo calculado. O pesquisador cuidadoso torna essa distinção perfeitamente clara; infelizmente, na imprensa popular, bem como em pesquisas acadêmicas, o tipo de variação registrada frequentemente é incerto.

EXEMPLO A.3 Aumento do imposto sobre vendas em Michigan

Em março de 1994, os eleitores de Michigan aprovaram um aumento de 4% para 6% do imposto sobre vendas. Nas propagandas políticas, os apoiadores da medida se referiam a isso como um aumento de dois pontos percentuais, ou um aumento de dois centavos sobre um dólar. Os que se opunham ao aumento do imposto alegavam que era um aumento de 50% do imposto sobre vendas. Ambas as afirmações estão corretas; elas são simplesmente formas diferentes de medir o aumento do imposto sobre vendas. Naturalmente, cada grupo divulgou a medida que tornava sua posição mais favorável.

Para uma variável como salário, não faz sentido falar em “variação de ponto percentual” porque o salário não é medido como porcentagem. Podemos descrever uma mudança salarial em termos financeiros ou de porcentagem.

A.4 Algumas funções especiais e suas propriedades

Na Seção A.2, revisamos as propriedades básicas das funções lineares. Já indicamos uma característica importante de funções como $y = \beta_0 + \beta_1 x$: uma mudança de uma unidade em x resulta na *mesma* variação em y , independentemente do valor inicial de x . Conforme notamos anteriormente, isso é o mesmo que dizer que o efeito marginal de x sobre y é constante, o que não é realista para muitas relações econômicas. Por exemplo, a importante noção econômica de *rendimentos marginais decrescentes* não condiz com uma relação linear.

Para modelar uma variedade de fenômenos econômicos, precisamos estudar diversas funções não lineares. Uma **função não linear** é caracterizada pelo fato de que a variação em y para determinada mudança em x depende do valor inicial de x . Certas funções não lineares surgem com frequência na economia empírica, por isso é importante saber como interpretá-las. Um entendimento completo das funções não lineares nos leva ao reino do cálculo. Aqui, vamos somente resumir os aspectos mais significativos das funções, deixando os detalhes de algumas derivações para a Seção A.5.

A.4a Funções quadráticas

Uma forma simples de capturar retornos decrescentes é adicionar um termo quadrático a uma relação linear. Considere a equação

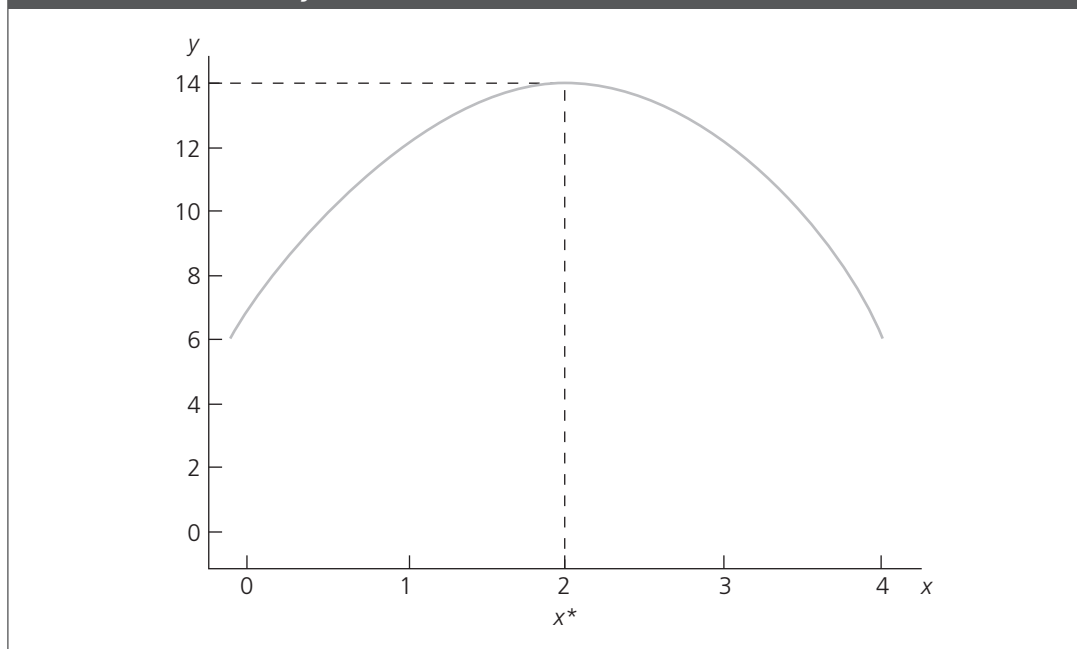
$$y = \beta_0 + \beta_1 x + \beta_2 x^2, \quad (\text{A.16})$$

em que β_0 , β_1 e β_2 são parâmetros. Quando $\beta_1 > 0$ e $\beta_2 < 0$, a relação entre y e x tem a forma parabólica dada na Figura A.3, onde $\beta_0 = 6$, $\beta_1 = 8$ e $\beta_2 = -2$.

Quando $\beta_1 > 0$ e $\beta_2 < 0$, pode-se mostrar (usando os cálculos da seção seguinte) que o *máximo* da função ocorre no ponto

$$x^* = \beta_1 / (-2\beta_2). \quad (\text{A.17})$$

Por exemplo, se $y = 6 + 8x - 2x^2$ (assim, $\beta_1 = 8$ e $\beta_2 = -2$), o maior valor de y ocorre em $x^* = 8/4 = 2$, e este valor é $6 + 8(2) - 2(2)^2 = 14$ (ver Figura A.3).

FIGURA A.3 Gráfico de $y = 6 + 8x - 2x^2$.

O fato de a equação (A.16) indicar um **efeito marginal decrescente** de x sobre y é facilmente visto neste gráfico. Suponha que comecemos com um valor baixo de x e o aumentemos em certa quantidade, digamos, c . Isso terá um efeito maior sobre y do que se começássemos com um valor alto de x e impuséssemos um aumento da mesma quantia c . De fato, uma vez que $x > x^*$, um aumento em x na verdade diminuirá y .

A afirmação de que x tem um efeito marginal decrescente sobre y é o mesmo que dizer que a inclinação da função na Figura A.3 diminui à medida que x aumenta. Embora isso fique claro ao olhar para o gráfico, geralmente queremos quantificar quão rapidamente a inclinação está mudando. Uma aplicação de cálculo mostra a inclinação aproximada da função quadrática como

$$\text{inclinação} = \frac{\Delta y}{\Delta x} \approx \beta_1 + 2\beta_2 x, \quad (\text{A.18})$$

para alterações “pequenas” em x . [O lado direito da equação (A.18) é a **derivada** da função na equação (A.16) em relação a x .] Outra forma de escrever isso é

$$\Delta y \approx (\beta_1 + 2\beta_2 x)\Delta x \text{ para “pequena” } \Delta x. \quad (\text{A.19})$$

Para ver como essa aproximação funciona bem, considere novamente a função $y = 6 + 8x - 2x^2$. Assim, de acordo com a equação (A.19), $\Delta y \approx (8 - 4x)\Delta x$. Agora, suponha que comecemos com $x = 1$ e alteremos x por $\Delta x = 0,1$. Usando a equação (A.19), $\Delta y \approx (8 - 4)(0,1) = 0,4$. É claro, podemos calcular a variação exata descobrindo os valores de y quando $x = 1$ e $x = 1,1$: $y_0 = 6 + 8(1) - 2(1)^2 = 12$ e $y_1 = 6 + 8(1,1) - 2(1,1)^2 = 12,38$, portanto, a variação exata de y é 0,38. A aproximação é bem próxima neste caso.

Agora, suponha que comecemos com $x = 1$, mas alteremos x em uma quantia maior: $\Delta x = 0,5$. Assim, a aproximação dá $\Delta y \approx 4(0,5) = 2$. A variação exata é determinada encontrando a diferença em y quando $x = 1$ e $x = 1,5$. O valor anterior de y era 12, e o valor atual é $6 + 8(1,5) - 2(1,5)^2 = 13,5$, assim, a variação real é 1,5 (não 2). A aproximação é pior neste caso porque a variação em x é maior.

Em muitas aplicações, a equação (A.19) pode ser usada para calcular o efeito marginal aproximado de x sobre y para qualquer valor inicial de x e pequenas alterações. E podemos sempre calcular a variação exata, se for necessário.

EXEMPLO A.4 Função salarial quadrática

Suponha que a relação entre o salário por hora e os anos na força de trabalho (*exper*) seja dada por

$$\text{salário} = 5,25 + 0,48 \text{ exper} - 0,008 \text{ exper}^2. \quad (\text{A.20})$$

Esta função tem a mesma forma geral que a da Figura A.3. Usando a equação (A.17), *exper* tem um efeito positivo sobre salário até o ponto de inflexão, $\text{exper}^* = 0,48/[2(0,008)] = 30$. O primeiro ano de experiência vale aproximadamente 0,48, ou 48 centavos [ver equação (A.19) com $x = 0$, $\Delta x = 1$]. Cada ano de experiência adicional aumenta o salário menos do que o ano anterior – refletindo um retorno marginal decrescente em relação à experiência. Com 30 anos, um ano adicional de experiência, na verdade, diminuirá o salário. Isso não é muito realista, mas é uma das consequências de usar uma função quadrática para capturar um efeito marginal decrescente: em algum ponto a função precisa atingir um máximo e se curvar para baixo. Para propósitos práticos, o ponto no qual isso acontece geralmente é grande o bastante para ser irrelevante, mas nem sempre.

O gráfico da função quadrática da equação (A.16) tem um formato de “U” se $\beta_1 < 0$ e $\beta_2 > 0$, caso em que há um retorno marginal crescente. O mínimo da função está no ponto $-\beta_1/(2\beta_2)$.

A.4b Logaritmo natural

A função não linear que desempenha o papel mais importante na análise econométrica é o **logaritmo natural**. Neste texto, vamos representar o logaritmo natural, ao qual nos referiremos com frequência como **função log**, como

$$y = \log(x). \quad (\text{A.21})$$

Você deve se lembrar de ter aprendido diferentes símbolos para o log natural; $\ln(x)$ ou $\log_e(x)$ são os mais comuns. Essas notações distintas são úteis quando logaritmos com várias bases diferentes estão sendo usados. Para nossos propósitos, apenas o logaritmo natural é importante, assim, $\log(x)$ indica o logaritmo natural ao longo do texto. Isso corresponde às sinalizações de muitos pacotes estatísticos, embora alguns usem $\ln(x)$ [e a maioria das calculadoras use $\ln(x)$]. Economistas usam tanto $\log(x)$ quanto $\ln(x)$, o que é útil para quando se trata de trabalhos de economia aplicada.

A função $y = \log(x)$ é definida apenas para $x > 0$, e isso está esboçado na Figura A.4. Não é muito importante saber como os valores de $\log(x)$ são obtidos. Para nossos

propósitos, a função pode ser pensada como uma caixa preta: podemos pegar qualquer $x > 0$ e obter $\log(x)$ em uma calculadora ou computador.

Várias coisas são aparentes na Figura A.4. Primeiro, quando $y = \log(x)$, a relação entre y e x apresenta retornos marginais decrescentes. Uma diferença importante entre o log e a função quadrática da Figura A.3 é que, quando $y = \log(x)$, o efeito de x sobre y nunca se torna negativo: a inclinação da função fica cada vez mais próxima de zero à medida que x aumenta, mas a inclinação nunca atinge zero e certamente não se torna negativa.

Os itens a seguir também surgem pela Figura A.4:

$$\log(x) < 0 \text{ para } 0 < x < 1$$

$$\log(1) = 0$$

$$\log(x) > 0 \text{ para } x > 1.$$

Em particular, $\log(x)$ pode ser positivo ou negativo. Alguns fatos algébricos úteis a respeito da função log são

$$\log(x_1 \cdot x_2) = \log(x_1) + \log(x_2), x_1, x_2 > 0$$

$$\log(x_1/x_2) = \log(x_1) - \log(x_2), x_1, x_2 > 0$$

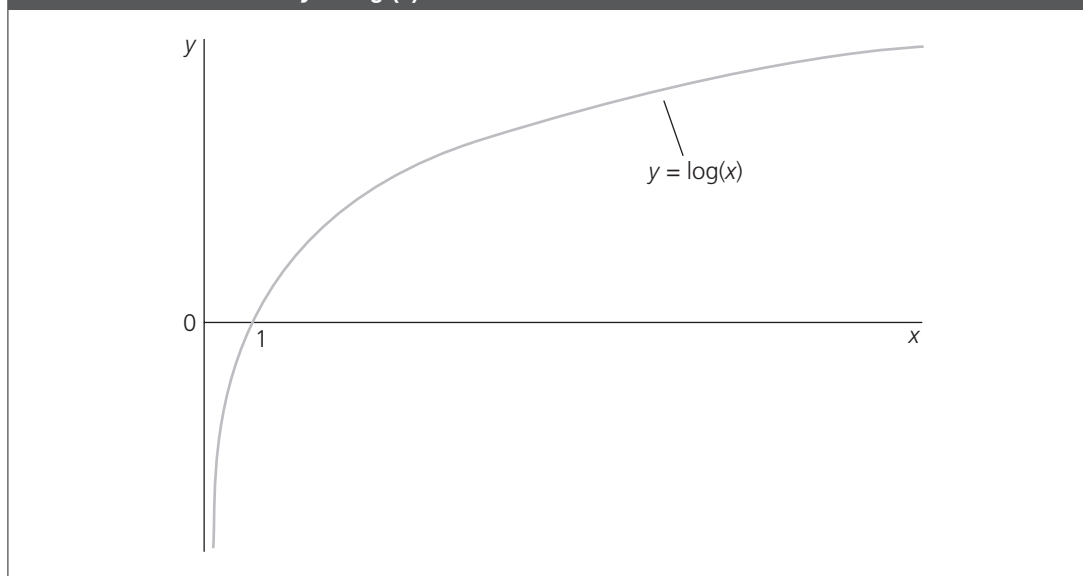
$$\log(x^c) = c \log(x), x > 0, \text{ qualquer número } c.$$

Às vezes precisaremos contar com essas propriedades.

O logaritmo pode ser usado para várias aproximações que surgem em aplicações econométricas. Primeiro, $\log(1 + x) \approx x$ para $x \approx 0$. Você pode testar isso com $x = 0,02, 0,1$ e $0,5$ para ver como a qualidade da aproximação se deteriora à medida que x aumenta. Ainda mais útil é o fato de que a diferença nos logs pode ser usada para aproximar variações proporcionais. Defina x_0 e x_1 como valores positivos. Então, pode ser mostrado (usando cálculos) que

$$\log(x_1) - \log(x_0) \approx (x_1 - x_0)/x_0 = \Delta x/x_0 \quad (\text{A.22})$$

FIGURA A.4 Gráfico de $y = \log(x)$.



Para pequenas alterações em x . Se multiplicarmos a equação (A.22) por 100 e escrevermos $\Delta \log(x) = \log(x_1) - \log(x_0)$, então

$$100 \cdot \Delta \log(x) \approx \% \Delta x \quad (\text{A.23})$$

para pequenas alterações em x . O significado de “pequenas” depende do contexto, e vamos encontrar vários exemplos ao longo deste texto.

Por que aproximamos a alteração percentual usando (A.23) quando a variação percentual exata é tão fácil de calcular? Em um momento vamos ver por que a aproximação da equação (A.23) é útil em econometria. Primeiro, vejamos quão boa é a aproximação em dois exemplos.

Suponha que $x_0 = 40$ e $x_1 = 41$. Então, a variação percentual de x ao mudar de x_0 para x_1 é 2,5%, usando $100(x_1 - x_0)/x_0$. Agora, $\log(41) - \log(40) = 0,0247$ (para quatro casas decimais), que, quando multiplicado por 100, fica bem próximo de 2,5. A aproximação funciona muito bem. Agora, considere uma variação muito maior: $x_0 = 40$ e $x_1 = 60$. A variação percentual exata é 50%. No entanto, $\log(60) - \log(40) \approx 0,4055$, então a aproximação dá 40,55%, que é bem distante.

Por que a aproximação da equação (A.23) será útil somente se for satisfatória para pequenas alterações? Para construir a resposta, primeiro definimos a **elasticidade** de y em relação a x como

$$\frac{\Delta y}{\Delta x} \cdot \frac{x}{y} = \frac{\% \Delta y}{\% \Delta x}. \quad (\text{A.24})$$

Em outras palavras, a elasticidade de y em relação a x é a variação percentual de y quando x aumenta 1%. Essa noção deve ser familiar da economia introdutória.

Se y é uma função linear de x , $y = \beta_0 + \beta_1 x$, então, a elasticidade é

$$\frac{\Delta y}{\Delta x} \cdot \frac{x}{y} = \beta_1 \cdot \frac{x}{y} = \beta_1 \cdot \frac{x}{\beta_0 + \beta_1 x}, \quad (\text{A.25})$$

que claramente depende do valor de x . (Essa é uma generalização do famoso resultado da teoria básica da demanda: a elasticidade não é constante ao longo de uma curva de demanda linear.)

Elasticidades têm uma importância crítica em muitas áreas da economia aplicada, não só na teoria da demanda. É conveniente em muitas situações ter modelos de elasticidade *constante*, e a função log nos permite especificar esses modelos. Se usarmos aproximação na equação (A.23) para x e y , a elasticidade será aproximadamente igual a $\Delta \log(y) / \Delta \log(x)$. Portanto, um modelo de elasticidade constante é aproximado pela equação

$$\log(y) = \beta_0 + \beta_1 \log(x), \quad (\text{A.26})$$

e β_1 é a elasticidade de y em relação a x (supondo que $x, y > 0$).

EXEMPLO A.5 Função de demanda com elasticidade constante

Se q é a quantia demandada e p é preço e essas variáveis se relacionam por

$$\log(q) = 4,7 - 1,25 \log(p),$$

então, a elasticidade-preço da demanda é $-1,25$. Aproximadamente um aumento de 1% no preço leva a uma queda de 1,25% da quantidade demandada.

Para nossos propósitos, o fato de β_1 na equação (A.26) ser apenas próximo da elasticidade não é importante. De fato, quando a elasticidade é definida usando cálculo – como na Seção A.5 –, a definição é exata. Para propósitos de análises econômicas, a equação (A.26) define um **modelo de elasticidade constante**. Este tipo de modelo tem um grande papel na economia empírica.

Outras possibilidades para usar a função log surgem com frequência no trabalho empírico. Suponha que $y > 0$ e

$$\log(y) = \beta_0 + \beta_1 x. \quad (\text{A.27})$$

Então, $\Delta \log(y) = \beta_1 \Delta x$, assim, $100 \cdot \Delta \log(y) = (100 \cdot \beta_1) \Delta x$. Verifica-se que, quando y e x são relacionados pela equação (A.27),

$$\% \Delta y \approx (100 \cdot \beta_1) \Delta x. \quad (\text{A.28})$$

EXEMPLO A.6**Equação logarítmica de salário**

Suponha que o salário por hora e os anos de educação estejam relacionados por

$$\log(\text{salário}) = 2,78 + 0,094 \text{ educ}.$$

Então, usando a equação (A.28),

$$\% \Delta \text{salário} \approx 100(0,094) \Delta \text{educ} = 9,4 \Delta \text{educ}.$$

Verifica-se que um ano a mais de educação aumenta o salário por hora em cerca de 9,4%.

Geralmente, a quantidade $\% \Delta y / \Delta x$ é chamada de **semielasticidade** de y em relação a x . A semielasticidade é a variação percentual em y quando x aumenta uma *unidade*. O que acabamos de mostrar é que, no modelo (A.27), a semielasticidade é constante e igual a $100 \cdot \beta_1$. No Exemplo A.6, podemos resumir convenientemente a relação entre salário e educação dizendo que um ano a mais de escolaridade – partindo de qualquer quantidade de educação – aumenta o salário em cerca de 9,4%. É por isso que esses modelos têm papel importante na economia.

Outra relação interessante para a economia aplicada é

$$y = \beta_0 + \beta_1 \log(x), \quad (\text{A.29})$$

onde $x > 0$. Como podemos interpretar esta equação? Se usarmos a variação em y , obtemos $\Delta y = \beta_1 \Delta \log(x)$, que pode ser reescrito como $\Delta y = (\beta_1/100)[100 \cdot \Delta \log(x)]$. Assim, usando a aproximação de (A.23), temos

$$\Delta y \approx (\beta_1/100)(\% \Delta x). \quad (\text{A.30})$$

Em outras palavras, $\beta_1/100$ é a variação unitária em y quando x aumenta 1%.

EXEMPLO A.7 Função de oferta de mão de obra

Suponha que a oferta de mão de obra de um trabalhador possa ser descrita como

$$\text{horas} = 33 + 45,1 \log(\text{salário}),$$

em que *salário* é a remuneração por hora e *horas* são as horas trabalhadas por semana. Então, a partir da equação (A.30),

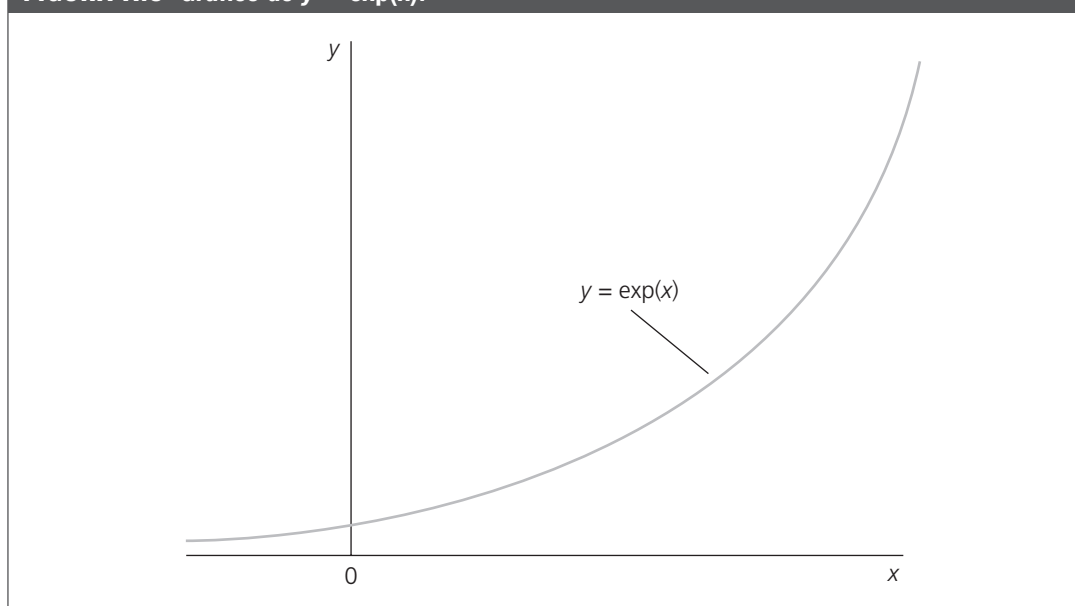
$$\Delta \text{horas} \approx (45,1/100)(\% \Delta \text{salário}) = 0,451 \% \Delta \text{salário}.$$

Em outras palavras, um aumento de 1% no *salário* aumenta as horas trabalhadas por semana em cerca de 0,45, ou pouco menos do que meia hora. Se o salário aumentasse 10%, então $\Delta \text{horas} = 0,451(10) = 4,51$, ou cerca de quatro horas e meia. Não gostaríamos de usar essa aproximação para variações percentuais muito grandes nos salários.

A.4c Função exponencial

Antes de finalizar esta seção, precisamos discutir uma função especial relacionada ao log. Como motivação, considere a equação (A.27). Lá, $\log(y)$ é uma função linear de x . Mas como encontramos y como uma função de x ? A resposta é dada pela **função exponencial**.

Escreveremos a função exponencial como $y = \exp(x)$, que foi colocada no gráfico da Figura A.5. A partir da Figura A.5, vemos que $\exp(x)$ é definida para qualquer valor de x e é sempre maior do que zero. Às vezes, a função exponencial é escrita como $y = e^x$, mas não usaremos essa notação. Dois valores importantes da função exponencial são $\exp(0) = 1$ e $\exp(1) = 2,7183$ (para quatro casas decimais).

FIGURA A.5 Gráfico de $y = \exp(x)$.

A função exponencial é o inverso da função log no seguinte sentido: $\log[\exp(x)] = x$ para todo x , e $\exp[\log(x)] = x$ para $x > 0$. Em outras palavras, log “cancela” a exponencial, e vice-versa. (É por isso que a função exponencial às vezes é chamada de função *antilog*.) Em especial, note que $\log(y) = \beta_0 + \beta_1 x$ é equivalente a

$$y = \exp(\beta_0 + \beta_1 x).$$

Se $\beta_1 > 0$, a relação entre x e y tem a mesma forma que a Figura A.5. Assim, se $\log(y) = \beta_0 + \beta_1 x$ com $\beta_1 > 0$, então x tem um efeito marginal *crescente* sobre y . No Exemplo A.6, isso quer dizer que um ano a mais de educação leva a uma mudança maior no salário do que a do número de anos de escolaridade anterior.

Dois fatos úteis sobre a função exponencial são $\exp(x_1 + x_2) = \exp(x_1)\exp(x_2)$ e $\exp[c \cdot \log(x)] = x^c$.

A.5 Cálculos diferenciais

Na seção anterior, expressamos uma série de aproximações fundamentadas em cálculo. Seja $y = f(x)$ para alguma função f . Assim, para pequenas alterações em x ,

$$\Delta y \approx \frac{df}{dx} \cdot \Delta x, \quad (\text{A.31})$$

em que df/dx é a derivada da função f , avaliada no ponto inicial x_0 . Também escrevemos a derivada como dy/dx .

Por exemplo, se $y = \log(x)$, então $dy/dx = 1/x$. Usando a equação (A.31), com dy/dx avaliada em x_0 , temos $\Delta y \approx (1/x_0)\Delta x$ ou $\Delta \log(x) \approx \Delta x/x_0$, que é a aproximação dada em (A.22).

Ao aplicar a econometria, ajuda se você se lembrar da derivada de uma porção de funções, porque usamos a derivada para definir a inclinação de uma função em um dado ponto. Podemos usar (A.31) para encontrar a variação aproximada de y para pequenas alterações em x . No caso linear, a derivada é simplesmente a inclinação da reta, como poderíamos esperar: se $y = \beta_0 + \beta_1 x$, então $dy/dx = \beta_1$.

Se $y = x^c$, então $dy/dx = cx^{c-1}$. A derivada de uma soma de duas funções é a soma das derivadas: $d[f(x) + g(x)]/dx = df(x)/dx + dg(x)/dx$. A derivada de uma constante vezes qualquer função é aquela mesma constante vezes a derivada da função: $d[cf(x)]/dx = c[df(x)/dx]$. Essas simples regras nos permitem encontrar derivadas de funções mais complicadas. Outras regras, como as de produto, quociente e cadeia, serão familiares para aqueles que já estudaram cálculo, mas não as revisaremos aqui.

Algumas funções que são usadas com frequência em economia, ao lado de suas derivadas, são

$$y = \beta_0 + \beta_1 x + \beta_2 x^2; \quad dy/dx = \beta_1 + 2\beta_2 x$$

$$y = \beta_0 + \beta_1/x; \quad dy/dx = -\beta_1/(x^2)$$

$$y = \beta_0 + \beta_1 \sqrt{x}; \quad dy/dx = (\beta_1/2)x^{-1/2}$$

$$y = \beta_0 + \beta_1 \log(x); \quad dy/dx = \beta_1/x$$

$$y = \exp(\beta_0 + \beta_1 x); \quad dy/dx = \beta_1 \exp(\beta_0 + \beta_1 x).$$

Se $\beta_0 = 0$ e $\beta_1 = 1$ nesta última expressão, obtemos $dy/dx = \exp(x)$, quando $y = \exp(x)$.

Na Seção A.4, notamos que a equação (A.26) define um modelo de elasticidade constante quando é usado cálculo. A definição da elasticidade em cálculo é $(dy/dx) \cdot (x/y)$. Pode ser demonstrado usando propriedades de logs e exponenciais que, quando (A.26) se mantém, $(dy/dx) \cdot (x/y) = \beta_1$.

Quando y é uma função de variáveis múltiplas, a noção de **derivada parcial** se torna importante. Suponha que

$$y = f(x_1, x_2). \quad (\text{A.32})$$

Assim, existem duas derivadas parciais, uma em relação a x_1 e outra em relação a x_2 . A derivada parcial de y em relação a x_1 , chamada aqui de $\partial y / \partial x_1$, é apenas a derivada usual de (A.32) em relação a x_1 , em que x_2 é tratada como uma *constante*. De forma similar, $\partial y / \partial x_2$ é apenas a derivada de (A.32) em relação a x_2 , mantendo x_1 fixo.

Derivadas parciais são úteis por quase as mesmas razões que derivadas comuns. Podemos aproximar a alteração em y como

$$\Delta y \approx \frac{\partial y}{\partial x_1} \cdot \Delta x_1, \text{ mantendo } x_2 \text{ fixo.} \quad (\text{A.33})$$

Assim, o cálculo nos permite definir efeitos parciais em modelos não lineares da mesma forma que em modelos lineares.

Na verdade, se

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

então,

$$\frac{\partial y}{\partial x_1} = \beta_1, \quad \frac{\partial y}{\partial x_2} = \beta_2.$$

Isso pode ser reconhecido como os efeitos parciais definidos na Seção A.2.

Um exemplo mais complicado é

$$y = 5 + 4x_1 + x_1^2 - 3x_2 + 7x_1 \cdot x_2. \quad (\text{A.34})$$

Agora, a derivada de (A.34), em relação a x_1 (tratando x_2 como uma constante), é simplesmente

$$\frac{\partial y}{\partial x_1} = 4 + 2x_1 + 7x_2;$$

note como ela depende de x_1 e x_2 . A derivada de (A.34), em relação a x_2 , é $\partial y / \partial x_2 = -3 + 7x_1$, assim, ela depende somente de x_1 .

EXEMPLO A.8 Função salarial com interação

Uma função que relaciona salário aos anos de educação e de experiência é

$$\begin{aligned} \text{salário} = & 3,10 + 0,41 \text{ educ} + 0,19 \text{ exper} - 0,004 \text{ exper}^2 \\ & + 0,007 \text{ educ} \cdot \text{exper}. \end{aligned} \quad (\text{A.35})$$

O efeito parcial de *exper* sobre *salário* é a derivada parcial de (A.35):

$$\frac{\partial \text{salário}}{\partial \text{exper}} = 0,19 - 0,008 \text{ exper} + 0,007 \text{ educ}.$$

Essa é a variação aproximada do salário devido ao aumento de um ano da experiência. Note que este efeito parcial depende do nível inicial de *exper* e de *educ*. Por exemplo, para um trabalhador que está começando com *educ* = 12 e *exper* = 5, o ano seguinte de experiência aumentará o salário em cerca de $0,19 - 0,008(5) + 0,007(12) = 0,234$ ou 23,4 centavos por hora. A variação exata pode ser calculada pela equação (A.35) com *exper* = 5, *educ* = 12 e *exper* = 6 e *educ* = 12, e então obtendo a diferença. O resultado é 0,23, muito perto da aproximação.

O cálculo diferencial desempenha o papel importante de minimizar e maximizar funções de uma ou mais variáveis. Se $f(x_1, x_2, \dots, x_k)$ é uma função diferenciável com k variáveis, então uma condição necessária para $x_1^*, x_2^*, \dots, x_k^*$ para minimizar ou maximizar f sobre todos os valores possíveis de x_j é

$$\frac{\partial f}{\partial x_j}(x_1^*, x_2^*, \dots, x_k^*) = 0, j = 1, 2, \dots, k. \quad (\text{A.36})$$

Em outras palavras, todas as derivadas parciais de f devem ser zero quando forem avaliadas em x_h^* . Elas são chamadas de *condições de primeira ordem* para minimizar ou maximizar uma função. De forma prática, esperamos resolver a equação (A.36) para x_h^* . Assim, podemos usar outros critérios para determinar se minimizamos ou maximizamos a função. Não precisaremos deles aqui. [Ver Sydsaeter e Hammond (1995) para uma discussão sobre cálculo multivariado e sua utilização para otimizar funções.]

Resumo

As ferramentas matemáticas revisadas aqui são cruciais para entender a análise de regressão e as probabilidades e estatísticas cobertas nos Apêndices B e C. O material sobre funções não lineares – especialmente funções quadráticas, logarítmicas e exponenciais – é crítico para entender a pesquisa econômica aplicada moderna. O nível de compreensão exigido dessas funções não inclui um conhecimento profundo de cálculo, embora ele seja necessário para algumas derivações.

Termos-chave

<i>Ceteris paribus</i>	Função linear	Modelo de elasticidade
Derivada	Função log	constante
Derivada parcial	Função não linear	Operador de soma
Efeito marginal	Inclinação	Semielasticidade
Efeito marginal decrescente	Intercepto	Varição de ponto percentual
Elasticidade	Logaritmo natural	Varição percentual
Estatística descritiva	Média	Varição proporcional
Função exponencial	Mediana	Varição relativa

Problemas

1. A tabela a seguir contém gastos mensais com habitação de 10 famílias.

Família	Gastos mensais com habitação (dólares)
1	300
2	440
3	350
4	1.100
5	640
6	480
7	450
8	700
9	670
10	530

- Encontre o gasto médio mensal com habitação.
 - Encontre o gasto mediano mensal com habitação.
 - Se os gastos mensais com habitação fossem medidos em centenas de dólares, em vez de em dólares, qual seriam os gastos médio e mediano?
 - Suponha que a família número 8 aumente seu gasto mensal com habitação para US\$ 900, mas os gastos de todas as outras famílias permaneçam os mesmos. Calcule os gastos médio e mediano com habitação.
2. Suponha que a seguinte equação descreva a relação entre o número médio de aulas perdidas durante um semestre (*faltas*) e a distância até a escola (*distância*, medida em milhas):

$$faltas = 3 + 0,2 \text{ distância}.$$

- Trace essa linha, certificando-se de indicar os eixos. Como você interpreta o coeficiente desta equação?
 - Qual é o número médio de aulas perdidas para alguém que mora a cinco milhas de distância?
 - Qual é a diferença entre o número médio de aulas perdidas por alguém que mora a 10 milhas e alguém que mora a 20 milhas de distância?
3. No Exemplo A.2, a quantidade de CDs foi relacionada ao preço e à renda por $quantidade = 120 - 9,8 \text{ preço} + 0,03 \text{ renda}$. Qual é a demanda por CDs se $preço = 15$ e $renda = 200$? O que isso sugere a respeito do uso de funções lineares para descrever curvas de demanda?
4. Suponha que a taxa de desemprego nos Estados Unidos tenha alterado de 6,4% em um ano para 5,6% no ano seguinte.
- Qual foi a queda em pontos percentuais na taxa de desemprego?
 - Em que porcentagem a taxa de desemprego caiu?
5. Suponha que o retorno por manter ações de determinada empresa vá de 15% em um ano para 18% no ano seguinte. O sócio majoritário alega que o “retorno das ações só cresceu 3%”, enquanto o CEO diz que o “retorno sobre as ações da empresa aumentou 20%”. Concilie essa divergência.
6. Suponha que a Pessoa A ganhe US\$ 35.000 por ano e a Pessoa B ganhe US\$ 42.000.
- Encontre a porcentagem exata pela qual o salário da Pessoa B excede o da Pessoa A.

- (ii) Agora, use a diferença em logs naturais para encontrar a diferença percentual aproximada.

7. Suponha que o modelo a seguir descreva a relação entre o salário anual (*salário*) e o número de anos anteriores de experiência no mercado de trabalho (*exper*):

$$\log(\text{salário}) = 10,6 + 0,027 \text{ exper}.$$

- (i) Qual é o *salário* quando *exper* = 0? E quando *exper* = 5? (Dica: Você precisará transformar em exponencial.)
- (ii) Use a equação (A.28) para aproximar o aumento percentual de *salário* quando *exper* aumenta cinco anos.
- (iii) Use os resultados do item (i) para calcular a diferença percentual exata do salário quando *exper* = 5 e *exper* = 0. Comente sobre como isso se compara com a aproximação do item (ii).
8. Seja *crescemp* o crescimento proporcional dos empregos, no âmbito do condado, de 1990 a 1995, e *impvendas* indica o imposto sobre vendas do condado, apresentado como uma proporção. Interprete o intercepto e a inclinação da equação

$$\text{crescemp} = 0,043 - 0,78 \text{ impvendas}.$$

9. Suponha que a produção de determinada colheita (em bushels por acre) esteja relacionada à quantidade de fertilizante (em libras por acre) como

$$\text{produção} = 120 + 0,19 \sqrt{\text{fertilizante}}.$$

- (i) Desenhe o gráfico desta relação adicionando diversos valores para *fertilizante*.
- (ii) Descreva como a forma dessa relação se compara com uma relação linear entre *produção* e *fertilizante*.
10. Suponha que, em determinado estado, um teste padrão seja aplicado a todos os formandos. Defina *nota* como a nota de um estudante no teste. Alguém descobriu que o desempenho no teste está relacionado com o tamanho da classe em que o estudante se formou no ensino médio. A relação é quadrática:

$$\text{nota} = 45,6 + 0,082 \text{ classe} - 0,000147 \text{ classe}^2,$$

onde *classe* é o número de alunos na turma do formando.

- (i) Como você interpreta literalmente o valor 45,6 da equação? Por si só, ele atrai muito interesse? Explique.
- (ii) A partir da equação, qual é o tamanho ideal da turma de formandos (o tamanho que maximiza a nota do teste)? (Arredonde sua resposta para o número inteiro mais próximo.) Qual é a maior nota possível para o teste?
- (iii) Esboce um gráfico que ilustre sua solução para o item (ii).
- (iv) Parece provável que *nota* e *classe* tenham uma relação determinística? Isto é, é realista pensar que, uma vez que saiba o tamanho da classe em que o aluno se formou, você saberá, com certeza, sua nota no teste? Explique.
11. Considere a linha

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

- (i) Faça com que (x_1, y_1) e (x_2, y_2) sejam dois pontos na linha. Mostre que $(\bar{x}, \bar{y})$ também está na linha, onde $\bar{x} = (x_1 + x_2)/2$ é a média dos dois valores e $\bar{y} = (y_1 + y_2)/2$.
- (ii) Amplie o resultado do item (i) para *n* pontos na linha, $\{(x_i, y_i): i = 1, \dots, n\}$.

APÊNDICE B

Fundamentos da probabilidade

Este apêndice engloba os conceitos-chave da probabilidade básica. Os Apêndices B e C são principalmente para revisão; eles não pretendem substituir um curso de probabilidade e estatística. No entanto, todos os conceitos de probabilidade e estatística que usamos no livro são abordados nestes apêndices.

A probabilidade por si só é interessante para estudantes de negócios, economia e outras ciências sociais. Por exemplo, considere o problema de uma companhia aérea que tenta decidir quantas reservas deve aceitar para um voo com 100 poltronas disponíveis. Se menos de 100 pessoas quiserem reservas, então todas devem ser aceitas. Mas e se mais de 100 pessoas solicitarem reservas? Uma solução segura é aceitar no máximo 100 reservas. Entretanto, como algumas pessoas fazem reserva e não comparecem para o voo, existem chances de o voo não estar cheio mesmo que as 100 reservas sejam vendidas. Isso resulta em perda de receita para a companhia. Uma estratégia diferente é agendar mais de 100 reservas e esperar que algumas pessoas não compareçam, assim, o número final de passageiros ficará o mais próximo possível de 100. Essa política traz o risco de a companhia ter de compensar as pessoas que forem necessariamente excluídas de um voo com *overbooking*.

Uma questão natural neste contexto é: Podemos decidir o número ideal (ou melhor) de reservas que a companhia aérea deve fazer? Este não é um problema trivial. No entanto, com determinadas informações (sobre custos da companhia e a frequência de comparecimento às reservas), podemos usar probabilidade básica para chegar a uma solução.

B.1 Variáveis aleatórias e suas distribuições de probabilidade

Suponha que joguemos uma moeda para o alto 10 vezes e contemos o número de vezes em que ela cai com o lado “cara” para cima. Este é um

exemplo de **experimento**. Geralmente, um experimento é qualquer procedimento que pode, pelo menos em teoria, ser infinitamente repetido e tem um conjunto de resultados bem definido. Podemos, a princípio, fazer o procedimento de jogada da moeda repetidas vezes. Antes de jogá-la, sabemos que o número de “caras” que vai surgir é um inteiro de 0 a 10, portanto, os resultados do experimento são bem definidos.

Uma **variável aleatória** é aquela que assume valores numéricos e tem um resultado determinado por um experimento. No exemplo do “cara ou coroa”, o número de “caras” em 10 jogadas de uma moeda é um exemplo de variável aleatória. Antes de jogar a moeda 10 vezes, não sabemos quantas vezes ela ficará com a “cara” para cima. Uma vez que tenhamos jogado a moeda 10 vezes e contado o número de caras, obtemos o resultado da variável aleatória para este determinado teste do experimento. Outro teste pode produzir um resultado diferente.

No exemplo das reservas da companhia aérea mencionado anteriormente, o número de pessoas que comparecem para embarcar no voo é uma variável aleatória: antes de qualquer voo em particular, não sabemos quantas pessoas vão comparecer.

Para analisar dados coletados em negócios e ciências sociais, é importante ter um entendimento básico das variáveis aleatórias e suas propriedades. Seguindo as convenções usuais de probabilidade e estatística ao longo dos Apêndices B e C, estipulamos variáveis aleatórias como letras maiúsculas, geralmente W , X , Y e Z ; resultados especiais de variáveis aleatórias são indicados pelas letras minúsculas correspondentes, w , x , y e z . Por exemplo, no experimento de “cara ou coroa”, faça X representar o número de “caras” em 10 jogadas da moeda. Dessa forma, X não está associada com algum valor em particular, mas sabemos que X vai assumir um valor no conjunto $\{0, 1, 2, \dots, 10\}$. Um resultado especial é, digamos, $x = 6$.

Indicamos grandes coleções de variáveis aleatórias usando subscrições. Por exemplo, se registramos a renda do último ano de 20 domicílios escolhidos aleatoriamente nos Estados Unidos, podemos indicar essas variáveis aleatórias como $X_1, X_2, \dots, X_{20}$; os resultados em particular serão marcados como $x_1, x_2, \dots, x_{20}$.

Como mostrado em sua definição, as variáveis aleatórias são sempre definidas para assumir valores numéricos, mesmo quando descrevem eventos qualitativos. Por exemplo, considere o arremesso de uma única moeda, em que os dois resultados são cara e coroa. Podemos definir uma variável aleatória como: $X = 1$ se a moeda cair com a “cara” para cima, e $X = 0$ se o resultado for “coroa”.

Uma variável aleatória que só pode assumir os valores zero e um é chamada de **variável aleatória de Bernoulli** (ou **binária**). Em probabilidade básica, é tradicional chamar o evento $X = 1$ de “sucesso” e o evento $X = 0$ de “fracasso”. Para aplicações em particular, a nomenclatura sucesso-fracasso pode não corresponder à nossa noção de sucesso ou fracasso, mas é uma terminologia útil que vamos adotar.

B.1a Variáveis aleatórias discretas

Uma **variável aleatória discreta** é aquela que assume apenas um número finito ou infinito contável de valores. A noção de “infinito contável” significa que, mesmo que um número infinito de valores possa ser assumido por uma variável aleatória, esses valores podem ser colocados em uma correspondência personalizada com os inteiros positivos. Como a distinção entre “infinito contável” e “infinito incontável” é algo sutil, vamos nos concentrar em variáveis aleatórias discretas que assumem somente um

número finito de valores. Larsen e Marx (1986, Capítulo 3) apresentam um tratamento detalhado.

Uma variável aleatória de Bernoulli é o exemplo mais simples de variável aleatória discreta. A única coisa que precisamos para descrever completamente o comportamento de uma variável aleatória de Bernoulli é a probabilidade que ela tem de assumir o valor um. No exemplo do “cara ou coroa”, se a moeda for “imparcial”, então $P(X = 1) = 1/2$ (leia como “a probabilidade de que X seja igual a um é meio”). Uma vez que as probabilidades precisam somar um, $P(X = 0)$ também é igual a $1/2$.

Cientistas sociais estão interessados em mais do que jogar moedas, por isso, precisamos levar em conta situações mais gerais. Novamente, considere o exemplo em que a companhia aérea deve decidir quantas pessoas agendar para um voo com 100 poltronas disponíveis. Este problema pode ser analisado no contexto de diversas variáveis aleatórias de Bernoulli da seguinte forma: para um consumidor selecionado aleatoriamente, defina uma variável aleatória de Bernoulli como $X = 1$ se a pessoa comparecer para a reserva, e $X = 0$ se não.

Não há motivo para pensar que a probabilidade de determinado consumidor aparecer é $1/2$; a princípio, a probabilidade pode ser qualquer número entre 0 e 1. Chame esse número de θ , de modo que

$$P(X = 1) = \theta \quad (\text{B.1})$$

$$P(X = 0) = 1 - \theta. \quad (\text{B.2})$$

Por exemplo, se $\theta = 0,75$, há uma chance de 75% de que o consumidor compareça depois de fazer a reserva e uma chance de 25% de ele não comparecer. De forma intuitiva, o valor de θ é crucial para determinar a estratégia da companhia aérea em relação às reservas. Métodos para *estimar* θ , com base em dados históricos das reservas da companhia, constituem um tema da estatística matemática, algo ao qual vamos nos voltar no Apêndice C.

De modo mais geral, qualquer variável aleatória discreta é completamente descrita ao listar seus possíveis valores e a probabilidade associada de que ela assuma cada um deles. Se X assume os k possíveis valores $\{x_1, \dots, x_k\}$, então as probabilidades $p_1, p_2, \dots, p_k$ são definidas por

$$p_j = P(X = x_j), j = 1, 2, \dots, k, \quad (\text{B.3})$$

em que cada p_j está entre 0 e 1 e

$$p_1 + p_2 + \dots + p_k = 1. \quad (\text{B.4})$$

A equação (B.3) é lida como: “A probabilidade de X assumir o valor x_j é igual a p_j ”.

As equações (B.1) e (B.2) mostram que as probabilidades de sucesso e fracasso de uma variável aleatória de Bernoulli são determinadas totalmente pelo valor de θ . Como as variáveis aleatórias de Bernoulli são tão prevalentes, temos uma notação especial para elas: $X \sim \text{Bernoulli}(\theta)$ é lida como “ X tem uma distribuição de Bernoulli com probabilidade de sucesso igual a θ ”.

A **função de densidade de probabilidade (fdp)** de X resume a informação em relação aos possíveis resultados de X e as probabilidades correspondentes:

$$f(x_j) = p_j, j = 1, 2, \dots, k, \quad (\text{B.5})$$

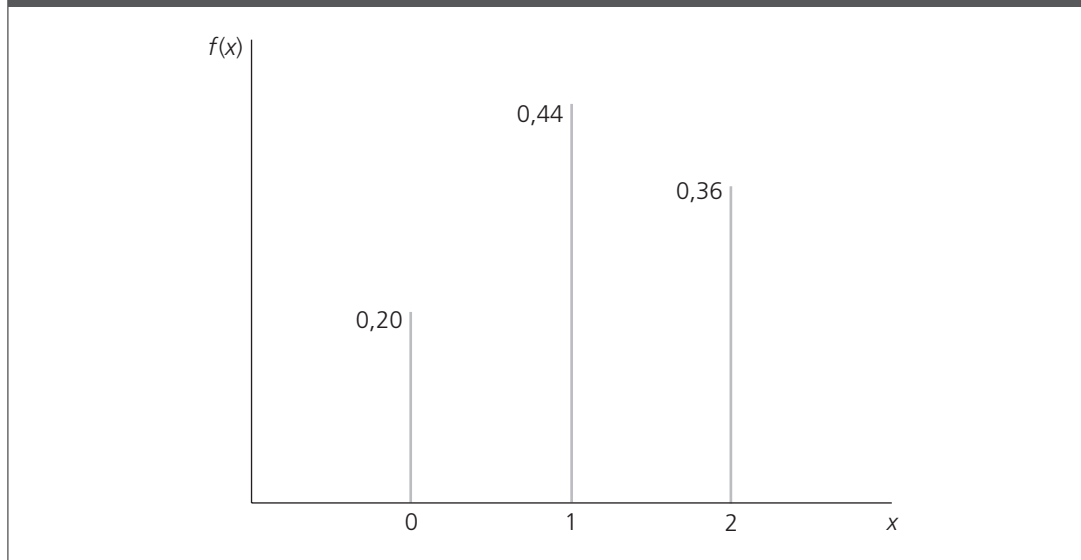
com $f(x) = 0$ para qualquer x que não for igual a x_j para algum j . Em outras palavras, para qualquer número real de x , $f(x)$ é a probabilidade de que a variável aleatória X assumira o valor determinado de x . Ao lidar com mais de uma variável aleatória, às vezes é útil subscrever a fdp em questão: f_X é a fdp de X , f_Y é a fdp de Y , e assim por diante.

Dada a fdp de qualquer variável aleatória discreta, é simples calcular a probabilidade de qualquer evento que envolva aquela variável aleatória. Por exemplo, suponha que X seja o número de lances livres feitos por um jogador de basquete em duas tentativas, de forma que X possa assumir os três valores $\{0, 1, 2\}$. Suponha que a fdp de X seja dada por

$$f(0) = 0,20, f(1) = 0,44, \text{ e } f(2) = 0,36.$$

As três probabilidades somam um, como deve ser. Usando essa fdp, podemos calcular a probabilidade de que o jogador converta *pelo menos* um lance livre: $P(X \geq 1) = P(X = 1) + P(X = 2) = 0,44 + 0,36 = 0,80$. A fdp de X é mostrada na Figura B.1.

FIGURA B.1 A fdp do número de lances livres feitos em duas tentativas.



B.1b Variáveis aleatórias contínuas

Uma variável X é uma **variável aleatória contínua** se assumir qualquer valor real com *zero* probabilidade. Essa definição é, de certa forma, contraintuitiva, porque em qualquer aplicação acabamos observando algum resultado para uma variável aleatória. A ideia é que uma variável aleatória contínua X possa assumir tantos valores possíveis que não somos capazes de contar ou conectá-los com os inteiros positivos, assim, a coerência lógica sentencia que X pode assumir cada valor com probabilidade zero. Enquanto medidas são sempre discretas na prática, variáveis aleatórias que assumem inúmeros valores são mais bem tratadas como contínuas. Por exemplo, a medida mais refinada do preço de um bem é em termos de centavos. Podemos imaginar uma lista de todos os possíveis valores de preço em ordem (embora a lista possa continuar indefinidamente), o que tecnicamente torna o preço uma variável aleatória discreta. No

entanto, existem tantos valores possíveis de preço que usar a mecânica de variáveis aleatórias discretas não é viável.

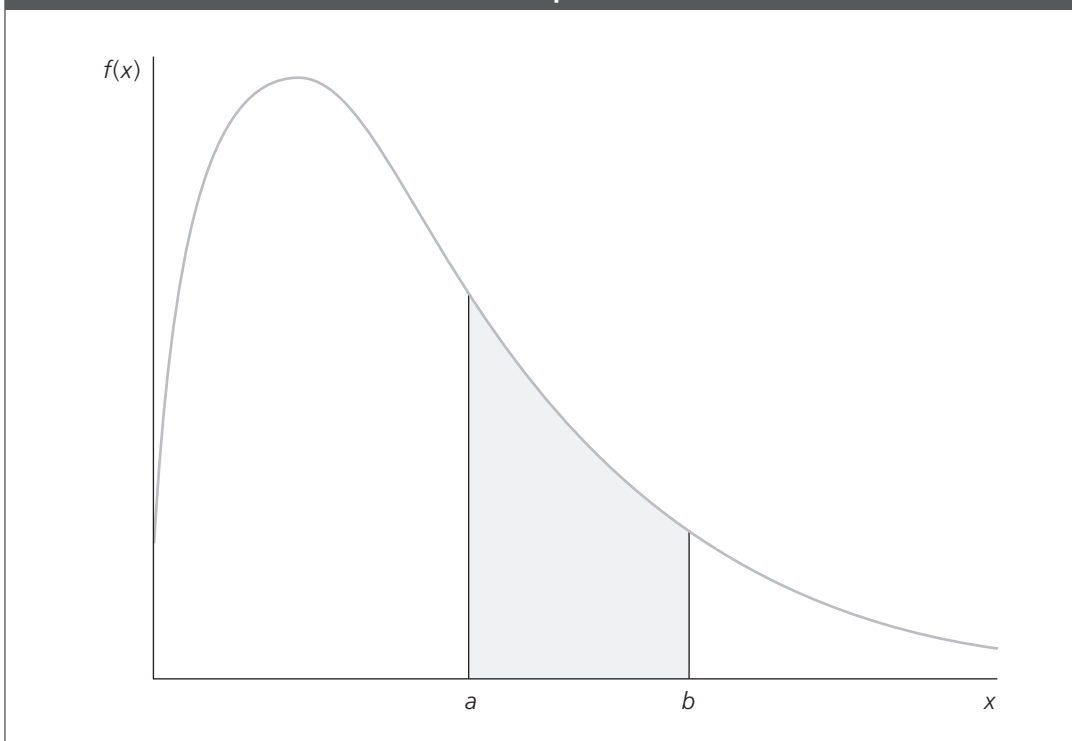
Podemos definir uma função de densidade de probabilidade para variáveis aleatórias contínuas e, assim como com as variáveis aleatórias discretas, a fdp proporciona informações sobre os possíveis resultados da variável aleatória. Entretanto, como não faz sentido discutir a probabilidade de uma variável aleatória contínua assumir determinado valor, usamos a fdp de uma variável aleatória contínua somente para calcular eventos que envolvam uma série de valores. Por exemplo, se a e b são constantes, onde $a < b$, a probabilidade de X estar entre os números a e b , $P(a \leq X \leq b)$, é a área abaixo da fdp entre os pontos a e b , como mostra a Figura B.2. Se estiver familiarizado com cálculo, você reconhecerá isso como a *integral* da função f entre os pontos a e b . Toda a área abaixo da fdp deve sempre ser igual a um.

Quando calculamos probabilidades para variáveis aleatórias contínuas, é mais fácil trabalhar com a **função de distribuição cumulativa (fdc)**. Se X for qualquer variável aleatória, então, sua fdc é definida para qualquer número real x por

$$F(x) \equiv P(X \leq x). \quad (\text{B.6})$$

Para variáveis aleatórias discretas, (B.6) é obtida somando a fdp de todos os valores x_j de forma que $x_j \leq x$. Para uma variável aleatória contínua, $F(x)$ é a área abaixo da fdp, f , à esquerda do ponto x . Como $F(x)$ é apenas uma probabilidade, está sempre entre 0 e 1. Além disso, se $x_1 < x_2$, então $P(X \leq x_1) \leq P(X \leq x_2)$, isto é, $F(x_1) \leq F(x_2)$. Isso quer dizer que uma fdc é uma função crescente (ou pelo menos não decrescente) de x .

FIGURA B.2 Probabilidade de X estar entre os pontos a e b .



Duas importantes propriedades das fdcs que são úteis para calcular probabilidades são as seguintes:

$$\text{Para qualquer número } c, P(X > c) = 1 - F(c). \quad (\text{B.7})$$

$$\text{Para quaisquer números } a < b, P(a < X \leq b) = F(b) - F(a). \quad (\text{B.8})$$

Em nossos estudos de econometria, usaremos fdc para calcular probabilidade apenas de variáveis aleatórias contínuas, caso em que não importa se as diferenças nas afirmações de probabilidade são exatas ou não. Isto é, para uma variável aleatória contínua X ,

$$P(X \geq c) = P(X > c), \quad (\text{B.9})$$

e

$$P(a < X < b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b). \quad (\text{B.10})$$

Combinadas com (B.7) e (B.8), as equações (B.9) e (B.10) expandem muito os cálculos de probabilidade que podem ser realizados usando fdc contínuas.

Funções de distribuição cumulativas para todas as distribuições contínuas importantes da probabilidade e estatística têm sido apuradas. A mais conhecida delas é a distribuição normal, que abordaremos ao lado de algumas distribuições relacionadas na Seção B.5.

B.2 Distribuições conjuntas, distribuições condicionais e independência

Na economia, geralmente estamos interessados na ocorrência de eventos que envolvem mais de uma variável aleatória. Por exemplo, no caso das reservas da companhia aérea citado anteriormente, a companhia pode estar interessada na probabilidade de que uma pessoa que faz uma reserva compareça e seja um viajante a negócios; este é um exemplo de *probabilidade conjunta*. Ou a companhia pode estar interessada na seguinte *probabilidade condicional*: com a condição de que a pessoa seja um viajante a negócios, qual é a probabilidade de ela comparecer? Nas próximas duas subseções, formalizaremos as noções de distribuições conjunta e condicional e a importante noção de *independência* de variáveis aleatórias.

B.2a Distribuições conjuntas e independência

Suponha que X e Y sejam variáveis aleatórias discretas. Então, (X, Y) tem uma **distribuição conjunta**, que é totalmente descrita pela *função de densidade de probabilidade conjunta* de (X, Y) :

$$f_{X,Y}(x, y) = P(X = x, Y = y), \quad (\text{B.11})$$

onde o lado direito é a probabilidade de $X = x$ e $Y = y$. Quando X e Y são contínuas, uma fdp conjunta também pode ser definida, mas não abordaremos estes detalhes porque fdps conjuntas para variáveis aleatórias contínuas não são explicitamente usadas neste livro.

Em alguns casos, é fácil obter a fdp conjunta se tivermos as fdps de X e Y . Em especial, variáveis aleatórias X e Y são classificadas como independentes se, e somente se,

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \quad (\text{B.12})$$

para todo x e y , onde f_X é a fdp de X e f_Y é a fdp de Y . No contexto de mais de uma variável aleatória, as fdps f_X e f_Y normalmente são chamadas de *funções de densidade de probabilidade marginal* para diferenciá-las da fdp conjunta $f_{X,Y}$. Essa definição de independência é válida para variáveis aleatórias discretas e contínuas.

Para entender o significado de (B.12), é mais fácil lidar com o caso discreto. Se X e Y são discretas, então (B.12) é o mesmo que

$$P(X = x, Y = y) = P(X = x)P(Y = y); \quad (\text{B.13})$$

em outras palavras, a probabilidade de que $X = x$ e $Y = y$ é o produto das duas probabilidades $P(X = x)$ e $P(Y = y)$. Uma implicação de (B.13) é que as probabilidades conjuntas são bem fáceis de calcular, já que elas exigem apenas o conhecimento de $P(X = x)$ e $P(Y = y)$.

Se as variáveis aleatórias não forem independentes, então são chamadas de *dependentes*.

EXEMPLO B.1

Arremesso de lances livres

Considere um jogador de basquete arremessando dois lances livres. Defina X como uma variável aleatória de Bernoulli igual a um se ele converter o primeiro lance livre, e zero no caso contrário. Defina Y como uma variável aleatória de Bernoulli igual a um se ele converter o segundo lance livre. Suponha que o atleta seja um arremessador de lances livres de 80%, de forma que $P(X = 1) = P(Y = 1) = 0,8$. Qual é a probabilidade de o jogador converter os dois lances livres?

Se X e Y forem independentes, conseguimos responder a essa questão facilmente: $P(X = 1, Y = 1) = P(X = 1)P(Y = 1) = (0,8)(0,8) = 0,64$. Assim, há 64% de chance de converter os dois lances livres. Se a chance de converter o segundo lance livre depender de como o primeiro foi feito – isto é, X e Y não forem independentes – este cálculo simples não é válido.

A independência de variáveis aleatórias é um conceito muito importante. Na próxima subseção, mostraremos que, se X e Y forem independentes, conhecer o resultado de X não muda as probabilidades dos possíveis resultados de Y , e vice-versa. Um fato útil sobre a independência é que se X e Y fossem independentes e definíssemos novas variáveis aleatórias $g(X)$ e $h(Y)$ para quaisquer funções g e h , essas novas variáveis também seriam independentes.

Não há necessidade de parar em duas variáveis aleatórias. Se $X_1, X_2, \dots, X_n$ são variáveis aleatórias discretas, sua fdp conjunta é $f(x_1, x_2, \dots, x_n) = P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$. As variáveis aleatórias $X_1, X_2, \dots, X_n$ são **variáveis aleatórias independentes** se, e somente se, sua fdp conjunta for o produto das fdps individuais para qualquer $(x_1, x_2, \dots, x_n)$. Essa definição de independência também se mantém para variáveis aleatórias contínuas.

A noção de independência desempenha importante papel na obtenção de algumas das distribuições clássicas da probabilidade e da estatística. Anteriormente, definimos uma variável aleatória de Bernoulli como uma variável aleatória zero-um, que indica se algum evento ocorreu ou não. Com frequência estamos interessados no número de sucessos em uma sequência de testes de Bernoulli *independentes*. Um exemplo padrão de teste de Bernoulli independente é jogar uma moeda repetidas vezes. Como o resultado de uma jogada não tem nada a ver com os resultados das outras jogadas, a independência é uma hipótese apropriada.

A independência geralmente é uma aproximação razoável em situações mais complicadas. No exemplo das reservas da companhia aérea, suponha que a empresa aceite n reservas para determinado voo. Para cada $i = 1, 2, \dots, n$, defina Y_i como a variável aleatória de Bernoulli que indica se o consumidor i vai comparecer: $Y_i = 1$ se ele comparecer, e $Y_i = 0$ no caso contrário. Deixando que θ represente a probabilidade de sucesso (usando reservas), cada Y_i tem uma distribuição de Bernoulli (θ). Como aproximação, podemos presumir que as Y_i sejam independentes umas das outras, embora isso não seja exatamente verdade: algumas pessoas viajam em grupos, o que quer dizer que o fato de uma pessoa comparecer (ou não) não é verdadeiramente independente do fato de todos os outros comparecerem. No entanto, criar modelos para esse tipo de dependência é complexo, por isso preferimos usar a independência como uma aproximação.

A variável de principal interesse é o número total de consumidores que vai comparecer entre as n reservas; chame essa variável de X . Como cada Y_i é unitária quando uma pessoa comparece, podemos escrever $X = Y_1 + Y_2 + \dots + Y_n$. Agora, supondo que cada Y_i tenha probabilidade θ de sucesso e que as Y_i sejam independentes, pode-se mostrar que X tem uma **distribuição binomial**. Ou seja, a função de densidade de probabilidade de X é

$$f(x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}, x = 0, 1, 2, \dots, n, \quad (\text{B.14})$$

onde $\binom{n}{x} = \frac{n!}{x!(n-x)!}$, e para qualquer inteiro n , $n!$ (ler “ n fatorial”) é definido como $n! = n \cdot (n-1) \cdot (n-2) \cdots 1$. Por convenção, $0! = 1$. Quando uma variável aleatória X tem a fdp dada em (B.14), escrevemos $X \sim \text{Binomial}(n, \theta)$. A equação (B.14) pode ser usada para calcular $P(X = x)$ para qualquer valor de x de 0 a n .

Se o voo tem 100 poltronas disponíveis, a companhia aérea está interessada em $P(X > 100)$. Suponha, inicialmente, que $n = 120$, portanto, a companhia aceitou 120 reservas, e a probabilidade de cada pessoa comparecer é $\theta = 0,85$. Assim, $P(X > 100) = P(X = 101) + P(X = 102) + \dots + P(X = 120)$, e cada uma das probabilidades da soma pode ser encontrada pela equação (B.14) com $n = 120$, $\theta = 0,85$ e o valor adequado de x (101 a 120). Este é um cálculo manual difícil, mas muitos pacotes estatísticos têm comandos para calcular esse tipo de probabilidade. Neste caso, a probabilidade de que mais de 100 pessoas compareçam é de cerca de 0,659, que provavelmente é um risco de *overbooking* maior do que a empresa deseja tolerar. Se, em vez disso, o número de reservas for 110, a probabilidade de mais de 100 passageiros comparecer é de apenas 0,024.

B.2b Distribuições condicionais

Em econometria, normalmente estamos interessados na forma como uma variável aleatória, chamemos de Y , está relacionada com uma ou mais outras variáveis. Por agora, suponha que haja apenas uma variável em cujos efeitos estamos interessados, chamada X . O máximo que conseguimos saber sobre a forma como X afeta Y está contido na **distribuição condicional** de Y , dado X . Essa informação é resumida pela *função de densidade de probabilidade condicional*, definida por

$$f_{Y|X}(y|x) = f_{X,Y}(x,y)/f_X(x) \quad (\text{B.15})$$

para todos os valores de x , de modo que $f_X(x) > 0$. A interpretação de (B.15) é vista mais facilmente quando X e Y são discretas. Assim,

$$f_{Y|X}(y|x) = P(Y = y|X = x), \quad (\text{B.16})$$

onde o lado direito é lido como “a probabilidade de $Y = y$, visto que $X = x$ ”. Quando Y é contínua, $f_{Y|X}(y|x)$ não pode ser interpretado diretamente como uma probabilidade, pelas razões discutidas anteriormente, mas as probabilidades condicionais são encontradas calculando as áreas abaixo da fdp condicional.

Uma característica importante das distribuições condicionais é que, se X e Y forem variáveis aleatórias independentes, saber o valor assumido por X não nos diz nada a respeito da probabilidade que Y tem de assumir vários valores (e vice-versa). Ou seja, $f_{Y|X}(y|x) = f_Y(y)$ e $f_{X|Y}(x|y) = f_X(x)$.

EXEMPLO B.2 Arremesso de lances livres

Considere novamente o exemplo dos arremessos de basquete, em que são tentados dois lances livres. Suponha que a densidade condicional seja

$$f_{Y|X}(1|1) = 0,85, f_{Y|X}(0|1) = 0,15$$

$$f_{Y|X}(1|0) = 0,70, f_{Y|X}(0|0) = 0,30.$$

Isso significa que a probabilidade de o jogador converter o segundo lance livre depende de como o primeiro lance foi feito: se o primeiro lance livre for convertido, a chance de fazer o segundo é 0,85; se o primeiro lance livre for perdido, a chance de converter o segundo é 0,70. Isso indica que X e Y não são independentes; mas sim dependentes.

Ainda podemos calcular $P(X = 1, Y = 1)$ já que conhecemos $P(X = 1)$. Suponha que a probabilidade de converter o primeiro lance livre seja 0,8, ou seja, $P(X = 1) = 0,8$. Assim, usando (B.15), temos

$$P(X = 1, Y = 1) = P(Y = 1|X = 1) \cdot P(X = 1) = (0,85)(0,8) = 0,68$$

B.3 Características das distribuições de probabilidade

Para muitos propósitos, estaremos interessados em apenas alguns aspectos das distribuições de variáveis aleatórias. As características de interesse podem ser colocadas em três categorias: medidas de tendência central, medidas de variabilidade de dispersão e medidas de associação entre duas variáveis aleatórias. Trataremos da última delas na Seção B.4.

B.3a Uma medida de tendência central: o valor esperado

O valor esperado é um dos conceitos probabilísticos mais importantes que encontraremos em nosso estudo de econometria. Se X é uma variável aleatória, o **valor esperado** (ou expectativa) de X , designado por $E(X)$ e às vezes μ_X ou simplesmente μ , é uma média ponderada de todos os valores possíveis de X . Os pesos são determinados pela função de densidade de probabilidade. Às vezes, o valor esperado é chamado de *média populacional*, especialmente quando queremos enfatizar que X representa alguma variável em uma população.

A definição precisa de um valor esperado é a mais simples no caso de X ser uma variável aleatória discreta que assume um número finito de valores, digamos, $\{x_1, \dots, x_k\}$. Seja $f(x)$ a indicação da função de densidade de probabilidade de X . O valor esperado de X é a média ponderada

$$E(X) = x_1 f(x_1) + x_2 f(x_2) + \dots + x_k f(x_k) \equiv \sum_{j=1}^k x_j f(x_j). \quad (\text{B.17})$$

Isso é facilmente calculado dados os valores da fdp em cada possível resultado de X .

EXEMPLO B.3

Calculando um valor esperado

Suponha que X assumira os valores -1 , 0 e 2 com probabilidades $1/8$, $1/2$ e $3/8$, respectivamente. Assim,

$$E(X) = (-1) \cdot (1/8) + 0 \cdot (1/2) + 2 \cdot (3/8) = 5/8.$$

Este exemplo ilustra algo curioso sobre valores esperados: o valor esperado de X pode ser um número que não é sequer um possível resultado de X . Sabemos que X assume os valores -1 , 0 ou 2 , apesar disso, seu valor esperado é $5/8$. Isso torna o valor esperado deficiente para resumir a tendência central de certas variáveis aleatórias discretas, mas cálculos como aqueles mencionados anteriormente podem ser úteis, como veremos mais adiante.

Se X é uma variável aleatória contínua, então $E(X)$ é definida como uma integral:

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx, \quad (\text{B.18})$$

que supomos ser bem definida. Isso ainda pode ser interpretado como uma média ponderada. Para as distribuições contínuas mais comuns, $E(X)$ é um número que é um possível resultado de X . Neste texto, não precisaremos calcular os valores esperados usando integração, embora recorramos a alguns resultados conhecidos da probabilidade para valores esperados de variáveis aleatórias especiais.

Dada uma variável aleatória X e uma função $g(\cdot)$, podemos criar uma nova variável aleatória $g(X)$. Por exemplo, se X é uma variável aleatória, então X^2 e $\log(X)$ (se $X > 0$) também são. O valor esperado de $g(X)$ é, de novo, simplesmente uma média ponderada:

$$E[g(X)] = \sum_{j=1}^k g(x_j) f_X(x_j) \quad (\text{B.19})$$

ou, para uma variável aleatória contínua,

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)dx. \quad (\text{B.20})$$

EXEMPLO B.4 Valor esperado de X^2

Para a variável aleatória do Exemplo B.3, defina $g(X) = X^2$. Então,

$$E(X^2) = (-1)^2(1/8) + (0)^2(1/2) + (2)^2(3/8) = 13/8.$$

No Exemplo B.3, calculamos $E(X) = 5/8$, de forma que $[E(X)]^2 = 25/64$. Isso mostra que $E(X)^2$ não é o mesmo que $[E(X)]^2$. Na verdade, para uma função não linear $g(X)$, $E[g(X)] \neq g[E(X)]$ (exceto em casos muito especiais).

Se X e Y são variáveis aleatórias, então $g(X, Y)$ é uma variável aleatória para qualquer função g , e, portanto, podemos definir sua expectativa. Quando X e Y são discretas, assumindo valores $\{x_1, x_2, \dots, x_k\}$ e $\{y_1, y_2, \dots, y_m\}$, respectivamente, o valor esperado é

$$E[g(X, Y)] = \sum_{h=1}^k \sum_{j=1}^m g(x_h, y_j)f_{X,Y}(x_h, y_j),$$

onde $f_{X,Y}$ é a fdp conjunta de (X, Y) . A definição é mais complicada para variáveis aleatórias contínuas, já que envolve integração; não precisamos dela aqui. A extensão para mais de duas variáveis é direta.

B.3b Propriedades de valores esperados

Em econometria, não estamos tão preocupados com o cálculo de valores esperados de várias distribuições; os principais cálculos já foram feitos muitas vezes, e vamos confiar neles cegamente. Precisaremos manipular alguns valores esperados usando algumas regras simples. Elas são tão importantes que vamos dar-lhes rótulos:

Propriedade E.1: Para qualquer constante c , $E(c) = c$.

Propriedade E.2: Para quaisquer constantes a e b , $E(aX + b) = aE(X) + b$.

Uma implicação útil de E.2 é que, se $\mu = E(X)$, e definirmos uma nova variável aleatória como $Y = X - \mu$, então $E(Y) = 0$; em E.2, use $a = 1$ e $b = -\mu$.

Como um exemplo da Propriedade E.2, defina X como a temperatura medida em Celsius ao meio-dia em determinado dia, em um dado local; suponha que a temperatura esperada seja $E(X) = 25$. Se Y é a temperatura medida em Fahrenheit, então $Y = 32 + (9/5)X$. Segundo a Propriedade E.2, a temperatura esperada em Fahrenheit é $E(Y) = 32 + (9/5) \cdot E(X) = 32 + (9/5) \cdot 25 = 77$.

Geralmente, é fácil calcular o valor esperado de uma função linear com muitas variáveis aleatórias.

Propriedade E.3: Se $\{a_1, a_2, \dots, a_n\}$ são constantes e $\{X_1, X_2, \dots, X_n\}$ são variáveis aleatórias, então

$$E(a_1X_1 + a_2X_2 + \dots + a_nX_n) = a_1E(X_1) + a_2E(X_2) + \dots + a_nE(X_n).$$

Ou, usando a notação de soma,

$$E\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i E(X_i). \quad (\text{B.21})$$

Como um caso especial dela, temos (com cada $a_i = 1$)

$$E\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n E(X_i), \quad (\text{B.22})$$

de forma que o valor esperado da soma é a soma dos valores esperados. Essa propriedade é usada com frequência para derivações em estatísticas matemáticas.

EXEMPLO B.5

Descobrimos a receita esperada

Defina X_1 , X_2 e X_3 o número de pizzas pequenas, médias e grandes, respectivamente, vendidas durante o dia em uma pizzaria. Elas são variáveis aleatórias com valores esperados $E(X_1) = 25$, $E(X_2) = 57$ e $E(X_3) = 40$. Os preços das pizzas pequenas, médias e grandes são US\$ 5,50, US\$ 7,60 e US\$ 9,15. Assim, a receita esperada pela venda de pizzas em certo dia é

$$\begin{aligned} E(5,50 X_1 + 7,60 X_2 + 9,15 X_3) &= 5,50 E(X_1) + 7,60 E(X_2) + 9,15 E(X_3) \\ &= 5,50(25) + 7,60(57) + 9,15(40) = 936,70, \end{aligned}$$

isto é, US\$ 936,70. A receita real de determinado dia geralmente vai diferir deste valor, mas esta é a receita *esperada*.

Também podemos usar a Propriedade E.3 para mostrar que, se $X \sim \text{Binomial}(n, \theta)$, então $E(X) = n\theta$. Ou seja, o número esperado de sucessos em n testes de Bernoulli é simplesmente o número de testes vezes a probabilidade de sucesso em qualquer teste em particular. Isso é visto com facilidade ao escrever X como $X = Y_1 + Y_2 + \dots + Y_n$, onde cada $Y_i \sim \text{Bernoulli}(\theta)$. Assim,

$$E(X) = \sum_{i=1}^n E(Y_i) = \sum_{i=1}^n \theta = n\theta.$$

Podemos aplicar isso ao exemplo das reservas aéreas, onde a companhia aérea faz $n = 120$ reservas e a probabilidade de comparecimento é $\theta = 0,85$. O número *esperado* de pessoas que vão comparecer é $120(0,85) = 102$. Portanto, se existem 100 poltronas disponíveis, o número esperado de pessoas que vão comparecer é muito grande; isso influencia a decisão de ser uma boa ideia para a companhia aérea fazer 120 reservas.

Na verdade, o que a companhia deve fazer é definir uma função de lucro que leve em conta a receita líquida obtida por assento vendido e o custo por passageiro excluído do voo. Essa função de lucro é aleatória porque o número real de pessoas que comparecerão é aleatório. Seja r a receita líquida para cada passageiro. (Você pode pensar nisso como o preço da passagem para simplificar.) Defina c como a compensação devida a qualquer passageiro excluído do voo. Nem r nem c são aleatórias;

supõe-se que elas sejam conhecidas pela empresa. Faça com que Y indique os lucros para voo. Então, com 100 poltronas disponíveis,

$$\begin{aligned} Y &= rX \text{ se } X \leq 100 \\ &= 100r - c(X - 100) \text{ se } X > 100. \end{aligned}$$

A primeira equação dá lucro se nenhuma pessoa a mais do que 100 comparecer para o voo; a segunda equação é lucrativa se mais de 100 pessoas comparecerem. (No segundo caso, a receita líquida pela venda de passagens é $100r$, já que todos os 100 assentos foram vendidos, assim, $c(X - 100)$ é o custo de fazer mais de 100 reservas.) Usando o fato de que X tem uma distribuição Binomial($n, 0,85$), em que n é o número de reservas realizadas, os lucros esperados, $E(Y)$, podem ser encontrados como uma função de n (e de r , e de c). Calcular $E(Y)$ de forma direta seria bem difícil, mas isso pode ser encontrado rapidamente usando um computador. Uma vez que os valores de r e c sejam dados, o valor de n que maximiza os lucros esperados pode ser encontrado ao examinar diferentes valores de n .

B.3c Outra medida de tendência central: a mediana

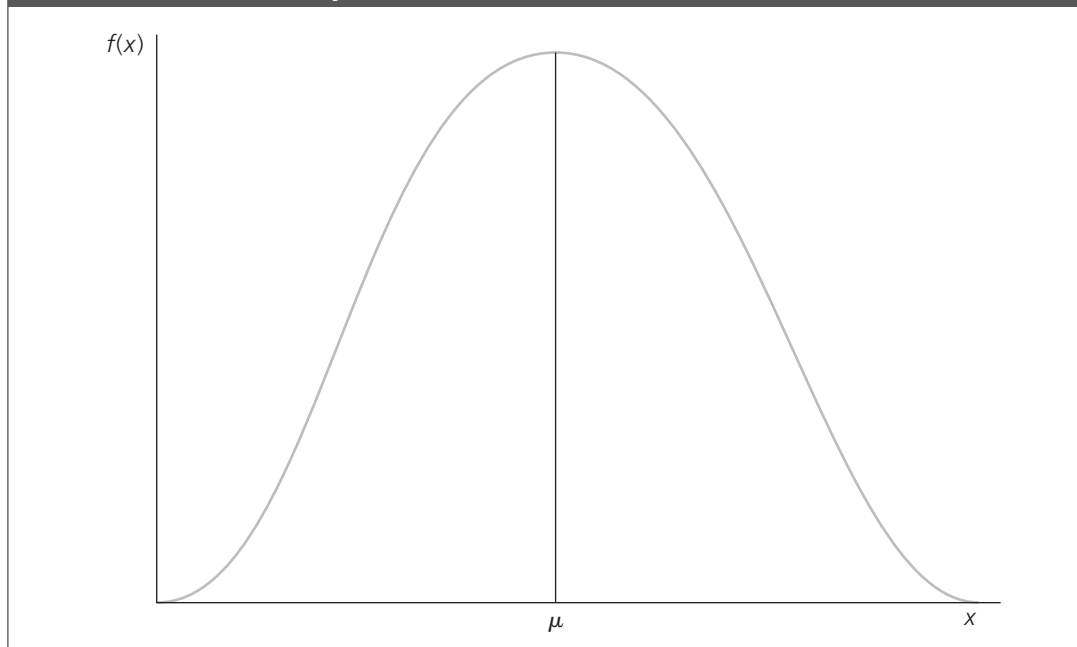
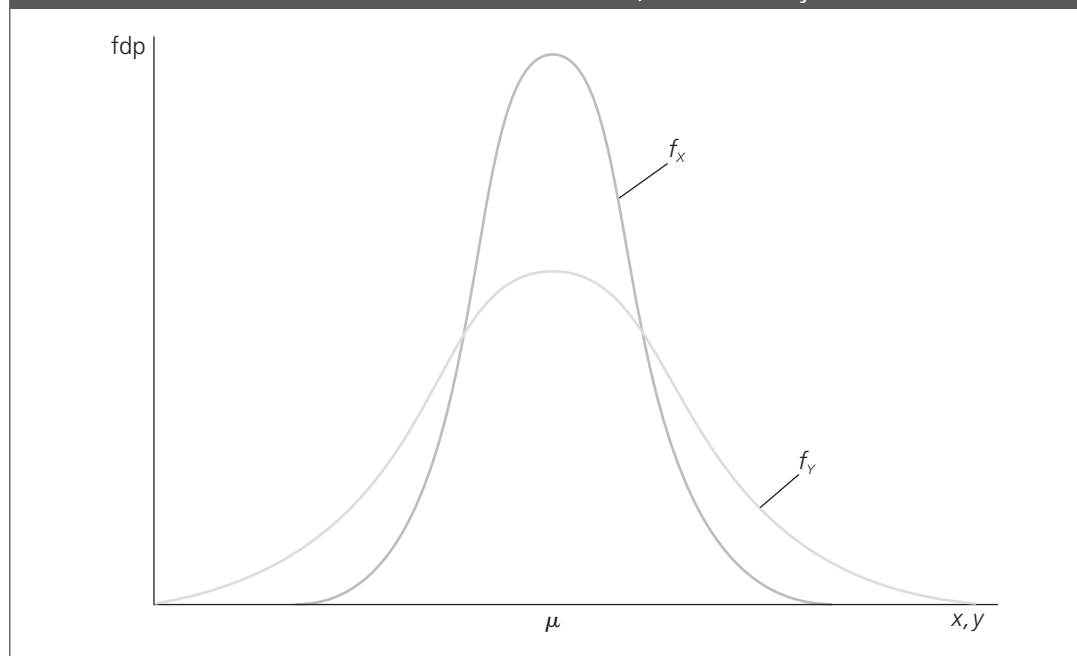
O valor esperado é apenas uma possibilidade para definir a tendência central de uma variável aleatória. Outra medida de tendência central é a **mediana**. Uma definição geral de *mediana* é muito complicada para nossos propósitos. Se X é contínuo, então a mediana de X , digamos, m , é o valor que faz com que uma metade da área abaixo da fdp esteja à esquerda de m e a outra à direita de m .

Quando X é discreta e assume um número finito ímpar de valores, a mediana é obtida ordenando os possíveis valores de X e selecionando o valor do meio. Por exemplo, se X pode assumir os valores $\{-4, 0, 2, 8, 10, 13, 17\}$, então o valor mediano de X é 8. Se X assumir um número par de valores, existem, na verdade, dois valores medianos; às vezes tira-se a média desses valores para obter um único valor mediano. Assim, se X tiver os valores $\{-5, 3, 9, 17\}$, os valores medianos são 3 e 9; se calcularmos a média, temos uma mediana igual a 6.

Em geral, a mediana, às vezes indicada como $\text{Med}(X)$, e o valor esperado, $E(X)$, são diferentes. Nenhum deles é “melhor” do que o outro como medida da tendência central; ambos são formas válidas para medir o centro de distribuição de X . Em um caso especial, a mediana e o valor esperado (ou média) têm o mesmo valor. Se X tiver uma **distribuição simétrica** próxima do valor de μ , então μ será tanto o valor esperado quanto a mediana. Matematicamente, a condição é $f(\mu + x) = f(\mu - x)$ para todo x . Este caso é ilustrado na Figura B.3.

B.3d Medidas de variabilidade: variância e desvio padrão

Embora a tendência central de uma variável aleatória seja valiosa, ela não nos diz tudo que queremos saber a respeito da distribuição de uma variável aleatória. A Figura B.4 mostra as fdps de duas variáveis aleatórias com a mesma média. Claramente, a distribuição de X é mais centrada sobre sua média do que a distribuição de Y . Gostaríamos de ter uma maneira simples de resumir as diferenças nas dispersões das distribuições.

FIGURA B.3 Uma distribuição de probabilidade simétrica.**FIGURA B.4** Variáveis aleatórias com a mesma média, mas distribuições diferentes.

B.3e Variância

Para uma variável aleatória X , defina $\mu = E(X)$. Existem várias formas de medir quão distante está X de seu valor esperado, mas a mais simples para trabalhar algebricamente é diferença ao quadrado, $(X - \mu)^2$. (Elevar ao quadrado elimina o sinal da

medida de distância; o valor positivo resultante corresponde à nossa noção intuitiva da distância e trata valores acima e abaixo de μ simetricamente.) Essa distância é uma variável aleatória, já que pode mudar com todo resultado de X . Assim como precisamos de um número para resumir a tendência central de X , precisamos de um número que nos diga quão distante X está de μ , *em média*. Um destes números é a **variância**, que nos diz a distância esperada de X até sua média:

$$\text{Var}(X) \equiv E[(X - \mu)^2]. \quad (\text{B.23})$$

Às vezes a variância é indicada como σ_X^2 , ou simplesmente σ^2 , quando o contexto for claro. A partir de (B.23), concluímos que a variância é sempre não negativa.

Como dispositivo computacional, é útil observar que

$$\sigma^2 = E(X^2 - 2X\mu + \mu^2) = E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2. \quad (\text{B.24})$$

Ao usar (B.23) ou (B.24), não precisamos distinguir variáveis aleatórias discretas e contínuas: a definição de variância é a mesma em ambos os casos. Mais frequentemente, calculamos primeiro $E(X)$, depois $E(X^2)$, e então usamos a fórmula da equação (B.24). Por exemplo, se $X \sim \text{Bernoulli}(\theta)$, então $E(X) = \theta$ e, já que $X^2 = X$, $E(X^2) = \theta$. Como resultado da equação (B.24), $\text{Var}(X) = E(X^2) - \mu^2 = \theta - \theta^2 = \theta(1 - \theta)$.

Duas propriedades importantes da variância são mostradas abaixo.

Propriedade VAR.1: $\text{Var}(X) = 0$ se, e somente se, houver uma constante c de forma que $P(X = c) = 1$, no caso em que $E(X) = c$.

Essa primeira propriedade diz que a variância de qualquer constante é zero e, se uma variável aleatória tem variância zero, ela é necessariamente constante.

Propriedade VAR.2: Para quaisquer constantes a e b , $\text{Var}(aX + b) = a^2\text{Var}(X)$.

Isso significa que somar uma constante a uma variável aleatória não muda a variância, mas multiplicar uma variável aleatória por uma constante aumenta a variância em um fator igual ao *quadrado* daquela constante. Por exemplo, se X simbolizar a temperatura em Celsius e $Y = 32 + (9/5)X$ é a temperatura em Fahrenheit, $\text{Var}(Y) = (9/5)^2\text{Var}(X) = (81/25)\text{Var}(X)$.

B.3f Desvio padrão

O **desvio padrão** de uma variável aleatória, chamado de $dp(X)$, é simplesmente a raiz quadrada positiva da variância: $dp(X) \equiv +\sqrt{\text{Var}(X)}$. O desvio padrão às vezes é indicado como σ_X , ou simplesmente σ , quando a variável aleatória é conhecida. Duas propriedades do desvio padrão derivam imediatamente das Propriedades VAR.1 e VAR.2.

Propriedade DP.1: Para qualquer constante c , $dp(c) = 0$.

Propriedade DP.2: Para quaisquer constantes a e b ,

$$dp(aX + b) = |a|dp(X).$$

Em especial se $a > 0$, assim, $dp(aX) = a \cdot dp(X)$.

Essa última propriedade torna o desvio padrão mais natural de trabalhar do que a variância. Por exemplo, suponha que X seja uma variável aleatória medida em

milhares de dólares, por exemplo, renda. Se definirmos $Y = 1.000X$, então Y é a renda medida em dólares. Suponha que $E(X) = 20$ e $dp(X) = 6$. Então, $E(Y) = 1.000E(X) = 20.000$ e $dp(Y) = 1.000 \cdot dp(X) = 6.000$, de forma que o valor esperado e o desvio padrão aumentam pelo mesmo fator, 1.000. Se tivéssemos trabalhado com a variância, teríamos $\text{Var}(Y) = (1.000)^2 \text{Var}(X)$, de modo que a variância de Y é um milhão de vezes maior do que a variância de X .

B.3g Padronizando uma variável aleatória

Como uma aplicação das propriedades da variância e do desvio padrão – e um tópico de interesse prático por si só –, suponha que, dada uma variável aleatória X , definimos uma nova variável aleatória subtraindo sua média e dividindo por seu desvio padrão σ :

$$Z \equiv \frac{X - \mu}{\sigma}, \quad (\text{B.25})$$

que podemos escrever como $Z = aX + b$, onde $a \equiv (1/\sigma)$ e $b \equiv -(\mu/\sigma)$. Assim, com base na Propriedade E.2,

$$E(Z) = aE(X) + b = (\mu/\sigma) - (\mu/\sigma) = 0.$$

Pela propriedade VAR.2,

$$\text{Var}(Z) = a^2 \text{Var}(X) = (\sigma^2/\sigma^2) = 1.$$

Dessa forma, a variável aleatória Z tem uma média zero e uma variância (e, portanto, um desvio padrão) igual a um. Este procedimento é conhecido como *padronização* da variável aleatória X , e Z é chamada de **variável aleatória padronizada**. (Nos cursos de estatística introdutória, isso às vezes é chamado de *transformação z de X* .) É importante lembrar que o desvio padrão, e não a variância, aparece no denominador de (B.25). Como veremos mais adiante, essa transformação é usada com frequência na inferência estatística.

Como um exemplo específico, suponha $E(X) = 2$ e $\text{Var}(X) = 9$. Então, $Z = (X - 2)/3$ tem um valor esperado zero e variância um.

B.3h Assimetria e curtose

Podemos usar a versão padronizada de uma variável aleatória para definir outras características da distribuição de uma variável aleatória. Essas características são descritas pelo uso do que são chamados de *momentos de ordem mais elevada*. Por exemplo, o terceiro momento da variável aleatória Z em (B.25) é usado para determinar se uma distribuição é simétrica perto de sua média. Podemos escrever

$$E(Z^3) = E[(X - \mu)^3]/\sigma^3.$$

Se X tem uma distribuição simétrica em torno de μ , então Z tem uma distribuição simétrica em torno de zero. (A divisão por σ^3 não muda se a distribuição for simétrica.) Isso quer dizer que a densidade de Z em quaisquer dois pontos z e $-z$ é a mesma, o que significa que, ao calcular $E(Z^3)$, valores positivos z^3 quando $z > 0$ são exatamente anulados com o valor negativo $(-z^3) = -z^3$. O resultado é que, se X é simétrico próximo a zero,

então $E(Z) = 0$. Geralmente, $E[(X - \mu)^3]/\sigma^3$ é vista como uma medida da **assimetria** da distribuição de X . Em um ambiente estatístico, podemos usar dados para estimar $E(Z^3)$ para determinar se uma distribuição populacional subjacente parece ser simétrica. (O Exercício em computador C5.4, do Capítulo 5, apresenta um exemplo.)

Também pode ser informativo calcular o quarto momento de Z .

$$E(Z^4) = E[(X - \mu)^4]/\sigma^4.$$

Como $Z^4 \geq 0$, $E(Z^4) \geq 0$ (e, em qualquer caso interessante, estritamente maior do que zero). Sem ter um valor de referência, é difícil interpretar valores de $E(Z^4)$, mas valores maiores significam que as extremidades da distribuição de X são mais densas. O quarto momento $E(Z^4)$ é chamado de medida de **curtose** da distribuição de X . Na Seção B.5, obteremos $E(Z^4)$ da distribuição normal.

B.4 Características das distribuições conjunta e condicional

B.4a Medidas de associação: covariância e correlação

Embora a fdp conjunta de duas variáveis aleatórias descreva completamente a relação entre elas, é útil ter medidas resumidas sobre como, em média, duas variáveis aleatórias variam umas em relação às outras. Assim como o valor esperado e a variância, isso é parecido com usar um único número para resumir algo a respeito de uma distribuição inteira, que, neste caso, é uma distribuição conjunta de duas variáveis aleatórias.

B.4b Covariância

Defina $\mu_X = E(X)$ e $\mu_Y = E(Y)$ e considere a variável aleatória $(X - \mu_X)(Y - \mu_Y)$. Agora, se X está acima de sua média e Y está acima de sua média, então $(X - \mu_X)(Y - \mu_Y) > 0$. Isso também é verdadeiro se $X < \mu_X$ e $Y < \mu_Y$. Por outro lado, se $X > \mu_X$ e $Y < \mu_Y$, ou vice-versa, $(X - \mu_X)(Y - \mu_Y) < 0$. Como, então, esse produto pode nos dizer algo sobre a relação entre X e Y ?

A **covariância** entre duas variáveis aleatórias X e Y , às vezes chamada de *covariância populacional* para enfatizar que ela envolve a relação entre duas variáveis que descrevem uma população, é definida como o valor esperado do produto $(X - \mu_X)(Y - \mu_Y)$:

$$\text{Cov}(X, Y) \equiv E[(X - \mu_X)(Y - \mu_Y)], \quad (\text{B.26})$$

que às vezes é simbolizada como σ_{XY} . Se $\sigma_{XY} > 0$, então, em média, quando X estiver acima de sua média, Y também estará. Se $\sigma_{XY} < 0$, então, em média, quando X estiver acima de sua média, Y estará abaixo de sua média.

Várias expressões úteis para calcular $\text{Cov}(X, Y)$ são as seguintes:

$$\begin{aligned} \text{Cov}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] = E[(X - \mu_X)Y] \\ &= E[X(Y - \mu_Y)] = E(XY) - \mu_X\mu_Y. \end{aligned} \quad (\text{B.27})$$

Como resultado de (B.27), se $E(X) = 0$ ou $E(Y) = 0$, então $\text{Cov}(X, Y) = E(XY)$.

A covariância mede a quantidade de dependência *linear* entre duas variáveis aleatórias. Uma covariância positiva indica que duas variáveis aleatórias se movem na mesma direção, enquanto uma covariância negativa indica que elas se movem em

direções opostas. Interpretar a *magnitude* de uma covariância pode ser um pouco complicado, como veremos em breve.

Já que a covariância é uma medida de como duas variáveis estão relacionadas, é natural questionar como a covariância está relacionada com a noção de independência. Isso é dado pela propriedade abaixo.

Propriedade Cov.1: Se X e Y são independentes, então $\text{Cov}(X, Y) = 0$.

Essa propriedade surge a partir da equação (B.27) e do fato de que $E(XY) = E(X)E(Y)$ quando X e Y são independentes. É importante se lembrar que o inverso de Cov.1 *não* é verdadeiro: covariância zero entre X e Y não indica que X e Y são independentes. Na verdade, existem variáveis aleatórias X que, se $Y = X^2$, $\text{Cov}(X) = 0$. [Qualquer variável aleatória com $E(X) = 0$ e $E(X^3) = 0$ tem essa propriedade.] Se $Y = X^2$, então X e Y claramente não são independentes: uma vez que conheçamos X , conheceremos Y . Parece bem estranho que X e X^2 possam ter zero covariância, e isso revela uma fragilidade da covariância como medida geral de associação entre variáveis aleatórias. A covariância é útil em contextos nos quais as relações são pelo menos aproximadamente lineares.

A segunda principal propriedade da covariância envolve covariâncias entre funções lineares.

Propriedade Cov.2: Para quaisquer constantes a_1 , b_1 , a_2 e b_2 ,

$$\text{Cov}(a_1X + b_1, a_2Y + b_2) = a_1a_2\text{Cov}(X, Y). \quad (\text{B.28})$$

Uma implicação importante da Cov.2 é que a covariância entre duas variáveis aleatórias pode ser alterada pela multiplicação de uma ou ambas as variáveis aleatórias por uma constante. Isso é importante na economia porque variáveis monetárias, taxas de inflação, e assim por diante, podem ser definidas com diferentes unidades de medidas sem mudar seu significado.

Por fim, é útil saber que o valor absoluto da covariância entre quaisquer duas variáveis é limitado pelo produto de seus desvios padrão; isso é conhecido como *desigualdade de Cauchy-Schwartz*.

Propriedade Cov.3: $|\text{Cov}(X, Y)| \leq \text{dp}(X)\text{dp}(Y)$.

B.4c Coeficiente de correlação

Suponha que desejemos conhecer a relação entre uma medida de educação e os rendimentos anuais da população trabalhadora. Podemos fazer com que X indique educação e Y , os rendimentos, e assim calcular sua covariância. No entanto, a resposta que obteremos vai depender de como escolhemos medir a educação e os rendimentos.

A Propriedade Cov.2 indica que a covariância entre educação e rendimentos depende do fato de a renda ser medida em dólares ou milhares de dólares, ou se a educação é medida em meses ou anos. É bem claro que a forma como medimos essas variáveis não tem influência sobre quão fortemente elas são relacionadas. Mas a covariância entre elas depende das unidades de medida.

O fato de a covariância depender de unidades de medida é uma deficiência que é superada pelo **coeficiente de correlação** entre X e Y :

$$\text{Corr}(X, Y) \equiv \frac{\text{Cov}(X, Y)}{\text{dp}(X) \cdot \text{dp}(Y)} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}; \quad (\text{B.29})$$

o coeficiente de correlação entre X e Y às vezes é indicado por ρ_{XY} (e pode ser chamado de *correlação populacional*).

Como σ_X e σ_Y são positivos, $\text{Cov}(X, Y)$ e $\text{Corr}(X, Y)$ sempre terão o mesmo sinal, e $\text{Corr}(X, Y) = 0$ se, e somente se, $\text{Cov}(X, Y) = 0$. Algumas das propriedades da covariância são transferidas para a correlação. Se X e Y são independentes, então $\text{Corr}(X, Y) = 0$, mas zero correlação não indica independência. (Assim como a covariância, o coeficiente de correlação também é uma medida de dependência linear.) No entanto, a magnitude do coeficiente de correlação é mais fácil de interpretar do que o tamanho da covariância devido à seguinte propriedade.

Propriedade Corr.1: $-1 \leq \text{Corr}(X, Y) \leq 1$.

Se $\text{Corr}(X, Y) = 0$, ou, de forma equivalente, $\text{Cov}(X, Y) = 0$, não há relação linear entre X e Y e elas são chamadas de **variáveis aleatórias não correlacionadas**; caso contrário, X e Y são *correlacionadas*. $\text{Corr}(X, Y) = 1$ indica uma relação linear positiva perfeita, o que significa que podemos escrever $Y = a + bX$ para uma constante a e uma constante $b > 0$. $\text{Corr}(X, Y) = -1$ indica uma relação linear negativa perfeita, de modo que $Y = a + bX$ para $b < 0$. Os casos extremos de 1 positivo ou negativo ocorrem raramente. Valores de ρ_{XY} próximos a 1 ou -1 indicam relações lineares mais fortes.

Como mencionado anteriormente, a correlação entre X e Y é invariável para as unidades de medida de X ou Y . Isso é declarado de forma mais geral conforme segue.

Propriedade Corr.2: Para constantes a_1, b_1, a_2 e b_2 , com $a_1 a_2 > 0$,

$$\text{Corr}(a_1 X + b_1, a_2 Y + b_2) = \text{Corr}(X, Y).$$

Se $a_1 a_2 < 0$, então

$$\text{Corr}(a_1 X + b_1, a_2 Y + b_2) = -\text{Corr}(X, Y).$$

Como um exemplo, suponha que a correlação entre rendimentos e educação na população trabalhadora seja 0,15. Essa medida não depende de os rendimentos serem medidos em dólares, milhares de dólares ou qualquer outra unidade; ela também não depende de a educação ser medida em anos, trimestres, meses, e assim por diante.

B.4d Variância de somas de variáveis aleatórias

Agora que definimos covariância e correlação, podemos completar nossa lista de grandes propriedades da variância.

Propriedade VAR.3: Para constantes a e b ,

$$\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{Cov}(X, Y).$$

Como consequência imediata, se X e Y forem não correlacionadas – de modo que $\text{Cov}(X, Y) = 0$ – então

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \quad (\text{B.30})$$

e

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y). \quad (\text{B.31})$$

No segundo caso, note que a variância da diferença é a *soma das variâncias*, e não a diferença delas.

Como exemplo de (B.30), defina X para indicar lucros obtidos por um restaurante durante uma sexta-feira à noite e Y para indicar os lucros obtidos na noite seguinte, de sábado. Então, $Z = X + Y$ é o lucro das duas noites. Suponha que X e Y tenham um valor esperado de US\$ 300 cada e um desvio padrão de US\$ 15 (de modo que a variância é 225). O lucro esperado para as duas noites é $E(Z) = E(X) + E(Y) = 2 \cdot (300) = 600$ dólares. Se X e Y são independentes e, portanto, não correlacionados, a variância do lucro total é a soma das variâncias: $\text{Var}(Z) = \text{Var}(X) + \text{Var}(Y) = 2 \cdot (225) = 450$. Como consequência, o desvio padrão do lucro total é $\sqrt{450}$ ou cerca de US\$ 21,21.

As expressões (B.30) e (B.31) se estendem para mais de duas variáveis aleatórias. Para expressar essa extensão, precisamos de uma definição. As variáveis aleatórias $\{X_1, \dots, X_n\}$ são **variáveis aleatórias não correlacionadas em pares** se cada variável do conjunto não for correlacionada com nenhuma outra variável do conjunto. Isto é, $\text{Cov}(X_i, X_j) = 0$ para todo $i \neq j$.

Propriedade Var.4: Se $\{X_1, \dots, X_n\}$ são variáveis aleatórias não correlacionadas em pares e a_i , $i = 1, \dots, n$ são constantes, então

$$\text{Var}(a_1X_1 + \dots + a_nX_n) = a_1^2\text{Var}(X_1) + \dots + a_n^2\text{Var}(X_n).$$

Em notação de soma, podemos escrever

$$\text{Var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \text{Var}(X_i). \quad (\text{B.32})$$

Um caso especial da Propriedade VAR.4 ocorre quando obtemos $a_i = 1$ para todos os i . Assim, para variáveis aleatórias não correlacionadas em pares, a variância da soma é a soma das variâncias:

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i). \quad (\text{B.33})$$

Como variáveis aleatórias independentes são não correlacionadas (ver Propriedade COV.1), a variância da soma de variáveis aleatórias independentes é a soma das variâncias.

Se X_i não for não correlacionada em par, então a expressão para $\text{Var}(\sum_{i=1}^n a_i X_i)$ será muito mais complicada; precisamos adicionar ao lado direito de (B.32) os termos $2a_i a_j \text{Cov}(x_i, x_j)$ para todo $i > j$.

Podemos usar (B.33) para derivar a variância para uma variável aleatória binomial. Defina $X \sim \text{Binomial}(n, \theta)$ e escreva $X = Y_1 + \dots + Y_n$, onde os Y_i são variáveis aleatórias independentes de Bernoulli (θ). Assim, a partir de (B.33), $\text{Var}(X) = \text{Var}(Y_1) + \dots + \text{Var}(Y_n) = n\theta(1 - \theta)$.

No exemplo das reservas da companhia aérea com $n = 120$ e $\theta = 0,85$, a variância do número de passageiros que comparecem para a reserva é $120(0,85)(0,15) = 15,3$, assim, o desvio padrão é cerca de 3,9.

B.4.e Expectativa condicional

Covariância e correlação medem a relação linear entre duas variáveis aleatórias e as trata de forma simétrica. Com mais frequência nas ciências sociais, gostaríamos de explicar uma variável, chamada de Y , em termos de outra variável, digamos, X . Além disso, se Y está relacionada a X de forma não linear, gostaríamos de saber isso. Chame Y de variável explicada e X de variável explicativa. Por exemplo, Y pode ser o salário por hora e X pode ser os anos de educação formal.

Já apresentamos a noção de função de densidade de probabilidade condicional de Y dado X . Dessa forma, podemos querer ver como a distribuição de salários muda com o nível de educação. Entretanto, geralmente queremos uma forma simples de resumir a distribuição. Um único número não será suficiente, já que a distribuição de Y dado $X = x$ geralmente depende do valor de x . No entanto, podemos resumir a relação entre Y e X olhando para a **expectativa condicional** de Y dado X , às vezes chamada de *média condicional*. A ideia é essa. Suponha que saibamos que X assumiu determinado valor, por exemplo, x . Então, podemos calcular o valor esperado de Y , visto que conhecemos esse resultado de X . Representamos esse valor esperado por $E(Y|X = x)$, ou às vezes $E(Y|x)$ para abreviar. Geralmente, à medida que x muda, $E(Y|x)$ também muda.

Quando Y é uma variável aleatória discreta que assume os valores $\{y_1, \dots, y_m\}$, então

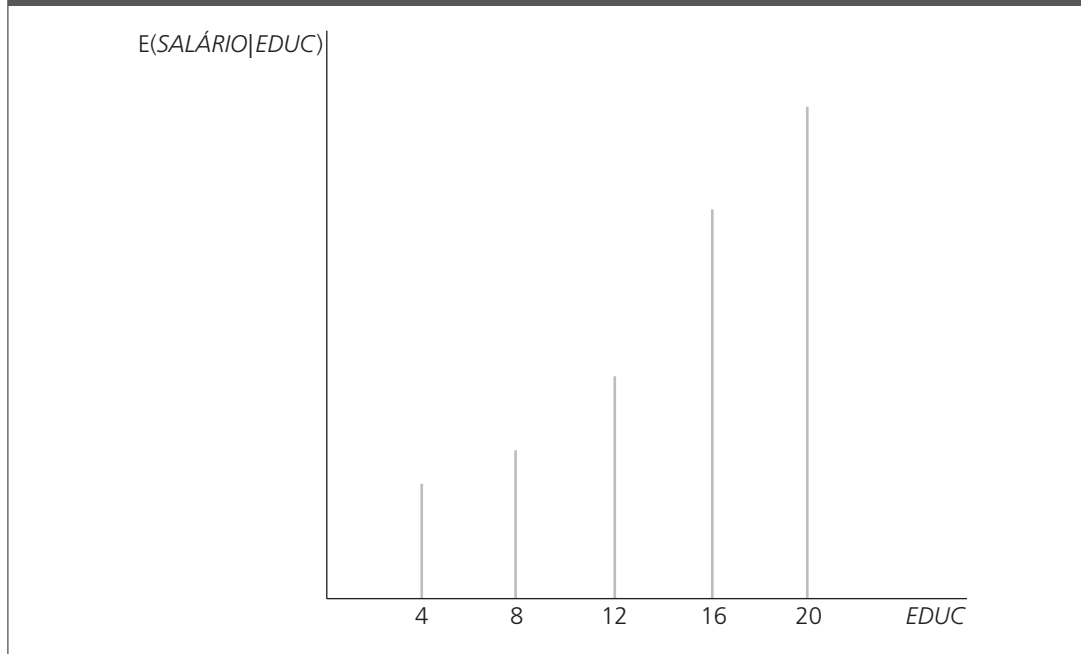
$$E(Y|x) = \sum_{j=1}^m y_j f_{Y|X}(y_j|x).$$

Quando Y é contínuo, $E(Y|x)$ é definido integrando $y f_{Y|X}(y|x)$ em todos os possíveis valores de y . Do mesmo modo que as expectativas incondicionais, a expectativa condicional é uma média ponderada dos possíveis valores de Y , mas agora os pesos refletem o fato de que X assumiu um valor específico. Portanto, $E(Y|x)$ é apenas uma função de x , que nos diz como o valor esperado de Y varia com x .

Como um exemplo, deixe (X, Y) representar a população de todos os indivíduos trabalhadores, onde X são os anos de educação e Y é o salário por hora. Assim, $E(Y|X = 12)$ é a média de salário por hora para todas as pessoas da população com 12 anos de educação (aproximadamente o ensino médio). $E(Y|X = 16)$ é o salário médio por hora das pessoas com 16 anos de educação. Traçar o valor esperado para vários níveis de educação apresenta informações importantes sobre como o salário e a escolaridade estão relacionados. Ver Figura B.5 para uma ilustração.

A princípio, o valor esperado do salário por hora pode ser encontrado para cada um dos níveis de escolaridade, e essas expectativas podem ser resumidas em uma tabela. Como a educação pode variar amplamente – e pode até ser medida em frações de um ano –, essa é uma forma complicada de mostrar a relação entre o salário médio e a medida relativa à quantidade de educação. Em econometria, geralmente especificamos funções simples que capturam essa relação. Como exemplo, suponha que o valor esperado de *SALÁRIO* dado *EDUC* seja a função linear

$$E(\text{SALÁRIO}|\text{EDUC}) = 1,05 + 0,45 \text{ EDUC}.$$

FIGURA B.5 O valor esperado do salário por hora para vários níveis de educação.

Se essa relação se mantiver na população de trabalhadores, o salário médio de uma pessoa com oito anos de educação será $1,05 + 0,45(8) = 4,65$, ou US\$ 4,65. O salário médio de uma pessoa com 16 anos de educação será 8,25, ou US\$ 8,25. O coeficiente de *EDUC* indica que cada ano de educação aumenta o salário esperado por hora em 0,45, ou 45 centavos.

Expectativas condicionais também podem ser funções não lineares. Por exemplo, suponha que $E(Y|x) = 10/x$, onde X é uma variável aleatória que é sempre maior do que zero. Essa função foi representada no gráfico da Figura B.6. Isso pode representar uma função de demanda, onde Y é a quantidade demandada e X é o preço. Se Y e X são relacionadas dessa forma, uma análise de associação linear, como a análise de correlação, seria incompleta.

B.4f Propriedades da expectativa condicional

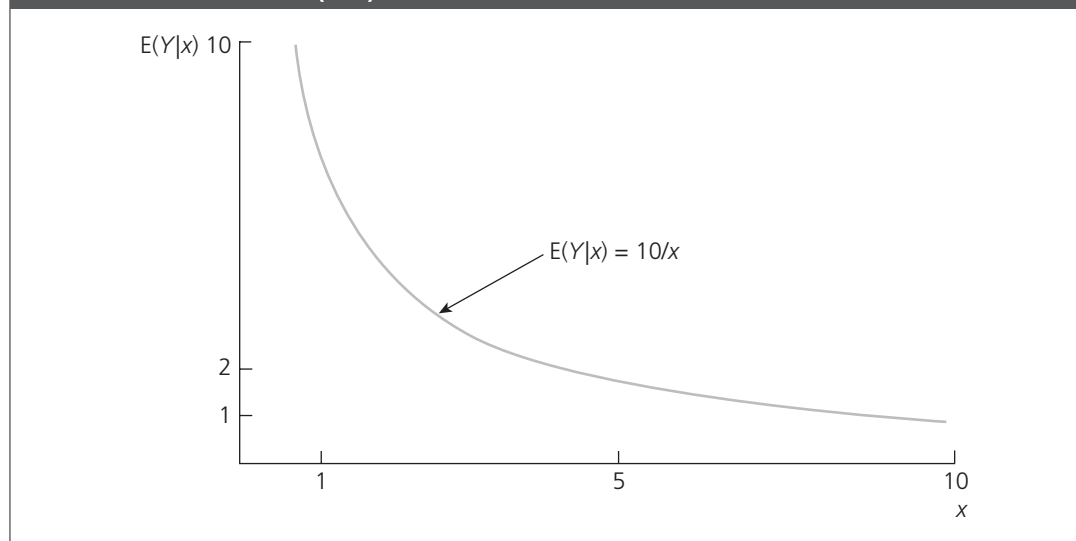
Várias propriedades básicas das expectativas condicionais são úteis para derivações na análise econométrica.

Propriedade EC.1: $E[c(X)|X] = c(X)$, para qualquer função $c(X)$.

Essa primeira propriedade quer dizer que funções de X agem como constantes quando calculamos expectativas condicionais sobre X . Por exemplo, $E(X^2|X) = X^2$. De forma intuitiva, isso quer dizer simplesmente que se conhecermos X , também conheceremos X^2 .

Propriedade EC.2: Para funções $a(X)$ e $b(X)$,

$$E[a(X)Y + b(X)|X] = a(X)E(Y|X) + b(X).$$

FIGURA B.6 Gráfico de $E(Y|X) = 10/x$.

Por exemplo, podemos calcular facilmente a expectativa condicional de uma função como $XY + 2X^2$: $E(XY + 2X^2|X) = XE(Y|X) + 2X^2$.

A propriedade abaixo une as noções de independência e expectativas condicionais.

Propriedade EC.3: Se X e Y são independentes, então $E(Y|X) = E(Y)$.

Essa propriedade significa que, se X e Y forem independentes, então o valor esperado de Y , dado X , não dependerá de X , caso em que $E(Y|X)$ sempre será igual ao valor esperado (incondicional) de Y . No exemplo de salário e escolaridade, se os salários fossem independentes da educação, o salário médio de graduados no ensino médio e no ensino superior seria o mesmo. Já que isso é quase com certeza falso, não podemos supor que salário e escolaridade sejam independentes.

Um caso especial da Propriedade EC.3 é o seguinte: se U e X são independentes e $E(U) = 0$, então $E(U|X) = 0$.

Também existem propriedades da expectativa condicional que têm a ver com o fato de que $E(Y|X)$ é uma função de X , por exemplo, $E(Y|X) = \mu(X)$. Como X é uma variável aleatória, $\mu(X)$ também é uma variável aleatória. Além disso, $\mu(X)$ tem uma distribuição de probabilidade e, portanto, um valor esperado. Geralmente, o valor esperado de $\mu(X)$ pode ser muito difícil de calcular diretamente. A **lei das expectativas iteradas** diz que o valor esperado de $\mu(X)$ é simplesmente igual ao valor esperado de Y . Escrevemos isso da seguinte maneira.

Propriedade EC.4: $E[E(Y|X)] = E(Y)$.

Essa propriedade é um pouco difícil de compreender de primeira. Ela significa que, se obtivermos primeiro $E(Y|X)$ como uma função de X e calcularmos seu valor esperado (em relação à distribuição de X , é claro), acabaremos com $E(Y)$. Isso é pouco óbvio, mas pode ser derivado usando a definição de valores esperados.

Como um exemplo de como usar a Propriedade EC.4, defina $Y = \text{SALÁRIO}$ e $X = \text{EDUC}$, onde SALÁRIO é medido em horas e EDUC em anos. Suponha que o valor

esperado de $SALÁRIO$, dado $EDUC$, seja $E(SALÁRIO|EDUC) = 4 + 0,60 EDUC$. Além disso, $E(EDUC) = 11,5$. Assim, a lei das expectativas iteradas indica que $E(SALÁRIO) = E(4 + 0,60 EDUC) = 4 + 0,60 E(EDUC) = 4 + 0,60(11,5) = 10,90$, ou US\$ 10,90 por hora.

A propriedade a seguir estabelece uma versão mais geral da lei das expectativas iteradas.

Propriedade EC.4': $E(Y|X) = E[E(Y|X,Z)|X]$.

Em outras palavras, podemos descobrir $E(Y|X)$ em duas etapas. Primeiro, encontramos $E(Y|X,Z)$ para qualquer outra variável aleatória Z . Em seguida, calculamos o valor esperado de $E(Y|X,Z)$, condicional em X .

Propriedade EC.5: Se $E(Y|X) = E(Y)$, então, $Cov(X, Y) = 0$ (bem como $Corr(X, Y) = 0$). Na verdade, *toda* função de X é não correlacionada a Y .

Essa propriedade significa que, se o conhecimento de X não muda o valor esperado de Y , então X e Y *devem ser* não correlacionadas, o que indica que, se X e Y forem correlacionadas, então $E(Y|X)$ deve depender de X . O inverso da Propriedade EC.5 não é verdadeiro: se X e Y forem não correlacionadas, $E(Y|X)$ *ainda pode* depender de X . Por exemplo, suponha que $Y = X^2$. Dessa forma, $E(Y|X) = X^2$, que é claramente uma função de X . No entanto, como mencionamos em nossa discussão sobre covariância e correlação, é possível que X e X^2 sejam não correlacionadas. A expectativa condicional captura a relação não linear entre X e Y que uma análise de correlação deixaria passar.

As propriedades EC.4 e EC.5 têm duas implicações importantes: se U e X são variáveis aleatórias de modo que $E(U|X) = 0$, então $E(U) = 0$, e U e X são não correlacionadas.

Propriedade EC.6: Se $E(Y^2) < \infty$ e $E[g(X)^2] < \infty$ para algumas funções g , então $E\{[Y - \mu(X)]^2|X\} \leq E\{[Y - g(X)]^2|X\}$ e $E\{[Y - \mu(X)]^2\} \leq E\{[Y - g(X)]^2\}$.

A propriedade EC.6 é muito útil em contextos de previsão. A primeira desigualdade diz que, se medirmos a imprecisão de uma previsão como o erro de previsão *esperado* ao quadrado, condicional em X , a média condicional será melhor do que qualquer outra função de X para prever Y . A média condicional também minimiza o quadrado do erro de previsão esperado não condicional.

B.4g Variância condicional

Dadas as variáveis aleatórias X e Y , a variância de Y , condicional em $X = x$, é simplesmente a variância associada com a distribuição condicional de Y , dado $X = x$: $E\{[Y - E(Y|x)]^2|x\}$. A fórmula

$$\text{Var}(Y|X = x) = E(Y^2|x) - [E(Y|x)]^2$$

geralmente é usada para cálculos. Ocasionalmente, teremos de calcular uma variância condicional. Porém, vamos ter de criar hipóteses sobre isso e manipular as variâncias condicionais para determinados tópicos da análise de regressão.

Como exemplo, defina $Y = \text{POUPANÇA}$ e $X = \text{RENDA}$ (ambas medidas anualmente para a população de todas as famílias). Suponha que $\text{Var}(\text{POUPANÇA}|\text{RENDA})$

$= 400 + 0,25 \text{ RENDA}$. Isso quer dizer que, à medida que a renda aumenta, a variância dos níveis de poupança também aumenta. É importante ver que a relação entre a variância de POUPANÇA e RENDA é totalmente separada daquela entre os *valores esperados* de POUPANÇA e RENDA.

Expressamos uma propriedade útil a respeito da variância condicional.

Propriedade VC.1: Se X e Y são independentes, então $\text{Var}(Y|X) = \text{Var}(Y)$.

Essa propriedade é bem clara, já que a distribuição de Y , dado X , não depende de X , e $\text{Var}(Y|X)$ é apenas uma característica dessa distribuição.

B.5 Distribuições normal e relacionada

B.5a Distribuição normal

A distribuição normal e aquelas derivadas dela são as distribuições mais usadas em estatística e econometria. Supor que as variáveis aleatórias definidas para populações sejam normalmente distribuídas simplifica os cálculos de probabilidade. Além disso, contaremos muito com as distribuições normal e relacionada para conduzir inferência em estatística e econometria – mesmo quando a população subjacente não for necessariamente normal. Devemos adiar os detalhes, mas tenha certeza de que essas distribuições vão surgir muitas vezes ao longo do texto.

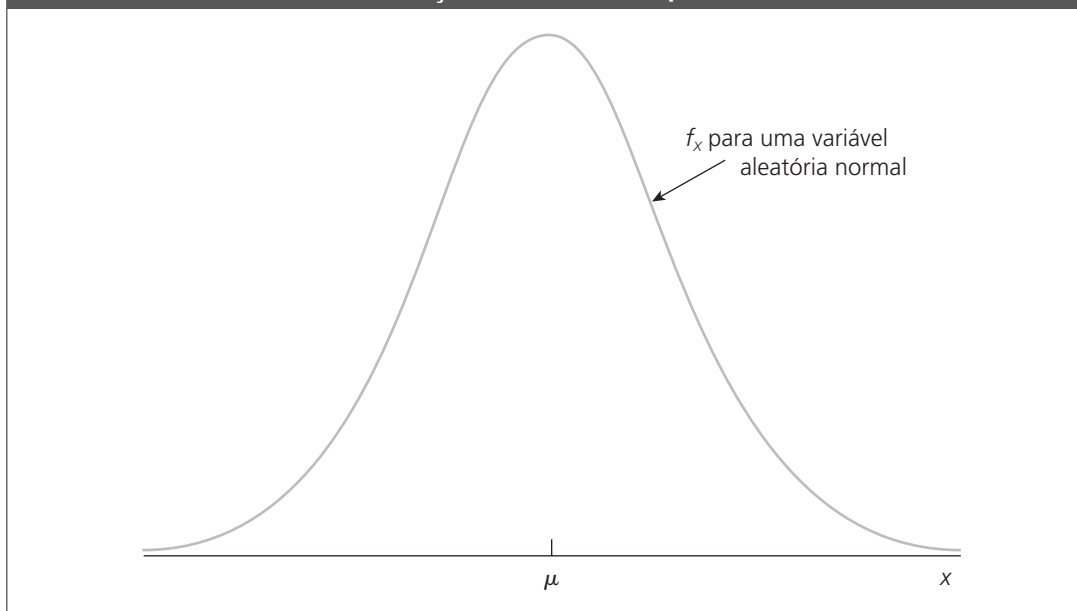
Uma variável aleatória normal é uma variável aleatória contínua que pode assumir qualquer valor. Sua função de densidade de probabilidade tem o formato familiar de um sino, mostrado na Figura B.7.

Matematicamente, a fdp de X pode ser escrita como

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-(x - \mu)^2/2\sigma^2\right], \quad -\infty < x < \infty, \quad (\text{B.34})$$

onde $\mu = E(X)$ e $\sigma^2 = \text{Var}(X)$. Dizemos que X tem uma **distribuição normal** com valor esperado μ e variância σ^2 , escrita como $X \sim \text{Normal}(\mu, \sigma^2)$. Como a distribuição normal é simétrica em relação a μ , μ também é a mediana de X . A distribuição normal às vezes é chamada de *distribuição gaussiana*, em homenagem ao famoso matemático C. F. Gauss.

Certas variáveis aleatórias parecem seguir aproximadamente uma distribuição normal. Altura e peso humano, notas de provas e taxas de desemprego de um condado têm fdp's com um formato aproximado ao da Figura B.7. Outras distribuições, como distribuições de renda, não parecem seguir a função normal de probabilidade. Na maioria dos países, a renda não é simetricamente distribuída em torno de qualquer valor; a distribuição é oblíqua na direção da extremidade superior. Em alguns casos, uma variável pode ser transformada para alcançar a normalidade. Uma transformação popular é o log natural, que faz sentido para variáveis aleatórias positivas. Se X é uma variável aleatória positiva, como renda, e $Y = \log(X)$ tem uma distribuição normal, dizemos que X tem uma *distribuição log-normal*. O resultado é que a distribuição log-normal enquadra a renda muito bem em vários países. Outras variáveis, como preço de mercadorias, parecem ser bem descritas como distribuídas log-normalmente.

FIGURA B.7 O formato usual da função de densidade de probabilidade normal.

B.5b Distribuição normal padrão

Um caso especial da distribuição normal ocorre quando sua média é zero e a variância (e, portanto, o desvio padrão) é um. Se uma variável aleatória Z tem uma distribuição Normal(0,1), dizemos que ela tem uma **distribuição normal padrão**. A fdp de uma variável aleatória normal padrão é indicada por $\phi(z)$; com base em (B.34), com $\mu = 0$ e $\sigma^2 = 1$, ela é dada por

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2), \quad -\infty < z < \infty. \quad (\text{B.35})$$

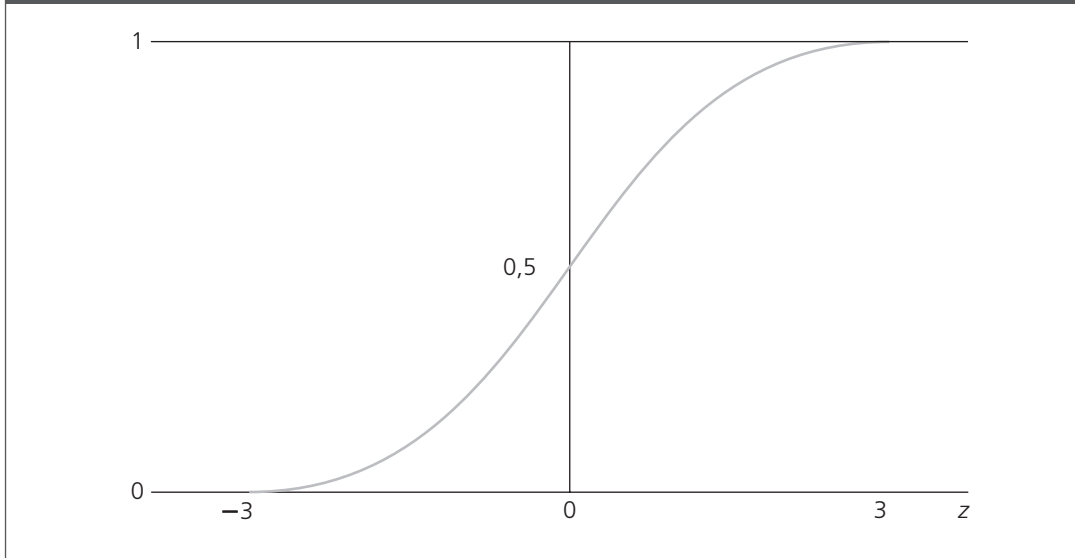
A função de distribuição normal cumulativa padrão é indicada por $\Phi(z)$ e obtida como a área abaixo de ϕ , à esquerda de z ; ver Figura B.8. Lembre-se de que $\Phi(z) = P(Z \leq z)$; já que Z é contínua, $\Phi(z) = P(Z < z)$ também é.

Não há uma fórmula simples que possa ser usada para obter os valores de $\Phi(z)$ [porque $\Phi(z)$ é a integral da função de (B.35), e essa integral não tem forma fechada]. No entanto, os valores para $\Phi(z)$ são facilmente tabulados; eles são dados para z entre $-3,1$ e $3,1$ na Tabela G.1, do Apêndice G. Para $z \leq -3,1$, $\Phi(z)$ é menor do que 0,001, e para $z \geq 3,1$, $\Phi(z)$ é maior do que 0,999. A maioria dos programas de estatística e econometria inclui comandos simples para calcular valores da fdc normal padrão, assim podemos evitar totalmente as tabelas impressas e obter as probabilidades de qualquer valor de z .

Usando fatos básicos da probabilidade – e, em especial, as propriedades (B.7) e (B.8) em relação às fdcs – podemos utilizar a fdc normal padrão para calcular a probabilidade de qualquer evento que envolva uma variável aleatória normal padrão. As fórmulas mais importantes são

$$P(Z > z) = 1 - \Phi(z), \quad (\text{B.36})$$

$$P(Z < -z) = P(Z > z), \quad (\text{B.37})$$

FIGURA B.8 Função de distribuição cumulativa normal padrão.

e

$$P(a \leq Z \leq b) = \Phi(b) - \Phi(a). \quad (\text{B.38})$$

Como Z é uma variável aleatória contínua, todas as três fórmulas se mantêm se as desigualdades forem ou não estritas. Alguns exemplos incluem $P(Z > 0,44) = 1 - 0,67 = 0,33$, $P(Z < -0,92) = P(Z > 0,92) = 1 - 0,821 = 0,179$ e $P(-1 < Z \leq 0,5) = 0,692 - 0,159 = 0,533$.

Outra expressão útil é que, para qualquer $c > 0$,

$$\begin{aligned} P(|Z| > c) &= P(Z > c) + P(Z < -c) \\ &= 2 \cdot P(Z > c) = 2[1 - \Phi(c)]. \end{aligned} \quad (\text{B.39})$$

Dessa forma, a probabilidade de que o valor absoluto de Z seja maior do que alguma constante positiva c é simplesmente duas vezes a probabilidade $P(Z > c)$; isso reflete a simetria da distribuição normal padrão.

Na maioria das aplicações, começamos com uma variável aleatória normalmente distribuída, $X \sim \text{Normal}(\mu, \sigma^2)$, em que μ é diferente de zero e $\sigma^2 \neq 1$. Qualquer variável aleatória normal pode ser transformada em normal padrão usando a propriedade a seguir.

Propriedade Normal.1: Se $X \sim \text{Normal}(\mu, \sigma^2)$, então $(X - \mu)/\sigma \sim \text{Normal}(0, 1)$.

A Propriedade Normal.1 mostra como transformar qualquer variável aleatória normal em uma normal padrão. Assim, suponha que $X \sim \text{Normal}(3, 4)$, e que desejemos calcular $P(X \leq 1)$. As etapas sempre envolvem a normalização de X para uma normal padrão:

$$\begin{aligned} P(X \leq 1) &= P(X - 3 \leq 1 - 3) = P\left(\frac{X - 3}{2} \leq -1\right) \\ &= P(Z \leq -1) = \Phi(-1) = 0,159. \end{aligned}$$

EXEMPLO B.6**Probabilidades de uma variável aleatória normal**

Primeiro, vamos calcular $P(2 < X \leq 6)$ quando $X \sim \text{Normal}(4, 9)$ (usar $<$ ou $\leq$ é irrelevante porque X é uma variável aleatória contínua). Agora,

$$\begin{aligned} P(2 < X \leq 6) &= P\left(\frac{2-4}{3} < \frac{X-4}{3} \leq \frac{6-4}{3}\right) = P(-2/3 < Z \leq 2/3) \\ &= \Phi(0,67) - \Phi(-0,67) = 0,749 - 0,251 = 0,498. \end{aligned}$$

Agora, vamos calcular $P(|X| > 2)$:

$$\begin{aligned} P(|X| > 2) &= P(X > 2) + P(X < -2) \\ &= P[(X-4)/3 > (2-4)/3] + P[(X-4)/3 < (-2-4)/3] \\ &= 1 - \Phi(-2/3) + \Phi(-2) \\ &= 1 - 0,251 + 0,023 = 0,772. \end{aligned}$$

B.5c Propriedades adicionais da distribuição normal

Terminaremos essa subseção reunindo diversos outros fatos a respeito das distribuições normais que usaremos mais adiante.

Propriedade Normal.2: $X \sim \text{Normal}(\mu, \sigma^2)$, então $aX + b \sim \text{Normal}(a\mu + b, a^2\sigma^2)$.

Assim, se $X \sim \text{Normal}(1, 9)$, então $Y = 2X + 3$ é distribuída como normal com média $2E(X) + 3 = 5$ e variância $2^2 \cdot 9 = 36$; $\text{dp}(Y) = 2\text{dp}(X) = 2 \cdot 3 = 6$.

Anteriormente, discutimos como, em geral, zero correlação e independência não são a mesma coisa. No caso de variáveis aleatórias normalmente distribuídas, a correlação zero é suficiente para independência.

Propriedade Normal.3: Se X e Y são distribuídas normalmente de forma conjunta, então elas serão independentes se, e somente se, $\text{Cov}(X, Y) = 0$.

Propriedade Normal.4: Qualquer combinação de variáveis aleatórias independentes e normalmente distribuídas de forma idêntica tem uma distribuição normal.

Por exemplo, sejam X_i , para $i = 1, 2$ e 3 , variáveis aleatórias independentes distribuídas como $\text{Normal}(\mu, \sigma^2)$. Defina $W = X_1 + 2X_2 - 3X_3$. Então, W é normalmente distribuída; devemos simplesmente encontrar sua média e variância. Agora,

$$E(W) = E(X_1) + 2E(X_2) - 3E(X_3) = \mu + 2\mu - 3\mu = 0.$$

Além disso,

$$\text{Var}(W) = \text{Var}(X_1) + 4\text{Var}(X_2) + 9\text{Var}(X_3) = 14\sigma^2.$$

A Propriedade Normal.4 também indica que a média de variáveis aleatórias independentes e normalmente distribuídas tem uma distribuição normal. Se $Y_1, Y_2, \dots, Y_n$ são variáveis independentes e cada uma delas é distribuída como $\text{Normal}(\mu, \sigma^2)$, então

$$\bar{Y} \sim \text{Normal}(\mu, \sigma^2/n). \quad (\text{B.40})$$

Esse resultado é crítico para a inferência estatística sobre a média em uma população normal.

Outras características da distribuição normal valem a pena ser conhecidas, embora elas não desempenhem papel central neste livro. Como uma variável aleatória normal é simétrica em torno de sua média, ela tem zero assimétrica, isto é, $E[(X - \mu)^3] = 0$. Além disso, podemos demonstrar que

$$E[(X - \mu)^4]/\sigma^4 = 3,$$

Ou $E(Z^4) = 3$, onde Z tem uma distribuição normal padrão. Já que a distribuição normal é tão prevalente em probabilidade e em estatística, a medida de curtose para qualquer determinada variável aleatória X (cujo quarto momento exista) geralmente é definida como $E[(X - \mu)^4]/\sigma^4 - 3$, ou seja, em relação ao valor da distribuição normal padrão. Se $E[(X - \mu)^4]/\sigma^4 > 3$, a distribuição de X terá caudas mais largas do que a distribuição normal (uma ocorrência comum, de certa forma, como aquela com a distribuição t , a ser introduzida em breve); se $E[(X - \mu)^4]/\sigma^4 < 3$, a distribuição terá extremidades mais leves do que a normal (uma situação mais rara).

B.5d Distribuição qui-quadrado

A distribuição qui-quadrada é obtida diretamente de variáveis aleatórias normais padrão independentes. Sejam Z_i , $i = 1, 2, \dots, n$, variáveis aleatórias independentes, cada uma delas distribuída como uma normal padrão. Defina uma nova variável aleatória como a soma dos quadrados de Z_i :

$$X = \sum_{i=1}^n Z_i^2. \quad (\text{B.41})$$

Assim, X tem o que é conhecido como **distribuição qui-quadrado** com n **graus de liberdade** (ou *gl*, para abreviar). Escrevemos isso como $X \sim X_n^2$. O *gl* em uma distribuição qui-quadrado corresponde ao número de termos na soma de (B.41). O conceito de graus de liberdade terá papel importante em nossas análises estatísticas e econométricas.

A fdp para distribuições qui-quadrados com variados graus de liberdade é dada na Figura B.9; não precisaremos de fórmula para essa fdp, por isso, não vamos reproduzi-la aqui. Com base na equação (B.41), fica claro que uma variável aleatória qui-quadrado é sempre não negativa e que, diferentemente da distribuição normal, a distribuição qui-quadrado não é simétrica em torno de nenhum ponto. Pode-se demonstrar que, se $X \sim \chi_n^2$, o valor esperado de X é n [o número de termos em (B.41)] e a variância de X é $2n$.

B.5e Distribuição t

A distribuição t é o carro-chefe da estatística clássica e da análise de regressão múltipla. Obtemos uma distribuição t a partir de uma variável aleatória normal padrão e uma qui-quadrado.

Considere que Z tenha uma distribuição normal padrão e X , uma distribuição qui-quadrado com n graus de liberdade. Além disso, suponha que X e Z sejam independentes. Assim, a variável aleatória

$$T = \frac{Z}{\sqrt{X/n}} \quad (\text{B.42})$$

tem uma **distribuição t** com n graus de liberdade. Vamos simbolizar isso com $T \sim t_n$. O número de graus de liberdade de distribuição t é obtido a partir da variável aleatória qui-quadrado do denominador de (B.42).

A fdp da distribuição t tem um formato parecido com a da distribuição normal padrão, exceto por ser mais aberta e, portanto, ter mais área dentro das caudas. O valor esperado de uma variável aleatória t distribuída é zero (falando de forma absoluta, o valor esperado existe somente para $n > 1$), e a variância é $n/(n-2)$ para $n > 2$. (A variância não existe para $n \leq 2$ porque a distribuição é muito aberta.) A fdp da distribuição t foi esboçada na Figura B.10 para vários graus de liberdade. À medida que os graus de liberdade aumentam, a distribuição t se aproxima da distribuição normal padrão.

B.5f Distribuição F

Outra distribuição importante para a estatística e para a econometria é a distribuição F . Em especial, a distribuição F será usada para testar hipóteses no contexto da análise de regressão múltipla.

Para definir uma variável aleatória F , defina $X_1 \sim X_{k_1}^2$ e $X_2 \sim X_{k_2}^2$ e suponha que X_1 e X_2 sejam independentes. Assim, a variável aleatória

$$F = \frac{(X_1/k_1)}{(X_2/k_2)} \quad (\text{B.43})$$

tem uma **distribuição F** com (k_1, k_2) graus de liberdade. Simbolizamos isso como $F \sim F_{k_1, k_2}$. A fdp da distribuição F com diferentes graus de liberdade é dada na Figura B.11.

FIGURA B.9 Distribuição qui-quadrado com vários graus de liberdade.

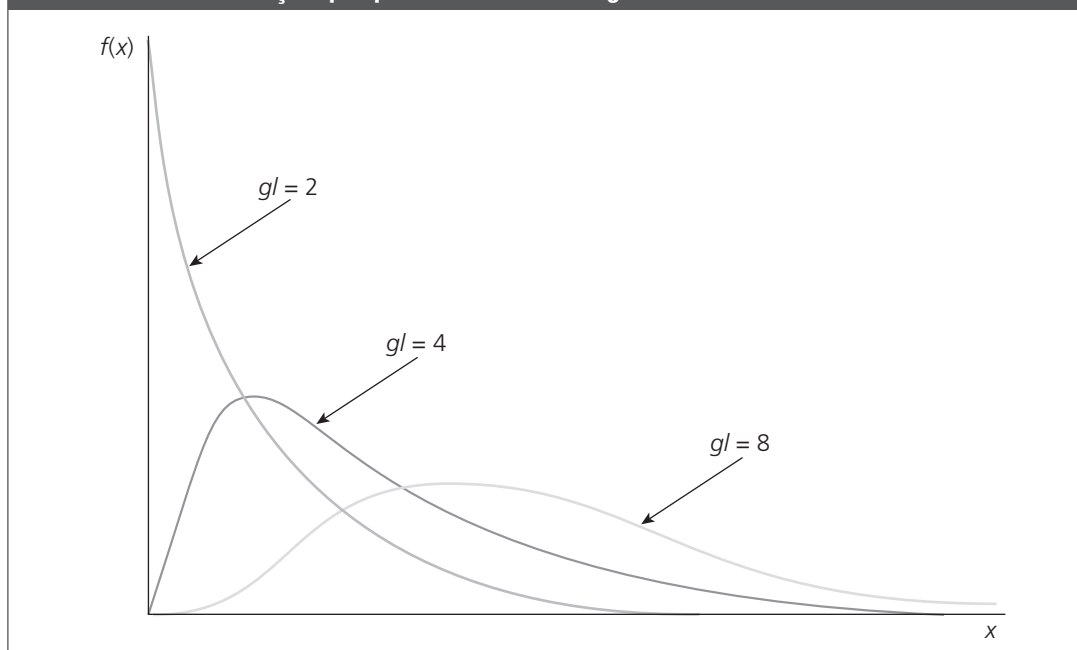
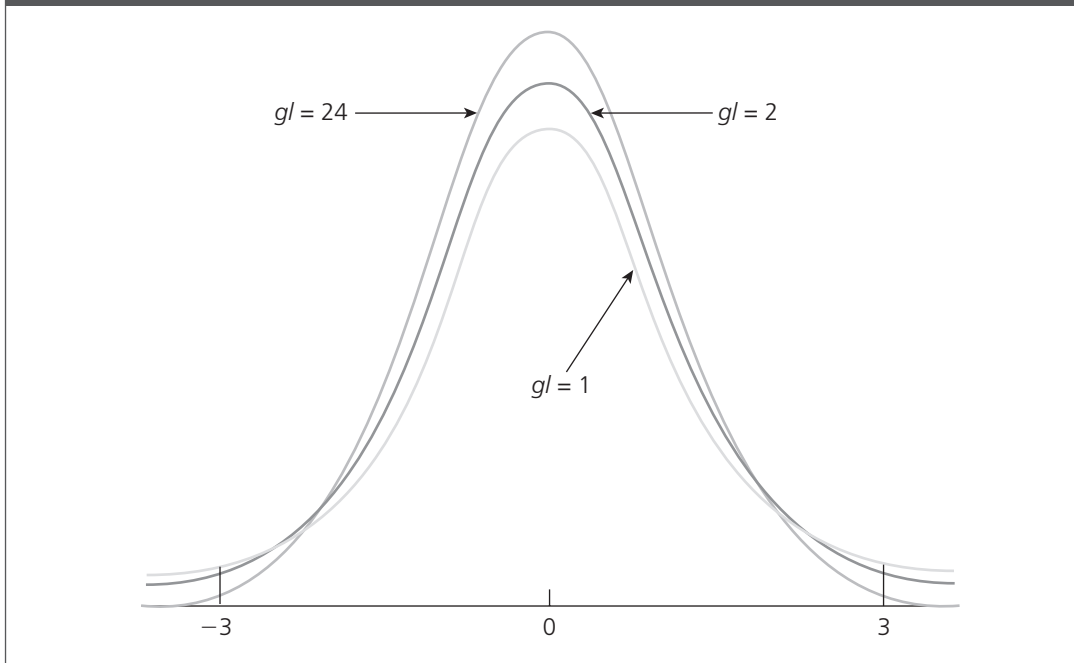
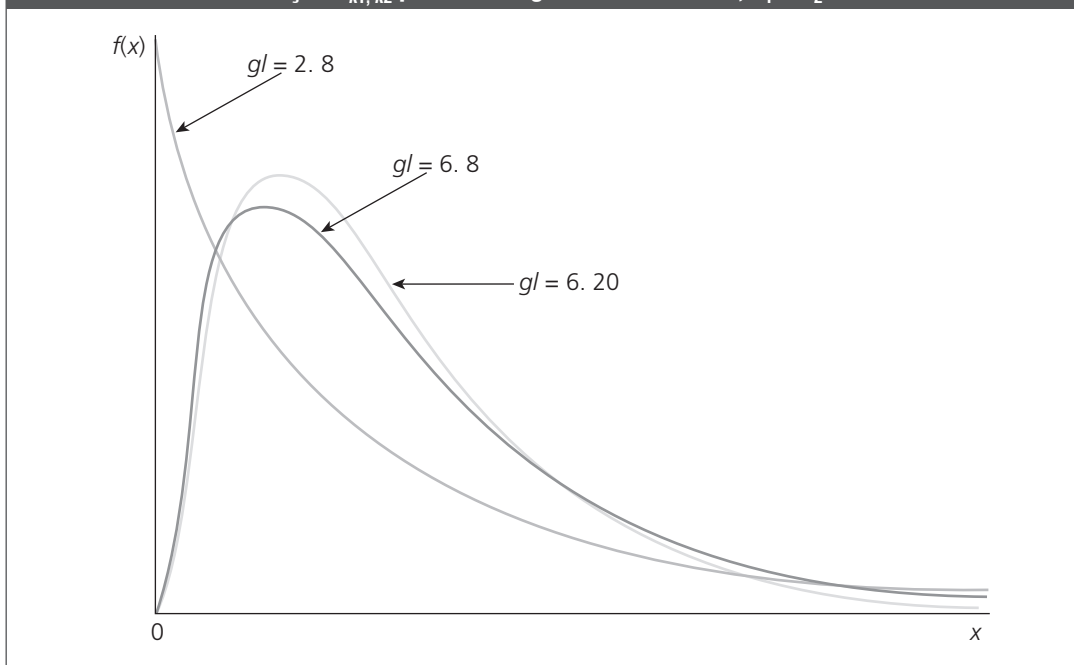


FIGURA B.10 Distribuição t com vários graus de liberdade.**FIGURA B.11** Distribuição F_{k_1, k_2} para vários graus de liberdade, k_1 e k_2 .

A ordem dos graus de liberdade em F_{k_1, k_2} é crítica. O inteiro k_1 é chamado de *grau de liberdade do numerador* porque está associado à variável qui-quadrado do numerador. Da mesma forma, o inteiro k_2 é chamado de *grau de liberdade do denominador* porque está associado à variável qui-quadrado do denominador. Isso pode ser

um pouco complicado porque (B.43) também pode ser escrita como $(X_1 k_2)/(X_2 k_1)$, de forma que k_1 aparece no denominador. É só se lembrar de que o gl do numerador é o inteiro associado com a variável qui-quadrado do numerador de (B.43), e da mesma forma com o gl do denominador.

Resumo

Neste apêndice, revisamos os conceitos de probabilidade que são necessários na econometria. A maioria dos conceitos deve ser familiar ao curso introdutório de probabilidade e estatística. Alguns dos tópicos mais avançados, como as características das expectativas condicionais, não precisam ser dominados agora – ainda há tempo até que esses conceitos surjam no contexto da análise de regressão da Parte 1.

Em um curso introdutório de estatística, o foco está no cálculo das médias, variâncias, covariâncias, e por aí vai, de determinadas distribuições. Na Parte 1, não precisaremos destes cálculos: contaremos, na maior parte do tempo, com as *propriedades* das expectativas, variâncias e outras que foram expostas neste apêndice.

Termos-chave

Assimetria	Distribuição t	Variáveis aleatórias independentes
Coefficiente de correlação	Expectativa condicional	Variáveis aleatórias não correlacionadas
Covariância	Experimento	Variáveis aleatórias não correlacionadas em pares
Curtose	Função de densidade de probabilidade (fdp)	Variável aleatória
Desvio padrão	Função de distribuição cumulativa (fdc)	Variável aleatória contínua
Distribuição binomial	Graus de liberdade	Variável aleatória de Bernoulli (ou binária)
Distribuição condicional	Lei das expectativas iteradas	Variável aleatória discreta
Distribuição conjunta	Mediana	Variável aleatória padronizada
Distribuição F	Valor esperado	
Distribuição normal	Variância	
Distribuição normal padrão		
Distribuição qui-quadrado		
Distribuição simétrica		

Problemas

1. Suponha que um estudante do ensino médio esteja se preparando para fazer o teste SAT. Explique porque sua eventual nota do SAT é adequadamente vista como uma variável aleatória.
2. Defina X como uma variável aleatória distribuída como uma Normal(5, 4). Encontre as probabilidades dos seguintes eventos:
 - (i) $P(X \leq 6)$.
 - (ii) $P(X > 4)$.
 - (iii) $P(|X - 5| > 1)$.
3. Muito tem sido feito a respeito do fato de determinados fundos mútuos superarem o desempenho do mercado depois de um ano (ou seja, o retorno por manter ações do fundo mútuo é maior do que o retorno por manter uma carteira como S&P 500). Por materialidade, considere um período de 10 anos e defina a população como os 4.170 fundos mútuos registrados no *The Wall Street Journal* de 1 de janeiro de 1995. Ao

dizer que o desempenho em relação ao mercado é aleatório, significa que cada fundo tem uma chance 50-50 de superar o mercado em qualquer ano e que o desempenho é independente ano a ano.

- (i) Se o desempenho em relação ao mercado é realmente aleatório, qual é a probabilidade de que qualquer fundo em particular supere o mercado em todos os 10 anos?
 - (ii) Dos 4.170 fundos mútuos, qual é o número esperado de fundos que vai superar o mercado em todos os 10 anos?
 - (iii) Encontre a probabilidade de que *pelo menos* um fundo entre os 4.170 supere o desempenho do mercado em todos os 10 anos. O que você conclui dessa resposta?
 - (iv) Se você tiver um pacote estatístico que calcule probabilidades binomiais, encontre a probabilidade de que pelo menos cinco fundos superem o desempenho do mercado em todos os 10 anos.
4. Para um condado selecionado aleatoriamente nos Estados Unidos, defina X como a proporção de adultos com mais de 65 anos que estão empregados, ou a taxa de emprego na terceira idade. Assim, X está restrito a um valor entre zero e um. Suponha que a função de distribuição cumulativa de X seja dada por $F(x) = 3x^2 - 2x^3$ para $0 \leq x \leq 1$. Encontre a probabilidade de a taxa de emprego na terceira idade ser, no mínimo, 0,6 (60%).
5. Pouco antes da seleção do júri para o julgamento de assassinato de O. J. Simpson, em 1995, uma pesquisa revelou que cerca de 20% da população adulta acreditava que Simpson era inocente (depois de grande parte das evidências físicas do caso ter sido revelada ao público). Ignore o fato de que 20% é uma estimativa baseada em uma subamostra da população; para ilustração, considere o valor uma porcentagem real de pessoas que acreditavam que Simpson era inocente antes da seleção do júri. Suponha que os 12 jurados tenham sido selecionados de forma aleatória e independente da população (embora isso tenha se provado não verdadeiro).
- (i) Encontre a probabilidade de o júri ter pelo menos um membro que acreditava na inocência de Simpson antes da seleção do júri. [Dica: Defina a variável aleatória Binomial(12, 0,20) X como o número de jurados que acredita na inocência de Simpson.]
 - (ii) Encontre a probabilidade de o júri ter pelo menos dois membros que acreditavam na inocência de Simpson. [Dica: $P(X \geq 2) = 1 - P(X \leq 1)$ e $P(X \leq 1) = P(X = 0) + P(X = 1)$.]
6. (Exige cálculo) Defina X como a sentença de prisão, em anos, para pessoas condenadas por roubo de veículos em determinado estado norte-americano. Suponha que a fdp de X seja dada por

$$f(x) = (1/9)x^2, 0 < x < 3.$$

Use integração para encontrar a sentença esperada.

7. Se um jogador de basquete é um arremessador de lances livres de 74%, então, em média, quantos lances livres ele converterá em um jogo com oito tentativas?
8. Suponha que um estudante universitário esteja fazendo três cursos: um curso de dois créditos, um de três créditos e um de quatro créditos. A nota esperada no curso de dois créditos é 3,5, enquanto a nota esperada para os cursos de três e quatro créditos é 3. Qual é a pontuação média geral esperada para o semestre? (Lembre-se de que a nota de cada curso é ponderada por sua parcela do número total de unidades.)

9. Seja X o salário anual de professores universitários dos Estados Unidos, medido em milhares de dólares. Suponha que o salário médio seja 52,3, com um desvio padrão de 14,6. Encontre a média e o desvio padrão quando o salário for medido em dólares.
10. Suponha que, em uma grande universidade, a nota média, GPA , e a pontuação no SAT, SAT , sejam relacionadas pela expectativa condicional $E(GPA|SAT) = 0,70 + 0,002 SAT$.
- (i) Encontre a GPA esperada quando $SAT = 800$. Encontre $E(GPA|SAT = 1.400)$. Comente sobre a diferença.
 - (ii) Se a SAT média na universidade é 1.100, qual é a GPA média? (Dica: Use a Propriedade EC.4.)
 - (iii) Se a pontuação SAT de um estudante é 1.100, isso significa que ele terá a GPA descoberta no item (ii)? Explique.
11. (i) Defina X como uma variável aleatória que assume os valores -1 e 1 , cada um com probabilidade $1/2$. Encontre $E(X)$ e $E(X^2)$.
- (ii) Agora, seja X uma variável aleatória que assume os valores 1 e 2 , cada um com probabilidade $1/2$. Encontre $E(X)$ e $E(1/X)$.
- (iii) Conclua a partir dos itens (i) e (ii) que, em geral,

$$E[g(X)] \neq g[E(X)]$$

para uma função não linear $g(\cdot)$.

- (iv) Dada a definição da variável aleatória F na equação (B.43), demonstre que

$$E(F) = E\left[\frac{1}{(X_2/k_2)}\right].$$

É possível concluir que $E(F) = 1$?

Fundamentos da estatística matemática

C.1 Populações, parâmetros e amostragem aleatória

A inferência estatística envolve descobrir algo sobre uma população, dada a disponibilidade de uma amostra dessa população. Por **população** queremos dizer qualquer grupo bem definido de sujeitos, que podem ser indivíduos, empresas, cidades ou muitas outras possibilidades. “Descobrir” pode significar várias coisas, que são amplamente divididas entre as categorias de *estimação* e *teste de hipótese*.

Alguns exemplos podem ajudar a entender esses termos. Na população de todos os adultos trabalhadores dos Estados Unidos, os economistas do trabalho estão interessados em descobrir sobre o retorno à educação, medido pelo aumento percentual médio nos rendimentos com um ano a mais de educação. Seria pouco prático e muito custoso obter informações sobre rendimentos e educação de toda a população trabalhadora dos Estados Unidos, mas podemos obter dados de um subconjunto da população. Usando os dados coletados, um economista do trabalho pode registrar que sua melhor estimativa do retorno para outro ano de educação é 7,5%. Esse é um exemplo de uma *estimativa pontual*. Ou, ele poderia registrar uma escala, como “o retorno à educação é algo entre 5,6% e 9,4%”. Este é um exemplo de *estimativa intervalar*.

Um economista urbano pode saber se programas de vigilância de bairro estão associados a taxas de criminalidade mais baixas. Depois de comparar as taxas de criminalidade de bairros com e sem esses programas em uma amostra da população, ele pode traçar uma ou duas conclusões: que programas de vigilância de bairro afetam ou não a criminalidade. Este exemplo entra na categoria de teste de hipótese.

O primeiro passo na inferência estatística é identificar a população de interesse. Isso pode parecer óbvio, mas é importante ser muito específico. Uma vez identificada a população, conseguimos especificar um modelo para a relação populacional de interesse. Esses modelos envolvem distribuições de probabilidade ou recursos de distribuições de probabilidade, e

dependem de parâmetros desconhecidos. Parâmetros são simplesmente constantes que determinam as direções e as forças de relações entre variáveis. No exemplo da economia trabalhista apresentado há pouco, o parâmetro de interesse é o retorno à educação na população.

C.1a Amostragem

Para revisar a inferência estatística, focaremos no ambiente mais simples possível. Defina Y como uma variável aleatória que representa uma população com uma função de densidade de probabilidade $f(y; \theta)$, que depende de um único parâmetro θ . A função de densidade de probabilidade (fdp) de Y é supostamente conhecida, exceto pelo valor de θ ; diferentes valores de θ implicam diferentes distribuições populacionais e, portanto, estamos interessados no valor de θ . Se pudermos obter certos tipos de amostra da população, conseguiremos descobrir algo sobre θ . O esquema de amostragem mais fácil de trabalhar é a amostragem aleatória.

Amostragem aleatória. Se $Y_1, Y_2, \dots, Y_n$ são variáveis aleatórias independentes com uma função de densidade de probabilidade comum $f(y; \theta)$, então $\{Y_1, \dots, Y_n\}$ é chamada de **amostra aleatória** de $f(y; \theta)$; [ou amostra aleatória da população representada por $f(y; \theta)$].

Quando $\{Y_1, \dots, Y_n\}$ é uma amostra aleatória da densidade $f(y; \theta)$, também dizemos que Y_i são variáveis aleatórias *independentes e identicamente distribuídas* (ou *i.i.d.*) de $f(y; \theta)$. Em alguns casos, não precisaremos especificar totalmente o que é a distribuição comum.

A natureza aleatória de $Y_1, Y_2, \dots, Y_n$ na definição de amostragem aleatória reflete o fato de que muitos resultados diferentes são possíveis antes de a amostragem ser realmente posta em prática. Por exemplo, se a renda familiar de uma amostra de $n = 100$ famílias dos Estados Unidos for obtida, as rendas que observarmos geralmente vão diferir para cada amostra diferente de 100 famílias. Uma vez obtida a amostra, teremos um conjunto de números, digamos $(y_1, y_2, \dots, y_n)$, que constitui os dados com os quais vamos trabalhar. Saber se isso é ou não adequado para supor que a amostra veio de um esquema de amostragem aleatória exige conhecimento a respeito do processo real de amostragem.

Amostras aleatórias de uma distribuição de Bernoulli normalmente são usadas para ilustrar conceitos estatísticos, e elas também surgem em aplicações empíricas. Se $Y_1, Y_2, \dots, Y_n$ são variáveis aleatórias independentes e cada uma delas é distribuída como Bernoulli(θ), de forma que $P(Y_i = 1) = \theta$ e $P(Y_i = 0) = 1 - \theta$, então $\{Y_1, Y_2, \dots, Y_n\}$ constitui uma amostra aleatória da distribuição Bernoulli(θ). Como ilustração, considere o exemplo de reservas da companhia aérea usado no Apêndice B. Cada Y_i indica se o consumidor i comparecerá para sua reserva; $Y_i = 1$ se o passageiro i comparecer, e $Y_i = 0$, caso contrário. Aqui, θ é a probabilidade de uma pessoa aleatoriamente selecionada da população de todas as pessoas que fizeram reservas aéreas comparecer para sua reserva.

Para muitas outras aplicações, pode-se supor que amostras aleatórias sejam extraídas de uma distribuição normal. Se $\{Y_1, \dots, Y_n\}$ é uma amostra aleatória da população Normal (μ, σ^2) , a população será caracterizada por dois parâmetros, a média μ e a variância σ^2 . O principal interesse em geral está em μ , mas σ^2 é interessante por si só porque fazer inferências sobre μ normalmente exige aprender sobre σ^2 .

C.2 Propriedades de amostras finitas de estimadores

Nesta seção, estudaremos o que são chamadas de propriedades de amostras finitas de estimadores. O termo “amostra finita” vem do fato de que as propriedades se mantêm para amostras de qualquer tamanho, não importa quão grande ou pequena. Às vezes, elas são chamadas de propriedades de amostras pequenas. Na Seção C.3, trataremos de “propriedades assintóticas”, que têm a ver com o comportamento dos estimadores à medida que o tamanho da amostra aumenta sem limite.

C.2a Estimadores e estimativas

Para estudar propriedades de estimadores, precisamos definir o que queremos dizer com estimador. Dada uma amostra aleatória $\{Y_1, Y_2, \dots, Y_n\}$ retirada de uma distribuição populacional que depende de um parâmetro desconhecido θ , um **estimador** de θ é uma regra que atribui a cada possível resultado da amostra um valor de θ . A regra é especificada antes de qualquer amostra ser colocada em prática; em particular, a regra é a mesma independentemente dos dados realmente obtidos.

Como exemplo de estimador, defina $\{Y_1, \dots, Y_n\}$ como uma amostra aleatória de uma população com média μ . Um estimador natural de μ é a média da amostra aleatória:

$$\bar{Y} = n^{-1} \sum_{i=1}^n Y_i. \quad (\text{C.1})$$

$\bar{Y}$ é chamada de **média amostral**, mas, diferentemente do Apêndice A, em que definimos a média amostral de um conjunto de números como uma estatística descritiva, $\bar{Y}$ agora é vista como um estimador. Com quaisquer resultados das variáveis aleatórias $Y_1, \dots, Y_n$, usaremos a mesma regra para estimar μ : simplesmente calcularemos sua média. Para resultados de dados reais $\{y_1, \dots, y_n\}$, a **estimativa** é apenas a média da amostra: $\bar{y} = (y_1 + y_2 + \dots + y_n)/n$.

EXEMPLO C.1 Taxas de desemprego da cidade

Suponha que obtivemos a seguinte amostra de taxas de desemprego de 10 cidades dos Estados Unidos:

Cidade	Taxa de desemprego
1	5,1
2	6,4
3	9,2
4	4,1
5	7,5
6	8,3
7	2,6
8	3,5
9	5,8
10	7,5

Nossa estimativa da taxa de desemprego média das cidades dos Estados Unidos é $\bar{y} = 6,0$. Cada amostra geralmente resulta em uma estimativa diferente. Mas a *regra* para obter a estimativa é a mesma, independentemente de qual cidade apareça na amostra, ou quantas delas.

De forma mais geral, um estimador W de um parâmetro θ pode ser expresso como uma fórmula matemática abstrata:

$$W = h(Y_1, Y_2, \dots, Y_n), \quad (\text{C.2})$$

para qualquer função conhecida h das variáveis aleatórias $Y_1, Y_2, \dots, Y_n$. Assim como no caso especial da média amostral, W é uma variável aleatória porque depende da amostra aleatória: já que obtemos diferentes amostras aleatórias da população, o valor de W pode mudar. Quando determinado conjunto de números, por exemplo, $\{y_1, y_2, \dots, y_n\}$, está inserido na função h , obtemos uma *estimativa* de θ , indicada por $w = h(y_1, \dots, y_n)$. Às vezes, W é chamada de estimador pontual e w de estimativa pontual para distingui-las dos estimadores e das estimativas *do intervalo*, que veremos na Seção C.5.

Para avaliar procedimentos de estimação, estudamos várias propriedades da distribuição de probabilidade da variável aleatória W . A distribuição de um estimador geralmente é chamada de sua **distribuição por amostragem**, porque essa distribuição descreve a probabilidade de vários resultados de W em diferentes amostras aleatórias. Como existem regras ilimitadas para combinar dados para estimar parâmetros, precisamos de critérios conscientes para escolher entre os estimadores, ou pelo menos para eliminar alguns estimadores da consideração. Assim, teremos de deixar a estatística descritiva, em que calculamos coisas como a média amostral, para simplesmente resumir um corpo de dados. Na estatística matemática, estudamos a distribuição por amostragem dos estimadores.

C.2b Ausência de viés

A princípio, a distribuição por amostragem completa de W pode ser obtida dada a distribuição de probabilidade de Y_i e a função h . Normalmente é mais fácil focar em algumas características da distribuição de W ao avaliá-la como um estimador de θ . A primeira propriedade importante de um estimador envolve seu valor esperado.

Estimador não viesado. Um estimador, W de θ , será um **estimador não viesado** se

$$E(W) = \theta, \quad (\text{C.3})$$

para todos os possíveis valores de θ .

Se o estimador for não viesado, então sua distribuição de probabilidade terá um valor esperado igual ao parâmetro que deveria estar estimando. A ausência de viés *não* significa que a estimativa que obtivermos com qualquer amostra em particular será igual a θ , ou mesmo muito próxima a θ . Em vez disso, se pudermos extrair *indefinidamente* amostras aleatórias de Y da população, calcular uma estimativa em cada uma das vezes e obter uma média dessas estimativas em todas as amostras aleatórias, obteremos θ . Essa noção de experimento é abstrata porque, na maioria das aplicações, temos somente uma variável aleatória para trabalhar.

Para um estimador que é viesado, definimos seu **viés** da seguinte forma.

Viés de um estimador. Se W é um **estimador viesado** de θ , seu viés é definido

$$\text{Bias}(W) \equiv E(W) - \theta. \quad (\text{C.4})$$

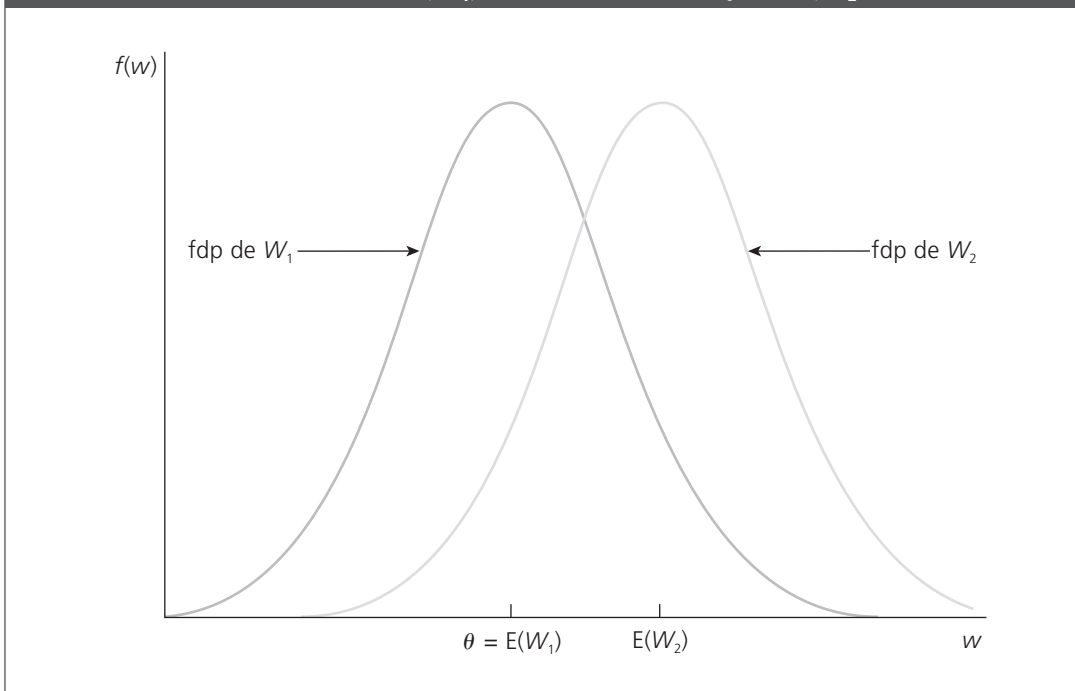
A Figura C.1 mostra dois estimadores; o primeiro é não viesado e o segundo tem um viés positivo.

A ausência de viés de um estimador e o tamanho de qualquer viés possível dependem da distribuição de Y e da função h . A distribuição de Y normalmente está além do nosso controle (embora escolhamos com frequência um *modelo* para a distribuição): ela pode ser determinada pela natureza ou por forças sociais. Mas a escolha da regra h é nossa, e se queremos um estimador não viesado, precisamos escolher h de forma apropriada.

Em geral, alguns estimadores podem ser mostrados como não viesados. Mostraremos agora que a média amostral $\bar{Y}$ é um estimador não viesado da média populacional μ , independentemente da distribuição populacional subjacente. Usaremos as propriedades dos valores esperados (E.1 e E.2) que abordamos na Seção B.3:

$$\begin{aligned} E(\bar{Y}) &= E\left((1/n) \sum_{i=1}^n Y_i\right) = (1/n)E\left(\sum_{i=1}^n Y_i\right) = (1/n)\left(\sum_{i=1}^n E(Y_i)\right) \\ &= (1/n)\left(\sum_{i=1}^n \mu\right) = (1/n)(n\mu) = \mu. \end{aligned}$$

FIGURA C.1 Estimador não viesado, W_1 , e estimador com viés positivo, W_2 .



Para teste de hipóteses, precisaremos estimar a variância σ^2 de uma população com média μ . Se $\{Y_1, \dots, Y_n\}$ indicar a amostra aleatória da população com $E(Y) = \mu$ e $\text{Var}(Y) = \sigma^2$, defina o estimador como

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2, \quad (\text{C.5})$$

que normalmente é chamada de **variância amostral**. Pode-se demonstrar que S^2 é não viesada para σ^2 : $E(S^2) = \sigma^2$. A divisão por $n-1$, em vez de n , corresponde ao fato de que a média μ é estimada, e não conhecida. Se μ fosse conhecida, um estimador não viesado de σ^2 seria $n^{-1} \sum_{i=1}^n (Y_i - \mu)^2$, mas μ raramente é conhecida na prática.

Embora a ausência de viés tenha certo apelo como propriedade para um estimador – na verdade, seu antônimo, “viesado”, tem, decididamente, conotações negativas –, ela tem seus problemas. Uma fraqueza da ausência de viés é que alguns estimadores razoáveis, e até alguns muito bons, são não viesados. Veremos um exemplo em breve.

Outra importante fraqueza da ausência de viés é que existem estimadores viesados que são estimadores realmente ruins. Considere estimar a média μ de uma população. Em vez de usar a média $\bar{Y}$ amostral para estimar μ , suponha que, depois de coletar uma amostra de tamanho n , descartemos todas as observações, exceto a primeira. Ou seja, nosso estimador de μ é simplesmente $W \equiv Y_1$. Este estimador é não viesado porque $E(Y_1) = \mu$. Por sorte, você percebe que ignorar todas as observações menos a primeira não é uma abordagem prudente para estimação: ela rejeita a maior parte da informação da amostra. Por exemplo, com $n = 100$, obtemos 100 resultados para a variável aleatória Y , mas usamos só a primeira delas para estimar $E(Y)$.

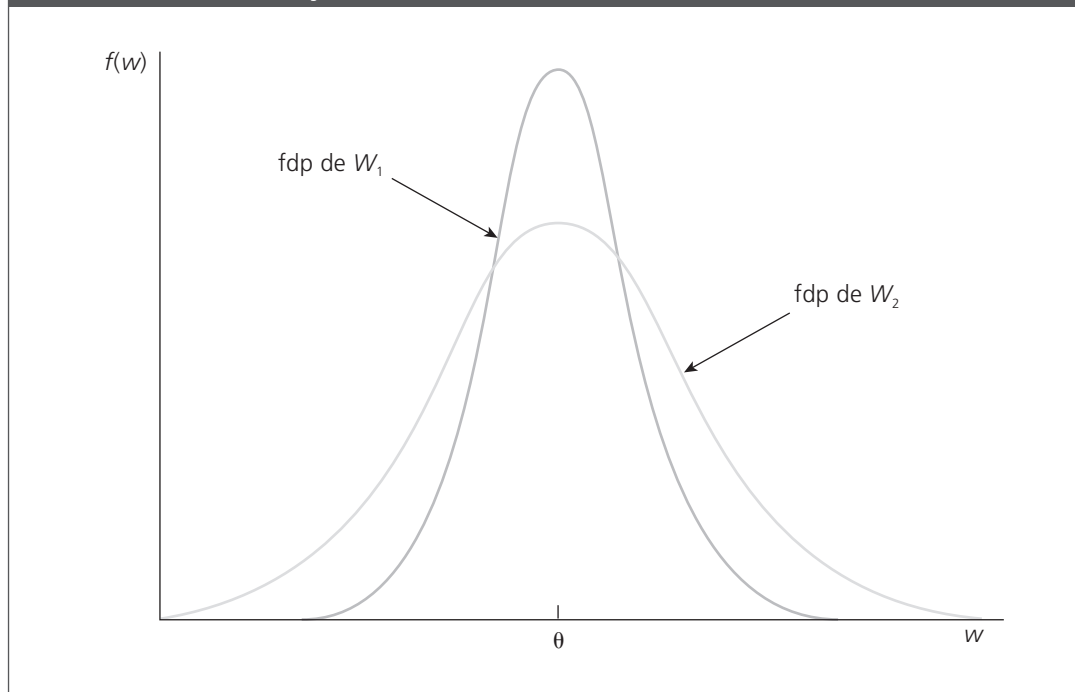
C2.d Variância por amostragem dos estimadores

O exemplo do fim da subseção anterior mostra que precisamos de critérios adicionais para avaliar estimadores. A ausência de viés só garante que a distribuição por amostragem de um estimador tem um valor médio igual ao parâmetro que supostamente está estimando. Isso é bom, mas também precisamos saber quão estendida é a distribuição de um estimador. Um estimador pode ser igual a θ , em média, mas também pode estar bem longe com grande probabilidade. Na Figura C.2, W_1 e W_2 são estimadores não viesados de θ . No entanto, a distribuição de W_1 é mais intimamente centrada ao redor de θ ; a probabilidade de que W_1 seja maior do que qualquer dada distância de θ é menor do que a probabilidade de W_2 ser maior do que a mesma distância de θ . Usar W_1 como nosso estimador significa que é menos provável que obtenhamos uma amostra aleatória que gere uma estimativa muito distante de θ .

Para resumir a situação mostrada na Figura C.2, contamos com a variância (ou desvio padrão) de um estimador. Recorde que isso dá uma única medida da dispersão na distribuição. A variância de um estimador geralmente é chamada de **variância por amostragem** porque é a variância associada com uma distribuição por amostragem. Lembre-se, a variância por amostragem não é uma variável aleatória; ela é uma constante, mas deve ser desconhecida.

Obteremos agora a variância da média amostral para estimar a média μ de uma população:

$$\begin{aligned} \text{Var}(\bar{Y}) &= \text{Var}\left(\left(\frac{1}{n}\right) \sum_{i=1}^n Y_i\right) = \left(\frac{1}{n^2}\right) \text{Var}\left(\sum_{i=1}^n Y_i\right) = \left(\frac{1}{n^2}\right) \left(\sum_{i=1}^n \text{Var}(Y_i)\right) \\ &= \left(\frac{1}{n^2}\right) \left(\sum_{i=1}^n \sigma^2\right) = \left(\frac{1}{n^2}\right) (n\sigma^2) = \sigma^2/n. \end{aligned} \quad (\text{C.6})$$

FIGURA C.2 As distribuições por amostragem de dois estimadores não viesados de θ .

Note a forma como usamos as propriedades da variância das Seções B.3 e B.4 (VAR.2 e VAR.4), assim como a independência de Y_i . Para resumir: Se $\{Y_i: i = 1, 2, \dots, n\}$ é uma amostra aleatória de uma população com média μ e variância σ^2 , então $\bar{Y}$ tem a mesma média que a população, mas sua variância por amostragem é igual à variância populacional, σ^2 , dividida pelo tamanho da amostra.

Uma implicação importante de $\text{Var}(\bar{Y}) = \sigma^2/n$ é que ela pode se tornar muito próxima de zero aumentando o tamanho da amostra n . Essa é uma das principais características de um estimador razoável, e voltaremos a ela na Seção C.3.

Como sugerido pela Figura C.2, entre estimadores não viesados, preferimos o estimador com a menor variância. Isso nos permite eliminar certos estimadores do interesse. Para uma amostra aleatória de uma população com média μ e variância σ^2 , sabemos que $\bar{Y}$ é não viesada e que $\text{Var}(\bar{Y}) = \sigma^2/n$. E quanto ao estimador Y_1 , que é a primeira observação traçada? Como Y_1 é uma extração aleatória da população, $\text{Var}(Y_1) = \sigma^2$. Assim, a diferença entre $\text{Var}(Y_1)$ e $\text{Var}(\bar{Y})$ pode ser grande mesmo em tamanhos de amostra pequenos. Se $n = 10$, então $\text{Var}(Y_1)$ é 10 vezes maior do que $\text{Var}(\bar{Y}) = \sigma^2/10$. Isso nos dá uma maneira formal de excluir Y_1 como um estimador de μ .

Para enfatizar este ponto, a Tabela C.1 contém o resultado de um pequeno estudo simulatório. Usando o pacote estatístico Stata®, 20 amostras aleatórias de tamanho 10 foram geradas a partir de uma distribuição normal, com $\mu = 2$ e $\sigma^2 = 1$; estamos interessados em estimar μ aqui. Para cada uma das 20 amostras aleatórias, calculamos dois estimadores, y_1 e $\bar{y}$; esses valores foram listados na Tabela C.1. Como pode ser visto na tabela, os valores de y_1 são muito mais estendidos do que os de $\bar{y}$: y_1 varia de $-0,64$ a $4,27$, enquanto $\bar{y}$ varia apenas de $1,16$ a $2,58$. Além disso, em 16 dos 20 casos, $\bar{y}$ está mais próximo do que y_1 de $\mu = 2$. A média de y_1 ao longo das simulações é

de cerca de 1,89, enquanto a de $\bar{y}$ é 1,96. O fato de essas médias estarem próximas de 2 ilustra a ausência de viés de ambos os estimadores (e podemos deixar essas médias ainda mais próximas de 2 fazendo mais de 20 réplicas). Mas a comparação somente dos resultados médios entre as extrações aleatórias mascara o fato de que a média amostral $\bar{Y}$ é muito superior a Y_1 como um estimador de μ .

C.2e Eficiência

A comparação de $\bar{Y}$ e Y_1 da subseção anterior é um exemplo de uma abordagem geral para comparar diferentes estimadores não viesados.

Eficiência relativa. Se W_1 e W_2 são dois estimadores não viesados de θ , W_1 é eficiente relativo a W_2 quando $\text{Var}(W_1) \leq \text{Var}(W_2)$ para todo θ , com desigualdade estrita para pelo menos um valor de θ .

TABELA C.1 Simulação de estimadores para uma distribuição Normal(μ , 1) com $\mu = 2$.

Réplica	y_1	$\bar{y}$
1	-0,64	1,98
2	1,06	1,43
3	4,27	1,65
4	1,03	1,88
5	3,16	2,34
6	2,77	2,58
7	1,68	1,58
8	2,98	2,23
9	2,25	1,96
10	2,04	2,11
11	0,95	2,15
12	1,36	1,93
13	2,62	2,02
14	2,97	2,10
15	1,93	2,18
16	1,14	2,10
17	2,08	1,94
18	1,52	2,21
19	1,33	1,16
20	1,21	1,75

Anteriormente, mostramos que, para estimar a média populacional μ , $\text{Var}(\bar{Y}) < \text{Var}(Y_1)$ para qualquer valor de σ^2 quando $n > 1$. Assim, $\bar{Y}$ é eficiente relativo a Y_1 para estimar μ . Não é sempre que podemos escolher entre estimadores não viesados com base no critério de menor variância: dados dois estimadores não

viesados de θ , um pode ter menor variância para alguns valores de θ , enquanto o outro pode ter menor variância para os outros valores de θ .

Se restringirmos nossa atenção a uma determinada classe de estimadores, podemos mostrar que a média amostral tem a menor variância. O Problema C.2 pede que você mostre que $\bar{Y}$ tem a menor variância entre todos os estimadores não viesados que também são funções lineares de $Y_1, Y_2, \dots, Y_n$. As hipóteses são que Y_i têm média e variância comuns, e que são não correlacionados em pares.

Se não restringirmos nossa atenção a estimadores não viesados, a comparação de variâncias não faz sentido. Por exemplo, quando estimamos a média populacional μ , podemos usar um estimador trivial que é igual a zero, independentemente da amostra que extraímos. Naturalmente, a variância deste estimador é zero (já que é o mesmo valor para toda amostra aleatória). Mas o viés do estimador é $-\mu$, portanto é um estimador muito fraco quando $|\mu|$ é grande.

Uma forma de comparar estimadores que não são necessariamente não viesados é calcular o **erro quadrático médio (EQM)** dos estimadores. Se W é um estimador de θ , então o EQM de W é definido como $\text{EQM}(W) = E[(W - \theta)^2]$. O EQM mede quão distante, em média, o estimador está de θ . Pode-se demonstrar que $\text{EQM}(W) = \text{Var}(W) + [\text{Viés}(W)]^2$, de forma que $\text{EQM}(W)$ depende da variância e do viés (se houver algum presente). Isso nos permite comparar dois estimadores quando um ou dois são viesados.

C.3 Propriedades dos estimadores assintóticos ou de amostras grandes

Na Seção C.2, encontramos o estimador Y_1 para a média populacional μ e vimos que, embora ele seja não viesado, é um estimador fraco porque sua variância pode ser muito maior do que a da média amostral. Uma característica notável de Y_1 é que ele tem a mesma variância para qualquer tamanho de amostra. Parece razoável exigir que um procedimento de estimação melhore à medida que o tamanho da amostra aumenta. Para estimar uma média populacional μ , $\bar{Y}$ melhora no sentido que sua variância fica menor à medida que n aumenta; Y_1 não melhora nesse sentido.

Podemos excluir certos estimadores absurdos ao estudar as propriedades *assintóticas* ou *de amostras grandes* dos estimadores. Além disso, podemos dizer algo positivo a respeito dos estimadores que são não viesados e cujas variâncias não são facilmente encontradas.

A análise assintótica envolve a aproximação das características da distribuição por amostragem de um estimador. Essas aproximações dependem do tamanho da amostra. Infelizmente, somos necessariamente limitados no que podemos dizer sobre quão “grande” é preciso que uma amostra seja para a análise assintótica ser adequada; isso depende da distribuição da população subjacente. Porém, sabe-se que aproximações de amostras grandes funcionam bem em tamanhos de amostra tão pequenas quanto $n = 20$.

C.3a Consistência

A primeira propriedade assintótica de estimadores se relaciona com a distância que o estimador é propenso a estar do parâmetro que se presume estimar à medida que deixamos o tamanho da amostra aumentar indefinidamente.

Consistência. Defina W_n como um estimador de θ baseado em uma amostra $Y_1, Y_2, \dots, Y_n$ de tamanho n . Assim, W_n é um **estimador consistente** de θ para todo $\varepsilon > 0$.

$$P(|W_n - \theta| > \varepsilon) \rightarrow 0 \text{ à medida que } n \rightarrow \infty. \quad (\text{C.7})$$

Se W_n não for consistente para θ , dizemos que ele é **inconsistente**.

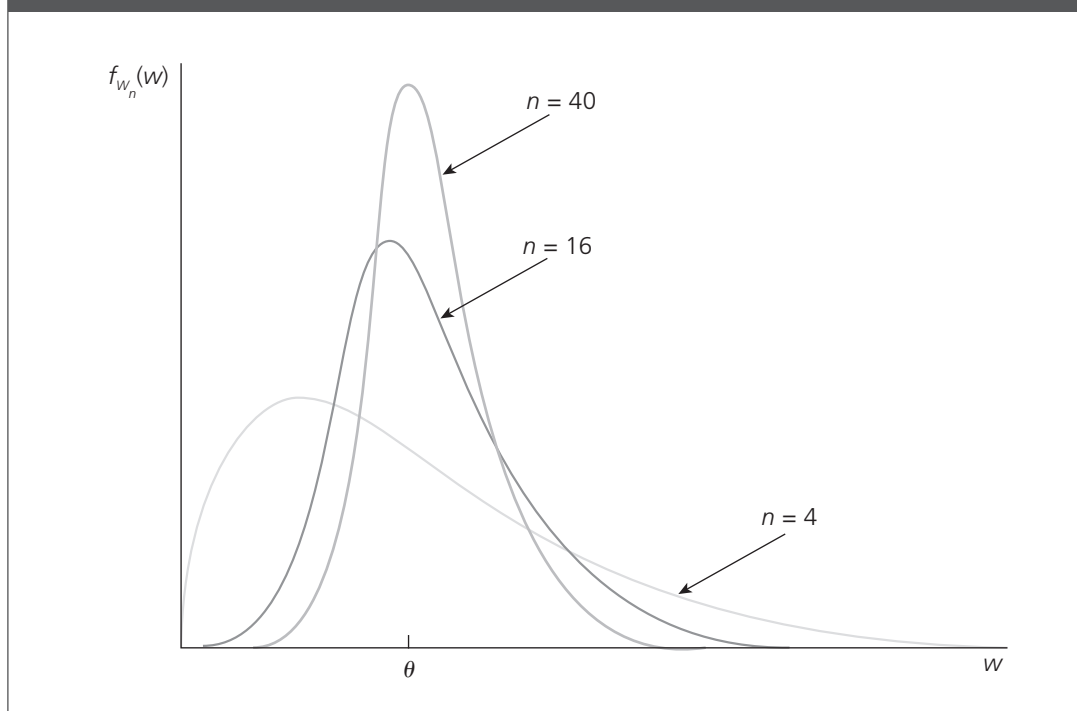
Quando W_n é consistente, também dizemos que θ é o **limite de probabilidade** de W_n , escrito como $\text{plim}(W_n) = \theta$.

Diferentemente da ausência de viés – que é característica de um estimador para determinado tamanho de amostra – a consistência envolve o comportamento da distribuição por amostragem do estimador à medida que o tamanho da amostra n aumenta. Para enfatizar isso, indexamos o estimador por tamanho da amostra ao expressar essa definição, e continuaremos com essa convenção ao longo da seção.

A equação (C.7) parece técnica e pode ser bem difícil de se estabelecer com base nos princípios fundamentais da probabilidade. Em contrapartida, interpretar (C.7) é simples. Ela quer dizer que a distribuição de W_n se torna cada vez mais concentrada ao redor de θ , o que, em termos gerais, significa que, para tamanhos maiores de amostras, será cada vez menos provável que W_n esteja muito longe de θ . Essa tendência foi ilustrada na Figura C.3.

Se um estimador não for consistente, ele não nos ajudará a aprender sobre θ , mesmo com uma quantidade ilimitada de dados. Por essa razão, a consistência é uma exigência mínima de um estimador usado em estatística ou econometria. Encontraremos estimadores que são consistentes sob determinadas hipóteses e inconsistentes quando elas fracassarem. Quando estimadores forem inconsistentes, geralmente

FIGURA C.3 Distribuições por amostragem de um estimador consistente para três tamanhos de amostra.



poderemos encontrar seus limites de probabilidade, e será importante saber quão distantes esses limites de probabilidade estão de θ .

Como notamos antes, estimadores não viesados não são necessariamente consistentes, mas aqueles cujas variâncias diminuem a zero quando o tamanho da amostra aumenta *são* consistentes. Isso pode ser expresso formalmente: Se W_n for um estimador não viesado de θ e $\text{Var}(W_n) \rightarrow 0$ à medida que $n \rightarrow \infty$, então $\text{plim}(W_n) = \theta$. Estimadores não viesados que usam a amostra de dados completa normalmente terão uma variância que diminui a zero à medida que a amostra aumenta, sendo, com isso, consistentes.

Um bom exemplo de estimador consistente é a média de uma extração amostral aleatória de uma população com média μ e variância σ^2 . Já mostramos que a média amostral é não viesada para μ . Na equação (C.6), derivamos $\text{Var}(\bar{Y}_n) = \sigma^2/n$ para qualquer tamanho de amostra n . Assim, $\text{Var}(\bar{Y}_n) \rightarrow 0$ à medida que $n \rightarrow \infty$, de forma que $\bar{Y}_n$ é um estimador consistente de μ (além de ser não viesado).

A conclusão de que $\bar{Y}_n$ é consistente para μ se mantém mesmo se $\text{Var}(\bar{Y}_n)$ não existir. Esse resultado clássico é conhecido como **lei dos grandes números (LGN)**.

Lei dos grandes números. Sejam $Y_1, Y_2, \dots, Y_n$ variáveis aleatórias independentes e identicamente distribuídas com média μ . Assim,

$$\text{plim}(\bar{Y}_n) = \mu. \quad (\text{C.8})$$

A lei dos grandes números significa que, se estivermos interessados em estimar a média populacional μ , poderemos chegar de maneira arbitrária perto de μ escolhendo uma amostra suficientemente grande. Esse resultado fundamental pode ser combinado com propriedades básicas de plims para mostrar que estimadores razoavelmente complicados são consistentes.

Propriedade PLIM.1: Determine θ como um parâmetro e defina um novo parâmetro, $\gamma = g(\theta)$, para uma *função contínua* $g(\theta)$. Suponha que $\text{plim}(W_n) = \theta$. Defina um estimador de γ por $G_n = g(W_n)$. Assim,

$$\text{plim}(G_n) = \gamma. \quad (\text{C.9})$$

Isso é frequentemente declarado como

$$\text{plim } g(W_n) = g(\text{plim } W_n) \quad (\text{C.10})$$

para uma função contínua $g(\theta)$.

A hipótese de que $g(\theta)$ é contínua é uma exigência técnica que em muitos casos foi descrita de forma não técnica como “uma função que pode ser grafada sem levantar seu lápis do papel”. Como todas as funções que encontrarmos neste texto são contínuas, não apresentaremos uma definição formal de função contínua. Exemplos de funções contínuas são $g(\theta) = a + b\theta$ para constantes a e b , $g(\theta) = \theta^2$, $g(\theta) = 1/\theta$, $g(\theta) = \sqrt{\theta}$, $g(\theta) = \exp(\theta)$ e muitas variações delas. Não precisaremos mencionar a hipótese de continuidade novamente.

Como um importante exemplo de estimador consistente, mas viesado, considere estimar o desvio padrão, σ , de uma população com média μ e variância σ^2 . Já afirmamos que a variância amostral $S_n^2 = (n-1)^{-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$ é não viesada para σ^2 .

Usando a lei dos grandes números e um pouco de álgebra, S_n^2 também pode ser mostrado como consistente para σ^2 . O estimador natural de $\sigma = \sqrt{\sigma^2}$ é $S_n = \sqrt{S_n^2}$ (onde a raiz quadrada é sempre a positiva). S_n , que é chamado de **desvio padrão amostral**, não é um estimador não viesado porque o valor esperado da raiz quadrada não é a raiz quadrada do valor esperado (ver Seção B.3). No entanto, segundo PLIM.1, $\text{plim } S_n = \sqrt{\text{plim } S_n^2} = \sqrt{\sigma^2} = \sigma$, assim, S_n é um estimador consistente de σ .

Aqui estão outras propriedades úteis do limite de probabilidade:

Propriedade PLIM.2: Se $\text{plim}(T_n) = \alpha$ e $\text{plim}(U_n) = \beta$, então

- (i) $\text{plim}(T_n + U_n) = \alpha + \beta$
- (ii) $\text{plim}(T_n U_n) = \alpha\beta$;
- (iii) $\text{plim}(T_n/U_n) = \alpha/\beta$, desde que $\beta \neq 0$.

Esses três fatos sobre limites de probabilidade nos permitem combinar estimadores consistentes em uma variedade de formas para obter outros estimadores consistentes. Por exemplo, defina $\{Y_1, \dots, Y_n\}$ como uma amostra aleatória de tamanho n sobre rendimentos anuais da população de trabalhadores com escolaridade de nível médio e indique a média populacional por μ_Z . Seja $\{Z_1, \dots, Z_n\}$ como a amostra aleatória sobre rendimentos anuais da população de trabalhadores com escolaridade de nível superior e indique a média populacional por μ_Y . Queremos estimar a diferença percentual nos rendimentos anuais entre os dois grupos, que é $\gamma = 100 \cdot (\mu_Z - \mu_Y)/\mu_Y$. (Essa é a porcentagem pela qual os rendimentos médios de graduados em universidades diferem dos rendimentos médios de graduados no ensino médio.) Como $\bar{Y}_n$ é consistente para μ_Y e $\bar{Z}$ é consistente para μ_Z , o resultado, a partir de PLIM.1 e da parte (iii) de PLIM.2, é que

$$G_n \equiv 100 \cdot (\bar{Z}_n - \bar{Y}_n)/\bar{Y}_n$$

é um estimador consistente de γ . G_n é apenas a diferença percentual entre $\bar{Z}_n$ e $\bar{Y}_n$ na amostra, por isso, é um estimador natural. G_n não é um estimador não viesado de γ , mas ainda assim é um bom estimador, exceto, possivelmente, quando n for pequeno.

C.3b Normalidade assintótica

A consistência é uma propriedade de estimadores pontuais. Embora ela nos diga que a distribuição de um estimador entre em colapso ao redor do parâmetro à medida que o tamanho da amostra aumenta, não nos diz basicamente nada sobre o *formato* dessa distribuição para determinado tamanho de amostra. Para construir estimadores de intervalo e testar hipóteses, precisamos de uma forma de aproximar a distribuição de nossos estimadores. A maioria dos estimadores econométricos tem distribuições que são bem aproximadas por uma distribuição normal para amostras grandes, o que motiva a definição a seguir.

Normalidade assintótica. Seja $\{Z_n: n = 1, 2, \dots\}$ uma sequência de variáveis aleatórias, dessa forma para todos os números z ,

$$P(Z_n \leq z) \rightarrow \Phi(z) \text{ à medida que } n \rightarrow \infty, \quad (\text{C.11})$$

onde $\Phi(z)$ é a função de distribuição normal padrão cumulativa. Assim, diz-se que Z_n tem uma *distribuição normal padrão assintótica*. Neste caso, geralmente

escrevemos $Z_n \overset{a}{\sim} \text{Normal}(0, 1)$. (O “a” acima do til significa “assimptoticamente” ou “aproximadamente”).

A propriedade (C.11) significa que a função de distribuição cumulativa para Z_n fica cada vez mais próxima da fdc da distribuição normal padrão à medida que o tamanho da amostra n aumenta. Quando a **normalidade assintótica** se mantém, para grandes n temos a aproximação $P(Z_n \leq z) \approx \Phi(z)$. Assim, as probabilidades que envolvem Z_n podem ser aproximadas por probabilidades normais padrão.

O **teorema do limite central (TLC)** é um dos resultados mais poderosos da probabilidade e da estatística. Ele afirma que a média de uma amostra aleatória para *qualquer* população (com variância finita), quando padronizada, tem uma distribuição normal padrão assintótica.

Teorema do limite central. Defina $\{Y_1, Y_2, \dots, Y_n\}$ como uma amostra aleatória com média μ e variância σ^2 . Assim,

$$Z_n = \frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}} \quad (\text{C.12})$$

tem uma distribuição normal padrão assintótica.

A variável Z_n em (C.12) é a versão padronizada de $\bar{Y}_n$: subtraímos $E(\bar{Y}_n) = \mu$ e dividimos por $\text{dp}(\bar{Y}_n) = \sigma/\sqrt{n}$. Assim, independentemente da distribuição populacional de Y , Z_n terá média zero e variância um, o que coincide com a média e a variância da distribuição normal padrão. O interessante é que toda a distribuição de Z_n se torna arbitrariamente próxima da distribuição normal à medida que n aumenta.

Podemos escrever a variável padronizada na equação (C.12) como $\sqrt{n}(\bar{Y}_n - \mu)/\sigma$, que mostra que devemos multiplicar a diferença entre a média amostral e a média populacional pela raiz quadrada do tamanho da amostra, com o objetivo de obter uma distribuição limitadora útil. Sem a multiplicação por $\sqrt{n}$, teríamos apenas $(\bar{Y}_n - \mu)/\sigma$, que converge em probabilidade a zero. Em outras palavras, a distribuição de $(\bar{Y}_n - \mu)/\sigma$ simplesmente entra em colapso para um único ponto à medida que $n \rightarrow \infty$, que sabemos não ser uma boa aproximação para a distribuição de $(\bar{Y}_n - \mu)/\sigma$ para tamanhos de amostra razoáveis. A multiplicação por $\sqrt{n}$ garante que a variância de Z_n permaneça constante. De maneira prática, muitas vezes tratamos $\bar{Y}_n$ como normalmente distribuída com média μ e variância σ^2/n , e isso nos dá os procedimentos estatísticos corretos porque nos leva à variável padronizada da equação (C.12).

A maioria dos estimadores encontrados em estatística e econometria pode ser escrita como funções de médias amostrais, caso em que podemos aplicar a lei dos grandes números e o teorema do limite central. Quando dois estimadores consistentes têm distribuições normais assintóticas, escolhemos o estimador com a menor variância assintótica.

Além da média amostral padronizada em (C.12), muitas outras estatísticas que dependem de médias amostrais acabam sendo assintoticamente normais. Uma importante é obtida substituindo σ por seu estimador consistente S_n na equação (C.12):

$$\frac{\bar{Y}_n - \mu}{S_n/\sqrt{n}} \quad (\text{C.13})$$

também tem uma distribuição normal padrão aproximada para n grande. As distribuições exatas (amostra finita) de (C.12) e (C.13) não são, definitivamente, as mesmas, mas a diferença muitas vezes é pequena o suficiente para ser ignorada com grandes n .

Ao longo desta seção, cada estimador foi subscrito por n para enfatizar a natureza da análise assintótica ou de grandes amostras. Continuar com essa convenção confunde a notação sem proporcionar uma percepção adicional, uma vez que os fundamentos da análise assintótica foram entendidos. Daqui em diante, eliminaremos o subscrito n e contaremos com você para se lembrar de que os estimadores dependem do tamanho da amostra, e que propriedades como consistência e normalidade assintótica normalmente se referem ao crescimento sem limites do tamanho da amostra.

C.4 Abordagens gerais para estimação de parâmetro

Até este ponto, usamos a média amostral para ilustrar propriedades de amostras finitas e grandes de estimadores. É natural perguntar: Existem abordagens gerais para estimação que produzem estimadores com boas propriedades, como ausência de viés, consistência e eficiência?

A resposta é sim. Uma explicação detalhada das várias abordagens para estimação está além do escopo deste livro; aqui, apresentamos apenas uma discussão informal. Uma discussão completa é oferecida em Larsen e Marx (1986, Capítulo 5).

C.4a Método dos momentos

Dado um parâmetro θ que surge em uma distribuição populacional, geralmente existem muitas formas de obter estimadores de θ não viesados e consistentes. Experimentar todas as diferentes possibilidades e compará-las com base nos critérios das Seções C.2 e C.3 não é prático. Felizmente, alguns métodos mostraram ter boas propriedades gerais e, em sua maioria, a lógica por detrás deles tem apelo intuitivo.

Nas seções anteriores, estudamos a média amostral como estimador não viesado da média populacional e a variância amostral como um estimador não viesado da variância populacional. Esses estimadores são exemplos de estimadores do **método dos momentos**. Geralmente, a estimação do método dos momentos ocorre da seguinte forma. Mostra-se que parâmetro θ é relacionado com algum valor esperado da distribuição de Y , normalmente $E(Y)$ ou $E(Y^2)$ (embora opções mais exóticas sejam usadas às vezes). Suponha, por exemplo, que o parâmetro de interesse, θ , esteja relacionado com a média populacional como $\theta = g(\mu)$ para uma função g . Como a média amostral $\bar{Y}$ é um estimador não viesado e consistente de μ , é natural substituir μ por $\bar{Y}$, o que nos dá o estimador $g(\bar{Y})$ de θ . O estimador $g(\bar{Y})$ é consistente para θ e, se $g(\mu)$ for uma função linear de μ , então $g(\bar{Y})$ também será não viesado. O que fizemos foi substituir o momento populacional, μ , por sua contraparte amostral, $\bar{Y}$. É daí que vem o nome “método dos momentos”.

Abordamos dois estimadores adicionais de método dos momentos que serão úteis para nossa discussão de análise de regressão. Lembre-se de que a covariância entre duas variáveis aleatórias X e Y é definida como $\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$. O método dos momentos sugere estimar σ_{XY} por $n^{-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$. Esse é um estimador consistente de σ_{XY} , mas acaba sendo viesado essencialmente pelo mesmo motivo que a variância amostral é viesada se n , e não $n - 1$, for usado como divisor. A **covariância amostral** é definida como

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}). \quad (\text{C.14})$$

Pode-se mostrar que esse é um estimador não viesado de σ_{XY} . (A substituição de n por $n - 1$ não faz diferença à medida que o tamanho da amostra cresce indefinidamente, portanto, este estimador ainda é consistente.)

Como discutimos na Seção B.4, a covariância entre duas variáveis é, muitas vezes, difícil de interpretar. Geralmente, estamos mais interessados na correlação. Já que a correlação populacional é $\rho_{XY} = \sigma_{XY}/(\sigma_X\sigma_Y)$, o método dos momentos sugere estimar ρ_{XY} como

$$R_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\left(\sum_{i=1}^n (X_i - \bar{X})^2\right)^{1/2} \left(\sum_{i=1}^n (Y_i - \bar{Y})^2\right)^{1/2}}, \quad (\text{C.15})$$

que é chamado de **coeficiente de correlação amostral** (ou *correlação amostral*, para abreviar). Note que cancelamos a divisão por $n - 1$ da covariância amostral e desvios padrão amostrais. Na verdade, podemos dividir cada uma delas por n e chegaremos à mesma fórmula final.

Podemos demonstrar que o coeficiente de correlação amostral está sempre no intervalo $[-1, 1]$, como deve ser. Como S_{XY} , S_X e S_Y são consistentes para o parâmetro populacional correspondente, R_{XY} é um estimador consistente da correlação populacional, ρ_{XY} . No entanto, R_{XY} é um estimador viesado por duas razões. Primeiro, S_X e S_Y são estimadores viesados de σ_X e σ_Y , respectivamente. Segundo, R_{XY} é uma relação de estimadores, por isso não deve ser não viesada, mesmo que S_X e S_Y sejam. Para nossos propósitos, isso não é importante, embora o fato de não existir nenhum estimador não viesado de ρ_{XY} seja um resultado clássico da estatística matemática.

C.4b Máxima verossimilhança

Outra abordagem geral para estimação é o método da *máxima verossimilhança*, um tópico abordado em muitos cursos introdutórios de estatística. Um breve resumo do caso mais simples será suficiente aqui. Defina $\{Y_1, Y_2, \dots, Y_n\}$ como uma amostra aleatória da distribuição populacional $f(y; \theta)$. Devido à hipótese de amostragem aleatória, a distribuição conjunta de $\{Y_1, Y_2, \dots, Y_n\}$ é simplesmente o produto das densidades: $f(y_1; \theta)f(y_2; \theta) \dots f(y_n; \theta)$. No caso discreto, ela é $P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n)$. Agora, defina a *função de verossimilhança* como

$$L(\theta; Y_1, \dots, Y_n) = f(Y_1; \theta)f(Y_2; \theta) \dots f(Y_n; \theta),$$

que é uma variável aleatória porque depende do resultado da amostra aleatória $\{Y_1, Y_2, \dots, Y_n\}$. O **estimador de verossimilhança máxima** de θ , chamado de W , é o valor de θ que maximiza a função de verossimilhança. (É por isso que escrevemos L como uma função de θ , seguida pela amostra aleatória.) É claro que esse valor depende da amostra aleatória. O princípio da máxima verossimilhança diz que, de todos os possíveis valores de θ , o valor que torna maior a verossimilhança dos dados observados deve ser escolhido. Intuitivamente, essa é uma abordagem razoável para estimar θ .

Em geral, é mais conveniente trabalhar com a *função de log-verossimilhança*, que é obtida calculando o log natural da função de verossimilhança:

$$\log[L(\theta; Y_1, \dots, Y_n)] = \sum_{i=1}^n \log[f(Y_i; \theta)], \quad (\text{C.16})$$

onde usamos o fato de que o log do produto é a soma dos logs. Como (C.16) é a soma de variáveis aleatórias independentes e identicamente distribuídas, a análise de estimadores que derivam de (C.16) é relativamente fácil.

A estimação de máxima verossimilhança (EMV) geralmente é consistente e às vezes é não viesada. Mas assim também são muitos outros estimadores. O forte apelo do EMV é que ele é, em geral, o estimador mais assintoticamente eficiente quando o modelo populacional $f(y; \theta)$ é corretamente especificado. Além disso, o EMV às vezes é o **estimador não viesado com variância mínima**; ou seja, ele tem a menor variância entre todos os estimadores não viesados de θ . [Ver Larsen e Marx (1986, Capítulo 5) para verificar essas afirmações.]

No Capítulo 17, precisaremos da verossimilhança máxima para estimar os parâmetros de modelos econométricos mais avançados. Na econometria, quase sempre estamos interessados na distribuição de Y condicional a um conjunto de variáveis explicativas, digamos, $X_1, X_2, \dots, X_n$. Assim, substituímos a densidade em (C.16) por $f(Y_i | X_{i1}, \dots, X_{in}; \theta_1, \dots, \theta_p)$, onde se permite que a densidade dependa de p parâmetros, $\theta_1, \dots, \theta_p$. Felizmente, para aplicações bem-sucedidas de métodos de verossimilhança máxima, não precisamos nos aprofundar muito nas questões computacionais ou na teoria estatística de grandes amostras. Wooldridge (2010, Capítulo 13) aborda a teoria de EMV.

C.4c Mínimos quadrados

Um terceiro tipo de estimador, e aquele que tem o papel principal ao longo do texto, é chamado de **estimador de mínimos quadrados**. Já vimos um exemplo de mínimos quadrados: a média amostral, $\bar{Y}$, é um estimador de mínimos quadrados da média populacional, μ . Já sabemos que $\bar{Y}$ é um estimador do método de momentos. O que o torna um estimador de mínimos quadrados? Pode-se demonstrar que o valor de m que torna a soma dos desvios quadrados

$$\sum_{i=1}^n (Y_i - m)^2$$

A menor possível é $m = \bar{Y}$. Mostrar isso não é difícil, mas omitiremos a álgebra.

Para algumas distribuições importantes, incluindo normal e de Bernoulli, a média amostral $\bar{Y}$ também é o estimador de verossimilhança máxima da média populacional μ . Assim, os princípios dos mínimos quadrados, método dos momentos e verossimilhança máxima geralmente resultam no *mesmo* estimador. Em outros casos, os estimadores são similares, mas não idênticos.

C.5 Estimação por intervalo e intervalos de confiança

C.5a Natureza da estimação por intervalo

Uma estimativa pontual obtida a partir de determinada amostra não apresenta, por si só, informações suficientes para testar teorias econômicas ou para informar discussões

de políticas públicas. Uma estimativa pontual pode ser a melhor aposta do pesquisador em relação ao valor da população, mas, por sua natureza, ela não traz informações sobre quão próxima essa estimativa é “propensa” a estar do parâmetro populacional. Como exemplo, suponha que um pesquisador registre, com base em uma amostra aleatória de trabalhadores, que subsídios para treinamentos corporativos aumentam o salário por hora em 6,4%. Como poderíamos saber se isso está ou não próximo do efeito sobre a população de trabalhadores que poderiam ser treinados? Como não sabemos o valor populacional, não conseguimos saber quão próxima uma estimativa é para determinada amostra. No entanto, podemos fazer afirmações matemáticas envolvendo probabilidade, e é aí que entra a estimação por intervalo.

Já conhecemos uma forma de avaliar a incerteza de um estimador: encontrando seu desvio padrão por amostragem. O registro do desvio padrão do estimador, ao lado da estimativa pontual, traz informações sobre a acurácia de nossa estimativa. Entretanto, mesmo que o problema da dependência de parâmetros populacionais desconhecidos do desvio padrão seja ignorado, registrar o desvio padrão ao lado da estimativa pontual não faz nenhuma afirmação direta sobre onde o valor populacional provavelmente estaria em relação à estimativa. Essa limitação é superada com a construção de um **intervalo de confiança**.

Ilustramos o conceito de intervalo de confiança com um exemplo. Suponha que a população tenha uma distribuição Normal(μ , 1) e defina $\{Y_1, \dots, Y_n\}$ como uma amostra aleatória dessa população. (Assumimos que a variância da população é conhecida e igual a um para efeito de ilustração; mostramos, então, o que fazer no caso mais realista em que a variância é desconhecida.) A média amostral, $\bar{Y}$, tem uma distribuição normal com média μ e variância $1/n$: $\bar{Y} \sim \text{Normal}(\mu, 1/n)$. A partir disso, podemos padronizar $\bar{Y}$ e, como a versão padronizada de $\bar{Y}$ tem uma distribuição normal padrão, temos

$$P\left(-1,96 < \frac{\bar{Y} - \mu}{1/\sqrt{n}} < 1,96\right) = 0,95.$$

O evento entre parêntesis é idêntico ao evento $\bar{Y} - 1,96/\sqrt{n} < \mu < \bar{Y} + 1,96/\sqrt{n}$, portanto,

$$P(\bar{Y} - 1,96/\sqrt{n} < \mu < \bar{Y} + 1,96/\sqrt{n}) = 0,95. \quad (\text{C.17})$$

A equação (C.17) é interessante porque ela nos diz que a probabilidade de o intervalo aleatório $[\bar{Y} - 1,96/\sqrt{n}, \bar{Y} + 1,96/\sqrt{n}]$ conter a média populacional μ é 0,95 ou 95%. Essa informação nos permite construir uma *estimativa de intervalo* de μ , que é obtida ao adicionar o resultado amostral da média, $\bar{y}$. Assim,

$$[\bar{y} - 1,96/\sqrt{n}, \bar{y} + 1,96/\sqrt{n}] \quad (\text{C.18})$$

é um exemplo de estimativa de intervalo de μ . Ele também é chamado de intervalo de confiança de 95%. Uma notação abreviada para este intervalo é $\bar{y} \pm 1,96/\sqrt{n}$.

O intervalo de confiança da equação (C.18) é fácil de calcular, uma vez que os dados amostrais $\{y_1, y_2, \dots, y_n\}$ tenham sido observados; $\bar{y}$ é o único fator que depende dos dados. Por exemplo, suponha que $n = 16$ e que a média dos 16 pontos de dados seja 7,3. Então, o intervalo de confiança de 95% para μ é $7,3 \pm 1,96/\sqrt{16} = 7,3 \pm 0,49$, que podemos escrever em forma de intervalo como $[6,81, 7,79]$. Por construção, $\bar{y} = 7,3$ está no centro deste intervalo.

Diferentemente de seu cálculo, o significado de um intervalo de confiança é mais difícil de compreender. Quando dizemos que a equação (C.18) é um intervalo de confiança de 95% para μ , queremos dizer que o intervalo *aleatório*

$$[\bar{Y} - 1,96/\sqrt{n}, \bar{Y} + 1,96/\sqrt{n}] \quad (\text{C.19})$$

contém μ com probabilidade 0,95. Em outras palavras, *antes* de a amostra aleatória ser traçada, há 95% de chance de que (C.19) contenha μ . A equação (C.19) é um exemplo de **estimador de intervalo**. É um intervalo aleatório, já que suas extremidades mudam com diferentes amostras.

Um intervalo de confiança muitas vezes é interpretado da seguinte forma: “A probabilidade de μ estar no intervalo (C.18) é 0,95”. Isso é incorreto. Uma vez que a amostra tenha sido observada e $\bar{y}$ calculado, os limites do intervalo de confiança são simplesmente números (6,81 e 7,79 no exemplo dado há pouco). O parâmetro populacional, μ , apesar de desconhecido, também é apenas um número. Assim, μ pode estar ou não no intervalo (C.18) (e nunca saberemos com certeza qual é o caso). A probabilidade não tem papel nenhum, já que o intervalo de confiança é calculado em particular para os dados existentes. The probabilistic interpretation comes from the fact that for 95% of all random samples, the constructed confidence interval will contain μ .

Para enfatizar o significado de um intervalo de confiança, a Tabela C.2 contém cálculos para 20 amostras aleatórias (ou réplicas) da distribuição Normal(2, 1) com tamanho de amostra $n = 10$. Para cada uma das 20 amostras, $\bar{y}$ foi obtida, e (C.18) é

TABELA C.2 Intervalos de confiança simulados para uma distribuição Normal ($\mu, 1$) com $\mu = 2$

Réplica	$\bar{y}$	Intervalo de 95%	Contém μ ?
1	1,98	(1,36,2,60)	Sim
2	1,43	(0,81,2,05)	Sim
3	1,65	(1,03,2,27)	Sim
4	1,88	(1,26,2,50)	Sim
5	2,34	(1,72,2,96)	Sim
6	2,58	(1,96,3,20)	Sim
7	1,58	(0,96,2,20)	Sim
8	2,23	(1,61,2,85)	Sim
9	1,96	(1,34,2,58)	Sim
10	2,11	(1,49,2,73)	Sim
11	2,15	(1,53,2,77)	Sim
12	1,93	(1,31,2,55)	Sim
13	2,02	(1,40,2,64)	Sim
14	2,10	(1,48,2,72)	Sim
15	2,18	(1,56,2,80)	Sim
16	2,10	(1,48,2,72)	Sim
17	1,94	(1,32,2,56)	Sim
18	2,21	(1,59,2,83)	Sim
19	1,16	(0,54,1,78)	Não
20	1,75	(1,13,2,37)	Sim

calculada como $\bar{y} \pm 1,96/\sqrt{10} = \bar{y} \pm 0,62$ (todas arredondadas para duas casas decimais). Como você pode ver, o intervalo muda com cada amostra aleatória. Dezenove dos vinte intervalos contêm o valor populacional de μ . Somente na réplica número 19, μ não está no intervalo de confiança. Em outras palavras, 95% das amostras resultam em um intervalo de confiança que contém μ . Esse não tem de ser o caso com somente 20 réplicas, mas funcionou dessa forma nesta simulação.

C.5b Intervalos de confiança para a média de uma população normalmente distribuída

O intervalo de confiança derivado da equação (C.18) ajuda a ilustrar como construir e interpretar intervalos de confiança. Na prática, a equação (C.18) não é muito útil para a média de uma população normal porque ela supõe que a variância seja sabidamente unitária. É fácil estender (C.18) para o caso em que sabemos que o desvio padrão pode ser qualquer valor: o intervalo de confiança de 95% será

$$[\bar{y} - 1,96\sigma/\sqrt{n}, \bar{y} + 1,96\sigma/\sqrt{n}]. \quad (\text{C.20})$$

Portanto, desde que σ seja conhecido, um intervalo de confiança para μ será prontamente construído. Para permitir σ desconhecido, precisaremos usar uma estimativa. Defina

$$s = \left(\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \right)^{1/2} \quad (\text{C.21})$$

como o desvio padrão amostral. Em seguida, obtemos um intervalo de confiança que depende totalmente dos dados observados substituindo σ na equação (C.20) pela estimativa, s . Infelizmente, isso não preserva o nível de 95% de confiança porque s depende da amostra em particular. Em outras palavras, o intervalo aleatório $[\bar{Y} \pm 1,96(S/\sqrt{n})]$ não contém mais μ com probabilidade 0,95 porque a constante σ foi substituída pela variável aleatória S .

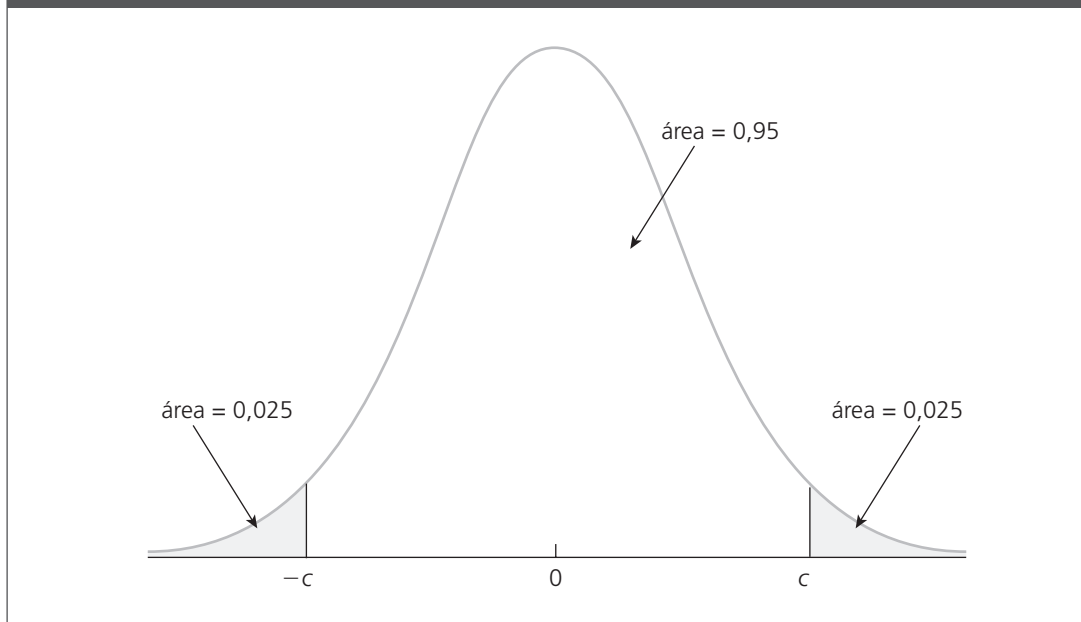
Como devemos agir? Em vez de usar a distribuição normal padrão, devemos contar com a distribuição t . A distribuição t surge do fato de que

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}, \quad (\text{C.22})$$

onde $\bar{Y}$ é a média amostral e S é o desvio padrão amostral da amostra aleatória $\{Y_1, \dots, Y_n\}$. Não vamos provar (C.22); uma prova cuidadosa pode ser encontrada em uma variedade de lugares [por exemplo, Larsen e Marx (1986, Capítulo 7)].

Para construir um intervalo de confiança de 95%, permita que c indique o 97,5° percentil da distribuição t_{n-1} . Em outras palavras, c é o valor para o qual 95% da área de t_{n-1} estará entre $-c$ e c : $P(-c < t_{n-1} < c) = 0,95$. (O valor de c depende dos graus de liberdade $n-1$, mas não deixaremos isso explícito.) A escolha de c foi ilustrada na Figura C.4. Uma vez que c tenha sido escolhido adequadamente, o intervalo aleatório $[\bar{Y} - c \cdot S/\sqrt{n}, \bar{Y} + c \cdot S/\sqrt{n}]$ conterá μ com probabilidade 0,95. Para uma amostra em particular, o intervalo de confiança de 95% é calculado como

$$[\bar{y} - c \cdot s/\sqrt{n}, \bar{y} + c \cdot s/\sqrt{n}]. \quad (\text{C.23})$$

FIGURA C.4 O 97,5º percentil, c , em uma distribuição t .

Os valores de c para vários graus de liberdade podem ser obtidos da Tabela G.2 do Apêndice G. Por exemplo, se $n = 20$, de modo que os gl sejam $n - 1 = 19$, então $c = 2,093$. Assim, o intervalo de confiança de 95% é $[\bar{y} \pm 2,093(s/\sqrt{20})]$, onde $\bar{y}$ e s são os valores obtidos a partir da amostra. Mesmo que $s = \sigma$ (o que é muito improvável), o intervalo de confiança de (C.23) será mais amplo do que o de (C.20) porque $c > 1,96$. Para graus de liberdade pequenos, (C.23) é muito mais ampla.

De forma mais geral, defina c_α como o percentil $100(1 - \alpha)$ da distribuição t_{n-1} . Assim, um intervalo de confiança de $100(1 - \alpha)\%$ é obtido por

$$[\bar{y} - C_{\alpha/2}S/\sqrt{n}, \bar{y} + C_{\alpha/2}S/\sqrt{n}]. \quad (\text{C.24})$$

Obter $c_{\alpha/2}$ exige que você escolha α e conheça os graus de liberdade $n - 1$; assim, a Tabela G.2 pode ser usada. Na maior parte, nos concentraremos nos intervalos de confiança de 95%.

Existe uma forma simples de se lembrar como construir um intervalo de confiança para a média de uma distribuição normal. Lembre-se de que $\text{dp}(\bar{Y}) = \sigma/\sqrt{n}$. Assim, $s/\sqrt{n}$ é a estimativa pontual de $\text{dp}(\bar{Y})$. A variável aleatória associada, $S/\sqrt{n}$, às vezes é chamada de **erro padrão** de $\bar{Y}$. Como o que aparece nas fórmulas é a estimativa pontual $s/\sqrt{n}$, definimos o erro padrão de $\bar{y}$ como $\text{ep}(\bar{y}) = s/\sqrt{n}$. Dessa forma, (C.24) pode ser escrita de forma abreviada como

$$[\bar{y} \pm c_{\alpha/2} \cdot \text{ep}(\bar{y})]. \quad (\text{C.25})$$

Essa equação mostra por que a noção de erro padrão de uma estimativa tem papel importante na econometria.

EXEMPLO C.2**Efeito de subsídios de treinamento corporativo sobre a produtividade do trabalhador**

Holzer, Block, Cheatham e Knott (1993) estudaram os efeitos de subsídios de treinamentos corporativos sobre a produtividade dos trabalhadores por meio da coleta de informações sobre “taxas de refugo” de uma amostra de empresas manufatureiras de Michigan que receberam subsídios para treinamentos corporativos em 1988. A Tabela C.3 lista as taxas de refugo – medidas como o número de itens, a cada 100 produzidos, que não são utilizáveis e, portanto, precisam ser descartados – de 20 empresas. Cada uma dessas empresas recebeu um subsídio para treinamento profissional em 1988; não houve subsídios em 1987. Estamos interessados em construir um intervalo de confiança para a mudança na taxa de refugo de 1987 a 1988 da população de todas as empresas manufatureiras que podem ter recebido subsídios.

Partimos do pressuposto de que a mudança nas taxas de refugo tem uma distribuição normal. Já que $n = 20$, um intervalo de confiança de 95% para a alteração média das taxas de refugo μ é $[\bar{y} \pm 2,093 \cdot \text{ep}(\bar{y})]$, onde $\text{ep}(\bar{y}) = s/\sqrt{n}$. O valor 2,093 é o 97,5º percentil em uma distribuição t_{19} . Para os valores da amostra, em especial, $\bar{y} = -1,15$ e $\text{ep}(\bar{y}) = 0,54$ (arredondados para duas casas decimais), portanto, o intervalo de confiança de 95% é $[-2,28, -0,02]$. O valor zero é excluído deste intervalo; por isso, concluímos que, com confiança de 95%, a alteração média das taxas de refugo na população não é zero.

TABELA C.3 Taxas de refugo de 20 empresas manufatureiras de Michigan

Empresa	1987	1988	Alteração
1	10	3	-7
2	1	1	0
3	6	5	-1
4	0,45	0,5	0,05
5	1,25	1,54	0,29
6	1,3	1,5	0,2
7	1,06	0,8	-0,26
8	3	2	-1
9	8,18	0,67	-7,51
10	1,67	1,17	-0,5
11	0,98	0,51	-0,47
12	1	0,5	-0,5
13	0,45	0,61	0,16
14	5,03	6,7	1,67
15	8	4	-4
16	9	7	-2
17	18	19	1
18	0,28	0,2	-0,08
19	7	5	-2
20	3,97	3,83	-0,14
Média	4,38	3,23	-1,15

Neste ponto, o Exemplo C.2 é bem ilustrativo porque tem falhas potencialmente sérias como análise econométrica. Mais importante do que isso, ele supõe que qualquer redução sistemática das taxas de refugo ocorra devido aos subsídios de capacitação profissional. Porém, muitas coisas podem acontecer no curso de um ano para mudar a produtividade de um trabalhador. A partir dessa análise, não temos como saber se a queda nas taxas médias de refugo é atribuível aos subsídios de treinamento profissional ou se, pelo menos parcialmente, alguma força externa foi responsável.

C.5c Uma regra de ouro simples para um intervalo de confiança de 95%

O intervalo de confiança em (C.25) pode ser calculado para qualquer tamanho de amostra e qualquer nível de confiança. Como vimos na Seção B.5, a distribuição t se aproxima da distribuição normal à medida que os graus de liberdade aumentam. Em especial, para $\alpha = 0,05$, $c_{\alpha/2} \rightarrow 1,96$ à medida que $n \rightarrow \infty$, embora $c_{\alpha/2}$ seja sempre maior que 1,96 para todo n . Uma *regra de ouro* para um intervalo de confiança de 95% aproximado é

$$[\bar{y} \pm 2 \cdot \text{ep}(\bar{y})]. \quad (\text{C.26})$$

Em outras palavras, obtemos $\bar{y}$ e seu erro padrão, em seguida, calculamos $\bar{y}$ mais ou menos duas vezes seu erro padrão para chegar ao intervalo de confiança. Isso é um pouco amplo demais para um n muito grande, e estreito demais para um n pequeno. Como podemos ver no Exemplo C.2, mesmo para um n pequeno como 20, (C.26) é uma aproximação do intervalo de confiança de 95% para a média de uma distribuição normal. Isso quer dizer que conseguimos chegar bem perto de um intervalo de confiança de 95% sem termos de consultar tabelas t .

C.5d Intervalos de confiança assintóticos para populações não normais

Em algumas aplicações, a população claramente é não normal. Um dos principais casos é a distribuição de Bernoulli, em que as variáveis aleatórias assumem apenas os valores zero e um. Em outros casos, a população não normal não tem distribuição padrão. Isso não importa, desde que o tamanho da amostra seja suficientemente grande para que o teorema do limite central dê uma boa aproximação para a distribuição da média amostral $\bar{Y}$. Para um n grande, um intervalo de confiança de 95% *aproximado* é

$$[\bar{y} \pm 1,96 \cdot \text{ep}(\bar{y})], \quad (\text{C.27})$$

onde o valor 1,96 é o 97,5º percentil na distribuição normal padrão. De forma mecânica, calcular um intervalo de confiança aproximado não difere do caso normal. Uma pequena diferença é que o número multiplicado pelo erro padrão vem da distribuição normal padrão, em vez da distribuição t , porque estamos usando assintóticos. Como a distribuição t se aproxima da distribuição normal à medida que os gl aumentam, a equação (C.25) também é perfeitamente legítima como um intervalo de 95% aproximado; alguns preferem essa equação à (C.27) porque a anterior é exata para populações normais.

EXEMPLO C.3 Discriminação racial em contratações

O Urban Institute conduziu um estudo em Washington, D.C., em 1988, para examinar a extensão da discriminação racial nas contratações. Cinco duplas de pessoas foram entrevistadas para vários cargos. Em cada dupla, uma pessoa era negra e a outra era branca. Foram entregues currículos que indicavam que eles eram praticamente iguais em termos de experiência, escolaridade e outros fatores que determinam qualificação profissional. A ideia era tornar os indivíduos o mais similares possível, com exceção da raça. Cada pessoa da dupla foi entrevistada para a mesma vaga e os pesquisadores registraram qual candidato recebeu a proposta de emprego. Esse é um exemplo de *análise de pares combinados*, em que cada teste consiste em dados de duas pessoas (ou empresas, cidades, e assim por diante) consideradas similares em muitos aspectos, mas diferentes em uma característica importante.

Seja θ_N a probabilidade de o negro ser contratado e θ_B a probabilidade de o branco ser contratado. Estamos interessados principalmente na diferença, $\theta_N - \theta_B$. Defina N_i como uma variável de Bernoulli igual a um se a pessoa negra receber uma oferta do empregador i , e zero no caso contrário. De forma similar, $B_i = 1$ se a pessoa branca receber uma oferta do empregador i e zero no caso contrário. Combinando as cinco duplas de pessoas, houve um total de $n = 241$ testes (pares de entrevistas com empregadores). Estimadores não viesados de θ_N e θ_B são $\bar{N}$ e $\bar{B}$, as frações de entrevistas para as quais negros e brancos receberam ofertas de emprego, respectivamente.

Para colocar esses dados em uma estrutura de cálculo de um intervalo de confiança para a média populacional, defina uma nova variável $Y_i = N_i - B_i$. Agora, Y_i pode assumir três valores: -1 se o negro não conseguiu o trabalho, mas o branco sim; 0 se ambos conseguiram ou não o emprego; e 1 se o negro conseguiu o trabalho e o branco não. Assim, $\mu \equiv E(Y_i) = E(N_i) - E(B_i) = \theta_N - \theta_B$.

A distribuição de Y_i certamente não é normal – ela é discreta e assume apenas três valores. No entanto, um intervalo de confiança aproximado para $\theta_N - \theta_B$ pode ser obtido usando métodos de grandes amostras.

Os dados do estudo do Urban Institute estão no arquivo AUDIT. Usando os 241 pontos de dados observados, $\bar{N} = 0,224$ e $\bar{B} = 0,357$, assim, $\bar{Y} = 0,224 - 0,357 = -0,133$. Consequentemente, 22,4% dos candidatos negros tiveram cargos oferecidos, enquanto 35,7% dos brancos receberam as propostas. Isso é uma prova *prima facie* de discriminação contra negros, mas podemos descobrir muito mais calculando um intervalo de confiança para μ . Para calcular um intervalo de confiança aproximado de 95%, precisamos do desvio padrão amostral. O resultado é $s = 0,482$ [usando a equação (C.21)]. Usando (C.27), obtemos um IC de 95% para $\mu = \theta_N - \theta_B$ como $-0,133 \pm 1,96(0,482/\sqrt{241}) = -0,133 \pm 0,031 = [-0,164, -0,102]$. O IC aproximado de 99% é $-0,133 \pm 2,58(0,482/\sqrt{241}) = [-0,213, -0,053]$. Naturalmente, ele contém uma amplitude maior de valores do que o IC de 95%. Porém, mesmo o IC de 99% não contém o valor zero. Assim, estamos muito confiantes de que a diferença populacional $\theta_N - \theta_B$ não é zero.

Antes de nos voltarmos ao teste de hipóteses, é útil revisar as várias quantidades de população e amostra que medem a dispersão das distribuições populacionais e distribuições por amostragem dos estimadores. Essas quantidades aparecem

frequentemente nas análises estatísticas, e extensões delas são importantes para a análise de regressão do texto principal. A quantidade σ é o desvio padrão populacional (desconhecido); é uma medida da dispersão da distribuição de Y . Quando dividimos σ por $\sqrt{n}$, obtemos o **desvio padrão por amostragem** de $\bar{Y}$ (a média amostral). Embora σ seja uma característica fixa da população $\text{dp}(\bar{Y}) = \sigma/\sqrt{n}$ diminui a zero à medida que $n \rightarrow \infty$: nosso estimador de μ fica cada vez mais preciso à medida que o tamanho da amostra aumenta.

A estimativa de σ para determinada amostra, s , é chamada de desvio padrão amostral porque é obtido a partir da amostra. (Também chamamos a variável aleatória subjacente, S , que muda ao longo de diferentes amostras, de desvio padrão amostral.) Da mesma forma que $\bar{y}$ como uma estimativa de μ , s é nossa “melhor aposta” de σ dada a amostra à mão. A quantidade $s/\sqrt{n}$ é o que chamamos de erro padrão de $\bar{y}$, e é nossa melhor estimativa de $\sigma/\sqrt{n}$. Intervalos de confiança para o parâmetro populacional μ dependem diretamente de $\text{ep}(\bar{y}) = s/\sqrt{n}$. Como este erro padrão diminui a zero à medida que o tamanho da amostra aumenta, uma amostra maior geralmente significa um intervalo de confiança menor. Assim, vemos claramente que um benefício de ter mais dados é que eles resultam em intervalos de confiança mais estreitos. A noção de erro padrão de uma estimativa, que, na grande maioria dos casos, diminui a zero em uma taxa $1/\sqrt{n}$, tem papel fundamental no teste de hipóteses (como veremos na próxima seção) e em intervalos de confiança e testes no contexto de regressão múltipla (como discutido no Capítulo 4).

C.6 Teste de hipóteses

Até aqui, revisamos como avaliar estimadores pontuais e vimos – no caso de uma média populacional – como construir e interpretar intervalos de confiança. Porém, às vezes, a questão na qual estamos interessados tem uma resposta definitiva sim ou não. Estes são alguns exemplos: (1) Um programa de capacitação profissional aumenta a produtividade média do trabalhador? (ver Exemplo C.2); (2) Negros são discriminados no momento da contratação? (ver Exemplo C.3); (3) As leis mais rígidas contra direção sob efeito de álcool reduzem o número de prisões por dirigir embriagado? Um método elaborado para responder a essas questões, usando uma amostra de dados, é conhecido como teste de hipóteses.

C.6a Fundamentos do teste de hipóteses

Para ilustrar as questões envolvidas no teste de hipóteses, considere um exemplo de eleição. Suponha que haja dois candidatos em uma eleição, Candidatos A e B. Foi registrado que o Candidato A recebeu 42% dos votos populares, enquanto o Candidato B recebeu 58%. Supõe-se que esses dados representem as porcentagens reais da população eleitora, e os trataremos como tal.

O Candidato A está convencido de que mais pessoas devem ter votado nele, por isso gostaria de investigar se a eleição foi manipulada. Conhecendo um pouco de estatística, o Candidato A contrata uma agência consultora para reunir uma amostra aleatória de 100 eleitores para registrar se essas pessoas votaram ou não nele. Suponha que, na amostra coletada, 53 pessoas votaram no Candidato A. Essa estimativa amostral de 53% claramente excede o valor populacional registrado de 42%. O Candidato A deve concluir que a eleição foi, de fato, uma fraude?

Embora pareça que os votos no Candidato A foram subestimados, não podemos ter certeza. Mesmo que só 42% da população tenha votado no Candidato A, é possível que, em uma amostra de 100, observemos 53 pessoas que votaram nele. A questão é: Quão forte é a evidência amostral em relação à porcentagem oficial de 42%?

Uma forma de prosseguir é definindo um **teste de hipótese**. Defina θ como a proporção real da população que vota no Candidato A. A hipótese de que os resultados registrados são precisos pode ser declarada como

$$H_0: \theta = 0,42 \quad (\text{C.28})$$

Este é um exemplo de **hipótese nula**. Sempre simbolizamos a hipótese nula com H_0 . No teste de hipóteses, a hipótese nula tem um papel parecido com o de um réu em um tribunal em muitos sistemas judiciais: assim como se pressupõe que o réu seja inocente até que se prove o contrário, a hipótese nula é considerada verdadeira até que dados sugiram fortemente o contrário. No exemplo atual, o Candidato A precisa apresentar evidências bem fortes contra (C.28) para ganhar uma recontagem.

A **hipótese alternativa** no exemplo da eleição é que a proporção real que votou no Candidato A na eleição é maior do que 0,42:

$$H_1: \theta > 0,42. \quad (\text{C.29})$$

Para concluir que H_0 é falsa e H_1 é verdadeira, precisamos ter evidências “além da dúvida razoável” contra H_0 . Quantos votos dos 100 seriam necessários antes de sentirmos que as evidências estão fortemente contra H_0 ? A maioria concordaria que observar 43 votos em uma amostra de 100 não é suficiente para derrubar o resultado original da eleição; este resultado está dentro da variação amostral esperada. Por outro lado, não precisamos observar 100 votos no Candidato A para lançar dúvidas sobre H_0 . Saber se 53 de 100 é o bastante para rejeitar H_0 é muito menos claro. A resposta depende de como quantificamos “além da dúvida razoável”.

Antes de nos voltarmos à questão da quantificação da incerteza no teste de hipóteses, precisamos impedir algumas possíveis confusões. Você deve ter notado que as hipóteses das equações (C.28) e (C.29) não esgotam todas as possibilidades: pode ser que θ seja menor do que 0,42. Para a aplicação que temos, não estamos especialmente interessados nessa possibilidade; ela não tem nada a ver com a anulação dos resultados da eleição. Portanto, podemos declarar a princípio que estamos ignorando alternativas θ com $\theta < 0,42$. No entanto, alguns autores preferem declarar hipóteses nula e alternativa até que elas estejam esgotadas, caso em que nossa hipótese nula deveria ser $H_0: \theta \leq 0,42$. Expressa dessa forma, a hipótese nula é uma hipótese nula *composta* porque permite que haja mais de um valor sob H_0 . [Em contrapartida, a equação (C.28) é um exemplo de hipótese nula *simples*.] Para estes tipos de exemplo, não importa se declaramos a hipótese nula como em (C.28) ou como uma nula composta: o valor mais difícil de rejeitar se $\theta \leq 0,42$ é $\theta = 0,42$. (Ou seja, se rejeitarmos o valor $\theta = 0,42$ contra $\theta > 0,42$, logicamente precisaremos rejeitar qualquer valor menor do que 0,42.) Portanto, nosso procedimento de testes baseado em (C.28) leva ao mesmo teste que teríamos se $H_0: \theta \leq 0,42$. Neste texto, sempre declaramos a hipótese nula como uma hipótese nula simples.

No teste de hipóteses, podemos cometer dois tipos de erros. Primeiro, podemos rejeitar a hipótese nula quando ela for, de fato, verdadeira. Isso é chamado de **erro Tipo I**. No exemplo da eleição, um erro Tipo I ocorre se rejeitarmos H_0 quando a

proporção real de pessoas que votaram no Candidato A for, na verdade, 0,42. O segundo tipo de erro é não conseguir rejeitar H_0 quando ela for realmente falsa. Isso é chamado de **erro Tipo II**. No exemplo da eleição, um erro Tipo II ocorre se $\theta > 0,42$, mas fracassarmos ao rejeitar H_0 .

Depois de tomarmos a decisão de rejeitar ou não a hipótese nula, teremos decidido corretamente ou cometido um erro. Nunca saberemos ao certo se um erro foi cometido. Entretanto, podemos calcular a *probabilidade* de cometermos um erro Tipo I ou Tipo II. As regras de teste de hipótese são construídas para tornar a probabilidade de cometer um erro Tipo I bem pequena. Geralmente, definimos o **nível de significância** (ou simplesmente *nível*) de um teste como a probabilidade de um erro Tipo I; ele é tipicamente indicado por α . Simbolicamente, temos

$$\alpha = P(\text{Rejeitar } H_0 | H_0). \quad (\text{C.30})$$

O lado direito é lido como: “A probabilidade de rejeitar H_0 sendo H_0 verdadeira”.

O teste de hipóteses clássico exige que especifiquemos inicialmente um nível de significância para um teste. Quando especificamos um valor para α , estamos basicamente quantificando nossa tolerância para um erro Tipo I. Valores comuns de α são 0,10, 0,05 e 0,01. Se $\alpha = 0,05$, então o pesquisador está propenso a rejeitar falsamente H_0 5% do tempo, com o objetivo de detectar desvios de H_0 .

Uma vez escolhido o nível de significância, desejaríamos minimizar a probabilidade de um erro Tipo II. Como alternativa, gostaríamos de maximizar o **poder de um teste** contra todas as alternativas relevantes. O poder de um teste é simplesmente um menos a probabilidade de um erro Tipo II. Matematicamente,

$$\pi(\theta) = P(\text{Rejeitar } H_0 | \theta) = 1 - P(\text{Tipo II} | \theta),$$

onde θ simboliza o valor real do parâmetro. Naturalmente, o desejo é que o poder fosse igual a um sempre que a hipótese nula fosse falsa. Porém, isso é impossível de se obter enquanto mantivermos o nível de significância pequeno. Em vez disso, escolhemos nossos testes para maximizar o poder para determinado nível de significância.

C.6b Testando hipóteses sobre a média em uma população normal

Com o objetivo de testar uma hipótese nula contra uma alternativa, precisamos escolher uma estatística teste (ou estatística, para abreviar) e um valor crítico. As opções de estatística e valor crítico são baseadas em conveniência e no desejo de maximizar o poder dado um nível de significância para o teste. Nesta subseção, revisaremos como testar hipóteses para a média de uma população normal.

Uma **estatística teste**, chamada de T , é uma função da amostra aleatória. Quando calculamos a estatística de determinado resultado, obtemos um resultado da estatística teste, que indicaremos por t .

Dada uma estatística teste, podemos definir uma regra de rejeição que determine quando H_0 é rejeitada em favor de H_1 . Neste texto, todas as regras de rejeição são baseadas na comparação do valor de uma estatística teste, t , com um **valor crítico**, c . Os valores de t que resultam em rejeição da hipótese nula são conjuntamente conhecidos como **região de rejeição**. Para determinar o valor crítico, precisamos decidir primeiro um nível de significância para o teste. Em seguida, dado α , o valor crítico associado a α é determinado pela distribuição de T , supondo que H_0 seja verdadeira. Escreveremos esse valor crítico como c , suprimindo o fato de que ele depende de α .

O teste de hipóteses sobre a média μ de uma população Normal(μ, σ^2) é simples. A hipótese nula é declarada como

$$H_0: \mu = \mu_0, \quad (C.31)$$

onde μ_0 é um valor que especificamos. Na maior parte das aplicações, $\mu_0 = 0$, mas o caso geral não é mais difícil.

A regra de rejeição que escolhemos depende da natureza da hipótese alternativa. As três alternativas de interesse são

$$H_1: \mu > \mu_0, \quad (C.32)$$

$$H_1: \mu < \mu_0, \quad (C.33)$$

e

$$H_1: \mu \neq \mu_0. \quad (C.34)$$

A equação (C.32) dá uma **alternativa unilateral**, assim como (C.33). Quando a hipótese alternativa é (C.32), a hipótese nula será efetivamente $H_0: \mu \leq \mu_0$, desde que rejeitemos H_0 somente quando $\mu > \mu_0$. Isso é adequado quando estivermos interessados no valor de μ apenas quando μ for pelo menos tão grande quanto μ_0 . A equação (C.34) é uma **alternativa bilateral**. Ela é adequada quando estivermos interessados em qualquer desvio da hipótese nula.

Considere a primeira alternativa em (C.32). Intuitivamente, devemos rejeitar H_0 em favor de H_1 quando o valor da média amostral, $\bar{y}$, for “suficientemente” maior do que μ_0 . Mas como podemos determinar quando $\bar{y}$ é grande o bastante para H_0 ser rejeitada no nível de significância escolhido? Isso requer que saibamos a probabilidade de rejeitar a hipótese nula quando ela for verdadeira. Em vez de trabalhar diretamente com $\bar{y}$, usaremos sua versão padronizada, onde σ é substituído pelo desvio padrão amostral, s :

$$t = \sqrt{n}(\bar{y} - \mu_0)/s = (\bar{y} - \mu_0)/\text{ep}(\bar{y}), \quad (C.35)$$

onde $\text{ep}(\bar{y}) = s/\sqrt{n}$ é o erro padrão de $\bar{y}$. Dada a amostra de dados, é fácil obter t . Trabalhamos com t porque, sob a hipótese nula, a variável aleatória

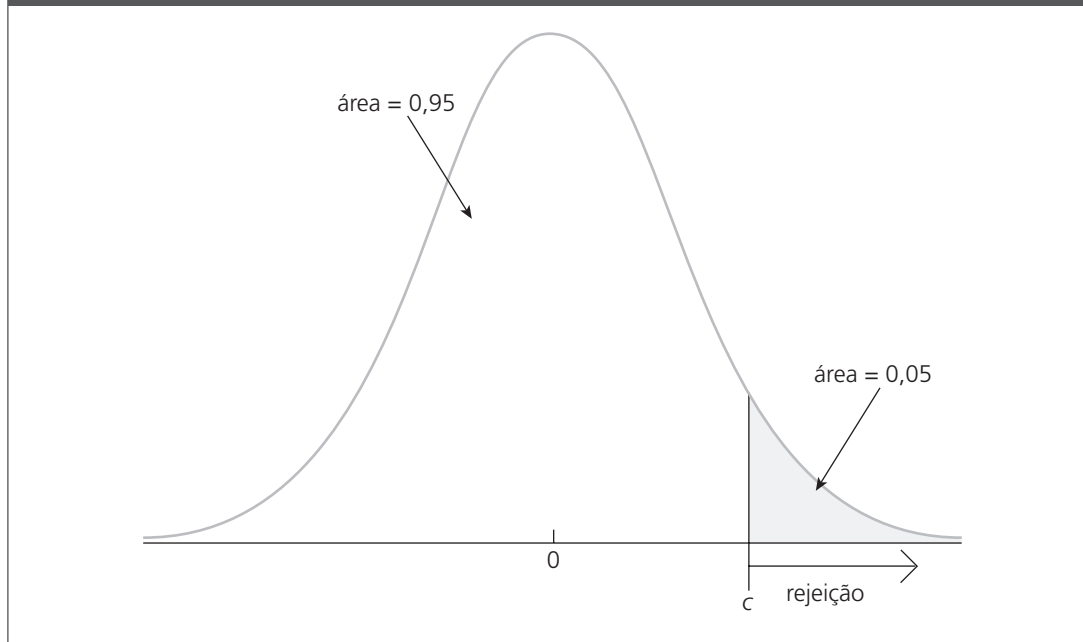
$$T = \sqrt{n}(\bar{Y} - \mu_0)/S$$

tem uma distribuição t_{n-1} . Agora, suponha que tenhamos definido um nível de significância de 5%. Assim, o valor crítico c é escolhido de forma que $P(T > c | H_0) = 0,05$; isto é, a probabilidade de um erro Tipo I é 5%. Uma vez encontrado c , a regra de rejeição é

$$t > c, \quad (C.36)$$

onde c é o $100(1 - \alpha)$ percentil em uma distribuição t_{n-1} ; como porcentagem, o nível de significância é $100 \cdot \alpha\%$. Este é um exemplo de **teste unicaudal** porque a região de rejeição está em uma “cauda” da distribuição t . Para um nível de significância de 5%, c é o 95º percentil da distribuição t_{n-1} ; isso é ilustrado na Figura C.5. Um nível de significância diferente leva a um valor crítico diferente.

FIGURA C.5 Região de rejeição para um teste com nível de significância de 5% contra a alternativa unilateral $\mu > \mu_0$.



A estatística da equação (C.35) muitas vezes é chamada de **estatística t** para testar $H_0: \mu = \mu_0$. A estatística t mede a distância de $\bar{y}$ a μ_0 em relação ao erro padrão de $\bar{y}$, $ep(\bar{y})$.

EXEMPLO C.4

Efeitos das zonas empresariais sobre investimentos em negócios

Na população de cidades outorgadas zonas empresariais em determinado estado [ver Papp (1994) para Indiana], seja Y a alteração percentual em investimentos do ano anterior até um ano depois que a cidade se tornou uma zona empresarial. Suponha que Y tenha uma distribuição Normal(μ, σ^2). A hipótese nula de que a zona empresarial não tem efeitos sobre investimentos empresariais é $H_0: \mu = 0$; a alternativa de que elas têm um efeito positivo é $H_1: \mu > 0$. (Partimos do pressuposto de que elas não têm efeito negativo.) Suponha que desejemos testar H_0 a um nível de 5%. A estatística teste neste caso é

$$t = \frac{\bar{y}}{s/\sqrt{n}} = \frac{\bar{y}}{ep(\bar{y})}. \quad (C.37)$$

Suponha que tenhamos uma amostra de 36 cidades outorgadas zonas empresariais. Assim, o valor crítico é $c = 1,69$ (ver Tabela G.2), e rejeitamos H_0 em favor de H_1 se $t > 1,69$. Suponha que a amostra gere $\bar{y} = 8,2$ e $s = 23,9$. Então, $t \approx 2,06$ e H_0 é, portanto, rejeitada ao nível de 5%. Assim, concluímos que, a um nível de 5%, zonas empresariais têm um efeito sobre os investimentos médios. O valor crítico de 1% é 2,44, assim H_0 não é rejeitada ao nível de 1%. O mesmo processo se mantém aqui como no Exemplo C.2: não controlamos outros fatores que poderiam afetar investimentos em cidades ao longo do tempo, por isso não podemos afirmar que o efeito é causal.

A regra de rejeição é similar para a alternativa unilateral (C.33). Um teste com um nível de significância de $100 \cdot \alpha\%$ rejeita H_0 contra (C.33) sempre que

$$t < -c; \quad (\text{C.38})$$

em outras palavras, estamos procurando valores negativos da estatística t – o que indica que $\bar{y} < \mu_0$ – que estejam suficientemente longe de zero para rejeitar H_0 .

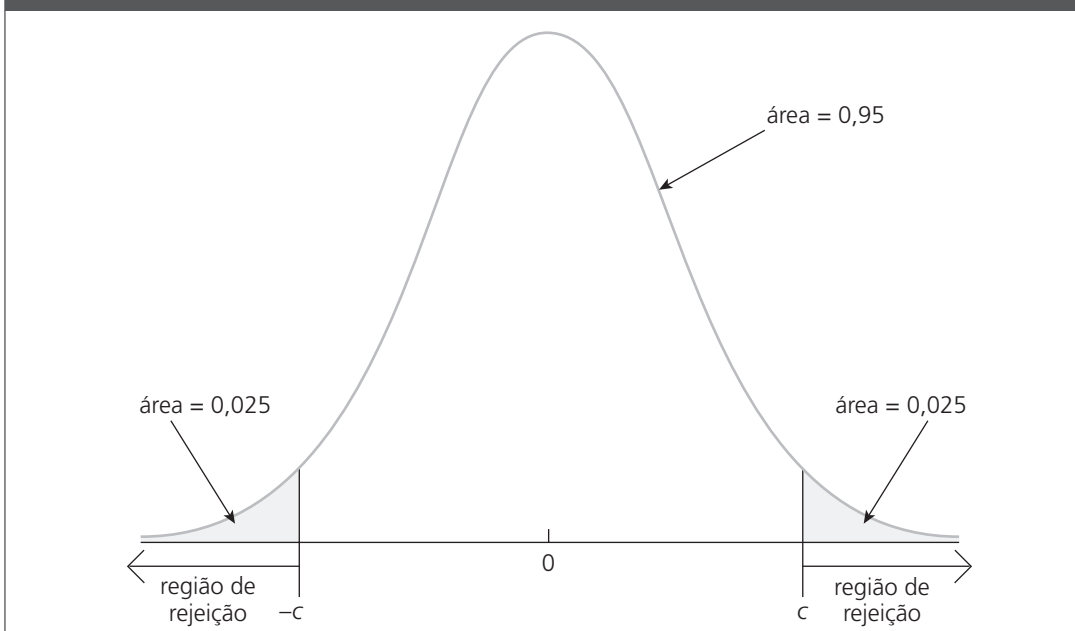
Com alternativas bilaterais, precisamos ser cuidadosos ao escolher o valor crítico, de forma que o nível de significância do teste ainda seja α . Se H_1 é dada por $H_1: \mu \neq \mu_0$, então rejeitamos H_0 se $\bar{y}$ estiver distante de μ_0 em *valor absoluto*: um $\bar{y}$ muito maior ou muito menor do que μ_0 cria evidências contra H_0 em favor de H_1 . Um teste de nível $100 \cdot \alpha\%$ é obtido a partir da regra de rejeição

$$|t| > c, \quad (\text{C.39})$$

onde $|t|$ é o valor absoluto da estatística t em (C.35). Isso nos dá um **teste bicaudal**. Agora precisaremos ter cuidado ao escolher o valor crítico: c é o $100(1 - \alpha/2)$ percentil na distribuição t_{n-1} . Por exemplo, se $\alpha = 0,05$, o valor crítico é o 97,5º percentil da distribuição t_{n-1} . Isso garante que H_0 seja rejeitada somente 5% do tempo quando ela for verdadeira (ver Figura C.6). Por exemplo, se $n = 22$, o valor crítico é $c = 2,08$, o 97,5º percentil em uma distribuição t_{n-1} (ver Tabela G.2). O valor absoluto da estatística t deve superar 2,08 com o objetivo de rejeitar H_0 contra H_1 no nível de 5%.

É importante conhecer a linguagem própria do teste de hipóteses. Às vezes, a frase adequada “fracassamos ao rejeitar H_0 em favor de H_1 a um nível de significância de 5%” é substituída por “aceitamos H_0 a um nível de significância de 5%”. A segunda frase é incorreta. Com o mesmo conjunto de dados, normalmente existem muitas

FIGURA C.6 Região de rejeição para um teste de nível de significância de 5% contra a alternativa bilateral $H_1: \mu \neq \mu_0$.



hipóteses que não podem ser rejeitadas. No exemplo anterior das eleições, seria logicamente inconsistente dizer que $H_0: \theta = 0,42$ e $H_0: \theta = 0,43$ são ambas “aceitas”, já que apenas uma delas pode ser verdadeira. Mas é totalmente possível que nenhuma dessas hipóteses seja rejeitada. Por essa razão, sempre dizemos “fracassamos ao rejeitar H_0 ” em vez de “aceitamos H_0 ”.

C.6c Testes assintóticos para populações não normais

Se o tamanho da amostra for grande o suficiente para invocar o teorema do limite central (ver Seção C.3), a mecânica do teste de hipóteses para médias populacionais será a *mesma*, sendo a distribuição da população normal ou não. A justificativa teórica vem do fato de que, sob a hipótese nula,

$$T = \sqrt{n}(\bar{Y} - \mu_0)/S \stackrel{a}{\sim} \text{Normal}(0,1).$$

Portanto, com um n maior, podemos comparar a estatística t em (C.35) com os valores críticos de uma distribuição normal padrão. Como a distribuição t_{n-1} converge para a distribuição normal padrão à medida que n aumenta, os valores críticos t e normal padrão serão bem próximos para um n extremamente grande. Já que a teoria assintótica é baseada em n aumentando sem limites, ela não nos diz se os valores críticos normal padrão ou t são melhores. Para valores moderados de n , digamos, entre 30 e 60, é tradicional usar a distribuição t porque sabemos que ela é correta para populações normais. Para $n > 120$, a opção entre as distribuições t e normal padrão é irrelevante porque os valores críticos são praticamente os mesmos.

Como os valores críticos escolhidos usando a distribuição normal padrão ou a distribuição t são válidos apenas aproximadamente para populações não normais, nossos níveis de significância escolhidos também são apenas aproximados; assim, para populações não normais, nossos níveis de significância são realmente níveis de significância *assintóticos*. Dessa forma, se escolhermos um nível de significância de 5%, mas nossa população não for normal, o nível de significância real será maior ou menor do que 5% (e não saberemos qual é o caso). Quando o tamanho da amostra for grande, o nível de significância real será bem próximo de 5%. Falando de forma prática, a distinção não é importante, por isso abandonaremos o qualificador “assintótico”.

EXEMPLO C.5

Discriminação racial em contratações

No estudo do Urban Institute sobre discriminação nas contratações (ver Exemplo C.3) usando dados do arquivo AUDIT, estamos principalmente interessados em testar $H_0: \mu = 0$ contra $H_1: \mu < 0$, onde $\mu = \theta_N - \theta_B$ é a diferença nas probabilidades de que negros e brancos recebam ofertas de emprego. Lembre-se de que μ é a média populacional da variável $Y = N - B$, onde N e B são indicadores binários. Usando as comparações pareadas $n = 241$ do arquivo de dados AUDIT obtivemos $\bar{y} = -0,133$ e $\text{ep}(\bar{y}) = 0,482/\sqrt{241} \approx 0,031$. A estatística t do teste $H_0: \mu = 0$ é $t = -0,133/0,031 \approx -4,29$. Você vai se lembrar do Apêndice B que dizia que a distribuição normal padrão é, para propósitos práticos, indistinguível da distribuição t com 240 graus de liberdade. O valor $-4,29$ está tão distante na extremidade esquerda da distribuição que rejeitamos H_0 em qualquer nível de significância razoável. Na verdade, o valor crítico 0,005 (meio por cento) (para o teste unilateral) é cerca de $-2,58$. Um t valor de $-4,29$ é evidência *muito* forte contra H_0 em favor de H_1 . Portanto, concluímos que há discriminação nas contratações.

C.6d Calculando e usando p -valores

A exigência tradicional de escolher um nível de significância com antecedência significa que pesquisadores diferentes, usando os mesmos dados e os mesmos procedimentos para testar as mesmas hipóteses, podem acabar com diferentes conclusões. Registrar o nível de significância com o qual estamos realizando o teste resolve esse problema até certo grau, mas não o elimina completamente.

Para gerar mais informações, você pode fazer a seguinte pergunta: Qual é o *maior* nível de significância no qual podemos realizar o teste e ainda assim fracassar ao rejeitar a hipótese nula? Esse valor é conhecido como **p -valor** de um teste (às vezes chamado de *prob-valor*). Comparado com a escolha de um nível de significância com antecedência e a obtenção de um valor crítico, o cálculo de um p -valor é, de certa forma, mais difícil. Mas, com o advento de cálculos rápidos e baratos, os p -valores agora são bem fáceis de obter.

Como ilustração, considere o problema de testar $H_0: \mu = 0$ em uma população Normal(μ, σ^2). Nossa estatística teste neste caso é $T = \sqrt{n} \cdot \bar{Y}/S$, e presumimos que n é grande o bastante para tratar T como se tivesse uma distribuição normal padrão sob H_0 . Suponha que o valor observado de T para nossa amostra seja $t = 1,52$. (Note como pulamos a etapa de escolher um nível de significância.) Agora que vimos o valor t , podemos encontrar o maior nível de significância ao qual podemos fracassar ao rejeitar H_0 . Este é o nível de significância associado com o uso de t como nosso valor crítico. Como nossa estatística teste T tem uma distribuição normal padrão sob H_0 , temos

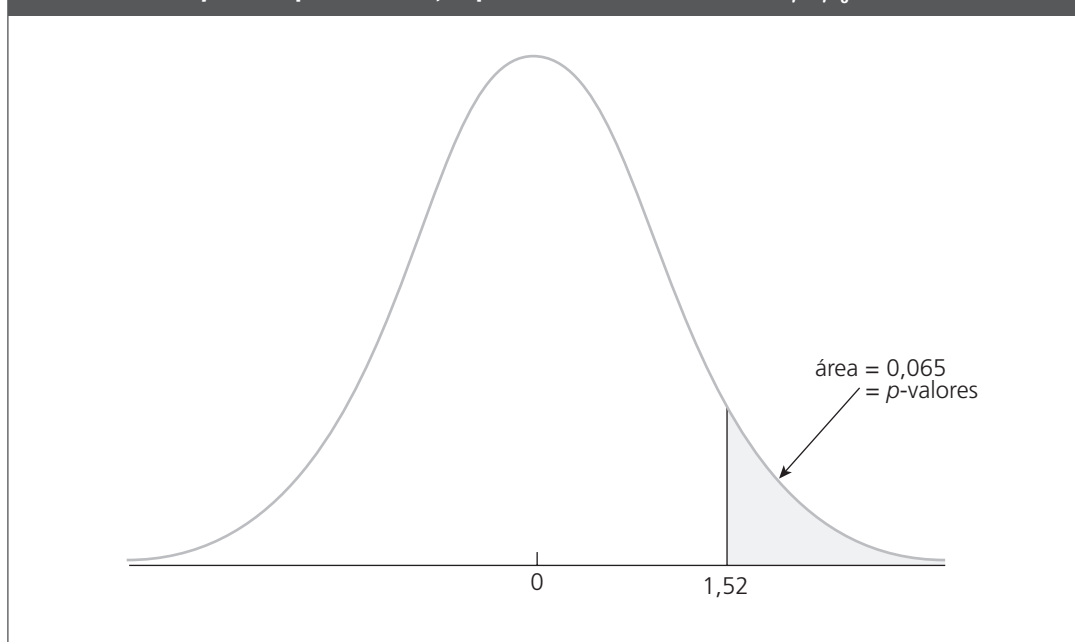
$$p\text{-value} = P(T > 1,52 | H_0) = 1 - \Phi(1,52) = 0,065, \quad (\text{C.40})$$

onde $\Phi(\cdot)$ indica a fdc normal padrão. Em outras palavras, o p -valor deste exemplo é simplesmente a área à direita de 1,52, o valor observado na estatística teste, em uma distribuição normal padrão. Ver Figura C.7 para ilustração.

Devido ao p -valor = 0,065, o maior nível de significância com o qual podemos realizar este teste e fracassar ao rejeitar é 6,5%. Se executarmos o teste com um nível abaixo de 6,5% (como 5%), fracassaremos ao rejeitar H_0 . Se executarmos o teste com um nível mais alto do que 6,5% (como 10%), rejeitaremos H_0 . Com o p -valor à mão, podemos fazer o teste em qualquer nível.

O p -valor neste exemplo tem outra interpretação útil: a probabilidade de observarmos um valor de T tão grande quanto 1,52 quando a hipótese nula for verdadeira. Se a hipótese nula for realmente verdadeira, poderíamos observar um valor de T tão grande quanto 1,52 casualmente apenas em 6,5% do tempo. O fato de ela ser pequena o suficiente para rejeitarmos H_0 depende de nossa tolerância a um erro Tipo I. O p -valor tem uma interpretação similar em todos os outros casos, como veremos.

Geralmente, p -valores pequenos são evidências *contra* H_0 , já que indicam que o resultado dos dados ocorre com pequena probabilidade se H_0 for verdadeira. No exemplo anterior, se t tivesse um valor maior, digamos, $t = 2,85$, então o p -valor seria $1 - \Phi(2,85) \approx 0,002$. Isso quer dizer que, se a hipótese nula fosse verdadeira, observaríamos um valor de T tão grande quanto 2,85 com probabilidade 0,002. Como interpretamos isso? Podemos ter obtido uma amostra muito incomum ou a hipótese nula é falsa. A não ser que tenhamos uma tolerância *muito* pequena a erros Tipo I, rejeitaremos a hipótese nula. Por outro lado, um p -valor grande é uma evidência fraca contra H_0 . Se tivéssemos obtido $t = 0,47$ no exemplo anterior, o p -valor = $1 - \Phi(0,47) = 0,32$.

FIGURA C.7 O p -valor quando $t = 1,52$ para a alternativa unilateral $\mu > \mu_0$.

Observar um valor de T maior do que 0,47 ocorre com probabilidade 0,32 mesmo quando H_0 é verdadeira; ele é grande o suficiente, portanto, as dúvidas em relação a H_0 são insuficientes, a não ser que tenhamos uma tolerância muito alta a erros Tipo I.

Em testes de hipóteses sobre uma média populacional usando a distribuição t , precisamos de tabelas detalhadas para calcular os p -valores. A Tabela G.2 só permite que coloquemos limites nos p -valores. Felizmente, muitos programas estatísticos e econométricos agora calculam p -valores rotineiramente, e também oferecem cálculos de fdc's para distribuições t e outras usadas para calcular os p -valores.

EXEMPLO C.6**Efeito de subsídios de treinamento profissional sobre a produtividade do trabalhador**

Considere novamente os dados de Holzer et al. (1993) do Exemplo C.2. Na perspectiva de políticas públicas, existem duas questões de interesse. Primeira, qual é nossa melhor estimativa da alteração média nas taxas de refugo, μ ? Já obtivemos isso para a amostra de 20 empresas listada na Tabela C.3: a média amostral de alteração nas taxas de refugo é de $-1,15$. Em relação à taxa de refugo inicial média de 1987, isso representa uma queda de cerca de 26,3% ($-1,15/4,38 \approx -0,263$), que é um efeito incomum.

Também gostaríamos de saber se a amostra apresenta fortes evidências para um efeito sobre a população de empresas manufatureiras que podem ter recebido subsídios. A hipótese nula é $H_0: \mu = 0$, e nós testamos isto contra $H_1: \mu < 0$, onde μ é a alteração média nas taxas de refugo. Sob a hipótese nula, os subsídios de capacitação profissional não tiveram efeito sobre as taxas de refugo médias. A alternativa afirma que há um efeito. Não nos importamos com a alternativa $\mu > 0$, por isso a hipótese nula é efetivamente $H_0: \mu \geq 0$.

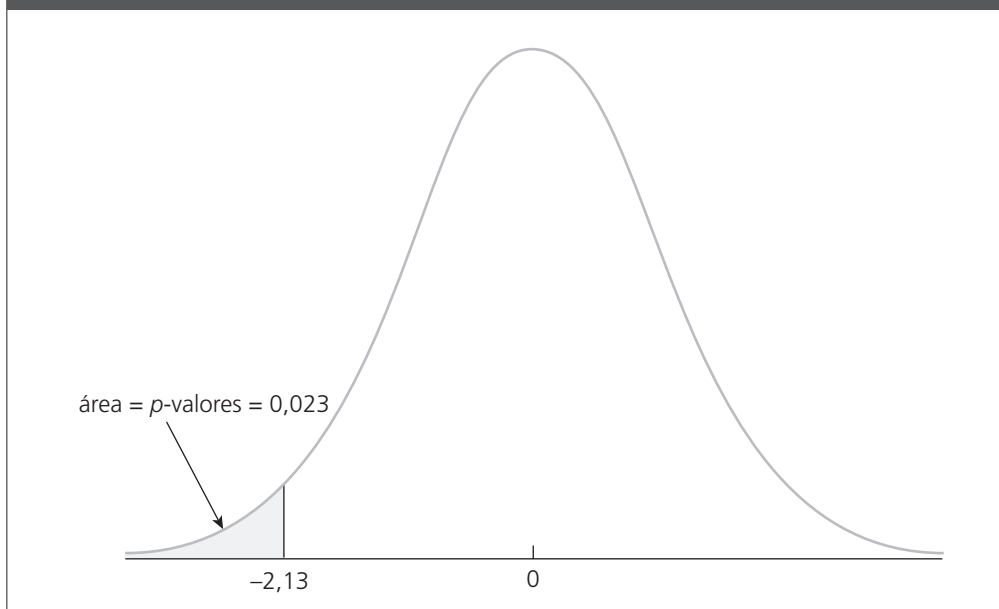
Como $\bar{y} = -1,15$ e $ep(\bar{y}) = 0,54$, $t = -1,15/0,54 = -2,13$. Isso está abaixo do valor crítico de 5% de $-1,73$ (de uma distribuição t_{19}), mas acima do valor crítico de 1%, $-2,54$. O p -valor neste caso é calculado como

$$p\text{-value} = P(T_{19} < -2,13), \quad (\text{C.41})$$

onde T_{19} representa uma variável aleatória distribuída como t com 19 graus de liberdade. A desigualdade é inversa a partir de (C.40) porque a alternativa tem a forma de (C.33). A probabilidade em (C.41) é a área à esquerda de $-2,13$ em uma distribuição t_{19} (ver Figura C.8).

Usando a Tabela G.2, o máximo que podemos dizer é que o p -valor está entre 0,025 e 0,01, mas está mais próximo de 0,025 (já que o 97,5º percentil é cerca de 2,09). Usando um pacote estatístico, como Stata, podemos calcular o p -valor exato. O resultado é 0,023, que é uma evidência razoável contra H_0 . Isso é, com certeza, evidência suficiente para rejeitar a hipótese nula de que os subsídios de treinamento não tinham efeito a um nível de significância de 2,5% (e, conseqüentemente, a um nível de 5%).

FIGURA C.8 O p -valor quando $t = -2,13$ com 19 graus de liberdade para a alternativa unilateral $\mu < 0$.



Calcular um p -valor para um teste bilateral é parecido, mas temos de levar em conta a natureza bilateral da regra de rejeição. Para testes t sobre médias populacionais, o p -valor é calculado como

$$P(|T_{n-1}| > |t|) = 2P(T_{n-1} > |t|), \quad (\text{C.42})$$

onde t é o valor da estatística teste e T_{n-1} é uma variável aleatória t . (Para um n maior, substitua T_{n-1} por uma variável aleatória normal padrão.) Assim, calcule o valor

absoluto da estatística t , encontre a área à direita deste valor em uma distribuição t_{n-1} e multiplique a área por dois.

Em populações não normais, o p -valor exato pode ser difícil de obter. No entanto, podemos encontrar p -valores *assimptóticos* usando os mesmos cálculos. Esses p -valores são válidos para tamanhos de amostras grandes. Para n maior que, digamos, 120, também podemos usar a distribuição normal padrão. A Tabela G.1 é detalhada o bastante para obtermos p -valores exatos, mas também podemos usar um programa estatístico ou econométrico.

EXEMPLO C.7**Discriminação racial em contratações**

Usando os dados de pares combinados do Urban Institute que estão no arquivo AUDIT ($n = 241$), obtivemos $t = -4,29$. Se Z é uma variável aleatória normal padrão, $P(Z < -4,29)$ será, para propósitos práticos, zero. Em outras palavras, o p -valor (assimptótico) para este exemplo é essencialmente zero. Isso é uma evidência muito forte contra H_0 .

Resumo sobre como usar p -valores:

(i) Escolha uma estatística teste T e decida sobre a natureza da alternativa. Isso determina se a regra de rejeição será $t > c$, $t < -c$ ou $|t| > c$.

(ii) Use o valor observado da estatística t como valor crítico e calcule o nível de significância correspondente do teste. Este é o p -valor. Se a regra de rejeição tem a forma $t > c$, então o p -valor = $P(T > t)$. Se a regra de rejeição for $t < -c$, o p -valor = $P(T < t)$; se a regra de rejeição for $|t| > c$, então o p -valor = $P(|T| > |t|)$.

(iii) Se um nível de significância α foi escolhido, rejeitamos H_0 ao nível $100 \cdot \alpha\%$ se o p -valor $< \alpha$. Se o p -valor $\geq \alpha$, fracassaremos na rejeição de H_0 ao nível de $100 \cdot \alpha\%$. Portanto, é um p -valor pequeno que leva à rejeição da hipótese nula.

C.6e A relação entre intervalos de confiança e testes de hipóteses

Como a construção de intervalos de confiança e testes de hipóteses envolvem afirmações de probabilidade, é natural pensar que eles sejam, de certa forma, conectados. E são mesmo. Depois que um intervalo de confiança foi construído, podemos executar uma variedade de testes de hipóteses.

Os intervalos de confiança que discutimos eram todos bilaterais por natureza. (Neste livro, não teremos necessidade de construir intervalos de confiança unilaterais.) Assim, intervalos de confiança podem ser usados para testar alternativas *bilaterais*. No caso da média populacional, a hipótese nula é dada por (C.31), e a alternativa é (C.34). Suponha que tenhamos construído um intervalo de confiança de 95% para μ . Dessa forma, se o valor hipotético de μ sob H_0 , μ_0 , não é o intervalo de confiança, $H_0: \mu = \mu_0$ é rejeitada contra $H_1: \mu \neq \mu_0$ a um nível de 5%. Se μ_0 estiver neste intervalo, fracassaremos ao rejeitar H_0 a um nível de 5%. Note como qualquer valor de μ_0 pode ser testado uma vez que o intervalo de confiança esteja construído, e, desde que o intervalo de confiança contenha mais de um valor, existirão muitas hipóteses nulas que não serão rejeitadas.

EXEMPLO C.8**Subsídios de treinamento e produtividade do trabalhador**

No exemplo de Holzer et al., construímos um intervalo de confiança de 95% para a alteração média da taxa de refugo μ como $[-2,28, -0,02]$. Já que zero está excluído deste intervalo, rejeitamos $H_0: \mu = 0$ contra $H_1: \mu \neq 0$ a um nível de 5%. Esse intervalo de confiança de 95% também significa que não conseguimos rejeitar $H_0: \mu = -2$ a um nível de 5%. Na verdade, há uma série de hipóteses nulas que não são rejeitadas, dado esse intervalo de confiança.

C.6f Significância prática versus estatística

Nos exemplos abordados até aqui, produzimos três tipos de evidências em relação a parâmetros populacionais: estimativas pontuais, intervalos de confiança e teste de hipóteses. Essas ferramentas para aprendizado sobre parâmetros populacionais são igualmente importantes. Há uma tendência compreensível de estudantes focarem em intervalos de confiança e testes de hipóteses porque são coisas às quais podemos atribuir níveis de confiança ou de significância. Porém, em um estudo, também precisamos interpretar as *magnitudes* das estimativas pontuais.

O sinal e a magnitude de $\bar{y}$ determinam sua **significância prática** e permitem que discutamos a direção de uma intervenção ou efeito de uma política pública, e se o efeito estimado é “grande” ou “pequeno”. Por outro lado, a **significância estatística** de $\bar{y}$ depende da magnitude de sua estatística t . Para testar $H_0: \mu = 0$, a estatística t seria simplesmente $t = \bar{y}/\text{ep}(\bar{y})$. Em outras palavras, a significância estatística depende da relação de $\bar{y}$ com seu erro padrão. Consequentemente, uma estatística t pode ser grande porque $\bar{y}$ é grande ou $(\bar{y})$ é pequeno. Em aplicações, é importante discutir a significância estatística e a significância prática, estando ciente de que uma estimativa pode ser estatisticamente significativa, sem ser especialmente grande em sentido prático. Para uma estimativa ser ou não importante na prática vai depender tanto do contexto quanto do julgamento do indivíduo, por isso não existem regras definidas para determinar a significância prática.

EXEMPLO C.9**Efeito da largura de uma rodovia sobre o tempo de viagem**

Seja Y a variação no tempo de viagem, medida em minutos, de trabalhadores de uma área metropolitana, desde antes de a rodovia ser ampliada até depois dessas obras. Suponha que $Y \sim \text{Normal}(\mu, \sigma^2)$. A hipótese nula de que a ampliação não reduziu o tempo de viagem médio é $H_0: \mu = 0$; a alternativa que diz que ela reduziu o tempo médio de viagem é $H_1: \mu < 0$. Suponha que uma amostra aleatória de trabalhadores com tamanho $n = 900$ tenha sido obtida para determinar a eficácia do projeto da rodovia. A variação média do tempo de viagem é calculada como $\bar{y} = -3,6$ e o desvio padrão da amostra é $s = 32,7$; assim, $\text{se}(\bar{y}) = 32,7/\sqrt{900} = 1,09$. A estatística t é $t = -3,6/1,09 \approx -3,30$, que é muito significativa estatisticamente; o p -valor é de cerca de 0,0005. Dessa forma, concluímos que a ampliação da rodovia teve um efeito estatisticamente significativo sobre o tempo de viagem médio.

Se o resultado do teste de hipótese fosse tudo o que registramos a partir do estudo, ele seria equivocado. Registrar apenas a significância estatística mascara o fato de que a redução estimada do tempo de viagem médio, 3,6 minutos, parece muito pequena, embora ela dependa até certo ponto de qual era o tempo de viagem médio antes da ampliação da rodovia. Para sermos honestos, devemos registrar a estimativa pontual de $-3,6$, ao lado de seu teste de significância.

Pode acontecer de encontrarmos estimativas pontuais que são estatisticamente significantes sem serem significantes na prática, quando estivermos trabalhando com amostras grandes. Para discutir por que isso acontece, é útil conhecer a seguinte definição.

Consistência de um teste. Um teste consistente rejeita H_0 com probabilidade se aproximando de um à medida que o tamanho da amostra aumenta sempre que H_1 for verdadeira.

Outra forma de dizer que um teste é consistente é que, à medida que o tamanho da amostra tende ao infinito, o poder do teste fica cada vez mais próximo de um quando H_1 for verdadeira. Todos os testes que apresentamos no texto têm essa propriedade. No caso do teste de hipóteses sobre a média populacional, a consistência do teste ocorre porque a variância de $\bar{Y}$ converge a zero à medida que o tamanho da amostra aumenta. A estatística t para testar $H_0: \mu = 0$ é $T = \bar{Y}/(S/\sqrt{n})$. Já que $\text{plim}(\bar{Y}) = \mu$ e $\text{plim}(S) = \sigma$, verifica-se que, se, por exemplo, $\mu > 0$, T fica cada vez maior (com alta probabilidade) à medida que $n \rightarrow \infty$. Em outras palavras, não importa quão próxima m esteja de zero, podemos ter quase certeza para rejeitar $H_0: \mu = 0$ com um tamanho de amostra grande o suficiente. Isso não diz nada a respeito de μ ser grande em sentido prático.

C.7 Observações sobre notações

Em nossa revisão de probabilidade e estatística aqui e no Apêndice B, fomos cuidadosos ao usar convenções padrão para indicar variáveis aleatórias, estimadores e estatísticas teste. Por exemplo, usamos W para indicar um estimador (variável aleatória) e w para indicar determinada estimativa (resultado da variável aleatória W). Diferenciar um estimador e uma estimativa é importante para entender vários conceitos da estimação e do teste de hipóteses. No entanto, fazer essa distinção rapidamente se torna uma dificuldade na análise econométrica porque os modelos são mais complicados: muitas variáveis aleatórias e parâmetros serão envolvidos, e estar em conformidade com as convenções usuais da probabilidade e estatística exige muitos símbolos extras.

No texto principal, usamos uma convenção mais simples e que é amplamente usada em econometria. Se θ é um parâmetro populacional, a notação $\hat{\theta}$ (“teta chapéu”) será utilizada para indicar um estimador e uma estimativa de θ . Essa notação é útil porque proporciona uma forma simples de ligar um estimador ao parâmetro populacional que ele supostamente está estimando. Assim, se o parâmetro populacional for β , então $\hat{\beta}$ indicará um estimador ou estimativa de β ; se o parâmetro for σ^2 , $\hat{\sigma}^2$ será um estimador ou estimativa de σ^2 ; e assim por diante. Às vezes, discutiremos dois estimadores do mesmo parâmetro, caso em que vamos precisar de uma notação diferente, como $\tilde{\theta}$ (“teta til”).

Embora a eliminação de convenções de probabilidade e estatística para indicar estimadores, variáveis aleatórias e estatísticas teste coloque maior responsabilidade sobre você, ela não será um grande problema se a diferença entre um estimador e uma estimativa for compreendida. Se estivermos discutindo propriedades *estatísticas* de $\hat{\theta}$ – como derivar se ele é não viesado ou consistente –, estaremos necessariamente vendo $\hat{\theta}$ como um estimador. Por outro lado, se escrevermos algo como $\hat{\theta} = 1,73$, claramente estamos indicando uma estimativa pontual para determinada amostra de dados. A confusão que pode surgir ao usar $\hat{\theta}$ para indicar ambos pode ser mínima se você tiver um bom entendimento de probabilidade e estatística.

Resumo

Discutimos tópicos da estatística matemática que são grandes bases da análise econométrica. A noção de um estimador, que é simplesmente uma regra para combinar dados para estimar um parâmetro populacional, é fundamental. Abordamos várias propriedades dos estimadores. As mais importantes das propriedades de amostras pequenas são a ausência de viés e a eficiência, que depende da comparação de variâncias quando estimadores forem não viesados. Propriedades de amostras grandes dizem respeito à sequência de estimadores obtidos à medida que o tamanho da amostra aumenta, e também são bases da econometria. Qualquer estimador útil é consistente. O teorema do limite central indica que, em amostras grandes, a distribuição por amostragem da maioria dos estimadores é aproximadamente normal.

A distribuição por amostragem de um estimador pode ser usada para construir intervalos de confiança. Vimos isso para estimar a média de uma distribuição normal e para calcular intervalos de confiança aproximados em casos não normais. Testes clássicos de hipóteses, que exigem a especificação de uma hipótese nula, uma hipótese alternativa e um nível de significância, são realizados comparando uma estatística teste a um valor crítico. De forma alternativa, um p -valor pode ser calculado para permitir que realizemos um teste em qualquer nível de significância.

Termos-chave

Alternativa bilateral	Estimador de mínimos	Normalidade assintótica
Alternativa unilateral	quadrados	Poder de um teste
Amostra aleatória	Estimador de verossimilhança	População
Coefficiente de correlação	máxima	p -valor
amostral	Estimador não viesado	Região de rejeição
Covariância amostral	Estimador não viesado com	Significância estatística
Desvio padrão amostral	variância mínima	Significância prática
Desvio padrão por	Estimador viesado	Teorema do limite central
amostragem	Estimativa	(TLC)
Distribuição por amostragem	Hipótese alternativa	Teste bicaudal
Erro padrão	Hipótese nula	Teste consistente
Erro quadrático médio (EQM)	Inconsistente	Teste de hipótese
Erro Tipo I	Intervalo de confiança	Teste unicaudal
Erro Tipo II	Lei dos grandes números	Valor crítico
Estatística t	(LGN)	Variância amostral
Estatística teste	Limite de probabilidade	Variância por amostragem
Estimador	Média amostral	Viés
Estimador consistente	Método dos momentos	
Estimador de intervalo	Nível de significância	

Problemas

- Defina Y_1, Y_2, Y_3 e Y_4 como variáveis aleatórias independentes e identicamente distribuídas de uma população com média μ e variância σ^2 . Permita que $\bar{y} = \frac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4)$. Indique a média dessas quatro variáveis aleatórias.
 - Quais são o valor esperado e a variância de $\bar{Y}$ em termos de μ e σ^2 ?
 - Agora, considere um estimador diferente de μ

$$W = \frac{1}{8}Y_1 + \frac{1}{8}Y_2 + \frac{1}{4}Y_3 + \frac{1}{2}Y_4.$$

Este é um exemplo de média *ponderada* de Y_i . Mostre que W também é um estimador não viesado de μ . Encontre a variância de W .

- Com base em suas respostas dos itens (i) e (ii), qual estimador de m você prefere, $\bar{Y}$ ou W ?
- Essa é uma versão mais geral do Problema C.1. Permita que $Y_1, Y_2, \dots, Y_n$ indiquem n variáveis aleatórias não correlacionadas em pares com média comum m e variância comum σ^2 . Faça com que $\bar{Y}$ represente a média amostral.
 - Defina a classe de *estimadores lineares* de μ por

$$W_a = a_1Y_1 + a_2Y_2 + \dots + a_nY_n,$$

onde os a_i são constantes. Qual restrição sobre a_i é necessária para que W_a seja um estimador não viesado de μ ?

- Encontre $\text{Var}(W_a)$.
 - Para quaisquer números $a_1, a_2, \dots, a_n$ a seguinte desigualdade se mantém: $(a_1 + a_2 + \dots + a_n)^2/n \leq a_1^2 + a_2^2 + \dots + a_n^2$. Use a desigualdade, junto com os itens (i) e (ii), para mostrar que $\text{Var}(W_a) \geq \text{Var}(\bar{Y})$ sempre que W_a for não viesado, de forma que $\bar{Y}$ seja o *melhor estimador linear não viesado*. [Dica: O que a desigualdade se torna quando a_i satisfaz a restrição do item (i)?]
- Deixe $\bar{Y}$ indicar a média amostral de uma amostra aleatória com média μ e variância σ^2 . Considere dois estimadores alternativos de μ : $W_1 = [(n-1)/n]\bar{Y}$ e $W_2 = \bar{Y}/2$.
 - Mostre que W_1 e W_2 são ambos estimadores viesados de μ e encontre os vieses. O que acontece com os vieses à medida que $n \rightarrow \infty$? Comente sobre qualquer diferença importante no viés dos dois estimadores à medida que o tamanho da amostra aumenta.
 - Encontre os limites de probabilidade de W_1 e W_2 . {Dica: Use as Propriedades PLIM.1 e PLIM.2; para W_1 , note que $\text{plim}[(n-1)/n] = 1$ }. Qual estimador é consistente?
 - Descubra $\text{Var}(W_1)$ e $\text{Var}(W_2)$.
 - Defenda que W_1 é um estimador melhor do que $\bar{Y}$ se μ for “próxima” de zero. (Considere viés e variância.)
 - Para variáveis aleatórias positivas X e Y , suponha que o valor esperado de Y , dado X , seja $E(Y|X) = \theta X$. O parâmetro desconhecido θ mostra como o valor esperado de Y muda com X .
 - Defina a variável aleatória $Z = Y|X$. Mostre que $E(Z) = \theta$. [Dica: Use a Propriedade EC.2 junto com a lei das expectativas iteradas, Propriedade EC.4. Em particular, mostre primeiro que $E(Z|X) = \theta$ e depois use EC.4.]

- (ii) Use o item (i) para provar que o estimador $W_1 = n^{-1} \sum_{i=1}^n (Y_i/X_i)$ é não viesado para θ , onde $\{(X_i, Y_i): i = 1, 2, \dots, n\}$ é uma variável aleatória.
- (iii) Explique por que o estimador $W_2 = \bar{Y}/\bar{X}$, em que as barras superiores indicam médias amostrais, não é o mesmo que W_1 . Apesar disso, mostre que W_2 também é não viesado para θ .
- (iv) A tabela a seguir contém dados sobre a produção de milho em vários condados de Iowa. O USDA prevê o número de hectares de milho em cada condado com base em fotos de satélite. Pesquisadores contam o número de “pixels” de milho na imagem de satélite (em oposição ao número, por exemplo, de pixels de grãos de soja ou de terra não cultivada) e usam o valor para prever o número real de hectares. Para desenvolver uma equação de previsão para ser usada para condados em geral, o USDA fez uma pesquisa com fazendeiros nos condados selecionados para obter a produção de milho em hectares. Que Y_i seja a produção de milho no condado i e X_i o número de pixels de milho na foto de satélite do condado i . Existem $n = 17$ observações para oito condados. Use essa amostra para calcular as estimativas de θ desenvolvidas nos itens (ii) e (iii). As estimativas são similares?

Terreno	Produção de milho	Pixels de milho
1	165,76	374
2	96,32	209
3	76,08	253
4	185,35	432
5	116,43	367
6	162,08	361
7	152,04	288
8	161,75	369
9	92,88	206
10	149,94	316
11	64,75	145
12	127,07	355
13	133,55	295
14	77,70	223
15	206,39	459
16	108,33	290
17	118,17	307

5. Seja Y uma variável aleatória de Bernoulli(θ) com $0 < \theta < 1$. Suponha que estejamos interessados em estimar a *razão de probabilidade*, $\gamma = \theta/(1 - \theta)$, que é a probabilidade de sucesso sobre a probabilidade de fracasso. Dada uma amostra aleatória $\{Y_1, \dots, Y_n\}$, sabemos que um estimador consistente e não viesado de θ é $\bar{Y}$, a proporção de sucessos em n testes. Um estimador natural de γ é $G = \bar{Y}/(1 - \bar{Y})$, a proporção de sucessos sobre a proporção de fracassos na amostra.
- (i) Por que G não é um estimador não viesado de γ ?
- (ii) Use PLIM.2 (iii) para mostrar que G é um estimador consistente de γ .
6. Você foi contratado pelo governador para estudar se um imposto sobre bebidas alcoólicas diminuiu o consumo médio de álcool em seu estado. Você conseguiu obter, para uma amostra de indivíduos selecionados aleatoriamente, a diferença no consumo de bebidas alcoólicas (em onças) para os anos antes e depois do imposto. Para a pessoa i ,

que foi selecionada de forma aleatória na população, Y_i indica a mudança no consumo de álcool. Trate-as como uma amostra aleatória de uma distribuição Normal(μ, σ^2).

- (i) A hipótese nula é que não houve mudança no consumo de bebidas alcoólicas. Declare isso formalmente em termos de μ .
 - (ii) A alternativa é que houve um declínio no consumo de álcool; expresse a alternativa em termos de μ .
 - (iii) Agora, suponha que o tamanho de sua amostra seja $n = 900$ e que você obteve as estimativas $\bar{y} = -32,8$ e $s = 466,4$. Calcule a estatística t para testar H_0 contra H_1 ; obtenha o p -valor do teste. (Devido ao tamanho grande da amostra, use somente a distribuição normal padrão da Tabela G.1.) Você rejeita H_0 a um nível de 5%? E a um nível de 1%?
 - (iv) Você diria que a queda estimada no consumo é grande em magnitude? Comente a significância prática versus a significância estatística dessa estimativa.
 - (v) O que foi implicitamente suposto em sua análise a respeito de outros determinantes do consumo de bebidas alcoólicas durante o período de dois anos com o objetivo de inferir causalidade entre a mudança nos impostos e o consumo de álcool?
- 7** A nova gerência de uma padaria alega que os trabalhadores agora estão mais produtivos do que eram sob a antiga gestão, motivo pelo qual os salários foram “aumentados de forma geral”. Que W_i^b seja o salário do trabalhador i sob a antiga gestão e W_i^a o salário do trabalhador i depois da mudança. A diferença é $D_i \equiv W_i^a - W_i^b$. Suponha que D_i seja uma amostra aleatória de uma distribuição Normal(μ, σ^2).
- (i) Usando os dados abaixo sobre 15 trabalhadores, construa um intervalo de confiança exato de 95% para μ .
 - (ii) Expresse formalmente a hipótese nula de que não houve mudança nos salários médios. Em especial, qual é $E(D_i)$ sob H_0 ? Se você tivesse sido contratado para examinar a validade da afirmação do novo gerente, qual seria a hipótese alternativa relevante em termos de $\mu = E(D_i)$?
 - (iii) Teste a hipótese nula do item (ii) contra a alternativa expressa em níveis de 5% e 1%.
 - (iv) Obtenha o p -valor do teste do item (iii).

Trabalhador	Salário antes	Salário depois
1	8,30	9,25
2	9,40	9,00
3	9,00	9,25
4	10,50	10,00
5	11,40	12,00
6	8,75	9,50
7	10,00	10,25
8	9,50	9,50
9	10,80	11,50
10	12,55	13,10
11	12,00	11,50
12	8,65	9,00
13	7,75	7,75
14	11,25	11,50
15	12,65	13,00

- 8 O *New York Times* (05/02/1990) registrou o desempenho de arremessos de três pontos dos 10 principais arremessadores da NBA. A tabela abaixo resume esses dados:

Jogador	LR - LC
Mark Price	429-188
Trent Tucker	833-345
Dale Ellis	1.149-472
Craig Hodges	1.0116-396
Danny Ainge	1.051-406
Byron Scott	676-260
Reggie Miller	416-159
Larry Bird	1.206-455
Jon Sundvold	440-166
Brian Taylor	417-157

Nota: LR = lançamentos realizados e LC = lançamentos convertidos.

Para determinado jogador, o resultado de um dado arremesso pode ser modelado como uma variável de Bernoulli (zero-um): se Y_i é o resultado do arremesso i , então $Y_i = 1$ se o lançamento for convertido, e $Y_i = 0$ se ele for perdido. Seja θ a probabilidade de converter qualquer tentativa de arremesso de três pontos. O estimador natural de θ é $\bar{Y} = LC/LR$.

- (i) Estime θ de Mark Price.
 - (ii) Encontre o desvio padrão do estimador $\bar{Y}$ em termos de θ e o número de tentativas de arremessos, n .
 - (iii) A distribuição assintótica de $(\bar{Y} - \theta)/\text{ep}(\bar{Y})$ é normal padrão, onde $\text{ep}(\bar{Y}) = \sqrt{\bar{Y}(1 - \bar{Y})/n}$. Use esse fato para testar $H_0: \theta = 0,5$ contra $H_1: \theta < 0,5$ para Mark Price. Use um nível de significância de 1%.
- 9 Suponha que um ditador militar de um país não nomeado realize um plebiscito (um voto de confiança sim/não) e alegue ter sido apoiado por 65% dos eleitores. Um grupo de direitos humanos suspeita de uma jogada desonesta e contrata você para testar a validade da afirmação do ditador. Você tem um orçamento que lhe permite fazer uma amostragem aleatória de 200 eleitores do país.
- (i) Defina X como o número de votos sim obtidos em uma amostra aleatória de 200 pessoas entre toda a população de eleitores. Qual é o valor esperado de X se, na verdade, 65% de todos os eleitores apoiaram o ditador?
 - (ii) Qual é o desvio padrão de X , novamente supondo que a fração real de votos sim no plebiscito seja 0,65?
 - (iii) Agora, você coletou a amostra de 200 e descobriu que 115 pessoas na verdade votaram sim. Use o TLC para aproximar a probabilidade de que você poderia encontrar 115 ou menos votos sim em uma amostra aleatória de 200 se, na verdade, 65% da população votou sim.
 - (iv) Como você explicaria a relevância do número da parte (iii) para alguém que não tenha treinamento de estatística?
- 10 Antes de uma greve encerrar prematuramente a temporada da principal liga de beisebol em 1994, Tony Gwynn, do San Diego Padres, tinha 165 rebatidas em 419 oportunidades no bastão, para uma média de rebatidas de 0,394. Havia discussões sobre

o fato de Gwynn ser um possível batedor de 0,400 naquele ano. Essa questão pode ser formulada em termos da probabilidade de Gwynn conseguir uma rebatida em determinada oportunidade no bastão, chame-a de θ . Defina Y_i como o indicador de Bernoulli(θ) igual a um se Gwynn conseguir uma rebatida durante sua i^{a} oportunidade no bastão, e zero caso contrário. Assim, $Y_1, Y_2, \dots, Y_n$ é uma amostra aleatória de uma distribuição de Bernoulli(θ), onde θ é a probabilidade de sucesso, e $n = 419$.

Nossa melhor estimativa pontual de θ é a média de rebatidas de Gwynn, que é a proporção de sucessos: $\bar{y} = 0,394$. Usando o fato de que $\text{ep}(\bar{y}) = \sqrt{\bar{y}(1 - \bar{y})/n}$, construa um intervalo de confiança aproximado de 95% para θ , usando a distribuição normal padrão. Você diria que há fortes evidências contra Gwynn ser um possível rebatedor de 0,400? Explique.

- 11** Suponha que entre o primeiro e segundo ano na universidade, 400 estudantes foram selecionados aleatoriamente e receberam um subsídio da instituição para comprar um novo computador. Para o estudante i , y_i representa a mudança do GPA do primeiro ao segundo ano. Se a alteração média for $\bar{y} = 0,132$ com desvio padrão $s = 1,27$, a alteração média no GPA é estatisticamente maior do que zero?

Resumo de álgebra matricial

Este apêndice resume os conceitos de álgebra matricial, incluindo a álgebra da probabilidade, necessários para o estudo de modelos de regressão linear múltipla com matrizes do Apêndice E. Nenhum destes materiais é usado no texto principal.

D.1 Definições básicas

Definição D.1 (Matriz). Uma **matriz** é uma formação retangular de números. Mais precisamente, uma matriz $m \times n$ tem m linhas e n colunas. O inteiro positivo m é chamado de *dimensão da linha* e n é chamado de *dimensão da coluna*.

Utilizamos letras maiúsculas em negrito para indicar matrizes. Podemos escrever uma matriz $m \times n$ de forma genérica como

$$\mathbf{A} = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \vdots & & & & \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix},$$

onde a_{ij} representa o elemento na i^{a} linha e j^{a} coluna. Por exemplo, a_{25} representa o número na segunda linha e quinta coluna de $\mathbf{A}$. Um exemplo específico de matriz 2×3 é

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 7 \\ -4 & 5 & 0 \end{bmatrix}, \quad (\text{D.1})$$

onde $a_{13} = 7$. A abreviação $\mathbf{A} = [a_{ij}]$ é usada com frequência para definir operações matriciais.

Definição D.2 (Matriz quadrada). Uma **matriz quadrada** tem o mesmo número de linhas e de colunas. A dimensão de uma matriz quadrada é seu número de linhas e colunas.

Definição D.3 (Vetores).

(i) Uma matriz $1 \times m$ é chamada de **vetor linha** (da dimensão m) e pode ser escrita como $\mathbf{x} \equiv (x_1, x_2, \dots, x_m)$.

(ii) Uma matriz $n \times 1$ é chamada de **vetor coluna** e pode ser escrita como

$$\mathbf{y} \equiv \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

Definição D.4 (Matriz diagonal). Uma matriz quadrada $\mathbf{A}$ é uma **matriz diagonal** quando todos os seus elementos fora da diagonal forem zero, isto é, $a_{ij} = 0$ para todo $i \neq j$. Podemos sempre escrever uma matriz diagonal como

$$\mathbf{A} = \begin{bmatrix} a_{11} & 0 & 0 & \dots & 0 \\ 0 & a_{22} & 0 & \dots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \dots & a_{nn} \end{bmatrix}.$$

Definição D.5 (Matriz identidade e matriz nula).

(i) A **matriz identidade** $n \times n$, sinalizada por $\mathbf{I}$, ou às vezes $\mathbf{I}_n$ para enfatizar sua dimensão, é a matriz diagonal com unidade (um) em cada posição diagonal e zero no restante:

$$\mathbf{I} \equiv \mathbf{I}_n \equiv \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}.$$

(ii) A **matriz nula** $m \times n$, indicada por $\mathbf{0}$, é a matriz $m \times n$ com zero em todas as entradas. Ela não precisa ser uma matriz quadrada.

D.2 Operações matriciais

D.2a Soma de matrizes

Duas matrizes $\mathbf{A}$ e $\mathbf{B}$, cada uma com dimensão $m \times n$, podem ser somadas elemento a elemento: $\mathbf{A} + \mathbf{B} = [a_{ij} + b_{ij}]$. Mais precisamente

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2n} + b_{2n} \\ \vdots & & & \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \dots & a_{mn} + b_{mn} \end{bmatrix}.$$

Por exemplo,

$$\begin{bmatrix} 2 & -1 & 7 \\ -4 & 5 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & -4 \\ 4 & 2 & 3 \end{bmatrix} = \begin{bmatrix} 3 & -1 & 3 \\ 0 & 7 & 3 \end{bmatrix}.$$

Matrizes com dimensões diferentes não podem ser somadas.

D.2b Multiplicação escalar

Dado qualquer número real γ (frequentemente chamado de escalar), a **multiplicação escalar** é definida como $\gamma\mathbf{A} \equiv [\gamma a_{ij}]$, ou

$$\gamma\mathbf{A} = \begin{bmatrix} \gamma a_{11} & \gamma a_{12} & \dots & \gamma a_{1n} \\ \gamma a_{21} & \gamma a_{22} & \dots & \gamma a_{2n} \\ \vdots & & & \\ \gamma a_{m1} & \gamma a_{m2} & \dots & \gamma a_{mn} \end{bmatrix}.$$

Por exemplo, se $\gamma = 2$ e $\mathbf{A}$ é a matriz da equação (D.1), então

$$\gamma\mathbf{A} = \begin{bmatrix} 4 & -2 & 14 \\ -8 & 10 & 0 \end{bmatrix}.$$

D.2c Multiplicação de matrizes

Para multiplicar a matriz $\mathbf{A}$ pela matriz $\mathbf{B}$ e formar o produto $\mathbf{AB}$, a dimensão da *coluna* de $\mathbf{A}$ deve ser igual à dimensão da *linha* de $\mathbf{B}$. Portanto, seja $\mathbf{A}$ uma matriz $m \times n$ e $\mathbf{B}$ uma matriz $n \times p$. Assim, a **multiplicação de matrizes** é definida como

$$\mathbf{AB} = \left[\sum_{k=1}^n a_{ik} b_{kj} \right].$$

Em outras palavras, o $(i, j)^\circ$ elemento da nova matriz $\mathbf{AB}$ é obtido pela multiplicação de cada elemento da i° linha de $\mathbf{A}$ pelo elemento correspondente da j° coluna de $\mathbf{B}$ e pela soma desses n produtos. Um esquema pode ajudar a tornar esse processo mais claro:

$$\begin{array}{ccc} \mathbf{A} & \mathbf{B} & \mathbf{AB} \\ i^\circ \text{ linha} \rightarrow \begin{bmatrix} a_{i1} & a_{i2} & a_{i3} & \dots & a_{in} \end{bmatrix} & \begin{bmatrix} b_{1j} \\ b_{2j} \\ b_{3j} \\ \vdots \\ b_{nj} \end{bmatrix} & = \begin{bmatrix} \sum_{k=1}^n a_{ik} b_{kj} \end{bmatrix} \\ & \uparrow & \uparrow \\ & j^\circ \text{ coluna} & (i, j)^\circ \text{ elemento} \end{array}$$

onde, pela definição do operador de soma do Apêndice A,

$$\sum_{k=1}^n a_{ik} b_{kj} = a_{i1} b_{1j} + a_{i2} b_{2j} + \dots + a_{in} b_{nj}.$$

Por exemplo,

$$\begin{bmatrix} 2 & -1 & 0 \\ -4 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 6 & 0 \\ -1 & 2 & 0 & 1 \\ 3 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 12 & -1 \\ -1 & -2 & -24 & 1 \end{bmatrix}.$$

Também podemos multiplicar uma matriz e um vetor. Se $\mathbf{A}$ é uma matriz $n \times m$ e $\mathbf{y}$ é um vetor $m \times 1$, $\mathbf{Ay}$ será um vetor $n \times 1$. Se $\mathbf{x}$ for um vetor $1 \times n$, $\mathbf{xA}$ será um vetor $1 \times m$.

Soma de matrizes, multiplicação escalar e multiplicação de matrizes podem ser combinadas de diversas formas, e essas operações satisfazem várias regras que são familiares nas operações numéricas básicas. Na lista de propriedades a seguir, $\mathbf{A}$, $\mathbf{B}$ e $\mathbf{C}$ são matrizes com dimensões apropriadas para cada operação, e α e β são números reais. A maioria dessas propriedades é fácil de ilustrar a partir das definições.

Propriedades de operações matriciais. (1) $(\alpha + \beta)\mathbf{A} = \alpha\mathbf{A} + \beta\mathbf{A}$; (2) $\alpha(\mathbf{A} + \mathbf{B}) = \alpha\mathbf{A} + \alpha\mathbf{B}$; (3) $(\alpha\beta)\mathbf{A} = \alpha(\beta\mathbf{A})$; (4) $\alpha(\mathbf{AB}) = (\alpha\mathbf{A})\mathbf{B}$; (5) $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$; (6) $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$; (7) $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$; (8) $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$; (9) $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$; (10) $\mathbf{IA} = \mathbf{AI} = \mathbf{A}$; (11) $\mathbf{A} + \mathbf{0} = \mathbf{0} + \mathbf{A} = \mathbf{A}$; (12) $\mathbf{A} - \mathbf{A} = \mathbf{0}$; (13) $\mathbf{A0} = \mathbf{0A} = \mathbf{0}$; e (14) $\mathbf{AB} \neq \mathbf{BA}$, mesmo quando ambos os produtos forem definidos.

A última propriedade merece comentários adicionais. Se $\mathbf{A}$ for $n \times m$ e $\mathbf{B}$ for $m \times p$, então $\mathbf{AB}$ será definida, mas $\mathbf{BA}$ só seria definida se $n = p$ (a dimensão da linha de $\mathbf{A}$ fosse igual à dimensão da coluna de $\mathbf{B}$). Se $\mathbf{A}$ fosse $m \times n$ e $\mathbf{B}$ fosse $n \times m$, então $\mathbf{AB}$ e $\mathbf{BA}$ seriam ambas definidas, mas normalmente elas não são as mesmas; na verdade, elas terão dimensões diferentes, a não ser que $\mathbf{A}$ e $\mathbf{B}$ sejam duas matrizes quadradas. Mesmo quando $\mathbf{A}$ e $\mathbf{B}$ forem quadradas, $\mathbf{AB} \neq \mathbf{BA}$, exceto sob circunstâncias especiais.

D.2d Transposição

Definição D.6 (Transposição). Seja $\mathbf{A} = [a_{ij}]$ uma matriz $m \times n$. A **transposição** de $\mathbf{A}$, indicada por $\mathbf{A}'$ (chamada de **$\mathbf{A}$ primária**), é a matriz $n \times m$ obtida pela troca de linhas e colunas de $\mathbf{A}$. Podemos escrevê-la como $\mathbf{A}' \equiv [a_{ji}]$.

Por exemplo,

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 7 \\ -4 & 5 & 0 \end{bmatrix}, \quad \mathbf{A}' = \begin{bmatrix} 2 & -4 \\ -1 & 5 \\ 7 & 0 \end{bmatrix}.$$

Propriedades da transposição. (1) $(\mathbf{A}')' = \mathbf{A}$; (2) $(\alpha\mathbf{A})' = \alpha\mathbf{A}'$ para qualquer escalar α ; (3) $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$; (4) $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$, onde $\mathbf{A}$ é $m \times n$ e $\mathbf{B}$ é $n \times k$; (5) $\mathbf{x}'\mathbf{x} = \sum_{i=1}^n x_i^2$, onde $\mathbf{x}$ é um vetor $n \times 1$; e (6) se $\mathbf{A}$ for uma matriz $n \times k$ com linhas dadas pelos vetores $1 \times k$ $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$, de modo que possamos escrever

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{bmatrix},$$

então, $\mathbf{A}' = (\mathbf{a}'_1 \mathbf{a}'_2 \dots \mathbf{a}'_n)$.

Definição D.7 (Matriz simétrica). Uma matriz quadrada $\mathbf{A}$ será uma **matriz simétrica** se, e somente se, $\mathbf{A}' = \mathbf{A}$.

Se $\mathbf{X}$ for qualquer matriz $n \times k$, então $\mathbf{X}'\mathbf{X}$ será sempre definida e uma matriz simétrica, como podemos ver ao aplicar a primeira e a quarta propriedades de transposição (ver Problema 3).

D.2e Multiplicação de matriz particionada

Seja $\mathbf{A}$ uma matriz $n \times k$ com linhas dadas pelos vetores $1 \times k$ $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ e $\mathbf{B}$ uma matriz $n \times m$ com linhas dadas pelos vetores $1 \times m$ $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$:

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_n \end{bmatrix}.$$

Então,

$$\mathbf{A}'\mathbf{B} = \sum_{i=1}^n \mathbf{a}_i' \mathbf{b}_i,$$

onde, para cada i , $\mathbf{a}_i' \mathbf{b}_i$ é uma matriz $k \times m$. Dessa forma, $\mathbf{A}'\mathbf{B}$ pode ser escrita como a soma de n matrizes, cada uma delas $k \times m$. Como caso especial, temos

$$\mathbf{A}'\mathbf{A} = \sum_{i=1}^n \mathbf{a}_i' \mathbf{a}_i,$$

onde $\mathbf{a}_i' \mathbf{a}_i$ é uma matriz $k \times k$ para todo i .

Uma forma mais geral de multiplicação de matrizes particionadas se mantém quando tivermos matrizes $\mathbf{A}$ ($m \times n$) e $\mathbf{B}$ ($n \times p$) escritas como

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix},$$

onde $\mathbf{A}_{11}$ é $m_1 \times n_1$, $\mathbf{A}_{12}$ é $m_1 \times n_2$, $\mathbf{A}_{21}$ é $m_2 \times n_1$, $\mathbf{A}_{22}$ é $m_2 \times n_2$, $\mathbf{B}_{11}$ é $n_1 \times p_1$, $\mathbf{B}_{12}$ é $n_1 \times p_2$, $\mathbf{B}_{21}$ é $n_2 \times p_1$ e $\mathbf{B}_{22}$ é $n_2 \times p_2$. Naturalmente, $m_1 + m_2 = m$, $n_1 + n_2 = n$ e $p_1 + p_2 = p$.

Quando formamos o produto $\mathbf{AB}$, a expressão fica igual a quando as entradas são escalares:

$$\mathbf{AB} = \begin{pmatrix} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} \end{pmatrix}.$$

Note que cada uma das multiplicações da matriz que formam a partição à direita é bem definida porque as dimensões linha e coluna são compatíveis para multiplicação.

D.2f Traço

O traço de uma matriz é uma operação muito simples, definida apenas para matrizes *quadradas*.

Definição D.8 (Traço). Para qualquer matriz $n \times n$, o **traço de uma matriz** $\mathbf{A}$, indicado por $\text{tr}(\mathbf{A})$, será a soma de seus elementos diagonais. Matematicamente,

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^n a_{ii}.$$

Propriedades do Traço. (1) $\text{tr}(\mathbf{I}_n) = n$; (2) $\text{tr}(\mathbf{A}') = \text{tr}(\mathbf{A})$; (3) $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B})$; (4) $\text{tr}(\alpha \mathbf{A}) = \alpha \text{tr}(\mathbf{A})$ para qualquer escalar α ; e (5) $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$, onde $\mathbf{A}$ é $m \times n$ e $\mathbf{B}$ é $n \times m$.

D.2g Inverso

A noção de inverso de uma matriz é muito importante para matrizes quadradas.

Definição D.9 (Inverso). Uma matriz $n \times n$ $\mathbf{A}$ terá um **inverso**, indicado por $\mathbf{A}^{-1}$, desde que $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$ e $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_n$. Neste caso, $\mathbf{A}$ é dita *invertível* ou *não singular*. No caso contrário, é chamada de *não invertível* ou *singular*.

Propriedades do inverso. (1) Se existir um inverso, ele será único; (2) $(\alpha\mathbf{A})^{-1} = (1/\alpha)\mathbf{A}^{-1}$, se $\alpha \neq 0$ e $\mathbf{A}$ for invertível; (3) $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$, se $\mathbf{A}$ e $\mathbf{B}$ forem ambos $n \times n$ e invertíveis; e (4) $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$.

Não nos preocuparemos com os mecanismos de cálculo de uma matriz inversa. Qualquer texto sobre álgebra matricial contém exemplos detalhados destes cálculos.

D.3 Independência linear e posto matricial

Para um conjunto de vetores que têm a mesma dimensão, é importante saber se um vetor pode ser expresso como uma combinação linear dos vetores restantes.

Definição D.10 (Independência linear). Defina $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_r\}$ como um conjunto de vetores $n \times 1$. Eles serão **vetores linearmente independentes** se, e somente se,

$$\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \dots + \alpha_r\mathbf{x}_r = \mathbf{0} \quad (\text{D.2})$$

indicar que $\alpha_1 = \alpha_2 = \dots = \alpha_r = 0$. Se (D.2) se mantiver para um conjunto de escalares que não sejam todos zero, então, $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_r\}$ será *linearmente dependente*.

A afirmação de que $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_r\}$ é linearmente dependente é equivalente a dizer que pelo menos um vetor nesse conjunto pode ser escrito como uma combinação linear dos outros.

Definição D.11 (Posto)

(i) Seja $\mathbf{A}$ uma matriz $n \times m$. O **posto de uma matriz $\mathbf{A}$** , sinalizado por **posto ($\mathbf{A}$)**, é o número máximo de colunas linearmente independentes de $\mathbf{A}$.

(ii) Se $\mathbf{A}$ é $n \times m$ e **posto ($\mathbf{A}$)** = m , então $\mathbf{A}$ tem um *posto de coluna completo*.

Se $\mathbf{A}$ for $n \times m$, seu posto pode ser, no máximo, m . Uma matriz tem posto de coluna completo se suas colunas formarem um conjunto linearmente independente. Por exemplo, a matriz 3×2

$$\begin{bmatrix} 1 & 3 \\ 2 & 6 \\ 0 & 0 \end{bmatrix}$$

pode ter, no máximo, posto dois. Na verdade, seu posto é somente um porque a segunda coluna é três vezes a primeira coluna.

Propriedades de Posto. (1) **posto($\mathbf{A}'$)** = **posto($\mathbf{A}$)**; (2) Se $\mathbf{A}$ for $n \times k$, então **posto ($\mathbf{A}$)** $\leq \min(n, k)$; e (3) Se $\mathbf{A}$ for $k \times k$ e **posto($\mathbf{A}$)** = k , então $\mathbf{A}$ será invertível.

D.4 Formas quadráticas e matrizes positivas definidas

Definição D.12 (Forma quadrática). Seja $\mathbf{A}$ uma matriz $n \times n$ simétrica. A **forma quadrática** associada com a matriz $\mathbf{A}$ é a função de valor real definida por todos os vetores $\mathbf{x} \ n \times 1$:

$$f(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^n a_{ii}x_i^2 + 2 \sum_{i=1}^n \sum_{j>i}^n a_{ij}x_i x_j.$$

Definição D.13 (Positiva definida e positiva semidefinida)

(i) Uma matriz simétrica $\mathbf{A}$ é dita **positiva definida (p.d)** se

$$\mathbf{x}'\mathbf{A}\mathbf{x} > 0 \text{ para todo vetor } \mathbf{x} \ n \times 1, \text{ exceto } \mathbf{x} = \mathbf{0}.$$

(ii) Uma matriz simétrica $\mathbf{A}$ é **positiva semidefinida (p.s.d)** se

$$\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0 \text{ para todo vetor } \mathbf{x} \ n \times 1.$$

Se uma matriz é positiva definida ou positiva semidefinida, ela é automaticamente suposta como simétrica.

Propriedades de matrizes positivas definidas e positivas semidefinidas. (1) Uma matriz p.d. tem elementos diagonais estritamente positivos, enquanto uma matriz p.s.d. tem elementos diagonais não negativos; (2) Se $\mathbf{A}$ é p.d, então $\mathbf{A}^{-1}$ existe e é p.d; (3) Se $\mathbf{X}$ é $n \times k$, então $\mathbf{X}'\mathbf{X}$ e $\mathbf{X}\mathbf{X}'$ são p.s.d.; e (4) Se $\mathbf{X}$ é $n \times k$ e $\text{posto}(\mathbf{X}) = k$, então $\mathbf{X}'\mathbf{X}$ é p.d. (e, portanto, não singular).

D.5 Matrizes idempotentes

Definição D.14 (Matriz idempotente). Seja $\mathbf{A}$ uma matriz simétrica $n \times n$. Dessa forma, $\mathbf{A}$ é dita como uma **matriz idempotente** se, e somente se, $\mathbf{A}\mathbf{A} = \mathbf{A}$.

Por exemplo,

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

é uma matriz idempotente, como verificado pela multiplicação direta.

Propriedades de matrizes idempotentes. Defina $\mathbf{A}$ como uma matriz idempotente $n \times n$. (1) $\text{posto}(\mathbf{A}) = \text{tr}(\mathbf{A})$ e (2) $\mathbf{A}$ é positiva semidefinida.

Podemos construir matrizes idempotentes muito geralmente. Seja $\mathbf{X}$ uma matriz $n \times n$ com $\text{posto}(\mathbf{X}) = k$. Defina

$$\mathbf{P} \equiv \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

$$\mathbf{M} \equiv \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{I}_n - \mathbf{P}.$$

Assim, $\mathbf{P}$ e $\mathbf{M}$ são matrizes simétricas idempotentes com $\text{posto}(\mathbf{P}) = k$ e $\text{posto}(\mathbf{M}) = n - k$. Os postos são obtidos mais facilmente usando a Propriedade 1: $\text{tr}(\mathbf{P}) = \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}]$ (da Propriedade 5 do traço) $= \text{tr}(\mathbf{I}_k) = k$ (a partir da Propriedade 1 do traço). Isso resulta facilmente em $\text{tr}(\mathbf{M}) = \text{tr}(\mathbf{I}_n) - \text{tr}(\mathbf{P}) = n - k$.

D.6 Diferenciação das formas linear e quadrática

Para determinado vetor $\mathbf{a} \, n \times 1$, considere a função linear definida por

$$f(\mathbf{x}) = \mathbf{a}'\mathbf{x},$$

para todos os vetores $\mathbf{x} \, n \times 1$. A derivada de f em relação a $\mathbf{x}$ é o vetor $1 \times n$ das derivadas parciais, que é simplesmente

$$\partial f(\mathbf{x})/\partial \mathbf{x} = \mathbf{a}'.$$

Para qualquer matriz simétrica $\mathbf{A} \, n \times n$, defina a forma quadrática

$$g(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x}.$$

Assim,

$$\partial g(\mathbf{x})/\partial \mathbf{x} = 2\mathbf{x}'\mathbf{A},$$

que é um vetor $1 \times n$.

D.7 Momentos e distribuições de vetores aleatórios

Com o objetivo de derivar o valor esperado e a variância dos estimadores de MQO usando matrizes, precisamos definir o valor esperado e a variância de um **vetor aleatório**. Como o próprio nome sugere, um vetor aleatório é simplesmente um vetor de variáveis aleatórias. Também precisamos definir a distribuição normal multivariada. Esses conceitos são somente extensões daqueles que foram tratados no Apêndice B.

D.7a Valor esperado

Definição D.15 (Valor esperado)

(i) Se $\mathbf{y}$ é um vetor aleatório $n \times 1$, o **valor esperado** de $\mathbf{y}$, sinalizado por $E(\mathbf{y})$, é o vetor dos valores esperados: $E(\mathbf{y}) = [E(y_1), E(y_2), \dots, E(y_n)]'$.

(ii) Se $\mathbf{Z}$ é uma matriz aleatória $n \times m$, $E(\mathbf{Z})$ é a matriz $n \times m$ dos valores esperados: $E(\mathbf{Z}) = [E(z_{ij})]$.

Propriedades do valor esperado. (1) Se $\mathbf{A}$ é uma matriz $m \times n$ e $\mathbf{b}$ é um vetor $n \times 1$, ambos não aleatórios, então $E(\mathbf{A}\mathbf{y} + \mathbf{b}) = \mathbf{A}E(\mathbf{y}) + \mathbf{b}$; e (2) Se $\mathbf{A}$ é $p \times n$ e $\mathbf{B}$ é $m \times k$, ambas não aleatórias, $E(\mathbf{AZB}) = \mathbf{A}E(\mathbf{Z})\mathbf{B}$.

D.7b Matriz de variância-covariância

Definição D.16 (Matriz de variância-covariância). Se $\mathbf{y}$ for um vetor aleatório $n \times 1$, sua **matriz de variância-covariância**, indicada por $\text{Var}(\mathbf{y})$, será definida como

$$\text{Var}(\mathbf{y}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{bmatrix},$$

onde $\sigma_j^2 = \text{Var}(y_j)$ e $\sigma_{ij} = \text{Cov}(y_i, y_j)$. Em outras palavras, a matriz de variância-covariância tem as variâncias de cada elemento de $\mathbf{y}$ em sua diagonal, com os termos de covariância fora das diagonais. Como $\text{Cov}(y_i, y_j) = \text{Cov}(y_j, y_i)$, verifica-se imediatamente que uma matriz de variância-covariância é simétrica.

Propriedades da variância. (1) Se $\mathbf{a}$ é um vetor não aleatório $n \times 1$, $\text{Var}(\mathbf{a}'\mathbf{y}) = \mathbf{a}'[\text{Var}(\mathbf{y})\mathbf{a}] \geq 0$; (2) Se $\text{Var}(\mathbf{a}'\mathbf{y}) > 0$ para todo $\mathbf{a} \neq \mathbf{0}$, $\text{Var}(\mathbf{y})$ é positiva definida; (3) $\text{Var}(\mathbf{y}) = E[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})']$, onde $\boldsymbol{\mu} = E(\mathbf{y})$; (4) Se os elementos de $\mathbf{y}$ forem não correlacionados, $\text{Var}(\mathbf{y})$ será uma matriz diagonal. Se, além disso, $\text{Var}(y_j) = \sigma^2$ para $j = 1, 2, \dots, n$, então $\text{Var}(\mathbf{y}) = \sigma^2 \mathbf{I}_n$; e (5) Se $\mathbf{A}$ é uma matriz não aleatória $m \times n$ e $\mathbf{b}$ é um vetor não aleatório $n \times 1$, $\text{Var}(\mathbf{A}\mathbf{y} + \mathbf{b}) = \mathbf{A}[\text{Var}(\mathbf{y})]\mathbf{A}'$.

D.7c Distribuição normal multivariada

A distribuição normal de uma variável aleatória foi discutida amplamente no Apêndice B. Precisamos estender a distribuição normal a vetores aleatórios. Não forneceremos uma expressão para a função de distribuição de probabilidade, já que não precisamos de uma. É importante saber que um vetor aleatório normal multivariado é completamente caracterizado por sua média e sua matriz de variância-covariância. Assim, se $\mathbf{y}$ for um vetor aleatório normal multivariado $n \times 1$ com média $\boldsymbol{\mu}$ e matriz de variância-covariância $\boldsymbol{\Sigma}$, escreveremos $\mathbf{y} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Expressaremos agora várias propriedades úteis da **distribuição normal multivariada**.

Propriedades da distribuição normal multivariada. (1) Se $\mathbf{y} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, cada elemento de $\mathbf{y}$ será normalmente distribuído; (2) Se $\mathbf{y} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, y_i e y_j , quaisquer dois elementos de $\mathbf{y}$ serão independentes se, e somente se, forem não correlacionados, isto é $\sigma_{ij} = 0$; (3) Se $\mathbf{y} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{A}\mathbf{y} + \mathbf{b} \sim \text{Normal}(\mathbf{A}\boldsymbol{\mu}, \mathbf{b} + \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$, onde $\mathbf{A}$ e $\mathbf{B}$ são não aleatórias; (4) Se $\mathbf{y} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, então, para matrizes não aleatórias $\mathbf{A}$ e $\mathbf{B}$, $\mathbf{A}\mathbf{y}$ e $\mathbf{B}\mathbf{y}$ serão independentes se, e somente se, $\mathbf{A}\boldsymbol{\Sigma}\mathbf{B}' = \mathbf{0}$. Em particular, se $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_n$, então $\mathbf{A}\mathbf{B}' = \mathbf{0}$ é necessário e suficiente para a independência de $\mathbf{A}\mathbf{y}$ e $\mathbf{B}\mathbf{y}$; (5) Se $\mathbf{y} \sim \text{Normal}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, $\mathbf{A}$ é uma matriz não aleatória $k \times n$ e $\mathbf{B}$ é uma matriz simétrica idempotente $n \times n$, $\mathbf{A}\mathbf{y}$ e $\mathbf{y}'\mathbf{B}\mathbf{y}$ serão independentes se, e somente se, $\mathbf{A}\mathbf{B} = \mathbf{0}$; e (6) Se $\mathbf{y} \sim \text{Normal}(\mathbf{0}, \boldsymbol{\Sigma}^2 \mathbf{I}_n)$ e $\mathbf{A}$ e $\mathbf{B}$ são matrizes simétricas não aleatórias e idempotentes, $\mathbf{y}'\mathbf{A}\mathbf{y}$ e $\mathbf{y}'\mathbf{B}\mathbf{y}$ serão independentes se, e somente se, $\mathbf{A}\mathbf{B} = \mathbf{0}$.

D.7d Distribuição qui-quadrada

No Apêndice B, definimos uma **variável aleatória qui-quadrada** como a soma dos *quadrados* de variáveis aleatórias normais padrão independentes. Em notação vetorial, se $\mathbf{u} \sim \text{Normal}(\mathbf{0}, \mathbf{I}_n)$, então $\mathbf{u}'\mathbf{u} \sim \chi_n^2$.

Propriedades da distribuição qui-quadrada. (1) Se $\mathbf{u} \sim \text{Normal}(\mathbf{0}, \mathbf{I}_n)$ e $\mathbf{A}$ é uma matriz idempotente e simétrica $n \times n$ com $\text{posto}(\mathbf{A}) = q$, então $\mathbf{u}'\mathbf{A}\mathbf{u} \sim \chi_q^2$; (2) Se $\mathbf{u} \sim \text{Normal}(\mathbf{0}, \mathbf{I}_n)$ e $\mathbf{A}$ e $\mathbf{B}$ são matrizes idempotentes simétricas $n \times n$ de modo que $\mathbf{A}\mathbf{B} = \mathbf{0}$, $\mathbf{u}'\mathbf{A}\mathbf{u}$ e $\mathbf{u}'\mathbf{B}\mathbf{u}$ serão variáveis aleatórias independentes e qui-quadradas; e (3) Se $\mathbf{z} \sim \text{Normal}(\mathbf{0}, \mathbf{C})$, em que $\mathbf{C}$ é uma matriz não singular $m \times m$, $\mathbf{z}'\mathbf{C}^{-1}\mathbf{z} \sim \chi_m^2$.

D.7e Distribuição t

Também definimos a **distribuição t** no Apêndice B. Agora, acrescentamos uma importante propriedade.

Propriedade da distribuição t . Se $\mathbf{u} \sim \text{Normal}(\mathbf{0}, \mathbf{I}_n)$, $\mathbf{c}$ for um vetor não aleatório $n \times 1$, $\mathbf{A}$ for uma matriz não aleatória, simétrica $n \times n$ e idempotente com posto q , e $\mathbf{Ac} = \mathbf{0}$, então $\{\mathbf{c}'\mathbf{u}/(\mathbf{c}'\mathbf{c})^{1/2}\}/(\mathbf{u}'\mathbf{A}\mathbf{u}/q)^{1/2} \sim t_q$.

D.7f Distribuição F

Lembre-se de que uma **variável aleatória F** é obtida ao selecionar duas variáveis aleatórias qui-quadradas *independentes* e ao encontrar a razão de cada uma delas, padronizada por graus de liberdade.

Propriedade da distribuição F . Se $\mathbf{u} \sim \text{Normal}(\mathbf{0}, \mathbf{I}_n)$ e $\mathbf{A}$ e $\mathbf{B}$ forem matrizes $n \times n$ não aleatórias, simétricas e idempotentes com $\text{posto}(\mathbf{A}) = k_1$, $\text{posto}(\mathbf{B}) = k_2$ e $\mathbf{AB} = \mathbf{0}$, então $(\mathbf{u}'\mathbf{A}\mathbf{u}/k_1)/(\mathbf{u}'\mathbf{B}\mathbf{u}/k_2) \sim F_{k_1, k_2}$.

Resumo

Este apêndice contém uma forma condensada das informações básicas necessárias para estudar o modelo linear clássico utilizando matrizes. Embora seja independente, este material foi elaborado principalmente como uma revisão para leitores que já estão familiarizados com álgebra matricial e estatística multivariada, e será usado amplamente no Apêndice E.

Termos-chave

Distribuição normal multivariada	Matriz nula	Variável aleatória F
Distribuição t	Matriz quadrada	Variável aleatória qui-quadrada
Forma quadrática	Matriz simétrica	Vetor aleatório
Inverso	Multiplicação de matrizes	Vetor coluna
Matriz	Multiplicação escalar	Vetor linha
Matriz de variância-covariância	Positiva definida (p.d)	Vetores linearmente independentes
Matriz diagonal	Positiva semidefinida (p.s.d)	
Matriz idempotente	Posto de uma matriz $\mathbf{A}$	
Matriz identidade	Traço de uma matriz	
	Transposição	
	Valor esperado	

Problemas

- 1** (i) Encontre o produto $\mathbf{AB}$ usando

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 7 \\ -4 & 5 & 0 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 0 & 1 & 6 \\ 1 & 8 & 0 \\ 3 & 0 & 0 \end{bmatrix}.$$

- (ii) O produto $\mathbf{BA}$ existe?

- 2 Se $\mathbf{A}$ e $\mathbf{B}$ são matrizes diagonais $n \times n$, mostre que $\mathbf{AB} = \mathbf{BA}$.
- 3 Seja $\mathbf{X}$ qualquer matriz $n \times k$. Mostre que $\mathbf{X}'\mathbf{X}$ é uma matriz simétrica.
- 4 (i) Use as propriedades de traço para defender que $\text{tr}(\mathbf{A}'\mathbf{A}) = \text{tr}(\mathbf{AA}')$ em qualquer matriz $\mathbf{A}$ $n \times m$.
(ii) Para $\mathbf{A} = \begin{bmatrix} 2 & 0 & -1 \\ 0 & 3 & 0 \end{bmatrix}$, verifique se $\text{tr}(\mathbf{A}'\mathbf{A}) = \text{tr}(\mathbf{AA}')$.
- 5 (i) Use a definição de inverso para provar o seguinte: se $\mathbf{A}$ e $\mathbf{B}$ são matrizes $n \times n$ não singulares, então $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.
(ii) Se $\mathbf{A}$, $\mathbf{B}$ e $\mathbf{C}$ são matrizes $n \times n$ não singulares, descubra $(\mathbf{ABC})^{-1}$ em termos de $\mathbf{A}^{-1}$, $\mathbf{B}^{-1}$ e $\mathbf{C}^{-1}$.
- 6 (i) Mostre que, se $\mathbf{A}$ for uma matriz positiva definida $n \times n$ simétrica, então $\mathbf{A}$ deverá ter elementos diagonais estritamente positivos.
(ii) Desenhe uma matriz simétrica 2×2 com elementos diagonais estritamente positivos que *não* seja positiva definida.
- 7 Seja $\mathbf{A}$ uma matriz positiva definida $n \times n$ simétrica. Mostre que, se $\mathbf{P}$ for uma matriz não singular $n \times n$, $\mathbf{P}'\mathbf{AP}$ será positiva definida.
- 8 Prove a Propriedade 5 das variâncias para vetores usando a Propriedade 3.
- 9 Defina $\mathbf{a}$ como um vetor não aleatório $n \times 1$ e $\mathbf{u}$ como um vetor aleatório $n \times 1$ com $E(\mathbf{uu}') = \mathbf{I}_n$. Mostre que $E[\text{tr}(\mathbf{auu}'\mathbf{a}')] = \sum_{i=1}^n a_i^2$.
- 10 Partindo das propriedades da distribuição qui-quadrada listadas no texto, mostre que essas propriedades, ao lado da definição de uma variável aleatória F , implicam a propriedade fixa da distribuição F (em relação às razões das formas quadráticas).
- 11 Seja $\mathbf{X}$ uma matriz $n \times k$ particionada como

$$\mathbf{X} = (\mathbf{X}_1 \ \mathbf{X}_2),$$

em que $\mathbf{X}_1$ é $n \times k_1$ e $\mathbf{X}_2$ é $n \times k_2$.

(i) Mostre que

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix}.$$

Quais são as dimensões de cada uma das matrizes?

(ii) Seja $\mathbf{b}$ um vetor $k \times 1$, particionado como

$$\mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix},$$

em que $\mathbf{b}_1$ é $k_1 \times 1$ e $\mathbf{b}_2$ é $k_2 \times 1$. Mostre que

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \begin{pmatrix} (\mathbf{X}_1'\mathbf{X}_1)\mathbf{b}_1 + (\mathbf{X}_1'\mathbf{X}_2)\mathbf{b}_2 \\ (\mathbf{X}_2'\mathbf{X}_1)\mathbf{b}_1 + (\mathbf{X}_2'\mathbf{X}_2)\mathbf{b}_2 \end{pmatrix}.$$

0 modelo de regressão linear em forma de matriz

Este apêndice deriva vários resultados para a estimação de mínimos quadrados ordinários do modelo de regressão linear múltiplo, utilizando notação de matriz e álgebra matricial (veja um resumo sobre o assunto no Apêndice D).

E-1 Modelo e estimação de mínimos quadrados ordinários

Por meio deste apêndice, usamos o t subscrito às observações do índice e um n para denotar o tamanho do amostra. É útil escrever o modelo de regressão linear múltipla com parâmetros k , como se segue:

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \dots + \beta_k x_{tk} + u_t, t = 1, 2, \dots, n, \quad (\text{E.1})$$

em que y_t é a variável dependente para observação t e x_{tj} , $j = 1, 2, \dots, k$, são as variáveis independentes. Como de hábito, β_0 é o intercepto e $\beta_1, \dots, \beta_k$ denota os parâmetros de inclinação.

Para cada t , defina um vetor $1 \times (k + 1)$, $\mathbf{x}_t = (1, x_{t1}, \dots, x_{tk})$, e considere que $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)'$ seja o vetor $(k + 1) \times 1$ de todos os parâmetros. Depois, podemos escrever (E.1) como

$$y_t = \mathbf{x}_t \boldsymbol{\beta} + u_t, t = 1, 2, \dots, n. \quad (\text{E.2})$$

[Alguns autores preferem definir $\mathbf{x}_t$ como um vetor de coluna, caso $\mathbf{x}_t$ seja substituído por $\mathbf{x}_t'$ em (E.2). Matematicamente, faz mais sentido defini-lo como um vetor coluna]. Podemos escrever (E.2) em notação matricial integral ao definir de forma adequada os vetores de dados e matrizes. Considere que $\mathbf{y}$ denote o vetor de observações $n \times 1$ em y : o elemento t^{th} de $\mathbf{y}$ é y_t . Considere que $\mathbf{X}$ seja o vetor de observações $n \times (k + 1)$ sobre as variáveis explicativas. Em outras palavras, a coluna t^{th} de $\mathbf{X}$ consiste no vetor $\mathbf{x}_t$. Descrito em detalhes,

$$\mathbf{X}_{n \times (k+1)} \equiv \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}.$$

Por fim, consideremos $\mathbf{u}$ o vetor de erros ou interferências não observáveis $n \times 1$. Sendo assim, podemos escrever (E.2) para todas as observações n em **notação de matriz**:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \quad (\text{E.3})$$

Lembre-se, já que $\mathbf{X}$ é $n \times (k+1)$ e $\boldsymbol{\beta}$ é $(k+1) \times 1$, $\mathbf{X}\boldsymbol{\beta}$ é $n \times 1$.

A estimação de $\boldsymbol{\beta}$ funciona por meio da minimização da soma de resíduos quadrados, conforme a Seção 3.2. Defina a soma da função de resíduos quadrados para qualquer vetor $\mathbf{b}$ de parâmetro possível $(k+1) \times 1$, conforme

$$\text{SQR}(\mathbf{b}) \equiv \sum_{t=1}^n (y_t - \mathbf{x}_t \mathbf{b})^2.$$

O vetor de estimativas de mínimos quadrados ordinários $(k+1) \times 1$, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)'$, minimiza $\text{SQR}(\mathbf{b})$ sobre todos os possíveis vetores $\mathbf{b}$ $(k+1) \times 1$. Isso é um problema que ocorre em cálculo multivariável. Para que $\hat{\boldsymbol{\beta}}$ minimize a soma dos resíduos quadrados, é necessário resolver a **condição de primeira ordem**

$$\partial \text{SQR}(\hat{\boldsymbol{\beta}}) / \partial \mathbf{b} \equiv 0. \quad (\text{E.4})$$

Levando em conta o fato de que o derivativo de $(y_t - \mathbf{x}_t \mathbf{b})^2$ em relação a $\mathbf{b}$ é o $1 \times (k+1)$ vetor $-2(y_t - \mathbf{x}_t \mathbf{b})\mathbf{x}_t$, (E.4) é equivalente a

$$\sum_{t=1}^n \mathbf{x}_t' (y_t - \mathbf{x}_t \hat{\boldsymbol{\beta}}) \equiv 0. \quad (\text{E.5})$$

(Dividimos por -2 e fizemos a transposição). Podemos escrever essa condição de primeira ordem como

$$\begin{aligned} \sum_{t=1}^n (y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t1} - \dots - \hat{\beta}_k x_{tk}) &= 0 \\ \sum_{t=1}^n x_{t1} (y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t1} - \dots - \hat{\beta}_k x_{tk}) &= 0 \\ \vdots \\ \sum_{t=1}^n x_{tk} (y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t1} - \dots - \hat{\beta}_k x_{tk}) &= 0, \end{aligned}$$

que é idêntica às condições de primeira ordem na equação (3.13). Queremos escrevê-las na forma de matriz para tornar mais fácil de manipulá-las. Usando a fórmula para multiplicação particionada do Apêndice D, vemos que (E.5) é equivalente a

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0} \quad (\text{E.6})$$

ou

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}. \quad (\text{E.7})$$

Pode ser mostrado que (E.7) sempre teve ao menos uma solução. Múltiplas soluções não nos ajudam caso estejamos procurando um conjunto único de estimativas MQO, visto nosso conjunto de dados. Presumindo que a matriz simétrica $(k + 1) \times (k + 1)$ $\mathbf{X}'\mathbf{X}$ não seja singular, podemos multiplicar previamente ambos os lados de (E.7) por $(\mathbf{X}'\mathbf{X})^{-1}$ para resolver o estimador MQO $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \quad (\text{E.8})$$

Essa é a fórmula essencial para a análise de matriz do modelo de regressão linear múltipla. A hipótese de que $\mathbf{X}'\mathbf{X}$ seja invertível equivale à hipótese de classificação $\text{rank}(\mathbf{X}) = (k + 1)$, o que significa que as colunas de $\mathbf{X}$ devem ser linearmente independentes. Essa é a versão matricial de RLM.3 do Capítulo 3.

Antes de continuarmos, é necessário deixar um alerta sobre (E.8). É tentador simplificar a fórmula para $\hat{\boldsymbol{\beta}}$, conforme segue:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}^{-1}(\mathbf{X}')^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}^{-1}\mathbf{y}.$$

A falha nesse raciocínio é que $\mathbf{X}$ geralmente não é uma matriz quadrada, ou seja, não pode ser invertida. Em outras palavras, não podemos escrever $(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{X}^{-1}(\mathbf{X}')^{-1}$ a menos que $n = (k + 1)$, um caso que quase nunca surge na prática.

Os vetores de valores ajustados e os resíduos de MQO $n \times 1$ são dados por

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}, \text{ respectivamente.}$$

Com (E.6) e a definição de $\hat{\mathbf{u}}$, podemos ver que a condição de primeira ordem para $\hat{\boldsymbol{\beta}}$ é a mesma que

$$\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}. \quad (\text{E.9})$$

Já que a primeira coluna de $\mathbf{X}$ é composta inteiramente de uns, (E.9) implica que os resíduos MQO sempre somatizam zero quando um intercepto é incluído na equação e que a covariância de amostra entre cada variável independente e os resíduos MQO é zero (discutimos ambas as propriedades no Capítulo 3).

A soma dos resíduos quadrados pode ser escrita como

$$\text{SQR} = \sum_{i=1}^n \hat{u}_i^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \quad (\text{E.10})$$

Todas as propriedades algébricas do Capítulo 3 podem ser derivadas usando álgebra matricial. Por exemplo, podemos mostrar que a soma total de quadrados é igual à soma explicada de quadrados, mais a soma dos resíduos quadrados [ver (3.27)]. O uso das matrizes não fornece uma prova mais simples do que a notação somatória, dessa forma, não fornecemos outra derivação.

A abordagem de matriz para regressão múltipla pode ser usada como base para a interpretação geométrica da regressão. Isso envolve conceitos matemáticos ainda mais avançados do que os abordados no Apêndice D [Ver Goldberger (1991) ou Greene (1997)].

E.1a Teorema Frisch-Waugh

Na Seção 3.2, descrevemos uma interpretação “parcialmente ausente” das estimativas de mínimos quadrados ordinários. Podemos estabelecer a interpretação parcialmente ausente de forma muito geral, usando notação matricial. Dividindo a matriz $\mathbf{X} \, n \times (k + 1)$ como

$$\mathbf{X} = (\mathbf{X}_1 | \mathbf{X}_2),$$

em que $\mathbf{X}_1$ é $n \times (k_1 + 1)$ e inclui o intercepto – mesmo que não seja exigido para que o resultado se mantenha – e $\mathbf{X}_2$ é $n \times k_2$. Ainda supomos que $\mathbf{X}$ tenha classificado $k + 1$, o que significa que $\mathbf{X}_1$ classificou $k_1 + 1$ e que $\mathbf{X}_2$ classificou k_2 .

Considere as estimativas MQO $\hat{\beta}_1$ e $\hat{\beta}_2$ da (longa) regressão

$$\mathbf{y} \text{ em } \mathbf{X}_1, \mathbf{X}_2.$$

Como sabemos, os coeficientes de regressão múltipla em $\mathbf{X}_2$, $\hat{\beta}_2$, geralmente diferem de $\tilde{\beta}_2$ da regressão $\mathbf{y}$ em $\mathbf{X}_2$. Uma forma de descrever a diferença é compreender que podemos obter $\hat{\beta}_2$ de uma regressão menor, mas que primeiro devemos “parcelar de forma ausente” $\mathbf{X}_1$ de $\mathbf{X}_2$. Considere o seguinte método, de duas etapas:

(i) Regrida (cada coluna de) $\mathbf{X}_2$ em $\mathbf{X}_1$ e obtenha a matriz dos resíduos, digamos $\ddot{\mathbf{X}}_2$. Podemos escrever $\ddot{\mathbf{X}}_2$ como

$$\ddot{\mathbf{X}}_2 = [\mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1']\mathbf{X}_2 = (\mathbf{I}_n - \mathbf{P}_1)\mathbf{X}_2 = \mathbf{M}_1\mathbf{X}_2,$$

em que $\mathbf{P}_1 = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$ e $\mathbf{M}_1 = \mathbf{I}_n - \mathbf{P}_1$ são $n \times n$ matrizes idempotentes simétricas.

(ii) Regrida $\mathbf{y}$ em $\ddot{\mathbf{X}}_2$ e preveja o vetor $k_2 \times 1$ de coeficiente $\hat{\beta}_2$.

O teorema de Frisch-Waugh (FW) declara que

$$\hat{\beta}_2 = \hat{\beta}_2.$$

Consideravelmente, o teorema FW em geral não diz nada sobre a igualdade das estimativas de regressão longa, $\hat{\beta}_2$, e sobre as da regressão curta, $\tilde{\beta}_2$. Geralmente, $\hat{\beta}_2 \neq \tilde{\beta}_2$. Contudo, se $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$, então $\ddot{\mathbf{X}}_2 = \mathbf{M}_1\mathbf{X}_2$, em cujo caso $\hat{\beta}_2 = \tilde{\beta}_2$; logo, $\hat{\beta}_2 = \tilde{\beta}_2$ resulta de FW. Também vale a pena ressaltar que vamos obter $\hat{\beta}_2$ se também favorecermos $\mathbf{X}_1$ em detrimento de $\mathbf{y}$. Em outras palavras, consideremos $\ddot{\mathbf{y}}$ os resíduos de regredir $\mathbf{y}$ em $\mathbf{X}_1$, de forma que

$$\ddot{\mathbf{y}} = \mathbf{M}_1\mathbf{y}.$$

Logo, $\hat{\beta}_2$ é obtido da regressão de $\ddot{\mathbf{y}}$ em $\ddot{\mathbf{X}}_2$. É importante entender que não é suficiente apenas favorecer $\mathbf{X}_1$ em detrimento de $\mathbf{y}$. Um passo importante é favorecer $\mathbf{X}_1$ em detrimento de $\mathbf{X}_2$. O Problema 6, no final deste capítulo, pede que você derive o teorema FW e investigue algumas questões relacionadas.

Outro resultado algébrico útil ocorre quando regredimos $\ddot{\mathbf{y}}$ em $\ddot{\mathbf{X}}_2$ e poupamos os resíduos, por exemplo, $\ddot{\mathbf{u}}$, estes são idênticos aos resíduos MQO da regressão (longa) original:

$$\ddot{\mathbf{y}} = \ddot{\mathbf{X}}_2\hat{\beta}_2 = \ddot{\mathbf{u}} = \mathbf{u} = \mathbf{y} - \mathbf{X}_1\hat{\beta}_1 - \mathbf{X}_2\hat{\beta}_2,$$

em que usamos o resultado FW $\tilde{\beta}_2 = \hat{\beta}_2$. Não obtemos os resíduos MQO originais se regredirmos $\mathbf{y}$ em $\ddot{\mathbf{X}}_2$ (mas obtemos $\tilde{\beta}_2$).

Antes dos computadores eficazes, o resultado Frisch-Waugh era às vezes usado como dispositivo computacional. Atualmente, o resultado possui interesse mais teórico, e é muito útil para compreender os mecanismos de MQO. Por exemplo, lembre-se de que no Capítulo 10 usamos o teorema FW para estabelecer que adicionar uma tendência temporal a uma regressão múltipla é algebricamente equivalente a primeiro “detrend” linearmente todas as variáveis explicativas antes de estabelecer a regressão. O teorema FW também foi usado no Capítulo 14 para estabelecer que o estimador de efeitos fixos, que apresentamos como obtido de dados temporais reduzidos, também pode ser obtido da regressão de variável *dummy* (longa).

E.2 Propriedades finitas da amostra de MQO

Derivar o valor e a variação esperados do estimador MQO $\hat{\beta}$ é facilitado por álgebra matricial, mas devemos demonstrar algum cuidado ao declarar as hipóteses.

Hipótese E.1 Linear em parâmetros

O modelo pode ser escrito como em (E.3), em que $\mathbf{y}$ é um vetor observado $n \times 1$, $\mathbf{X}$ é uma matriz observada $n \times (k + 1)$ e $\mathbf{u}$ é um vetor de erros ou distúrbios não observados $n \times 1$.

Hipótese E.2 Colinearidade imperfeita

A matriz $\mathbf{X}$ classifica $(k + 1)$.

Essa é uma declaração cuidadosa da hipótese, que descarta dependências lineares entre as variáveis explicativas. Sob a Hipótese E.2, $\mathbf{X}'\mathbf{X}$ não é singular, ou seja, $\hat{\beta}$ é única e pode ser escrita como em (E.8).

Hipótese E.3 Medida condicional zero

Condicional à matriz $\mathbf{X}$ inteira, cada erro u_t tem medida zero: $E(u_t|\mathbf{X}) = 0$, $t = 1, 2, \dots, n$.

Na forma de vetor, a Hipótese E.3 pode ser escrita como

$$E(\mathbf{u}|\mathbf{X}) = \mathbf{0}. \quad (\text{E.11})$$

Essa hipótese é implicada por RLM.4 sob a hipótese de amostra aleatória RLM.2. Em aplicações de série temporal, a Hipótese E.3 impõe estrita exogeneidade nas variáveis explicativas, algo discutido longamente no Capítulo 10. Isso exclui variáveis explicativas cujos valores futuros sejam correlacionados com u_t ; em especial, isso elimina variáveis dependentes defasadas. Sob a Hipótese E.3, podemos condicionar em x_{ij} quando calculamos o valor esperado de $\hat{\beta}$.

**TEOREMA
E.1****INEXISTÊNCIA DE VIÉS DO MQO**

Sob as Hipóteses E.1, E.2 e E.3, o estimador MQO $\hat{\beta}$ é não viesado para β .

Demonstração: Utilize as Hipóteses E.1, E.2 e álgebra simples para escrever

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u},\end{aligned}\quad (\text{E.12})$$

em que usamos o fato de que $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X}) = \mathbf{I}_{k+1}$. Colocar a expectativa condicional em $\mathbf{X}$ nos dá

$$\begin{aligned}E(\hat{\beta}|\mathbf{X}) &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{u}|\mathbf{X}) \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{0} = \beta,\end{aligned}$$

logo, $E(\mathbf{u}|\mathbf{X}) = \mathbf{0}$, sob a hipótese E.3. Esse argumento claramente não depende do valor de β , então mostramos que $\hat{\beta}$ é não viesado.

Para obter a forma mais simples de matriz de variação-covariação de $\hat{\beta}$, introduzimos as hipóteses de homoscedasticidade e nenhuma correlação serial.

Hipótese E.4. Homoscedasticidade sem correlação serial

(i) $\text{Var}(u_t|\mathbf{X}) = \sigma^2$, $t = 1, 2, \dots, n$. (ii) $\text{Cov}(u_t, u_s|\mathbf{X}) = 0$, para todos $t \neq s$. Em forma de matriz, podemos escrever essas duas hipóteses como

$$\text{Var}(\mathbf{u}|\mathbf{X}) = \sigma^2 \mathbf{I}_n, \quad (\text{E.13})$$

em que $\mathbf{I}_n$ é a matriz de identidade $n \times n$.

A parte (i) da Hipótese E.4 é a hipótese de homoscedasticidade: a variação de u_t não pode depender de qualquer elemento de $\mathbf{X}$, e a variação deve ser constante nas variações, t . A parte (ii) é a hipótese de não correlação serial: os erros não podem ser correlacionados nas observações. Sob a amostragem aleatória, e em quaisquer outros esquemas de amostragem de corte transversal com observações independentes, a parte (ii) da Hipótese E.4 automaticamente se sustenta. Para aplicações de séries temporais, a parte (ii) exclui correlação nos erros com o tempo (ambos condicionais em $\mathbf{X}$ e incondicionalmente).

Por causa de (E.13), frequentemente dizemos que $\mathbf{u}$ tem uma **matriz de variação-covariação escalar** caso a Hipótese E.4 se mantenha. Podemos agora derivar **matriz de variação-covariação do estimador MQO**.

**TEOREMA
E.2****MATRIZ DE VARIAÇÃO-COVARIAÇÃO DO ESTIMADOR MQO**

Sob a Hipótese E.1 em E.4,

$$\text{Var}(\hat{\beta}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \quad (\text{E.14})$$

Prova: Da última fórmula da equação (E.12), temos

$$\text{Var}(\hat{\beta}|\mathbf{X}) = \text{Var}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\text{Var}(\mathbf{u}|\mathbf{X})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

Agora, usamos a Hipótese E.4 para obter

$$\begin{aligned}\text{Var}(\hat{\beta}|\mathbf{X}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I}_n)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

A fórmula (E.14) significa que a variação de $\hat{\beta}_j$ (condicional em $\mathbf{X}$) é obtida pela multiplicação de σ^2 pelo elemento diagonal j^{th} de $(\mathbf{X}'\mathbf{X})^{-1}$. Para os coeficientes de inclinação, damos uma fórmula interpretável à equação (3.51). A equação (E.14) também nos diz como obter a covariação entre qualquer uma das duas estimativas MQO: multiplicar σ^2 pelo elemento de $(\mathbf{X}'\mathbf{X})^{-1}$ fora da diagonal. No Capítulo 4, mostramos como evitar explicitamente encontrar covariações para obter intervalos de confiança e testes de hipótese ao reescrever o modelo de forma apropriada.

O Teorema de Gauss-Markov, em sua plena generalidade, pode ser provado.

TEOREMA E.3

TEOREMA DE GAUSS-MARKOV

Sob a Hipótese E.1 em E.4, $\hat{\beta}$ é o melhor estimador linear não viesado.

Prova: Qualquer outro estimador linear de β pode ser escrito como

$$\tilde{\beta} = \mathbf{A}'\mathbf{y}, \quad (\text{E.15})$$

em que $\mathbf{A}$ é uma matriz $n \times (k + 1)$. Para que $\tilde{\beta}$ seja um condicional não viesado em $\mathbf{X}$, $\mathbf{A}$ pode consistir em números e funções não aleatórios de $\mathbf{X}$ (por exemplo, $\mathbf{A}$ não pode ser uma função de $\mathbf{y}$). Para ver que restrições adicionais de $\mathbf{A}$ são necessárias, escreva

$$\tilde{\beta} = \mathbf{A}'(\mathbf{X}\beta + \mathbf{u}) = (\mathbf{A}'\mathbf{X})\beta + \mathbf{A}'\mathbf{u}. \quad (\text{E.16})$$

Logo,

$$\begin{aligned}E(\tilde{\beta}|\mathbf{X}) &= \mathbf{A}'\mathbf{X}\beta + E(\mathbf{A}'\mathbf{u}|\mathbf{X}) \\ &= \mathbf{A}'\mathbf{X}\beta + \mathbf{A}'E(\mathbf{u}|\mathbf{X}) \text{ porque } \mathbf{A} \text{ é uma função de } \mathbf{X} \\ &= \mathbf{A}'\mathbf{X}\beta \text{ porque } E(\mathbf{u}|\mathbf{X}) = \mathbf{0}.\end{aligned}$$

Para que $\tilde{\beta}$ seja um estimador não viesado de β , é preciso que $E(\tilde{\beta}|\mathbf{X}) = \beta$ para todos os vetores $(k + 1) \times 1$ β , ou seja,

$$\mathbf{A}'\mathbf{X}\beta = \beta \text{ para todos } (k + 1) \times 1 \text{ vetores } \beta. \quad (\text{E.17})$$

Como $\mathbf{A}'\mathbf{X}$ é uma matriz $(k + 1) \times (k + 1)$, (E.17) se sustenta apenas se $\mathbf{A}'\mathbf{X} = \mathbf{I}_{k+1}$. As equações (E.15) e (E.17) caracterizam a classe de estimadores não viesados, lineares para β .

Na sequência, de (E.16), temos

$$\text{Var}(\tilde{\beta}|\mathbf{X}) = \mathbf{A}'[\text{Var}(\mathbf{u}|\mathbf{X})]\mathbf{A} = \sigma^2\mathbf{A}'\mathbf{A},$$

pela Hipótese E.4. Portanto,

$$\begin{aligned}\text{Var}(\tilde{\beta}|\mathbf{X}) - \text{Var}(\hat{\beta}|\mathbf{X}) &= \sigma^2[\mathbf{A}'\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}] \\ &= \sigma^2[\mathbf{A}'\mathbf{A} - \mathbf{A}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}] \text{ porque } \mathbf{A}'\mathbf{X} = \mathbf{I}_{k+1} \\ &= \sigma^2\mathbf{A}'[\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{A} \\ &\equiv \sigma^2\mathbf{A}'\mathbf{M}\mathbf{A},\end{aligned}$$

em que $\mathbf{M} \equiv \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Como $\mathbf{M}$ é simétrico e idempotente, $\mathbf{A}'\mathbf{M}\mathbf{A}$ é positivo e semidefinitivo para qualquer matriz $\mathbf{A}$ $n \times (k+1)$. Isso estabelece que o estimador MQO $\hat{\beta}$ é BLUE. Por que isso é importante? Leve em conta que $\mathbf{c}$ é um vetor $k+1 \times 1$ qualquer e considere a combinação linear $\mathbf{c}'\beta = c_0\beta_0 + c_1\beta_1 + \dots + c_k\beta_k$ como escalar. Os estimadores não viesados de $\mathbf{c}'\beta$ são $\mathbf{c}'\tilde{\beta}$ e $\mathbf{c}'\hat{\beta}$. Contudo

$$\text{Var}(\mathbf{c}'\tilde{\beta}|\mathbf{X}) - \text{Var}(\mathbf{c}'\hat{\beta}|\mathbf{X}) = \mathbf{c}'[\text{Var}(\tilde{\beta}|\mathbf{X}) - \text{Var}(\hat{\beta}|\mathbf{X})]\mathbf{c} \geq 0,$$

porque $[\text{Var}(\tilde{\beta}|\mathbf{X}) - \text{Var}(\hat{\beta}|\mathbf{X})]$ é p.s.d. Logo, quando é usado para estimar qualquer combinação linear de β , MQO rende a menor variação. Em particular, $\text{Var}(\hat{\beta}_j|\mathbf{X}) \leq \text{Var}(\tilde{\beta}_j|\mathbf{X})$ para qualquer outro estimador linear, não viesado de β_j .

O estimador não viesado da variação de erro σ^2 pode ser escrito como

$$\hat{\sigma}^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}}/(n - k - 1),$$

significando o mesmo que a equação (3.56).

TEOREMA E.4

INEXISTÊNCIA DE VIÉS DE $\hat{\sigma}^2$

Segundo as Hipóteses de E.1 a E.4, $\hat{\sigma}^2$ é não viesada para σ^2 : $E(\hat{\sigma}^2|\mathbf{X}) = \sigma^2$ para todo $\sigma^2 > 0$.

Prova: Escreva $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\beta} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{M}\mathbf{y} = \mathbf{M}\mathbf{u}$, em que $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ e a última equação ocorre porque $\mathbf{M}\mathbf{X} = \mathbf{0}$. Como $\mathbf{M}$ é simétrico e idempotente,

$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{u}'\mathbf{M}'\mathbf{M}\mathbf{u} = \mathbf{u}'\mathbf{M}\mathbf{u}.$$

Como $\mathbf{u}'\mathbf{M}$ é escalar, se iguala a seu traço. Portanto,

$$\begin{aligned}E(\mathbf{u}'\mathbf{M}\mathbf{u}|\mathbf{X}) &= E[\text{tr}(\mathbf{u}'\mathbf{M}\mathbf{u})|\mathbf{X}] = E[\text{tr}(\mathbf{M}\mathbf{u}\mathbf{u}')|\mathbf{X}] \\ &= \text{tr}[E(\mathbf{M}\mathbf{u}\mathbf{u}'|\mathbf{X})] = \text{tr}[\mathbf{M}E(\mathbf{u}\mathbf{u}'|\mathbf{X})] \\ &= \text{tr}(\mathbf{M}\sigma^2\mathbf{I}_n) = \sigma^2\text{tr}(\mathbf{M}) = \sigma^2(n - k - 1).\end{aligned}$$

A última igualdade ocorre a partir de $\text{tr}(\mathbf{M}) = \text{tr}(\mathbf{I}_n) - \text{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = n - \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] = n - \text{tr}(\mathbf{I}_{k+1}) = n - (k+1) = n - k - 1$. Logo,

$$E(\hat{\sigma}^2|\mathbf{X}) = E(\mathbf{u}'\mathbf{M}\mathbf{u}|\mathbf{X})/(n - k - 1) = \sigma^2.$$

E.3. Inferência estatística

Quando adicionamos a hipótese do modelo linear clássico, $\hat{\beta}$ apresenta distribuição normal multivariada, o que nos leva às distribuições t e F para as estatísticas de teste padrão abordadas no Capítulo 4.

Hipótese E.5 Normalidade de erros

Condicional em $\mathbf{X}$, o u_t é independente e identicamente distribuído como Normal $(0, \sigma^2)$. De forma equivalente, visto $\mathbf{X}$, $\mathbf{u}$ é distribuído como multivariado normal com significado zero e matriz de variação-covariação $\sigma^2 \mathbf{I}_n$: $\mathbf{u} \sim \text{Normal}(0, \sigma^2 \mathbf{I}_n)$.

Sob a Hipótese E.5, cada u_t é independente das variáveis explicativas para todos os t . Ao se estabelecer séries temporais, essa é a hipótese de exogeneidade estrita.

TEOREMA E.5**NORMALIDADE DE $\hat{\beta}$**

Sob as hipóteses de modelo linear clássico de E.1 a E.5, o condicional $\hat{\beta}$ em $\mathbf{X}$ é distribuído como normal multivariado com significado β e matriz de variação-covariação $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

O Teorema E.5 é a base para inferência estatística envolvendo β . De fato, juntamente com as propriedades de qui-quadrado e as distribuições t e F resumidas no Apêndice D, podemos usar o Teorema E.5 para estabelecer que estatísticas t têm uma distribuição t segundo as Hipóteses de E.1 a E.5 (sob a hipótese nula) e igualmente para estatísticas F . Ilustramos essa situação com uma demonstração de estatísticas t .

TEOREMA E.6**DISTRIBUIÇÃO DE ESTATÍSTICA t**

Sob as Hipóteses de E.1 a E.5,

$$(\hat{\beta}_j - \beta_j)/\text{ep}(\hat{\beta}_j) \sim t_{n-k-1}, j = 0, 1, \dots, k.$$

Prova: A prova exige várias etapas; as seguintes declarações são inicialmente condicionais em $\mathbf{X}$. Primeiro, pelo Teorema E.5, $(\hat{\beta}_j - \beta_j)/\text{dp}(\hat{\beta}_j) \sim \text{Normal}(0, 1)$, em que $\text{dp}(\hat{\beta}_j) = \sigma \sqrt{C_{jj}}$, e C_{jj} são os j^{th} elementos diagonais de $(\mathbf{X}'\mathbf{X})^{-1}$. A seguir, de acordo com as Hipóteses de E.1 a E.5, condicionais em $\mathbf{X}$,

$$(n - k - 1)\hat{\sigma}^2/\sigma^2 \sim \chi_{n-k-1}^2. \quad (\text{E.18})$$

Isso ocorre porque $(n - k - 1)\hat{\sigma}^2/\sigma^2 = (\mathbf{u}/\sigma)' \mathbf{M}(\mathbf{u}/\sigma)$, em que $\mathbf{M}$ é a matriz $n \times n$, simétrica, idempotente, definida no Teorema E.4. Contudo, $\mathbf{u}/\sigma \sim \text{Normal}(\mathbf{0}, \mathbf{I}_n)$, de acordo com a Hipótese E.5. Ocorre na Propriedade 1 para distribuição qui-quadrada do Apêndice D que $(\mathbf{u}/\sigma)' \mathbf{M}(\mathbf{u}/\sigma) \sim \chi_{n-k-1}^2$ (já que $\mathbf{M}$ avaliou $n - k - 1$).

Também precisamos mostrar que $\hat{\beta}$ e $\hat{\sigma}^2$ são independentes. Contudo, $\hat{\beta} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$ e $\hat{\sigma}^2 = \mathbf{u}'\mathbf{M}\mathbf{u}/(n - k - 1)$. Agora, $[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{M} = \mathbf{0}$, já que $\mathbf{X}'\mathbf{M} = \mathbf{0}$. Ocorre, na Propriedade 5 da distribuição normal multivariada do Apêndice D, que $\hat{\beta}$ e $\mathbf{M}\mathbf{u}$ são independentes. Já que $\hat{\sigma}^2$ é uma função de $\mathbf{M}\mathbf{u}$, $\hat{\beta}$ e $\hat{\sigma}^2$ são também independentes.

$$(\hat{\beta}_j - \beta_j)/\text{ep}(\hat{\beta}_j) = [(\hat{\beta}_j - \beta_j)/\text{dp}(\hat{\beta}_j)]/(\hat{\sigma}^2/\sigma^2)^{1/2},$$

que é a proporção de uma variável aleatória normal padrão e a raiz quadrada de uma variável aleatória $\chi_{n-k-1}^2/(n - k - 1)$. Apenas mostramos que ela é independente, então, por definição de uma variável aleatória t , $(\hat{\beta}_j - \beta_j)/\text{ep}(\hat{\beta}_j)$ possui a distribuição t_{n-k-1} . Como essa distribuição não depende de $\mathbf{X}$, também é a distribuição incondicional de $(\hat{\beta}_j - \beta_j)/\text{ep}(\hat{\beta}_j)$.

Podemos associar a partir desse teorema qualquer valor hipotético para β_j e usar a estatística t para testar hipóteses, como habitualmente.

Segundo as Hipóteses de E.1 a E.5, podemos calcular o que se conhece como limite inferior de *Cramer-Rao* para matriz de variação-covariação de estimadores não viesados de β (novamente, condicional em $\mathbf{X}$) [ver Greene (1997, Capítulo 4)]. Isso pode ser mostrado como $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, exatamente a matriz de variação-covariação do estimador MQO. Isso implica que $\hat{\beta}$ é o **estimador não viesado de variância mínima** de β (condicional em $\mathbf{X}$): $\text{Var}(\tilde{\beta}|\mathbf{X}) - \text{Var}(\hat{\beta}|\mathbf{X})$ é positivo, semidefinitivo para qualquer outro estimador não viesado de β ; não precisamos mais restringir nossa atenção aos estimadores lineares em $\mathbf{y}$.

É fácil mostrar que os estimadores MQO são, de fato, os estimadores de máxima verossimilhança de β , segundo a Hipótese E.5. Para cada t , a distribuição de y_t , dado $\mathbf{X}$, é $\text{Normal}(\mathbf{x}_t\beta, \sigma^2)$. Como y_t é condicional independente em $\mathbf{X}$, a função de probabilidade para a amostra é obtida do produto das densidades:

$$\prod_{t=1}^n (2\pi\sigma^2)^{-1/2} \exp[-(y_t - \mathbf{x}_t\beta)^2/(2\sigma^2)],$$

em que Π denota produto. Maximizar essa função relacionando β e σ^2 é o mesmo que maximizar seu logaritmo natural:

$$\sum_{t=1}^n [-(1/2)\log(2\pi\sigma^2) - (y_t - \mathbf{x}_t\beta)^2/(2\sigma^2)].$$

Para obter $\hat{\beta}$, basta fazer o mesmo que minimizar $\sum_{t=1}^n (y_t - \mathbf{x}_t\beta)^2$ – a divisão por $2\sigma^2$ não afeta a otimização – justamente o problema que MQO resolve. O estimador de σ^2 que usamos, $\text{SQR}/(n - k)$, descobriu-se não ser o EMV de σ^2 ; o EMV é SQR/n , estimador viesado. Como o estimador não viesado de σ^2 resulta em estatísticas t e F , com exatas distribuições t e F sob a nula, ele é sempre usado em vez de MLE.

O fato de que o estimador MQO é o EMV, sob a Hipótese E.5, implica uma interessante propriedade de robustez do EMV com base em distribuição normal. O motivo é simples. Sabemos que o estimador MQO é não viesado sob as Hipóteses de E.1 a E.3; a normalidade dos erros não é usada em lugar algum na demonstração, assim como a Hipótese 4. Como a próxima seção mostra, o estimador MQO também é consistente sem normalidade, contanto que a lei de grandes quantidades se sustente (o que é amplamente verdade). Essas propriedades estatísticas do estimador MQO implicam que o EMV com base em função normal de verossimilhança logarítmica é robusto para especificação distribucional: a distribuição pode ser (quase) qualquer coisa e ainda assim obtemos um estimador consistente (e, de E.1 a E.3, não viesado). Como discutido na Seção 17.3, um estimador de máxima verossimilhança obtido ao se presumir que a distribuição esteja correta é frequentemente chamado de **estimador por quase máxima verossimilhança (EQMV)**.

Geralmente, a consistência do EMV depende de uma distribuição correta para concluir que é consistente para os parâmetros. Vimos que a distribuição normal é uma exceção notável. Existem outras distribuições que compartilham essa propriedade, incluindo a distribuição de Poisson – como discutido na Seção 17.3. Wooldridge (2010, Capítulo 18) discute outros exemplos úteis.

E.4 Algumas análises assintóticas

A abordagem de matriz para o modelo de regressão múltipla também pode fazer derivações de propriedades assintóticas mais concisas. De fato, fornecemos demonstrações gerais das alegações no Capítulo 11.

Começamos provando o resultado de consistência do Teorema 11.1. Lembre-se de que essas hipóteses contêm, como um caso especial, as hipóteses para análise de corte transversal sob amostragem aleatória.

Prova do Teorema 11.1. Como visto no Problema E.1 e usando a Hipótese TS.1', escrevemos o estimador MQO como

$$\begin{aligned}\hat{\beta} &= \left(\sum_{t=1}^n \mathbf{x}_t' \mathbf{x}_t \right)^{-1} \left(\sum_{t=1}^n \mathbf{x}_t' y_t \right) = \left(\sum_{t=1}^n \mathbf{x}_t' \mathbf{x}_t \right)^{-1} \left(\sum_{t=1}^n \mathbf{x}_t' (\mathbf{x}_t \beta + u_t) \right) \\ &= \beta + \left(\sum_{t=1}^n \mathbf{x}_t' \mathbf{x}_t \right)^{-1} \left(\sum_{t=1}^n \mathbf{x}_t' u_t \right) \\ &= \beta + \left(n^{-1} \sum_{t=1}^n \mathbf{x}_t' \mathbf{x}_t \right)^{-1} \left(n^{-1} \sum_{t=1}^n \mathbf{x}_t' u_t \right).\end{aligned}\tag{E.19}$$

Agora, segundo a lei de grandes quantidades,

$$n^{-1} \sum_{t=1}^n \mathbf{x}_t' \mathbf{x}_t \xrightarrow{p} \mathbf{A} \text{ e } n^{-1} \sum_{t=1}^n \mathbf{x}_t' u_t \xrightarrow{p} \mathbf{0},\tag{E.20}$$

em que $\mathbf{A} = E(\mathbf{x}_t' \mathbf{x}_t)$ é uma $(k+1) \times (k+1)$ matriz não singular sob a Hipótese TS.2', e usamos o fato de que $E(\mathbf{x}_t' u_t) = \mathbf{0}$ sob a Hipótese TS.3'. Agora, devemos usar a versão matricial de Propriedade PLIM.1 do Apêndice C. Especificamente, como $\mathbf{A}$ não é singular,

$$\left(n^{-1} \sum_{t=1}^n \mathbf{x}_t' \mathbf{x}_t \right)^{-1} \xrightarrow{p} \mathbf{A}^{-1}.\tag{E.21}$$

[Wooldridge (2010, Capítulo 3) apresenta uma discussão sobre esse tipo de resultado de convergência]. O resultado observado em (E.19), (E.20) e (E.21) é que

$$\text{plim}(\hat{\beta}) = \beta + \mathbf{A}^{-1} \cdot \mathbf{0} = \beta.$$

Isso completa a demonstração.

Na sequência, esboçamos uma demonstração do resultado da normalidade assintótica do Teorema 11.2.

Prova do Teorema 11.2. Segundo a equação (E.19), podemos escrever

$$\begin{aligned}\sqrt{n}(\hat{\beta} - \beta) &= \left(n^{-1} \sum_{t=1}^n \mathbf{x}_t' \mathbf{x}_t \right)^{-1} \left(n^{-1/2} \sum_{t=1}^n \mathbf{x}_t' u_t \right) \\ &= \mathbf{A}^{-1} \left(n^{-1/2} \sum_{t=1}^n \mathbf{x}_t' u_t \right) + o_p(1),\end{aligned}\tag{E.22}$$

em que o termo “ $O_p(1)$ ” é um termo remanescente que, em probabilidade, converge a zero. Esse termo é igual a $[(n^{-1} \sum_{t=1}^n \mathbf{x}_t' \mathbf{x}_t)^{-1} - \mathbf{A}^{-1}](n^{-1/2} \sum_{t=1}^n \mathbf{x}_t' u_t)$. O termo entre

parênteses converge em probabilidade a zero (pelo mesmo argumento usado na prova do Teorema 11.1), enquanto $(n^{-1/2} \sum_{t=1}^n \mathbf{x}'_t u_t)$ se limita em probabilidade porque converge a uma distribuição normal multivariada pelo teorema de limite central. Um resultado bem conhecido na teoria assintótica é que o produto desses termos converge em probabilidade a zero. Além disso, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ herda sua distribuição assintótica de $\mathbf{A}^{-1}(n^{-1/2} \sum_{t=1}^n \mathbf{x}'_t u_t)$. Veja Wooldridge (2010, Capítulo 3) para mais detalhes sobre os resultados convergentes usados nessa demonstração.

Segundo o teorema de limite central, $n^{-1/2} \sum_{t=1}^n \mathbf{x}'_t u_t$ possui uma distribuição normal assintótica com significado zero e, digamos, matriz $\mathbf{B}$ de variação-covariação $(k+1) \times (k+1)$. Logo, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ tem uma distribuição normal multivariada assintótica com significado zero e matriz de variação-covariação $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$. Mostramos agora que, de acordo com as Hipóteses TS.4' e TS.5', $\mathbf{B} = \sigma^2\mathbf{A}$. (O termo geral é útil porque sustenta erros padrão robustos com relação à heteroscedasticidade e correlação serial para MQO, do tipo que discutimos no Capítulo 12). Em primeiro lugar, de acordo com a Hipótese TS.5', $\mathbf{x}'_t u_t$ e $\mathbf{x}'_s u_s$ não são correlacionados para $t \neq s$. Por quê? Suponha $s < t$ para uma concretude. Então, pela lei das expectativas iteradas, $E(\mathbf{x}'_t u_t \mathbf{x}'_s) = E[E(u_t u_s | \mathbf{x}'_t \mathbf{x}'_s) \mathbf{x}'_t \mathbf{x}'_s] = E[E(u_t u_s | \mathbf{x}'_t \mathbf{x}'_s) \mathbf{x}'_t \mathbf{x}'_s] = E[0 \cdot \mathbf{x}'_t \mathbf{x}'_s] = 0$. As covariações zero implicam que a variação da soma é a soma das variações. Contudo, $\text{Var}(\mathbf{x}'_t u_t \mathbf{x}'_t) = E(\mathbf{x}'_t u_t u_t \mathbf{x}'_t) = E(u_t^2 \mathbf{x}'_t \mathbf{x}'_t)$. Pela lei das expectativas iteradas, $E(u_t^2 \mathbf{x}'_t \mathbf{x}'_t) = E[E(u_t^2 \mathbf{x}'_t \mathbf{x}'_t | \mathbf{x}'_t)] = E[E(u_t^2 | \mathbf{x}'_t) \mathbf{x}'_t \mathbf{x}'_t] = E[\sigma^2 \mathbf{x}'_t \mathbf{x}'_t] = \sigma^2 E(\mathbf{x}'_t \mathbf{x}'_t) = \sigma^2 \mathbf{A}$, em que usamos $E(u_t^2 | \mathbf{x}'_t) = \sigma^2$ de acordo com as Hipóteses TS.3' e TS.4'. Isso mostra que $\mathbf{B} = \sigma^2\mathbf{A}$ e então, de acordo com as Hipóteses de TS.1' a TS.5', temos

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \underset{d}{\sim} \text{Normal}(\mathbf{0}, \sigma^2 \mathbf{A}^{-1}). \quad (\text{E.23})$$

Isso completa a prova.

Segundo a equação (E.23), tratamos $\hat{\boldsymbol{\beta}}$ como se fosse aproximada e normalmente distribuído com significado $\boldsymbol{\beta}$ e matriz de variação-covariação $\sigma^2 \mathbf{A}^{-1}/n$. Espera-se aqui a divisão pelo tamanho da amostra, n : a aproximação para a matriz de variação-covariação de $\hat{\boldsymbol{\beta}}$ diminui a zero à proporção $1/n$. Quando substituímos σ^2 por seu estimador consistente, $\hat{\sigma}^2 = \text{SQR}/(n - k - 1)$, e substituímos $\mathbf{A}$ por seu estimador consistente, $n^{-1} \sum_{t=1}^n \mathbf{x}'_t \mathbf{x}'_t = \mathbf{X}'\mathbf{X}/n$, obtemos um estimador para a variação assintótica de $\hat{\boldsymbol{\beta}}$:

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1}. \quad (\text{E.24})$$

Perceba como as duas divisões por n terminam e a parte direita de (E.24) é simplesmente a forma habitual com que estimamos a matriz de variação do estimador MQO sob as Hipóteses de Gauss-Markov. Para resumir, mostramos que, sob as Hipóteses de TS.1' a TS.5' – que contêm RLM.1 a RLM.5 como casos especiais –, os erros padrões habituais e estatísticas t são assintoticamente válidos. É perfeitamente legítimo usar a distribuição t usual para obter valores críticos e p -valores para testar uma única hipótese. Curiosamente, na configuração geral do Capítulo 11, presumir a normalidade dos erros – digamos, u_t , visto que $\mathbf{x}_t, u_{t-1}, \mathbf{x}_{t-1}, \dots, u_1, \mathbf{x}_1$, é distribuído como $\text{Normal}(0, \sigma^2)$ – não ajuda necessariamente, já que as estatísticas t não teriam geralmente estatísticas t exatas, de acordo com esse tipo de hipótese de normalidade. Quando não presumimos exogeneidade estrita das variáveis explicativas, os resultados distributivos exatos são difíceis, se não impossíveis, de obter.

Se modificarmos os argumentos acima, podemos derivar uma matriz de variação-covariação robusta com relação à heteroscedasticidade. A chave é estimar $E(u_i^2 \mathbf{x}_i' \mathbf{x}_i)$ separadamente, porque a matriz não é mais igual a $\sigma^2 E(\mathbf{x}_i' \mathbf{x}_i)$. Contudo, se $\hat{u}_i$ for os resíduos MQO, um estimador consistente seria

$$(n - k - 1)^{-1} \sum_{i=1}^n \hat{u}_i^2 \mathbf{x}_i' \mathbf{x}_i, \quad (\text{E.25})$$

em que a divisão por $n - k - 1$, em vez de n , representa graus de ajuste de liberdade que tipicamente ajudam as propriedades de amostras finitas do estimador. Quando usamos a expressão na equação (E.25), obtemos

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\beta}}) = [n/(n - k - 1)](\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{i=1}^n \hat{u}_i^2 \mathbf{x}_i' \mathbf{x}_i \right) (\mathbf{X}'\mathbf{X})^{-1}. \quad (\text{E.26})$$

As raízes quadradas dos elementos diagonais dessa matriz são os mesmos erros padrão robustos com relação à heteroscedasticidade obtidos na Seção 8.2, para o caso de puro corte transversal. Uma extensão de matriz dos erros padrão robustos com relação à correlação serial (e heteroscedasticidade) obtidos na Seção 12.5 também está disponível, mas a matriz que deve substituir (E.25) é complicada por causa da correlação serial. Veja, por exemplo, Hamilton (1994, Seção 10.5).

E.4 Estatística de Wald para o teste de múltiplas hipóteses

Argumentos similares podem ser usados para obter a distribuição assintótica da **estatística Wald** para o teste de múltiplas hipóteses. Considere $\mathbf{R}$ uma matriz $q \times (k + 1)$, com $q \leq (k + 1)$. Presuma que as restrições q sobre o vetor de parâmetros $(k + 1) \times 1$, $\boldsymbol{\beta}$, possam ser expressas como $H_0: \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, em que $\mathbf{r}$ é um vetor $q \times 1$ de constantes conhecidas. Levando em conta as Hipóteses de TS.1' a TS.5', pode ser mostrado que, sob H_0 ,

$$[\sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})]'(\sigma^2 \mathbf{R}\mathbf{A}^{-1}\mathbf{R}')^{-1}[\sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})] \xrightarrow{d} \chi_q^2, \quad (\text{E.27})$$

em que $\mathbf{A} = (\mathbf{x}_i' \mathbf{x}_i)$, como nas demonstrações dos Teoremas 11.1 e 11.2. A intuição por trás da equação (E.25) é simples. Como $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ é aproximadamente distribuída como $\text{Normal}(\mathbf{0}, \sigma^2 \mathbf{A}^{-1})$, $\mathbf{R}[\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] = \sqrt{n}\mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ é aproximadamente $\text{Normal}(\mathbf{0}, \sigma^2 \mathbf{R}\mathbf{A}^{-1}\mathbf{R}')$ pela Propriedade 3 da distribuição normal multivariada do Apêndice D. Sob H_0 , $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, logo, $\sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) \stackrel{a}{\sim} \text{Normal}(\mathbf{0}, \sigma^2 \mathbf{R}\mathbf{A}^{-1}\mathbf{R}')$ sob H_0 . Segundo a Propriedade 3 da distribuição qui-quadrada, $z'(\sigma^2 \mathbf{R}\mathbf{A}^{-1}\mathbf{R}')^{-1}z \sim \chi_q^2$ se $z \sim \text{Normal}(\mathbf{0}, \sigma^2 \mathbf{R}\mathbf{A}^{-1}\mathbf{R}')$. Para obter formalmente o resultado final, precisamos usar uma versão assintótica dessa propriedade, que pode ser encontrada em Wooldridge (2010, Capítulo 3).

Visto o resultado em (E.25), obtemos uma estatística computável ao substituir $\mathbf{A}$ e σ^2 por seus estimadores consistentes; fazer isso não muda a distribuição assintótica. O resultado é a chamada estatística Wald, que, depois de anular os tamanhos de amostras e fazer uma pequena álgebra, pode ser descrita como

$$W = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})/\hat{\sigma}^2. \quad (\text{E.28})$$

De acordo com H_0 , $W \stackrel{a}{\sim} \chi_q^2$, em que lembramos que q é o número de restrições testadas. Se $\sigma^2 = \text{SQR}/(n - k - 1)$, pode ser demonstrado que W/q é exatamente a estatística F que obtemos no Capítulo 4 para o teste de restrições lineares múltiplas. [Ver, por exemplo, Greene (1997, Capítulo 7).] Portanto, de acordo com as hipóteses de modelo linear clássico de TS.1 a TS.6, do Capítulo 10, W/q apresentou uma exata distribuição $F_{q, n-k-1}$. Sob as Hipóteses de TS.1' a TS.5', temos somente o resultado assintótico em (E.26). Entretanto, é apropriado e comum tratar a estatística F usual como tendo uma distribuição $F_{q, n-k-1}$ aproximada.

Uma estatística Wald robusta com relação à heteroscedasticidade, de forma desconhecida, é obtida usando uma matriz em (E.26), no lugar de $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$, e, de forma similar, para uma estatística de teste robusta com relação à heteroscedasticidade e também à correlação serial. As versões robustas de estatísticas de teste não podem ser calculadas por meio da soma de resíduos quadrados ou R -quadrados de regressões restritas e irrestritas.

Resumo

Este Apêndice forneceu uma breve abordagem do modelo de regressão linear usando notação de matriz. Este material foi incluso para classes mais avançadas, que utilizam álgebra matricial; contudo, não é necessário ler o texto. De fato, este apêndice prova alguns dos resultados que ambos estabelecemos sem demonstração, provando apenas em casos especiais ou através de um método ou demonstração mais difícil. Outros tópicos – como propriedades assintóticas, estimação de variáveis instrumentais e modelos de dados em painel – podem receber tratamentos concisos usando matrizes. Para mais detalhes, consulte os textos avançados em econometria de Davidson e MacKinnon (1993), Greene (1997), Hayashi (2000) e Wooldridge (2010).

Termos-chave

Condição de primeira ordem	Estimador por quase máxima verossimilhança (EQMV)	Matriz de variação-covariação escalar
Estatística Wald		
Estimador não viesado de variância mínima	Matriz de variação-covariação do estimador MQO	Notação de matriz Teorema de Frisch-Waugh (FW)

Problemas

- 1 Considere $\mathbf{x}_t$ o vetor de variáveis explicativas $1 \times (k + 1)$ para a observação t . Mostre que o estimador MQO $\hat{\boldsymbol{\beta}}$ pode ser escrito como

$$\hat{\boldsymbol{\beta}} = \left(\sum_{t=1}^n \mathbf{x}_t' \mathbf{x}_t \right)^{-1} \left(\sum_{t=1}^n \mathbf{x}_t' y_t \right).$$

Dividir cada somatória por n mostra que $\hat{\boldsymbol{\beta}}$ é uma função de médias amostrais.

2. Considere $\hat{\boldsymbol{\beta}}$ o vetor de estimativas MQO $(k + 1) \times 1$.
 - (i) Demonstre que, para cada vetor $\mathbf{b}$ $(k + 1) \times 1$, podemos escrever a soma de resíduos quadrados como

$$\text{SQR}(\mathbf{b}) = \hat{\mathbf{u}}'\hat{\mathbf{u}} + (\hat{\boldsymbol{\beta}} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{b}).$$

{Dica: Escreva $(\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) = [\hat{\mathbf{u}} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{b})]'[\hat{\mathbf{u}} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{b})]$ e use o fato de que $\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}$.}

- (ii) Explique como a expressão para $\text{SQR}(\mathbf{b})$ da parte (i) prova que $\hat{\boldsymbol{\beta}}$ minimiza unicamente $\text{SQR}(\mathbf{b})$ sobre todos os valores possíveis de $\mathbf{b}$, presumindo que $\mathbf{X}$ classificou $k + 1$.
3. Considere $\hat{\boldsymbol{\beta}}$ o estimador MQO da regressão de $\mathbf{y}$ em $\mathbf{X}$. Suponha que $\mathbf{A}$ seja uma matriz não singular $(k + 1) \times (k + 1)$ e defina $\mathbf{z}_t \equiv \mathbf{x}_t\mathbf{A}$, $t = 1, \dots, n$. Portanto, $\mathbf{z}_t$ é $1 \times (k + 1)$ e é uma combinação não singular linear de $\mathbf{x}_t$. Considere $\mathbf{Z}$ a matriz $n \times (k + 1)$ com linhas $\mathbf{z}_t$. Considere que $\tilde{\boldsymbol{\beta}}$ denote a estimativa MQO para regressão de $\mathbf{y}$ em $\mathbf{Z}$.
- (i) Mostre que $\tilde{\boldsymbol{\beta}} = \mathbf{A}^{-1}\hat{\boldsymbol{\beta}}$.
- (ii) Considere $\hat{y}_t$ os valores ajustados da regressão original e $\tilde{y}_t$ os valores ajustados da regressão de $\mathbf{y}$ em $\mathbf{Z}$. Mostre que $\tilde{y}_t = \hat{y}_t$, para todos $t = 1, 2, \dots, n$. Como se comparam os resíduos das duas regressões?
- (iii) Mostre que a matriz de variação estimada para $\hat{\boldsymbol{\beta}}$ é $\hat{\sigma}^2\mathbf{A}^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}^{-1'}$, em que $\hat{\sigma}^2$ é a estimativa de variação usual da regressão de $\mathbf{y}$ em $\mathbf{X}$.
- (iv) Considere $\hat{\beta}_j$ as estimativas MQO da regressão de y_t em $x_{t1}, \dots, x_{tk}$, e que $\tilde{\beta}_j$ seja as estimativas MQO da regressão de y_t em $1, a_1x_{t1}, \dots, a_kx_{tk}$, em que $a_j \neq 0, j = 1, \dots, k$. Use os resultados do item (i) para encontrar a relação entre $\tilde{\beta}_j$ e $\hat{\beta}_j$.
- (v) Presuma a organização do item (iv), e use o item (iii) para mostrar que $(\hat{\beta}_j) = \text{ep}(\tilde{\beta}_j)/|a_j|$.
- (vi) Presuma a organização do item (iv), e mostre que os valores absolutos das estatísticas t para $\tilde{\beta}_j$ e $\hat{\beta}_j$ são idênticos.
4. Presuma que o modelo $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ satisfaça as hipóteses de Gauss-Markov, considere $\mathbf{G}$ uma matriz não singular e não aleatória $(k + 1) \times (k + 1)$, e defina $\boldsymbol{\delta} = \mathbf{G}\boldsymbol{\beta}$, de forma que $\boldsymbol{\delta}$ seja também um vetor $(k + 1) \times 1$. Considere $\hat{\boldsymbol{\beta}}$ o vetor de estimadores MQO $(k + 1) \times 1$ e defina $\hat{\boldsymbol{\delta}} = \mathbf{G}\hat{\boldsymbol{\beta}}$ como estimador MQO de $\boldsymbol{\delta}$.
- (i) Mostre que $E(\hat{\boldsymbol{\delta}}|\mathbf{X}) = \boldsymbol{\delta}$.
- (ii) Encontre $\text{Var}(\hat{\boldsymbol{\delta}}|\mathbf{X})$ em termos de σ^2 , $\mathbf{X}$ e $\mathbf{G}$.
- (iii) Use o problema E.3 para provar que $\hat{\boldsymbol{\delta}}$ e a estimativa adequada de $\text{Var}(\hat{\boldsymbol{\delta}}|\mathbf{X})$ são obtidos da regressão de $\mathbf{y}$ em $\mathbf{X}\mathbf{G}^{-1}$.
- (iv) Agora, considere $\mathbf{c}$ um vetor $(k + 1) \times 1$ com ao menos uma entrada com valor diferente de zero. Para concretude, presuma que $c_k \neq 0$. Defina $\theta = \mathbf{c}'\boldsymbol{\beta}$, de forma que θ seja escalar. Defina $\delta_j = \beta_j, j = 0, 1, \dots, k - 1$ e $\delta_k = \theta$. Mostre como definir uma matriz $\mathbf{G}$ não singular $(k + 1) \times (k + 1)$, de forma que $\boldsymbol{\delta} = \mathbf{G}\boldsymbol{\beta}$ (Dica: Cada uma das primeiras fileiras k de $\mathbf{G}$ contém zero k e um zero. Qual é a última fileira?)
- (v) Mostre que, para a escolha de $\mathbf{G}$ na parte (iv),

$$\mathbf{G}^{-1} = \begin{bmatrix} 1 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ \cdot & & & & & & \\ \cdot & & & & & & \\ \cdot & & & & & & \\ 0 & 0 & \cdot & \cdot & \cdot & 1 & 0 \\ -c_0/c_k & -c_1/c_k & \cdot & \cdot & \cdot & -c_{k-1}/c_k & 1/c_k \end{bmatrix}$$

Use essa expressão para $\mathbf{G}^{-1}$ e o item (iii) para concluir que x_{tk}/c_k e seu erro padrão são obtidos como coeficiente em x_{tk}/c_k , na regressão de

$$y_t \text{ on } [1 - (c_0/c_k)x_{tk}], [x_{t1} - (c_1/c_k)x_{tk}], \dots, [x_{t, k-1} - (c_{k-1}/c_k)x_{tk}], x_{tk}/c_k, t = 1, \dots, n.$$

Essa regressão é a exatamente a obtida quando escrevemos β_k em termos de θ e $\beta_0, \beta_1, \dots, \beta_{k-1}$, associando o resultado ao modelo original e rearranjando-o. Portanto, podemos justificar formalmente o truque que usamos no texto para obter o erro padrão de uma combinação linear de parâmetros.

5. Suponha que o modelo $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ satisfaça as hipóteses de Gauss-Markov e considere $\hat{\boldsymbol{\beta}}$ o estimador MQO de $\boldsymbol{\beta}$. Suponha que $\mathbf{Z} = \mathbf{G}(\mathbf{X})$ seja uma função de matriz de $\mathbf{X}$ $n \times (k+1)$ e presuma que a matriz $\mathbf{Z}'\mathbf{X}[a(k+1) \times (k+1)]$ seja não singular. Defina um novo estimador de $\boldsymbol{\beta}$ por $\tilde{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$.
 - (i) Mostre que $E(\tilde{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta}$, de forma que $\tilde{\boldsymbol{\beta}}$ também seja um condicional não viesado em $\mathbf{X}$.
 - (ii) Encontre $\text{Var}(\tilde{\boldsymbol{\beta}}|\mathbf{X})$. Certifique-se de que é uma matriz $(k+1) \times (k+1)$ simétrica que depende de $\mathbf{Z}$, $\mathbf{X}$ e σ^2 .
 - (iii) Que estimador você prefere, $\hat{\boldsymbol{\beta}}$ ou $\tilde{\boldsymbol{\beta}}$? Explique.
6. Considere a organização do Teorema de Frisch-Waugh.
 - (i) Usando matrizes separadas, mostre que as condições de primeira ordem $(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ podem ser escritas como

$$\begin{aligned}\mathbf{X}'_1\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}'_1\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 &= \mathbf{X}'_1\mathbf{y} \\ \mathbf{X}'_2\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}'_2\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 &= \mathbf{X}'_2\mathbf{y}.\end{aligned}$$

- (ii) Multiplique o primeiro conjunto de equações por $\mathbf{X}'_2\mathbf{X}_1(\mathbf{X}'_2\mathbf{X}_2)^{-1}$ e subtraia o resultado do segundo conjunto de equações para mostrar que

$$(\mathbf{X}'_2\mathbf{M}_1\mathbf{X}_2)\hat{\boldsymbol{\beta}}_2 = \mathbf{X}'_2\mathbf{M}_1\mathbf{y},$$

em que $\mathbf{I}_n - \mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1$. Determine que

$$\hat{\boldsymbol{\beta}}_2 = (\ddot{\mathbf{X}}'_2\ddot{\mathbf{X}}_2)^{-1}\ddot{\mathbf{X}}'_2\ddot{\mathbf{y}}.$$

- (iii) Use a parte (ii) para mostrar que

$$\hat{\boldsymbol{\beta}}_2 = (\ddot{\mathbf{X}}'_2\ddot{\mathbf{X}}_2)^{-1}\ddot{\mathbf{X}}'_2\ddot{\mathbf{y}}.$$

- (iv) Use o fato de que $\mathbf{M}_1\mathbf{X}_1 = \mathbf{0}$ para mostrar que os resíduos $\ddot{\mathbf{u}}$ da regressão de $\ddot{\mathbf{y}}$ em $\ddot{\mathbf{X}}_2$ são idênticos aos resíduos $\hat{\mathbf{u}}$ da regressão $\mathbf{y}$ em $\mathbf{X}_1, \mathbf{X}_2$. [Dica: Por definição e pelo teorema FW,

$$\ddot{\mathbf{u}} = \ddot{\mathbf{y}} - \ddot{\mathbf{X}}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{M}_1(\mathbf{y} - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2) = \mathbf{M}_1(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2).$$

Agora faça o resto.]

7. Considere que o modelo linear, escrito em notação de matriz,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

satisfaça as Hipóteses E.1, E.2 e E.3. Divida o modelo como

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u},$$

em que $\mathbf{X}_1$ é $n \times (k_1 + 1)$ e $\mathbf{X}_2$ é $n \times k_2$.

- (i) Considere a seguinte proposta de estimação de β_2 . Primeiro regrida y em X_1 e obtenha os resíduos, digamos, $\check{y}$. Depois, regrida $\check{y}$ em X_2 para obter $\check{\beta}_2$. Mostre que $\check{\beta}_2$ é geralmente viesado e mostre qual é o viés. [Você deve encontrar $E(\check{\beta}_2|\mathbf{X})$ em termos de β_2 , X_2 e a matriz $\mathbf{M}_1$ de realização residual].
- (ii) Como um caso especial, escreva

$$y = X_1\beta_1 + \beta_k X_k + u,$$

em que X_k é um vetor $n \times 1$ na variável x_{ik} . Mostre que

$$E(\check{\beta}_k|\mathbf{X}) = \left(\frac{SQR_k}{\sum_{t=1}^n x_{tk}^2} \right) \beta_k,$$

em que SQR_k é a soma dos resíduos quadrados da regressão de x_{ik} em $1, x_{i1}, x_{i2}, \dots, x_{i, k-1}$. Como o fator multiplicador de β_k nunca é maior que um?

- (iii) Suponha que você conheça β_1 . Mostre que a regressão $y - X_1\beta_1$ em X_1 produz um estimador não viesado de β_2 (condicional em $\mathbf{X}$).

Respostas das questões dos capítulos

Capítulo 2

Questão 2.1. Quando a habilidade do aluno, sua motivação, a idade e outros fatores em u não se relacionam com assiduidade, a equação (2.6) se sustenta. Esse não parece ser o caso.

Questão 2.2. Cerca de \$ 11,05. Para obter esse resultado, extraído dos salários médios em dólares, computados em 1976 e 2003, podemos chegar ao deflator de CPI, $19,06/5,90 \approx 3,23$. Quando multiplicamos 3,42 por 3,23, obtemos cerca de 11,05.

Questão 2.3. 54,65, como podemos ver associando $shareA = 60$ à equação (2.28). Isso não parece injustificado: se o candidato A gasta 60% do total de dinheiro investido, prevê-se que ele receba quase 55% de votos.

Questão 2.4. A equação será $\widehat{salary}_{hun} = 9.631,91 + 185,01 roe$, como é possível ver claramente ao multiplicarmos a equação (2.39) por 10.

Questão 2.5. A equação (2.58) pode ser escrita como $Var(\hat{\beta}_0) = (\sigma^2 n^{-1}) (\sum_{i=1}^n x_i^2) / (\sum_{i=1}^n (x_i - \bar{x})^2)$, em que o termo multiplicador $\sigma^2 n^{-1}$ é maior ou igual a um, mas é igual a um se, e apenas se, $\bar{x} = 0$. Nesse caso, a variação é a menor possível: $Var(\hat{\beta}_0) = \sigma^2/n$.

Capítulo 3

Questão 3.1. Apenas poucos fatores incluem idade e distribuição por gênero, tamanho da força policial (ou, de forma mais geral, os recursos destinados à luta contra o crime), população e fatores históricos gerais. Esses fatores certamente podem estar correlacionados com $prbcond$ e $senmed$, o que significa que a equação (3.5) não se sustenta. Por exemplo, o tamanho da força policial possivelmente é correlacionado com $prbconv$ e $avgsen$, já que algumas cidades dedicam-se mais à prevenção de crimes e manutenção da lei. Devemos tentar trazer o máximo possível desses fatores para a equação.

Questão 3.2. Usamos a terceira propriedade de MQO em relação a valores e resíduos previstos: quando associamos os valores médios de todas as variáveis independentes à linha de regressão MQO, obtemos o valor médio da variável dependente. Então, $\overline{colGPA} = 1,29 + 0,453 \overline{hsGPA} + 0,0094 \overline{ACT} = 1,29 + 0,453(3,4) + 0,0094(24,2) \approx 3,06$. Você pode checar a média para verificar a segunda casa decimal.

Questão 3.3. Não. A variável *shareA* não é uma função linear exata de *expendA* e *expendB*, mesmo se for uma função não linear exata: $shareA = 100 \cdot [expendA / (expendA + expendB)]$. Portanto, é legítimo ter *expendA*, *expendB* e *shareA* como variáveis explicativas.

Questão 3.4. Como discutido na Seção 3.4, se estivermos atentos ao efeito de x_1 em y , a correlação entre as outras variáveis explicativas (x_2 , x_3 , e assim por diante) não afetará $\text{Var}(\hat{\beta}_1)$. Essas variáveis são incluídas como controles, dessa forma, não precisamos nos preocupar com colinearidade entre as variáveis de controle. É claro, nós as controlamos principalmente porque achamos que estão correlacionadas com assiduidade, mas isso é necessário para desempenhar uma análise *ceteris paribus*.

Capítulo 4

Questão 4.1. Sob essas hipóteses, as hipóteses de Gauss-Markov são demonstradas: u é independente das variáveis explicativas, então $E(u|x_1, \dots, x_k) = E(u)$ e $E(u|x_1, \dots, x_k) = \text{Var}(u)$. Além disso, pode-se ver facilmente que $E(u) = 0$. Portanto, RLM.4 e RLM.5 se sustentam. As hipóteses do modelo linear clássico não são demonstradas porque u não é distribuído normalmente (o que é uma violação de RLM.6).

Questão 4.2. $H_0: \beta_1 = 0$, $H_1: \beta_1 < 0$.

Questão 4.3. Como $\hat{\beta}_1 = 0,56 > 0$ e estamos testando-a com relação a $H_1: \beta_1 > 0$, o p -valor unilateral é metade do valor bilateral do p -valor, ou 0,043.

Questão 4.4. $H_0: \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$. $k = 8$ e $q = 4$. A versão restrita do modelo é

$$score = \beta_0 + \beta_1 classize + \beta_2 expend + \beta_3 tchcomp + \beta_4 enroll + u.$$

Questão 4.5. A estatística F para testar a exclusão de *ACT* é $[(0,291 - 0,183)/(1 - 0,291)](680 - 3) \approx 103,13$. Portanto, o valor absoluto da estatística t é cerca de 10,16. A estatística t em *ACT* é negativa, porque $\hat{\beta}_{ACT}$ é negativo, de forma que $t_{ACT} = -10,16$.

Questão 4.6. Nem tanto. O teste F para significância conjunta de *droprate* e *gradrate* é facilmente computável de acordo com os R -quadrados da tabela: $F = [(0,361 - 0,353)/(1 - 0,361)](402/2) \approx 2,52$. O valor crítico de 10% é obtido da Tabela G.3a como 2,30, enquanto o valor crítico 5% da Tabela G.3b é 3. O p -valor é cerca de 0,082. Logo, *droprate* e *gradrate* são significativamente conjuntos em um nível de 10%, mas não em um nível de 5%. Em todo caso, controlar essas variáveis afeta pouco o coeficiente b/s .

Capítulo 5

Questão 5.1. Isso exige algumas hipóteses. Parece razoável presumir que $\beta_2 > 0$ (*score* depende positivamente de *priGPA*) e $\text{Cov}(skipped, priGPA) < 0$ (*skipped* e *priGPA* são

negativamente correlacionados). Isso significa que $\beta_2\delta_1 < 0$, ou seja, $\tilde{\beta}_1 < \beta_1$. Como β_1 é idealizado para ser negativo (ou ao menos não positivo), é possível que uma regressão simples superestime a importância de “pular” aulas.

Questão 5.2. $\hat{\beta}_j \pm 1,96\text{ep}(\hat{\beta}_j)$ é o intervalo de confiança assintótico de 95%. Ou podemos substituir 1,96 por 2.

Capítulo 6

Questão 6.1. Como $\text{fincdol} = 1.000 \cdot \text{faminc}$, o coeficiente em *fincdol* será o coeficiente em *faminc* dividido por 1.000 ou $0,0927/1.000 = 0,0000927$. O erro padrão também diminui por um fator de 1.000, dessa forma, a estatística *t* não muda, nem nenhuma das outras estatísticas de MQO. Para legibilidade, é melhor medir renda familiar em milhares de dólares.

Questão 6.2. Podemos fazer isso de forma geral. A equação é

$$\log(y) = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \dots,$$

em que x_2 é uma proporção, em vez de um percentual. Logo, *ceteris paribus*, $\Delta \log(y) = \beta_2 \Delta x_2$, $100 \cdot \Delta \log(y) = \beta_2 (100 \cdot \Delta x_2)$ ou $\% \Delta y \approx \beta_2 (100 \cdot \Delta x_2)$. Agora, já que Δx_2 é a mudança na proporção, $100 \cdot \Delta x_2$ é a mudança de ponto percentual. Em particular, se $\Delta x_2 = 0,01$, então $100 \cdot \Delta x_2 = 1$, que corresponde a uma mudança de ponto percentual. Contudo, logo, β_2 é a mudança percentual em *y* quando $100 \cdot \Delta x_2 = 1$.

Questão 6.3. O novo modelo seria $\text{stndfnl} = \beta_0 + \beta_1 \text{atndrte} + \beta_2 \text{priGPA} + \beta_3 \text{ACT} + \beta_4 \text{priGPA}^2 + \beta_5 \text{ACT}^2 + \beta_6 \text{priGPA} \cdot \text{atndrte} + \beta_7 \text{ACT} \cdot \text{atndrte} + u$. Portanto, o efeito parcial de $\beta_1 + \beta_6 \text{priGPA} + \beta_7 \text{ACT}$. Isso é o que multiplicamos por $\Delta \text{atndrte}$ para obter a mudança *ceteris paribus* em *stndfnl*.

Questão 6.4. Da equação (6.21), $\bar{R}^2 = 1 - \hat{\sigma}^2 / [\text{SQT}/(n - 1)]$. Para uma dada amostra e variável dependente, fixa-se $\text{SQT}/(n - 1)$. Quando usamos diferentes conjuntos de variáveis explicativas, apenas $\hat{\sigma}^2$ muda. Conforme $\hat{\sigma}^2$ diminui, $\bar{R}^2$ aumenta. Se fizermos $\hat{\sigma}$, e, logo, $\hat{\sigma}^2$, o menor possível, estamos tornando $\bar{R}^2$ o maior possível.

Questão 6.5. Uma possibilidade é coletar dados sobre rendimentos anuais para uma amostra de atores, juntamente com a rentabilidade dos filmes nos quais trabalharam. Em uma análise de regressão simples, podemos relatar os resultados para rentabilidade. Mas provavelmente devemos controlar outros fatores que possam afetar salário, como idade, gênero e tipo de filme em que os atores trabalham. São analisados, no Capítulo 7, métodos para incluir fatores qualitativos nos modelos de regressão.

Capítulo 7

Questão 7.1. Não, porque não ficaria claro quando *partido* fosse um e quando fosse zero. Um nome melhor seria algo como *Dem*, ou seja, um para candidatos democratas e zero para republicanos. Ou *Rep*, um para republicanos e zero para democratas.

Questão 7.2. Com *outfield* como grupo de base, podemos incluir as variáveis *dummy frstbase*, *scndbase*, *thrdbase*, *shrtstop* e *catcher*.

Questão 7.3. A nula, nesse caso, é $H_0: \delta_1 = \delta_2 = \delta_3 = \delta_4 = 0$, de forma que há quatro restrições. Como de hábito, usáramos um teste F (em que $q = 4$ e em que k depende do número de outras variáveis explicativas).

Questão 7.4. Como *tenure* parece ser quadrática, deveríamos permitir quadráticas separadas para homens e mulheres. Ou seja, teríamos que adicionar as variáveis explicativas *female · tenure* e *female · tenure*².

Questão 7.5. Associamos $pcnv = 0$, $avgse = 0$, $totttime = 0$, $pstime86 = 0$, $qemp86 = 4$, $black = 1$ e $hispan = 0$ na equação (7.31): $\widehat{arr86} = 0,380 - 0,038(4) + 0,170 = 0,398$ ou quase 0,4. É difícil dizer se isso é “aceitável”. Para alguém sem convicções prévias, que conseguiu emprego durante o ano, essa estimativa pode parecer alta, mas lembre-se de que a população consiste em homens que já foram presos pelo menos uma vez antes de 1986.

Capítulo 8

Questão 8.1. Essa declaração é claramente falsa. Por exemplo, na equação (8.7), o erro padrão habitual para *black* é 0,147, enquanto o erro padrão robusto com relação à heteroscedasticidade é 0,118.

Questão 8.2. O teste F seria obtido regredindo $\hat{u}^2$ on *marrmale*, *marrfem*, e *singfem* (*singmale* é o grupo-base). Com $n = 526$ e três variáveis independentes nessa regressão, o gl é 3 e 522.

Questão 8.3. Certamente, o resultado do teste de estatística sugere algum motivo para preocupação. Uma estatística t de 2,96 é muito significativa e implica que existe heteroscedasticidade nos recursos da equação. Como uma questão prática, sabemos que o erro padrão de MQP, 0,063, está substancialmente abaixo do erro padrão robusto com relação à heteroscedasticidade para MQO, 0,104, e logo, a heteroscedasticidade parece ser importante, de forma prática (além disso, o erro padrão MQO não robusto é 0,061, o que é muito animador). Portanto, mesmo que simplesmente ajustemos o erro padrão MQO para heteroscedasticidade de forma desconhecida, não há implicações triviais).

Questão 8.4. O 1% de valor crítico na distribuição F com $(2, \infty)$ gl é 4,61. Uma estatística F de 11,15 está bem acima do 1% de valor crítico; dessa forma, rejeitamos fortemente a hipótese nula de que os erros transformados, $u_i/\sqrt{h_i}$, são homoscedásticos (na verdade, o p -valor é menor que 0,00002, que é obtido da distribuição $F_{2,804}$. Isso significa que nosso modelo de $\text{Var}(u|\mathbf{x})$ é inadequado para eliminar totalmente a heteroscedasticidade em u .

Capítulo 9

Questão 9.1. São variáveis binárias e nivelá-las não causa nenhum efeito: $black^2 = black$, e $hispan^2 = hispan$.

Questão 9.2. Quando $educ \cdot IQ$ está na equação, o coeficiente em $educ$, digamos, β_1 , mede o efeito de $educ$ em $\log(wage)$, quando $IQ = 0$ (o efeito parcial de educação é $\beta_1 + \beta_1 IQ$). Não há ninguém na população de interesse com um QI próximo de zero. Levando em conta o QI médio da população, que é 100, o retorno estimado da

educação da coluna (3) é $0,018 + 0,00034(100) = 0,052$, que é quase o que obtemos como o coeficiente em *educ* na coluna (2).

Questão 9.3. Não. Se $educ^*$ for um número inteiro – o que significa que alguém não possui formação além do primeiro grau completo – o erro de medida é zero. Se $educ^*$ não for um número inteiro, $educ < educ^*$, dessa forma, o erro de medida será negativo. No mínimo, e_1 não pode ser zero, além disso, e_1 e $educ^*$ provavelmente são correlacionados.

Questão 9.4. A decisão do candidato de não concorrer pode estar sistematicamente relacionada a como ele espera se sair nas eleições. Logo, podemos ter apenas uma amostra de candidatos mais fortes, na média, do que todos os possíveis concorrentes. Isso resulta em um problema de amostra de seleção, caso a população de interesse inclua todos os candidatos. Se nos concentramos apenas nos efeitos dos gastos de campanha sobre os resultados da eleição para candidatos que buscam a reeleição, não existe um problema de seleção de amostra.

Capítulo 10

Questão 10.1. A propensão de impacto é 0,48, enquanto a propensão de longo prazo é $0,48 - 0,15 + 0,32 = 0,65$.

Questão 10.2. As variáveis explicativas são $x_{t1} = z_t$ e $x_{t2} = z_{t-1}$. A ausência de uma colinearidade perfeita significa que eles não podem ser constantes e que não há uma relação linear exata entre eles na amostra. Isso exclui a possibilidade de que todos os $z_1, \dots, z_n$ assumam o mesmo valor ou de que $z_0, z_1, \dots, z_{n-1}$ assumam o mesmo valor. Contudo, isso também elimina outros padrões. Por exemplo, se $z_t = a + bt$ para constantes a e b , logo $z_{t-1} = a + b(t-1) = (a + bt) - b = z_t - b$, o que é uma função linear perfeita de z_t .

Questão 10.3. Se $\{z_t\}$ se move lentamente com o tempo – como é o caso de níveis ou logs de muitas séries temporais econômicas – logo, z_t e z_{t-1} podem ser altamente correlacionados. Por exemplo, a correlação entre $unem_t$ e $unem_{t-1}$ em PHILLIPS é 0,75.

Questão 10.4. Não, porque uma tendência temporal linear com $\alpha_1 < 0$ se torna cada vez mais negativa conforme t fica maior. Já que *gfr* não pode ser negativa, uma tendência temporal linear com coeficiente de tendência negativo não pode representar *gfr* em todos os futuros períodos temporais.

Questão 10.5. O intercepto para março é $\beta_0 + \delta_2$. Variáveis *dummy* sazonais são estritamente exógenas porque seguem determinado padrão. Por exemplo, os meses não mudam com base na mudança de variáveis explicativas ou de variáveis dependentes.

Capítulo 11

Questão 11.1. (i) Não, porque $E(y_t) = \delta_0 + \delta_1 t$ depende de t . (ii) Sim, porque $y_t - E(y_t) = e_t$ é uma sequência i.i.d.

Questão 11.2. Associamos $\inf_t^e = (1/2)\inf_{t-1} + (1/2)\inf_{t-2}$ com $\inf_t - \inf_t^e = \beta_1 (unem_t - \mu_0) + e_t$ e rearranjamos: $\inf_t - (1/2)(\inf_{t-1} + \inf_{t-2}) = \beta_0 + \beta_1 unem_t + e_t$, em que $\beta_0 = -\beta_1$, como anteriormente. Portanto, poderíamos regredir y_t em $unem_t$, em que $y_t = \inf_t - (1/2)(\inf_{t-1} + \inf_{t-2})$. Note que perdemos as primeiras duas observações ao construir y_t .

Questão 11.3. Não, porque u_t e u_{t-1} são correlacionados. Em particular, $\text{Cov}(u_t, u_{t-1}) = E[(e_t + \alpha_1 e_{t-1})(e_{t-1} + \alpha_1 e_{t-2})] = \alpha_1 E(e_{t-1}^2) = \alpha_1 \sigma_e^2 \neq 0$ se $\alpha_1 \neq 0$. Se os erros estiverem serialmente correlacionados, o modelo não pode ser dinamicamente completo.

Capítulo 12

Questão 12.1. Usamos a equação (12.4). Agora, apenas termos adjacentes são correlacionados. Em particular, a covariação entre $x_t u_t$ e $x_{t+1} u_{t+1}$ é $x_t x_{t+1} \text{Cov}(u_t, u_{t+1}) = x_t x_{t+1} \alpha \sigma_e^2$. Logo, a fórmula é

$$\begin{aligned} \text{Var}(\hat{\beta}_1) &= \text{SQT}_x^{-2} \left(\sum_{t=1}^n x_t^2 \text{Var}(u_t) + 2 \sum_{t=1}^{n-1} x_t x_{t+1} E(u_t u_{t+1}) \right) \\ &= \sigma^2 / \text{SQT}_x + (2 / \text{SQT}_x^2) \sum_{t=1}^{n-1} \alpha \sigma_e^2 x_t x_{t+1} \\ &= \sigma^2 / \text{SQT}_x + \alpha \sigma_e^2 (2 / \text{SQT}_x^2) \sum_{t=1}^{n-1} x_t x_{t+1}, \end{aligned}$$

em que $\sigma^2 = \text{Var}(u_t) = \sigma_e^2 + \sigma_1^2 \sigma_e^2 = \sigma_e^2 (1 + \sigma_1^2)$. A menos que x_t e x_{t+1} sejam correlacionados na amostra, o segundo termo não resulta em zero, sempre que $\alpha_1 \neq 0$. Perceba que, se x_t e x_{t+1} forem positivamente correlacionados e $\alpha < 0$, a verdadeira variação será, de fato, *menor* do que a variação habitual. Quando a equação estiver em níveis (como oposta à diferenciação), o caso típico é $\alpha > 0$, com correlação positiva entre x_t e x_{t+1} .

Questão 12.2. $\hat{\rho} \pm 1,96 \text{ep}(\hat{\rho})$, em que $\text{ep}(\hat{\rho})$ é o erro padrão reportado na regressão. Ou podemos usar o erro padrão robusto em relação à heteroscedasticidade. Mostrar que isso é assintoticamente válido é complicado, já que os resíduos MQO dependem de $\hat{\beta}_j$; contudo, isso pode ser feito.

Questão 12.3. O modelo que temos em mente é $u_t = \rho_1 u_{t-1} + \rho_4 u_{t-4} + e_t$ e queremos testar $H_0: \rho_1 = 0, \rho_4 = 0$ contra a alternativa de que H_0 é falsa. Organizamos a regressão de $\hat{u}_t$ em $\hat{u}_{t-1}$ e $\hat{u}_{t-4}$ para obter a estatística F usual para significância conjunta de duas defasagens (estamos testando duas restrições).

Questão 12.4. Provavelmente estimaríamos a equação usando primeiras diferenças, conforme $\hat{\rho} = 0,92$ for próxima o suficiente de um para levantar questões sobre regressão de níveis. Ver o Capítulo 18 para mais discussões.

Questão 12.5. Como só existe uma variável explicativa, o teste White é fácil de calcular. Simplesmente regreda $\hat{u}_t^2$ em return_{t-1} e return_{t-1}^2 (com um intercepto, como sempre) e calcule o teste F para significância conjunta de return_{t-1} e return_{t-1}^2 . Se forem

significativamente conjuntos em um nível de significância pequeno o bastante, rejeitaremos a hipótese nula de homoscedasticidade.

Capítulo 13

Questão 13.1. Sim, presumindo que controlamos todos os fatores relevantes. O coeficiente em *black* é 1,076 e, com um erro padrão de 0,174, não é estatisticamente diferente de 1. O intervalo de confiança de 95% varia de 0,735 a 1,417.

Questão 13.2. O coeficiente em *highearn* mostra que, na ausência de qualquer mudança no limite de ganhos, pessoas com rendimentos elevados gastam muito mais tempo – da ordem de 29,2% em média [porque $\exp(0,256) - 1 \approx 0,292$] – da compensação dos trabalhadores.

Questão 13.3. Em primeiro lugar, $E(v_{i1}) = E(a_i + u_{i1}) = E(a_i) + E(u_{i1}) = 0$. Similarmente, $E(v_{i2}) = 0$. Portanto, a covariância entre v_{i1} e v_{i2} é simplesmente $E(v_{i1} v_{i2}) = E[(a_i + u_{i1})(a_i + u_{i2})] = E(a_i^2) + E(a_i u_{i1}) + E(a_i u_{i2}) + E(u_{i1} u_{i2}) = E(a_i^2)$, já que todos os termos de covariância são zero, por hipótese. Contudo, $E(a_i^2) = \text{Var}(a_i)$, porque $E(a_i) = 0$. Isso causa correlação serial positiva com o tempo nos erros dentro de cada i , que enviesará os erros padrão de MQO usuais em uma regressão de MQO agrupado.

Questão 13.4. Como $\Delta \text{admn} = \text{admn}_{90} - \text{admn}_{85}$ é a diferença em indicadores binários, pode ser -1 se, e apenas se, $\text{admn}_{90} = 0$ e $\text{admn}_{85} = 1$. Em outras palavras, o estado de Washington teve uma lei administrativa *per se* em 1985, que foi revogada em 1990.

Questão 13.5. Não, assim como não causa polarização e inconsistência em uma regressão de série temporal com variáveis explicativas estritamente exógenas. Há dois motivos para se preocupar com isso. Primeiro, a correlação serial nos erros em qualquer equação geralmente enviesará os erros padrão de MQO e estatísticas de teste usuais. Em segundo lugar, isso significa que MQO agrupados não é eficiente como estimador responsável pela correlação serial (como no Capítulo 12).

Capítulo 14

Questão 14.1. Mesmo se usarmos a primeira diferença ou a transformação intergrupo, teremos problema em estimar o coeficiente em $kids_{it}$. Por exemplo, usando a transformação intergrupo, se $kids_{it}$ não variar por família i , então, $\dot{kids}_{it} = kids_{it} - \bar{kids}_i = 0$ para $t = 1, 2, 3$. Desde que algumas famílias tenham variação em $kids_{it}$, poderemos, então, calcular o estimador de efeitos fixos; contudo, o coeficiente $kids$ pode ser estimado de forma muito imprecisa. Essa é uma forma de multicolinearidade na estimação de efeitos fixos (ou estimação de primeiras diferenças).

Questão 14.2. Se uma empresa não receber um subsídio no primeiro ano, poderá recebê-lo ou não no segundo ano. Mas, se uma empresa receber um subsídio no primeiro ano, pode não recebê-lo no segundo ano. Ou seja, se $grant_{-1} = 1$, logo $grant = 0$. Isso induz uma correlação negativa entre $grant$ e $grant_{-1}$. Podemos verificar isso ao calcular uma regressão de $grant$ em $grant_{-1}$, usando os dados em JTRAIN de 1989. Utilizando todas as empresas na amostra, obtemos

$$\widehat{grant} = 0,248 - 0,248 \text{ grant}_{-1}$$

$$(0,035) \quad (0,072)$$

$$n = 157, R^2 = 0,070.$$

O coeficiente em $grant_{-1}$ deve ser o negativo do intercepto, porque $\widehat{grant} = 0$ quando $grant_{-1} = 1$.

Questão 14.3. Isso sugere que o efeito não observado a_i é positivamente correlacionado com $union_{it}$. Lembre-se, o MQO agrupado deixa a_i no termo de erro, enquanto efeitos fixos removem a_i . Por definição, a_i tem efeito positivo em $\log(wage)$. Pela análise de variáveis omitidas padrão (ver Capítulo 3), MQO tem um viés ascendente quando a variável explicativa ($union$) é positivamente correlacionada com a variável omitida (a_i). Logo, fazer parte de um sindicato parece estar positivamente relacionado a constantes de tempo, fatores não observados que afetam salário.

Questão 14.4. A resposta é não, se todas as irmãs de uma família tiverem a mesma mãe e o mesmo pai. Logo, como as variáveis de raça dos pais não mudariam de acordo com cada irmã, seriam diferenciadas na equação (14.13).

Capítulo 15

Questão 15.1. Provavelmente não. Na Equação simples (15.18), anos de educação fazem parte do termo de erro. Se alguns homens que receberam números baixos no sorteio obtiverem escolaridade adicional, então esses números e a educação serão negativamente correlacionados, o que viola a primeira exigência para uma variável instrumental na equação (15.4).

Questão 15.2. (i) Para a equação (15.27), exigimos que os efeitos de grupos de colegas do ensino médio sejam transferidos para a universidade. Em outras palavras, para determinada nota no SAT, um estudante que frequentou o ensino médio em uma escola em que fumar maconha era popular fumaria maconha na faculdade. Mesmo se a equação de condição de identificação (15.27) se mantiver, a ligação pode ser fraca.

(ii) Temos de supor que a porcentagem de estudantes que fumam maconha no ensino médio não é correlacionada com fatores não observados que afetam a nota média na faculdade. Embora estejamos, de certa forma, controlando a qualidade no ensino médio ao incluir SAT na equação, isso pode não ser suficiente. Talvez escolas que tenham preparado melhor os alunos para a universidade também tenham menos estudantes fumando maconha. Ou o uso de maconha pode ser correlacionado com níveis médios de renda. Essas são, é claro, questões empíricas que podemos, ou não, ser capazes de responder.

Questão 15.3. Embora a prevalência da NRA e assinaturas de revistas sobre armas estejam provavelmente correlacionadas com a presença de legislação de controle de armas, não é óbvio que elas não são correlacionadas com fatores não observados que afetam a taxa de crimes violentos. Na verdade, podemos alegar que uma população interessada em armas seja um reflexo das altas taxas de criminalidade, e que controlar variáveis econômicas e demográficas não é suficiente para capturar isso. Seria difícil argumentar de forma persuasiva que isso é verdadeiramente exógeno na equação de crimes violentos.

Questão 15.4. Como sempre, há duas exigências. Primeiro, poderia ser o caso em que o crescimento dos gastos do governo estivesse sistematicamente relacionado ao partido do presidente, depois de compensados a taxa de investimentos e o crescimento na força de trabalho. Em outras palavras, o instrumento deve ser parcialmente correlacionado com a variável explicativa endógena. Embora possamos pensar que os gastos governamentais cresçam mais lentamente com presidentes Republicanos, isso certamente não foi sempre verdade nos Estados Unidos e teria que ser testado usando a estatística t em REP_{t-1} na forma reduzida $gGov_t = \pi_0 + \pi_1 REP_{t-1} + \pi_2 INVRAT_t + \pi_3 gLAB_t + v_t$. Precisamos supor que o partido do presidente não tem nenhum efeito separado sobre $gGDP$. Isso seria violado se, por exemplo, as políticas monetárias diferissem sistematicamente por partido político e teria um efeito separado sobre o crescimento do PIB.

Capítulo 16

Questão 16.1. Provavelmente não. É porque as empresas escolhem preço e gastos com propaganda conjuntamente que não estamos interessados no experimento em que, digamos, a propaganda muda de forma exógena, mas queremos saber os efeitos sobre o preço. Em vez disso, modelaríamos preço e propaganda separadamente como uma função de demanda e variáveis de custo. Isso é o que surge da teoria econômica.

Questão 16.2. Precisamos supor duas coisas. Primeiro, o crescimento da base monetária deve aparecer na equação (16.22), por isso, é parcialmente correlacionado com inf . Segundo, devemos supor que o crescimento da base monetária não apareça na equação (16.23). Se pensarmos que devemos incluir o crescimento da base monetária na equação (16.23), ainda assim teremos falta de um instrumento para inf . É claro que a hipótese de o crescimento da base monetária ser exógeno também pode ser questionada.

Questão 16.3. Use o teste de Hausman do Capítulo 15. Em particular, defina $\hat{v}_2$ como os resíduos de MQO da regressão de forma reduzida de $abertura$ sobre $\log(pcinc)$ e $\log(land)$. Em seguida use uma regressão por MQO de inf sobre $open$, $\log(pcinc)$ e $\hat{v}_2$ e calcule a estatística t para a significância de $\hat{v}_2$. Se $\hat{v}_2$ for significativo, as estimativas de MQ2E e MQO serão estatisticamente diferentes.

Questão 16.4. A equação de demanda é

$$\log(fish_t) = \beta_0 + \beta_1 \log(prcfish_t) + \beta_2 \log(inc_t) + \beta_3 \log(prchick_t) + \beta_4 \log(prcbeef_t) + u_{1t},$$

onde os logaritmos são usados para que todas as elasticidades sejam constantes. Por suposição, a função de demanda não contém sazonalidade, assim, a equação não contém variáveis *dummy* mensais (por exemplo, feb_t , mar_t , ..., dec_t , com janeiro como o mês-base). Além disso, por suposição, a oferta de peixes é sazonal, o que significa que a função de oferta depende de pelo menos uma das variáveis *dummy* mensais. Mesmo sem resolver a forma reduzida de $\log(prcfish)$, podemos concluir que ela depende das variáveis *dummy* mensais. Como são exógenas, elas podem ser usadas como instrumentos para $\log(prcfish)$ na equação de demanda. Assim, podemos estimar a equação de demanda por peixes usando *dummies* mensais como VIs de $\log(prcfish)$.

A identificação exige que pelo menos uma variável *dummy* mensal apareça com coeficiente diferente de zero na forma reduzida de $\log(\text{prcfish})$.

Capítulo 17

Questão 17.1. $H_0: \beta_4 = \beta_5 = \beta_6 = 0$, assim, existem três restrições e, portanto, três *gl* no teste RV ou de Wald.

Questão 17.2. Precisamos da derivada parcial de $\Phi(\hat{\beta}_0 + \hat{\beta}_1 \text{nwifcinc} + \hat{\beta}_2 \text{educ} + \hat{\beta}_3 \text{exper} + \hat{\beta}_4 \text{exper}^2 + \dots)$ em relação a *exper*, que é $\phi(\cdot) (\hat{\beta}_3 + 2\hat{\beta}_4 \text{exper})$, em que $\phi(\cdot)$ é avaliado pelos valores dados e pelo nível de experiência inicial. Assim, precisamos avaliar a densidade de probabilidade normal padrão em $0,270 - 0,012(20,13) + 0,131(12,3) + 0,123(10) - 0,0019(10^2) - 0,053(42,5) - 0,868(0) + 0,036(1) \approx 0,463$, onde inserimos o nível inicial de experiência (10). Mas, se $\phi(0,463) = (2\pi)^{-1/2} \exp[-(0,463^2)/2] \approx 0,358$. Em seguida, multiplicamos isso por $\hat{\beta}_3 + 2\hat{\beta}_4 \text{exper}$, que é avaliada com *exper* = 10. O efeito parcial usando aproximação de cálculo é $0,358[0,123 - 2(0,0019)(10)] \approx 0,030$. Em outras palavras, com os valores dados das variáveis explicativas, e partindo de *exper* = 10, o ano seguinte de experiência aumentará a probabilidade de participação da força de trabalho em cerca de 0,03.

Questão 17.3. Não. O número de relações extraconjugais é um inteiro não negativo, que supostamente assume o valor zero ou pequenos números para uma fração substancial da população. Não é realista usar um modelo tobit, que, embora permita um acúmulo em zero, trata *y* como continuamente distribuída ao longo de valores positivos. Formalmente, supor que $y = \max(0, y^*)$, onde y^* é normalmente distribuída, é incompatível com a natureza discreta do número de relações extraconjugais quando $y > 0$.

Questão 17.4. Os erros padrão ajustados são os erros padrão da EMV de Poisson multiplicados por $\hat{\sigma} = \sqrt{2} \approx 1,41$; assim, os erros padrão serão cerca de 41% mais altos. A estatística quase-RV é a estatística RV usual dividida por $\hat{\sigma}^2$, de forma que ela será metade da estatística RV comum.

Questão 17.5: Por suposição, $mvp_i = \beta_0 + \mathbf{x}_i \boldsymbol{\beta} + u_i$, onde, como ocorre normalmente, $\mathbf{x}_i \boldsymbol{\beta}$ indica uma função linear das variáveis exógenas. Agora, o salário observado é o maior dos salários mínimos e o produto de valor marginal, assim, $\text{wage}_i = \max(\text{minwage}_i, mvp_i)$, muito similar à equação (17.34), exceto porque o operador *max* substituiu o operador *min*.

Capítulo 18

Questão 18.1. Podemos inserir esses valores diretamente na equação (18.1) e obter expectativas. Primeiro, como $z_s = 0$, para todo $s < 0$, $y_{-1} = \alpha + u_{-1}$. Assim, $z_0 = 1$, então $y_0 = \alpha + \delta_0 + u_0$. Para $h \geq 1$, $y_h = \alpha + \delta_{h-1} + \delta_h + u_h$. Como os erros têm valores esperados zero, $E(y_{-1}) = \alpha$, $E(y_0) = \alpha + \delta_0$ e $E(y_h) = \alpha + \delta_{h-1} + \delta_h$, para todo $h \geq 1$. À medida que $h \rightarrow \infty$, $\delta_h \rightarrow 0$. Como consequência, $E(y_h) \rightarrow \alpha$ à medida que $h \rightarrow \infty$, ou seja, o valor esperado de y_h retorna ao valor esperado antes do aumento de z , no tempo zero. Isso faz sentido: embora o aumento de z tenha durado dois períodos, ele ainda é um aumento temporário.

Questão 18.2. Sob a configuração descrita, Δy_t e Δx_t são sequências i.i.d. que são independentes umas das outras. Δy_t e Δx_t são não correlacionadas. Se $\hat{\gamma}_t$ é o coeficiente de inclinação da regressão de Δy_t sobre Δx_t , $t = 1, 2, \dots, n$, então $\text{plim } \hat{\gamma}_t = 0$. É assim que deve ser, já que estamos regredindo um processo $I(0)$ sobre outro processo $I(0)$ e eles são não correlacionados. Escrevemos a equação $\Delta y_t = \gamma_0 + \gamma_1 \Delta x_t + e_t$, onde $\gamma_0 = \gamma_1 = 0$. Como $\{e_t\}$ é independente de $\{\Delta x_t\}$, a hipótese de exogeneidade estrita se mantém. Além disso, $\{e_t\}$ é serialmente não correlacionado e homoscedástico. Segundo o Teorema 11.2, do Capítulo 11, a estatística t de $\hat{\gamma}_t$ tem uma distribuição normal padrão aproximada. Se e_t é normalmente distribuído, as hipóteses do modelo linear clássico se mantêm, e a estatística t tem uma distribuição t exata.

Questão 18.3. Escreva $x_t = x_{t-1} + a_t$, onde $\{a_t\}$ é $I(0)$. Por suposição, há uma combinação linear, digamos, $s_t = y_t - \beta x_t$, que é $I(0)$. Agora, $y_t - \beta x_{t-1} = y_t - \beta(x_t - a_t) = s_t + \beta a_t$. Já que s_t e a_t são $I(0)$ por suposição, também o é $s_t + \beta a_t$.

Questão 18.4. Use somente a forma de soma dos quadrados dos resíduos do teste F e suponha homoscedasticidade. A SQR restrita é obtida a partir da regressão de $\Delta hy6_t - \Delta hy3_{t-1} + (hy6_{t-1} - hy3_{t-2})$ em uma constante. Note que a_0 é o único parâmetro a se estimar em $\Delta hy6_t = \alpha_0 + \gamma_0 \Delta hy3_{t-1} + \delta(hy6_{t-1} - hy3_{t-2})$ quando as restrições são impostas. A soma irrestrita dos quadrados dos resíduos é obtida pela equação (18.39).

Questão 18.5. Estamos ajustando duas equações: $\hat{y}_t = \hat{\alpha} + \hat{\beta}_t$ e $\hat{y}_t = \hat{\gamma} + \hat{\delta} year_t$. Podemos obter a relação entre os parâmetros notando que $year_t = t + 49$. Adicionar isso à segunda equação gera $\hat{y}_t = \hat{\gamma} + \hat{\delta}(t + 49) = (\hat{\gamma} + 49\hat{\delta}) + \hat{\delta}t$. A combinação da inclinação e do intercepto com a primeira equação gera $\hat{\delta} = \hat{\beta} -$ assim as inclinações de t e $year_t$ serão idênticas – e $\hat{\alpha} = \hat{\gamma} + 49\hat{\delta}$. Geralmente, quando usarmos $year$ em vez de t , o intercepto vai mudar, mas a inclinação não. (Você pode verificar isso usando um dos conjuntos de dados de séries temporais, como HSEINV ou INVEN) Usar t ou alguma medida de $year$ não muda os valores ajustados e, naturalmente, não muda previsões de valores futuros. O intercepto simplesmente se ajusta adequadamente às diferentes formas de incluir uma tendência na regressão.

APÊNDICE G

Tabelas estatísticas

TABELA G.1 Áreas acumuladas sob a distribuição normal padrão.

<i>z</i>	0	1	2	3	4	5	6	7	8	9
−3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
−2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
−2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
−2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
−2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
−2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
−2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
−2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
−2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
−2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
−2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
−1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
−1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
−1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
−1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
−1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
−1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
−1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
−1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
−1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
−1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
−0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
−0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
−0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
−0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
−0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
−0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
−0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483

(continua)

TABELA G.1 (Continuação) Áreas acumuladas sob a distribuição normal padrão.

<i>z</i>	0	1	2	3	4	5	6	7	8	9
-0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
-0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
-0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990

Exemplos: Se $Z \sim \text{Normal}(0,1)$, então $P(Z \leq -1,32) = 0,0934$ e $P(Z \leq 1,84) = 0,9671$.

Fonte: Esta tabela foi gerada usando a função normal do Stata®.

TABELA G.2 Valores críticos da distribuição *t*.

Nível de significância						
1-Monocaudal:		0,10	0,05	0,025	0,01	0,005
2-Bicaudal:		0,20	0,10	0,05	0,02	0,01
G r a u s d e L i b e r d a d e	1	3,078	6,314	12,706	31,821	63,657
	2	1,886	2,920	4,303	6,965	9,925
	3	1,638	2,353	3,182	4,541	5,841
	4	1,533	2,132	2,776	3,747	4,604
	5	1,476	2,015	2,571	3,365	4,032
	6	1,440	1,943	2,447	3,143	3,707
	7	1,415	1,895	2,365	2,998	3,499
	8	1,397	1,860	2,306	2,896	3,355
	9	1,383	1,833	2,262	2,821	3,250
	10	1,372	1,812	2,228	2,764	3,169
	11	1,363	1,796	2,201	2,718	3,106
	12	1,356	1,782	2,179	2,681	3,055
	13	1,350	1,771	2,160	2,650	3,012
	14	1,345	1,761	2,145	2,624	2,977
	15	1,341	1,753	2,131	2,602	2,947
	16	1,337	1,746	2,120	2,583	2,921
	17	1,333	1,740	2,110	2,567	2,898
	18	1,330	1,734	2,101	2,552	2,878
	19	1,328	1,729	2,093	2,539	2,861
	20	1,325	1,725	2,086	2,528	2,845
	21	1,323	1,721	2,080	2,518	2,831
	22	1,321	1,717	2,074	2,508	2,819
	23	1,319	1,714	2,069	2,500	2,807
	24	1,318	1,711	2,064	2,492	2,797
	25	1,316	1,708	2,060	2,485	2,787
	26	1,315	1,706	2,056	2,479	2,779
	27	1,314	1,703	2,052	2,473	2,771
	28	1,313	1,701	2,048	2,467	2,763
	29	1,311	1,699	2,045	2,462	2,756
	30	1,310	1,697	2,042	2,457	2,750
	40	1,303	1,684	2,021	2,423	2,704
	60	1,296	1,671	2,000	2,390	2,660
	90	1,291	1,662	1,987	2,368	2,632
	120	1,289	1,658	1,980	2,358	2,617
	∞	1,282	1,645	1,960	2,326	2,576

Exemplos: O valor crítico de 1% de um teste monocaudal com 25 *gl* é 2,485. O valor crítico de 5% de um teste bicaudal com *gl* grande (> 120) é 1,96.

Fonte: Esta tabela foi gerada usando a função *invttail* do Stata®.

TABELA G.3a Valores críticos de 10% da distribuição *F*.

		Graus de liberdade do numerador									
		1	2	3	4	5	6	7	8	9	10
G r a u s d e L i b e r d a d e d o D e n o m i n a d o r	10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32
	11	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25
	12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19
	13	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14
	14	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10
	15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06
	16	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03
	17	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00
	18	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00	1,98
	19	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98	1,96
	20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94
	21	2,96	2,57	2,36	2,23	2,14	2,08	2,02	1,98	1,95	1,92
	22	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93	1,90
	23	2,94	2,55	2,34	2,21	2,11	2,05	1,99	1,95	1,92	1,89
	24	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91	1,88
	25	2,92	2,53	2,32	2,18	2,09	2,02	1,97	1,93	1,89	1,87
	26	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88	1,86
	27	2,90	2,51	2,30	2,17	2,07	2,00	1,95	1,91	1,87	1,85
	28	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87	1,84
	29	2,89	2,50	2,28	2,15	2,06	1,99	1,93	1,89	1,86	1,83
	30	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82
	40	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76
	60	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74	1,71
	90	2,76	2,36	2,15	2,01	1,91	1,84	1,78	1,74	1,70	1,67
	120	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,68	1,65
	∞	2,71	2,30	2,08	1,94	1,85	1,77	1,72	1,67	1,63	1,60

Exemplos: O valor crítico de 10% dos *gl* do numerador = 2 e *gl* do denominador = 40 é 2,44.

Fonte: Esta tabela foi gerada usando a função invFtail do Stata®.

TABELA G.3b Valores críticos de 5% da distribuição *F*.

		Graus de Liberdade do Numerador									
		1	2	3	4	5	6	7	8	9	10
G r a u s d e L i b e r d a d e d o D e n o m i n a d o r	10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
	11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
	12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
	13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
	14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
	15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
	16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
	17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
	18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
	19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
	20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
	21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
	22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
	23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
	24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
	25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
	26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
	27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
	28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
	29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
	30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
	40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
	60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
	90	3,95	3,10	2,71	2,47	2,32	2,20	2,11	2,04	1,99	1,94
	120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91
	∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83

Exemplos: O valor crítico de 5% dos *gl* do numerador = 4 e *gl* do denominador grande (∞) é 2,37.
Fonte: Esta tabela foi gerada usando a função invFtail do Stata®.

TABELA G.3c Valores críticos de 1% da distribuição *F*.

		Graus de liberdade do numerador									
		1	2	3	4	5	6	7	8	9	10
G r a u s d e L i b e r d a d e d o D e n o m i n a d o r	10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85
	11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54
	12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30
	13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10
	14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94
	15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80
	16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69
	17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59
	18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51
	19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43
	20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37
	21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31
	22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26
	23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21
	24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17
	25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13
	26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09
	27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06
	28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03
	29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00
	30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98
	40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80
	60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63
	90	6,93	4,85	4,01	3,54	3,23	3,01	2,84	2,72	2,61	2,52
	120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47
	∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32

Exemplos: O valor crítico de 1% dos *gl* do numerador = 3 e *gl* do denominador = 60 é 4,13.

Fonte: Esta tabela foi gerada usando a função invFtail do Stata®.

TABELA G.4 Valores críticos da distribuição qui-quadrado.

		Nível de significância		
		0,10	0,05	0,01
G r a u s d e L i b e r d a d e	1	2,71	3,84	6,63
	2	4,61	5,99	9,21
	3	6,25	7,81	11,34
	4	7,78	9,49	13,28
	5	9,24	11,07	15,09
	6	10,64	12,59	16,81
	7	12,02	14,07	18,48
	8	13,36	15,51	20,09
	9	14,68	16,92	21,67
	10	15,99	18,31	23,21
	11	17,28	19,68	24,72
	12	18,55	21,03	26,22
	13	19,81	22,36	27,69
	14	21,06	23,68	29,14
	15	22,31	25,00	30,58
	16	23,54	26,30	32,00
	17	24,77	27,59	33,41
	18	25,99	28,87	34,81
	19	27,20	30,14	36,19
	20	28,41	31,41	37,57
	21	29,62	32,67	38,93
	22	30,81	33,92	40,29
	23	32,01	35,17	41,64
	24	33,20	36,42	42,98
	25	34,38	37,65	44,31
	26	35,56	38,89	45,64
	27	36,74	40,11	46,96
	28	37,92	41,34	48,28
	29	39,09	42,56	49,59
	30	40,26	43,77	50,89

Exemplo: O valor crítico de 5% com $gl = 8$ é 15,51

Fonte: Esta tabela foi gerada usando a função `invchi2tail` do Stata®.