de Y = preço de venda da casa (em dólares) sobre X = tamanho da casa (em pés quadrados). A equação de previsão obtida foi $\hat{y} = -50{,}926 + 126{,}6x$. Agora, consideramos o tamanho da casa como X_1 e também consideramos X_2 = se a casa é nova (sim ou não). A equação de previsão relacionando $\hat{y}$ a x_1 tem uma inclinação de 161 para casas novas e de 109 para casas mais antigas. Isto fornece evidência:

(a) de uma interação de X_1 e X_2 nos seus efeitos em Y.
(b) de uma associação espúria entre preço de venda e tamanho.
(c) de um relacionamento encadeado, pelo qual "novas" afeta "tamanho" que afeta o "preço de venda".
(d) de que o tamanho da casa não tem um efeito causal no preço.

10.45 Considere o relacionamento entre Y = preferência partidária (Democrata, Republicano) e X_1 = raça (negra, branca) e X_2 = gênero. Existe uma associação entre Y e X_1 e X_2, com a preferência pelos Democratas ser mais provável para os negros do que para os brancos e para as mulheres do que para os homens.

(a) X_1 e X_2 são provavelmente causas independentes de Y.
(b) A associação entre Y e X_1 é provavelmente espúria quando X_2 é controlado.
(c) Visto que ambas as variáveis afetam Y, provavelmente existe interação.
(d) As variáveis provavelmente satisfazem o relacionamento encadeado.
(e) A raça é provavelmente a variável supressora.
(f) Nenhuma das respostas acima.

NOTAS

[1] EAKER, E. D. et al. *American Journal of Epidemiology*, v. 135, p. 835-64, 1992.
[2] GUMP, B. B., MATTHEWS, K. A. *Psychosomatic Medicine*, v. 62, p. 608-12, 2000.
[3] MÜLLER, A. *British Medical Journal*, v. 324, p. 23-5, 2002.
[4] PRUS, S. G. *Canadian Journal* [illegible], v. 23, suplemento, p. S145-53, 2004.
[5] RADELET, M. *American Sociological Review*, v. 46, p. 918-27, 1981.
[6] WAINER H., BROWN, L. *American Statistician*, v. 58, p.119, 2004.
[7] KORAN et al. *American Journal of Psychiatry*, v. 163, p. 1806, 2006.
[8] RADELET, M. L., PIERCE, G. L. *Florida Law Review*, v. 43, 1991.

11 REGRESSÃO MÚLTIPLA E CORRELAÇÃO

O Capítulo 9 introduziu a modelagem por regressão do relacionamento entre duas variáveis quantitativas. Relacionamentos multivariados requerem modelos mais complexos contendo muitas variáveis explicativas. Algumas delas podem ser previsoras de interesse teórico e algumas podem ser variáveis controle.

Para prever y = GPA na universidade, é sensato usar vários previsores no mesmo modelo. As possibilidades incluem x_1 = GPA do ensino médio, x_2 = escore do exame de admissão de matemática da faculdade, x_3 = escore do exame de admissão em língua da faculdade e x_4 = avaliação do orientador educacional do ensino médio. Este capítulo apresenta modelos para o relacionamento entre uma variável resposta y e um grupo de variáveis explicativas.

Um modelo multivariado fornece previsões melhores de y do que um modelo com uma única variável explicativa. Tal modelo pode analisar, também, os relacionamentos entre variáveis enquanto controla outras variáveis. Isto é importante porque o Capítulo 10 mostrou que, após controlar uma variável, uma associação pode parecer bem diferente do que quando a variável é ignorada. Portanto, este modelo fornece informação não disponível com modelos simples que analisam somente duas variáveis de uma vez.

As Seções 11.1 e 11.2 estendem o modelo de regressão a um **modelo de regressão múltipla** que pode ter várias variáveis explicativas. A Seção 11.3 define a correlação e as medidas r ao quadrado que descrevem a associação entre y e um conjunto de variáveis explicativas. A Seção 11.4 apresenta procedimentos de inferência para a regressão múltipla. A Seção 11.5 mostra como permitir a *interação estatística* no modelo. As duas seções finais introduzem medidas que resumem a associação entre a variável resposta e uma variável explicativa enquanto controla outras variáveis.

11.1 O MODELO DE REGRESSÃO MÚLTIPLA

O Capítulo 9 modelou o relacionamento entre a variável explicativa x e a média da variável resposta y pela equação (linear) da linha reta $E(y) = \alpha + \beta x$. Referimo-nos a este modelo contendo um *único* previsor como um **modelo bivariado** porque somente contém duas variáveis.

A função de regressão múltipla

Suponha que existam duas variáveis explicativas, representadas por x_1 e x_2. Como nos capítulos anteriores, usamos a letra minúscula para representar observações ou valores particulares das variáveis. A função de regressão bivariada é generalizada para a **função de regressão múltipla**.

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$$

Nesta equação, α, β_1 e β_2 são parâmetros discutidos abaixo. Para valores em particular de x_1 e x_2, a equação especifica a média populacional de y para todos os sujeitos com esses valores de x_1 e x_2. Quando existem variáveis explicativas adicionais, cada uma tem um termo βx, por exemplo, $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ com quatro previsores.

A função de regressão múltipla é mais difícil de ser exibida graficamente do que a função de regressão bivariada. Com duas variáveis explicativas, os eixos x_1 e x_2 são perpendiculares, mas estão em um plano horizontal e o eixo y é vertical e perpendicular aos eixos x_1 e x_2. A equação $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$ representa um plano atravessando um espaço tridimensional, como o representado na Figura 11.1.

A interpretação mais simples trata todas exceto uma variável explicativa como variáveis controle e as fixa em níveis particulares. Isso deixa uma equação relacionando a média de y com a variável explicativa restante.

EXEMPLO 11.1 Altos níveis educacionais causam altas taxas de crimes?

O Exercício 9.39 na página 331 contém dados recentes de muitas variáveis para os 67 condados do estado da Flórida. Para cada condado, considere y = taxa de crimes (número anual de crimes por 1000 habitantes), x_1 = educação (percentual de adultos residentes que têm pelo menos o ensino médio completo) e x_2 = urbanização (percentual que vive em áreas urbanas).

O relacionamento bivariado entre taxa de crimes e educação é aproximado por $E(y) = -51{,}3 + 1{,}5x_1$. Surpreendentemente, a associação é moderadamente *positiva*, a correlação sendo $r = 0{,}47$. À medida que o percentual de residentes do condado que têm pelo menos o ensino médio completo aumenta, também aumenta a taxa de crimes.

Um olhar mais atento aos dados revela associações positivas fortes entre a taxa de crimes e urbanização ($r = 0{,}68$) e entre educação e urbanização ($r = 0{,}79$). Isso sugere que a associação entre taxa de crimes e educação pode ser espúria.

Talvez a urbanização seja um fator causal comum. Veja a Figura 11.2. À medida que a urbanização aumenta, tanto a taxa de crimes quanto a taxa de educação aumentam, resultando em uma correlação positiva entre crimes e educação.

A relação entre a taxa de crimes e os dois previsores considerados juntos é aproximada pela função de regressão múltipla:

$$E(y) = 58{,}9 - 0{,}6x_1 + 0{,}7x_2.$$

Por exemplo, a taxa de crimes esperada para um condado com níveis médios de educação ($\bar{x}_1 = 70$) e urbanização ($\bar{x}_2 = 50$) é $E(y) = 58{,}9 - 0{,}6(70) + 0{,}7(50) = 52$ crimes anuais por 1000 habitantes.

Vamos estudar o efeito de x_1 controlado por x_2. Primeiro, ajustamos x_2 no seu nível médio de 50. Então, o relacionamento entre taxa de crime e educação é:

$$E(y) = 58{,}9 - 0{,}6x_1 + 0{,}7(50) = 58{,}9 - 0{,}6x_1 + 35{,}0 = 93{,}9 - 0{,}6x_1.$$

A Figura 11.3 representa essa equação. Controlando x_2, fixando-o em 50, o relacionamento entre taxa de crimes e educação é negativo em vez de positivo. A inclinação diminuiu e mudou de sinal de 1,5 no relacionamento bivariado para $-0{,}6$. Neste nível fixo de urbanização, existe um relacionamento negativo entre educação e taxa de crimes. Usamos o termo equação de regressão *parcial* para distinguir a equação $E(y) = 93{,}9 - 0{,}6x_1$ da equação $E(y) = 51{,}3 + 1{,}5x_1$ para o relacionamento *bivariado* entre y e x_1. A equação de regressão *parcial* se refere a uma *parte* das observações potenciais, neste caso os condados tendo $x_2 = 50$.

A seguir, fixamos x_2 em um nível diferente, digamos $x_2 = 40$ em vez de 50. Então, podemos verificar que $E(y) = 86{,}9 - 0{,}6x_1$. Portanto, diminuindo x_2 em 10

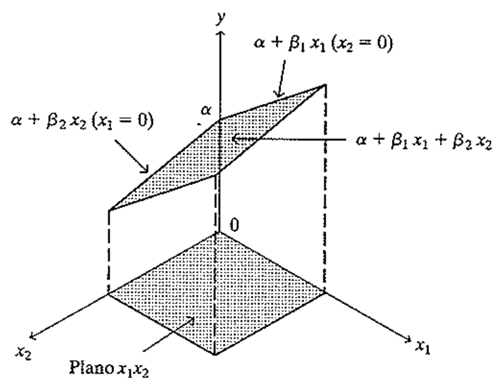


Figura 11.1 Representação gráfica de uma função de regressão múltipla com duas variáveis explicativas.

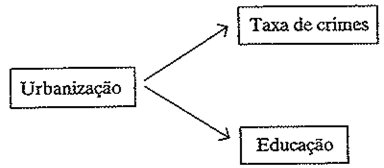


Figura 11.2 Relação positiva entre taxa de crimes e educação pode ser espúria, explicada pelos efeitos da urbanização em cada uma das variáveis.

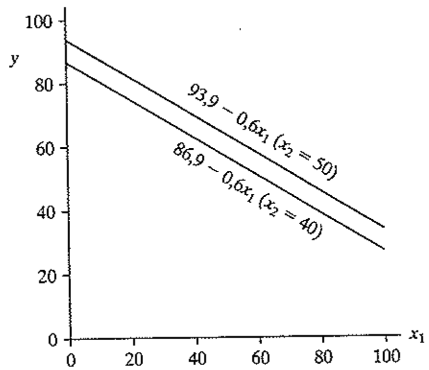


Figura 11.3 Relacionamentos parciais entre $E(y)$ e x_1 para a equação de regressão múltipla $E(y) = 58{,}9 - 0{,}6x_1 + 0{,}7x_2$. Estas equações de regressão parcial fixam x_2 igual a 50 ou 40.

unidades movemos a linha parcial relacionando y a x_1 para baixo por $10\beta_2 = 7{,}0$ unidades (veja Figura 11.3). A inclinação de $-0{,}6$ para o relacionamento parcial permanece a mesma, assim a linha é paralela à original. Fixar x_2 em outros valores gera um conjunto de linhas paralelas, cada uma tendo uma inclinação $\beta_1 = -0{,}6$.

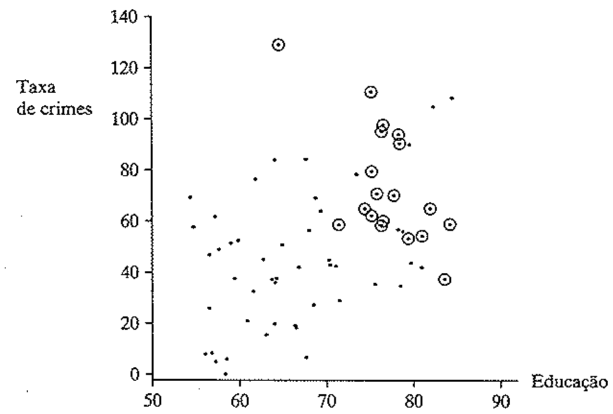
Da mesma forma, fixar x_1 em outros valores gera um conjunto de linhas paralelas, cada uma tendo uma inclinação de 0,7 relacionando a média y a x_2. Em outras palavras, controlando a educação, a inclinação do relacionamento parcial entre taxa de crimes e urbanização é $\beta_2 = 0{,}7$.

Em resumo, a educação tem um efeito positivo geral na taxa de crimes, mas ela tem um efeito negativo quando a urbanização é controlada. A associação parcial tem a direção oposta da associação bivariada. Este é o chamado de **paradoxo de Simpson.** A Figura 11.4 ilustra como isso acontece. Ela mostra o diagrama de dispersão relacionando a taxa de crimes à educação, exibindo uma associação global positiva entre essas variáveis. O diagrama circula os 19 condados que apresentam níveis de urbanização maiores. O subconjunto de pontos, para os quais a urbanização é aproximadamente constante, tem uma tendência negativa entre taxa de crimes e educação. A associação positiva alta entre educação e urbanização é refletida pelo fato de que a maioria das observações destacadas em que a urbanização é maior também tem valores maiores em educação. ■

Interpretação dos coeficientes de regressão

Vimos que, para um valor fixo de x_2, a equação $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$ fica simplificada a uma equação linear em x_1 com inclinação β_1. A inclinação é a mesma para cada valor fixo de x_2. Quando fixamos o valor de x_2, estamos mantendo-o constante: estamos *controlando* x_2. Essa é a diferença básica entre a interpretação das inclinações na regressão múltipla e na regressão bivariada:

- Na *regressão múltipla*, uma inclinação descreve o efeito de uma variável explicativa quando são *controlados* os efeitos das outras variáveis explicativas no modelo.
- A *regressão bivariada* tem somente uma única variável explicativa. Assim, uma inclinação na regressão bivariada descreve o efeito daquela variável enquanto *ignora* todas as outras variáveis explicativas possíveis.



☑ **Figura 11.4** Diagrama de dispersão relacionando taxa de crimes e educação. Os pontos circulados são os condados com maior urbanização. Uma linha de regressão ajustando os pontos circulados tem uma inclinação negativa, embora a linha de regressão que passa por todos os pontos tenha uma inclinação positiva (paradoxo de Simpson).

O parâmetro β_1 mensura o *efeito parcial* de x_1 em y, isto é, o efeito de um aumento de uma unidade em x_1, mantendo x_2 constante. O efeito parcial de x_2 em y, mantendo x_1 constante, tem uma inclinação β_2. Da mesma forma, para o modelo de regressão múltipla com *vários* previsores, o coeficiente beta de um previsor descreve a mudança na média de y para um aumento de uma unidade naquele previsor, controlando as outras variáveis no modelo. O parâmetro α representa a média de y quando cada variável explicativa é igual a 0.

Os parâmetros $\beta_1, \beta_2, \ldots$ são chamados de **coeficientes da regressão parcial**. O adjetivo *parcial* distingue estes parâmetros do coeficiente da regressão β no modelo *bivariado* $E(y) = \alpha + \beta x$, que *ignora* em vez de *controlar* os efeitos de outras variáveis explicativas.

Esse modelo de regressão múltipla presume que a inclinação do relacionamento parcial entre y e cada previsor é idêntica para *todas* as combinações de valores das outras variáveis explicativas. Isso significa que o modelo é apropriado quando *não existe interação estatística*, no sentido da Seção 10.3 (página 344). Se a inclinação parcial verdadeira entre y e x_1 é muito diferente em $x_2 = 50$ do que em $x_2 = 40$, por exemplo, precisamos de um modelo mais complexo. A Seção 11.5 mostrará este modelo.

Uma inclinação parcial em um modelo de regressão múltipla geralmente difere da inclinação do modelo bivariado para aquele previsor, mas não necessariamente. Com dois previsores, as inclinações parciais e as bivariadas são iguais se a correlação entre x_1 e x_2 é igual a 0. Quando x_1 e x_2 são causas independentes de y, o efeito de x_1 em y não muda quando controlamos x_2.

Equação de previsão e resíduos

Similar à equação de regressão múltipla, o *software* encontra uma equação de previsão estimando os parâmetros do modelo a partir de dados amostrais. Por clareza de notação, até agora usamos apenas dois previsores. Em geral, k representa o número de previsores do modelo.

> ☑ **Notação para a equação de previsão**
>
> A equação de previsão que estima a de regressão múltipla $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ é representada por $\hat{y} = a + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$.

Para a regressão múltipla, é quase imperativo usar um *software* para encontrar a equação de previsão. As fórmulas de cálculo são complexas e não são apresentadas no livro.

Conseguimos o valor previsto de y para um sujeito substituindo os valores-x para aquele sujeito na equação de previsão. Como o modelo bivariado, o modelo da regressão múltipla tem **resíduos** que mensuram os erros de previsão. Para um sujeito com uma resposta prevista $\hat{y}$ e resposta observada y, o resíduo é $y - \hat{y}$. A próxima seção mostra um exemplo.

A **soma dos quadrados do erro** (SQE)

$$SQE = \sum (y - \hat{y})^2$$

resume a proximidade do ajuste da equação de previsão aos dados da resposta. A maioria dos *softwares* chama o SQE de **soma dos quadrados dos resíduos**. A fórmula para a SQE é a mesma apresentada no Capítulo 9. A única diferença é que o valor previsto $\hat{y}$ resulta do uso de *diversas* variáveis explicativas em vez de apenas um único previsor.

As estimativas do parâmetro na equação de previsão satisfazem o critério dos

mínimos quadrados: a equação de previsão tem o menor valor da SQE de todas as equações possíveis da forma $\hat{y} = a + b_1x_1 + ... + b_kx_k$.

11.2 EXEMPLO COM UMA SAÍDA COMPUTACIONAL DA REGRESSÃO MÚLTIPLA

Ilustramos os métodos deste capítulo com os dados do exemplo seguinte:

EXEMPLO 11.2 Regressão múltipla para um estudo de saúde mental

Um estudo no condado de Alachua, Flórida, investigou o relacionamento entre certos índices de saúde mental e diversas variáveis explicativas. O interesse principal estava focado em um índice de distúrbio mental que incorporasse várias dimensões de sintomas psiquiátricos, incluindo aspectos de ansiedade e depressão. Esta medida, que é a variável resposta y, variava de 17 a 41 na amostra. Escores mais altos indicam maior distúrbio mental.

As duas variáveis explicativas usadas aqui x_1 = escore dos eventos vividos e x_2 = *status* socioeconômico (SES). Os escores dos eventos vividos é uma medida composta do número e severidade dos principais eventos vividos que o sujeito experienciou nos últimos três anos. Estes eventos variam de transtornos pessoais graves, como uma morte na família, sentença de prisão, ou um caso extraconjugal, para eventos menos graves, como conseguir um novo emprego, o nascimento de um filho, mudar-se dentro da mesma cidade ou ter um filho casando. Esta medida[1] variava de 3 a 97 na amostra. Um escore alto representa um número maior e/ou uma maior gravidade nos eventos vividos. O escore do SES é um índice composto baseado na ocupação, renda e educação. Mensurado em uma escala padrão que variava de 0 a 100. Quanto maior o escore, mais alto o *status*.

A Tabela 11.1 mostra os dados das três variáveis para uma amostra aleatória de 40 adultos do condado. (Estes dados são baseados em um levantamento maior de dados. O autor agradece ao Dr. Charles Holzer pela permissão do uso do estudo como base para este exemplo.) A Tabela 11.2 resume as médias amostrais e os desvios padrão das três variáveis. ■

Matriz de diagramas de dispersão para relacionamentos bivariados

Diagramas dos dados fornecem uma verificação informal para se verificar se os relacionamentos são lineares. A maioria dos *softwares* pode construir diagramas de dispersão em um único diagrama para cada par de variáveis. A Figura 11.5 mostra os diagramas para as variáveis da Tabela 11.1. Este tipo de diagrama é chamado de **matriz de diagramas de dispersão**. Como uma matriz de correlação, ele mostra cada par de variáveis duas vezes. Em um diagrama, uma variável está no eixo y e uma está no eixo x. O distúrbio mental (a variável resposta) está no eixo y para os diagramas da primeira linha da Figura 11.5, assim estes são os diagramas de interesse. Os diagramas não mostram evidência de não linearidade e modelos com efeitos lineares parecem apropriados. Os diagramas sugerem que os eventos vividos tinham um efeito positivo moderado e o SES tinha um efeito negativo moderado no distúrbio mental.

Diagramas parciais para relacionamentos parciais

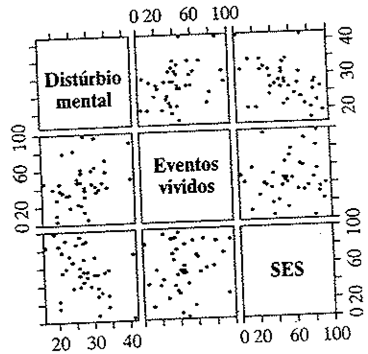
O modelo de regressão múltipla declara que cada previsor tem um efeito linear com uma inclinação comum, quando controlados os demais previsores. Para verificar isso, poderíamos usar um *software* para fazer um diagrama de y *versus* cada previsor, para subconjuntos de pontos que são aproximadamente constantes nos outros previsores. Com uma única variável controle poderíamos, por exemplo, dividir as observações entre quatro grupos usando os quartis como limites e, então, ou construir quatro diagramas de dispersão separados ou marcar as observações em um único diagrama de dispersão de acordo com o seu grupo. Com várias variáveis controle, entretanto, manter todas aproximadamente

☑ **Tabela 11.1** Escores em y = distúrbio mental, x_1 = eventos vividos e x_2 = *status* socioeconômico

y	x_1	x_2	y	x_1	x_2	y	x_1	x_2
17	46	84	26	50	40	30	44	53
19	39	97	26	48	52	31	35	38
20	27	24	26	45	61	31	95	29
20	3	85	27	21	45	31	63	53
20	10	15	27	55	88	31	42	7
21	44	55	27	45	56	32	38	32
21	37	78	27	60	70	33	45	55
22	35	91	28	97	89	34	70	58
22	78	60	28	37	50	34	57	16
23	32	74	28	30	90	34	40	29
24	33	67	28	13	56	41	49	3
24	18	39	28	40	56	41	89	75
25	81	87	29	5	40			
26	22	95	30	59	72			

☑ **Tabela 11.2** Médias estimadas e desvios padrão de distúrbio mental, eventos vividos e *status* socioeconômico (SES)

Variável	Média	Desvio padrão
Distúrbio mental	27,30	5,46
Eventos vividos	44,42	22,62
SES	56,60	25,28



☑ **Figura 11.5** Uma matriz de diagramas de dispersão para pares de variáveis da Tabela 11.1.

constantes pode reduzir a amostra a poucas observações. Uma única figura mais informativa é fornecida pelo **diagrama de regressão parcial**, que exibe o relacionamento entre a variável resposta e a variável explicativa após remover os efeitos dos outros previsores do modelo de regressão múltipla. O modelo faz isto traçando um gráfico dos resíduos dos modelos usando essas duas variáveis como respostas e as outras variáveis explicativas como previsoras.

Por exemplo, aqui está como encontrar o digrama de regressão parcial para o efeito de x_1 quando o modelo de regressão múltipla também tem as variáveis explicativas x_2 e x_3. Encontre os resíduos do modelo usando x_2 e x_3 para prever y. Também encontre os resíduos do modelo usando x_2 e x_3 para prever x_1. Faça, então, um diagrama dos resíduos da primeira análise (no eixo y) *versus* os resíduos da segunda análise. Para estes resíduos, os efeitos de x_2 e x_3 são removidos. A inclinação obtida pelos mínimos quadrados para os pontos nesse diagrama é necessariamente a mesma da inclinação parcial estimada b_1 para o modelo de regressão múltipla.

A Figura 11.6 mostra um diagrama de regressão parcial (do SPSS) para y = distúrbio mental e x_1 = eventos vividos, controlados por x_2 = SES. Ele faz uma representação gráfica dos resíduos no eixo y do modelo $\hat{y} = 32{,}2 - 0{,}086x_2$ usando x_2 para prever y *versus* os resíduos no eixo x do modelo $\hat{x}_1 = 38{,}2 + 0{,}110x_2$ usando x_2 para prever x_1. Os dois eixos têm valores negativos e positivos porque eles se referem aos resíduos. Lembre que os resíduos (erros de previsão) podem ser positivos ou negativos e têm média igual a 0. A Figura 11.6 sugere que o efeito parcial dos eventos vividos é aproximadamente linear e positivo.

A Figura 11.7 mostra a regressão parcial para o SES. Ela mostra que seu efeito parcial é também aproximadamente linear, mas negativo. É simples obter diagramas de regressão parcial com *softwares* como o SPSS. (Veja o Apêndice A.)

Saídas computacionais de resultados amostrais

As Tabelas 11.3 e 11.4 são saídas do SPSS dos coeficientes para os relacionamentos bivariados entre distúrbio mental e as variáveis explicativas separadas. Os coeficientes de regressão estimados estão na coluna rotulada de "B". As equações de previsão são:

$\hat{y} = 23{,}31 + 0{,}090x_1$ e $\hat{y} = 32{,}17 - 0{,}086x_2$.

Na amostra, o distúrbio mental está positivamente relacionado aos eventos vividos, desde que os coeficientes de x_1 (0,090) sejam positivos. Quanto maior o número e a gravidade dos eventos vividos nos últimos três anos, maior a tendência de distúrbio mental (isto é, mais insatisfatória é a saúde mental). O distúrbio mental está relacionado negativamente ao *status* socioeconômico. Quanto maior o nível de SES, menor a tendência de distúrbio mental. As correlações entre o distúrbio mental e as variáveis explicativas são modestas, 0,372 para os eventos vividos e −0,399 para o SES (listado no SPSS como *Standardized coefficients* (coeficientes padronizados); o rótulo "beta" é equivocado e se refere ao termo alternativo *pesos beta* para os coeficientes de regressão padronizados).

A Tabela 11.5 mostra parte da saída do SPSS para o modelo de regressão múl-

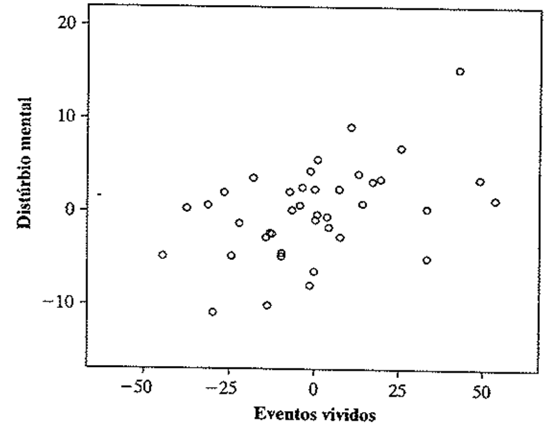


Figura 11.6 Um diagrama de dispersão parcial para distúrbio mental e eventos vividos, controlados pelo SES. Nele estão representados graficamente os resíduos da regressão do distúrbio mental explicados pelo SES *versus* os resíduos da regressão de eventos vividos explicados pelo SES.

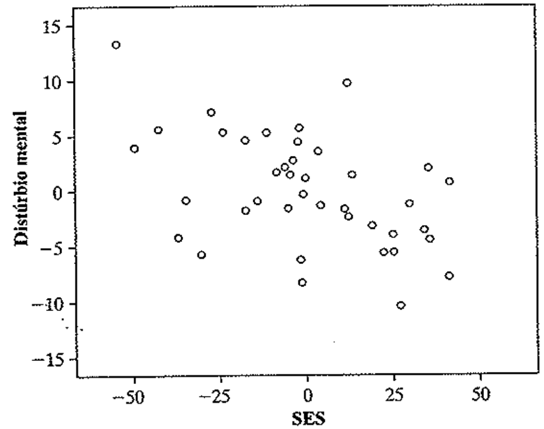


Figura 11.7 Um diagrama de dispersão parcial para distúrbio mental e SES, controlado pelos eventos vividos. O diagrama faz uma representação gráfica dos resíduos do distúrbio mental explicado pelos eventos vividos *versus* os resíduos da regressão do SES explicados pelos eventos vividos.

Tabela 11.3 Análise de regressão bivariada para y = distúrbio mental (IMPAIR) e x_1 = eventos vividos (LIFE)

Coeficientes (a)					
Modelo	Coeficientes não padronizados		Coeficientes padronizados		
	B	Erro Padão	Beta	t	Sig
1 (Constante)	23,309	1,807	12,901	0,0C0	
LIFE	0,090	0,036	0,372	2,472	0,018
a	Variável dependente: IMPAIR				

☑ **Tabela 11.4** Análise de regressão bivariada para y = distúrbio mental e x_2 = *status* socioeconômico (SES)

Coeficientes (a)					
Modelo	Coeficientes não padronizados		Coeficientes padronizados		
	B	Erro padrão	Beta	t	Sig
1 (Constante)	32,172	1,988		16,186	0,000
SES	−0,086	0,032	−0,399	−2,679	0,011
a	Variável Dependente: IMPAIR				

tipla $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$. A equação de previsão é

$$\hat{y} = a + b_1x_1 + b_2x_2 = 28{,}230 + 0{,}103x_1 - 0{,}097x_2.$$

Controlando o SES, o relacionamento amostral entre o distúrbio mental e eventos vividos é positivo, desde que o coeficiente dos eventos vividos ($b_1 = 0{,}103$) seja positivo. A média estimada de distúrbio mental aumenta aproximadamente 0,1 para cada unidade de aumento no escore dos eventos vividos, controlados pelo SES. Visto que $b_2 = -0{,}097$, existe uma associação negativa entre distúrbio mental e SES, controlados pelos eventos vividos. Por exemplo, sobre um intervalo de tamanho 100 de valores potenciais do SES (de um mínimo de 0 a um máximo de 100), a média estimada de distúrbio mental muda por $100(-0{,}097) = -9{,}7$. Visto que o distúrbio mental tem um intervalo de somente 17 a 41 com um desvio padrão de 5,5, uma diminuição de 9,7 pontos é digno de nota.

Da Tabela 11.1, o primeiro sujeito da amostra tinha $y = 17$, $x_1 = 46$ e $x_2 = 84$. O distúrbio mental previsto desse sujeito é:

$$\hat{y} = 28{,}230 + 0{,}103(46) - 0{,}097(84) = 24{,}8.$$

O erro de previsão (resíduo) é $y - \hat{y} = 17 - 24{,}8 = -7{,}8$.

A Tabela 11.6 resume alguns resultados da análise de regressão. Ela mostra os erros padrão em parênteses abaixo das estimativas do parâmetro. As inclinações parciais para o modelo de regressão múltipla são similares às inclinações para os modelos bivariados. Em cada caso, a introdução de um segundo previsor altera muito pouco o efeito de outro. Isto sugere que estes previsores podem ter aproximadamente efeitos amostrais independentes em y. Na verdade, a correlação amostral entre x_1 e x_2 é somente de 0,123. A próxima seção mostra como mensurar a associação conjunta de variáveis explicativas com a variável resposta e mostra como interpretar o valor R^2 listado para o modelo de regressão múltipla.

☑ **Tabela 11.5** Ajuste do modelo de regressão múltipla para y = distúrbio mental e x_1 = eventos vividos (LIFE) e x_2 = *status* socioeconômico (SES)

	Coeficientes não padronizados		Coeficientes padronizados		
	B	Erro padrão	Beta	t	Sig
(Constante)	28,230	2,174		12,984	0,000
LIFE	0,103	0,032	0,428	3,177	0,003
SES	−0,097	0,029	−0,451	−3,351	0,002
Variável dependente: IMPAIR					

11.3 CORRELAÇÃO MÚLTIPLA E R^2

A correlação r e seu quadrado descrevem a força da associação linear para relacionamentos bivariados. Esta seção apresenta medidas análogas para o modelo de regressão múltipla, que descrevem a força da associação entre y e o conjunto de variáveis explicativas agindo juntas como previsores no modelo.

A correlação múltipla

As variáveis explicativas, de maneira coletiva, estão fortemente associadas com y se os valores observados de y estão altamente correlacionados com os valores previstos $\hat{y}$ da equação de previsão. A correlação entre os valores observados e os previstos resume essa correlação.

> ☑ **Correlação múltipla**
>
> A **correlação múltipla** para um modelo de regressão é a correlação entre os valores observados de y e os valores previstos $\hat{y}$.

Para cada sujeito, a equação de previsão fornece um valor previsto $\hat{y}$. Assim, cada sujeito tem um valor-y e um valor-$\hat{y}$. Por exemplo, acima dissemos que o primeiro sujeito na amostra tinha $y = 17$ e $\hat{y} = 24{,}8$. Para os três primeiros sujeitos na Tabela 11.1, os valores observados e previstos de y são:

y	$\hat{y}$
17	24,8
19	22,8
20	28,7

A correlação amostral calculada entre os valores de y e $\hat{y}$ é a correlação múltipla, que é representada por R.

Os valores previstos não podem estar correlacionados negativamente com os valores observados. As previsões devem ser pelo menos tão boas quanto a média amostral $\bar{y}$, que é a previsão de quando todas as inclinações parciais são iguais a 0 e $\bar{y}$ tem correlação zero com y. Assim, R sempre está entre 0 e 1. Neste sentido, a correlação entre y e $\hat{y}$ difere da correlação entre y e um previsor x, que está entre −1 e +1. Quanto maior a correlação múltipla R, melhor será a previsão de y pelo conjunto de variáveis explicativas.

R^2: o coeficiente de determinação múltipla

Outra medida usa o conceito da *redução proporcional no erro*, generalizando r^2 para modelos bivariados. Esta medida resume a melhoria relativa nas previsões

☑ **Tabela 11.6** Resumo dos modelos de regressão para o distúrbio mental

	Previsores no modelo de regressão		
Efeito	**Múltiplo**	**Eventos vividos**	**SES**
Intercepto	28,230	23,309	32,172
Eventos vividos	0,103 (0,032)	0,090 (0,036)	–
SES	−0,097 (0,029)	–	−0,086 (0,032)
R^2	0,339	0,138	0,159
(n)	(40)	(40)	(40)

usando a equação de previsão em vez de $\bar{y}$ e tem os seguintes elementos:

Regra 1 (Prever y sem usar $x_1, ..., x_k$): O melhor previsor é, então, a média amostral, $\bar{y}$.

Regra 2 (Prever y usando $x_1, ..., x_k$): O melhor previsor é a equação de previsão $\hat{y} = a + b_1x_1 + b_2x_2 + ... + b_kx_k$.

Erros de previsão: o erro de previsão para um sujeito é a diferença entre os valores observados e previstos de y. Com a regra 1, o erro é $y - \bar{y}$. Com a regra 2, é o resíduo $y - \hat{y}$. Em qualquer um dos casos, resumimos o erro pela soma dos quadrados dos erros previstos. Para a regra 1, é SQT $= \sum(y - \bar{y})^2$, chamado de *soma dos quadrados total*. Para a regra 2, é SQE $= \sum(y - \hat{y})^2$, a soma dos quadrados dos erros usando a equação de previsão, chamada de *soma dos quadrados dos resíduos*.

Definição de medida: a redução proporcional no erro com o uso da equação de previsão $\hat{y} = a + b_1x_1 + b_2x_2, ..., + b_kx_k$ em vez de $\bar{y}$ para prever y é chamada de **coeficiente de determinação múltipla** ou, para simplificar, **R quadrado**.

☑ *R* quadrado: o coeficiente de determinação múltipla

$$R^2 = \frac{SQT - SQE}{SQT} = \frac{\Sigma(y - \bar{y})^2 - \Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}$$

R^2 mensura a proporção da variação total em y que é explicada pelo poder de previsão de todas as variáveis explicativas, por meio do modelo de regressão múltipla. O símbolo reflete que ele é o quadrado da correlação múltipla. A representação com a letra maiúscula de R^2 distingue esta medida do r^2 para o modelo bivariado. Suas fórmulas são idênticas e r^2 é o caso especial de R^2 aplicado a um modelo de regressão com uma variável explicativa. Para o modelo de regressão múltipla ser útil para a previsão, ele deve fornecer previsões aprimoradas relativas não somente a $\bar{y}$, mas também aos modelos bivariados separados para y e cada variável explicativa.

EXEMPLO 11.3 Correlação múltipla e R^2 para o distúrbio mental

Para os dados em y = distúrbio mental, x_1 = eventos vividos e x_2 = *status* socioeconômico, introduzidos no Exemplo 11.2, a equação de previsão é $\hat{y} = 28{,}23 + 0{,}103x_1 - 0{,}097x_2$. A Tabela 11.5 mostrou a saída para este modelo. O *software* também informa as tabelas da ANOVA (análise de variância) com as tabelas das somas dos quadrados e os valores de R e R^2. A Tabela 11.7 mostra a saída do SPSS.

Da coluna da "Soma dos Quadrados", a soma dos quadrados total é SQT $= \Sigma(y - \bar{y})^2 = 1162{,}4$ e a soma dos quadrados do resíduo do uso da equação de previsão para prever y é SQE $= \sum(y - \hat{y})^2 = 768{,}2$. Portanto,

$$R^2 = \frac{SQT - SQE}{SQT} = \frac{1162{,}4 - 768{,}2}{1162{,}4} = 0{,}339.$$

Usar os eventos vividos e SES juntos para prever o distúrbio mental fornece uma redução de 33,9% no erro de previsão comparado ao uso de apenas $\bar{y}$. O modelo de regressão múltipla fornece uma redução substancialmente grande no erro comparado ao modelo bivariado. (A Tabela 11.6 apresentou os valores r^2 de 0,138 e 0,159 para eles.) Neste caso, ele é mais útil do que esses modelos para fins de previsão.

A correlação múltipla entre o distúrbio mental e as duas variáveis explicativas é $R = +\sqrt{0{,}339} = 0{,}582$. Isso é igual à correlação entre os valores observados de y e os valores previstos de $\hat{y}$ para o modelo.

☑ **Tabela 11.7** Tabela da ANOVA e resumo do modelo para a regressão do distúrbio mental (IMPAIR) sobre os eventos vividos (LIFE) e o *status* socioeconômico (SES)

ANOVA

	Soma dos quadrados	gl	Média dos quadrados	F	Sig.
Regressão	394,238	2	197,119	9,495	0,000
Resíduo	768,162	37	20,761		
Total	1162,400	39			

Resumo do modelo

R	R ao quadrado	R ao quadrado Ajustado	Erro padrão da estimativa
0,582	0,339	0,303	4,556

Previsores: (Constante), SES, LIFE

Variável dependente: IMPAIR

O SPSS informa o R e o R^2 em uma tabela separada denominada *Model Summary* (Resumo do Modelo), como a Tabela 11.7 mostra. A maioria dos *softwares* também informa uma versão ajustada do R^2 que é uma estimativa menos tendenciosa do valor da população. O Exercício 11.61 define esta medida e a Tabela 11.7 informa o seu valor de 0,303. ■

Propriedades do R e R^2

As propriedades do R^2 são similares àquelas do r^2 para os modelos bivariados.

- O R^2 está entre 0 e 1.
- Quanto maior o valor do R^2, melhor o conjunto das variáveis explicativas $(x_1, ..., x_k)$ coletivamente prevê y.
- $R^2 = 1$ somente quando todos os resíduos são 0, isto é, quando todos $y = \hat{y}$, nesse caso o SQE = 0. Nesta situação a equação de previsão passa por todos os pontos dos dados.
- $R^2 = 0$ quando as previsões não variam tanto quanto os valores-x. Neste caso, $b_1 = b_2 = ... = b_k = 0$ e $\hat{y}$ é idêntico a $\bar{y}$, uma vez que as variáveis explicativas não adicionam poder de previsão. Quando isso acontece, a correlação entre y e cada variável explicativa é igual a 0.
- O R^2 não pode diminuir quando adicionamos uma variável explicativa ao modelo. É impossível explicar *menos* variação em y adicionando variáveis explicativas ao modelo de regressão.
- O R^2 para o modelo da regressão múltipla é pelo menos tão grande quanto os valores do r^2 para os modelos bivariados separados. Isto é, o R^2 para o modelo de regressão múltipla é pelo menos tão grande quanto $r^2_{yx_1}$ para y como uma função linear de x_1, $r^2_{yx_2}$ para y como uma função linear de x_2 e assim por diante.

As propriedades da correlação múltipla R seguem diretamente daquelas para R^2, desde que R seja a raiz quadrada positiva de R^2. Por exemplo, a correlação múltipla para o modelo $E(y) = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3$ é pelo menos tão grande quanto a correlação múltipla para o modelo $E(y) = \alpha + \beta_1x_1 + \beta_2x_2$.

O numerador de R^2, SQT − SQE, resume a variação em y explicada pelo modelo de regressão múltipla. Essa diferença, que é igual a $\Sigma(\hat{y} - \bar{y})^2$, é chamada de **soma dos quadrados da regressão**. Os valores da ANOVA na Tabela 11.7 apresentam a soma dos quadrados da regressão como 394,2. (Alguns *softwares*, como o SAS, rotulam esse valor como a soma dos quadrados do "modelo".) Apresenta ainda a soma dos quadrados total (SQT) dos valores de y sobre a partição de $\bar{y}$ nas variações explicadas pelo modelo de regressão (soma dos quadrados da regressão) mais a variação não explicada pelo modelo (a soma dos quadrados dos resíduos, SQE).

Multicolinearidade com muitas variáveis explicativas

Quando existem muitas variáveis explicativas com correlações entre elas fortes, uma vez que algumas forem incluídas no modelo, o R^2 geralmente não aumenta muito mais quando são colocadas as demais. Por exemplo, para o conjunto de dados *House selling price* (preço de venda das casas) no *site* do livro (introduzido no Exemplo 9.10 na página 310), o r^2 é 0,71 com o valor dos impostos sendo um previsor do preço de venda. Nesse caso, o R^2 aumenta para 0,77 quando acrescentamos o tamanho da casa como um segundo previsor. Mas ele aumenta somente para 0,79 quando acrescentamos o número de banheiros, o número de quartos e se a casa é nova ou não como previsores adicionais.

Quando o R^2 aumenta pouco, não quer dizer que as variáveis adicionais não estão correlacionadas com y. Simplesmente significa que elas não acrescentam muito poder adicional para prever y, considerando os previsores que já estão no modelo. Essas variáveis adicionais podem ter associações pequenas com y, dadas as variáveis que já estão no modelo. Isso geralmente acontece na pesquisa em Ciências Sociais quando as variáveis explicativas estão altamente correlacionadas e, assim, nenhuma tendo muito poder explicativo por si só. A Seção 14.3 discute essa condição, que é denominada de **multicolinearidade**.

A Figura 11.8, que exibe a porção da variabilidade total de y que é explicada por cada um dos três previsores, mostra uma ocorrência comum. A área do conjunto representando um previsor, nesta figura, representa o tamanho do seu valor de r^2 na previsão de y. O quanto a área de um previsor se sobrepõe à área de outro previsor representa a sua associação entre eles. A área de um conjunto representando um previsor que não se sobrepõe a outros conjuntos representa a parte da variabilidade de y explicada unicamente por esse previsor. Na Figura 11.8, todos os três previsores têm associações moderadas com y e, juntos, eles explicam uma parte considerável da variação de y. Contudo, se x_1 e x_2 estiverem no modelo, x_3 explicará pouca variação adicional de y em virtude de sua forte correlação com x_1 e x_2. Por causa dessa sobreposição, R^2 aumenta muito pouco quando x_3 é adicionado a um modelo que já contém x_1 e x_2.

Para propósitos de previsão, ganhamos pouco adicionando variáveis explicativas a um modelo, estando elas fortemente correlacionadas com outras que já estão no modelo, visto que R^2 irá aumentar pouco. Em condições ideais, devemos usar as variáveis explicativas que tenham correlações fracas entre si, mas fortes com y. Na prática, isso nem sempre é possível, especialmente quando queremos incluir certas variáveis no modelo por razões teóricas.

Na realidade, o tamanho da amostra que é necessário para realizar uma regressão múltipla satisfatória fica maior quando queremos usar muitas variáveis explicativas. As dificuldades técnicas causadas pela multicolinearidade serão menos severas para amostras grandes. Em condições ideais, o tamanho da amostra deve ser pelo menos de aproximadamente 10 vezes maior que o número de variáveis explicativas utilizadas (por exemplo, aproximadamente 40 com 4 variáveis explicativas).

11.4 INFERÊNCIA PARA OS COEFICIENTES DA REGRESSÃO MÚLTIPLA

A função da regressão múltipla

$$E(y) = \alpha + \beta_1 x_1 + ... + \beta_k x_k$$

descreve o relacionamento entre as variáveis explicativas e a média da variável resposta. Para valores particulares das variáveis explicativas, $\alpha + \beta_1 x_1 + ... + \beta_k x_k$ representa a média de y para a população que tem estes valores.

Para fazer inferências sobre os parâmetros, formulamos todo o *modelo de regressão múltipla*. Isto consiste nessa equação mais o seguinte conjunto de suposições:

- A distribuição da população y é normal para cada combinação de valores de $x_1, ..., x_k$.
- O desvio padrão, σ, da distribuição condicional das respostas em y é o mesmo para cada combinação de valores de $x_1, ..., x_k$.
- A amostra é selecionada aleatoriamente.

Sob estas suposições, a verdadeira distribuição amostral é exatamente igual a estas citadas nesta seção. Na prática, as suposições nunca são perfeitamente satisfeitas. As inferências bilaterais são robustas para a normalidade e a suposição de um mesmo σ. Mais importantes são as suposições de aleatorização e que a função de regressão descreva bem como a média de y depende das variáveis explicativas. Veremos formas de verificar a última suposição na Seção 14.2.

Dois tipos de testes de significância são usados na regressão múltipla. O primeiro é um teste global de independência. Ele verifica se *qualquer uma* das variáveis explicativas está estatisticamente relacionada a y. A segunda estuda os coeficientes da regressão parcial individualmente para avaliar quais variáveis têm efeitos parciais significativos em y.

Testando a influência coletiva das variáveis explicativas

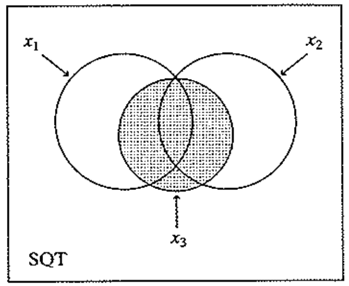
As variáveis explicativas coletivamente têm um efeito estatisticamente significativo na variável resposta? Verificamos isso testando:

$$H_0: \beta_1 = \beta_2 = ... = \beta_k = 0,$$

que afirma que a média de y não depende dos valores de $x_1, ..., x_k$. Sob as suposições da inferência, isso afirma que y é estatisticamente independente de todas as k variáveis explicativas.

A hipótese alternativa é:

$$H_a: \text{pelo menos um } \beta_i \neq 0.$$



☑ **Figura 11.8** O R^2 não aumenta muito quando x_3 é adicionado ao modelo que já contém x_1 e x_2.

Ela afirma que *pelo menos uma* variável explicativa está relacionada a y, controladas as demais. O teste julga se usar todos os $x_1, \ldots, x_k$ para prever y, com a equação de previsão $\hat{y} = a + b_1x_1 + \ldots + b_kx_k$, é melhor do que usar $\bar{y}$.

Essas hipóteses sobre os $\{\beta_i\}$ são equivalentes a:

H_0: A correlação múltipla na população = 0
H_a: A correlação múltipla na população > 0.

A equivalência ocorre porque a correlação múltipla é igual a 0 somente naquelas situações nas quais todos os coeficientes de regressão parciais são iguais a 0. Também, H_0 é equivalente a H_0: R quadrado na população = 0.

Para essas hipóteses sobre os k previsores, a estatística-teste é igual a:

$$F = \frac{R^2/k}{(1 - R^2)/[n - (k + 1)]}.$$

A distribuição amostral dessa estatística é denominada **distribuição *F***. A seguir, estudaremos essa distribuição e suas propriedades.

A distribuição *F*

O símbolo da estatística-teste F e de sua distribuição é uma homenagem ao estatístico mais eminente da história, Ronald A. Fisher, que determinou a distribuição F em 1922. Como a distribuição do qui-quadrado, a distribuição F pode assumir somente valores não negativos e é assimétrica à direita. A Figura 11.9 ilustra isso.

A forma da distribuição F é determinada por dois parâmetros ou graus de liberdade, representados por gl_1 e gl_2:

$gl_1 = k$, o número de variáveis explicativas do modelo.
$gl_2 = n - (k + 1) = n -$ número de parâmetros na equação de regressão.

O primeiro deles, $gl_1 = k$, é o divisor do termo do numerador (R^2) na estatística-teste F. O segundo, $gl_2 = n - (k + 1)$, é o divisor do termo do numerador $(1 - R^2)$. O número de parâmetros no modelo de regressão múltipla é $k + 1$, representando os k termos beta e o termo alfa.

A média da distribuição F é aproximadamente[2] igual a 1. Quanto maior o valor do R^2, maior a razão $R^2/(1 - R^2)$ e maior se torna a estatística-teste F. Portanto, valores maiores da estatística-teste F fornecem evidências mais fortes contra H_0. Sob a suposição de que H_0 é verdadeira, o valor-p é a probabilidade de que a estatística-teste F seja maior do que o valor observado de F. Isto é a probabilidade da cauda direita sob a distribuição F à direita do valor observado de F, como mostra a Figura 11.9.

A Tabela D (página 653) no final do livro lista os escores-F tendo valores-p de 0,05; 0,01 e 0,001, para várias combinações de gl_1 e gl_2. Esta tabela nos permite determinar se $p > 0{,}05$; $0{,}01 < p < 0{,}05$; $0{,}001 < p < 0{,}01$ ou $p < 0{,}001$. Um *software* para a regressão informa o valor-p real.

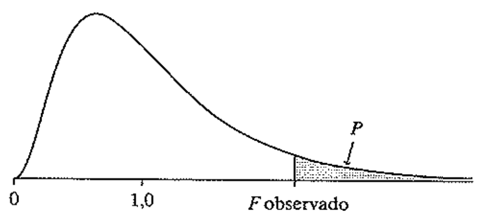


Figura 11.9 A distribuição F e o valor-p para o teste F. Valores maiores de F fornecem evidências mais fortes contra H_0.

EXEMPLO 11.4 O teste *F* para os dados do distúrbio mental

No Exemplo 11.2 (página 366), usamos a regressão múltipla para $n = 40$ observações em y = distúrbio mental, com $k = 2$ variáveis explicativas, os eventos vividos e o SES. A hipótese nula de que o distúrbio mental é estatisticamente dependente dos eventos vividos e do SES é H_0: $\beta_1 = \beta_2 = 0$.

No Exemplo 11.3 (página 372), encontramos que esse modelo tem $R^2 = 0{,}339$. O valor da estatística-teste F é:

$$F = \frac{R^2/k}{(1 - R^2)/[n - (k + 1)]} = \frac{0{,}339/2}{0{,}661/[40 - (2 + 1)]} = 9{,}5.$$

Os dois graus de liberdade da distribuição F são $gl_1 = k = 2$ e $gl_2 = n - (k + 1) = 40 - 3 = 37$, os dois divisores nesta estatística.

Da Tabela D, quando $gl_1 = 2$ e $gl_2 = 37$, o valor-F com probabilidade unicauda à direita de 0,001 está entre 8,25 e 8,77. Visto que a estatística-teste F observada de 9,5 está acima desses dois valores, ela está mais acima na cauda da distribuição e tem uma probabilidade unicaudal menor do que 0,001. Portanto, o valor-p é < 0,001. Parte da saída do SPSS na Tabela 11.7 mostra a tabela da ANOVA na qual vemos a estatística F. O valor-p, que foi arredondado para três casas decimais é igual a 0,000, e aparece sob o título "Sig." na tabela da ANOVA.

Este valor-p extremamente pequeno fornece forte evidência contra H_0. Ele sugere que, pelo menos, uma das variáveis explicativas está relacionada ao distúrbio mental. De forma equivalente, podemos concluir que a correlação múltipla na população e o R quadrado são positivos. Assim, obtemos significativamente melhores previsões de y usando a equação de regressão múltipla do que usando $\bar{y}$. ■

Normalmente, a não ser que o tamanho da amostra seja pequeno e as associações fracas, este teste F tem um valor-p pequeno. Se escolhermos com cuidado as variáveis, pelo menos uma delas deve ter *algum* poder explicativo.

Inferências para os coeficientes de regressão individual

Suponha que o valor-p é pequeno para o teste F em que todos os coeficientes de regressão são iguais a 0. Isso não implica que *cada* variável explicativa tenha um efeito em y (controlada pelas outras variáveis explicativas no modelo), mas meramente que *pelo menos uma* delas tem um efeito. As análises focadas de modo mais restrito julgam *quais* efeitos parciais são diferentes de zero e estimam os tamanhos desses efeitos. Essas inferências fazem as mesmas suposições do que o teste F, as mais importantes sendo a aleatorização e que a função da regressão descreve bem como a média de y depende das variáveis explicativas.

Considere uma variável explicativa arbitrária x_i, com coeficiente β_i no modelo de regressão múltipla. O teste para o seu efeito parcial em y é H_0: $\beta_i = 0$. Se $\beta_i = 0$, a média de y é idêntica para todos os valores de x_i, controlando as outras variáveis explicativas no modelo. A alternativa pode ser bilateral, H_a: $\beta_i \neq 0$, ou unilateral, H_a: $\beta_i > 0$ ou H_a: $\beta_i < 0$, para prever a direção do efeito parcial.

A estatística-teste para H_0: $\beta_i = 0$, usando a estimativa amostral b_i de β_i é:

$$t = \frac{b_i}{ep},$$

onde ep é o erro padrão de b_i. Como de costume, a estatística-teste t é a melhor estimativa (b_i) do parâmetro (β_i), subtrai o valor de H_0 do parâmetro (0) e divide pelo erro padrão. A fórmula para o *ep* é complexa, mas o *software* fornece seu valor. Se H_0 é verdadeira e as suposições do modelo se mantêm, a estatística *t* tem uma distribuição *t* com $gl = n - (k + 1)$. O valor *gl* é o mesmo do gl_2 no teste *F*.

É mais informativo estimar o tamanho do efeito parcial do que testar se ele é zero. Lembre que β_i representa a mudança na média de *y* para um aumento de uma unidade em x_i, controladas as demais variáveis. Um intervalo de confiança para β_i é:

$$b_i \pm t(ep).$$

O valor-*t* é obtido da tabela *t*, com $gl = n - (k + 1)$. Por exemplo, um intervalo de 95% de confiança para o efeito parcial de x_1 é $b_i \pm t_{0,025}(ep)$.

EXEMPLO 11.5 Inferências para previsores parciais do distúrbio mental

Para o modelo de regressão múltipla de y = distúrbio mental em x_1 = eventos vividos e x_2 = SES,

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2,$$

vamos considerar o efeito dos eventos vividos. A hipótese de que o distúrbio mental é estatisticamente independente dos eventos vividos, controlado o SES, é H_0: $\beta_1 = 0$. Se H_0 é verdadeira, a equação de regressão múltipla se reduz a $E(y) = \alpha + \beta_2 x_2$. Se H_0 é falsa, então $\beta_1 \neq 0$ e o modelo completo fornece um ajuste melhor do que o modelo bivariado.

Os resultados da Tabela 11.5 (página 370) nos dizem que a estimativa por ponto de β_1 é $b_1 = 0{,}103$ e o erro padrão $ep = 0{,}032$. A estatística-teste é igual a:

$$t = \frac{b_1}{ep} = \frac{0{,}103}{0{,}032} = 3{,}18.$$

Isto aparece na tabela sob o título "*t*" e na linha da variável "LIFE". A estatística tem um $gl = n - (k + 1) = 40 - 3 = 37$. O valor-*p* aparece na coluna "Sig" e na linha "LIFE". Seu valor é 0,003 e é a probabilidade de que a estatística *t* exceda 3,18 em valor absoluto. Existe uma forte evidência de que o distúrbio mental está relacionado aos eventos vividos, controlado o SES.

Um intervalo de 95% de confiança para β_1 usa $t_{0,025} = 2{,}026$, o valor-*t* com $gl = 37$ que tem uma probabilidade de $0{,}05/2 = 0{,}025$ em cada cauda. Este intervalo é igual a:

$$b_1 \pm t_{0,025}(ep) \text{ ou } 0{,}103 \pm 2{,}026(0{,}032) \text{ ou ainda } (0{,}04; 0{,}17).$$

Controlado o SES, estamos 95% confiantes de que, para uma mudança média de uma unidade no distúrbio mental, o aumento nos eventos vividos está entre 0,04 e 0,17. O intervalo não contém 0. Isto está de acordo com a rejeição de H_0: $\beta_1 = 0$ em favor de H_a: $\beta_1 \neq 0$ ao nível $\alpha = 0{,}05$.

Visto que este intervalo contém somente números positivos, o relacionamento entre o distúrbio mental e os eventos vividos é positivo, controlado o SES. Pode ser mais simples interpretar o intervalo (0,04; 0,17) observando que um aumento de 100 unidades nos eventos vividos corresponde a algo entre 100(0,04) = 4 a 100(0,17) = 17 unidades de aumento no distúrbio mental. O intervalo é relativamente grande em virtude do pequeno tamanho amostral. ■

Quão diferente é o teste *t* para um coeficiente de regressão parcial do teste *t* de H_0: $\beta = 0$ para o modelo bivariado, $E(y) = \alpha + \beta x$, estudado na Seção 9.5 (página 308)? Aquele teste *t* avaliava se *y* e *x* estão associados, *ignorando* as outras variáveis, porque ele se aplica ao modelo bivariado. Por outro lado, o teste recém apresentado avalia se as variáveis estão associadas, *controlando* as outras variáveis.

Uma nota de advertência: suponha que exista multicolinearidade, isto é, muita sobreposição entre as variáveis explicativas no sentido de que qualquer uma é bem prevista pelas outras. Então, possivelmente nenhum dos efeitos parciais individuais tenha um valor-*p* pequeno, mesmo quando o R^2 é grande e uma estatística *F* grande ocorre no teste geral para os βs. Qualquer variável, em particular, pode explicar por si só pouco da variação em *y*, embora juntas elas expliquem muita variação.

Variabilidade e média dos quadrados na Tabela da ANOVA*

A precisão das estimativas dos mínimos quadrados está relacionada ao tamanho do desvio padrão condicional σ que mensura a variabilidade de *y* para valores fixos dos previsores. Quanto menor a variabilidade dos valores de *y* em torno da equação de regressão, menor o erro padrão. A estimativa de σ é:

$$s = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - (k + 1)}} = \sqrt{\frac{SQE}{gl}}.$$

O valor dos graus de liberdade é também *gl* para inferências *t* para os coeficientes de regressão e é gl_2 para o teste *F* sobre o efeito coletivo dos previsores. (Quando um modelo tem somente $k = 1$ previsor, o *gl* é $n - 2$, o termo na fórmula do *s* da Seção 9.3, na página 297.)

Parte da saída do SPSS da Tabela 11.7 (página 373) mostrou a tabela da ANOVA contendo a soma dos quadrados do modelo de regressão múltipla com os dados do distúrbio mental. Vemos que SQE = 768,2. Visto que $n = 40$ para $k = 2$ previsores, temos $gl = n - (k + 1) = 40 - 3 = 37$ e

$$s = \sqrt{\frac{SQE}{gl}} = \sqrt{\frac{768{,}2}{37}} = \sqrt{20{,}76} = 4{,}56.$$

Se as distribuições condicionais têm distribuição aproximadamente normal, então quase todos os escores de distúrbio mental estão a aproximadamente 14 unidades (3 desvios padrão) da média especificada pela função da regressão.

O SPSS informa o desvio padrão condicional sob o título *Std. Error of the Estimate* (Erro Padrão da Estimativa) na tabela *Model Summary* (Resumo do Modelo), que apresenta, também, os valores de *R* e R^2 (veja Tabela 11.7). Esse é um nome mal escolhido pelo SPSS porque *s* se refere ao **erro quadrático médio**, geralmente abreviado por EQM. O *software* apresenta o seu valor na tabela da ANOVA na coluna denominada *Mean Square* (Média dos Quadrados) e na linha rotulada de *Residual* (ou *Error* em alguns *softwares*). Por exemplo, EQM = 20,76 na tabela. Alguns *softwares* (como o SAS) rotulam melhor a estimativa do desvio padrão condicional *s* como *Root MSE*, porque ele é a raiz quadrada do erro quadrático médio.

A estatística *F* é a razão entre as médias dos quadrados*

Uma fórmula alternativa da estatística-teste *F* para verificar H_0: $\beta_1 = ... = \beta_k = 0$ usa as médias dos quadrados da tabela da ANOVA. Especificamente:

$$F = \frac{\text{Média dos quadrados da regressão}}{\text{Média dos quadrados dos resíduos}} = \frac{197{,}1}{20{,}8} = 9{,}5.$$

Esse resultado é o mesmo obtido pela fórmula da estatística-teste *F* com base no R^2.

A média dos quadrados da regressão é igual à soma dos quadrados da regressão dividida pelos seus graus de liberdade. O *gl* é igual a *k*, o número de variáveis explicativas no modelo, que é gl_1 para o teste *F*. Na saída do computador mostrada, a média dos quadrados da regressão é igual a:

$$\frac{\text{SQ da Regressão}}{gl_1} = \frac{394{,}2}{2} = 197{,}1.$$

O relacionamento entre F e a estatística t*

Vimos que a distribuição F é usada para testar se todos os coeficientes de regressão parciais são iguais a 0. Alguns *softwares* de regressão também listam a estatística-teste F em vez das estatística-teste t para os testes sobre os coeficientes de regressão individuais. As duas estatísticas estão relacionadas e têm os mesmos valores-p. O quadrado da estatística t para verificar se um coeficiente de regressão parcial é igual a 0 tem uma distribuição F com $gl_1 = 1$ e $gl_2 = n - (k + 1)$.

Para ilustrar, no Exemplo 11.5 (página 378), para testar H_0: $\beta_1 = 0$ contra H_a: $\beta_1 \neq 0$, a estatística-teste era $t = 3{,}18$ com $gl = 37$. De forma alternativa, poderíamos usar $F = t^2 = 3{,}18^2 = 10{,}1$, que tem uma distribuição F com $gl_1 = 1$ e $gl_2 = 37$. O valor-p para esse valor F é 0,003, que é o mesmo que a Tabela 11.5 fornece para o teste bilateral t.

Em geral, se uma estatística tem a distribuição t com d graus de liberdade, então o quadrado dessa estatística tem a distribuição F com $gl_1 = 1$ e $gl_2 = d$. Uma desvantagem da abordagem F é que ela carece de informação sobre a direção da associação. Ela não pode ser usada para testar hipóteses alternativas unilaterais.

11.5 INTERAÇÃO ENTRE PREVISORES E SEUS EFEITOS

A equação de regressão múltipla

$$E(y) = \alpha + \beta_1x_1 + \beta_2x_2 + ... + \beta_kx_k$$

assume que o relacionamento parcial entre y e cada x_i é linear e que a inclinação β_i desse relacionamento é idêntico para todos os valores das demais variáveis explicativas. Isso implica um paralelismo de linhas relacionando as duas variáveis, para os vários valores das outras variáveis, como a Figura 11.3 ilustrou.

Este modelo é, algumas vezes, muito simples para ser adequado. Geralmente, existe uma **interação** com o relacionamento entre duas variáveis mudando de acordo com os valores de uma terceira variável. A Seção 10.3 (página 344) introduziu esse conceito.

> ☑ **Interação**
>
> Para variáveis quantitativas, existe uma interação entre duas variáveis explicativas nos seus efeitos em y quando o efeito de uma variável muda à medida que o nível de outra variável, também, muda.

Por exemplo, suponha que o relacionamento entre x_1 e a média de y é $E(y) = 2 + 5x_1$ quando $x_2 = 0$, é $E(y) = 4 + 15x_1$ quando $x_2 = 50$ e é $E(y) = 6 + 25x_1$ quando $x_2 = 100$. A inclinação para o efeito parcial de x_1 muda acentuadamente à medida que o valor para x_2 muda. Existe, então, uma interação entre x_1 e x_2 nos seus efeitos em y.

Termos dos produtos cruzados

Uma abordagem comum para permitir uma interação introduz os **termos do produto cruzado** de variáveis explicativas no modelo de regressão múltipla. Com duas variáveis explicativas, o modelo é:

$$E(y) = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2.$$

Este é um caso especial do modelo de regressão múltipla com três variáveis explicativas, no qual x_3 é uma variável artificial criada como o produto cruzado $x_3 = x_1x_2$ das duas variáveis explicativas principais.

Vamos ver por que este modelo permite a interação. Considere como y está relacionado a x_1, controlado x_2. Escrevemos a equação em termos de x_1 como:

$$E(y) = (\alpha + \beta_2x_2) + (\beta_1 + \beta_3x_2)x_1$$
$$= \alpha' + \beta'x_1,$$

onde:

$$\alpha' = \alpha + \beta_2x_2 \text{ e } \beta' = \beta_1 + \beta_3x_2.$$

Assim, para x_2 fixo, a média de y muda linearmente como uma função de x_1. A inclinação do relacionamento é $\beta' = (\beta_1 + \beta_3x_2)$. Este valor depende de x_2. À medida que x_2 muda, a inclinação para o efeito de x_1, também, muda. Em resumo, a média de y é uma função linear de x_1, mas a inclinação da linha depende do valor assumido por x_2.

Observe que, agora, podemos interpretar β_1 como o efeito de x_1 somente quando $x_2 = 0$. A menos que $x_2 = 0$ seja um valor particular de interesse para x_2, ele não é, nesse caso, útil para construir intervalos de confiança ou executar testes de significância sobre β_2 nesse modelo.

De forma similar, a média de y é uma função linear de x_2, mas a inclinação varia de acordo com o valor de x_1. O coeficiente β_2 de x_2 se refere ao efeito de x_2 somente quando $x_1 = 0$.

EXEMPLO 11.6 Modelo de interação para o exemplo do distúrbio mental

Para o conjunto de dados em y = distúrbio mental, x_1 = eventos vividos e x_2 = SES, criamos uma terceira variável explicativa x_3 que representa o produto cruzado de x_1 e x_2 para os 40 indivíduos. Para o primeiro sujeito, por exemplo, $x_1 = 46$, $x_2 = 84$, assim $x_3 = 46(84) = 3864$. O *software* torna fácil criar esta variável sem que você mesmo tenha que fazer os cálculos. A Tabela 11.8 mostra parte da saída para o modelo com interação. A equação de previsão é:

$$\hat{y} = 26{,}0 + 0{,}156x_1 - 0{,}060x_2 - 0{,}00087x_1x_2.$$

A Figura 11.10 exibe o relacionamento entre o distúrbio mental previsto e os eventos vividos para os alguns valores distintos do SES. Para um escore do SES de $x_2 = 0$, o relacionamento entre $\hat{y}$ e x_1 é:

$$\hat{y} = 26{,}0 + 0{,}156x_1 - 0{,}060(0) - 0{,}00087x_1(0)$$
$$= 26{,}0 + 0{,}156x_1.$$

Quando $x_2 = 50$, a equação de previsão é:

$$\hat{y} = 26{,}0 + 0{,}156x_1 - 0{,}060(50) - 0{,}00087(50)x_1$$
$$= 23{,}0 + 0{,}113x_1.$$

Quando $x_2 = 100$, a equação de previsão é:

$$y = 20{,}0 + 0{,}069x_1.$$

Quanto maior for o valor do SES menor será a inclinação entre o distúrbio mental previsto e os eventos vividos e, assim, mais fraco será o efeito dessa variável. Isto sugere que sujeitos que possuem maiores recursos, na forma de um SES mais alto, resistem melhor ao estresse mental de potenciais eventos traumáticos vividos. ■

Testando um termo interação

Para duas variáveis explicativas, o modelo que permite interação é:

$$E(y) = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2.$$

O modelo mais simples não assumindo interação é o caso especial de $\beta_3 = 0$. A hipótese de não existência de interação é H_0: $\beta_3 = 0$. Como sempre, a estatística-teste t divide a estimativa do parâmetro (β_3) pelo seu erro padrão.

Da Tabela 11.8, $t = -0{,}00087/0{,}0013 = -0{,}67$. O valor-$p$ para H_a: $\beta_3 \neq 0$ é 0,51. Existe pouca evidência de interação. A variação da inclinação do relacionamento entre o distúrbio mental e os eventos vividos para vários níveis do SES pode ser devido à variabilidade amostral. O

Tabela 11.8 Modelo com interação para y = distúrbio mental, x_1 = eventos vividos e x_2 = SES

	Soma dos quadrados	gl	Média dos quadrados	F	Sig.
Regressão	403,631	3	134,544	6,383	0,0014
Resíduo	758,769	36	21,077		
Total	1162,400	39			

R	R quadrado
0,589	0,347

	B	Erro padrão	t	Sig
(Constante)	26,036649	3,948826	6,594	0,0001
LIFE	0,155865	0,085338	1,826	0,0761
SES	−0,060493	0,062675	−0,965	0,3409
LIFE*SES	−0,000866	0,001297	−0,668	0,5087

tamanho da amostra aqui é pequeno, entretanto, e isto torna difícil estimar precisamente os efeitos. Estudos baseados em grandes amostras (por exemplo, Holzer, 1977) mostraram que realmente existe interação, do tipo visto nesse exemplo, para essas variáveis.

Na Tabela 11.8, nem o teste de H_0: $\beta_1 = 0$ ou H_0: $\beta_2 = 0$ tem valores-*p* pequenos. Mas os testes em H_0: $\beta_1 = 0$ e H_0: $\beta_2 = 0$ são altamente significativos para o modelo de "nenhuma interação" $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$; da Tabela 11.5, os valores-*p* são 0,003 e 0,002. Esta perda de significância ocorre porque $x_3 = x_1 x_2$ está fortemente correlacionado com x_1 e x_2, com $r_{x_1x_3} = 0{,}779$ e $r_{x_2x_3} = 0{,}646$. Essas correlações substanciais não são surpresa, visto que $x_3 = x_1x_2$ é determinada completamente por x_1 e x_2.

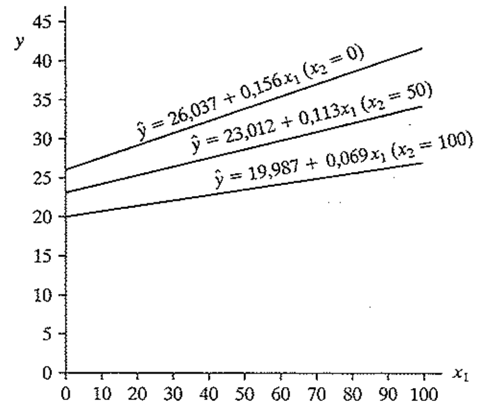


Figura 11.10 Representação da interação entre x_1 e x_2 nos seus efeitos em *y*.

Visto que existe uma sobreposição considerável na variação em *y* que é explicada por x_1 e x_1x_2 e, também, por x_2 e x_1x_2, a variabilidade *parcial* explicada por cada uma é relativamente pequena. Por exemplo, muito do poder previsto contido em x_1 está também contido em x_2 e x_1x_2. A *única* contribuição de x_1 (ou x_2) ao modelo é relativamente pequena e não significativa, quando x_2 (ou x_1) e x_1x_2 estão no modelo.

Quando a evidência de interação é fraca, como é o caso aqui, com o valor-*p* de 0,51, é melhor abandonar o termo de interação do modelo antes de testar a hipótese sobre os efeitos parciais como H_0: $\beta_1 = 0$ e H_0: $\beta_2 = 0$. Por outro lado, *se a evidência de interação é forte, não faz mais sentido testar essas hipóteses*. Se existe interação, então o efeito de cada variável existe e difere de acordo com o nível da outra variável.

Centrando as variáveis explicativas*

Para os dados da saúde mental, vimos que x_1 e x_2 são altamente significativos no modelo quando eles são os únicos previsores (veja a Tabela 11.5), mas perdem significância quando o termo de interação é adicionado, embora a interação não seja significativa (veja a Tabela 11.8). Vimos, também, que os coeficientes de x_1 e x_2 em um modelo de interação não são geralmente significativos porque eles se referem ao efeito de um previsor somente quando o outro previsor é igual a 0.

Existe uma forma alternativa de parametrizar a interação no modelo para que ele dê estimativas e significância para o efeito de x_1 e x_2 similar àquelas para o modelo sem interação. O método envolve *centrar* os escores de cada variável explicativa em torno de 0, subtraindo a média. Considere $x_1^C = x_1 - \mu_{x_1}$ e $x_2^C = x_2 - \mu_{x_2}$, assim cada nova variável explicativa terá média igual a 0. Então, expressamos o modelo com interação como:

$$E(y) = \alpha + \beta_1 x_1^C + \beta_2 x_2^C + \beta_3 x_1^C x_2^C$$
$$= \alpha + \beta_1(x_1 - \mu_{x_1}) + \beta_2(x_2 - \mu_{x_2}) + \beta_3(x_1 - \mu_{x_1})(x_2 - \mu_{x_2}).$$

Agora, β_1 se refere ao efeito de x_1 na média de x_2, e β_2 se refere ao efeito de x_2 na média de x_1. Suas estimativas são geralmente similares aos efeitos estimados para o modelo sem interação.

Quando executamos novamente o modelo de interação para os dados da saúde mental após centrar os previsores em torno de suas médias amostrais, isto é, com

LIFE_CEN = LIFE − 44,425 e SES_CEN = SES − 56,60,

obtemos a seguinte saída computacional

	B	Erro padrão	t	Sig
(Constante)	27,359555	0,731366	37,409	0,0001
LIFE_CEN	0,106850	0,033185	3,220	0,0027
SES_CEN	−0,098965	0,029390	−3,367	0,0018
LIFE_CEN* SES_CEN	−0,000866	0,001297	−0,668	0,5087

A estimativa para o termo de interação é a mesma para o modelo com previsores não centralizados. Embora, agora, as estimativas (e os erros padrão) somente para os efeitos de x_1 e x_2 sejam similares aos valores para o modelo sem interação. Isso acontece porque o coeficiente para uma variável representa seu efeito na média da outra variável, que é tipicamente similar ao efeito para o modelo sem interação. Da mesma forma, a significância estatística tanto de x_1 quanto de x_2 são similares ao modelo sem interação.

Centrar as variáveis previsoras, antes de usá-las, em um modelo que permite interação traz dois benefícios. Primeiro, as estimativas dos efeitos de x_1 e x_2 são mais significativas, tendo efeitos nos valores médios em vez de no valor em 0. Segundo, as estimativas e seus erros padrão são similares ao modelo sem interação. O termo do produto cruzado com variáveis centralizadas não se sobrepõe aos outros termos como acontece no modelo com as variáveis não centradas.

Generalizações e limitações*

Quando o número de variáveis explicativas é superior a duas, um modelo com interação tem produtos cruzados para cada par de variáveis. Por exemplo, com três variáveis explicativas, um modelo de interação é:

$$E(y) = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_1x_2 + \beta_5x_1x_3 + \beta_6x_2x_3.$$

Este é um caso especial de regressão múltipla com seis variáveis explicativas, identificando $x_4 = x_1x_2$, $x_5 = x_1x_3$ e $x_6 = x_2x_3$. Testes de significância podem julgar quais termos, se existe algum, dos produtos cruzados são necessários no modelo.

Quando existe uma interação e o modelo contém produtos de termos cruzados, é mais difícil resumir os relacionamentos de forma simples. Uma abordagem é traçar um conjunto de linhas como aquela da Figura 11.10 para descrever graficamente como o relacionamento entre duas variáveis muda de acordo com os valores das demais variáveis. Outra possibilidade é dividir os dados em grupos de acordo com o valor de uma variável controle (por exemplo, alto, médio e baixo em x_2) e relatar a inclinação entre y e x_1 dentro de cada subconjunto como forma de descrever a interação.

As interações no modelo acima são chamados de **termos de segunda ordem**, para distingui-los dos termos de interação de ordem mais alta com produtos de mais de duas variáveis de uma vez. Tais termos são ocasionalmente usados em modelos mais complexos, mas não são considerados nesse capítulo.

11.6 COMPARANDO MODELOS DE REGRESSÃO

Quando o número de variáveis explicativas aumenta, o modelo de regressão múltipla se torna mais difícil de interpretar e algumas variáveis podem tornar-se redundantes. Isto é especialmente verdade quando algumas variáveis explicativas são produtos cruzados de outras para permitir interações. Nem todos os previsores podem ser necessários no modelo. A seguir apresentaremos um teste para verificar se um modelo se ajusta significativamente melhor do que um modelo mais simples contendo somente os previsores.

Modelos completos e reduzidos

O modelo com todos os previsores é denominado de **modelo completo**. O que contém somente alguns desses previsores é chamado de **modelo reduzido**. O modelo reduzido é dito estar *aninhado* dentro do modelo completo, o que significa que ele é um caso especial dele.

Os modelos completo e reduzido são idênticos se os coeficientes da regressão parcial para as variáveis extras do modelo completo são todas iguais a 0. Neste caso, nenhum dos previsores extras aumenta a variabilidade explicada em y, na população de interesse. Testar se o modelo completo é idêntico ao modelo reduzido é equivalente a testar se os parâmetros extras do modelo completo são iguais a 0. A hipótese alternativa é que pelo menos um destes parâmetros extras não é 0, nesse caso, o modelo completo é melhor do que o modelo reduzido.

Por exemplo, um modelo completo com três variáveis explicativas e todos os termos de interação de segunda ordem é:

$$E(y) = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_1x_2 + \beta_5x_1x_3 + \beta_6x_2x_3.$$

O modelo reduzido, sem os termos de interação, é:

$$E(y) = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3.$$

O teste que compara se o modelo completo é igual ao reduzido tem H_0: $\beta_4 = \beta_5 = \beta_6 = 0$.

Comparando modelos pela comparação do SQE ou os valores de R^2

A estatística-teste para comparar dois modelos de regressão compara a soma dos quadrados dos resíduos para os dois modelos. Represente a $SQE = \sum (y - \hat{y})^2$ para o modelo reduzido por SQE_r e para o completo por SQE_c. Agora, $SQE_r \geq SQE_c$ uma vez que o modelo reduzido tem menos previsores e tende a fazer previsões menos precisas. Mesmo se H_0 fosse verdadeira, não esperaríamos que as estimativas dos parâmetros extras e a diferença ($SQE_r - SQE_c$) fossem iguais a 0. Alguma redução no erro ocorre do ajuste de termos extras em virtude da variabilidade amostral.

A estatística-teste usa a redução no erro, $SQE_r - SQE_c$, que resulta da adição das variáveis extras. Uma estatística equivalente usa os valores de R^2 representados por R_c^2 se o modelo for completo e R_r^2 se for o reduzido. A estatística-teste é igual a:

$$F = \frac{(SQE_r - SQE_c)/gl_1}{SQE_c/gl_2} = \frac{(R_c^2 - R_r^2)/gl_1}{(1 - R_c^2)/gl_2}.$$

Aqui, gl_1 é o número extra de termos no modelo completo (por exemplo, 3 no exemplo acima que adiciona três termos de interação para conseguir o modelo completo) e gl_2 é o resíduo gl usual para o modelo completo, que é $gl_2 = n - (k + 1)$. Uma redução relativamente grande no erro (ou um aumento relativamente alto em R^2) gera uma estatística-teste F grande e um valor-p pequeno. Como de costume, para estatísticas F, o valor-p é a probabilidade da cauda direita.

EXEMPLO 11.7 Comparando modelos para o distúrbio mental

Para os dados de distúrbio mental, uma comparação do modelo completo

$$E(y) = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2$$

para o modelo reduzido

$$E(y) = \alpha + \beta_1x_1 + \beta_2x_2$$

analisa se a interação existe. O modelo completo tem apenas um termo adicional e a hipótese nula é H_0: $\beta_3 = 0$.

A soma dos quadrados dos erros para o modelo completo é $SQE_c = 758{,}8$ (Tabela 11.8), enquanto o reduzido é $SQE_r = 768{,}2$ (Tabela 11.7). A diferença

$$SQE_r - SQE_c = 768{,}2 - 758{,}8 = 9{,}4$$

tem $gl_1 = 1$ visto que o modelo completo tem um parâmetro a mais. Visto que o tamanho da amostra é $n = 40$, o $gl_2 = n - (k + 1) = 40 - (3 + 1) = 36$, que é o gl para a SQE na Tabela 11.8. A estatística-teste é igual a:

$$F = \frac{(SQE_r - SQE_c)/gl_1}{SQE_c/gl_2} = \frac{9{,}4/1}{758{,}8/36} = 0{,}45.$$

De forma equivalente, os valores de R^2 para os dois modelos são $R_r^2 = 0{,}339$ e $R_c^2 = 0{,}347$, assim:

$$F = \frac{\left(R_c^2 - R_r^2\right)/gl_1}{\left(1 - R_c^2\right)/gl_2} = \frac{(0{,}347 - 0{,}339)/1}{(1 - 0{,}347)/36} = 0{,}45.$$

Do *software*, o valor-p da distribuição F com $gl_1 = 1$ e $gl_2 = 36$ é 0,51. Existe pouca evidência de que o modelo completo seja melhor. A hipótese nula parece plausível, assim não é possível dizer que modelo reduzido não é adequado.

Quando H_0 contém um único parâmetro, o teste t está disponível. Na verdade, da seção anterior (e Tabela 11.8), a estatística t é igual a:

$$t = \frac{b_3}{ep} = \frac{-0{,}00087}{0{,}0013} = -0{,}67.$$

Ela também tem um valor-p de 0,51 para H_a: $\beta_3 \neq 0$. Obtemos com o teste t o mesmo resultado que foi obtido com o teste F para os modelos completo e reduzido. Na verdade, a estatística-teste F é igual ao quadrado da estatística t (veja à página 380). ■

O método t é limitado a testar um parâmetro por vez. O teste F pode testar *vários* parâmetros simultaneamente para verificar se, pelo menos, um deles não é zero, como no teste global F de H_0: $\beta_1 = ... = \beta_k = 0$ ou no teste comparando o modelo completo ao reduzido. O teste F é equivalente ao teste t somente quando H_0 contém um único parâmetro.

11.7 CORRELAÇÃO PARCIAL*

Os modelos de regressão múltipla descrevem o efeito de uma variável explicativa na variável resposta enquanto são controladas outras variáveis de interesse. Medidas relacionadas descrevem a força da associação. Por exemplo, para descrever a associação entre distúrbio mental e eventos vividos, controlados pelo SES, poderíamos perguntar: "Controlado o SES, que proporção da variação do distúrbio mental é explicada pelos eventos vividos?".

Essas medidas descrevem a associação parcial entre y e um previsor em particular, enquanto a correlação múltipla e o R^2 descrevem a associação entre y e todo o conjunto de previsores do modelo. A *correlação parcial* é baseada nas correlações ordinárias entre cada par de variáveis. Para uma única variável controle, ela é definida como segue:

> ☑ **Correlação parcial**
>
> A **correlação parcial** amostral entre y e x_1, controlada por x_2, é:
>
> $$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{\left(1 - r^2_{yx_2}\right)\left(1 - r^2_{x_1x_2}\right)}}.$$

No símbolo $r_{yx_1 \cdot x_2}$, a variável à direita do ponto representa a variável controlada. A fórmula análoga para $r_{yx_2 \cdot x_1}$ (isto é, controlando x_1) é:

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{\sqrt{\left(1 - r^2_{yx_1}\right)\left(1 - r^2_{x_1x_2}\right)}}.$$

Visto que uma variável é controlada, as correlações parciais $r_{yx_1 \cdot x_2}$ e $r_{yx_2 \cdot x_1}$ são chamadas de **correlações parciais de primeira ordem**.

EXEMPLO 11.8 Correlação parcial entre educação e taxa de crimes

O Exemplo 11.1 (página 362) discutiu um conjunto de dados para os condados da Flórida, com y = taxa de crimes, x_1 = educação e x_2 = urbanização. As correlações entre os pares de variáveis são r_{yx_1}= 0,468, r_{yx_2}= 0,678 e $r_{x_1x_2}$= 0,791. Foi surpreendente observar uma correlação positiva entre taxa de crimes e educação. Ela pode ser explicada por sua dependência conjunta da urbanização? Isto é plausível se a associação desaparece quando a urbanização é controlada.

A correlação parcial entre taxa de crimes e educação, controlada pela urbanização, é igual a:

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{(1 - r^2_{yx_2})(1 - r^2_{x_1x_2})}}$$

$$= \frac{0{,}468 - 0{,}678(0{,}791)}{\sqrt{(1 - 0{,}678^2)(1 - 0{,}791^2)}} = -0{,}152.$$

Não surpreendentemente, $r_{yx_1 \cdot x_2}$ é muito menor do que r_{yx_1}. Ela até mesmo tem uma direção diferente, ilustrando o paradoxo de Simpson. O relacionamento entre taxa de crimes e educação pode ser espúrio, refletindo sua dependência conjunta da urbanização. ■

Interpretando as correlações parciais

A correlação parcial tem propriedades similares àquelas da correlação usual entre duas variáveis, com valores variando de −1 a +1, sem unidades, onde valores absolutos maiores representam associações mais fortes. Listamos as propriedades abaixo para $r_{yx_1 \cdot x_2}$, mas propriedades análogas se aplicam a $r_{yx_2 \cdot x_1}$.

- $r_{yx_1 \cdot x_2}$ está entre −1 e +1.
- Quanto maior o valor absoluto de $r_{yx_1 \cdot x_2}$, mais forte é a associação entre y e x_1, controlado x_2.
- O valor de uma correlação parcial não depende das unidades de medidas das variáveis.
- $r_{yx_1 \cdot x_2}$ tem o mesmo sinal da inclinação parcial (b_1) para o efeito de x_1 na equação de previsão $\hat{y} = a + b_1x_1 + b_2x_2$. Isto acontece porque a mesma variável (x_2) é controlada tanto no modelo quanto na correlação.
- Sob as suposições da realização de inferência para a regressão múltipla (veja o início da Seção 11.4), $r_{yx_1 \cdot x_2}$ estima a correlação entre y e x_1 para cada valor *fixo* de x_2. Se pudéssemos controlar x_2 considerando uma subpopulação de sujeitos tendo, todos, o mesmo valor em x_2, então $r_{yx_1 \cdot x_2}$ estima a correlação entre y e x_1 para aquela subpopulação.
- A correlação amostral parcial é idêntica à correlação calculada para os pontos no *diagrama de regressão parcial* (Seção 11.2).

Interpretando as correlações parciais ao quadrado

Como o r^2 e o R^2, o quadrado de uma correlação parcial é interpretada como uma redução proporcional no erro (RPE). Por exemplo, o quadrado do valor de $r_{yx_2 \cdot x_1}$ nos diz que o $r^2_{yx_2 \cdot x_1}$ é a proporção da variação em y explicada por x_2, quando x_1 é controlado. Esta medida ao quadrado descreve o efeito de remover a porção da soma dos quadrados total (SQT) em y que é explicada por x_1 e, então, encontrar a proporção da variação inexplicada remanescente em y que é explicada por x_2.

> ☑ **Correlação parcial ao quadrado**
>
> O quadrado da correlação parcial $r_{yx_2 \cdot x_1}$ representa a proporção da variação de y que é explicada por x_2, menos aquela parte não explicada por x_1. Ele é igual a:
>
> $$r^2_{yx_2 \cdot x_1} = \frac{R^2 - r^2_{yx_1}}{1 - r^2_{yx_1}}$$
>
> $$= \frac{\text{proporção parcial explicada por } x_2}{\text{proporção não explicada por } x_1}.$$

Lembre, da Seção 9.4 (página 301), que $r^2_{yx_1}$ representa a proporção da variabilidade de y explicada por x_1. A proporção remanescente $(1 - r^2_{yx_1})$ representa a variação não explicada. Quando x_2 é adicionado ao modelo, ele é responsável por uma variação adicional. A proporção total da variação de y debitada conjuntamente a x_1 e x_2 é R^2 para o modelo com ambos, x_1 e x_2, como variáveis explicativas. Assim, $R^2 - r^2_{yx_1}$ é a proporção adicional da variabilidade de y explicada por x_2, após os efeitos de x_1 terem sido removidos ou controlados. O máximo que esta diferença poderia ser é $1 - r^2_{yx_1}$, a proporção da variação ainda a ser explicada após considerar a influência de x_1. A variação adicional explicada $R^2 - r^2_{yx_1}$ dividida por esta diferença máxima possível é a medida que tem um valor máximo possível de 1. Na verdade, como a fórmula acima sugere, esta razão é igual à correlação parcial ao quadrado entre y e x_2, controlados por x_1.

A Figura 11.11 ilustra essa propriedade da correlação parcial ao quadrado. Ela mostra a razão da contribuição parcial de x_2 além daquela de x_1, a saber, $R^2 - r^2_{yx_1}$, dividida pela proporção $(1 - r^2_{yx_1})$ não explicada por x_1. De forma similar, o quadrado de $r_{yx_1 \cdot x_2}$ é igual a:

$$r^2_{yx_1 \cdot x_2} = \frac{R^2 - r^2_{yx_2}}{1 - r^2_{yx_2}},$$

a proporção da variação de y explicada por x_1, menos aquela parte não explicada por x_2.

EXEMPLO 11.9 Correlação parcial dos eventos vividos com o distúrbio mental

Retornamos ao estudo da saúde mental, com y = distúrbio mental, x_1 = eventos vividos, x_2 = SES. O *software* informa a matriz de correlações:

	Distúrbio Mental (IMPAIR)	Eventos Vividos (LIFE)	SES
Distúrbio Mental (IMPAIR)	1,000	0,372	-0,399
Eventos Vividos (LIFE)	0,372	1,000	0,123
SES	-0,399	0,123	1,000

Assim, $r_{yx_1} = 0{,}372$, $r_{yx_2} = -0{,}339$ e $r_{x_1x_2} = 0{,}123$. Por sua definição, a correlação parcial entre o distúrbio mental e os eventos vividos, controlados pelo SES, é:

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2}r_{x_1x_2}}{\sqrt{\left(1 - r^2_{yx_2}\right)\left(1 - r^2_{x_1x_2}\right)}}$$

$$= \frac{0{,}372 - (-0{,}399)(0{,}123)}{\sqrt{[1 - (-0{,}399)^2](1 - 0{,}123^2)}} = 0{,}463.$$

A correlação parcial, como a correlação de 0,37 entre o distúrbio mental e os eventos vividos, é moderadamente positiva.

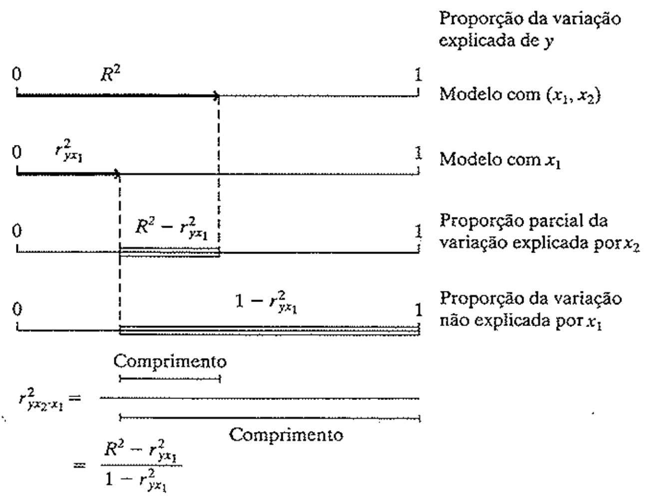


Figura 11.11 Representação de $r^2_{yx_2 \cdot x_1}$ como a proporção da variabilidade que pode ser explicada por x_2 da variabilidade não explicada por x_1.

Visto que $r^2_{yx_1 \cdot x_2} = (0{,}463)^2 = 0{,}21$, controlado o SES, 21% da variação do distúrbio mental são explicados pelos eventos vividos. De forma alternativa, visto que $R^2 = 0{,}339$ (Tabela 11.7),

$$r^2_{yx_1 \cdot x_2} = \frac{R^2 - r^2_{yx_2}}{1 - r^2_{yx_2}}$$

$$= \frac{0{,}339 - (-0{,}399)^2}{1 - (-0{,}399)^2} = 0{,}21. \quad ■$$

Correlações parciais de ordem mais alta

Uma razão por que mostramos a conexão entre os valores da correlação parcial ao quadrado e do R quadrado é que essa abordagem também funciona quando o número de variáveis controle é maior do que um. Por exemplo, com três previsores, considere $R^2_{y(x_1,x_2,x_3)}$ como a representação do valor de R^2. O quadrado da correlação parcial entre y e x_3, controlada por x_1 e x_2, se relaciona ao quão maior isso é do valor do R^2 para o modelo com somente x_1 e x_2 como previsores, que representamos por $R^2_{y(x_1,x_2)}$. A correlação parcial ao quadrado é:

$$r^2_{yx_3 \cdot x_1,x_2} = \frac{R^2_{y(x_1,x_2,x_3)} - R^2_{y(x_1,x_2)}}{1 - R^2_{y(x_1,x_2)}}.$$

Nesta expressão, $R^2_{y(x_1,x_2,x_3)} - R^2_{y(x_1,x_2)}$ é o aumento na proporção da variância explicada pela adição do x_3 ao modelo. O denominador $1 - R^2_{y(x_1,x_2)}$ é a proporção da variação não explicada quando x_1 e x_2 são os únicos previsores no modelo.

A correlação parcial $r_{yx_3 \cdot x_1,x_2}$ é chamada de **correlação parcial de segunda ordem**, visto que ela controla duas variáveis. Ela tem o mesmo sinal do que o b_3 na equação de previsão $\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3$, que também controla x_1 e x_2 na descrição do efeito de x_3.

Inferência para correlações parciais

Controlando certo número de variáveis, a inclinação do efeito parcial de um previsor é zero nas mesmas situações nas quais a correlação parcial entre y e aquele previsor é 0. Uma fórmula alternativa para o teste t para um efeito parcial usa a correlação parcial.

Com previsores k no modelo, a estatística-teste t é:

$$t = \frac{\text{correlação parcial}}{\sqrt{(1 - \text{correlação parcial ao quadrado})/[n - (k + 1)]}}.$$

A estatística tem uma distribuição t com $gl = n - (k + 1)$. Ela é igual à estatística t baseada na estimativa da inclinação parcial e, portanto, tem o mesmo valor-p.

Ilustramos testando se a correlação parcial populacional entre o distúrbio mental (IMPAIR) e os eventos vividos (LIFE), controlados pelo SES é 0. Do Exemplo 11.9, $r_{yx_1 \cdot x_2} = 0{,}463$. Existem $k = 2$ variáveis explicativas e $n = 40$ observações. A estatística-teste é igual a:

$$t = \frac{r_{yx_1 \cdot x_2}}{\sqrt{(1 - r^2_{yx_1 \cdot x_2})/[n - (k + 1)]}}$$

$$= \frac{0{,}463}{\sqrt{[1 - (0{,}463)^2]/37}} = 3{,}18.$$

Este resultado é igual à estatística-teste para H_0: $\beta_1 = 0$ na Tabela 11.5. Portanto, o valor-p é também o mesmo: 0,003.

Quando nenhuma variável é controlada (isto é, o número de variáveis explicativas é $k = 1$), a fórmula da estatística t é mais simples e igual a:

$$t = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}}.$$

Esta é a estatística para testar se a correlação bivariada populacional é igual a 0 (Seção 9.5). Os intervalos de confiança para correlações parciais são mais complexos. Eles requerem uma transformação logarítmica como a mostrada para a correlação no Exercício 9.64 do Capítulo 9.

11.8 COEFICIENTES DE REGRESSÃO PADRONIZADOS*

Como na regressão bivariada (lembre-se da Seção 9.4, na página 301), os tamanhos dos coeficientes de regressão nos modelos de regressão múltipla dependem das unidades de mensuração das variáveis. Para comparar os efeitos relativos de duas variáveis explicativas é apropriado comparar os seus coeficientes somente se as variáveis têm as mesmas unidades. Do contrário, as versões *padronizadas* dos coeficientes de regressão fornecem comparações apropriadas.

> ☑ **Coeficiente de regressão padronizado**
>
> **O coeficiente de regressão padronizado** para uma variável explicativa representa a mudança na média de y, em desvios padrão, para um aumento de um desvio padrão na variável considerada, controladas as demais variáveis explicativas do modelo. Nós os representamos por $\beta_1^*, \beta_2^*, \ldots$

Se $|\beta_2^*| > |\beta_1^*|$, por exemplo, então um aumento de um desvio padrão em x_2 tem um efeito parcial maior em y do que um aumento de um desvio padrão em x_1.

O mecanismo de padronização

Os coeficientes de regressão padronizados representam os valores que os coeficientes de regressão assumem quando as unidades são tais que y e todas as variáveis explicativas têm os mesmos desvios padrão. Padronizamos os coeficientes da regressão parcial ajustando o desvio padrão diferente de y em cada x_i. Considere s_y a representação do desvio padrão amostral de y e considere $s_{x_1}, s_{x_2}, \ldots, s_{x_k}$ a representação dos desvios padrão das variáveis explicativas.

> ☑ As estimativas dos coeficientes de regressão padronizados são:
>
> $$b_1^* = b_1\left(\frac{s_{x_1}}{s_y}\right), \quad b_2^* = b_2\left(\frac{s_{x_2}}{s_y}\right), \ldots$$

EXEMPLO 11.10 Coeficientes padronizados para o distúrbio mental

A equação de previsão relacionando o distúrbio mental aos eventos vividos e ao SES é:

$$\hat{y} = 28{,}23 + 0{,}103x_1 - 0{,}097x_2.$$

A Tabela 11.2 informou os desvios padrão amostrais $s_y = 5{,}5$, $s_{x_1} = 22{,}6$ e $s_{x_2} = 25{,}3$. Visto que o coeficiente não padronizado de x_1 é $b_1 = 0{,}103$, o coeficiente padronizado estimado é:

$$b_1^* = b_1\left(\frac{s_{x_1}}{s_y}\right) = 0{,}103\left(\frac{22{,}6}{5{,}5}\right) = 0{,}43.$$

Visto que $b_2 = -0{,}097$, o valor padronizado é igual a:

$$b_2^* = b_2\left(\frac{s_{x_2}}{s_y}\right) = -0{,}097\left(\frac{25{,}3}{5{,}5}\right) = -0{,}45.$$

A mudança estimada na média de y para um aumento de um desvio padrão em x_1, controlado por x_2, tem magnitude similar da mudança estimada para um aumento de um desvio padrão em x_2, controlado por x_1. Entretanto, o efeito parcial de x_1 é positivo, enquanto o efeito parcial de x_2 é negativo.

A Tabela 11.9, que repete a Tabela 11.5, mostra como o SPSS informa os coeficientes de regressão padronizados. Ele usa o título BETA, refletindo o nome alternativo **pesos beta** para estes coeficientes. ■

Propriedades dos coeficientes de regressão padronizados

Para a regressão bivariada, a padronização do coeficiente de regressão gera a correlação. Para o modelo de regressão múltipla, o coeficiente de regressão parcial padronizado se relaciona com a correlação parcial (Exercício 11.65) e ele geralmente assume um valor semelhante.

Distinto da correlação parcial, entretanto, os b_i^* não precisam estar entre -1 e $+1$. Um valor $|b_i^*| > 1$ ocorre, ocasionalmente, quando x_i está altamente correlacionado com o conjunto de outras variáveis explicativas no modelo. Em tais casos, os erros padrão são geralmente grandes e as estimativas não são confiáveis.

Visto que o coeficiente de regressão padronizado é um múltiplo do coeficiente não padronizado, se um é igual a zero o outro também será. O teste H_0: $\beta_i^* = 0$ é equivalente ao teste t de H_0: $\beta_i = 0$. Não é necessário ter testes separados para esses coeficientes. Na amostra, as magnitudes dos $\{b_i^*\}$ tem os mesmos tamanhos relativos que as estatísticas t daqueles testes. Por exemplo, o previsor com o efeito parcial padronizado maior é aquele que tem a estatística t maior em valor absoluto.

A forma padronizada da equação de previsão*

As equações de regressão têm uma expressão que usa os coeficientes de regressão padronizados. Nesta equação, as variáveis aparecem na forma padronizada.

> ☑ **Notação para variáveis padronizadas**
>
> Considere $z_y, z_{x_1}, \ldots, z_{x_k}$ a representação das versões padronizadas das variáveis $y, x_1, \ldots, x_k$. Por exemplo, $z_y = (y - \bar{y})/s_y$ representa o número de desvios padrão que uma observação em y está da sua média.

Cada escore de um sujeito em $y, x_1, \ldots, x_k$ tem escores-z correspondentes $z_y, z_{x_1}, \ldots, z_{x_k}$. Se o escore de um sujeito em x_1 é tal que $z_{x_1} = (x_1 - \bar{x}_1)/s_{x_1} = 2{,}0$, por exemplo, então aquele sujeito está dois desvios padrão acima da média $\bar{x}_1$ naquela variável.

Considere $\hat{z}_y = (\hat{y} - \bar{y})/s_y$ a representação do escore-z previsto para a variável resposta. Para as variáveis padronizadas e os coeficientes de regressão padronizados estimados, a equação de previsão é:

$$\hat{z}_y = b_1^* z_{x_1} + b_2^* z_{x_2} + \cdots + b_k^* z_{x_k}.$$

Esta equação prevê quão longe uma observação em y está da sua média, em unidades de desvios padrão, baseado em quão longe as variáveis explicativas estão das suas médias, em unidades de desvios padrão. Os coeficientes padronizados são os pesos agregados às variáveis explicativas padronizadas na contribuição à variável resposta padronizada prevista.

EXEMPLO 11.11 Equação de previsão padronizada para o distúrbio mental

O Exemplo 11.10 encontrou que os coeficientes de regressão padronizados estimados para os previsores eventos vividos e SES do distúrbio mental são $b_1^* = 0{,}43$ e $b_2^* = -0{,}45$. A equação de previsão re-

☑ **Tabela 11.9** Saída para o ajuste do modelo de regressão múltipla para os dados do distúrbio mental

	Coeficientes não padronizados		Coeficientes padronizados		
	B	Erro padrão	Beta	t	Sig.
(Constante)	28,230	2,174		12,984	0,000
LIFE	0,103	0,032	0,428	3,177	0,003
SES	−0,097	0,029	−0,451	−3,351	0,002

lacionando as variáveis padronizadas é, portanto:

$$\hat{z}_y = 0{,}43z_{x_1} - 0{,}45z_{x_2}.$$

Considere um sujeito que está dois desvios padrão acima da média em eventos vividos, mas dois desvios padrão abaixo da média em SES. Este sujeito tem um distúrbio mental padronizado previsto de:

$$\hat{z}_y = 0{,}43(2) - 0{,}45(-2) = 1{,}8.$$

O distúrbio mental previsto para aquele sujeito está 1,8 desvios padrão acima da média. Se a distribuição de distúrbio mental é aproximadamente normal, este sujeito pode muito bem ter problemas de saúde mental, visto que apenas aproximadamente 4% dos escores em uma distribuição normal estão, pelo menos, 1,8 desvios padrão acima da sua média. ■

Na equação de previsão com variáveis padronizadas, não aparece o termo intercepto. Por quê? Quando todas as variáveis explicativas padronizadas são iguais a 0, todas estas variáveis são iguais às suas médias. Então, $\hat{y} = \bar{y}$, tal que:

$$\hat{z}_y = \frac{\hat{y} - \bar{y}}{s_y} = 0.$$

Assim, isso simplesmente nos diz que um sujeito que está na média em cada variável explicativa tem previsão de estar na média na variável resposta.

Precauções na comparação dos coeficientes de regressão padronizados

Para avaliar qual previsor em um modelo de regressão múltipla tem o maior impacto na variável resposta, é tentador comparar seus coeficientes de regressão padronizados. Faça tais comparações com cuidado. Em alguns casos as diferenças observadas em b_i^* podem simplesmente refletir o erro amostral. Em particular, quando existe multicolinearidade, os erros padrão são altos e os coeficientes padronizados estimados podem ser instáveis.

Para que um coeficiente de regressão padronizado faça sentido, a variação na variável previsora deve ser representativa da variação na população de interesse. É inapropriado comparar o efeito padronizado de um previsor a outros se o estudo propositadamente amostrou valores daquele previsor em um intervalo pequeno. Esse comentário está relacionado a um aviso da Seção 9.6 (página 315) sobre a correlação: o valor depende fortemente do intervalo dos valores amostrados do previsor.

Tenha em mente, também, que os efeitos são parciais dependendo das outras variáveis que estão no modelo. Uma variável explicativa que parece importante em um sistema de variáveis pode parecer sem importância quando outras variáveis são controladas. Por exemplo, é possível que $|b_2^*| > |b_1^*|$ em um modelo com duas variáveis explicativas, mas quando uma terceira variável explicativa é adicionada ao modelo, $|b_2^*| < |b_1^*|$.

Não é necessário padronizar para comparar o efeito da mesma variável para dois grupos, como na comparação dos resultados de regressões separadas para mulheres e homens, visto que as unidades de mensuração são as mesmas em cada grupo. Na verdade, geralmente é imprudente padronizar neste caso porque os coeficientes padronizados são mais suscetíveis do que os coeficientes não padronizados para diferenças nos desvios padrão dos previsores. Dois grupos que têm os mesmos valores para um coeficiente de regressão estimado têm coeficientes padronizados diferentes se o desvio padrão do previsor difere para os dois grupos.

Finalmente, se uma variável explicativa estiver altamente correlacionada com um conjunto de outras variáveis explicativas, é artificial conceber a mudança daquela variável enquanto as outras permanecem fixas. Como um exemplo extremo, suponha que y = peso, x_1 = comprimento da perna esquerda, x_2 = comprimento da perna direita. A correlação entre x_1 e x_2 está muito próxima de um. Não faz muito sentido imaginar como y muda à medida que x_1 muda enquanto x_2 é controlado.

11.9 RESUMO DO CAPÍTULO

Este capítulo generalizou o modelo de regressão bivariada para incluir variáveis explicativas adicionais. A **equação de regressão múltipla** relacionando uma variável resposta y a um conjunto de variáveis explicativas k é:

$$E(y) = \alpha + \beta_1x_1 + \beta_2x_2 + \ldots + \beta_kx_k.$$

- Os $\{\beta_i\}$ são os **coeficientes de regressão parcial**. O valor β_i é a mudança na média de y para uma mudança de uma unidade em x_i, controlando as outras variáveis no modelo.
- A **correlação múltipla** R descreve a associação entre y e o conjunto de variáveis explicativas. Ele é igual à correlação entre os valores de y observados e previstos. Ele varia entre 0 e 1.
- R^2 = (SQT − SQE)/SQT representa a *redução proporcional no erro* da previsão de y usando a equação de previsão $\hat{y} = a + b_1x_1 + \ldots + b_kx_k$, em vez de $\bar{y}$. Ele é igual ao quadrado da correlação múltipla.
- Uma **correlação parcial**, como $r_{yx_1 \cdot x_2}$, descreve a associação entre duas variáveis, controladas as demais. Seu valor está entre −1 e +1.
- A correlação parcial ao quadrado entre y e x_i representa a proporção de variação em y que pode ser explicada por x_i, excluída a parte que não foi explicada pelo conjunto de variáveis controle.
- Uma **estatística** F testa H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$, isto é, que a variável resposta é independente de todos os previsores. Um valor-p pequeno sugere que pelo menos um previsor afeta a resposta.
- Testes t individuais e intervalos de confiança para $\{\beta_i\}$ analisam os efeitos parciais de cada previsor, controlando as outras variáveis no modelo.
- A **interação** entre x_1 e x_2 nos seus efeitos na média de y significa que o efeito de cada previsor muda à medida que o valor do outro previsor muda. Podemos permitir isto introduzindo o produto cruzado das variáveis explicativas ao modelo, como o termo $\beta_3(x_1x_2)$.
- Para **comparar os modelos de regressão**, *completo* e *reduzido*, o teste F compara os valores das SQE ou os valores dos R^2.
- Os **coeficientes de regressão padronizados** não dependem das unidades de medida. O coeficiente padronizado estimado b_i^* descreve a mudança em y, em unidades de desvios padrão, para um aumento de um desvio padrão em x_i, controlado pelas demais variáveis explicativas.

Como exemplo, com k = 2 variáveis explicativas, a equação de previsão é:

$$\hat{y} = a + b_1x_1 + b_2x_2.$$

Fixando x_2, uma linha reta descreve a relação entre y e x_1. Sua inclinação b_1 é a mudança em $\hat{y}$ para uma unidade de aumento em x_1, controlado x_2. A correlação múltipla R é, pelo menos, tão grande quanto a correlação entre y e cada previsor. A correlação parcial ao quadrado $r^2_{yx_2 \cdot x_1}$ é a proporção da variação de y que é explicada por x_2, exceto a parte da variação que não foi explicada por x_1. O coeficiente da regressão padronizado estimado $b_1^* = b_1(s_{x_1}/s_y)$ descreve o efeito da mudança em x_1 quando x_2 é controlado.

A Tabela 11.10 resume as propriedades básicas e os métodos de inferência para estas medidas e aquelas introduzidas no Capítulo 9 para a regressão bivariada.

O modelo estudado neste capítulo é ainda um pouco restrito no sentido de que todos os previsores são quantitativos. O próximo capítulo mostra como incluir previsores categóricos ao modelo.

☑ **Tabela 11.10** Resumo dos modelos de regressão múltipla e bivariada

	REGRESSÃO BIVARIADA	REGRESSÃO MÚLTIPLA	
Equação de previsão do modelo	$E(y) = \alpha + \beta x$ $\hat{y} = a + bx$	$E(y) = \alpha + \beta_1 x_1 + ... + \beta_k x_k$ $\hat{y} = a + b_1 x_1 + ... + b_k x_k$	
		Efeito simultâneo de $x_1, ..., x_k$	Efeito parcial de um x_i
Propriedades das medidas	b = inclinação r = correlação, inclinação padronizada, $-1 \le r \le 1$, r tem o mesmo sinal que b	R = correlação múltipla, $0 \le R \le 1$	b_i = inclinação parcial b_i^* = coeficiente padronizado da regressão
	r^2 = medida RPE, $0 \le r^2 \le 1$	R^2 = medida RPE, $0 \le R^2 \le 1$	Correlação parcial, $-1 \le r_{yx_1 \cdot x_2} \le 1$, mesmo sinal que b_i e b_i^*, $r^2_{yx_1 \cdot x_2}$ *é a medida* da RPE
Teste sem associação	H_0: $\beta = 0$ ou H_0: $\rho = 0$, y não está associado a x	H_0: $\beta_1 = ... = \beta_k = 0$ (y não está associado a $x_1, ...x_k$)	H_0: $\beta_i = 0$ ou H_0: corr. parcial da pop. = *0*, y não está associado a x_i, controladas as demais variáveis x
Estatística-teste	$t = \frac{b}{ep} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$ $gl = n - 2$	$F = \frac{\text{MQ da Regressão}}{\text{MQ dos Resíduos}} = \frac{R^2/k}{(1-R^2)/[n-(k+1)]}$, $gl_1 = k$, $gl_2 = n - (k+1)$	$t = \frac{b_i}{ep}$ $gl = n - (k + 1)$

EXERCÍCIOS

Praticando o básico

11.1 Para os estudantes da Walden University, o relacionamento entre y = GPA na universidade (no intervalo de 0 a 4,0), x_1 = GPA no ensino médio (no intervalo de 0 a 4,0) e x_2 = escore no SAT (no intervalo de 200 a 800) satisfaz $E(y) = 0{,}20 + 0{,}50x_1 + 0{,}002x_2$.

(a) Encontre o GPA universitário médio para estudantes que têm (i) GPA no ensino médio = 4,0 e escore no SAT = 800, (ii) $x_1 = 3{,}0$ e $x_2 = 300$.

(b) Mostre que o relacionamento entre y e x_1 para os estudantes com x_2 = 500 é $E(y) = 1{,}2 + 0{,}5x_1$.

(c) Mostre que quando $x_2 = 600$, $E(y) = 1{,}4 + 0{,}5x_1$. Portanto, aumentando x_2 por 100 desloca a linha relacionando y a x_1 para cima por $100\beta_2$ = 0,2 unidades.

(d) Mostre que ajustar x_1 em uma variedade de valores gera um conjunto de linhas paralelas, cada uma tendo uma inclinação de 0,002, relacionando a média de y a x_2.

11.2 Para dados recentes da Flórida sobre y = preço de venda de casas (em dólares), x_1 =tamanho da casa (em pés quadrados) e x_2 = tamanho do terreno (em pés quadrados), a equação de previsão é $\hat{y} = -10{,}536 + 53{,}8x_1 + 2{,}84x_2$.

(a) Uma casa em particular de 1240 pés quadrados com um terreno de 18000 pés quadrados foi vendida por \$145000. Encontre o preço de venda previsto, o resíduo e interprete.

(b) Para um tamanho de fixo de terreno, qual é o preço previsto de venda das casas para cada aumento de um pé quadrado no tamanho da casa? Por quê?

11.3 Considere o exercício anterior:

(a) Para um tamanho fixo de uma casa, em quanto o tamanho do terreno deveria aumentar para ter o mesmo impacto de um aumento de um pé quadrado no tamanho da casa?

(b) Suponha que os preços de venda das casas são trocados de dólares para milhares de dólares. Explique por que a equação de previsão muda para $\hat{y} = -10{,}536 + 0{,}0538x_1 + 0{,}00284x_2$.

11.4 Use um *software* com o arquivo de dados *2005 statewide crime* do *site* do livro com a taxa de assassinatos (número de assassinatos por 100000 pessoas) como a variável resposta e com o percentual de ensino médio completo e a taxa de pobreza (percentual da população com rendimento abaixo do índice de pobreza) como variáveis explicativas.

(a) Construa os diagramas de regressão parciais. Interprete. Você vê observações incomuns?

(b) Informe a equação de previsão. Explique como interpretar os coeficientes estimados.

(c) Refaça a análise após apagar a observação para o D.C. Descreva a influência desta observação no efeito previsto da taxa de pobreza. O que isso informa sobre o quanto os valores atípicos são influentes?

11.5 Uma análise de regressão com dados recentes das Nações Unidas sobre vários países com y = percentual de pessoas que usam a internet sobre x_1 = produto interno *per capita* (em milhares de dólares) e x_2 = percentual de pessoas que usam o telefone celular apresenta os resultados na Tabela 11.11.

(a) Escreva a equação de previsão.

(b) Encontre o uso da internet previsto para um país com o PIB *per capita* de \$10000 e 50% de uso do telefone celular.

(c) Encontre as equações de previsão para quando o uso do telefone celular é (i) 0%, (ii) 100% e use-as para interpretar o efeito do PIB.

(d) Use as equações em (c) para explicar a propriedade do modelo "sem interação".

☑ **Tabela 11.11**

	B	Erro padrão	t	Sig
(Constante)	−3,601	2,506	−1,44	0,159
PIB	1,2799	0,2703	4,74	0,000
CELULAR	0,1021	0,0900	1,13	0,264

R quadrado 0,796

ANOVA

	Soma dos quadrados	GL
Regressão	10316,8	2
Erro residual	2642,5	36
Total	12959,3	38

11.6 Considere o exercício anterior.

(a) Mostre como obter o R ao quadrado a partir das somas dos quadrados na tabela da ANOVA. Interprete-o.

(b) r^2 = 0,78 quando o PIB é o único previsor. Por que você acha que o R^2 não aumenta muito quando o uso do telefone celular é adicionado ao modelo, embora ele próprio esteja altamente associado a y (com r = 0,67)? (*Dica:* você esperaria que x_1 e x_2 estivessem altamente correlacionados? Se for assim, qual o efeito?)

11.7 A Tabela 9.16 na página 333 mostrou dados dos condados da Flórida para y = taxa de crimes (por 1000 residentes), x_1 = renda média (em milhares de dólares) e x_2 = percentual em ambiente urbano.

(a) A Figura 11.12 mostra um diagrama de dispersão relacionando y a x_1. Faça a previsão do sinal que o efeito estimado de x_1 tem na equação de previsão $\hat{y} = a + bx_1$. Explique.

(b) A Figura 11.13 mostra um diagrama de regressão parcial de y para x_1, controlado x_2. Faça uma previsão do sinal que o efeito estimado de x_1 tem na equação de previsão $\hat{y} = a + b_1x_1 + b_2x_2$. Explique.

(c) A Tabela 11.12 mostra parte de uma saída de um *software* para modelos de regressão múltipla e bivariada. Informe a equação de previsão relacionando y a x_1 e interprete a inclinação.

(d) Informe a equação de previsão relacionando y a ambos x_1 e x_2. Interprete o coeficiente de x_1 e compare-o a (c).

(e) As correlações são $r_{yx_1} = 0{,}43$, $r_{yx_2} = 0{,}68$, $r_{x_1x_2} = 0{,}73$. Use-as para explicar por que o efeito de x_1 parece tão diferente em (c) e (d).

(f) Informe as equações de previsão relacionando a taxa de crimes à renda nos níveis de urbanização de (i) 0, (ii) 50, (iii) 100. Interprete.

☑ **Tabela 11.12**

	B	Erro padrão	t	Sig
(Constante)	−11,526	16,834	−0,685	0,4960
RENDA	2,609	0,675	3,866	0,0003

	B	Erro padrão	t	Sig
(Constante)	40,261	16,365	2,460	0,0166
RENDA	−0,809	0,805	−1,005	0,3189
URBANO	0,646	0,111	5,811	0,0001

11.8 Considere o exercício anterior. Usando um *software* com o arquivo de dados *Florida crime* do *site* do livro:

(a) Construa diagramas de caixa e bigodes para cada variável, diagramas de dispersão e diagramas de regressão parciais entre y e x_1 e y e x_2. Interprete esses diagramas.

(b) Encontre as equações de previsão para os (i) efeitos bivariados de x_1 e x_2, (ii) o modelo de regressão múltipla. Interprete.

(c) Encontre R^2 para o modelo de regressão múltipla e mostre que ele não é muito maior do que o r^2 para o modelo que usa somente a urbanização como variável previsora. Interprete.

11.9 Dados recentes das Nações Unidas de vários países para y = taxa de natalidade (número de nascimentos por 1000 habitantes), x_1 = atividade econômica das mulheres (força de trabalho feminina como percentual da masculina) e x_2 = PIB (*per capita*, em milhares de dólares) tem a equação de previsão $\hat{y} = 34{,}53 - 0{,}13x_1 - 0{,}64x_2$.

(a) Interprete o coeficiente de x_1.

(b) Trace em um único gráfico o relacionamento entre y e x_1 quando $x_2 = 0$, $x_2 = 10$ e $x_2 = 20$. Interprete os resultados.

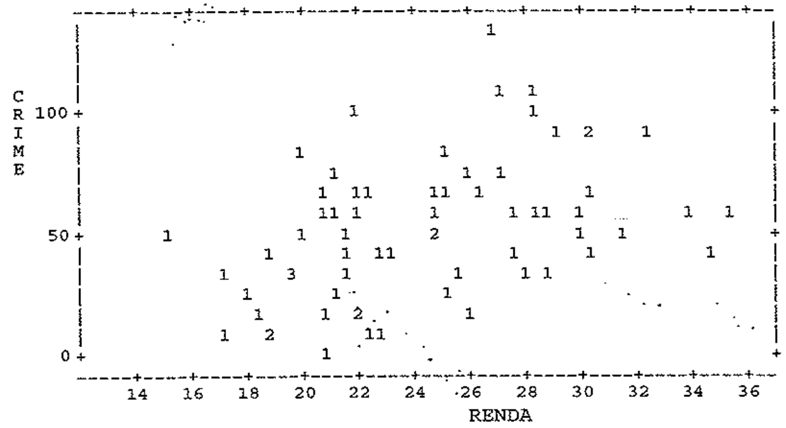
(c) A equação de previsão bivariada com x_1 é $\hat{y} = 37{,}65 - 0{,}31x_1$. As correlações são $r_{yx_1} = -0{,}58$, $r_{yx_2} = -0{,}72$, $r_{x_1x_2} = 0{,}58$. Explique por que o coeficiente de x_1 na equação bivariada é bem diferente daquele da equação de previsão múltipla.

11.10 Para dados recentes das Nações Unidas para vários países, a regressão do uso de dióxido de carbono (CO_2, uma medida da poluição do ar) sobre o produto interno bruto (PIB) tem uma correlação de 0,786 entre as variáveis envolvidas. Com a expectativa de vida como uma segunda variável explicativa, a correlação múltipla é 0,787.

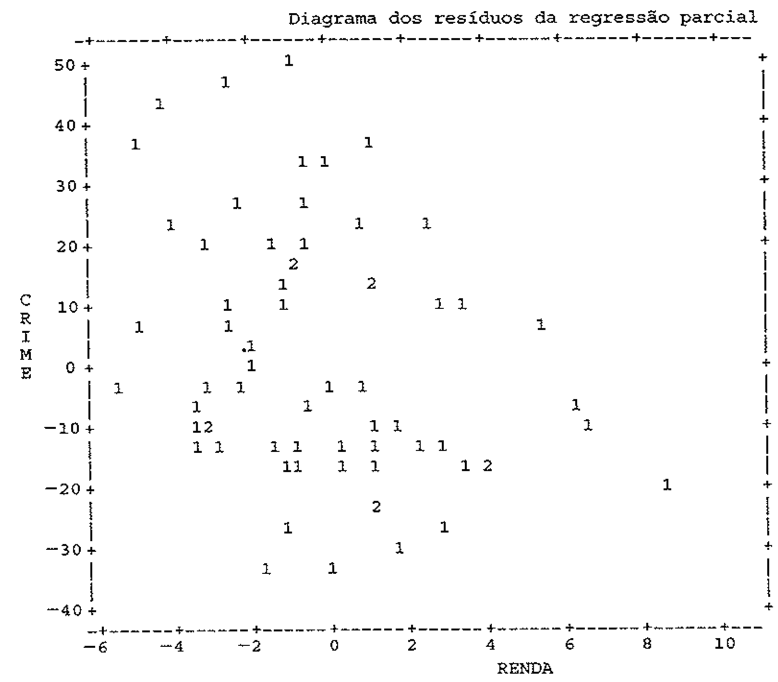
(a) Explique como interpretar a correlação múltipla.

(b) Para prever o CO_2, a adição da expectativa de vida ao modelo ajuda? Isto significa que a expectativa de vida está fracamente correlacionada ao CO_2? Explique.

11.11 A Tabela 11.13 mostra a saída de um *software* ajustando um modelo de regressão múltipla a dados recentes de todos os estados, exceto o D.C., de y = taxa de crimes violentos (por 100000 habitantes) sobre x_1 = taxa de pobreza (percentual com renda abaixo do nível de pobreza) e x_2 = percentual vivendo em áreas urbanas.



☑ **Figura 11.12**



☑ **Figura 11.13**

☑ Tabela 11.13

	Soma dos quadrados	GL	Média dos quadrados	F	Sig
Regressão	2448368,07	2	1224184,04	31,249	0,0001
Resíduo	1841257,15	47	39175,68		
Total	4289625,22	49			

R	R quadrado	Erro padrão da estimativa
0,7555	0,5708	197,928

	B	Erro Padrão	t	Sig
(Constante)	−498,683	140,988	−3,537	0,0009
POBREZA	32,622	6,677	4,885	0,0001
URBANO	9,112	1,321	6,900	0,0001

Correlações

	VIOLENTO	POBREZA	URBANO
VIOLENTO	1,0000	0,3688	0,5940
POBREZA	0,3688	1,0000	−0,1556
URBANO	0,5940	−0,1556	1,0000

(a) Determine a equação de previsão.
(b) Massachusetts tinha $y = 805$, $x_1 = 10{,}7$ e $x_2 = 96{,}2$. Encontre a taxa de crimes violentos prevista. Encontre o resíduo e interprete.
(c) Interprete o ajuste mostrando a equação de previsão relacionando $\hat{y}$ e x_1 para os estados com (i) $x_2 = 0$, (ii) $x_2 = 50$, (iii) $x_2 = 100$. Interprete.
(d) Interprete a matriz de correlações.
(e) Determine o R^2 e a correlação múltipla e interprete.

11.12 Considere o exercício anterior.
(a) Determine a estatística F testando H_0: $\beta_1 = \beta_2 = 0$, informe os valores do *gl* e do valor-*p* e interprete.
(b) Mostre como construir a estatística t para testar H_0: $\beta_1 = 0$, informe os valores do *gl* e do valor-*p* para testar H_0: $\beta_1 \neq 0$ e interprete.
(c) Construa um intervalo de 95% de confiança para β_1 e interprete.
(d) Visto que essas análises usam dados para todos os estados, que relevância, se alguma, as inferências têm em (a) − (c)?

11.13 Considere os dois exercícios anteriores. Quando acrescentamos x_3 = percentual de famílias com mãe ou pai solteiro(a) ao modelo, obtemos os resultados da Tabela 11.14.
(a) Determine a equação de previsão e interprete o coeficiente da taxa de pobreza.

☑ Tabela 11.14

Variável	Coeficiente	Erro padrão
Intercepto	−1197,538	
Pobreza	18,283	6,136
Urbano	7,712	1,109
Pai/mãe solteiro(a)	89,401	17,836
R^2	0,722	
n	50	

(b) Por que você acha que o efeito da taxa de pobreza é muito mais baixo após x_3 ser acrescentado ao modelo?

11.14 A Tabela 11.15 vem de uma análise de regressão de y = número de filhos em uma família em x_1 = nível de instrução da mãe em anos "MEDUC" e x_2 = *status* socioeconômico do pai "PSES", para uma amostra aleatória de 49 estudantes universitários da Texas A&M University.
(a) Escreva a equação de previsão. Interprete as estimativas dos parâmetros.
(b) Informe o SQE. Use-o para explicar a propriedade dos mínimos quadrados dessa equação de previsão.
(c) Explique por que não é possível que $r_{yx_1 \cdot x_2} = 0{,}40$.
(d) Você pode dizer a partir da tabela se r_{yx_1} é positivo ou negativo? Explique.

☑ Tabela 11.15

	Soma dos quadrados
Regressão	31,8
Resíduo	199,3

	b
(Constante)	5,25
MEDUC	−0,24
PSES	0,02

11.15 Uma PSG pediu aos sujeitos para avaliar vários grupos usando o "termômetro perceptivo". A avaliação utilizou uma escala entre 0 e 100, mais favorável à medida que o escore se aproxima de 100 e menos favorável à medida que o escore se aproxima de 0. Para um conjunto de dados pequeno da PSG, a Tabela 11.16 mostra os resultados do ajuste do modelo de regressão múltipla com sentimentos em relação aos liberais como a variável resposta, usando variáveis explicativas de ideologia política (escores 1 = extremamente liberal, 2 = liberal, 3 = moderadamente liberal, 4 = moderado, 5 = levemente conservador, 6 = conservador, 7 = extremamente conservador) e frequência religiosa, usando os escores (1 = nunca, 2 = menos de uma vez por ano, 3 = uma ou duas vezes por ano, 4 = muitas vezes ao ano, 5 = aproximadamente uma vez por mês, 8 = toda semana, 9 = várias vezes por semana). Os erros padrão são mostrados em parênteses.

☑ Tabela 11.16

Variável	Coeficiente
Intercepto	135,31
Ideologia	−14,07 (3,16)**
Religião	−2,95 (2,26)
F	13,93**
R^2	0,799
R^2 Ajust.	0,742
(n)	(10)

(a) Informe a equação de previsão e interprete o efeito parcial da ideologia política.
(b) Informe o valor previsto e o resíduo para a primeira observação, para a qual a ideologia = 7, religião = 9 e sentimentos = 10.
(c) Informe e explique como interpretar o R^2.
(d) As tabelas, deste tipo, geralmente colocam * para um efeito tendo valor-$p < 0{,}05$, ** para um efeito tendo valor-$p < 0{,}01$ e *** para um efeito tendo valor-$p < 0{,}001$. Mostre como isto foi determinado para o efeito da ideologia e discuta a desvantagem de resumir desta maneira.
(e) Explique como o valor-F pode ser obtido do valor R^2 informado. Informe os valores dos *gls* e explique como interpretar o resultado.
(f) Os coeficientes de regressão padronizados estimados são −0,79 para ideologia e −0,23 para religião. Interprete.

11.16 Considere a Tabela 11.5 da página 370. Teste H_0: $\beta_2 = 0$, isto é, que o distúrbio mental é independente de SES, controlado pelos eventos vividos. Informe a estatística-teste e interprete o valor-p para (a) H_a: $\beta_2 \neq 0$, (b) H_a: $\beta_2 < 0$.

11.17 Para uma amostra aleatória de 66 jurisdições do estado, os dados estão disponíveis para:

y = percentual de adultos residentes que estão registrados para votar.

x_1 = percentual de adultos residentes que são proprietários de casas.

x_2 = percentual de adultos residentes que não são brancos.

x_3 = renda média familiar (em milhares de dólares).

x_4 = idade média dos residentes.

x_5 = percentual de residentes que tem residido na jurisdição ao menos por 10 anos.

A Tabela 11.17 mostra uma parte da saída de um *software* para analisar esses dados.

- **(a)** Preencha todos os valores que faltam na saída.
- **(b)** Você acha que é necessário incluir todas as cinco variáveis explicativas no modelo? Explique.
- **(c)** A que teste o "F" se refere? Interprete o resultado do teste.
- **(d)** A que teste o valor-t oposto a x_1 se refere? Interprete o resultado do teste.

11.18 Considere o exercício anterior.

- **(a)** Encontre um intervalo de 95% de confiança para a mudança na média de y para um aumento de uma unidade no percentual de adultos proprietários de casas, controlado pelas demais variáveis. Interprete.
- **(b)** Encontre um intervalo de 95% de confiança para a mudança na média de y para um aumento de 50 unidades no percentual de adultos proprietários de casas, controlado pelas demais variáveis. Interprete.

11.19 Use um *software* com o arquivo de dados *House selling price* (preço de venda de casas) do *site* do livro para realizar uma análise de regressão múltipla de y = preço de venda da casa (em dólares) sobre x_1 = tamanho da casa (em pés quadrados), x_2 = número de quartos, x_3 = número de banheiros.

- **(a)** Use diagramas de dispersão para exibir os efeitos dos previsores em y. Interprete e explique como a natureza altamente discreta de x_2 e x_3 afeta os diagramas.
- **(b)** Informe a equação de previsão e interprete o efeito parcial estimado do tamanho da casa.
- **(c)** Inspecione a matriz de correlações e informe a variável que apresenta a (i) associação mais forte a y, (ii) associação mais fraca a y.
- **(d)** Informe o R^2 para este modelo e r^2 para o modelo mais simples usando somente x_1 como previsor. Interprete.

☑ **Tabela 11.17**

	Soma dos quadrados	GL	Média dos quadrados	F	Sig	R^2
Regressão	------------	----	---------	---	---	----
Resíduo	2940,0	----	---------	Raiz EQM		
Total	3753,3	----				
Variável	Estimativa do parâmetro	Erro padrão	t	Sig		
intercepto	70,0000					
x1	0,1000	0,0450	----	-----		
x2	−0,1500	0,0750	----	-----		
x3	0,1000	0,2000	----	-----		
x4	−0,0400	0,0500	----	-----		
x5	0,1200	0,0500	----	-----		

11.20 Considere o exercício anterior.

- **(a)** Teste o efeito parcial do número de banheiros e interprete.
- **(b)** Encontre a correlação parcial entre o preço de venda e o número de banheiros, controlado pelo número de quartos. Compare o resultado à correlação e interprete.
- **(c)** Encontre as estimativas dos coeficientes de regressão padronizados para o modelo e interprete.
- **(d)** Escreva a equação de previsão usando variáveis padronizadas. Interprete.

11.21 O Exercício 11.11 mostrou uma análise de regressão para os dados de todo o estado para y = taxa de crimes violentos, x_1 = taxa de pobreza e x_2 = percentual residindo em áreas urbanas. Quando acrescentamos um termo de interação, obtemos a equação de previsão $\hat{y}$ = 158,9 − 14,72x_1 − 1,29x_2 + 0,76x_1x_2.

- **(a)** À medida que o percentual residindo em áreas urbanas aumenta, o efeito da taxa de pobreza tende a aumentar ou diminuir? Explique.
- **(b)** Mostre como interpretar a equação de previsão encontrando como pode ser simplificada quando x_2 = 0, 50 e 100.

11.22 Um estudo analisa relacionamentos entre y = percentual que vota no candidato Democrata, x_1 = percentual de eleitores registrados que são Democratas e x_2 = percentual de eleitores registrados que votaram nas várias eleições para o congresso em 2006. Os pesquisadores esperam interação, visto que eles esperam uma curva mais alta entre y e x_1 para valores maiores de x_2 do que para valores menores. Eles obtêm a equação de previsão $\hat{y}$ = 20 + 0,30x_1 + 0,05x_2 + 0,005x_1x_2. Esta equação confirma a direção da sua previsão? Explique.

11.23 Use um *software* para o conjunto de dados *House selling price* (preço de vendas das casas) para permitir a interação entre o número de quartos e o número de banheiros nos efeitos sobre o preço de venda.

- **(a)** Interprete o ajuste mostrando a equação de previsão relacionando $\hat{y}$ e o número de quartos para casas com (i) dois banheiros, (ii) três banheiros.
- **(b)** Teste a significância do termo interação. Interprete.

11.24 Uma análise de regressão múltipla investiga o relacionamento entre y = GPA na universidade e várias variáveis usando uma amostra aleatória de 195 estudantes da Slippery Rock University. Primeiro, o GPA no ensino médio e o escore total no SAT são introduzidos no modelo. A soma dos quadradodos erros é SQE = 20. A seguir, o nível educacional e a renda dos pais são adicionados, para determinar se eles têm um efeito, controlado o GPA no ensino médio e o SAT. Para este modelo expandido SQE = 19. Teste se este modelo completo é significativamente melhor do que aquele contendo somente o GPA no ensino médio e o SAT. Informe e interprete o valor-p.

11.25 A Tabela 11.18 mostra os resultados da regressão de y = taxa de nascimento ("BIRTHS", número de nascimentos por 1000 habitantes) em x_1 = atividade econômica das mulheres ("ECON") e x_2 = taxa de analfabetismo ("LITERACY") usando dados das Nações Unidas de 23 países.

- **(a)** Informe o valor de cada um dos seguintes itens:
  (i) r_{yx_1} (ii) r_{yx_2} (iii) R^2
  (iv) SQT (v) SQE (vi) erro quadrático médio (vii) s (viii) s_y
  (ix) ep para b_1 (x) t para H_0: $\beta_1 = 0$
  (xi) O valor-p para H_0: $\beta_1 = 0$ contra H_a: $\beta_1 \neq 0$
  (xii) O valor-p para H_0: $\beta_1 = 0$ contra H_a: $\beta_1 < 0$
  (xiii) F para H_0: $\beta_1 = \beta_2 = 0$
  (xiv) O valor-p para H_0: $\beta_1 = \beta_2 = 0$
- **(b)** Informe a equação de previsão e interprete os sinais das estimativas dos coeficientes de regressão.
- **(c)** Interprete as correlações r_{yx_1} e r_{yx_2}.
- **(d)** Informe o R^2 e interprete o seu valor.
- **(e)** Informe a correlação múltipla e interprete.

(f) Embora a inferência não seja relevante para estes dados, informe a estatística F para H_0: $\beta_1 = \beta_2 = 0$, e o seu valor-p e interprete.
(g) Mostre como construir a estatística t para H_0: $\beta_1 = 0$ e informe ainda o gl e o valor-p para H_a: $\beta_1 \neq 0$ e interprete.

11.26 Considere o exercício anterior.
(a) Encontre a correlação parcial entre y e x_1, controlado x_2. Interprete a correlação parcial e seu quadrado.
(b) Encontre a estimativa do desvio padrão condicional e interprete o seu valor.
(c) Mostre como encontrar o coeficiente de regressão padronizado esperado para x_1 usando a estimativa não padronizada e os desvios padrão e interprete o seu valor.
(d) Escreva a equação de previsão usando variáveis padronizadas. Interprete.
(e) Encontre o escore-z previsto para um país que está um desvio padrão acima da média em ambos os previsores. Interprete.

11.27 Considere os Exemplos 11.1 (página 362) e 11.8 (página 386). Explique porque a correlação parcial entre a taxa de crimes e a taxa de ensino médio completo é tão diferente da correlação bivariada. (Isto é um exemplo do *paradoxo de Simpson*, que afirma que uma associação bivariada pode ter uma direção diferente de uma associação parcial.)

11.28 Para um grupo de 100 crianças com idades variando de 3 a 15 anos, a correlação entre o escore do vocabulário em um teste de desempenho e a altura da criança é de 0,65. A correlação entre o escore do vocabulário e a idade para esta amostra é de 0,85 e a correlação entre a altura e a idade é de 0,75.
(a) Mostre que a correlação parcial entre o vocabulário e altura, controlada pela idade, é 0,036. Interprete.
(b) Teste se essa correlação parcial é significativamente diferente de zero. Interprete.
(c) É plausível que o relacionamento entre a altura e o vocabulário seja espúrio, no sentido de que ele é devido à dependência conjunta da idade? Explique.

11.29 Um modelo de regressão múltipla descreve o relacionamento entre um conjunto de cidades utilizando y = taxa de assassinatos (número de assassinatos por 100000 habitantes) e:

x_1 = número de policiais (por 100000 habitantes).

x_2 = tempo médio da sentença de prisão dada a assassinos condenados (em anos).

x_3 = renda média dos habitantes da cidade (em milhares de dólares).

x_4 = taxa de desemprego na cidade.

Estas variáveis são observadas em uma amostra aleatória de 30 cidades com populações acima de 35000. Para o modelo com estes previsores, um *software* informa as estimativas dos coeficientes de regressão padronizados de −0,075 para x_1, − 0,125 para x_2, −0,30 para x_3, e 0,20 para x_4.
(a) Escreva a equação de previsão usando variáveis padronizadas.
(b) Que variável explicativa tem maior efeito parcial em y? Explique.
(c) Encontre o escore-z previsto da taxa de assassinatos para uma cidade que está um desvio padrão acima da média em x_1, x_2 e x_3 e um desvio padrão abaixo da média em x_4. Interprete.

11.30 O Exercício 11.11 mostrou uma regressão da taxa de crimes violentos sobre a taxa de pobreza e o percentual de residentes em áreas metropolitanas. Os coeficientes de regressão padronizados estimados são 0,473 para a taxa de pobreza e 0,668 para o percentual em áreas metropolitanas.
(a) Interprete os coeficientes de regressão padronizados estimados.
(b) Expresse a equação de previsão usando variáveis padronizadas e explique como ela é usada.

Conceitos e aplicações

11.31 Considere o arquivo de dados *Student survey* (Exercício 1.11 da página 25). Usando um *software*, execute uma análise de regressão usando y = ideologia política com previsores de número de vezes por semana de leitura de um jornal e religiosidade. Prepare um relatório, propondo um pergunta de pesquisa e resumindo sua análise gráfica, modelos bivariados e interpretações, inferências, verificações e efeitos dos valores atípicos e resumo geral dos relacionamentos.

11.32 Repita o exercício anterior usando y = GPA universitário com previsores GPA no ensino médio e número de horas semanais de exercícios físicos.

11.33 Considere o arquivo de dados que você criou no Exercício 1.12 (página 26). Para as variáveis escolhidas pelo seu professor, ajuste um modelo de regressão múltipla e conduza análises estatísticas inferenciais e descritivas. Interprete e resuma as suas descobertas.

Tabela 11.18

	Média	Desvio padrão	N
BIRTHS	22,117	10,469	23
ECON	47,826	19,872	23
LITERACY	77,696	17,665	23

Correlações

		BIRTHS	ECON	LITERACY
Correlação	BIRTHS	1,00000	−0,61181	−0,81872
	ECON	−0,61181	1,00000	0,42056
	LITERACY	−0,81872	0,42056	1,00000
Sig. (bilateral)	BIRTHS	.	0,0019	0,0001
	ECON	0,0019	.	0,0457
	LITERACY	0,0001	0,0457	.

	Soma dos Quadrados	GL	Média dos Quadrados	F	Sig
Regressão	1825,969	2	912,985	31,191	0,0001
Resíduo	585,424	20	29,271		
Total	2411,393	22			

Raiz EQM (Erro padrão da estimativa) 5,410 R quadrado 0,7572

	Coeficiente não padronizado B	Erro padrão	Coeficiente Padr. (Beta)	t	Sig
(Constante)	61,713	5,2453		11,765	0,0001
ECON	−0,171	0,0640	−0,325	−2,676	0,0145
LITERACY	−0,404	0,0720	−0,682	−5,616	0,0001

11.34 Considere o arquivo *OECD data* no *site* do livro mostrado na Tabela 3.11 na página 80. Proponha uma pergunta de pesquisa sobre como pelo menos duas das variáveis apresentadas naquela tabela se relacionam às emissões de dióxido de carbono. Execute análises apropriadas para responder a esta pergunta e prepare um relatório de duas páginas resumindo as suas análises e conclusões.

11.35 Usando um *software* com o arquivo de dados *2005 statewide crime* do *site* do livro, execute uma análise de regressão para a taxa de crimes violentos tendo como previsores a taxa de pobreza, o percentual residindo em áreas urbanas e o percentual com ensino médio completo. Prepare um relatório no qual você propõe uma pergunta de pesquisa que poderia responder com estes dados, forneça interpretações e resuma as suas conclusões.

11.36 Para o exercício anterior, repita a análise excluindo a observação para o D.C. Descreva o efeito dessa observação nas várias análises.

11.37 A Tabela 9.13 do Capítulo 9 (página 328) é o arquivo de dados *UN data* do *site* do livro. Construa um modelo de regressão múltipla contendo duas variáveis explicativas que forneçam boas previsões para a taxa de fertilidade. Como você selecionou esse modelo? (*Dica:* uma maneira é ter por base as entradas da matriz de correlações.)

11.38 Em aproximadamente 200 palavras, explique para alguém que nunca estudou estatística o que a regressão múltipla faz e como ela pode ser útil.

11.39 Analise o arquivo de dados *House selling price* do *site* do livro (que foi introduzido no Exemplo 9.10 da página 310 usando o preço de venda de uma casa, tamanho da casa, número de quartos e taxas. Prepare um breve relatório resumindo as suas análises e conclusões.

11.40 Para o Exemplo 11.2 sobre o distúrbio mental, a Tabela 11.19 mostra o resultado da adição da frequência religiosa como um previsor, mensurada como o número aproximado de vezes que o sujeito frequenta eventos religiosos ao longo de um ano. Escreva um breve relatório interpretando a informação dessa tabela.

☑ **Tabela 11.19**

Variável	Coeficiente
Intercepto	27,422
Eventos vividos	0,0935
	(0,0313)**
SES	−0,0958
	(0,0256)***
Frequência religiosa	−0,0370
	(0,0219)
R^2	0,358
(n)	(40)

11.41 Um estudo[3] das taxas de mortalidade verificou que, nos Estados Unidos, estados com maior desigualdade na renda tendem a ter taxas mais altas de mortalidade. O efeito da desigualdade de renda desaparecia quando o percentual de residentes que tinham pelo menos o ensino médio completo era controlado. Explique como estes resultados se relacionam às análises conduzidas usando a regressão bivariada e a regressão múltipla.

11.42 Um estudo de 2002[4] relacionando o percentual que uma criança viveu na pobreza ao número de anos de educação completados pela mãe e o percentual que uma criança viveu em um lar de mãe ou pai solteiro(a) apresentou os resultados mostrados na Tabela 11.20. Prepare um relatório de uma página explicando como interpretar os valores dessa tabela.

☑ **Tabela 11.20**

	Coeficientes não padronizados		Coeficientes padronizados		
	B	Erro padrão	Beta	t	Sig
(Constante)	56,401	2,121		12,662	0,000
% pai solteiro	0,323	0,014	0,295	11,362	0,000
escol. Mãe	−3,330	0,152	−0,290	−11,294	0,000

F = 611,6 (gl = 2 e 4731) Sig 0,000

R = 0,453 R quadrado 0,205

11.43 A revista *The economist*[5] desenvolveu um índice de qualidade de vida para nações como o valor previsto obtido regredindo uma média dos escores da satisfação com a vida de vários levantamentos de dados sobre o produto interno bruto (PIB, *per capita*, em dólares), expectativa de vida (em anos), um índice de liberdade política (de 1= completamente livre a 7 = sem liberdade), o percentual de desempregados, a taxa de divórcios (em uma escala de 1 para taxas mais baixas e 5 para mais altas), latitude (para distinguir entre climas mais quentes e mais frios), uma medida de estabilidade política, igualdade de gênero definida como a razão dos ganhos médios entre homens e mulheres e vida comunitária (1 se o país tem uma taxa alta de frequência à igreja ou associados a sindicatos, 0 caso contrário). A Tabela 11.21 mostra os resultados do modelo ajustado para 74 países, para os quais a correlação múltipla é de 0,92. O estudo usou a equação de regressão para prever a qualidade de vida, em 2005, para 111 países. Os 10 melhores avaliados foram Irlanda, Suíça, Noruega, Luxemburgo, Suécia, Austrália, Islândia, Itália, Dinamarca e Espanha. Outras posições incluíam 13 para os Estados Unidos, 14 para o Canadá, 15 para a Nova Zelândia, 16 para a Holanda e 29 para o Reino Unido.

☑ **Tabela 11.21**

	Coeficiente	Erro padrão	Estatística t
Constante	2,796	0,789	3,54
PIB por pessoa	0,00003	0,00001	3,52
Expectativa de vida	0,045	0,011	4,23
Liberdade política	−0,105	0,056	−1,87
Desemprego	−0,022	0,010	−2,21
Taxa de divórcios	−0,188	0,064	−2,93
Latitude	−1,353	0,469	−2,89
Estabilidade política	0,152	0,052	2,92
Igualdade de gênero	0,742	0,543	1,37
Vida comunitária	0,386	0,124	3,13

(a) Que variáveis você esperaria que tivessem efeitos negativos na qualidade de vida? Isto é sustentado pelos resultados?

(b) O estudo afirma que por si só "o PIB explica mais do que 50% da variação da satisfação com a vida". Como isto está relacionado a uma medida da associação?

(c) O estudo relatou que "usando os supostos coeficientes Betas da regressão para derivar os pesos de vários fatores, a expectativa de vida e o PIB foram os mais importantes". Explique o significado disto.

(d) Embora o PIB pareça ser um previsor importante, em um sentido bivariado e em um sentido parcial, a Tabela 11.21 apresenta um coeficiente, muito pequeno, de 0,00003. Por que você acha que isto acontece?

(e) O estudo mencionou outros previsores que não foram incluídos porque não forneciam mais poder de previsão. Por exemplo, o estudo afirmou que a educação parecia ter um efeito principalmente por meio de seus efeitos em outras variáveis no modelo, como o PIB, expectativa de vida e liberdade política. Isto significa que não existe associação entre educação e qualidade de vida? Explique.

11.44 Um artigo recente[6] usou a regressão múltipla para prever atitudes em relação à homossexualidade. Os pesquisadores descobriram que o efeito do número de anos de instrução em uma medida de tolerância em relação à homossexualidade varia de essencialmente sem efeito para políticas conservadoras a um efeito consideravelmente positivo para políticas liberais. Explique como isto é um exemplo de interação estatística e explique como isto seria tratado por um modelo de regressão múltipla.

11.45 No estudo mencionado no exercício anterior, um modelo separado não continha os termos de interação. O melhor previsor das atitudes em relação à homossexualidade era o nível educacional, com um coeficiente de regressão padronizado estimado de 0,21. Os autores também relataram: "Controlando as outras variáveis, um ano adicional de educação completada foi associado com um aumento de 0,09 na unidade de avaliação de atitudes em relação à homossexualidade". Comparando o efeito da educação com os efeitos dos outros previsores no modelo, como a idade do sujeito, explique o objetivo de estimar coeficientes padronizados. Explique como interpretar o determinado para a educação.

11.46 No Exercício 11.1 com y = GPA na universidade, x_1 = GPA no ensino médio e x_2 = escore no SAT, $E(y) = 0{,}20 + 0{,}50x_1 + 0{,}002x_2$. Verdadeiro ou falso: visto que $\beta_1 = 0{,}50$ é maior do que $\beta_2 = 0{,}002$, isto implica que x_1 tem o efeito parcial maior em y. Explique.

11.47 A Tabela 11.22 mostra os resultados do ajuste de vários modelos de regressão para dados de y = GPA na universidade, x_1 = GPA no ensino médio e x_2 = escore do exame de admissão de matemática e x_3 = escore do exame de admissão em língua. Indique quais das seguintes afirmações são falsas.

(a) A correlação entre y e x_1 é positiva.
(b) O aumento de uma unidade em x_1 corresponde a uma mudança de 0,45 na média estimada de y, controlados x_2 e x_3.
(c) Deduz-se dos tamanhos das estimativas para o terceiro modelo que x_1 tem o efeito parcial mais forte em y.
(d) O valor de $r^2_{yx_3}$ é 0,40.
(e) A correlação parcial $r_{yx_1 \cdot x_2}$ é positiva.
(f) Controlando x_1, um aumento de 100 unidades em x_2 corresponde a um aumento previsto de 0,3 no GPA na universidade.
(g) Para o primeiro modelo, o coeficiente estimado de regressão padronizado é igual a 0,50.

11.48 Na análise de regressão, quais das seguintes afirmações devem ser falsas? Por quê?
(a) $r_{yx_1} = 0{,}01$, $r_{yx_2} = -0{,}75$, $R = 0{,}2$
(b) O valor da soma dos quadrados do resíduo, SQE, pode aumentar à medida que acrescentamos variáveis adicionais ao modelo.
(c) Para o modelo $E(y) = a + \beta_1 x_1$, y está significativamente relacionado a x_1, ao nível de 0,05, mas, quando x_2 é adicionado ao modelo, y não está significativamente relacionado a x_1 ao nível de 0,05.
(d) O coeficiente estimado de x_1 é positivo no modelo bivariado, mas negativo no modelo de regressão múltipla.
(e) Quando o modelo é ajustado novamente após y ser multiplicado por 10, R^2, r_{yx_1}, $r_{yx_1 \cdot x_2}$, e as estatísticas F e t não mudam.
(f) A estatística F para testar que todos os coeficientes da regressão são iguais a 0 tem valor-p < 0,05, mas nenhum dos testes t individuais tem valor-p < 0,05.
(g) Se você calcular o coeficiente de regressão padronizado para um modelo bivariado, você sempre obtém a correlação.
(h) $r^2_{yx_1} = r^2_{yx_2} = 0{,}6$ e $R^2 = 1{,}2$.
(i) A correlação entre y e $\hat{y}$ é igual a $-0{,}10$.
(j) Para cada teste F, existe um teste equivalente usando a distribuição t.
(k) Quando $|b_1| > |b_2|$ em uma equação de previsão de regressão múltipla podemos concluir que x_1 tem um efeito mais forte do x_2 em y.
(l) O coeficiente estimado da regressão padronizado para um previsor, em um modelo de regressão múltipla, pode ser interpretado como o valor usual que a inclinação teria na equação de previsão linear se esse previsor e y fossem transformados de modo que ambos tivesse o mesmo desvio padrão.
(m) Se $\hat{y} = 31{,}3 + 0{,}15x_1 - 0{,}05x_2 - 0{,}002x_1x_2$, então o efeito estimado de x_1 em y diminui à medida que x_2 aumenta.
(n) Suponha que $\hat{y} = 31{,}3 + 0{,}15x_1 - 0{,}05x_2 - 0{,}002x_1x_2$, com x_1 e x_2 assumindo valores entre 0 e 100. Então, visto que o coeficiente de x_1x_2 é pequeno quando comparado aos coeficientes de x_1 e de x_2, podemos concluir que a quantidade da interação é desprezível.

Tabela 11.22

	Modelo		
Estimativas	$E(y) = \alpha + \beta x_1$	$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$	$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
Coeficiente de x_1	0,450	0,400	0,340
Coeficiente de x_2		0,003	0,002
Coeficiente de x_3			0,002
R^2	0,25	0,34	0,38

Para os Exercícios de 11.49 a 11.52, selecione a(s) resposta(s) correta(s) e indique por que as outras respostas são inadequadas: (Mais de uma resposta pode estar correta.)

11.49 Se $\hat{y} = 2 + 3x_1 - 5x_2 - 8x_3$, então, controlando x_2 e x_3, a média prevista de y muda, quando x_1 é aumentado de 10 para 20, para:

(a) 3 **(b)** 30 **(c)** 0,3 **(d)** Não pode ser determinada, pois depende dos valores específicos de x_2 e x_3.

11.50 Se $\hat{y} = 2 + 3x_1 - 5x_2 - 8x_3$, então:
(a) A correlação mais forte é entre y e x_3.
(b) A variável com a influência parcial mais forte em y é x_2.
(c) A variável com a influência parcial mais forte em y é x_3, mas não se pode dizer a partir dessa equação que par tem a correlação mais forte.

11.51 Se $\hat{y} = 2 + 3x_1 - 5x_2 - 8x_3$, então:
(a) $r_{yx_3} < 0$
(b) $r_{yx_3 \cdot x_1} < 0$
(c) $r_{yx_3 \cdot x_1, x_2} < 0$
(d) A informação é insuficiente para responder.
(e) As respostas (a), (b) e (c) estão corretas.

11.52 O teste F para comparar o modelo completo a um reduzido:
(a) Pode ser usado para testar a significância de um único parâmetro da regressão em um modelo de regressão múltipla.
(b) Pode ser usado para testar H_0: $\beta_1 = \ldots = \beta_k = 0$ em uma regressão múltipla.
(c) Pode ser usado para testar H_0: Nenhuma interação, no modelo:

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3$$

(d) Pode ser usado para testar se o modelo $E(y) = a + \beta_1 x_1 + \beta_2 x_2$ fornece um ajuste significativamente melhor do que o modelo $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_3$

11.53 Explique a diferença nos objetivos da correlação, da correlação múltipla e da correlação parcial.

11.54 Considere y = altura, x_1 = comprimento da perna direita, x_2 = comprimento da perna esquerda. Descreva o que você espera para os valores relativos de $r_{x_1x_2}$, r_{yx_2}, R e $r_{yx_2 \cdot x_1}$.

11.55 Dê um exemplo de três variáveis para as quais você espera $\beta \neq 0$ no modelo $E(y) = \alpha + \beta x_1$, mas $\beta_1 = 0$ no modelo $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$.

11.56 Para os modelos $E(y) = \alpha + \beta x$ e $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$, expresse a hipótese nula em termos de correlações que são equivalentes ao que segue:
(a) H_0: $\beta = 0$
(b) H_0: $\beta_1 = \beta_2 = 0$
(c) H_0: $\beta_2 = 0$

***11.57** Sempre que x_1 e x_2 não estão correlacionados, então R^2 para o modelo $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$ satisfaz $R^2 = r^2_{yx_1} + r^2_{yx_2}$. Neste caso, trace uma figura que mostre a variabilidade em y, a parte dessa variabilidade que é explicada por cada um dos x_1 e x_2 e a variabilidade total explicada por ambos.

***11.58** Quais dos seguintes conjuntos de correlações você esperaria que gerassem o maior valor de R^2? Por quê?
(a) $r_{yx_1} = 0{,}4$, $r_{yx_2} = 0{,}4$, $r_{x_1x_2} = 0{,}0$
(b) $r_{yx_1} = 0{,}4$, $r_{yx_2} = 0{,}4$, $r_{x_1x_2} = 0{,}5$
(c) $r_{yx_1} = 0{,}4$, $r_{yx_2} = 0{,}4$, $r_{x_1x_2} = 1{,}0$

***11.59** Suponha que a correlação entre y e x_1 é igual à correlação múltipla entre y, x_1 e x_2. O que isto implica em relação à correlação parcial $r_{yx_2 \cdot x_1}$? Interprete.

***11.60** Um *software* informa quatro tipos da soma dos quadrados nos modelos de regressão múltipla. A **soma dos quadrados do Tipo I** (algumas vezes chamado de *sequencial*) representa a variabilidade explicada por uma variável, controladas as variáveis que previamente entraram no modelo. A **soma dos quadrados do Tipo III** (algumas vezes chamada de *parcial*) representa a variabilidade explicada por aquela variável, controladas todas as outras variáveis no modelo.

(a) Para qualquer modelo de regressão múltipla, explique por que a soma dos quadrados do Tipo I para x_1 é a soma dos quadrados da regressão para o modelo bivariado com x_1 como previsor, enquanto a soma dos quadrados do Tipo I para x_2 é igual à quantidade pela qual a SQE diminui quando x_2 é adicionado ao modelo.

(b) Explique por que a soma dos quadrados do Tipo I para a última variável que entrou no modelo é a mesma da soma dos quadrados do Tipo III para aquela variável.

***11.61** O valor amostral de R^2 tende a superestimar o valor populacional porque os dados amostrais estão mais próximos da equação de previsão na amostra do que na verdadeira equação de regressão na população. Essa tendenciosidade é maior se n for pequeno ou o número de previsores k for grande. Uma estimativa melhor é o **R^2 ajustado**:

$$R^2_{\text{ajust}} = 1 - \frac{s^2}{s_y^2} = R^2 - \left[\frac{k}{n - (k+1)}\right](1 - R^2),$$

onde s^2 é a variância condicional estimada e s_y^2 é a variância amostral de y. Usaremos esta medida na Seção 14.1.

(a) Suponha que $R^2 = 0{,}339$ para um modelo com $k = 2$ variáveis explicativas, como na Tabela 11.5. Encontre R^2_{ajust} quando $n = 10, 40$ (como no exemplo do livro), 100 e 1000. Mostre que o R^2_{ajust} se aproxima do R^2 à medida que n aumenta.

(b) Mostre que $R^2_{\text{ajust}} < 0$ quando $R^2 < k/(n-1)$. Isto não é desejado e o R^2_{ajust} é igualado a 0 em tais casos. (Também, diferentemente do R^2, o R^2_{ajust} pode diminuir quando adicionamos uma variável explicativa ao modelo.)

***11.62** Considere $R^2_{y(x_1,\ldots,x_k)}$ a representação do R^2 para o modelo de regressão múltipla com k variáveis explicativas. Explique por que

$$r^2_{yx_k \cdot x_1,\ldots,x_{k-1}} = \frac{R^2_{y(x_1,\ldots,x_k)} - R^2_{y(x_1,\ldots,x_{k-1})}}{1 - R^2_{y(x_1,\ldots,x_{k-1})}}.$$

***11.63** O numerador $R^2 - r^2_{yx_1}$ da correlação parcial ao quadrado $r^2_{yx_2 \cdot x_1}$ fornece o aumento na proporção da variação explicada em virtude da adição do x_2 ao modelo. Este aumento, representado por $r^2_{y(x_2 \cdot x_1)}$, é denominado de correlação **semiparcial** ao quadrado. Podemos usar correlações semiparciais ao quadrado para particionar a variação da variável resposta. Por exemplo, para as três variáveis explicativas:

$$\begin{aligned}R^2_{y(x_1,x_2,x_3)} &= r^2_{yx_1} + (R^2_{y(x_1,x_2)} - r^2_{yx_1}) + \\ &\quad (R^2_{y(x_1,x_2,x_3)} - R^2_{y(x_1,x_2)}) \\ &= r^2_{yx_1} + r^2_{y(x_2 \cdot x_1)} + r^2_{y(x_3 \cdot x_1,x_2)}\end{aligned}$$

A variação total de y explicada por x_1, x_2 e x_3 juntos é particionada em (i) a proporção explicada por x_1 (isto é, $r^2_{yx_1}$), (ii) a proporção explicada por x_2 além daquela explicada por x_1 (isto é, $r^2_{y(x_2 \cdot x_1)}$) e (iii) a proporção explicada por x_3 além daquela explicada por x_1 e x_2 (isto é, $r^2_{y(x_3 \cdot x_1,x_2)}$). Essas correlações (cada uma obtida controlando todos os outros previsores no modelo) têm a mesma ordem do que as estatísticas t para testar os efeitos parciais, e alguns pesquisadores as usam como índices de importância dos previsores.

(a) No Exemplo 11.2 sobre o distúrbio mental, mostre que $r^2_{y(x_2 \cdot x_1)} = 0{,}20$ e $r^2_{y(x_1 \cdot x_2)} = 0{,}18$. Interprete.

(b) Explique por que a correlação semiparcial ao quadrado $r^2_{y(x_2 \cdot x_1)}$ não pode ser maior do que a correlação parcial ao quadrado $r^2_{yx_2 \cdot x_1}$.

***11.64** A equação de previsão dos mínimos quadrados fornece os valores previstos de $\hat{y}$ com a maior correlação possível com y, de todas as possíveis equações de previsões do mesmo tipo. Isto é, a equação dos mínimos quadrados determina a melhor previsão de y no sentido de que ele representa a redução linear de $x_1, \ldots, x_k$ à única variável que está mais fortemente correlacionada com y. Baseado nessa propriedade, explique por que a correlação múltipla não pode diminuir quando se adiciona uma variável a um modelo de regressão múltipla. (*Dica*: a equação de previsão para o modelo mais simples é um caso especial de uma equação de previsão para o modelo completo que tem coeficiente 0 para a variável adicionada.)

***11.65** Considere que $\overline{b_i^*}$ represente o coeficiente estimado da equação de regressão padronizada estimada quando x_i é tratado como a variável *resposta* e y como a variável *explicativa*, controlados pelo mesmo conjunto das demais variáveis. Então, $\overline{b_i^*}$ não precisa ser igual a b_i^*. A correlação parcial ao quadrado entre y e x_i, que *é* simétrica em relação as duas variáveis é igual a:

$$b_i^* \overline{b_i^*}$$

(a) A partir dessa fórmula, explique por que a correlação parcial deve estar entre b_i^* e $\overline{b_i^*}$. (*Nota:* quando $a = \sqrt{bc}$, *a é dito ser a média geométrica de b e c*).

(b) Embora b_i^* não esteja necessariamente entre -1 e $+1$, explique por que $b_i^* \overline{b_i^*}$ não pode exceder a 1.

***11.66** Os Capítulos 12 e 13 mostram como incorporar previsores categóricos nos modelos de regressão e este exercício fornece uma visão antecipada. A Tabela 11.23 mostra parte de uma saída de um modelo para o conjunto de dados *House selling price* do *site* do livro, com y = preço de venda das casas, x_1 = tamanho da casa e x_2 = se a casa é nova (1 = sim, 0 = não).

(a) Especifique a equação de previsão. Tornando $x_2 = 0$ e depois 1, construa duas equações lineares separadas para casas mais antigas e para casas novas. Observe que o modelo implica que o tamanho do efeito da inclinação no preço de venda é o mesmo para cada um.

Tabela 11.23

	B	Erro padrão	t	Sig
(Constante)	−26,089	5,977	−4,365	0,0001
TAMANHO	72,575	3,508	20,690	0,0001
NOVA	19,587	3,995	4,903	0,0001

(b) Visto que x_2 assume somente os valores 0 e 1, explique por que o coeficiente de x_2 estima a diferença dos preços de venda médios entre casas novas e antigas, controlado o tamanho da casa.

***11.67** Considere o exercício anterior. Quando adicionamos um termo de interação, obtemos $\hat{y} = -16{,}6 + 66{,}6x_1 - 31{,}8x_2 + 29{,}4(x_1x_2)$.

(a) Interprete o ajuste determinando a equação de previsão entre o preço de venda e o tamanho da casa separadamente para casas novas ($x_2 = 1$) e para casas antigas ($x_2 = 0$). Interprete. (Esse ajuste é equivalente a determinar as equações separadamente para casas novas e para casas antigas.)

(b) Um diagrama dos dados mostra um valor atípico, uma casa nova com o preço de venda muito alto. Quando aquela observação é removida do conjunto de dados e o modelo é ajustado novamente, $\hat{y} = -16{,}6 + 66{,}6x_1 + 9{,}0x_2 - 5{,}0(x_1x_2)$. Refaça (a) e explique como um valor atípico pode ter um grande impacto em uma análise de regressão.

NOTAS

[1] Desenvolvida por PAYKEL, E. et al. *Archives of General Psychiatry*, v. 75, p. 340-7, 1971.
[2] É igual a $gl_2(gl_2 - 2)$, que geralmente está próximo de 1 a não ser que n seja muito pequeno.
[3] MULLER, A. *British Medical Journal*, v. 324, p. 23-5, 2002.
[4] http://www.heritage.org/Research/Family/cda02-05.cfm.
[5] http://www.economist.com/media/pdf/QUALITYOFLIFE.pdf.
[6] SHACKELFORD, T., BESSER, A. *Individual Differences Research*, v. 5, p. 106-14, 2007.

12 COMPARANDO GRUPOS: MÉTODOS DE ANÁLISE DE VARIÂNCIA (ANOVA)

O Capítulo 7 apresentou métodos para comparar as médias de dois grupos. Nesse capítulo veremos como esses métodos podem ser estendidos para comparar as médias de *vários* grupos.

O Capítulo 8 apresentou métodos para analisar associações entre duas variáveis *categóricas*. Os Capítulos 9 e 11 apresentaram métodos de regressão para analisar a associação entre variáveis *quantitativas*. Os métodos para comparar médias para vários grupos relacionam a associação entre uma variável resposta *quantitativa* e uma variável explicativa *categórica*. A média da variável resposta quantitativa é comparada entre os grupos que são categorias da variável explicativa. Por exemplo, para uma comparação da renda média anual entre negros, brancos e hispânicos, a variável resposta quantitativa é a renda anual e a variável explicativa categórica é o *status* étnico-racial.

O método inferencial para comparar várias médias é denominado de **análise de variância** e é abreviado por **ANOVA**. A Seção 12.1 mostra que o nome se refere à forma como um teste de significância se concentra em dois tipos de variabilidade dos dados. A Seção 12.2 apresenta os intervalos de confiança comparando as médias dos grupos. A Seção 12.3 mostra que as inferências são casos especiais de uma análise de regressão múltipla. As Seções 12.4 e 12.5 estendem esses métodos para incorporar variáveis explicativas adicionais – por exemplo, para comparar a renda média por meio das categorias do *status* étnico-racial e o gênero.

As Seções 12.1 a 12.5 apresentam análises para *amostras independentes*. Como a Seção 7.1 explicou, quando cada amostra tem os mesmos sujeitos, em vez de amostras não emparelhadas, as amostras são *dependentes* e diferentes métodos são aplicados. As Seções 12.6 e 12.7 apresentam esses métodos.

12.1 COMPARANDO VÁRIAS MÉDIAS: O TESTE *F* DA ANÁLISE DE VARIÂNCIA

O notável estatístico britânico Ronald A. Fisher desenvolveu o método de análise de variância em torno de 1920. O ponto principal dessa análise é o teste de significância, usando a distribuição F para detectar as diferenças entre um conjunto de médias populacionais.

Suposições do teste *F* para comparar médias

Considere g a representação do número de grupos a serem comparados, como $g = 3$, como acima, na comparação entre negros, brancos e hispânicos. As médias da variável resposta para as populações correspondentes são $\mu_1, \mu_2, \dots, \mu_g$, como μ_1 para a renda média anual de negros, μ_2 para dos