CAPÍTULO 5

Regressão Linear com Múltiplos Regressores

O Capítulo 4 terminou com uma observação preocupante. Embora diretorias regionais de ensino com razões aluno-professor menores tendam a ter pontuações nos exames maiores na base de dados da Califórnia, talvez os alunos de diretorias com turmas menores tenham outras vantagens que os ajudem a obter um bom desempenho em exames padronizados. Será que isso pode ter gerado resultados enganosos e, se for esse o caso, o que pode ser feito?

Fatores omitidos, como características dos alunos, podem na verdade tornar enganador ou, mais precisamente, viesado o estimador de mínimos quadrados ordinários (MQO) do efeito do tamanho das turmas sobre a pontuação nos exames. Neste capítulo, explicamos o "viés de omissão de variáveis" e apresentamos a regressão múltipla, um método que pode eliminar esse viés. A idéia principal da regressão múltipla é que, se tivermos dados sobre essas variáveis omitidas, poderemos incluí-las como regressores adicionais e, dessa forma, estimar o efeito de um regressor (a razão aluno-professor), mantendo constantes as outras variáveis (como as características dos alunos).

Neste capítulo, explicamos como estimar os coeficientes do modelo de regressão linear múltipla. Mostramos como realizar inferência estatística, isto é, como testar hipóteses sobre os coeficientes da regressão múltipla e construir intervalos de confiança para esses coeficientes. Muitos aspectos da regressão múltipla são semelhantes aos da regressão com um único regressor, estudada no Capítulo 4. Os coeficientes do modelo de regressão múltipla podem ser estimados a partir dos dados pelo uso de MQO; os estimadores de MQO na regressão múltipla são variáveis aleatórias porque dependem de dados de uma amostra aleatória; para amostras grandes, as distribuições amostrais dos estimadores de MQO são aproximadamente normais, e esses estimadores podem ser utilizados para testar hipóteses e construir intervalos de confiança para os coeficientes de regressão da população. Uma hipótese que pode ser testada é de que a redução da razão aluno-professor não possui nenhum efeito sobre a pontuação nos exames, mantendo constantes as características mensuráveis dos alunos da diretoria.

5.1 Viés de Omissão de Variáveis

Ao se concentrar apenas na razão aluno-professor, a análise empírica do Capítulo 4 ignorou alguns determinantes potencialmente importantes da pontuação nos exames ao agrupar suas influências no termo de erro da regressão. Esses fatores omitidos incluem características da escola, como a qualificação dos professores e a utilização de computadores, e características do aluno, como a situação econômica da família. Começamos pela consideração de uma característica omitida do aluno que é particularmente relevante na Califórnia em virtude da grande população de imigrantes: a prevalência na diretoria regional de ensino de alunos que ainda estão aprendendo inglês.

Ao ignorar a porcentagem de alunos que está aprendendo inglês na diretoria, o estimador de MQO da declividade na regressão da pontuação nos exames sobre a razão aluno-professor pode estar viesado; isto é, a média da distribuição amostral do estimador de MQO pode não ser igual ao efeito verdadeiro de uma variação unitária na razão aluno-professor sobre a pontuação nos exames. O raciocínio é o seguinte: os alunos que ainda estão aprendendo inglês podem ter um desempenho inferior nos exames padronizados em relação àqueles cujo inglês é o idioma nativo. Se as diretorias com turmas grandes também tivessem muitos alunos aprendendo inglês, a regressão de MQO da pontuação nos exames sobre a razão aluno-professor poderia encontrar erroneamente uma correlação e produzir um coeficiente estimado grande, quando na realidade o verdadeiro efeito causal da redução

do tamanho das turmas sobre a pontuação nos exames é pequeno, ou mesmo nulo. Assim, com base na análise do Capítulo 4, a superintendente poderá contratar um número suficiente de novos professores para reduzir a razão aluno-professor em dois, porém a melhoria esperada na pontuação dos exames poderá não se concretizar se o coeficiente verdadeiro for pequeno ou nulo.

Um exame dos dados da Califórnia dá credibilidade a essa preocupação. A correlação entre a razão aluno-professor e a porcentagem de alunos aprendendo inglês (alunos para os quais o inglês não é a língua materna e que ainda não se tornaram fluentes) na diretoria é de 0,19. Essa correlação pequena mas positiva sugere que as diretorias com mais alunos aprendendo inglês tendem a apresentar uma razão aluno-professor mais alta (turmas maiores). Se a razão aluno-professor não tivesse relação com a porcentagem de alunos aprendendo inglês, poderíamos seguramente ignorar a proficiência do inglês na regressão da pontuação nos exames contra a razão aluno-professor. Mas, como a razão aluno-professor e a porcentagem de alunos que está aprendendo inglês estão correlacionadas, é possível que o coeficiente da regressão de MQO reflita essa influência.

Definição do Viés de Omissão de Variáveis

Se o regressor (a razão aluno-professor) estiver correlacionado com uma variável que foi omitida da análise (a porcentagem de alunos aprendendo inglês), mas que determina, em parte, a variável dependente (pontuação nos exames), o estimador de MQO terá um **viés de omissão de variáveis**.

Esse viés ocorre quando duas condições são verdadeiras: se a variável omitida está correlacionada com o regressor incluído e se é um determinante da variável dependente. Para ilustrar essas condições, considere três exemplos de variáveis omitidas da regressão da pontuação nos exames sobre a razão aluno-professor.

Exemplo nº 1: Porcentagem de alunos que está aprendendo inglês. Como a porcentagem de alunos que está aprendendo inglês está correlacionada com a razão aluno-professor, a primeira condição para o viés de omissão de variáveis é válida. É plausível que alunos que ainda estejam aprendendo inglês tenham um desempenho inferior nos exames padronizados do que aqueles para os quais o inglês é a língua materna, caso em que a porcentagem de alunos aprendendo inglês é um determinante da pontuação nos exames e a segunda condição para o viés de omissão de variáveis é válida. Desse modo, o estimador de MQO na regressão da pontuação nos exames sobre a razão professor-aluno pode refletir incorretamente a influência da variável omitida, a porcentagem de alunos que está aprendendo inglês. Isto é, omitir a porcentagem de alunos aprendendo inglês pode introduzir um viés de omissão de variáveis.

Exemplo nº 2: Horário do exame. Outra variável omitida da análise é o horário em que o exame é aplicado. Para essa variável omitida, é plausível que a primeira condição para o viés de omissão de variáveis não seja válida, mas que a segunda o seja. Por exemplo, se o horário do exame varia de uma diretoria para a seguinte de uma forma que não esteja relacionada ao tamanho da turma, o horário do exame e o tamanho da turma não estão correlacionados, de modo que a primeira condição não é válida. Por outro lado, o horário do exame pode influenciar a pontuação (a percepção varia ao longo do dia), de modo que a segunda condição é válida. Contudo, como neste exemplo o horário em que o exame é aplicado não está correlacionado com a razão aluno-professor, essa razão não pode estar captando incorretamente o efeito "horário do exame". Portanto, a omissão do horário do exame não resulta em um viés de omissão de variáveis.

Exemplo nº 3: Área de estacionamento por aluno. Outra variável omitida é a área de estacionamento por aluno (área de estacionamento reservada aos professores dividida pelo número de alunos). Essa variável satisfaz a primeira condição para o viés de omissão de variáveis, mas não a segunda. Em especial, escolas com mais professores por aluno provavelmente têm uma área de estacionamento reservada aos professores maior, de modo que a primeira condição é válida. Contudo, sob a hipótese de que o aprendizado ocorre na sala de aula, e não no estacionamento, temos que a área de estacionamento não exerce um efeito direto sobre o aprendizado; desse modo, a segunda condição não é válida. Como a área de estacionamento por aluno não é um determinante da pontuação nos exames, excluí-la da análise não leva a um viés de omissão de variáveis.

O viés de omissão de variáveis está resumido no Conceito-Chave 5.1.

Viés de Omissão de Variáveis em uma Regressão com um Único Regressor

Conceito-Chave 5.1

O **viés de omissão de variáveis** é um viés do estimador de MQO que surge quando o regressor, X, está correlacionado com uma variável omitida. Duas condições devem ser verdadeiras para que ocorra um viés de omissão de variáveis:

1. X deve estar correlacionado com a variável omitida; e
2. a variável omitida deve ser um determinante da variável dependente, Y.

Viés de omissão de variáveis e a primeira hipótese dos mínimos quadrados. O viés de omissão de variáveis indica que a primeira hipótese dos mínimos quadrados — de que $E(u_i \mid X_i) = 0$, conforme relacionado no Conceito-Chave 4.3 — está incorreta. Para ver o porquê, lembre-se de que o termo de erro u_i no modelo de regressão linear com um único regressor representa todos os fatores — exceto X_i — que são determinantes de Y_i. Se um desses outros fatores está correlacionado com X_i, isso significa que o termo de erro (que contém esse fator) está correlacionado com X_i. Em outras palavras, se uma variável omitida é um determinante de Y_i, ela está no termo de erro e, se está correlacionada com X_i, o termo de erro está correlacionado com X_i. Como u_i e X_i estão correlacionados, a média condicional de u_i dado X_i é diferente de zero. Essa correlação, portanto, viola a primeira hipótese dos mínimos quadrados e a conseqüência é séria: o estimador de MQO é viesado. Esse viés não desaparece mesmo em amostras muito grandes, e o estimador de MQO é inconsistente.

Uma Fórmula para o Viés de Omissão de Variáveis

A discussão da seção anterior pode ser resumida matematicamente por uma fórmula para esse viés. Seja a correlação entre X_i e u_i dada por $\text{corr}(X_i, u_i) = \rho_{Xu}$. Suponha que a segunda e a terceira hipóteses de mínimos quadrados sejam válidas, mas que a primeira não seja válida, uma vez que ρ_{Xu} é diferente de zero. Então, o estimador de MQO tem o limite (derivado no Apêndice 5.1),

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{Xu}\frac{\sigma_u}{\sigma_X}. \tag{5.1}$$

Isto é, à medida que o tamanho da amostra aumenta, $\hat{\beta}_1$ fica próximo de $\beta_1 + \rho_{Xu}(\sigma_u/\sigma_X)$ com uma probabilidade cada vez maior.

A fórmula da Equação (5.1) resume várias idéias discutidas anteriormente sobre o viés de omissão de variáveis:

1. O viés de omissão de variáveis é um problema, seja para amostras grandes ou pequenas. Como $\hat{\beta}_1$ não converge em probabilidade para o valor verdadeiro de β_1, $\hat{\beta}_1$ é inconsistente; isto é, $\hat{\beta}_1$ não é um estimador consistente de β_1 quando existe um viés de omissão de variáveis. O termo $\rho_{Xu}(\sigma_u/\sigma_X)$ na Equação (5.1) é o viés em $\hat{\beta}_1$ que persiste mesmo em amostras grandes.

2. Na prática, o fato de o viés ser grande ou pequeno depende da correlação ρ_{Xu} entre o regressor e o termo de erro. Quanto maior for $|\rho_{Xu}|$, maior será o viés.

3. A direção do viés em $\hat{\beta}_1$ depende da existência de correlação positiva ou negativa entre X e u. Por exemplo, suponha que a porcentagem de alunos que está aprendendo inglês tenha um efeito *negativo* sobre a pontuação nos exames da diretoria (alunos que ainda estão aprendendo inglês têm notas menores), de modo que a porcentagem de alunos aprendendo inglês entra no termo de erro com um sinal negativo. Em nossos dados, a fração de alunos que está aprendendo inglês está *positivamente* correlacionada com a razão aluno-professor (diretorias com mais alunos aprendendo inglês possuem turmas maiores). Assim, a razão aluno-professor (X) estaria *negativamente* correlacionada com o termo de erro (u), de modo que

$\rho_{Xu} < 0$ e o coeficiente da razão aluno-professor $\hat{\beta}_1$ estaria viesado na direção de um número negativo. Em outras palavras, uma porcentagem pequena de alunos que está aprendendo inglês está associada tanto com *alta* pontuação nos exames quanto com *baixa* razão aluno-professor, de modo que um dos motivos pelo qual o estimador de MQO sugere que turmas pequenas aumentam a pontuação nos exames pode ser o fato de que as diretorias com turmas menores têm menos alunos aprendendo inglês.

Tratando do Viés de Omissão de Variáveis pela Divisão dos Dados em Grupos

O que você pode fazer quanto ao viés de omissão de variáveis? Nossa superintendente está considerando um aumento do número de professores em sua diretoria, mas ela não tem controle sobre a fração de imigrantes em sua comunidade. Conseqüentemente, ela está interessada no efeito da razão aluno-professor sobre a pontuação nos exames, *mantendo constantes* os demais fatores, incluindo a porcentagem de alunos aprendendo inglês. Essa nova maneira de colocar sua questão sugere que, em vez de utilizar dados de todas as diretorias, talvez devamos nos concentrar naquelas com porcentagens de alunos que estão aprendendo inglês comparáveis às da diretoria da superintendente. Desse subconjunto de diretorias, aquelas com turmas menores têm melhor desempenho nos exames padronizados?

A Tabela 5.1 apresenta a evidência sobre a relação entre o tamanho da turma e a pontuação nos exames das diretorias com porcentagens comparáveis de alunos que estão aprendendo inglês. As diretorias foram divididas em oito grupos. Em primeiro lugar, elas foram divididas em quatro categorias que correspondem aos quartis da distribuição da porcentagem de alunos aprendendo inglês entre as diretorias. Em segundo lugar, dentro dessas quatro categorias, as diretorias foram subdivididas em dois grupos, dependendo de a razão aluno-professor ser pequena ($RAP < 20$) ou grande ($RAP \geq 20$).

O Efeito Mozart: um Caso de Viés de Variáveis?

Um estudo publicado na revista *Nature* em 1993 (Rauscher, Shaw e Ky, 1993) sugeriu que ouvir Mozart por 10 a 15 minutos poderia elevar temporariamente seu QI em 8 ou 9 pontos. Esse estudo provocou um grande impacto — e políticos e pais vislumbraram uma maneira fácil de tornar as crianças mais inteligentes. Durante algum tempo, o Estado da Geórgia distribuiu CDs de música clássica para todas as crianças do Estado.

Qual é a evidência do "efeito Mozart"? Uma resenha de um grande número de estudos constatou que alunos que freqüentam cursos opcionais de música ou artes durante o ensino médio apresentam pontuações maiores nos exames de inglês e matemática do que aqueles que não os freqüentam.[1] Um exame mais detalhado desses estudos, contudo, sugere que o verdadeiro motivo para o melhor desempenho nos exames tem pouco a ver com esses cursos. Os autores desse ensaio sugeriram que a correlação entre o bom desempenho nos exames e as aulas de artes ou música poderia ser o resultado de vários fatores. Por exemplo, alunos com melhor desempenho acadêmico têm mais tempo para freqüentar cursos opcionais de música, ou mais interesse em fazê-lo, ou ainda as escolas com um currículo de música mais extenso podem simplesmente ser melhores em comparação com as demais.

Na terminologia da regressão, a relação estimada entre pontuações nos exames e freqüência nos cursos de música parece ter um viés de variável omitida. Ao omitir fatores tais como a habilidade inata do estudante ou a qualidade da escola como um todo, o estudo de música parece exercer um efeito sobre as pontuações nos exames — quando na verdade isso não ocorre.

Então existe um efeito Mozart? Uma maneira de descobrir isso é conduzir um experimento controlado aleatório. (Conforme a discussão adicional do Capítulo 11, experimentos controlados aleatórios eliminam o viés de omissão de variáveis ao atribuir aleatoriamente participantes a grupos de "tratamento" e de "controle".) Os vários experimentos controlados sobre o efeito Mozart não conseguiram mostrar que ouvir Mozart eleva o QI ou o desempenho geral nos exames. Por motivos ainda não totalmente esclarecidos, contudo, parece que ouvir música clássica *de fato* favorece temporariamente uma área restrita: a relacionada a dobradura de papel e visualização de formas. Portanto, da próxima vez que você tiver de estudar muito para um exame de origami, aproveite para ouvir um pouco de Mozart também.

[1] Veja o *Journal of Aesthetic Education*, v. 34, n. 3-4 (Outono/Inverno 2000), em especial o artigo por Ellen Winner e Monica Cooper (p. 11-76) e o de Lois Hetland (p. 105-148).

TABELA 5.1 Diferenças entre as Pontuações nos Exames das Diretorias Regionais de Ensino da Califórnia com Razões Aluno-Professor Baixas e Altas, por Porcentagem de Alunos que Está Aprendendo Inglês na Diretoria

	Aluno-professor Razão < 20		Aluno-professor Razão ≥ 20		Diferença entre a pontuação, nos exames, *RAP* baixa *versus* alta	
	Pontuação média	*n*	Pontuação média	*n*	Diferença	Estatística *t*
Todas as diretorias	657,4	238	650,0	182	7,4	4,04
Porcentagem aprendendo inglês						
< 2,2%	664,1	78	665,4	27	−1,3	−0,44
2,2–8,8%	666,1	61	661,8	44	4,3	1,44
8,8–23,0%	654,6	55	649,7	50	4,9	1,64
> 23,0%	636,7	44	634,8	61	1,9	0,68

A primeira linha da Tabela 5.1 apresenta a diferença total entre a pontuação média dos exames de diretorias com razões aluno-professor altas e baixas, isto é, a diferença entre a pontuação nos exames desses dois grupos sem dividi-los adicionalmente em quartis de alunos que estão aprendendo inglês. (Lembre-se de que essa diferença foi apresentada anteriormente na forma de regressão na Equação (4.33) como o estimador de MQO do coeficiente de D_i na regressão de *PontExame* sobre D_i, em que D_i é um regressor binário igual a um se $RAP_i < 20$ e igual a zero nos demais casos.) Na amostra completa das 420 diretorias, a pontuação média nos exames é 7,4 pontos maior naquelas com razão aluno-professor baixa do que naquelas em que a razão é alta; a estatística *t* é 4,04, de modo que a hipótese nula de que a pontuação média nos exames é a mesma nos dois grupos é rejeitada ao nível de significância de 1 por cento.

As últimas quatro linhas da Tabela 5.1 apresentam a diferença entre a pontuação nos exames de diretorias com razões aluno-professor altas e baixas subdivididas nos quartis de porcentagem de alunos aprendendo inglês. Essa evidência empírica mostra um quadro diferente. Das diretorias com o menor número de alunos aprendendo inglês (< 2,2 por cento), a pontuação média nos exames para as 78 diretorias com razões aluno-professor baixas é 664,1, e a média para as 27 com razões aluno-professor altas é 665,4. Portanto, para as diretorias com um número menor de alunos aprendendo inglês, a pontuação nos exames foi em média 1,3 ponto *mais baixa* nas diretorias com razões aluno-professor menores! No segundo quartil, diretorias com razões aluno-professor baixas apresentaram pontuação nos exames em média 4,3 pontos mais alta do que aquelas com razões aluno-professor elevadas; essa diferença foi de 4,9 pontos para o terceiro quartil e de apenas 1,9 ponto para o quartil de diretorias com a maioria dos alunos aprendendo inglês. Uma vez mantida constante a porcentagem de alunos que está aprendendo inglês, a diferença de desempenho entre as diretorias com razões aluno-professor altas e baixas talvez seja a metade (ou menos) da estimativa total de 7,4 pontos.

A princípio esse resultado pode parecer um mistério. Como o efeito total da pontuação nos exames pode ser o dobro do efeito da pontuação nos exames dentro de qualquer quartil? A resposta é a seguinte: as diretorias em que a maioria dos alunos está aprendendo inglês tendem a ter *tanto* razões aluno-professor mais elevadas *quanto* pontuações menores nos exames. A diferença na pontuação média nos exames entre as diretorias situadas no quartil mais baixo e no mais alto da porcentagem de alunos aprendendo inglês é grande, de aproximadamente 30 pontos. As diretorias com poucos alunos aprendendo inglês tendem a ter razões aluno-professor menores: 74

por cento (78 de 105) das diretorias no primeiro quartil de alunos aprendendo inglês possuem turmas pequenas ($RAP < 20$), ao passo que somente 42 por cento (44 de 105) das diretorias no quartil com a maioria de alunos aprendendo inglês possuem turmas pequenas. Desse modo, diretorias em que a maioria dos alunos está aprendendo inglês apresentam pontuações mais baixas nos exames, bem como razões aluno-professor maiores do que as outras diretorias.

Essa análise reforça a preocupação da superintendente de que haja um viés de variável omitida na regressão da pontuação nos exames contra a razão aluno-professor. Examinando dentro dos quartis da porcentagem de alunos que está aprendendo inglês, as diferenças entre as pontuações nos exames na segunda parte da Tabela 5.1 melhoram em relação à análise simples da diferença entre médias da primeira linha da Tabela 5.1. Ainda assim, essa análise não fornece para a superintendente uma estimativa útil do efeito da variação no tamanho da turma sobre a pontuação nos exames, mantendo constante a fração de alunos que está aprendendo inglês. Contudo, é possível obter essa estimativa utilizando o método da regressão múltipla.

5.2 O Modelo de Regressão Múltipla

O **modelo de regressão múltipla** estende o modelo de regressão com uma única variável do Capítulo 4 para incluir variáveis adicionais como regressores. Esse modelo permite estimar o efeito da variação em uma variável (X_{1i}) sobre Y_i mantendo constantes os outros regressores (X_{2i}, X_{3i} e assim por diante). No problema do tamanho da turma, o modelo de regressão múltipla fornece uma maneira de isolar o efeito da razão aluno-professor (X_{1i}) sobre a pontuação nos exames (Y_i), mantendo constante a porcentagem de alunos que está aprendendo inglês na diretoria (X_{2i}).

A Reta de Regressão da População

Suponha por enquanto que existam apenas duas variáveis independentes, X_{1i} e X_{2i}. No modelo de regressão linear múltipla, a relação média entre essas duas variáveis independentes e a variável dependente, Y, é dada pela função linear

$$E(Y_i | X_{1i} = x_1, X_{2i} = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \tag{5.2}$$

onde $E(Y_i | X_{1i} = x_1, X_{2i} = x_2)$ é a expectativa condicional de Y_i dados $X_{1i} = x_1$ e $X_{2i} = x_2$. Isto é, se a razão aluno-professor na i-ésima diretoria (X_{1i}) for igual a um valor x_1 e a porcentagem de alunos que está aprendendo inglês na i-ésima (X_{2i}) for igual a x_2, o valor esperado de Y_i dada a razão aluno-professor e a porcentagem de alunos que está aprendendo inglês será dado pela Equação (5.2).

A Equação (5.2) é a **reta de regressão da população** ou **função de regressão da população** no modelo de regressão múltipla. O coeficiente β_0 é o **intercepto**, o coeficiente β_1 é o **coeficiente da declividade de X_{1i}** ou, simplificando, o **coeficiente de X_{1i}**, e o coeficiente β_2 é o **coeficiente da declividade de X_{2i}** ou, simplificando, o **coeficiente de X_{2i}**. Uma ou mais das variáveis independentes no modelo de regressão múltipla são algumas vezes chamadas de **variáveis de controle**.

A interpretação do coeficiente β_1 na Equação (5.2) é diferente daquela em que X_{1i} é o único regressor: na Equação (5.2), β_1 é o efeito de uma variação unitária em X_1 sobre Y, **mantendo X_2 constante** ou **controlando a influência de X_2**.

Essa interpretação de β_1 segue-se da definição de que o efeito esperado de uma variação em X_1, ΔX_1, sobre Y, mantendo X_2 constante, é a diferença entre o valor esperado de Y quando as variáveis independentes assumem os valores $X_1 + \Delta X_1$ e X_2 e o valor esperado de Y quando as variáveis independentes assumem os valores X_1 e X_2. Da mesma forma, escreva a função de regressão da população da Equação (5.2) como $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ e imagine uma variação de X_1 no montante ΔX_1 sem qualquer variação de X_2, isto é, mantendo X_2 constante. Como X_1 variou, Y terá uma variação de um montante, digamos, ΔY. Após essa variação, o novo valor de Y, $Y + \Delta Y$, é

$$Y + \Delta Y = \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2 X_2. \tag{5.3}$$

Uma equação para ΔY em termos de ΔX_1 é obtida subtraindo-se a equação $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ da Equação (5.3), resultando em $\Delta Y = \beta_1 \Delta X_1$. Isto é,

$$\beta_1 = \frac{\Delta Y}{\Delta X_1}, \text{ mantendo } X_2 \text{ constante.} \tag{5.4}$$

O coeficiente β_1 é o efeito de uma variação unitária em X_1 sobre Y (a variação esperada em Y), mantendo X_2 fixo. Outra expressão utilizada para descrever β_1 é o **efeito parcial** de X_1 sobre Y, mantendo X_2 fixo.

A interpretação do intercepto, β_0, no modelo de regressão múltipla é semelhante à interpretação do intercepto no modelo com um único regressor: é o valor esperado de Y_i quando X_{1i} e X_{2i} são iguais a zero. Simplificando, o intercepto β_0 determina em que ponto do eixo Y a reta de regressão da população se inicia.

O Modelo de Regressão Múltipla da População

A reta de regressão da população na Equação (5.2) é a relação entre Y e X_1 e X_2 válida em média para a população. Contudo, assim como na regressão com um único regressor, essa relação não é precisamente válida, uma vez que muitos outros fatores influenciam a variável dependente. Por exemplo, além da razão aluno-professor e da fração de alunos aprendendo inglês, a pontuação nos exames também recebe influência das características da escola, das características de outros alunos e da sorte. Portanto, a função de regressão da população na Equação (5.2) precisa ser aumentada para incorporar esses fatores adicionais.

Assim como na regressão com um único regressor, os fatores que determinam Y_i além de X_{1i} e X_{2i} são incorporados à Equação (5.2) como um termo de "erro" u_i. Esse termo de erro é o desvio de uma observação em particular (pontuação nos exames na i-ésima diretoria, em nosso exemplo) referente à relação média na população. Portanto, temos

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \; i = 1, \ldots, n, \tag{5.5}$$

onde o subscrito i indica a i-ésima das n observações (diretorias) da amostra.

A Equação (5.5) é o **modelo de regressão múltipla da população** quando existem dois regressores, X_{1i} e X_{2i}.

Em uma regressão com regressores binários, pode ser útil tratar β_0 como o coeficiente de um regressor que sempre é igual a um; pense em β_0 como o coeficiente de X_{0i}, onde $X_{0i} = 1$ para $i = 1, \ldots, n$. Portanto, o modelo de regressão múltipla da população na Equação (5.5) pode ser escrito alternativamente como

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \text{ onde } X_{0i} = 1, \; i = 1, \ldots, n. \tag{5.6}$$

As duas formas de representar o modelo de regressão múltipla, as equações (5.5) e (5.6), são equivalentes.

A discussão até o momento se concentrou no caso de uma única variável adicional, X_2. Na prática, contudo, múltiplos fatores podem ser omitidos do modelo com um único regressor. Por exemplo, ignorar a condição financeira dos alunos pode resultar em um viés de omissão de variáveis, assim como aconteceu ao ignorar a fração de alunos que está aprendendo inglês. Esse raciocínio nos leva a considerar um modelo com três regressores ou, generalizando, um modelo que inclui k regressores. O modelo de regressão múltipla com k regressores, X_{1i}, X_{2i}, ..., X_{ki}, está resumido no Conceito-Chave 5.2.

As definições de homoscedasticidade e heteroscedasticidade no modelo de regressão múltipla são semelhantes às definições correspondentes para o modelo com um único regressor. O termo de erro u_i no modelo de regressão múltipla é **homoscedástico** se a variância da distribuição condicional de u_i dado X_{1i}, ..., X_{ki}, var($u_i | X_{1i}$, ..., X_{ki}), é constante para $i = 1, \ldots, n$ e, portanto, não depende de valores de X_{1i}, ..., X_{ki}. Caso contrário, o termo de erro é **heteroscedástico**.

O modelo de regressão múltipla promete fornecer exatamente o que a superintendente quer saber: o efeito da variação da razão aluno-professor mantendo constantes os demais fatores que estão além de seu controle. Esses fatores abrangem não apenas a porcentagem de alunos que está aprendendo inglês, mas outros fatores mensuráveis que podem influenciar o desempenho nos exames, incluindo a situação econômica dos alunos. Contudo, para podermos ajudar a superintendente na prática, precisamos fornecer a ela estimativas dos coeficientes da população desconhecidos β_0, ..., β_k do modelo de regressão da população calculados utilizando uma amostra de dados. Felizmente, esses coeficientes podem ser estimados utilizando o método dos mínimos quadrados ordinários.

Conceito-Chave 5.2

Modelo de Regressão Múltipla

O modelo de regressão múltipla é

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i,\ i = 1, \ldots, n. \tag{5.7}$$

onde:

- Y_i é a i-ésima observação da variável dependente; X_{1i}, X_{2i}, ..., X_{ki} são as i-ésimas observações sobre cada um dos k regressores e u_i é o termo de erro.
- A reta de regressão da população é a relação válida entre Y e X em média na população:

$$E(Y|X_{1i} = x_1, X_{2i} = x_2, \ldots, X_{ki} = x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

- β_1 é o coeficiente da declividade de X_1, β_2 é o coeficiente de X_2 etc. O coeficiente β_1 é a variação esperada em Y resultante da variação unitária em X_{1i}, mantendo constantes X_{2i}, ..., X_{ki}. Os coeficientes dos outros Xs são interpretados do mesmo modo.
- O intercepto β_0 é o valor esperado de Y quando todos os Xs são iguais a zero. O intercepto pode ser imaginado como o coeficiente de um regressor, X_{0i}, igual a um para todo i.

5.3 Estimador de MQO na Regressão Múltipla

Nesta seção, explicamos como os coeficientes do modelo de regressão múltipla podem ser estimados utilizando MQO.

Estimador de MQO

Na Seção 4.2, mostramos como estimar os coeficientes do intercepto e da declividade no modelo com um único regressor aplicando o método de MQO a uma amostra de observações de Y e X. A idéia principal é que esses coeficientes podem ser estimados pela minimização da soma dos quadrados dos erros de previsão, isto é, pela escolha dos estimadores b_0 e b_1 de modo que minimize $\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$; os estimadores que fazem isso são os estimadores de MQO $\hat{\beta}_0$ e $\hat{\beta}_1$.

O método de MQO também pode ser utilizado para estimar os coeficientes β_0, β_1, ..., β_k no modelo de regressão múltipla. Sejam b_0, b_1, ..., b_k os estimadores de β_0, β_1, ..., β_k. O valor previsto de Y_i, calculado pelo uso desses estimadores, é $b_0 + b_1 X_{1i} + \cdots + b_k X_{ki}$, e o erro ao se prever Y_i é $Y_i - (b_0 + b_1 X_{1i} + \cdots + b_k X_{ki}) = Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki}$. A soma dos quadrados desses erros de previsão entre todas as n observações é, portanto,

$$\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki})^2. \tag{5.8}$$

A soma dos quadrados dos erros para o modelo de regressão linear na Expressão (5.8) é a extensão da soma dos quadrados dos erros dada na Equação (4.6) para o modelo de regressão linear com um único regressor.

Os estimadores dos coeficientes β_0, β_1, ..., β_k que minimizam a soma dos quadrados dos erros na Expressão (5.8) são chamados de **estimadores de mínimos quadrados ordinários (MQO)** de β_0, β_1, ..., β_k. Os estimadores de MQO são representadas por $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_k$.

A terminologia de MQO no modelo de regressão linear múltipla é a mesma do modelo de regressão linear com um único regressor. A **reta de regressão de MQO** é a linha reta construída pelo uso dos estimadores de MQO, isto é, $\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k$. O **valor previsto** de Y_i dado X_{1i}, ..., X_{ki}, com base na reta de regressão de MQO é $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki}$. O **resíduo de MQO** para a i-ésima observação é a diferença entre Y_i e seu valor previsto de MQO, isto é, o resíduo de MQO é $\hat{u}_i = Y_i - \hat{Y}_i$.

Os estimadores de MQO poderiam ser calculados por tentativa e erro: você pode testar valores diferentes de b_0, ..., b_k até que esteja satisfeito por ter minimizado a soma dos quadrados total na Expressão (5.8). É muito mais fácil, contudo, utilizar fórmulas explícitas para os estimadores de MQO que derivamos com o uso de cálculo. As fórmulas para os estimadores de MQO no modelo de regressão múltipla são semelhantes àquelas do Conceito-Chave 4.2 para o modelo com um único regressor. Essas fórmulas estão incluídas nos pacotes estatísticos modernos. No modelo de regressão múltipla, as fórmulas são mais bem expressas e discutidas utilizando-se a notação matricial, de modo que sua apresentação foi postergada para a Seção 16.1.

As definições e a terminologia de MQO na regressão múltipla estão resumidas no Conceito-Chave 5.3.

Aplicação para a Pontuação nos Exames e a Razão Aluno-Professor

Na Seção 4.2, utilizamos MQO para estimar o intercepto e o coeficiente da declividade da regressão que relaciona a pontuação nos exames (*PontExame*) à razão aluno-professor (RAP)a partir das observações das 420 diretorias regionais de ensino da Califórnia; a reta de regressão de MQO estimada, descrita na Equação (4.7), é

$$\widehat{PontExame} = 698{,}9 - 2{,}28 \times RAP. \tag{5.9}$$

Preocupa-nos o fato de que essa relação seja enganosa, uma vez que a razão aluno-professor pode estar captando o efeito de ter muitos alunos aprendendo inglês nas diretorias com turmas grandes. Ou seja, é possível que o estimador de MQO esteja sujeito a um viés de omissão de variáveis.

Agora temos condição de nos dedicar a essa preocupação utilizando MQO para estimar uma regressão múltipla em que a variável dependente é a pontuação nos exames (Y_i) e há dois regressores: a razão aluno-professor (X_{1i}) e a porcentagem de alunos aprendendo inglês na diretoria regional de ensino (X_{2i}) para as 420 diretorias ($i = 1, \ldots, 420$). A reta de regressão de MQO estimada para essa regressão múltipla é

$$\widehat{PontExame} = 686{,}0 - 1{,}10 \times RAP - 0{,}65 \times \%AI, \tag{5.10}$$

onde *%AI* é a porcentagem de alunos da diretoria que está aprendendo inglês. A estimativa de MQO do intercepto ($\hat{\beta}_0$) é 686,0, a estimativa de MQO do coeficiente da razão aluno-professor ($\hat{\beta}_1$) é – 1,10 e a estimativa de MQO do coeficiente da porcentagem de alunos aprendendo inglês ($\hat{\beta}_2$) é – 0,65.

Na regressão múltipla, o efeito estimado de uma variação na razão aluno-professor sobre a pontuação nos exames é aproximadamente a metade do efeito observado quando a razão aluno-professor é o único regressor: na equação com um único regressor (Equação (5.9)), estima-se que uma redução unitária de *RAP* deve aumentar 2,28 pontos na pontuação nos exames, enquanto na equação de regressão múltipla (Equação (5.10)) estima-se que o aumento na pontuação nos exames seja de apenas 1,10 pontos. Essa diferença ocorre porque o coeficiente de *RAP* na regressão múltipla é o efeito de uma variação em *RAP* mantendo *%AI* constante (ou controlando sua influência), ao passo que na regressão simples *%AI* não é mantida constante.

Essas duas estimativas podem ser reconciliadas pela conclusão de que existe um viés de omissão de variáveis na estimativa do modelo com um único regressor na Equação (5.9). Na Seção 5.1, vimos que diretorias com uma porcentagem alta de alunos aprendendo inglês tendem a apresentar não só pontuação baixa nos exames mas também razão aluno-professor alta. Se a fração de alunos que está aprendendo inglês for omitida da regressão, estima-se que a redução da razão aluno-professor deva produzir um efeito maior sobre a pontuação nos exames, porém essa estimativa reflete *tanto* o efeito da variação na razão aluno-professor *quanto* o efeito omitido de um número menor de alunos aprendendo inglês na diretoria.

Chegamos à mesma conclusão de que existe viés de omissão de variáveis na relação entre a pontuação nos exames e a razão aluno-professor por meio de dois caminhos diferentes: o enfoque tabular por meio da divisão dos dados em dois grupos (veja a Seção 5.1) e o enfoque da regressão múltipla (veja a Equação (5.10)). Dentre esses dois métodos, a regressão múltipla possui duas vantagens importantes. Em primeiro lugar, fornece uma estimativa quantitativa do efeito de uma diminuição unitária na razão aluno-professor, que é o que a superintendente precisa saber para tomar sua decisão. Em segundo lugar, ela facilmente se estende para mais de dois regressores, de modo que é possível utilizar a regressão múltipla para controlar outros fatores mensuráveis além da porcentagem de alunos que está aprendendo inglês.

Conceito-Chave 5.3

Estimadores de MQO, Valores Previstos e Resíduos no Modelo de Regressão Múltipla

Os estimadores de MQO, $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_k$, são os valores de b_0, b_1, ..., b_k que minimizam a soma dos quadrados dos erros de previsão $\sum_{i=1}^{n}(Y_i - b_0 - b_1X_{1i} - \cdots - b_kX_{ki})^2$. Os valores previstos de MQO $\hat{Y}_i$ e os resíduos $\hat{u}_i$ são:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1X_{1i} + \cdots + \hat{\beta}_kX_{ki},\ i = 1, ..., n \text{ e} \tag{5.11}$$

$$\hat{u}_i = Y_i - \hat{Y}_i,\ i = 1, ..., n. \tag{5.12}$$

Os estimadores de $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_k$ e o resíduo $\hat{u}_i$ são calculados a partir de uma amostra de n observações de $(X_{1i}, ..., X_{ki}, Y_i)$, $i = 1, ..., n$. Estes são estimadores dos verdadeiros coeficientes da população β_0, β_1, ..., β_k e do termo de erro, u_i.

O restante deste capítulo é dedicado à compreensão e ao uso de MQO no modelo de regressão múltipla. Muito do que você aprendeu sobre o estimador de MQO com um único regressor se aplica à regressão múltipla com poucas ou nenhuma modificação, de modo que nos concentraremos no que é novo nessa regressão. Comecemos com a extensão das hipóteses de mínimos quadrados para o modelo de regressão múltipla.

5.4 Hipóteses de Mínimos Quadrados na Regressão Múltipla

Existem quatro hipóteses de mínimos quadrados no modelo de regressão múltipla. As três primeiras são aquelas da Seção 4.3 para o modelo com um único regressor (veja o Conceito-Chave 4.3) estendidas para permitir múltiplos regressores e que serão discutidas sucintamente. A quarta hipótese é nova e será discutida de maneira mais detalhada.

Hipótese nº 1: a Distribuição Condicional de u_i Dados X_{1i}, X_{2i}, ..., X_{ki} Possui uma Média Igual a Zero

A primeira hipótese é de que a distribuição condicional de u_i dados X_{1i}, ..., X_{ki} possui uma média igual a zero. Essa hipótese é a extensão da primeira hipótese dos mínimos quadrados com um único regressor para o caso de múltiplos regressores. A hipótese indica que algumas vezes Y_i situa-se acima da reta de regressão da população e algumas vezes Y_i está abaixo da reta de regressão da população, mas, na média, entre a população, Y_i situa-se sobre a reta de regressão da população. Portanto, para qualquer valor dos regressores, o valor esperado de u_i é zero. Assim como na regressão com um único regressor, essa é a hipótese principal que torna os estimadores de MQO não viesados. Voltaremos ao viés de omissão de variáveis na regressão múltipla na Seção 5.11.

Hipótese nº 2: $(X_{1i}, X_{2i}, ..., X_{ki}, Y_i)$, $i = 1, ..., n$ São i.i.d.

A segunda hipótese é de que $(X_{1i}, ..., X_{ki}, Y_i)$, $i = 1, ..., n$ são variáveis aleatórias independente e identicamente distribuídas (i.i.d.). Essa hipótese torna-se automaticamente válida se os dados são coletados por amostragem aleatória simples. Os comentários da Seção 4.3 sobre essa hipótese para um único regressor também se aplicam a regressores múltiplos.

Hipótese nº 3: X_{1i}, X_{2i}, ..., X_{ki} e u_i Possuem Quatro Momentos

A terceira hipótese é de que X_{1i}, ..., X_{ki}, e u_i possuem quatro momentos. Assim como a terceira hipótese para o modelo com um único regressor, essa hipótese limita a probabilidade de observância de valores extremamente grandes de X_{1i}, ..., X_{ki}, ou u_i. Essa hipótese é uma condição técnica utilizada nas provas das propriedades das estatísticas de MQO para amostras grandes.

Hipótese nº 4: Não Ocorre Multicolinearidade Perfeita

A quarta hipótese é nova no modelo de regressão múltipla. Ela descarta uma situação inconveniente, chamada multicolinearidade perfeita, em que é impossível calcular o estimador de MQO. Os regressores são considerados **perfeitamente multicolineares** (ou apresentam **multicolinearidade perfeita**) se um dos regressores é uma função linear perfeita dos outros regressores. A quarta hipótese dos mínimos quadrados afirma que os regressores não são perfeitamente multicolineares.

Para ilustrar o que é a multicolinearidade perfeita e mostrar por que ela representa um problema, considere três exemplos de regressões hipotéticas em que um terceiro regressor é adicionado à regressão da pontuação nos exames sobre a razão aluno-professor e a porcentagem de alunos que está aprendendo inglês na Equação (5.10).

Exemplo nº 1: Fração de alunos que está aprendendo inglês. Seja $FrAI_i$ a *fração* de alunos que está aprendendo inglês na i-ésima diretoria, que varia entre zero e um. Se a variável $FrAI_i$ fosse incluída como um terceiro regressor além de RAP_i e $\%AI_i$, os regressores seriam perfeitamente multicolineares. Isso porque $\%AI$ é a *porcentagem* de alunos aprendendo inglês, de modo que $\%AI_i = 100 \times FrAI_i$ para cada diretoria. Portanto, um dos regressores ($\%AI_i$) pode ser escrito como uma função linear perfeita de outro regressor ($FrAI_i$).

Em razão dessa multicolinearidade perfeita, é impossível calcular o estimador de MQO da regressão de $PontExame_i$ sobre RAP_i, $\%AI_i$ e $FrAI_i$. Dependendo do modo como seu pacote econométrico lida com a multicolinearidade perfeita, quando você tentar estimar essa regressão, o pacote fará uma destas três coisas: eliminará uma das variáveis (fazendo arbitrariamente a escolha de qual eliminar); se recusará a calcular o estimador de MQO, apresentando uma mensagem de erro, ou irá travar. O motivo matemático para isso é que a multicolinearidade perfeita produz uma divisão por zero nas fórmulas de MQO.

Intuitivamente, a multicolinearidade perfeita é um problema porque você está pedindo que a regressão responda a uma pergunta sem lógica. Lembre-se de que o coeficiente de $\%AI_i$ é o efeito de uma variação unitária em $\%AI$ sobre a pontuação nos exames, mantendo constantes as outras variáveis. Se uma das outras variáveis for $FrAI$, você estará perguntando sobre o efeito da variação de uma variação unitária na *porcentagem* de alunos que está aprendendo inglês, mantendo constante a *fração* de alunos aprendendo inglês. Como a porcentagem de alunos aprendendo inglês e a fração de alunos aprendendo inglês variam juntas em uma relação linear perfeita, essa pergunta não faz sentido e MQO não pode responder a ela.

Exemplo nº 2: Turmas "não muito pequenas". Seja NMP_i uma variável binária que é igual a um se a razão aluno-professor na i-ésima diretoria for "não muito pequena"; especificamente, NMP_i é igual a um se $RAP_i \geq 12$ e igual a zero nos demais casos. Essa regressão também apresenta multicolinearidade perfeita, mas por um motivo mais sutil do que a regressão do exemplo anterior. Não existe, de fato, em nossa base de dados, diretorias com $RAP_i < 12$; como você pode ver no gráfico de dispersão da Figura 4.2, o menor valor de RAP é 14. Assim, $RAP_i = 1$ para todas as observações. Agora, lembre-se de que podemos considerar para o modelo de regressão linear com um intercepto a inclusão de um regressor X_{0i} que é igual a um para todo i, conforme mostrado na Equação (5.6). Portanto, podemos escrever $NMP_i = 1 \times X_{0i}$ para todas as observações em nossa base de dados; isto é, NMP_i pode ser escrita como uma combinação linear perfeita dos regressores; especificamente, ela é igual a X_{0i}.

Isso ilustra dois pontos importantes sobre a multicolinearidade perfeita. Primeiro, quando a regressão inclui um intercepto, um dos regressores que pode estar envolvido em multicolinearidade perfeita é o regressor "constante" X_{0i}. Segundo, a multicolinearidade perfeita é uma afirmação sobre a base de dados que você tem em mãos.

Conceito-Chave 5.4

Hipóteses de Mínimos Quadrados no Modelo de Regressão Múltipla

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i,\ i = 1, \ldots, n, \text{ onde:}$$

1. u_i possui uma média condicional zero, dados X_{1i}, X_{2i}, ..., X_{ki}, isto é, $E(u_i | X_{1i}, X_{2i}, \ldots, X_{ki}) = 0$;
2. $(X_{1i}, X_{2i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$ são seleções independente e identicamente distribuídas (i.i.d.) de sua distribuição conjunta;
3. $(X_{1i}, X_{2i}, \ldots, X_{ki}, u_i)$ possuem quartos momentos finitos e diferentes de zero, e
4. não existe multicolinearidade perfeita.

Embora seja possível imaginar uma diretoria regional de ensino com menos de 12 alunos por professor, elas não existem em nossa base de dados, de modo que não podemos analisá-las em nossa regressão.

Exemplo nº 3: Porcentagem de alunos que falam inglês. Seja $\%FI_i$ a porcentagem de alunos que "falam inglês" da i-ésima diretoria, definida como a porcentagem de alunos que *não* está aprendendo inglês. Novamente, os regressores serão perfeitamente multicolineares. Como no exemplo anterior, a relação linear perfeita entre os regressores envolve o regressor "constante" X_{0i}: para cada diretoria, $\%FI_i = 100 \times X_{0i} - \%AI_i$.

Este exemplo ilustra outro ponto: a multicolinearidade perfeita é uma característica do conjunto completo de regressores. Se o intercepto (isto é, o regressor X_{0i}) ou $\%AI_i$ fossem excluídos dessa regressão, os regressores não seriam perfeitamente multicolineares.

Soluções para a multicolinearidade perfeita. A multicolinearidade perfeita normalmente ocorre quando há um erro na especificação da regressão. Algumas vezes o erro é fácil de ser detectado (como no primeiro exemplo), mas outras vezes não (como no segundo exemplo). Seja qual for o caso, seu pacote econométrico lhe informará se você cometer esse tipo de erro, uma vez que não poderá calcular o estimador de MQO se houver erro.

Quando seu pacote o informar que há multicolinearidade perfeita, é importante que você modifique sua regressão para eliminá-la. Alguns pacotes perdem a confiabilidade na presença de multicolinearidade perfeita e, no mínimo, você estará delegando o controle de sua escolha de regressores para seu computador se eles forem perfeitamente multicolineares.

Multicolinearidade imperfeita. Apesar do nome parecido, a multicolinearidade imperfeita é, do ponto de vista conceitual, completamente diferente da multicolinearidade perfeita. **Multicolinearidade imperfeita** significa que dois ou mais regressores são altamente correlacionados no sentido de que existe uma função linear dos regressores que é altamente correlacionada com outro regressor. A multicolinearidade imperfeita não suscita nenhum problema para a teoria dos estimadores de MQO; na realidade, um objetivo de MQO é identificar as influências independentes dos vários regressores quando estes são potencialmente correlacionados.

O Conceito-Chave 5.4 resume as hipóteses de mínimos quadrados para o modelo de regressão múltipla.

5.5 Distribuição dos Estimadores de MQO na Regressão Múltipla

Como os dados diferem de uma amostra para a seguinte, amostras diferentes produzem valores diferentes de estimadores de MQO. Essa variação entre amostras possíveis resulta na incerteza associada aos estimadores de MQO dos coeficientes de regressão da população, β_0, β_1, ..., β_k. Assim como na regressão com um único regressor, essa variação está resumida na distribuição amostral dos estimadores de MQO.

Conceito-Chave 5.5

Distribuição de $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ para Amostras Grandes

Se as hipóteses de mínimos quadrados (veja o Conceito-Chave 5.4) são válidas, para amostras grandes, os estimadores de MQOs $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ possuem distribuição normal conjunta e cada $\hat{\beta}_j$ é distribuído $N(\beta_j, \sigma^2_{\hat{\beta}_j})$, $j = 0, \ldots, k$.

Na Seção 4.4, você aprendeu que, sob a hipótese dos mínimos quadrados, os estimadores de MQO ($\hat{\beta}_0$ e $\hat{\beta}_1$) são estimadores não viesados e consistentes dos coeficientes desconhecidos (β_0 e β_1) no modelo de regressão linear com um único regressor. Além disso, em amostras grandes, a distribuição amostral de $\hat{\beta}_0$ e $\hat{\beta}_1$ tem uma aproximação boa por meio de uma distribuição normal bivariada.

Podemos estender esses resultados para a análise de regressão múltipla. Isto é, sob as hipóteses de mínimos quadrados do Conceito-Chave 5.4, os estimadores de MQO $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ são estimadores não viesados e consistentes de $\beta_0, \beta_1, \ldots, \beta_k$ no modelo de regressão linear múltipla. Para amostras grandes, a distribuição amostral conjunta de $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ tem uma aproximação boa por meio de uma distribuição normal multivariada, que é a extensão da distribuição normal bivariada para o caso geral de duas ou mais variáveis aleatórias normais conjuntas (veja a Seção 2.4).

Embora a álgebra seja mais complicada quando existem múltiplos regressores, o teorema central do limite se aplica aos estimadores de MQO no modelo de regressão múltipla pelo mesmo motivo que se aplica a $\overline{Y}$ e aos estimadores de MQO quando há um único regressor: os estimadores de MQO $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ são médias de dados com amostragem aleatória e, se o tamanho da amostra é suficientemente grande, a distribuição amostral dessas médias torna-se normal. Como a distribuição normal multivariada é mais bem tratada matematicamente pelo uso de álgebra matricial, deixamos as expressões para a distribuição conjunta dos estimadores de MQO para o Capítulo 16.

O Conceito-Chave 5.5 resume o resultado de que, para amostras grandes, a distribuição dos estimadores de MQO na regressão múltipla é aproximadamente normal conjunta. De modo geral, os estimadores de MQO são correlacionados; essa correlação surge da correlação entre os regressores. A distribuição amostral conjunta dos estimadores de MQO é discutida de maneira mais detalhada no Apêndice 5.2 para os casos em que há dois regressores e erros homoscedásticos e na Seção 16.2 para o caso geral.

Erros Padrão para os Estimadores de MQO

Lembre-se de que, no caso de um único regressor, foi possível estimar a variância do estimador de MQO substituindo-se as expectativas por médias da amostra, que levaram ao estimador $\hat{\sigma}^2_{\hat{\beta}_1}$ dado na Equação (4.19). Sob as hipóteses de mínimos quadrados, a lei dos grandes números implica que essas médias da amostra convergem para suas correspondentes na população, de modo que, por exemplo, $\hat{\sigma}^2_{\hat{\beta}_1} / \sigma^2_{\hat{\beta}_1} \xrightarrow{p} 1$. A raiz quadrada de $\hat{\sigma}^2_{\hat{\beta}_1}$ é o erro padrão de $\hat{\beta}_1$, $EP(\hat{\beta}_1)$, um estimador do desvio padrão da distribuição amostral de $\hat{\beta}_1$.

Tudo isso se estende diretamente para a regressão múltipla. O estimador de MQO $\hat{\beta}_j$ do j-ésimo coeficiente de regressão possui um desvio padrão que é estimado por seu erro padrão, $EP(\hat{\beta}_j)$. A fórmula para o erro padrão é expressa mais facilmente com o uso de matrizes, de modo que ela é dada na Seção 16.2. O ponto importante é que, no que diz respeito ao erro padrão, não há nada conceitualmente diferente entre os casos de um único regressor e de múltiplos regressores. As idéias principais — a normalidade dos estimadores para amostras grandes e a capacidade de estimar de maneira consistente o desvio padrão de sua distribuição amostral — são as mesmas na presença de um, dois ou 12 regressores.

5.6 Testes de Hipótese e Intervalos de Confiança para um Único Coeficiente

Nesta seção, descrevemos como testar hipóteses e construir intervalos de confiança para um único coeficiente em uma equação de regressão múltipla.

Testes de Hipótese para um Único Coeficiente

Suponha que você queira testar a hipótese de que uma variação na razão aluno-professor não tem nenhum efeito sobre a pontuação nos exames, mantendo constante a porcentagem de alunos que está aprendendo inglês na diretoria. Isso corresponde a supor que a hipótese de que o verdadeiro coeficiente β_1 da razão aluno-professor é zero na regressão da população de pontuação nos exames sobre *RAP* e *%AI*. Generalizando, podemos querer testar a hipótese de que o verdadeiro coeficiente β_j do j-ésimo regressor assume um valor específico, $\beta_{j,0}$. O valor nulo $\beta_{j,0}$ vem da teoria econômica ou, como no exemplo da razão aluno-professor, do contexto de tomada de decisão da aplicação. Se a hipótese alternativa é bicaudal, as duas hipóteses podem ser escritas matematicamente como

$$H_0: \beta_j = \beta_{j,0} \text{ vs. } H_1: \beta_j \neq \beta_{j,0} \text{ (hipótese alternativa bicaudal).} \tag{5.13}$$

Por exemplo, se o primeiro regressor é *RAP*, a hipótese nula de que uma variação na razão aluno-professor não tem efeito sobre o tamanho da turma corresponde à hipótese nula de que $\beta_1 = 0$ (de modo que $\beta_{1,0} = 0$). Nossa tarefa é testar a hipótese nula H_0 contra a alternativa H_1 utilizando uma amostra de dados.

O Conceito-Chave 4.6 fornece um procedimento para testar essa hipótese nula quando há somente um único regressor. O primeiro passo desse procedimento é calcular o erro padrão do coeficiente. O segundo é calcular a estatística t utilizando a fórmula geral do Conceito-Chave 4.5. O terceiro é calcular o valor p do teste usando a distribuição normal acumulada da Tabela 1 do Apêndice ou, alternativamente, comparar a estatística t ao valor crítico correspondente ao nível de significância desejado para o teste. A base teórica desse procedimento é a seguinte: o estimador de MQO possui uma distribuição normal para amostras grandes que, sob a hipótese nula, tem como média o valor verdadeiro da hipótese e a variância dessa distribuição pode ser estimada de maneira consistente.

Essa base também está presente na regressão múltipla. Conforme expresso no Conceito-Chave 5.5, a distribuição amostral de $\hat{\beta}_j$ é aproximadamente normal. Sob a hipótese nula, a média dessa distribuição é $\beta_{j,0}$. A variância dessa distribuição pode ser estimada de maneira consistente. Portanto, podemos simplesmente usar o mesmo procedimento do caso com um único regressor para testar a hipótese nula na Equação (5.13).

O Conceito-Chave 5.6 resume o procedimento para o teste de uma hipótese para um único coeficiente na regressão múltipla. A estatística t efetivamente calculada é representada no Conceito-Chave por t^{ef}. Contudo, é comum representá-la simplesmente por t, e será esta a notação que adotaremos no restante do livro.

Intervalos de Confiança para um Único Coeficiente

O método para se construir um intervalo de confiança na regressão múltipla é o mesmo do modelo com um único regressor. O Conceito-Chave 5.7 resume esse método.

O método para a condução de um teste de hipótese no Conceito-Chave 5.6 e o método para a construção de um intervalo de confiança no Conceito-Chave 5.7 se apóiam na aproximação normal para amostras grandes da distribuição do estimador de MQO $\hat{\beta}_j$. Desse modo, você deve ter em mente que esses métodos que quantificam a incerteza da amostragem só têm garantia de funcionamento para amostras grandes.

Testando a Hipótese $\beta_j = \beta_{j,0}$ contra a Hipótese Alternativa $\beta_j \neq \beta_{j,0}$

Conceito-Chave 5.6

1. Calcule o erro padrão de $\hat{\beta}_j$, $EP(\hat{\beta}_j)$.
2. Calcule a estatística t

$$t = \frac{\hat{\beta}_j - \beta_{j,0}}{EP(\hat{\beta}_j)}. \tag{5.14}$$

3. Calcule o valor p

$$\text{valor } p = 2\Phi(-|t^{ef}|), \tag{5.15}$$

onde t^{ef} é o valor da estatística t efetivamente calculado. Rejeite a hipótese ao nível de significância de 5 por cento se o valor p for menor do que 0,05 ou, de forma equivalente, se $|t^{ef}| > 1{,}96$.

O erro padrão e (normalmente) a estatística t e o valor p para testar $\beta_j = 0$ são calculados automaticamente pelo pacote de regressão.

Aplicação à Pontuação nos Exames e à Razão Aluno-Professor

Podemos rejeitar a hipótese nula de que uma variação na razão aluno-professor não tem efeito sobre a pontuação nos exames, uma vez que controlemos a porcentagem de alunos que está aprendendo inglês na diretoria? O que é um intervalo de confiança de 95 por cento para o efeito de uma variação na razão aluno-professor sobre a pontuação nos exames, controlando-se a porcentagem de alunos que está aprendendo inglês? Agora já somos capazes de responder a essa questão. A regressão da pontuação nos exames contra *RAP* e *%AI*, estimada por MQO, foi dada na Equação (5.10) e é novamente expressa aqui — os erros padrão estão entre parênteses abaixo dos coeficientes:

$$\widehat{PontExame} = 686{,}0 - 1{,}10 \times RAP - 0{,}650 \times \%AI. \tag{5.16}$$
$$(8{,}7)\ \ (0{,}43)\qquad\quad (0{,}031)$$

Para testar a hipótese de que o verdadeiro coeficiente sobre *RAP* é 0, primeiro precisamos calcular a estatística t na Equação (5.14). Como a hipótese nula diz que o valor verdadeiro desse coeficiente é zero, a estatística t é $t = (-1{,}10 - 0)/0{,}43 = -2{,}54$. O valor p associado a ela é $2\Phi(-2{,}54) = 1{,}1$ por cento; isto é, o menor nível de significância para o qual podemos rejeitar a hipótese nula é 1,1 por cento. Como o valor p é menor do que 5 por cento, a hipótese nula pode ser rejeitada ao nível de significância de 5 por cento (mas decididamente não ao nível de significância de 1 por cento).

Um intervalo de confiança de 95 por cento para o coeficiente da população de *RAP* é $-1{,}10 \pm 1{,}96 \times 0{,}43 = (-1{,}95, -0{,}26)$; isto é, podemos estar 95 por cento confiantes de que o verdadeiro valor do coeficiente está entre $-1{,}95$ e $-0{,}26$. Interpretado no contexto do interesse da superintendente de diminuir a razão aluno-professor em 2, o intervalo de confiança de 95 por cento para o efeito dessa redução sobre a pontuação nos exames é $(-1{,}95 \times 2, -0{,}26 \times 2) = (-3{,}90, -0{,}52)$.

Adicionando despesas por aluno à equação. Sua análise da regressão múltipla na Equação (5.16) convenceu a superintendente de que, com base na evidência até o momento, a redução do tamanho da turma ajudará a melhorar a pontuação nos exames em sua diretoria. Agora, contudo, ela parte para uma questão mais sutil. Se contratar mais professores, poderá pagá-los por meio de cortes de outros itens do orçamento (descartando novos computadores, reduzindo gastos com manutenção etc.) ou da solicitação de um aumento em seu orçamento, o que desagrada os contribuintes. Ela pergunta: qual é o efeito de uma redução da razão aluno-professor sobre a pontuação nos exames, mantendo constantes o gasto por aluno e a porcentagem de alunos que está aprendendo inglês?

Conceito-Chave 5.2

Intervalos de Confiança para um Único Coeficiente na Regressão Múltipla

Um intervalo de confiança bicaudal de 95 por cento para o coeficiente β_j é um intervalo que contém o valor verdadeiro de β_j com uma probabilidade de 95 por cento; isto é, contém o valor verdadeiro de β_j em 95 por cento de todas as amostras possíveis selecionadas aleatoriamente. De modo equivalente, ele também é o conjunto de valores de β_j que não podem ser rejeitados por um teste de hipótese bicaudal a um nível de 5 por cento. Quando o tamanho da amostra é grande, o intervalo de confiança de 95 por cento é:

intervalo de confiança de 95 por cento para $\beta_j = (\hat{\beta}_j - 1{,}96EP(\hat{\beta}_j), \hat{\beta}_j + 1{,}96EP(\hat{\beta}_j))$. (5.17)

Um intervalo de confiança de 90 por cento é obtido substituindo-se 1,96 na Equação (5.17) por 1,645.

Essa questão pode ser considerada estimando-se uma regressão da pontuação nos exames sobre a razão aluno-professor, o gasto total por aluno e a porcentagem de alunos que está aprendendo inglês. A reta de regressão de MQO é

$$\widehat{PontExame} = 649{,}6 - 0{,}29 \times RAP + 3{,}87 \times Gasto - 0{,}656 \times \%AI, \quad (5.18)$$
$$(15{,}5)\ (0{,}48) \qquad (1{,}59) \qquad (0{,}032)$$

onde *Gasto* é o gasto total anual (em milhares de dólares) por aluno na diretoria.

O resultado é surpreendente. Mantendo constantes o gasto por aluno e a porcentagem de alunos aprendendo inglês, estima-se que uma variação na razão aluno-professor tenha um efeito muito pequeno sobre a pontuação nos exames: o coeficiente estimado de *RAP* na Equação (5.16) é de –1,10, mas, adicionando-se *Gasto* como regressor na Equação (5.18), ele é de apenas –0,29. Além disso, a estatística *t* que testa se o valor verdadeiro do coeficiente é zero agora é $t = (-0{,}29 - 0)/0{,}48 = -0{,}60$, de modo que a hipótese de que o valor do coeficiente da população é de fato zero não pode ser rejeitada mesmo ao nível de significância de 10 por cento ($|-0{,}60| < 1{,}645$). Portanto, a Equação (5.18) não fornece nenhuma evidência de que a contratação de mais professores irá melhorar a pontuação nos exames se o gasto total por aluno for mantido constante.

Observe que, quando *Gasto* foi incluído, o erro padrão de *RAP* aumentou de 0,43 na Equação (5.16) para 0,48 na Equação (5.18). Isso ilustra o ponto geral de que uma correlação entre regressores (a correlação entre *RAP* e *Gasto* é –0,62) pode tornar os estimadores de MQO menos precisos (veja o Apêndice 5.2 para uma discussão mais detalhada).

E quanto a nosso contribuinte irritado? Ele afirma que os valores da população *tanto* do coeficiente da razão aluno-professor (β_1) *quanto* do coeficiente do gasto por aluno (β_2) são iguais a zero, isto é, sua hipótese é de que $\beta_1 = 0$ e $\beta_2 = 0$. Embora pareça que podemos rejeitar essa hipótese, uma vez que a estatística *t* para o teste de $\beta_2 = 0$ na Equação (5.18) é $t = 3{,}87/1{,}59 = 2{,}43$, esse raciocínio é falho. A hipótese do contribuinte é uma hipótese conjunta e para testá-la precisamos de uma ferramenta nova, a estatística *F*.

5.7 Testes de Hipóteses Conjuntas

Esta seção descreve como formular hipóteses conjuntas sobre coeficientes de regressão múltipla e como testá-las utilizando uma estatística *F*.

Teste de Hipóteses sobre Dois ou Mais Coeficientes

Hipótese nula conjunta. Considere a regressão na Equação (5.18) da pontuação nos exames contra a razão aluno-professor, o gasto por aluno e a porcentagem de alunos que está aprendendo inglês. A hipótese de nosso contribuinte irritado é de que nem a razão aluno-professor nem o gasto por aluno possuem nenhum efeito sobre a pontuação nos exames, uma vez que controlemos a porcentagem de alunos que está aprendendo inglês. Como *RAP* é o primeiro regressor na Equação (5.18) e *Gasto* é o segundo, podemos escrever matematicamente essa hipótese como

$$H_0: \beta_1 = 0 \text{ e } \beta_2 = 0 \quad versus \quad H_1: \beta_1 \neq 0 \text{ e/ou } \beta_2 \neq 0 \quad (5.19)$$

A hipótese de que *tanto* o coeficiente da razão aluno-professor (β_1) *quanto* o coeficiente do gasto por aluno (β_2) são iguais a zero é um exemplo de uma hipótese conjunta sobre os coeficientes no modelo de regressão múltipla. Nesse caso, a hipótese nula restringe o valor de dois dos coeficientes, de modo que no que diz respeito à terminologia podemos dizer que a hipótese nula na Equação (5.19) impõe duas **restrições** sobre o modelo de regressão múltipla: que $\beta_1 = 0$ *e* que $\beta_2 = 0$.

Em geral, uma **hipótese conjunta** é uma hipótese que impõe duas ou mais restrições sobre os coeficientes da regressão. Consideramos as hipóteses conjuntas nula e alternativa da seguinte forma:

$$H_0: \beta_j = \beta_{j,0}, \beta_m = \beta_{m,0} \text{ etc., para um total de } q \text{ restrições, } versus$$
$$H_1: \text{uma ou mais das } q \text{ restrições sob } H_0 \text{ não são válidas,} \quad (5.20)$$

onde β_j, β_m etc. referem-se a coeficientes de regressão diferentes e $\beta_{j,0}$, $\beta_{m,0}$ etc. referem-se ao valor desses coeficientes sob a hipótese nula. A hipótese nula na Equação (5.19) é um exemplo da Equação (5.20). Outro exemplo é o seguinte: em uma regressão com $k = 6$ regressores, a hipótese nula é de que os coeficientes do 2º, 4º e 5º regressores são nulos; isto é, $\beta_2 = 0$, $\beta_4 = 0$ e $\beta_5 = 0$, de modo que $q = 3$ restrições. Em geral, sob a hipótese nula H_0, existem q dessas restrições.

Se qualquer uma (ou mais de uma) das igualdades sob a hipótese nula H_0 na Equação (5.20) for falsa, a hipótese nula conjunta em si será falsa. Portanto, a hipótese alternativa é de que pelo menos uma das igualdades na hipótese nula H_0 não é válida.

Por que não posso testar somente um coeficiente individual de cada vez? Embora pareça possível testar uma hipótese conjunta utilizando a estatística *t* usual para testar uma restrição de cada vez, o cálculo a seguir mostra que esse enfoque não é confiável. Suponha que você esteja especificamente interessado em testar a hipótese nula conjunta na Equação (5.18) de que $\beta_1 = 0$ e $\beta_2 = 0$. Seja t_1 a estatística *t* para o teste da hipótese nula de que $\beta_1 = 0$ e seja t_2 a estatística *t* para o teste da hipótese nula de que $\beta_2 = 0$. O que acontece quando você usa o procedimento de teste "uma de cada vez": rejeita a hipótese nula conjunta se t_1 ou t_2 excedem 1,96 em valor absoluto?

Como essa pergunta envolve as duas variáveis aleatórias t_1 e t_2, responder a ela requer a caracterização da distribuição amostral conjunta de t_1 e t_2. Conforme mencionado na Seção 5.5, para amostras grandes, $\hat{\beta}_1$ e $\hat{\beta}_2$ possuem uma distribuição normal conjunta, de modo que, sob a hipótese nula conjunta, as estatísticas t t_1 e t_2 possuem uma distribuição normal bivariada, em que cada estatística *t* possui média igual a zero e variância igual a um.

Em primeiro lugar, considere o caso especial em que as estatísticas *t* são não-correlacionadas e, portanto, independentes. Qual é o tamanho do procedimento de teste "uma de cada vez", isto é, qual é a probabilidade de você rejeitar a hipótese nula quando ela for verdadeira? Mais de 5 por cento! Nesse caso especial, podemos calcular a probabilidade de rejeição desse método com exatidão. A hipótese nula *não* será rejeitada somente se $|t_1| \leq 1{,}96$ e $|t_2| \leq 1{,}96$. Como as estatísticas *t* são independentes, $P(|t_1| \leq 1{,}96 \text{ e } |t_2| \leq 1{,}96) = P(|t_1| \leq 1{,}96) \times P(|t_2| \leq 1{,}96) = 0{,}95^2 = 0{,}9025 = 90{,}25\%$. Desse modo, a probabilidade de rejeição da hipótese nula quando ela for verdadeira será de $1 - 0{,}95^2 = 9{,}75$ por cento. O método "uma de cada vez" rejeita a hipótese nula com muita freqüência, porque dá a você muitas oportunidades: se você deixa de rejeitar utilizando a primeira estatística *t*, pode tentar novamente utilizando a segunda.

Se os regressores são correlacionados, a situação fica ainda mais complicada. O tamanho do procedimento "uma de cada vez" depende do valor da correlação entre os regressores. Como o enfoque de teste "uma de cada vez" tem o tamanho errado — isto é, sua taxa de rejeição sob a hipótese nula não é igual ao nível de significância desejado —, é necessário um novo enfoque.

Um enfoque é modificar o método "uma de cada vez" de modo que sejam utilizados diferentes valores críticos que assegurem que seu tamanho seja igual ao nível de significância. Esse método, denominado método de Bonferroni, é descrito no Apêndice 5.3. A vantagem desse método é que ele se aplica de modo bastante geral. Sua desvantagem é ser pouco eficiente: freqüentemente deixa de rejeitar a hipótese nula quando na verdade a hipótese alternativa é verdadeira.

Felizmente, existe outro enfoque mais eficiente para o teste de hipóteses conjuntas, especialmente quando os regressores são altamente correlacionados. Esse enfoque baseia-se na estatística *F*.

Estatística *F*

A **estatística *F*** é utilizada para testar hipóteses conjuntas sobre coeficientes de regressão. As fórmulas para a estatística *F* foram incorporadas aos pacotes de regressão modernos. Discutiremos primeiro o caso de duas restrições para então voltarmos para o caso geral de q restrições.

Estatística F com q = 2 restrições. Quando a hipótese nula conjunta tem duas restrições — $\beta_1 = 0$ e $\beta_2 = 0$ —, a estatística *F* combina *(as duas estatísticas t)* t_1 e t_2 utilizando a fórmula

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}^2}\right), \tag{5.21}$$

onde $\hat{\rho}_{t_1,t_2}$ é um estimador da correlação entre as duas estatísticas *t*.

Para entender a estatística *F* na Equação (5.21), primeiro suponha que sabemos que as estatísticas *t* são não-correlacionadas, de modo que podemos excluir os termos que envolvem $\hat{\rho}_{t_1,t_2}$. Se for esse o caso, a Equação (5.21) é simplificada e $F = \frac{1}{2}(t_1^2 + t_2^2)$, isto é, a estatística *F* é a média dos quadrados das estatísticas *t*. Sob a hipótese nula, t_1 e t_2 são variáveis aleatórias normais padrão independentes (pois as estatísticas *t* são não-correlacionadas por hipótese), de modo que, sob a hipótese nula, *F* possui uma distribuição $F_{2,\infty}$ (veja a Seção 2.4). Sob a hipótese alternativa de que ou β_1 é diferente de zero ou β_2 é diferente de zero (ou ambos), então ou t_1^2 ou t_2^2 (ou ambos) serão grandes, o que leva o teste a rejeitar a hipótese nula.

Em geral, a estatística *t* é correlacionada, e a fórmula para a estatística *F* na Equação (5.21) ajusta-se a essa correlação. Esse ajuste é feito de modo que, sob a hipótese nula, a estatística *F* possua uma distribuição $F_{2,\infty}$ para amostras grandes, sejam as estatísticas *t* correlacionadas ou não.

Estatística* F *com* q *restrições. A fórmula para a estatística *F* testar as q restrições da hipótese nula conjunta na Equação (5.20) é dada na Seção 16.3. Os pacotes de regressão incorporam essa fórmula, o que facilita o cálculo da estatística *F* na prática.

Sob a hipótese nula, a estatística *F* tem uma distribuição amostral que, para amostras grandes, é dada pela distribuição $F_{q,\infty}$. Isto é, para amostras grandes, sob a hipótese nula,

$$\text{a estatística } F \text{ é distribuída como } F_{q,\infty}. \tag{5.22}$$

Assim, os valores críticos para a estatística *F* podem ser obtidos nas tabelas da distribuição $F_{q,\infty}$ na Tabela 4 do Apêndice para o valor apropriado de q e o nível de significância desejado.

***Calculando o valor* p *utilizando a estatística* F.** O valor *p* da estatística *F* pode ser calculado utilizando-se a aproximação qui-quadrado de sua distribuição para amostras grandes. Seja F^{ef} o valor da estatística *F* efetivamente calculado. Como a estatística *F* tem uma distribuição $F_{q,\infty}$, para amostras grandes, sob a hipótese nula, o valor *p* é

$$\text{valor } p = \mathrm{P}[F_{q,\infty} > F^{ef}]. \tag{5.23}$$

O valor *p* na Equação (5.23) pode ser avaliado pelo uso da tabela da distribuição $F_{q,\infty}$ (ou, de forma alternativa, uma tabela da distribuição χ_q^2, pois uma variável aleatória com distribuição χ_q^2 é q vezes uma variável aleatória com distribuição $F_{q,\infty}$). Alternativamente, o valor *p* pode ser avaliado por meio de um computador, uma vez que as fórmulas para as distribuições acumuladas qui-quadrado e *F* foram incorporadas aos pacotes estatísticos modernos.

Estatística F "global" da regressão. A estatística *F* "global" da regressão testa a hipótese conjunta de que *todos* os coeficientes de declividade são iguais a zero. Isto é, as hipóteses nula e alternativa são

$$H_0: \beta_1 = 0, \beta_2 = 0, \ldots, \beta_k = 0 \textit{ versus } H_1: H_1: \beta_j \neq 0, \text{ pelo menos um } j, j = 1, \ldots, k. \tag{5.24}$$

Sob essa hipótese nula, nenhum dos regressores explica qualquer variação em Y_i, embora o intercepto (que sob a hipótese nula é a média de Y_i) possa ser diferente de zero. A hipótese nula na Equação (5.24) é um caso especial da hipótese nula geral da Equação (5.20), e a estatística *F* global da regressão é a estatística *F* calculada para a hipótese nula na Equação (5.24). Para amostras grandes, a estatística *F* global da regressão tem uma distribuição $F_{k,\infty}$.

Estatística F quando q = 1. Quando $q = 1$, a estatística *F* testa uma única restrição. Então, a hipótese nula conjunta fica reduzida a uma hipótese nula com um único coeficiente de regressão e a estatística *F* é o quadrado da estatística *t*.

Heteroscedasticidade e homoscedasticidade novamente. Na Seção 4.9, você viu que, por razões históricas, os pacotes estatísticos às vezes calculam os erros padrão somente homoscedásticos como opção padrão, de modo que o usuário deve especificar que os erros padrão robustos quanto à heteroscedasticidade devem ser utilizados, no lugar da opção padrão. Esse aviso se aplica também à estatística *F*: para assegurar que você utiliza a estatística *F* robusta quanto à heteroscedasticidade, em alguns pacotes de regressão você deve selecionar a opção "robusto", de modo que sejam utilizadas estimativas robustas da "matriz de co-variância". Se a versão somente homoscedástica da estatística *F* (discutida no Apêndice 5.3) for utilizada, mas os erros forem heteroscedásticos, a estatística *F* não terá a distribuição $F_{q,\infty}$ na Equação (5.22) sob a hipótese nula e levará a inferências estatísticas enganosas.

Aplicação à Pontuação nos Exames e à Razão Aluno-Professor

Agora estamos capacitados para testar a hipótese nula de que os coeficientes *tanto* da razão aluno-professor *quanto* do gasto por aluno são iguais a zero contra a alternativa de que pelo menos um coeficiente é diferente de zero, controlando a porcentagem de alunos que está aprendendo inglês na diretoria.

Para testar essa hipótese, precisamos calcular a estatística *F* do teste em que $\beta_1 = 0$ e $\beta_2 = 0$ utilizando a regressão de *PontExame* sobre *RAP*, *Gasto* e *%AI* apresentada na Equação (5.18). O valor da estatística *F* é 5,43. Sob a hipótese nula, para amostras grandes, a estatística possui uma distribuição $F_{2,\infty}$. O valor crítico de 5 por cento da distribuição $F_{2,\infty}$ é 3,00 (veja a Tabela 4 do Apêndice) e o valor crítico de 1 por cento é 4,61. O valor da estatística *F* calculado a partir dos dados, 5,43, é maior do que 4,61, de modo que a hipótese nula é rejeitada ao nível de 1 por cento. É muito pouco provável que tenhamos selecionado uma amostra que produzisse uma estatística *F* tão grande quanto 5,43 se a hipótese nula fosse de fato verdadeira (o valor *p* é 0,005). Com base na evidência da Equação (5.18) e no que foi resumido na estatística *F*, podemos rejeitar a hipótese do contribuinte de que *nem* a razão aluno-professor *nem* o gasto por aluno tem um efeito sobre a pontuação nos exames (mantendo constante a porcentagem de alunos que está aprendendo inglês).

5.8 Testando Restrições Únicas que Envolvem Coeficientes Múltiplos

Às vezes a teoria econômica sugere uma única restrição que envolve dois ou mais coeficientes de regressão. Por exemplo, a teoria pode sugerir uma hipótese nula da forma $\beta_1 = \beta_2$, isto é, os efeitos do primeiro e do segundo regressor são iguais. Nesse caso, a tarefa é testar essa hipótese nula contra a alternativa de que os dois coeficientes são diferentes, isto é,

$$H_0: \beta_1 = \beta_2 \text{ versus } H_1: \beta_1 \neq \beta_2. \tag{5.25}$$

Essa hipótese nula tem uma restrição única, de modo que $q = 1$, porém a restrição envolve múltiplos coeficientes (β_1 e β_2). Precisamos modificar os métodos apresentados até o momento para testar essa hipótese. Existem dois enfoques; qual deles será mais fácil dependerá de seu pacote.

Enfoque nº 1: Teste a restrição diretamente. Alguns pacotes estatísticos possuem um comando especial destinado a testar restrições como a Equação (5.25), e o resultado é uma estatística F que, como $q = 1$, possui uma distribuição $F_{1,\infty}$ sob a hipótese nula. (Lembre-se da Seção 2.4, em que você viu que o quadrado de uma variável aleatória normal padrão possui uma distribuição $F_{1,\infty}$, de modo que o percentil de 95 por cento da distribuição $F_{1,\infty}$ é $1{,}96^2 = 3{,}84$.)

Enfoque nº 2: Transforme a regressão. Se o seu pacote estatístico não pode testar a restrição diretamente, a hipótese na Equação (5.25) pode ser testada utilizando-se um truque em que a equação da regressão original é reescrita para transformar a restrição na Equação (5.25) em uma restrição sobre um único coeficiente de regressão. Para dar um exemplo concreto, suponha que haja somente dois regressores na regressão, X_{1i} e X_{2i}, de modo que a regressão da população tem a forma

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i. \tag{5.26}$$

Aqui está o truque: ao subtrairmos e adicionarmos $\beta_2 X_{1i}$, temos que $\beta_1 X_{1i} + \beta_2 X_{2i} = \beta_1 X_{1i} - \beta_2 X_{1i} + \beta_2 X_{1i} + \beta_2 X_{2i} = (\beta_1 - \beta_2) X_{1i} + \beta_2 (X_{1i} + X_{2i}) = \gamma_1 X_{1i} + \beta_2 W_i$, onde $\gamma_1 = \beta_1 - \beta_2$ e $W_i = X_{1i} + X_{2i}$. Desse modo, a regressão da população na Equação (5.26) pode ser reescrita como

$$Y_i = \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i. \tag{5.27}$$

Como o coeficiente γ_1 nessa equação é $\gamma_1 = \beta_1 - \beta_2$, sob a hipótese nula, na Equação (5.25), $\gamma_1 = 0$, ao passo que, sob a alternativa, $\gamma_1 \neq 0$. Portanto, ao transformar a Equação (5.26) na Equação (5.27), transformamos uma restrição sobre dois coeficientes de regressão em uma restrição sobre um único coeficiente de regressão.

Como a restrição agora envolve o coeficiente único γ_1, a hipótese nula na Equação (5.25) pode ser testada usando o método da estatística t da Seção 5.6. Na prática, isso é feito construindo-se em primeiro lugar o novo regressor W_i como a soma dos dois regressores originais para que então a regressão de Y_i sobre X_{1i} e W_i seja estimada. Um intervalo de confiança de 95 por cento para a diferença entre os coeficiente $\beta_1 - \beta_2$ pode ser calculado como $\hat{\gamma}_1 \pm 1{,}96 EP(\hat{\gamma}_1)$.

Esse método pode ser estendido para outras restrições sobre equações de regressão utilizando-se o mesmo truque (veja o Exercício 5.8).

Os dois métodos (enfoques nº 1 e nº 2) são equivalentes, no sentido de que a estatística F do primeiro método é igual ao quadrado da estatística t do segundo método.

Extensão para q > 1. Em geral, é possível ter q restrições sob a hipótese nula em que algumas ou todas essas restrições envolvem múltiplos coeficientes. A estatística F da Seção 5.7 pode ser estendida para esse tipo de hipótese conjunta. A estatística F pode ser calculada por qualquer um dos dois métodos que acabamos de discutir para $q = 1$. A melhor forma de fazer isso na prática dependerá do pacote de regressão específico que está sendo utilizado.

5.9 Conjuntos de Confiança para Múltiplos Coeficientes

Nesta seção, explicamos como construir um conjunto de confiança para dois ou mais coeficientes de regressão múltipla. O método é conceitualmente similar ao método da Seção 5.6 para a construção de um conjunto de confiança para um único coeficiente pelo uso da estatística t, exceto pelo fato de que o conjunto de confiança para múltiplos coeficientes baseia-se na estatística F.

Um **conjunto de confiança de 95 por cento** para dois ou mais coeficientes é um conjunto que contém os verdadeiros valores da população desses coeficientes em 95 por cento das amostras selecionadas aleatoriamente. Portanto, um conjunto de confiança é a generalização para dois ou mais coeficientes de um intervalo de confiança para um único coeficiente.

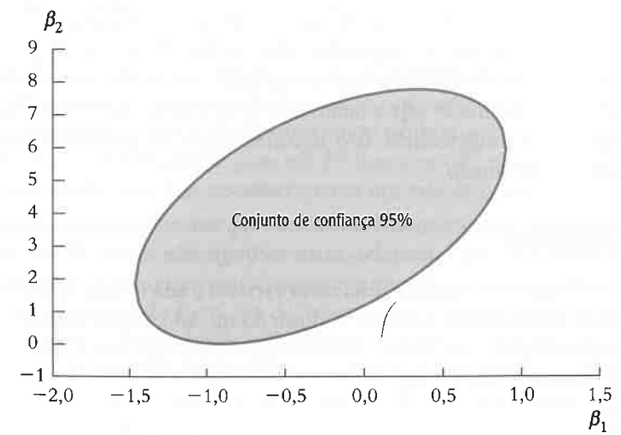
Lembre-se de que um intervalo de confiança de 95 por cento é calculado encontrando-se o conjunto de valores dos coeficientes que não são rejeitados pelo uso de uma estatística t a um nível de significância de 5 por cento. Esse enfoque pode ser estendido para o caso de múltiplos coeficientes. Para tornar isso concreto, suponha que você esteja interessado na construção de um conjunto de confiança para dois coeficientes, β_1 e β_2. Na Seção 5.7, mostramos como utilizar a estatística F para testar uma hipótese nula conjunta de que $\beta_1 = \beta_{1,0}$ e $\beta_2 = \beta_{2,0}$. Suponha que você queira testar todos os valores possíveis de $\beta_{1,0}$ e $\beta_{2,0}$ ao nível de 5 por cento. Para cada par de candidatos ($\beta_{1,0}$, $\beta_{2,0}$), você constrói a estatística F e rejeita-a se ela exceder o valor crítico a 5 por cento de 3,00. Como o teste tem um nível de significância de 5 por cento, os verdadeiros valores da população de β_1 e β_2 não serão rejeitados em 95 por cento de todas as amostras. Assim, o conjunto de valores não rejeitados ao nível de 5 por cento por essa estatística F constitui um conjunto de confiança de 95 por cento para β_1 e β_2.

Embora esse método de tentar todos os valores possíveis de $\beta_{1,0}$ e $\beta_{2,0}$ funcione na teoria, na prática é muito mais simples utilizar uma fórmula explícita para o conjunto de confiança. Essa fórmula para um número arbitrário de coeficientes baseia-se na fórmula para a estatística F da Seção 16.3. Quando há dois coeficientes, os conjuntos de confiança resultantes têm o formato de elipse.

Para fins de ilustração, a Figura 5.1 mostra um conjunto de confiança de 95 por cento (elipse de confiança) para os coeficientes da razão aluno-professor e do gasto por aluno, mantendo constante a porcentagem de alunos que está aprendendo inglês, com base na regressão estimada na Equação (5.18). Essa elipse não inclui o ponto (0,0). Isso significa que a hipótese nula de que esses dois coeficientes são iguais a zero é rejeitada pelo uso da estatística F a um nível de significância de 5 por cento, que já conhecíamos da Seção 5.7. A elipse de confiança é uma lingüiça gorda e sua parte longa é orientada para a direção abaixo-esquerda/acima-direita. Isso porque a correlação estimada entre $\hat{\beta}_1$ e $\hat{\beta}_2$ é positiva, o que por sua vez surge em razão da correlação entre os regressores *RAP* e *Gasto* (escolas que gastam mais por aluno tendem a ter menos alunos por professor).

FIGURA 5.1 Conjunto de Confiança de 95 por cento para β_1 e β_2

O conjunto de confiança de 95 por cento para β_1 e β_2 é uma elipse. A elipse contém os pares de valores de β_1 e β_2 que não podem ser rejeitados pelo uso da estatística F ao nível de significância de 5 por cento.



5.10 Estatísticas de Regressão Adicionais

Três estatísticas-resumo comumente utilizadas na regressão múltipla são o erro padrão da regressão, o R^2 da regressão e o R^2 ajustado (também conhecido como $\overline{R}^2$). As três estatísticas medem quão bem a estimativa de MQO da reta de regressão múltipla descreve ou "se ajusta" aos dados.

Erro Padrão da Regressão (EPR)

O erro padrão da regressão estima o desvio padrão do termo de erro u_i. Portanto, o *EPR* é uma medida da dispersão da distribuição de Y em torno da reta de regressão. Na regressão múltipla, o *EPR* é

$$EPR = s_{\hat{u}}, \text{ onde } s_{\hat{u}}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{SQR}{n-k-1}, \tag{5.28}$$

onde SQR é a soma dos quadrados dos resíduos, $SQR = \sum_{i=1}^{n}\hat{u}_i^2$.

A única diferença entre a definição na Equação (5.28) e a definição de *EPR* na Seção 4.8 para o modelo com um único regressor é que aqui o divisor é $n-k-1$ em vez de $n-2$. Na Seção 4.8, o divisor $n-2$ (em vez de n) ajusta o viés para baixo introduzido na estimativa de dois coeficientes (a declividade e o intercepto da reta de regressão). Aqui o divisor $n-k-1$ ajusta o viés para baixo introduzido na estimativa de $k+1$ coeficientes (k coeficientes de declividade e um intercepto). Assim como na Seção 4.8, o uso de $n-k-1$ em vez de n é chamado de ajuste de graus de liberdade. Se há um único regressor, então $k=1$, de modo que a fórmula na Seção 4.8 é igual à da Equação (5.28). Quando n é grande, o efeito do ajuste de graus de liberdade é desprezível.

O R^2

O R^2 da regressão é a fração da variância da amostra de Y_i explicada (ou prevista) pelos regressores. De modo equivalente, o R^2 é igual a um menos a fração da variância de Y_i *não* explicada pelos regressores.

A definição matemática de R^2 é a mesma da regressão com um único regressor:

$$R^2 = \frac{SQE}{SQT} = 1 - \frac{SQR}{SQT}, \tag{5.29}$$

onde a soma dos quadrados explicada é $SQE = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ e a soma dos quadrados total é $SQT = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$.

Na regressão múltipla, o R^2 aumenta quando um regressor é adicionado, a menos que o novo regressor seja perfeitamente multicolinear com os regressores originais. Para visualizar isso, considere o início com um regressor e então a adição de um segundo. Quando você utiliza MQO para estimar o modelo com os dois regressores, MQO encontra os valores dos coeficientes que minimizam a soma dos quadrados dos resíduos. Se MQO determina que o coeficiente do novo regressor é exatamente zero, a SQR será a mesma se a segunda variável for incluída ou não na regressão. Mas, se MQO escolhe qualquer valor diferente de zero, esse valor deve ter reduzido a SQR relativamente à regressão que exclui esse regressor. Na prática, é bastante incomum que um coeficiente estimado seja exatamente igual a zero, de modo que, em geral, a SQR diminuirá quando um novo regressor for adicionado. Isso significa que o R^2 geralmente aumenta (e nunca diminui) quando um novo regressor é adicionado.

O "R^2 Ajustado"

Como o R^2 aumenta quando uma nova variável é adicionada, um aumento de R^2 não significa que a adição de uma variável efetivamente melhora o ajuste do modelo. Nesse sentido, o R^2 dá uma estimativa inflada de quão bem a regressão se ajusta aos dados. Uma forma de corrigir isso é deflacionar ou reduzir o R^2 em algum fator; é o que o R^2 ajustado, ou $\overline{R}^2$, faz.

O **R^2 ajustado**, ou $\overline{R}^2$, é uma versão modificada do R^2 que não necessariamente aumenta quando um novo regressor é adicionado. O $\overline{R}^2$ é

$$\overline{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SQR}{SQT} = 1 - \frac{s_{\hat{u}}^2}{s_Y^2}. \tag{5.30}$$

A diferença entre essa fórmula e a segunda definição do R^2 na Equação (5.29) é que a razão da soma dos quadrados dos resíduos pela soma dos quadrados total é multiplicada pelo fator $(n-1)/(n-k-1)$. Como mostra a segunda expressão da Equação (5.30), isso torna o R^2 ajustado igual a um menos a razão da variância da amostra dos resíduos de MQO (com a correção de graus de liberdade da Equação (5.28)) pela variância da amostra de Y.

Há três coisas úteis para saber sobre o $\overline{R}^2$. Em primeiro lugar, $(n-1)/(n-k-1)$ é sempre maior do que um, de modo que $\overline{R}^2$ é sempre menor do que R^2.

Em segundo lugar, a adição de um regressor tem dois efeitos opostos sobre o $\overline{R}^2$. Por um lado, SQR diminui, o que aumenta o $\overline{R}^2$. Por outro lado, o fator $(n-1)/(n-k-1)$ aumenta. O aumento ou a diminuição de $\overline{R}^2$ depende de qual desses efeitos é o mais forte.

Por último, o $\overline{R}^2$ pode ser negativo. Isso acontece quando os regressores, tomados em conjunto, reduzem a soma dos quadrados dos resíduos em um montante tão pequeno que essa redução não consegue compensar o fator $(n-1)/(n-k-1)$.

Interpretando o R^2 e o R^2 Ajustado na Prática

Um R^2 ou um $\overline{R}^2$ próximos de um significa que os regressores são bons em prever os valores da variável dependente na amostra, e um R^2 ou um $\overline{R}^2$ próximos de zero significa que os regressores não são bons. Isso faz com que essas estatísticas sejam resumos úteis da capacidade de previsão da regressão. Contudo, podemos ser levados a dar mais crédito a elas do que mereceriam.

Há quatro armadilhas potenciais das quais temos de nos proteger quando utilizamos o R^2 ou o $\overline{R}^2$:

1. ***Um aumento no R^2 ou no $\overline{R}^2$ não significa necessariamente que uma variável adicionada seja estatisticamente significante.*** O R^2 aumenta quando você adiciona um regressor, seja ele estatisticamente significante ou não. O $\overline{R}^2$ nem sempre aumenta, mas se isso acontece não significa necessariamente que o coeficiente do regressor adicionado seja estatisticamente significante. Para ter certeza de que uma variável adicionada é estatisticamente significante, você deve realizar um teste de hipótese utilizando a estatística t.
2. ***Um R^2 alto ou um $\overline{R}^2$ alto não significa que os regressores sejam a causa verdadeira da variável dependente.*** Imagine uma regressão da pontuação nos exames contra a área de estacionamento por aluno. A área de estacionamento está correlacionada com a razão aluno-professor, com a localização da escola (subúrbio ou centro) e, possivelmente, com a renda da diretoria — fatores que estão correlacionados com a pontuação nos exames. Portanto, a regressão da pontuação nos exames sobre a área de estacionamento por aluno poderia ter um R^2 alto ou um $\overline{R}^2$ alto, embora a relação não seja causal (tente dizer para a superintendente que a forma de aumentar a pontuação nos exames é aumentar a área de estacionamento!).
3. ***Um R^2 alto ou um $\overline{R}^2$ alto não significa que não exista viés de omissão de variáveis.*** Lembre-se da discussão da Seção 5.1 sobre o viés de omissão de variáveis na regressão da pontuação nos exames sobre a razão aluno-professor. O R^2 da regressão nunca foi mencionado porque não desempenhou nenhum papel lógico nessa discussão. O viés de omissão de variáveis pode ocorrer em regressões com um R^2 baixo, um R^2 moderado ou um R^2 alto. De maneira inversa, um R^2 baixo não implica que haja necessariamente um viés de omissão de variáveis.
4. ***Um R^2 alto ou um $\overline{R}^2$ alto não significa necessariamente que você tenha o conjunto mais apropriado de regressores, assim como um R^2 baixo ou um $\overline{R}^2$ baixo não significa necessariamente que você tenha um conjunto não apropriado de regressores.*** A pergunta sobre o que constitui o conjunto certo de regressores na regressão múltipla é difícil e voltaremos a ela ao longo do livro. As decisões sobre regressores devem ponderar as questões sobre viés de omissão de variáveis, a disponibilidade dos dados, a qualidade dos dados e, mais importante, a teoria econômica e a natureza das questões importantes a serem tratadas. Nenhuma dessas questões pode ser respondida simplesmente por ter um R^2 alto (ou baixo) ou um $\overline{R}^2$ alto (ou baixo).

Esses pontos estão resumidos no Conceito-Chave 5.8.

Conceito-Chave 5.8

R^2 e $\overline{R}^2$: O que Eles Dizem a Você – e o que Não Dizem

O R^2 e o $\overline{R}^2$ dizem a você se os regressores são bons para prever, ou "explicar", os valores da variável dependente na amostra de dados em uso. Se o R^2 (ou $\overline{R}^2$) for próximo de um, os regressores produzirão boas previsões da variável dependente nessa amostra, no sentido de que a variância do resíduo de MQO é pequena se comparada com a variância da variável independente. Se o R^2 (ou $\overline{R}^2$) for próximo de zero, o oposto será verdadeiro.

O R^2 e o $\overline{R}^2$ NÃO dizem a você se:

1. uma variável incluída é estatisticamente significante;
2. os regressores são a causa verdadeira dos movimentos na variável dependente;
3. há um viés de omissão de variáveis; ou
4. você escolheu o conjunto mais apropriado de regressores.

5.11 Viés de Omissão de Variáveis e Regressão Múltipla

Os estimadores de MQO dos coeficientes na regressão múltipla apresentarão viés de omissão de variáveis se um determinante omitido de Y_i for correlacionado com pelo menos um dos regressores. Por exemplo, alunos de famílias ricas freqüentemente dispõem de mais oportunidades de aprendizado do que seus colegas de famílias menos ricas, o que poderia levá-los a melhores pontuações nos exames. Além disso, se a diretoria é rica, as escolas tendem a ter orçamentos maiores e razões aluno-professor menores. Se for esse o caso, a riqueza dos alunos e a razão aluno-professor estarão negativamente correlacionadas e a estimativa de MQO do coeficiente da razão aluno-professor captará o efeito da renda média da diretoria, mesmo após o controle da porcentagem de alunos que está aprendendo inglês. Em suma, a omissão da situação econômica dos alunos poderia levar a um viés de omissão de variáveis na regressão da pontuação nos exames sobre a razão aluno-professor e a porcentagem de alunos que está aprendendo inglês.

As condições gerais para o viés de omissão de variáveis na regressão múltipla são semelhantes àquelas com um único regressor: se uma variável omitida é um determinante de Y_i e é correlacionada com pelo menos um dos regressores, os estimadores de MQO têm um viés de omissão de variáveis. Como foi discutido na Seção 5.6, os estimadores de MQO são correlacionados, de modo que em geral os estimadores de MQO para todos os coeficientes são viesados. As duas condições para o viés de omissão de variáveis na regressão múltipla estão resumidas no Conceito-Chave 5.9.

Em termos matemáticos, se as duas condições para o viés de omissão de variáveis são satisfeitas, pelo menos um dos regressores está correlacionado com o termo de erro. Isso significa que a expectativa condicional de u_i dados X_{1i}, ..., X_{ki} é diferente de zero, de modo que a primeira hipótese de mínimos quadrados é violada. Como resultado, o viés de omissão de variáveis persiste mesmo que o tamanho da amostra seja grande, isto é, o viés de omissão de variáveis sugere que os estimadores de MQO são inconsistentes.

Especificação de Modelos na Teoria e na Prática

Na teoria, quando há dados disponíveis sobre a variável omitida, a solução para o viés de omissão de variáveis é incluir a variável omitida na regressão. Na prática, contudo, a decisão sobre a inclusão ou não de uma variável em particular pode ser difícil e requer um julgamento.

Conceito-Chave 5.9

Viés de Omissão de Variáveis na Regressão Múltipla

Viés de omissão de variáveis é o viés no estimador de MQO que surge quando um ou mais regressores incluídos estão correlacionados com uma variável omitida. Duas condições devem ser verdadeiras para que surja um viés de omissão de variáveis:

1. pelo menos um dos regressores incluídos deve estar correlacionado com a variável omitida; e
2. a variável omitida deve ser um determinante da variável dependente, Y.

Nosso enfoque ao desafio do viés potencial da variável omitida é duplo. Primeiro, um conjunto fundamental ou básico de regressores deveria ser escolhido utilizando-se uma combinação de julgamento cuidadoso, teoria econômica e conhecimento sobre como os dados foram coletados; a regressão que utiliza esse conjunto básico de regressores às vezes é identificada como uma **especificação de base**. Essa especificação deveria conter as variáveis de maior interesse e as variáveis de controle sugeridas pelo julgamento cuidadoso e pela teoria econômica. Contudo, um julgamento cuidadoso e a teoria econômica raramente são decisivos e com freqüência as variáveis sugeridas pela teoria econômica não são aquelas sobre as quais você possui dados. Portanto, o passo seguinte é desenvolver uma lista de **especificações alternativas** concorrentes, isto é, conjuntos de regressores alternativos. Se as estimativas dos coeficientes de interesse são numericamente semelhantes entre as especificações alternativas, isso fornece evidência de que as estimativas de sua especificação de base são confiáveis. Se, por outro lado, as estimativas dos coeficientes de interesse variam substancialmente entre as especificações, isso freqüentemente fornece evidência de que a especificação original apresentava um viés de omissão de variáveis. Detalharemos esse enfoque para especificação de modelos na Seção 7.2 após estudar algumas ferramentas para especificar regressões.

5.12 Análise do Conjunto de Dados de Pontuação nos Exames

Nesta seção, apresentamos uma análise do efeito da razão aluno-professor sobre a pontuação nos exames utilizando a base de dados sobre a Califórnia. Nosso objetivo principal é fornecer um exemplo em que a análise de regressão múltipla é utilizada para atenuar o viés de omissão de variáveis. Nosso objetivo secundário é demonstrar como utilizar uma tabela para resumir os resultados da regressão.

Essa análise se concentra na estimativa do efeito de uma variação na razão aluno-professor sobre a pontuação nos exames, mantendo constantes as características dos alunos que a superintendente não pode controlar. Anteriormente, estimamos regressões que incluíam tanto a razão aluno-professor quanto o gasto por aluno. Naquela regressão, coeficiente da razão aluno-professor era o efeito de uma variação na razão aluno-professor, mantendo constante o gasto por aluno; nossas estimativas sugeriram que esse efeito é pequeno e não é, em termos estatísticos, significativamente diferente de zero. As regressões relatadas aqui não incluem o gasto por aluno, de modo que o efeito estimado da razão aluno-professor *não* mantém o gasto por aluno constante.

Muitos fatores influenciam potencialmente a pontuação média nos exames em uma diretoria regional de ensino. Alguns fatores que poderiam influenciar a pontuação nos exames estão correlacionados com a razão aluno-professor, de modo que a omissão deles da regressão resultaria em um viés de omissão de variáveis. Se há dados disponíveis sobre as variáveis omitidas, a solução para esse problema é a inclusão dessas variáveis como regressores adicionais na regressão múltipla. Quando fazemos isso, o coeficiente da razão aluno-professor é o efeito de uma variação na razão aluno-professor, mantendo constantes os demais fatores.

Aqui consideramos três variáveis que controlam as características subjacentes dos alunos que poderiam influenciar a pontuação nos exames. Uma dessas variáveis de controle, utilizada anteriormente, é a fração de alunos que ainda está aprendendo inglês. As duas outras variáveis são novas e controlam a situação econômica dos alunos.

Como não existe uma medida perfeita da situação econômica na base de dados, utilizamos dois indicadores imperfeitos de baixa renda diretoria de ensino. A primeira variável nova é a porcentagem de alunos com direito a almoço subsidiado ou gratuito na escola. Os alunos têm direito a esse programa se a sua renda familiar está abaixo de um determinado patamar (aproximadamente 150 por cento da linha de pobreza). A segunda variável nova é a porcentagem de alunos da diretoria cujas famílias têm direito a um programa de auxílio à renda da Califórnia. O direito das famílias a esse programa depende em parte da renda familiar, porém o patamar estabelecido é mais baixo (mais estrito) do que o patamar para o programa de almoço subsidiado. Essas duas variáveis medem, portanto, a fração de crianças com dificuldades financeiras da diretoria; e, embora estejam relacionadas, elas não são perfeitamente correlacionadas (seu coeficiente de correlação é 0,74). Embora a teoria sugira que a situação econômica possa ser um fator omitido importante, a teoria e o julgamento cuidadoso realmente não nos ajudam a decidir qual dessas duas variáveis (porcentagem que tem direito a um almoço subsidiado ou porcentagem que tem direito a um auxílio à renda) é uma medida melhor da situação econômica. Para nossa especificação de base, escolhemos a porcentagem com direito a um almoço subsidiado como a variável da situação econômica, mas consideramos uma especificação alternativa que inclui também a outra variável.

A Figura 5.2 mostra gráficos de dispersão da pontuação nos exames e essas variáveis. Cada uma das variáveis exibe uma correlação negativa com a pontuação nos exames. A correlação entre a pontuação nos exames e a porcentagem de alunos que está aprendendo inglês é –0,64; entre a pontuação nos exames e a porcentagem com direito a um almoço subsidiado, a correlação é de –0,87; e entre a pontuação nos exames e a porcentagem com direito a um auxílio à renda ela é de –0,63.

Agora contemplamos um problema de comunicação. Qual é a melhor maneira de mostrar os resultados de diversas regressões múltiplas que contêm subconjuntos diferentes de regressores possíveis? Até aqui, apresentamos os resultados de uma regressão escrevendo as equações da regressão estimada, como na Equação (5.18). Isso é adequado quando há somente alguns poucos regressores e algumas poucas equações, porém esse método de apresentação pode ser confuso para um número maior de regressores e equações. Uma maneira mais eficiente de comunicar os resultados de várias regressões é por meio de uma tabela.

A Tabela 5.2 resume os resultados de regressões da pontuação nos exames sobre vários conjuntos de regressores. Cada coluna resume uma regressão. E cada uma tem a mesma variável dependente, pontuação nos exames. Nas primeiras cinco linhas, as entradas são os coeficientes de regressão estimados, com seus respectivos erros padrão abaixo deles, entre parênteses. Os asteriscos indicam se a estatística t, que testa a hipótese de que o coeficiente relevante é zero, é significante ao nível de 5 por cento (1 asterisco) ou ao nível de 1 por cento (2 asteriscos). As três últimas linhas contêm estatísticas-resumo da regressão (erro padrão da regressão, EPR e o R^2 ajustado, $\overline{R}^2$) e o tamanho da amostra (que é o mesmo para todas as regressões, 420 observações).

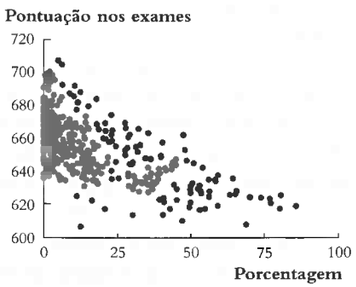
Todas as informações que apresentamos até aqui na forma de equações são mencionadas em uma coluna dessa tabela. Por exemplo, considere a regressão da pontuação nos exames contra a razão aluno-professor, sem variáveis de controle. Na forma de equação, essa regressão é

$$\widehat{PontExame} = \underset{(10,4)}{698,9} - \underset{(0,52)}{2,28} \times RAP,\ \overline{R}^2 = 0,049,\ EPR = 19,26,\ n = 420. \quad (5.31)$$

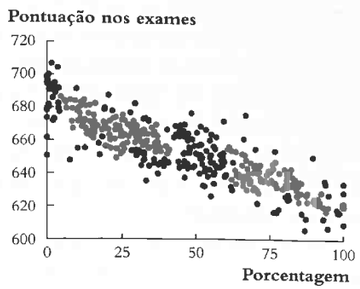
Toda essa informação está na coluna (1) da Tabela 5.2. O coeficiente estimado da razão aluno-professor (–2,28) está na primeira linha de entradas numéricas, e seu erro padrão (0,52) está entre parênteses logo abaixo do coeficiente estimado. O intercepto (698,9) e seu erro padrão (10,4) estão na linha chamada de "Intercepto". (Algumas vezes essa linha é chamada de "constante", uma vez que, conforme discutido na Seção 5.2, o intercepto pode ser visto como o coeficiente de um regressor que é sempre igual a um.) Da mesma forma, o $\overline{R}^2$ (0,049), o EPR (18,58) e o tamanho da amostra n (420) estão nas últimas linhas. As entradas em branco nas linhas dos outros regressores indicam que eles não estão incluídos nessa regressão.

FIGURA 5.2 Gráficos de Dispersão de Pontuação nos Exames *versus* Três Características de Alunos

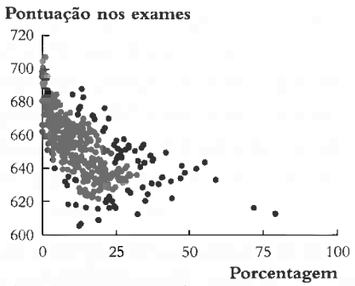
Os gráficos de dispersão mostram uma relação negativa entre a pontuação nos exames e (a) a porcentagem de alunos que está aprendendo inglês (correlação = –0,64); (b) a porcentagem de alunos com direito a almoço subsidiado (correlação = –0,87) e (c) a porcentagem com direito a auxílio à renda (correlação = –0,63).



(a) Porcentagem de alunos que está aprendendo inglês



(b) Porcentagem com direito a almoço subsidiado



(c) Porcentagem com direito a auxílio à renda

TABELA 5.2 Resultados de Regressões da Pontuação nos Exames sobre a Razão Aluno-Professor e sobre as Variáveis de Controle de Características de Alunos com a Utilização de Dados do Ensino Fundamental das Diretorias Regionais de Ensino da Califórnia

Variável dependente: pontuação média nos exames na diretoria					
Regressor	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
Razão aluno-professor (X_1)	–2,28** (0,52)	–1,10* (0,43)	–1,00** (0,27)	–1,31** (0,34)	–1,01** (0,27)
Porcentagem de alunos aprendendo inglês (X_2)		–0,650** (0,031)	–0,122** (0,033)	–0,488** (0,030)	–0,130** (0,036)
Porcentagem com direito a almoço subsidiado (X_3)			–0,547** (0,024)		–0,529** (0,038)
Porcentagem com direito a auxílio à renda (X_4)				–0,790** (0,068)	0,048 (0,059)
Intercepto	698,9** (10,4)	686,0** (8,7)	700,2** (5,6)	698,0** (6,9)	700,4** (5,5)
Estatísticas-resumo					
EPR	18,58	14,46	9,08	11,65	9,08
$\overline{R}^2$	0,049	0,424	0,773	0,626	0,773
n	420,0	420,0	420,0	420,0	420,0

Essas regressões foram estimadas com base nos dados sobre diretorias regionais de ensino K–8 na Califórnia, descritos no Apêndice 4.1. Os erros padrão aparecem entre parênteses abaixo dos coeficientes. O coeficiente individual é estatisticamente significante ao nível de *5 por cento ou de **1 por cento utilizando um teste bicaudal.

Embora a tabela não apresente as estatísticas *t*, estas podem ser calculadas a partir das informações fornecidas; por exemplo, a estatística *t* que testa a hipótese de que o coeficiente da razão aluno-professor na coluna (1) é zero é de 2,28/0,52 = –4,38. Essa hipótese é rejeitada ao nível de 1 por cento, que está indicado pelo asterisco duplo próximo do coeficiente estimado na tabela.

As colunas (2)-(5) mostram as regressões que incluem as variáveis de controle medindo as características dos alunos. A coluna (2), que mostra a regressão da pontuação nos exames sobre a razão aluno-professor e sobre a porcentagem de alunos que está aprendendo inglês, foi expressa anteriormente na Equação (5.16).

A coluna (3) apresenta a especificação de base, em que os regressores são a razão aluno-professor e duas variáveis de controle, a porcentagem de alunos que está aprendendo inglês e a porcentagem de alunos com direito a almoço subsidiado.

As colunas (4) e (5) mostram especificações alternativas que examinam o efeito de mudanças na forma como a situação econômica dos alunos é medida. Na coluna (4), a porcentagem de alunos com direito a auxílio à renda é incluída como regressor e na coluna (5) foram incluídas ambas as variáveis de situação econômica.

Os resultados sugerem três conclusões:

1. O controle das características dos alunos reduz o efeito da razão aluno-professor sobre a pontuação nos exames aproximadamente pela metade. Esse efeito estimado não é muito sensível às variáveis de controle específicas que são incluídas na regressão. Em todos os casos, o coeficiente da razão aluno-professor mantém-se estatisticamente significante ao nível de 5 por cento. Nas quatro especificações com variáveis de controle, as regressões (2)-(5), estima-se que a redução da razão aluno-professor em um aluno por professor aumenta a pontuação média nos exames em aproximadamente um ponto, mantendo constantes as características dos alunos.
2. As variáveis de características dos alunos são previsores muito úteis da pontuação nos exames. A razão aluno-professor sozinha explica apenas uma pequena fração da variação na pontuação nos exames: o $\overline{R}^2$ na coluna (1) é 0,049. Contudo, o $\overline{R}^2$ dá um salto quando as variáveis de características dos alunos são adicionadas. Por exemplo, o $\overline{R}^2$ na especificação de base, regressão (3), é 0,773. O sinal dos coeficientes das variáveis demográficas dos alunos é consistente com os padrões observados na Figura 5.2: diretorias com muitos alunos aprendendo inglês e diretorias com muitas crianças pobres apresentam pontuações menores nos exames.
3. As variáveis de controle nem sempre são estatisticamente significantes em termos individuais: na especificação (5), a hipótese de que o coeficiente da porcentagem com direito a auxílio à renda é zero e não é rejeitada ao nível de 5 por cento (a estatística *t* é –0,82). Como a adição dessa variável de controle à especificação de base (3) possui um efeito desprezível sobre o coeficiente estimado e seu erro padrão e como o coeficiente dessa variável de controle não é significativo na especificação (5), essa variável de controle adicional é redundante, pelo menos para esta análise.

5.13 Conclusão

Este capítulo começou com uma preocupação: na regressão da pontuação nos exames contra a razão aluno-professor, as características omitidas dos alunos que influenciam essa pontuação podem estar correlacionadas com a razão aluno-professor na diretoria e, se fosse esse o caso, a razão aluno-professor na diretoria captaria o efeito das características omitidas de alunos sobre a pontuação nos exames. Desse modo, o estimador de MQO teria um viés de omissão de variáveis. Para atenuarmos esse viés potencial da variável omitida, podemos aumentar a regressão incluindo variáveis que controlam várias características dos alunos (a porcentagem de alunos que está aprendendo inglês e duas medidas da situação econômica do aluno). Fazendo isso, o efeito estimado de uma variação unitária na razão aluno-professor é diminuído pela metade, embora continue sendo possível rejeitar a hipótese nula de que o efeito da pontuação nos exames na população — mantendo constantes essas variáveis de controle — é zero ao nível de significância de 5 por cento. Como eliminam o viés de omissão de variáveis

proveniente das características dos alunos, as estimativas dessas regressões múltiplas (e os intervalos de confiança associados a elas) são muito mais úteis para aconselhar a superintendente do que as estimativas com um único regressor do Capítulo 4.

A análise deste capítulo supôs que a função de regressão da população é linear sobre os regressores, isto é, que a expectativa condicional de Y_i dados os regressores é uma linha reta. Não há, contudo, nenhum motivo em particular para se pensar assim. Na verdade, o efeito de uma redução na razão aluno-professor pode ser completamente diferente nas diretorias com turmas grandes em relação às diretorias que já possuem turmas pequenas. Se for esse o caso, a reta de regressão da população não é linear nos *X*s, e sim uma função não-linear dos *X*s. Contudo, para estender nossa análise a funções de regressão que são não-lineares nos *X*s, precisamos das ferramentas que serão desenvolvidas no próximo capítulo.

Resumo

1. O viés de omissão de variáveis ocorre quando uma variável omitida (1) está correlacionada com um regressor incluído e (2) é um determinante de *Y*.
2. O modelo de regressão múltipla é um modelo de regressão linear que inclui múltiplos regressores, X_1, X_2, ..., X_k. Associado a cada regressor há um coeficiente de regressão, β_1, β_2, ..., β_k. O coeficiente β_1 é a variação esperada em *Y* associada a uma variação unitária em X_1, mantendo constantes os demais regressores. Os outros coeficientes da regressão possuem uma interpretação análoga.
3. Os coeficientes na regressão múltipla podem ser estimados por MQO. Quando as quatro hipóteses de mínimos quadrados do Conceito-Chave 5.4 são satisfeitas, os estimadores de MQO são não viesados, consistentes e normalmente distribuídos para amostras grandes.
4. Testes de hipótese e intervalos de confiança para um único coeficiente da regressão são implementados essencialmente pela utilização dos mesmos procedimentos utilizados no modelo de regressão linear com uma variável do Capítulo 4. Por exemplo, um intervalo de confiança de 95 por cento para β_1 é dado por $\hat{\beta}_1 \pm 1{,}96EP(\hat{\beta}_1)$.
5. Hipóteses que envolvem mais de uma restrição sobre os coeficientes são chamadas de hipóteses conjuntas. As hipóteses conjuntas podem ser testadas pela utilização de uma estatística *F*.
6. O erro padrão da regressão, o R^2, e o $\overline{R}^2$ são estatísticas-resumo para o modelo de regressão múltipla.

Termos-chave

viés de omissão de variáveis (98)
modelo de regressão múltipla (102)
reta de regressão da população (102)
função de regressão da população (102)
intercepto (102)
coeficiente de X_{1i} (102)
variável de controle (102)
efeito parcial (103)
modelo de regressão múltipla da população (103)
homoscedástico (103)
heteroscedástico (103)
estimadores de MQO de β_0, β_1, ..., β_k (104)
reta de regressão de MQO (104)
valor previsto (104)
resíduo de MQO (104)
multicolinearidade perfeita (107)
multicolinearidade imperfeita (108)
restrições (113)
hipótese conjunta (113)
estatística *F* (114)
conjunto de confiança de 95 por cento (117)
R^2 e R^2 ajustado ($\overline{R}^2$) (118, 119)
especificação de base (121)
especificações alternativas (121)
regra de bolso da estatística *F* (122)

Revisão dos Conceitos

5.1 Uma pesquisadora está interessada no efeito do uso do computador sobre a pontuação nos exames. Empregando dados das diretorias regionais de ensino semelhantes aos que foram utilizados neste capítulo, ela regride a pontuação média nos exames da diretoria sobre o número de computadores por aluno. Você acha que $\hat{\beta}_1$ será um estimador não viesado do efeito de um aumento do número de computadores por aluno sobre a pontuação nos exames? Justifique. Em caso afirmativo, trata-se de um viés para cima ou para baixo? Por quê?

5.2 Uma regressão múltipla inclui dois regressores: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. Qual será a variação esperada de Y se X_1 aumentar 3 unidades e X_2 permanecer constante? Qual será a variação esperada de Y se X_2 diminuir 5 unidades e X_1 permanecer constante? Qual será a variação esperada de Y se X_1 aumentar 3 unidades e X_2 diminuir 5 unidades?

5.3 Explique por que dois regressores perfeitamente multicolineares não podem ser incluídos em uma regressão linear múltipla. Dê dois exemplos de um par de regressores perfeitamente multicolineares.

5.4 Explique como você testaria a hipótese nula de que $\beta_1 = 0$ no modelo de regressão múltipla $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. Explique como testaria a hipótese nula de que $\beta_2 = 0$ e a hipótese conjunta de que $\beta_1 = 0$ e $\beta_2 = 0$. Por que o resultado do teste conjunto não é uma implicação dos resultados dos primeiros dois testes?

5.5 Dê um exemplo de uma regressão que provavelmente teria um valor alto de R^2 mas que produziria estimadores viesados e inconsistentes do(s) coeficiente(s) da regressão. Explique por que o R^2 provavelmente seria alto e por que os estimadores de MQO seriam viesados e inconsistentes.

Exercícios

Os sete primeiros exercícios referem-se à tabela de regressões estimadas a seguir, calculadas utilizando-se dados do CPS (Current Population Survey) de 1998. A base de dados consiste de informações sobre quatro mil trabalhadores empregados em período integral durante o ano inteiro. O grau de instrução mais elevado para cada trabalhador seria o ensino médio completo ou o superior completo. A faixa etária dos trabalhadores é de 25 a 34 anos. A base de dados também contém informações sobre a região do país onde a pessoa viveu, o estado civil e o número de filhos. Sejam para este exercício

SMH = salário médio por hora (em dólares de 1998)
Faculdade = variável binária (1 se curso superior, 0 se ensino médio)
Mulher = variável binária (1 se mulher, 0 se homem)
Idade = idade (em anos)
Nordeste = variável binária (1 se Região = Nordeste, 0 caso contrário)
Centro-Oeste = variável binária (1 se Região = Centro-Oeste, 0 caso contrário)
Sul = variável binária (1 se Região = Sul, 0 caso contrário)
Oeste = variável binária (1 se Região = Oeste, 0 caso contrário)

Resultados de Regressões do Salário Médio por Hora sobre as Variáveis Binárias Sexo e Grau de Instrução e outras Características Utilizando Dados do Current Population Survey de 1998.

Variável dependente: Salário Médio por Hora (SMH).			
Regressor	**(1)**	**(2)**	**(3)**
Faculdade (X_1)	5,46 (0,21)	5,48 (0,21)	5,44 (0,21)
Mulher (X_2)	–2,64 (0,20)	–2,62 (0,20)	–2,62 (0,20)
Idade (X_3)		0,29 (0,04)	0,29 (0,04)
Nordeste (X_4)			0,69 (0,30)
Centro-Oeste (X_5)			0,60 (0,28)
Sul (X_6)			–0,27 (0,26)
Intercepto	12,69 (0,14)	4,40 (1,05)	3,75 (1,06)
Estatísticas-resumo e Testes Conjuntos			
Estatística F para efeitos regionais = 0			6,10
EPR	6,27	6,22	6,21
R^2	0,176	0,190	0,194
$\overline{R}^2$			
n	4.000	4.000	4.000

5.1 Acrescente *(5 por cento) e **(1 por cento) à tabela para indicar a significância estatística dos coeficientes.

5.2 Calcule $\overline{R}^2$ para cada uma das regressões.

5.3 Com base nos resultados da regressão da coluna (1), responda:
***a.** Trabalhadores com curso superior têm salários maiores, em média, do que trabalhadores com apenas ensino médio? Em quanto são maiores? A diferença de salários estimada por essa regressão é estatisticamente significante ao nível de 5 por cento?
b. Os homens têm em média salários maiores do que as mulheres? Em quanto são maiores? A diferença de salários estimada por essa regressão é estatisticamente significante ao nível de 5 por cento?

5.4 Com base nos resultados da regressão da coluna (2), responda:
a. A idade é um determinante importante do salário? Explique.
b. Sally é uma mulher de 29 anos com curso superior. Betsy é uma mulher de 34 anos com curso superior. Faça uma previsão dos salários de Sally e de Betsy e construa um intervalo de confiança de 95 por cento para a diferença esperada entre seus salários.

5.5 Com base nos resultados da regressão da coluna (3), responda:
***a.** Existem diferenças regionais importantes?
b. Por que o regressor *Oeste* foi omitido da regressão? O que aconteceria se ele fosse incluído?

***c.** Juanita é uma mulher do Sul de 28 anos com curso superior. Molly é uma mulher do Oeste de 28 anos com curso superior. Jennifer é uma mulher do Centro-Oeste de 28 anos com curso superior.

ci. Construa um intervalo de confiança de 95 por cento para a diferença entre os salários esperados de Juanita e Molly.

cii. Calcule a diferença esperada entre os salários de Juanita e Jennifer.

ciii. Explique como você construiria um intervalo de confiança de 95 por cento para a diferença entre os salários esperados de Juanita e Jennifer. (*Dica*: O que aconteceria se você incluísse *Oeste* e excluísse *Centro-Oeste* da regressão?)

5.6 A regressão da coluna (2) foi estimada novamente utilizando-se dessa vez dados de 1992. (Quatro mil observações foram selecionadas aleatoriamente do CPS de março de 1993 e convertidas para dólares de 1998 por meio do Índice de Preços ao Consumidor.) Os resultados são

$$\widehat{SMH} = 0{,}77 + 5{,}29\mathit{Faculdade} - 2{,}59\mathit{Mulher} + 0{,}40\mathit{Idade},\ EPR = 5{,}85,\ \overline{R}^2 = 0{,}21$$
$$(0{,}98)\quad (0{,}20)\qquad (0{,}18)\qquad (0{,}03)$$

Comparando essa regressão à regressão para 1998 mostrada na coluna (2), é possível dizer que houve uma variação estatisticamente significante no coeficiente de *Faculdade*?

***5.7** Avalie a seguinte afirmação: "Em todas as regressões, o coeficiente de *Mulher* é negativo, grande e estatisticamente significante. Isso fornece uma evidência estatística forte da discriminação sexual no mercado de trabalho dos Estados Unidos".

5.8 Considere o modelo de regressão $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. Utilize o "Enfoque nº 2" da Seção 5.8 para transformar a regressão de modo que você possa utilizar uma estatística t para testar

a. $\beta_1 = \beta_2$;

b. $\beta_1 + a\beta_2 = 0$, onde a é uma constante;

c. $\beta_1 + \beta_2 = 1$. (*Dica*: Você deve redefinir a variável dependente na regressão.)

5.9 O Apêndice 5.3 fornece duas fórmulas para a regra de bolso da estatística F, as equações (5.38) e (5.39). Mostre que as duas fórmulas são equivalentes.

APÊNDICE 5.1 Derivação da Equação (5.1)

Este apêndice apresenta uma derivação da fórmula para o viés de omissão de variáveis da Equação (5.1). A Equação (4.51) do Apêndice 4.3 afirma que

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}. \tag{5.32}$$

Sob as hipóteses de mínimos quadrados no Conceito-Chave 5.4, $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \xrightarrow{p} \sigma_X^2$ e $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})u_i \xrightarrow{p} \text{cov}(u_i, X_i) = \rho_{X,u}\sigma_u\sigma_X$. A substituição desses limites na Equação (5.32) gera a Equação (5.1).

APÊNDICE 5.2 Distribuição dos Estimadores de MQO Quando Há Dois Regressores e Erros Homoscedásticos

Embora a fórmula geral para a variância dos estimadores de MQO em regressão múltipla seja complicada, se há dois regressores ($k = 2$) e os erros são homoscedásticos, a fórmula se torna simples o suficiente para fornecer alguma percepção da distribuição dos estimadores de MQO.

Como os erros são homoscedásticos, a variância condicional de u_i pode ser escrita como $\text{var}(u_i | X_{1i}, X_{2i}) = \sigma_u^2$. Quando há dois regressores, X_{1i} e X_{2i}, e o termo de erro é homoscedástico, para amostras grandes, a distribuição amostral de $\hat{\beta}_1$ é $N(\beta_1, \sigma_{\hat{\beta}_1}^2)$, onde a variância dessa distribuição, $\sigma_{\hat{\beta}_1}^2$, é

$$\sigma_{\hat{\beta}_1}^2 = \frac{1}{n}\left[\frac{1}{1 - \rho_{X_1,X_2}^2}\right]\frac{\sigma_u^2}{\sigma_{X_1}^2}, \tag{5.33}$$

onde ρ_{X_1,X_2} é a correlação da população entre os dois regressores X_1 e X_2 e $\sigma_{X_1}^2$ é a variância da população de X_1.

A variância de $\sigma_{\hat{\beta}_1}^2$ da distribuição amostral de $\hat{\beta}_1$ depende do quadrado da correlação entre os regressores. Se X_1 e X_2 são altamente correlacionados, seja positiva ou negativamente, ρ_{X_1,X_2}^2 é próximo de um e, portanto, o termo $1 - \rho_{X_1,X_2}^2$ no denominador da Equação (5.33) é pequeno e as variâncias de $\hat{\beta}_1$ e $\hat{\beta}_2$ são maiores do que seriam se ρ_{X_1,X_2} fosse próximo de zero. Isso requer uma interpretação intuitiva. Lembre-se de que o coeficiente de X_1 é o efeito de uma variação unitária no primeiro regressor, mantendo o segundo constante. Se os dois regressores são altamente correlacionados, é difícil estimar o efeito parcial do primeiro regressor, mantendo o segundo constante, uma vez que os dois regressores variam juntos na população.

Por exemplo, suponha que queremos estimar os efeitos separados sobre a pontuação nos exames de um número maior de professores (*RAP* menor), mantendo constante o gasto por aluno, e de ter um gasto maior por aluno, mantendo constante a *RAP*. Como os salários dos professores correspondem a uma parcela muito grande do orçamento de uma escola de ensino fundamental, *RAP* e gasto por aluno têm uma forte correlação negativa (mais professores significa uma *RAP* menor e gastos maiores por aluno). Como essas duas variáveis têm uma forte correlação negativa, seria difícil estimar os efeitos separados com precisão utilizando uma amostra de dados. Isso se reflete matematicamente em uma grande variância de $\hat{\beta}_1$.

Outra característica da distribuição conjunta normal para amostras grandes dos estimadores de MQO é que $\hat{\beta}_1$ e $\hat{\beta}_2$ em geral são correlacionados. Quando os erros são homoscedásticos, a correlação entre os estimadores de MQO $\hat{\beta}_1$ e $\hat{\beta}_2$ é o negativo da correlação entre os dois regressores:

$$\text{corr}(\hat{\beta}_1, \hat{\beta}_2) = -\rho_{X_1,X_2}. \tag{5.34}$$

APÊNDICE 5.3 Duas Outras Formas de Testar Hipóteses Conjuntas

O método da Seção 5.7 é a forma preferida de testar hipóteses conjuntas em regressão múltipla. Contudo, se o autor de um estudo apresentou resultados de regressão mas não testou uma restrição conjunta que lhe interessa e você não dispõe dos dados originais, a estatística F da Seção 5.7 não pode ser calculada.

Este apêndice descreve duas outras formas de testar hipóteses conjuntas que podem ser utilizadas quando você tem apenas uma tabela de resultados de regressão. A primeira, o teste de Bonferroni, é a aplicação de um enfoque de teste bastante geral baseado na desigualdade de Bonferroni. A segunda, uma regra de bolso da estatística F, é um enfoque especial para regressão múltipla que se justifica teoricamente apenas se os erros são homoscedásticos; a regra de bolso da estatística F é a estatística F correspondente à estatística t calculada utilizando-se erros padrão somente homoscedásticos.

O Teste de Bonferroni

O teste de Bonferroni é um teste de hipóteses conjuntas com base nas estatísticas t das hipóteses individuais; isto é, é o teste da estatística t "uma de cada vez" da Seção 5.7 conduzido de maneira apropriada. O **teste de Bonferroni** da hipótese nula conjunta $\beta_1 = \beta_{1,0}$ e $\beta_2 = \beta_{2,0}$ com base no valor crítico $c > 0$ utiliza a seguinte regra:

$$\begin{aligned}&\text{Aceite se } |t_1| \le c \text{ e se } |t_2| \le c\text{; caso contrário, rejeite}\\&\text{(teste de Bonferroni da estatística } t \text{ "uma de cada vez")},\end{aligned} \tag{5.35}$$

em que t_1 e t_2 são as estatísticas t que testam respectivamente as restrições sobre β_1 e β_2.

O truque é escolher o valor crítico c de tal forma que a probabilidade de que o teste "uma de cada vez" rejeite a hipótese nula quando esta for verdadeira não seja maior do que o nível de significância desejado, digamos, 5 por cento. Isso é feito utilizando-se a desigualdade de Bonferroni para escolher o valor crítico c que permita tanto o fato de duas restrições estarem sendo testadas quanto qualquer correlação possível entre t_1 e t_2.

A Desigualdade de Bonferroni

A desigualdade de Bonferroni é um resultado básico da teoria da probabilidade. Sejam A e B eventos. Seja $A \cap B$ o evento "A e B"(a intersecção entre A e B) e seja $A \cup B$ o evento "A ou B ou ambos"(a união de A e B). Então, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Como $P(A \cap B) \ge 0$, segue-se que $P(A \cup B) \le P(A) + P(B)$. Essa desigualdade, por sua vez, implica que $1 - P(A \cup B) \ge 1 - [P(A) + P(B)]$. Sejam A^c e B^c os complementos de A e B, isto é, os eventos "não A" e "não B". Como o complemento de $A \cup B$ é $A^c \cap B^c$, $1 - P(A \cup B) = P(A^c \cap B^c)$, que produz a desigualdade de Bonferroni, a saber, $P(A^c \cap B^c) \ge 1 - [P(A) + P(B)]$.

Agora seja A o evento $|t_1| > c$ e B o evento $|t_2| > c$. Então, a desigualdade $P(A \cup B) \le P(A) + P(B)$ produz

$$P(|t_1| > c \text{ ou } |t_2| > c \text{ ou ambos}) \le P(|t_1| > c) + P(|t_2| > c). \tag{5.36}$$

Testes de Bonferroni

Como o evento "$|t_1| > c$ ou $|t_2| > c$ ou ambos" é a região de rejeição do teste "uma de cada vez", a Equação (5.36) fornece uma forma de escolher o valor crítico c de modo que a estatística t "uma de cada vez" tenha o nível de significância desejado para amostras grandes. Sob a hipótese nula, para amostras grandes, $P(|t_1| > c) = P(|t_2| > c) = P(|Z| > c)$. Assim, a Equação (5.36) implica que, para amostras grandes, a probabilidade de que o teste "uma de cada vez" rejeite sob a hipótese nula é

$$P_{H_0}(\text{o teste "uma de cada vez" rejeite}) \le 2P(|Z| > c). \tag{5.37}$$

A desigualdade na Equação (5.37) fornece uma forma de escolher o valor crítico c de modo que a probabilidade de rejeição sob a hipótese nula seja igual ao nível de significância desejado. O enfoque de Bonferroni pode ser estendido para mais de dois coeficientes; se há q restrições sob a hipótese nula, o fator 2 do lado direito da Equação (5.37) é substituído por q.

A Tabela 5.3 apresenta valores críticos c para o teste "uma de cada vez" de Bonferroni para vários níveis de significância e $q = 2$, 3 e 4. Por exemplo, suponha que o nível de significância desejado seja 5 por cento e $q = 2$. De acordo com a Tabela 5.3, o valor crítico c é 2,241. Esse valor crítico é o percentil 1,25 por cento da distribuição normal padrão, de modo que $P(|Z| > 2{,}241) = 2{,}5$ por cento. Desse modo, a Equação (5.37) nos diz que, para amostras grandes, o teste "uma de cada vez" da Equação (5.35) rejeitará no máximo 5 por cento do tempo sob a hipótese nula.

Os valores críticos da Tabela 5.3 são maiores do que os valores críticos para testar uma restrição única. Por exemplo, para $q = 2$, o teste "uma de cada vez" rejeita se pelo menos uma estatística t exceder 2,241 em valor absoluto. Esse valor crítico é maior do que 1,96 porque corrige apropriadamente o fato de que, ao examinar duas estatísticas t, você tem uma segunda chance de rejeitar a hipótese nula conjunta, conforme discutido na Seção 5.7.

Se as estatísticas t individuais se basearem em erros padrão robustos quanto à heteroscedasticidade, o teste de Bonferroni será válido independentemente da presença de heteroscedasticidade, mas se as estatísticas t se basearem em erros padrão somente homoscedásticos, o teste só será válido na presença de homoscedasticidade.

TABELA 5.3 Valores Críticos de Bonferroni c para o Teste da Estatística t "Uma de Cada Vez" de uma Hipótese Conjunta

	Nível de significância		
Número de restrições (q)	10%	5%	1%
2	1,960	2,241	2,807
3	2,128	2,394	2,935
4	2,241	2,498	3,023

Aplicação para a Pontuação nos Exames

As estatísticas t que testam a hipótese nula conjunta de que os coeficientes verdadeiros sobre pontuação nos exames e gasto por aluno na Equação (5.18) são, respectivamente, $t_1 = -0{,}60$ e $t_2 = 2{,}43$. Ainda que $|t_1| < 2{,}241$, como $|t_2| > 2{,}241$, podemos rejeitar a hipótese nula conjunta ao nível de significância de 5 por cento utilizando o teste de Bonferroni. Contudo, tanto t_1 quanto t_2 são menores do que 2,807 em valor absoluto; desse modo, não podemos rejeitar a hipótese nula conjunta ao nível de significância de 1 por cento utilizando o teste. Mas, se utilizarmos a estatística F da Seção 5.7, seremos capazes de rejeitar essa hipótese ao nível de significância de 1 por cento.

Regra de Bolso da Estatística *F*

A regra de bolso da estatística F é calculada utilizando uma fórmula simples baseada na soma dos quadrados dos resíduos de duas regressões. Na primeira regressão, chamada de **regressão restrita**, a hipótese nula é forçada a ser verdadeira. Quando a hipótese nula é do tipo da Equação (5.20), em que todos os valores da hipótese são iguais a zero, a regressão restrita é aquela em que todos os coeficientes são fixados em zero, isto é, os regressores relevantes são excluídos da regressão. Na segunda regressão, chamada de **regressão irrestrita**, permite-se que a hipótese alternativa seja verdadeira. Se a soma dos quadrados dos resíduos for suficientemente menor na regressão irrestrita do que na regressão restrita, o teste rejeitará a hipótese nula.

A **regra de bolso da estatística *F*** é dada pela fórmula

$$F = \frac{(SQR_{restrito} - SSR_{irrestrito})/q}{SQR_{irrestrito}/(n - k_{irrestrito} - 1)}, \tag{5.38}$$

onde $SQR_{restrito}$ é a soma dos quadrados dos resíduos da regressão restrita, $SQR_{irrestrito}$ é a soma dos quadrados dos resíduos da regressão irrestrita, q é o número das restrições sob a hipótese nula e $k_{irrestrito}$ é o número de regressores na regressão irrestrita. Uma fórmula alternativa equivalente para a regra de bolso da estatística F baseia-se no R^2 das duas regressões:

$$F = \frac{(R^2_{irrestrito} - R^2_{restrito})/q}{(1 - R^2_{irrestrito})/(n - k_{irrestrito} - 1)}. \tag{5.39}$$

Se os erros são homoscedásticos, a diferença entre a regra de bolso da estatística F, calculada por meio da Equação (5.38), e a estatística F utilizada na Seção 5.7 desaparece à medida que o tamanho da amostra n aumenta. Desse modo, se os erros são homoscedásticos, a distribuição amostral da regra de bolso da estatística F sob a hipótese nula é, para amostras grandes, $F_{q,\infty}$.

Essas fórmulas de regras de bolso são fáceis de calcular e requerem uma interpretação intuitiva em termos de quão bem as regressões irrestrita e restrita se ajustam aos dados. Infelizmente, elas são válidas somente se os erros são homoscedásticos. Como a homoscedasticidade é um caso especial com o qual não se pode contar em aplicações com dados econômicos, ou, de forma mais geral, com bases de dados normalmente encontradas nas ciências sociais, a regra de bolso da estatística F não é um substituto satisfatório para a estatística F robusta quanto à heteroscedasticidade da Seção 5.7.

Aplicação para a Pontuação nos Exames e a Razão Aluno-Professor

Para testar a hipótese nula de que os coeficientes da população de *RAP* e *Gasto* são 0, controlando *%AI*, precisamos calcular a *SQR* (ou R^2) para as regressões restrita e irrestrita. A regressão irrestrita tem os regressores *RAP*, *Gasto* e *%AI* e é dada na Equação (5.18); seu R^2 é 0,4366; isto é, $R^2_{irrestrito} = 0,4366$. A regressão restrita impõe a hipótese nula conjunta de que os coeficientes verdadeiros de *RAP* e *Gasto* são iguais a zero; isto é, sob a hipótese nula, *RAP* e *Gasto* não entram na regressão da população, embora *%AI* entre (a hipótese nula não impõe restrição ao coeficiente de *%AI*). A regressão restrita estimada por MQO é

$$\widehat{PontExame} = 664,7 - 0,671 \times \%AI, \; R^2 = 0,4149. \quad (5.40)$$
$$(1,0) \; (0,032)$$

de modo que $R^2_{restrito} = 0,4149$. O número de restrições é $q = 2$, o número de observações é $n = 420$ e o número de regressores na regressão irrestrita é $k = 3$. A regra de bolso da estatística *F*, calculada utilizando-se a Equação (5.39), é

$$F = [(0,4366 - 0,4149)/2]/[(1 - 0,4366)/(420 - 3 - 1)] = 8,01.$$

Como 8,01 excede o valor crítico de 1 por cento de 4,61, a hipótese é rejeitada ao nível de 1 por cento utilizando o enfoque da regra de bolso.

Esse exemplo ilustra as vantagens e desvantagens da regra de bolso da estatística *F*. A vantagem é que ela pode ser calculada com o uso de uma calculadora. A desvantagem é que o valor da regra de bolso da estatística *F* pode ser muito diferente da estatística *F* robusta quanto à heteroscedasticidade utilizada na Seção 5.7: a estatística *F* robusta quanto à heteroscedasticidade que testa essa hipótese conjunta é 5,43, completamente diferente do valor menos confiável da regra de bolso somente homoscedástica de 8,01.

CAPÍTULO 6

Funções de Regressão Não-Lineares

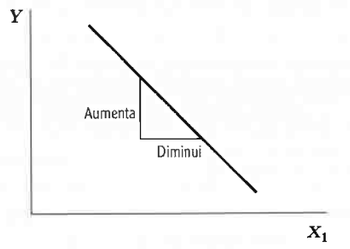
Nos capítulos 4 e 5, supôs-se que a função de regressão da população era linear. Em outras palavras, a declividade da função de regressão da população era constante, de modo que o efeito de uma variação unitária em X sobre Y não dependia em si do valor de X. Mas e se o efeito de uma variação em X sobre Y depender do valor de uma ou mais variáveis independentes? Se for esse o caso, a função de regressão da população é não-linear.

Neste capítulo, desenvolvemos dois grupos de métodos para detectar e modelar funções de regressão da população não-lineares. Os métodos do primeiro grupo são úteis quando o efeito de uma variação em uma variável independente, X_1, sobre Y depender do próprio valor de X_1. Por exemplo, a redução do tamanho das turmas em um aluno por professor pode ter um efeito maior em turmas relativamente pequenas do que em turmas tão grandes que o professor pouco pode fazer além de mantê-las sob controle. Nesse caso, a pontuação nos exames (Y) é uma função não-linear da razão aluno-professor (X_1), em que a função é mais inclinada quando X_1 é pequeno. A Figura 6.1 mostra um exemplo de função de regressão não-linear com essa característica. Enquanto a função de regressão da população linear da Figura 6.1a tem uma declividade constante, a função de regressão da população não-linear da Figura 6.1b apresenta uma declividade mais acentuada quando X_1 é pequeno do que quando é grande. Na Seção 6.2, apresentamos o primeiro grupo de métodos.

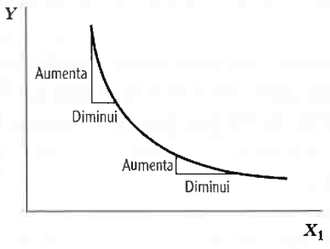
Os métodos do segundo grupo são úteis quando o efeito de uma variação em X_1 sobre Y depende do valor de outra variável independente, digamos X_2. Por exemplo, alunos que ainda estão aprendendo inglês podem se beneficiar em especial de uma atenção mais individualizada; se for esse o caso, o efeito de uma redução da razão aluno-professor sobre a pontuação nos exames será maior em diretorias com muitos alunos que ainda estão aprendendo inglês que em diretorias com poucos alunos aprendendo inglês. Nesse exemplo, o efeito de uma redução

FIGURA 6.1 Funções de Regressão da População com Declividades Diferentes

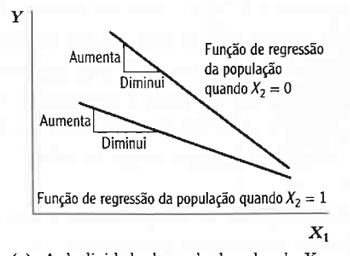
Na Figura 6.1a, a função de regressão da população tem uma declividade constante. Na Figura 6.1b, a declividade da função de regressão da população depende do valor de X_1. Na Figura 6.1c, a declividade da função de regressão da população depende do valor de X_2.



(a) Declividade constante



(b) A declividade depende do valor de X_1



(c) A declividade depende do valor de X_2

na razão aluno-professor (X_1) sobre a pontuação nos exames (Y) depende da porcentagem de alunos que está aprendendo inglês na diretoria (X_2). Como a Figura 6.1c mostra, a declividade desse tipo de função de regressão da população depende do valor de X_2. Na Seção 6.3, apresentamos o segundo grupo de métodos.

Nos modelos deste capítulo, a função de regressão da população é uma função não-linear das variáveis independentes, isto é, a expectativa condicional $E(Y_i \mid X_{1i}, \ldots, X_{ki})$ é uma função não-linear de um ou mais Xs. Embora os modelos sejam não-lineares nos Xs, eles são funções lineares dos coeficientes (ou parâmetros) desconhecidos do modelo de regressão da população e, portanto, são versões do modelo de regressão múltipla do Capítulo 5. Portanto, os coeficientes desconhecidos dessas funções de regressão não-lineares podem ser estimados e testados utilizando MQO e os métodos do Capítulo 5.

Nas seções 6.1 e 6.2, apresentamos as funções de regressão não-lineares no contexto da regressão com uma única variável independente e, na Seção 6.3, fizemos uma extensão para duas variáveis independentes. Para simplificar, as variáveis de controle adicionais são omitidas nos exemplos empíricos das seções 6.1-6.3. Na prática, contudo, é importante analisar as funções de regressão não-lineares em modelos que controlam o viés de omissão de variáveis ao incluir também as variáveis de controle. Na Seção 6.4, combinamos funções de regressão não-lineares e variáveis de controle adicionais quando olhamos de perto possíveis não-linearidades na relação entre a pontuação nos exames e a razão aluno-professor, mantendo constantes as características dos alunos.

6.1 Uma Estratégia Geral para Modelar Funções de Regressão Não-Lineares

Nesta seção, traçamos uma estratégia geral para modelar funções de regressão da população não-lineares. Em tal estratégia, os modelos não-lineares são extensões do modelo de regressão múltipla e, portanto, podem ser estimados e testados utilizando as ferramentas do Capítulo 5. Primeiro, contudo, voltamos aos dados sobre a pontuação nos exames da Califórnia e consideramos a relação entre a pontuação nos exames e a renda na diretoria.

Pontuação nos Exames e Renda na Diretoria

No Capítulo 5, constatamos que a situação econômica dos alunos é um fator importante para explicar o desempenho em exames padronizados. Aquela análise utilizou duas variáveis de situação econômica (a porcentagem de alunos com direito a almoço subsidiado e a porcentagem de famílias com direito a auxílio à renda) para medir a fração de alunos na diretoria proveniente de famílias pobres. Uma medida diferente, mais ampla, de situação econômica é a renda per capita anual média na diretoria regional de ensino ("renda na diretoria"). A base de dados da Califórnia inclui a renda na diretoria medida em milhares de dólares de 1998. A amostra contém uma grande gama de níveis de renda: nas 420 diretorias de nossa amostra, a renda mediana na diretoria é de 13,7 (isto é, US$ 13.700 por pessoa) e varia de 5,3 (US$ 5.300 por pessoa) a 55,3 (US$ 55.300 por pessoa).

A Figura 6.2 mostra um gráfico de dispersão da pontuação nos exames da 5ª série contra a renda na diretoria para a base de dados da Califórnia, juntamente com a reta de regressão de MQO que relaciona essas duas variáveis. A pontuação nos exames e a renda média têm uma forte correlação positiva, com um coeficiente de correlação de 0,71; alunos de diretorias ricas têm desempenho melhor nos exames do que alunos de diretorias pobres. Porém, esse gráfico de dispersão tem uma particularidade: a maioria dos pontos está abaixo da reta de MQO quando a renda é muito baixa (inferior a US$ 10.000) ou muito alta (superior a US$ 40.000), mas está acima da reta quando a renda se encontra entre US$ 15.000 e US$ 30.000. Parece haver alguma curvatura na relação entre pontuação nos exames e renda que não é captada pela regressão linear.

Em suma, parece que a relação entre renda na diretoria e pontuação nos exames não é uma linha reta. Ao contrário, é não-linear. Uma função não-linear é uma função com uma declividade que não é constante: a função $f(X)$ será linear se a declividade de $f(X)$ for a mesma para todos os valores de X, mas, se a declividade depender do valor de X, então $f(X)$ será não-linear.

Se uma linha reta não é uma descrição adequada da relação entre renda na diretoria e pontuação nos exames, o que seria adequado? Imagine o desenho de uma curva que se ajuste aos pontos da Figura 6.2. Essa curva seria inclinada para valores baixos de renda na diretoria e se tornaria cada vez mais plana à medida que a renda na diretoria aumentasse. Uma forma de aproximar essa curva matematicamente é modelar a relação como uma função quadrática. Isto é, podemos modelar a pontuação nos exames como uma função da renda *e* do quadrado da renda.

Um modelo de regressão quadrática da população relacionando a pontuação nos exames e a renda é escrito matematicamente como

$$PontExame_i = \beta_0 + \beta_1 Renda_i + \beta_2 Renda_i^2 + u_i, \tag{6.1}$$

onde β_0, β_1 e β_2 são coeficientes, $Renda_i$ é a renda na i-ésima diretoria, $Renda_i^2$ é o quadrado da renda na i-ésima diretoria e u_i é um termo de erro que, como sempre, representa todos os outros fatores que determinam a pontuação nos exames. A Equação (6.1) é chamada de **modelo de regressão quadrática** porque a função de regressão da população, $E(PontExame_i \mid Renda_i) = \beta_0 + \beta_1 Renda_i + \beta_2 Renda_i^2$, é uma função quadrática da variável independente, *Renda*.

Se você conhecesse os coeficientes da população β_0, β_1 e β_2 na Equação (6.1), poderia prever a pontuação nos exames de uma diretoria com base em sua renda média. Porém, esses coeficientes da população não são conhecidos e, portanto, devem ser estimados utilizando uma amostra de dados.

A princípio, pode parecer difícil encontrar os coeficientes da função quadrática que melhor se ajustem aos dados na Figura 6.2. Entretanto, se você comparar a Equação (6.1) com o modelo de regressão múltipla no Conceito-Chave 5.2, verá que essa equação é, na verdade, uma versão do modelo de regressão múltipla com dois regressores: o primeiro é *Renda* e o segundo é $Renda^2$. Assim, após a definição dos regressores como *Renda* e $Renda^2$, o modelo não-linear na Equação (6.1) é simplesmente um modelo de regressão múltipla com dois regressores!

Como o modelo de regressão quadrática é uma variante da regressão múltipla, seus coeficientes da população desconhecidos podem ser estimados e testados utilizando os métodos de MQO descritos no Capítulo 5. A estimação dos coeficientes da Equação (6.1) por meio de MQO para as 420 observações na Figura 6.2 produz

$$\widehat{PontExame} = \underset{(2,9)}{607,3} + \underset{(0,27)}{3,85}Renda - \underset{(0,0048)}{0,0423}Renda^2, \ \overline{R}^2 = 0,554, \tag{6.2}$$

onde (como sempre) os erros padrão dos coeficientes estimados estão entre parênteses. A Figura 6.3 mostra a função de regressão estimada (6.2) sobreposta ao gráfico de dispersão dos dados. A função quadrática capta a curvatura no gráfico de dispersão: é mais inclinada para valores baixos da renda na diretoria, mas torna-se menos inclinada à medida que a renda aumenta. Em suma, a função de regressão quadrática parece ajustar-se melhor aos dados do que a função linear.

FIGURA 6.2 Gráfico de Dispersão de Pontuação nos Exames *versus* Renda na Diretoria com uma Função de Regressão de MQO Linear

Existe uma correlação positiva entre pontuação nos exames e renda na diretoria (correlação = 0,71), porém a reta de regressão de MQO linear não descreve adequadamente a relação entre essas variáveis.

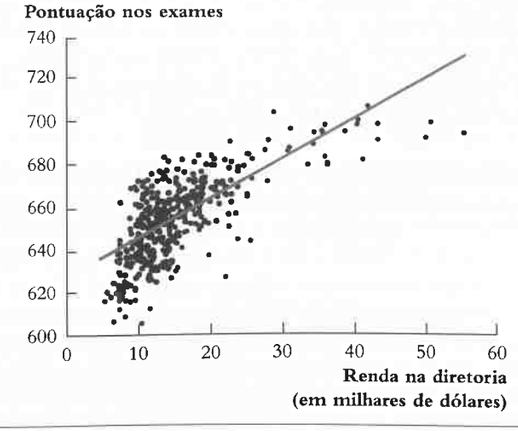
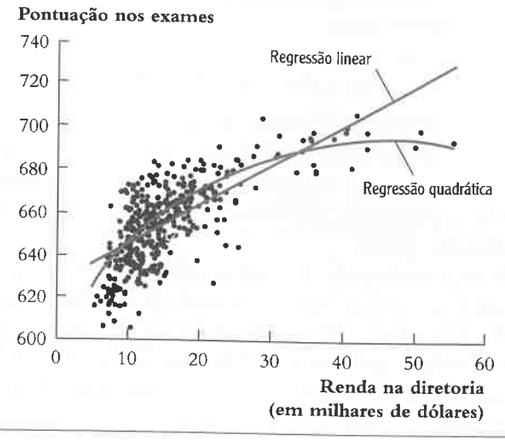


FIGURA 6.3 **Gráfico de Dispersão de Pontuação nos Exames *versus* Renda na Diretoria com Funções de Regressão Linear e Quadrática**

A função de regressão de MQO quadrática ajusta-se melhor aos dados do que a função de regressão de MQO linear.



Podemos ir um passo além dessa comparação visual e testar formalmente a hipótese de que a relação entre renda e pontuação nos exames é linear contra a alternativa de que a relação é não-linear. Se a relação é linear, a função de regressão está especificada corretamente como a Equação (6.1), exceto pelo regressor $Renda^2$, que está ausente; isto é, se a relação é linear, a equação é válida com $\beta_2 = 0$. Desse modo, podemos testar a hipótese nula de que a função de regressão da população é linear contra a alternativa de que ela é quadrática testando a hipótese nula de que $\beta_2 = 0$ contra a alternativa de que $\beta_2 \neq 0$.

Como a Equação (6.1) é somente uma variante do modelo de regressão múltipla, podemos testar a hipótese nula de que $\beta_2 = 0$ ao construirmos a estatística t para essa hipótese. A estatística t é $t = (\hat{\beta}_2 - 0)/EP(\hat{\beta}_2)$, que da Equação (6.2) é $t = -0{,}0423/0{,}0048 = -8{,}81$. Em valor absoluto, isso excede o valor crítico de 5 por cento desse teste (que é 1,96). De fato, o valor p para a estatística t é menor do que 0,01 por cento, de modo que podemos rejeitar a hipótese de que $\beta_2 = 0$ em todos os níveis de significância convencionais. Portanto, esse teste de hipótese formal sustenta nossa inspeção informal das figuras 6.2 e 6.3: o modelo quadrático se ajusta melhor aos dados do que o modelo linear.

Efeito de uma Variação em *X* sobre *Y* em Especificações Não-Lineares

Deixe de lado por um momento o exemplo da pontuação nos exames e considere um problema geral. Você quer saber qual é a variação esperada na variável dependente Y quando a variável independente X_1 varia em um montante ΔX_1, mantendo constantes as outras variáveis independentes X_2, ..., X_k. Quando a função de regressão da população é linear, esse efeito é fácil de calcular: conforme mostrado na Equação (5.4), a variação esperada em Y é $\Delta Y = \beta_1 \Delta X_1$, onde β_1 é o coeficiente da regressão da população multiplicando X_1. Contudo, quando a função de regressão é não-linear, a variação esperada em Y é mais difícil de calcular, uma vez que ela pode depender dos valores das variáveis independentes.

Uma fórmula geral para uma função de regressão da população não-linear.[1] Os modelos de regressão da população não-lineares considerados neste capítulo têm a seguinte forma:

$$Y_i = f(X_{1i}, X_{2i}, ..., X_{ki}) + u_i, \; i = 1, ..., n, \tag{6.3}$$

[1] O termo "regressão não-linear" se aplica a duas famílias de modelos conceitualmente diferentes. Na primeira família, a função de regressão da população é uma função não-linear dos Xs, mas é uma função linear dos parâmetros desconhecidos (os βs). Na segunda família, a função de regressão da população é uma função não-linear dos parâmetros desconhecidos e pode ou não ser uma função não-linear dos Xs. Todos os modelos deste capítulo pertencem à primeira família. Encontraremos modelos da segunda família no Capítulo 9 quando tratarmos da regressão com uma variável binária dependente.

onde $f(X_{1i}, X_{2i}, ..., X_{ki})$ é a **função de regressão não-linear** da população, uma função possivelmente não-linear das variáveis independentes X_{1i}, X_{2i}, ..., X_{ki}, e u_i é o termo de erro. Por exemplo, no modelo de regressão quadrática da Equação (6.1), somente uma variável independente está presente, de modo que X_1 é *Renda* e a função de regressão da população é $f(Renda_i) = \beta_0 + \beta_1 Renda_i + \beta_2 Renda_i^2$.

Como a função de regressão da população é a expectativa condicional de Y_i dados X_{1i}, X_{2i}, ..., X_{ki}, na Equação (6.3) admitimos a possibilidade de que essa expectativa condicional seja uma função não-linear de X_{1i}, X_{2i}, ..., X_{ki}, isto é, $E(Y_i | X_{1i}, X_{2i}, ..., X_{ki}) = f(X_{1i}, X_{2i}, ..., X_{ki})$, onde f pode ser uma função não-linear. Se a função de regressão da população é linear, então $f(X_{1i}, X_{2i}, ..., X_{ki}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$ e a Equação (6.3) torna-se o modelo de regressão linear do Conceito-Chave 5.2. Contudo, essa equação também admite funções de regressão não-lineares.

Efeito de uma variação em X_1 sobre Y. Conforme discutido na Seção 5.2, o efeito de uma variação em X_1, ΔX_1, sobre Y, mantendo constantes X_2, ..., X_k, é a diferença entre o valor esperado de Y quando as variáveis independentes assumem os valores $X_1 + \Delta X_1$, X_2, ..., X_k e o valor esperado de Y quando as variáveis independentes assumem os valores X_1, X_2, ..., X_k. A diferença entre esses dois valores esperados, digamos ΔY, é o que acontece com Y em média na população quando X_1 varia em um montante ΔX_1, mantendo constantes as outras variáveis X_2, ..., X_k. No modelo de regressão não-linear da Equação (6.3), esse efeito sobre Y é $\Delta Y = f(X_1 + \Delta X_1, X_2, ..., X_k) - f(X_1, X_2, ..., X_k)$.

Como a função de regressão f é desconhecida, o efeito da população de uma variação em X_1 sobre Y também é desconhecido. Para estimar o efeito da população, primeiro estime a função de regressão da população. Em um nível geral, represente essa função estimada por $\hat{f}$; um exemplo dessa função estimada é a função de regressão quadrática estimada na Equação (6.2). O efeito estimado de uma variação em X_1 sobre Y (representado por $\Delta\hat{Y}$) é a diferença entre o valor previsto de Y quando as variáveis independentes assumem os valores $X_1 + \Delta X_1$, X_2, ..., X_k e o valor previsto de Y quando as variáveis independentes assumem os valores X_1, X_2, ..., X_k.

O Conceito-Chave 6.1 resume o método para calcular o efeito esperado de uma variação em X_1 sobre Y.

Aplicação ao caso de pontuação nos exames e renda. Qual é a variação prevista da pontuação nos exames associada a uma variação da renda na diretoria de US$ 1.000,00, com base na função de regressão quadrática estimada na Equação (6.2)? Como a função de regressão é quadrática, esse efeito depende da renda inicial na diretoria. Consideremos, portanto, dois casos: um aumento da renda na diretoria de 10 para 11 (isto é, de US$ 10.000 per capita para US$ 11.000) e um aumento da renda na diretoria de 40 para 41.

Para calcularmos $\Delta\hat{Y}$ associado a uma variação da renda de 10 para 11, podemos aplicar a fórmula geral da Equação (6.6) ao modelo de regressão quadrático. Fazendo isso, temos

$$\Delta\hat{Y} = (\hat{\beta}_0 + \hat{\beta}_1 \times 11 + \hat{\beta}_2 \times 11^2) - (\hat{\beta}_0 + \hat{\beta}_1 \times 10 + \hat{\beta}_2 \times 10^2), \tag{6.4}$$

onde $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_2$ são os estimadores de MQO.

O termo no primeiro conjunto de parênteses da Equação (6.4) é o valor previsto de Y quando $Renda = 11$ e o termo no segundo conjunto de parênteses é o valor previsto de Y quando $Renda = 10$. Esses valores são calculados utilizando-se as estimativas de MQO dos coeficientes na Equação (6.2). Dessa forma, quando $Renda = 10$, o valor previsto da pontuação nos exames é $607{,}3 + 3{,}85 \times 10 - 0{,}0423 \times 10^2 = 641{,}57$. Quando $Renda = 11$, o valor previsto é $607{,}3 + 3{,}85 \times 11 - 0{,}0423 \times 11^2 = 644{,}53$. A diferença entre esses dois valores previstos é $\Delta\hat{Y} = 644{,}53 - 641{,}57 = 2{,}96$ pontos, isto é, a diferença prevista para a pontuação nos exames entre uma diretoria com renda média de US$ 11.000 e outra com renda média de US$ 10.000 é de 2,96 pontos.

No segundo caso, quando a renda passa de US$ 40.000 para US$ 41.000, a diferença nos valores previstos da Equação (6.4) é $\Delta\hat{Y} = (607{,}3 + 3{,}85 \times 41 - 0{,}0423 \times 41^2) - (607{,}3 + 3{,}85 \times 40 - 0{,}0423 \times 40^2) = 694{,}04 - 693{,}62 = 0{,}42$ pontos. Assim, uma variação da renda de US$ 1.000 está associada a uma variação maior da pontuação nos exames prevista se a renda inicial for US$ 10.000 do que se ela for US$ 40.000 (as variações previstas são 2,96 pontos e 0,42 pontos, respectivamente). Dito de outra forma, a declividade da função de regressão quadrática estimada da Figura 6.3 é mais acentuada para valores baixos de renda (como US$ 10.000) do que para valores mais altos (como US$ 40.000).

Conceito-Chave 6.1

Efeito Esperado de uma Variação em X_1 sobre Y no Modelo de Regressão Não-Linear

A variação esperada em Y, ΔY, associada a uma variação em X_1, ΔX_1, mantendo X_2, ..., X_k constantes, é a diferença entre o valor da função de regressão da população antes e depois da variação em X_1, mantendo X_2, ..., X_k constantes. Isto é, a variação esperada em Y é a diferença:

$$\Delta Y = f(X_1 + \Delta X_1, X_2, ..., X_k) - f(X_1, X_2, ..., X_k). \quad (6.5)$$

O estimador dessa diferença da população desconhecida é a diferença entre os valores previstos para esses dois casos. Seja $\hat{f}(X_1, X_2, ..., X_k)$ o valor previsto de Y com base no estimador $\hat{f}$ da função de regressão da população. Então a variação prevista em Y é

$$\Delta\hat{Y} = \hat{f}(X_1 + \Delta X_1, X_2, ..., X_k) - \hat{f}(X_1, X_2, ..., X_k). \quad (6.6)$$

Erros padrão dos efeitos estimados. O estimador do efeito de uma variação em X_1 sobre Y depende do estimador da função de regressão da população, $\hat{f}$, que varia de uma amostra para a seguinte. Portanto, o efeito estimado contém erro de amostragem. Uma forma de quantificar a incerteza com relação à amostragem associada ao efeito estimado é calcular um intervalo de confiança para o verdadeiro efeito da população. Para isso, precisamos calcular o erro padrão de $\Delta\hat{Y}$ na Equação (6.6).

É fácil calcular um erro padrão para $\Delta\hat{Y}$ quando a função de regressão é linear. O efeito estimado de uma variação em X_1 é $\hat{\beta}_1\Delta X_1$, de modo que um intervalo de confiança de 95 por cento para a variação estimada é $\hat{\beta}_1\Delta X_1 \pm 1{,}96EP(\hat{\beta}_1)\Delta X_1$.

Nos modelos de regressão não-linear deste capítulo, o erro padrão de $\Delta\hat{Y}$ pode ser calculado utilizando as ferramentas apresentadas na Seção 5.8 para o teste de uma única restrição que envolve múltiplos coeficientes. Para ilustrar esse método, considere a variação estimada da pontuação nos exames associada a uma variação na renda de 10 para 11 na Equação (6.4), que é $\Delta\hat{Y} = \hat{\beta}_1 \times (11 - 10) + \hat{\beta}_2 \times (11^2 - 10^2) = \hat{\beta}_1 + 21\hat{\beta}_2$. Portanto, o erro padrão da variação prevista é

$$EP(\Delta\hat{Y}) = EP(\hat{\beta}_1 + 21\hat{\beta}_2). \quad (6.7)$$

Assim, se conseguirmos calcular o erro padrão de $\hat{\beta}_1 + 21\hat{\beta}_2$, teremos calculado o erro padrão de $\Delta\hat{Y}$. Existem dois métodos para fazer isso utilizando pacotes de regressão padrão, que correspondem aos dois enfoques da Seção 5.8 para o teste de uma única restrição sobre múltiplos coeficientes.[2]

O primeiro método é utilizar o "Enfoque nº 1" da Seção 5.8, que consiste em calcular a estatística F que testa a hipótese de que $\beta_1 + 21\beta_2 = 0$. O erro padrão de $\Delta\hat{Y}$ então é dado por[3]

$$EP(\Delta\hat{Y}) = \frac{|\Delta\hat{Y}|}{\sqrt{F}}. \quad (6.8)$$

Quando aplicada à regressão quadrática na Equação (6.2), a estatística F que testa a hipótese de que $\beta_1 + 21\beta_2 = 0$ é $F = 299{,}94$. Como $\Delta\hat{Y} = 2{,}96$, a aplicação da Equação (6.8) gera $EP(\Delta\hat{Y}) = 2{,}96/\sqrt{299{,}94} = 0{,}17$. Assim, um intervalo de confiança de 95 por cento para a variação no valor esperado de Y é $2{,}96 \pm 1{,}96 \times 0{,}17$ ou (2,63, 3,29).

[2] Esses dois enfoques são formas diferentes de utilizar um pacote de regressão para implementar as fórmulas gerais para os erros padrão dos efeitos previstos apresentadas na Seção 16.2.

[3] A Equação (6.8) é derivada observando-se que a estatística F é o quadrado da estatística t que testa essa hipótese, isto é, $F = t^2 = [(\hat{\beta}_1 + 21\hat{\beta}_2)/EP(\hat{\beta}_1 + 21\hat{\beta}_2)]^2 = [\Delta\hat{Y}/EP(\Delta\hat{Y})]^2$ e resolvendo $EP(\Delta\hat{Y})$.

O segundo método é utilizar o "Enfoque nº 2" da Seção 5.8, que envolve a transformação dos regressores de modo que, na regressão transformada, um dos coeficientes seja $\beta_1 + 21\beta_2$. A demonstração dessa transformação é deixada como um exercício (veja o Exercício 6.4).

Um comentário sobre a interpretação de coeficientes em especificações não-lineares. No modelo de regressão múltipla do Capítulo 5, os coeficientes da regressão tinham uma interpretação natural. Por exemplo, β_1 é a variação esperada em Y associada a uma variação em X_1, mantendo constantes os demais regressores. Mas, como vimos, esse geralmente não é o caso em um modelo não-linear. Isto é, não ajuda muito pensar em β_1 na Equação (6.1) como o efeito de uma variação da renda na diretoria, mantendo constante o quadrado dessa renda. Isso significa que, em modelos não-lineares, a função de regressão é mais bem interpretada quando é mostrada em um gráfico e quando o efeito previsto de uma variação em uma ou mais das variáveis independentes sobre Y é calculado.

Enfoque Geral para a Modelagem de Não-Linearidades Utilizando Regressão Múltipla

O enfoque geral para a modelagem de funções de regressão não-lineares examinado neste capítulo possui cinco elementos:

1. ***Identifique uma possível relação não-linear.*** A melhor coisa a fazer é utilizar a teoria econômica e o que você sabe sobre a aplicação para sugerir uma possível relação não-linear. Antes mesmo de examinar os dados, pergunte a si mesmo se a declividade da função de regressão que relaciona Y e X teria motivos para depender do valor de X ou de outra variável independente. Por que essa dependência não-linear pode existir? Que formas não-lineares ela sugere? Por exemplo, pensar sobre a dinâmica em uma sala de aula com alunos de 11 anos de idade sugere que a redução do tamanho da turma de 18 para 17 alunos poderia ter um efeito maior do que a redução de 30 para 29 alunos.
2. ***Especifique uma função não-linear e estime seus parâmetros por MQO.*** As seções 6.2 e 6.3 contêm várias funções de regressão não-lineares que podem ser estimadas por MQO. Após estudar essas seções, você compreenderá as características de cada uma delas.
3. ***Determine se o modelo não-linear é melhor do que o modelo linear.*** O fato de você pensar que uma função de regressão é não-linear não significa que ela realmente o seja! Você deve determinar empiricamente se o seu modelo não-linear é apropriado. Na maioria das vezes, você pode utilizar a estatística t e a estatística F para testar a hipótese nula de que a função de regressão da população é linear contra a alternativa de que ela é não-linear.
4. ***Desenhe a função de regressão não-linear estimada.*** A função de regressão estimada descreve bem os dados? Um exame das figuras 6.2 e 6.3 sugere que o modelo quadrático se ajusta melhor aos dados do que o modelo linear.
5. ***Estime o efeito de uma variação em X sobre Y.*** O último passo é utilizar a regressão estimada para calcular o efeito de uma variação em um ou mais regressores X sobre Y utilizando o método do Conceito-Chave 6.1.

6.2 Funções Não-Lineares de uma Única Variável Independente

Nesta seção fornecemos dois métodos para modelar uma função de regressão não-linear. Para simplificar, desenvolvemos esses métodos para uma função de regressão não-linear que envolva somente uma variável independente X. Contudo, como veremos na Seção 6.4, esses modelos podem ser modificados para incluir múltiplas variáveis independentes.

O primeiro método discutido nesta seção é a regressão polinomial, uma extensão da regressão quadrática utilizada na seção anterior para modelar a relação entre pontuação nos exames e renda. O segundo método utiliza logaritmos de X e/ou de Y. Embora esses métodos sejam apresentados separadamente, eles podem ser utilizados de forma combinada.

Polinômios

Uma forma de especificar uma função de regressão não-linear é utilizar um polinômio em X. Em geral, represente por r a maior potência de X incluída na regressão. O **modelo de regressão polinomial** de grau r é

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_r X_i^r + u_i. \tag{6.9}$$

Quando $r = 2$, a Equação (6.9) é o modelo de regressão quadrática discutido na Seção 6.1. Quando $r = 3$, de modo que a maior potência de X incluída na regressão é X^3, a Equação (6.9) é chamada de **modelo de regressão cúbica**.

O modelo de regressão polinomial é semelhante ao modelo de regressão múltipla do Capítulo 5, exceto pelo fato de que nesse capítulo os regressores eram variáveis independentes distintas, ao passo que neste caso os regressores são potências da mesma variável independente, X, isto é, os regressores são X, X^2, X^3 etc. Assim, as técnicas para estimação e inferência desenvolvidas para a regressão múltipla podem ser aplicadas aqui. Em particular, os coeficientes desconhecidos β_0, β_1, ..., β_r da Equação (6.9) podem ser estimados pela regressão de MQO de Y_i sobre X_i, X_i^2, ..., X_i^r.

Testando a hipótese nula de que a função de regressão da população é linear. Se a função de regressão da população é linear, os termos quadráticos e de ordem maior não entram na função de regressão da população. Dessa forma, a hipótese nula (H_0) de que a regressão é linear e a alternativa (H_1) de que ela é um polinômio de grau r correspondem a

$$H_0: \beta_2 = 0, \beta_3 = 0, \ldots, \beta_r = 0 \textit{ versus } H_1: \text{pelo menos um } \beta_j \neq 0, j = 2, \ldots, r. \tag{6.10}$$

A hipótese nula de que a função de regressão da população é linear pode ser testada contra a alternativa de que ela é um polinômio de grau r pelo teste de H_0 contra H_1 na Equação (6.10). Como H_0 é uma hipótese nula conjunta com $q = r - 1$ restrições sobre os coeficientes do modelo de regressão polinomial da população, ela pode ser testada utilizando a estatística F descrita na Seção 5.7.

Que grau de polinômio deveria ser usado? Isto é, quantas potências de X deveriam ser incluídas em uma regressão polinomial? A resposta pondera um dilema entre flexibilidade e precisão estatística. O aumento do grau r introduz mais flexibilidade na função de regressão e permite que ela se ajuste a mais formas de curva; um polinômio de grau r pode ter até $r - 1$ pontos de inflexão em seu gráfico. Mas o aumento de r implica a adição de mais regressores, o que pode reduzir a precisão dos coeficientes estimados.

Assim, a resposta para a pergunta feita no parágrafo anterior é a seguinte: você deveria incluir o suficiente para modelar a função de regressão não-linear adequadamente, mas nada mais. Infelizmente, essa resposta não tem muita utilidade na prática!

Uma forma prática de determinar o grau do polinômio é perguntar se os coeficientes da Equação (6.9) associados aos valores maiores de r são iguais a zero. Se for esse o caso, esses termos poderão ser excluídos da regressão. Esse procedimento, chamado de teste seqüencial de hipótese, uma vez que as hipóteses individuais são testadas seqüencialmente, está resumido nos seguintes passos:

1. Escolha um valor máximo para r e estime a regressão polinomial para ele.
2. Utilize a estatística t para testar a hipótese de que o coeficiente de X^r (β_r na Equação (6.9)) é igual a zero. Se você rejeitar essa hipótese, X^r pertencerá à regressão; assim, utilize o polinômio de grau r.

3. Se você não rejeitou $\beta_r = 0$ no passo 2, elimine X^r da regressão e estime uma regressão polinomial de grau $r - 1$. Teste se o coeficiente de X^{r-1} é igual a zero. Se você rejeitou $\beta_r = 0$, utilize o polinômio de grau $r - 1$.
4. Se você não rejeitou $\beta_{r-1} = 0$ no passo 3, continue este procedimento até que o coeficiente da maior potência de seu polinômio seja estatisticamente significante.

Falta um ingrediente nesta receita: o grau inicial r do polinômio. Em muitas aplicações que envolvem dados econômicos, as funções não-lineares são suaves, isto é, não apresentam saltos abruptos ou "picos". Se for esse o caso, é apropriado escolher uma ordem máxima pequena para o polinômio, tal como 2, 3 ou 4; isto é, comece com $r = 2$, 3 ou 4 no passo 1.[4]

Aplicação ao caso de renda na diretoria e pontuação nos exames. A função de regressão cúbica estimada que relaciona renda na diretoria e pontuação nos exames é

$$\begin{aligned} \widehat{PontExame} &= 600{,}1 + 5{,}02 Renda - 0{,}096 Renda^2 + 0{,}00069 Renda^3, \\ &\quad (5{,}1)\ (0{,}71) \qquad (0{,}029) \qquad (0{,}00035) \\ \overline{R}^2 &= 0{,}555. \end{aligned} \tag{6.11}$$

A estatística t de $Renda^3$ é 1,97, de modo que a hipótese nula de que a função de regressão é quadrática é rejeitada contra a alternativa de que é cúbica ao nível de 5 por cento. Além disso, a estatística F que testa a hipótese nula conjunta de que os coeficientes de $Renda^2$ e $Renda^3$ são iguais a zero é 37,7, com um valor p inferior a 0,01 por cento, de modo que a hipótese nula de que a função de regressão é linear é rejeitada contra a alternativa de que ela é cúbica.

Interpretação dos coeficientes nos modelos de regressão polinomial. Os coeficientes nas regressões polinomiais não têm uma interpretação simples. A melhor forma de interpretar essas regressões é desenhar a função de regressão estimada e calcular o efeito estimado associado a uma variação em X sobre Y para um ou mais valores de X.

Logaritmos

Outra forma de especificar uma função de regressão não-linear é utilizar o logaritmo natural de Y e/ou de X. Os logaritmos convertem as variações nas variáveis em variações percentuais, e muitas relações são expressas naturalmente em termos de porcentagem. Aqui estão alguns exemplos:

- Na Seção 3.5, examinamos a diferença de salários entre homens e mulheres com curso superior. Naquela discussão, a diferença de salários foi medida em dólares. Contudo, é mais fácil comparar diferenças de salários entre profissões e ao longo do tempo quando elas são expressas em porcentagem.
- Na Seção 6.1, constatamos que a renda na diretoria e a pontuação nos exames estavam relacionadas de forma não-linear. Essa relação seria linear para variações percentuais? Isto é, uma variação da renda na diretoria de 1 por cento — em vez de US$ 1.000 — poderia estar associada a uma variação na pontuação nos exames que é aproximadamente constante para diferentes valores de renda?
- Na análise econômica da demanda do consumidor, freqüentemente se supõe que um aumento de 1 por cento nos preços leva a determinado *percentual* de queda na quantidade demandada. A variação percentual na demanda resultante de um aumento de 1 por cento no preço é chamada de **elasticidade-preço**.

[4] Uma forma diferente de escolher r é utilizar um "critério de informação", descrito no Capítulo 12 no contexto da análise de séries temporais. Na prática, o enfoque do critério de informação e o enfoque do teste seqüencial de hipótese descrito aqui freqüentemente geram resultados semelhantes.

As especificações de regressão que utilizam logaritmos naturais permitem que os modelos de regressão estimem relações de porcentagem como essas. Antes de apresentar essas especificações, revisaremos as funções exponencial e logaritmo natural.

A função exponencial e a função logarítmica (ou logaritmo natural). A função exponencial e seu inverso, o logaritmo natural, desempenham um papel importante na modelagem de funções de regressão não-lineares. A **função exponencial** de x é e^x, isto é, e elevado à potência x, onde e é a constante 2,71828 ...; a função exponencial também é expressa como $\exp(x)$. O **logaritmo natural** é o inverso da função exponencial, isto é, ele é a função para a qual $x = \ln(e^x)$ ou, de forma equivalente, $x = \ln[\exp(x)]$. A base do logaritmo natural é e. Embora haja logaritmos em outras bases, como base 10, neste livro consideraremos somente logaritmos na base e, isto é, o logaritmo natural, de modo que, quando utilizarmos o termo "logaritmo", estaremos nos referindo ao "logaritmo natural."

A Figura 6.4 mostra o gráfico da função logarítmica $y = \ln(x)$. Observe que ela só é definida para valores positivos de x. A função logarítmica possui uma declividade que é inicialmente acentuada e então se torna mais plana (embora a função continue a crescer). A declividade da função logarítmica $\ln(x)$ é $1/x$.

A função logarítmica tem as seguintes propriedades úteis:

$$\ln(1/x) = -\ln(x); \tag{6.12}$$

$$\ln(ax) = \ln(a) + \ln(x); \tag{6.13}$$

$$\ln(x/a) = \ln(x) - \ln(a); \text{ e} \tag{6.14}$$

$$\ln(x^a) = a\ln(x). \tag{6.15}$$

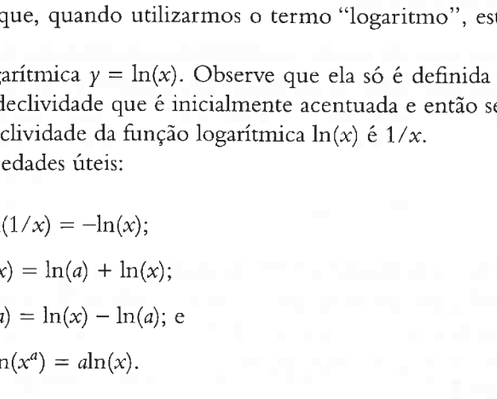
Logaritmos e porcentagem. A relação entre logaritmo e porcentagem baseia-se em um fato importante: quando Δx é pequeno, a diferença entre o logaritmo de $x + \Delta x$ e o logaritmo de x é aproximadamente $\frac{\Delta x}{x}$, a variação percentual em x dividida por 100. Isto é,

$$\ln(x + \Delta x) - \ln(x) \cong \frac{\Delta x}{x} \text{ (quando } \frac{\Delta x}{x} \text{ é pequeno)}, \tag{6.16}$$

onde $\cong$ significa "aproximadamente igual a". A derivação dessa aproximação baseia-se no cálculo, mas pode ser demonstrada rapidamente por meio da experimentação de alguns valores de x e Δx. Por exemplo, quando $x = 100$ e $\Delta x = 1$, então $\Delta x/x = 1/100 = 0{,}01$ (ou 1 por cento), ao passo que $\ln(x + \Delta x) - \ln(x) = \ln(101) - \ln(100) = 0{,}00995$ (ou 0,995 por cento). Assim, $\Delta x/x$ (que é 0,01) é muito próximo de $\ln(x + \Delta x) - \ln(x)$ (que é 0,00995). Quando $\Delta x = 5$, $\Delta x/x = 5/100 = 0{,}05$, ao passo que $\ln(x + \Delta x) - \ln(x) = \ln(105) - \ln(100) = 0{,}04879$.

FIGURA 6.4 A Função Logarítmica, $Y = \ln(X)$

A função logarítmica $Y = \ln(X)$ é mais inclinada para valores pequenos do que para valores grandes de X, é definida somente para $X > 0$ e tem declividade $1/X$.



Três modelos de regressão logarítmica. Existem três casos em que é possível utilizar logaritmos: quando X é transformado em seu logaritmo, mas Y não; quando Y é transformado em seu logaritmo, mas X não; e quando tanto X quanto Y são transformados em seus logaritmos. A interpretação dos coeficientes da regressão é diferente em cada caso. Discutiremos os três casos a seguir.

Caso I: X está em logaritmo, mas Y não. Neste caso, o modelo de regressão é

$$Y_i = \beta_0 + \beta_1\ln(X_i) + u_i,\ i = 1, ..., n. \tag{6.17}$$

Como Y não está em logaritmo, mas X está, esse modelo às vezes é chamado de **modelo linear-log**.

No modelo linear-log, uma variação de 1 por cento em X está associada a uma variação de $0{,}01\beta_1$ em Y. Para visualizar isso, considere a diferença entre a função de regressão da população para valores de X que diferem em ΔX: ela é $[\beta_0 + \beta_1\ln(X + \Delta X)] - [\beta_0 + \beta_1\ln(X)] = \beta_1[\ln(X + \Delta X) - \ln(X)] \cong \beta_1(\Delta X/X)$, onde o último passo utiliza a aproximação da Equação (6.16). Se X variar em 1 por cento, então $\Delta X/X = 0{,}01$; desse modo, nesse modelo, uma variação de 1 por cento em X está associada a uma variação de $0{,}01\beta_1$ em Y.

A única diferença entre o modelo de regressão da Equação (6.17) e o modelo de regressão do Capítulo 4 com um único regressor é que a variável do lado direito agora é o logaritmo de X, e não o próprio X. Para estimar o coeficientes β_0 e β_1 na Equação (6.17), primeiro calcule uma nova variável, $\ln(X)$; isso pode ser feito rapidamente utilizando uma planilha eletrônica ou um pacote estatístico. Então, β_0 e β_1 podem ser estimados pela regressão de MQO de Y_i sobre $\ln(X_i)$, hipóteses sobre β_1 podem ser testadas utilizando a estatística t e um intervalo de confiança de 95 por cento para β_1 pode ser construído como $\hat{\beta}_1 \pm 1{,}96EP(\hat{\beta}_1)$.

Como exemplo, volte à relação entre renda na diretoria e pontuação nos exames. No lugar da especificação quadrática, poderíamos utilizar a especificação linear-log da Equação (6.17). A estimação dessa regressão por MQO produz

$$\widehat{PontExame} = \underset{(3,8)}{557{,}8} + \underset{(1,40)}{36{,}42}\ln(Renda),\ \overline{R}^2 = 0{,}561. \tag{6.18}$$

De acordo com a Equação (6.18), um aumento de 1 por cento na renda está associado a um aumento na pontuação nos exames de $0{,}01 \times 36{,}42 = 0{,}36$ pontos.

Para estimarmos o efeito de uma variação em X sobre Y em suas unidades originais de milhares de dólares (não em logaritmos), podemos utilizar o método do Conceito-Chave 6.1. Por exemplo, qual é a diferença prevista na pontuação nos exames entre diretorias com rendas médias de US$ 10.000 *versus* US$ 11.000? O valor estimado de ΔY é a diferença entre os valores previstos: $\Delta\hat{Y} = [557{,}8 + 36{,}42\ln(11)] - [557{,}8 + 36{,}42\ln(10)] = 36{,}42 \times [\ln(11) - \ln(10)] = 3{,}47$. Do mesmo modo, a diferença prevista entre uma diretoria com renda média de US$ 40.000 e uma diretoria com renda média de US$ 41.000 é $36{,}42 \times [\ln(41) - \ln(40)] = 0{,}90$. Portanto, assim como a regressão quadrática, essa regressão prevê que um aumento de US$ 1.000 na renda tem um efeito maior na pontuação nos exames de diretorias pobres do que na de diretorias ricas.

A Figura 6.5 mostra a função de regressão linear-log estimada da Equação (6.18). Como o regressor na Equação (6.18) é o logaritmo natural da renda, e não a renda, a função de regressão estimada não é uma linha reta. Assim como a função de regressão quadrática da Figura 6.3, ela é inicialmente inclinada, mas torna-se mais plana para níveis mais altos de renda.

Caso II: Y está em logaritmo, mas X não. Neste caso, o modelo de regressão é

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i. \tag{6.19}$$

Como Y está em logaritmo, mas X não, ele é chamado de **modelo log-linear**.

No modelo log-linear, a variação unitária em X ($\Delta X = 1$) está associada a uma variação de $100 \times \beta_1$ por cento em Y. Para visualizar isso, compare os valores esperados de $\ln(Y)$ para valores de X que diferem em ΔX. O valor esperado de $\ln(Y)$ dado X é $\ln(Y) = \beta_0 + \beta_1 X$. Quando X é $X + \Delta X$, o valor esperado é dado por $\ln(Y + \Delta Y) = \beta_0 +$

$\beta_1(X + \Delta X)$. Assim, a diferença entre esses valores esperados é $\ln(Y + \Delta Y) - \ln(Y) = [\beta_0 + \beta_1(X + \Delta X)] - [\beta_0 + \beta_1 X] = \beta_1 \Delta X$. Da aproximação na Equação (6.16), contudo, se $\beta_1 \Delta X$ for pequeno, então $\ln(Y + \Delta Y) - \ln(Y) \cong \Delta Y/Y$. Assim, $\Delta Y/Y \cong \beta_1 \Delta X$. Se $\Delta X = 1$, de modo que X varia em uma unidade, $\Delta Y/Y$ varia em β_1. Traduzido em porcentagem, uma variação unitária em X está associada a uma variação de $100 \times \beta_1$ por cento em Y.

Para fins de ilustração, voltemos ao exemplo empírico da Seção 3.6, a relação entre idade e salário de indivíduos com curso superior. Muitos contratos de trabalho incluem uma cláusula segundo a qual, para cada ano adicional de serviço, um trabalhador ganha determinado percentual de aumento em seu salário. Essa relação percentual sugere que se estime a especificação log-linear da Equação (6.19) de modo que cada ano adicional de idade (X) está, em média, na população, associado a um aumento percentual constante no salário (Y). Calculando-se primeiro a nova variável dependente, $\ln(Salário_i)$, é possível estimar os coeficientes desconhecidos β_0 e β_1 pela regressão de MQO de $\ln(Salário_i)$ sobre $Idade_i$. A relação estimada utilizando as 12.077 observações sobre indivíduos com curso superior do Current Population Survey de 1999 (os dados estão descritos no Apêndice 3.1) é dada por

$$\widehat{\ln(Salário)} = 2{,}453 + 0{,}0128 Idade, \quad \overline{R}^2 = 0{,}0387. \qquad (6.20)$$
$$(0{,}024)\ (0{,}0006)$$

Segundo essa regressão, estima-se que o salário aumente em 1,28 por cento ((100 × 0,0128) por cento) para cada ano adicional de idade.

Caso III: Tanto X quanto Y estão em logaritmo. Neste caso, o modelo de regressão é

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i. \qquad (6.21)$$

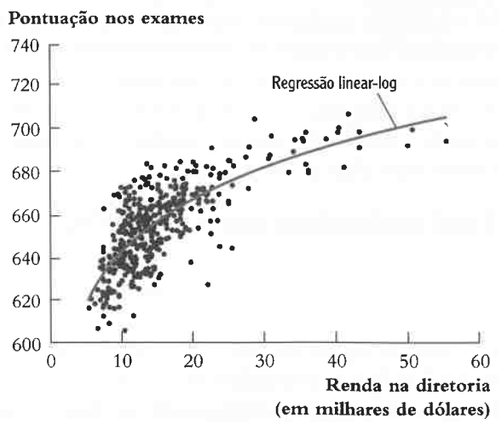
Como Y e X estão especificados em logaritmo, ele é identificado como **modelo log-log**.

No modelo log-log, uma variação de 1 por cento em X está associada a uma variação de β_1 por cento em Y. Assim, nessa especificação, β_1 é a elasticidade de Y com relação a X. Para visualizar isso, aplique novamente o Conceito-Chave 6.1; assim, $\ln(Y + \Delta Y) - \ln(Y) = [\beta_0 + \beta_1 \ln(X + \Delta X)] - [\beta_0 + \beta_1 \ln(X)] = \beta_1[\ln(X + \Delta X) - \ln(X)]$. A aplicação da aproximação na Equação (6.16) para os dois lados dessa equação produz

$$\frac{\Delta Y}{Y} \cong \beta_1 \frac{\Delta X}{X} \text{ ou} \qquad (6.22)$$

FIGURA 6.5 A Função de Regressão Linear-Log

A função de regressão linear-log estimada $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \ln(X)$ capta muito da relação não-linear entre pontuação nos exames e renda da diretoria.



$$\beta_1 = \frac{\Delta Y/Y}{\Delta X/X} = \frac{100 \times (\Delta Y/Y)}{100 \times (\Delta X/X)} = \frac{\text{Variação percentual de } Y}{\text{Variação percentual de } X}$$

Portanto, na especificação log-log, β_1 é a razão da variação percentual em Y associada à variação percentual em X. Se a variação percentual em X é de 1 por cento (isto é, se $\Delta X = 0{,}01X$), então β_1 é a variação percentual em Y associada a uma variação de 1 por cento em X. Isto é, β_1 é a elasticidade de Y com relação a X.

Para fins de ilustração, voltemos à relação entre renda e pontuação nos exames. Quando essa relação está especificada dessa forma, os coeficientes desconhecidos são estimados por uma regressão do logaritmo de pontuação nos exames contra o logaritmo da renda. A equação estimada resultante é

$$\widehat{PontExame} = 6{,}336 + 0{,}0554 \ln(Renda), \quad \overline{R}^2 = 0{,}557. \qquad (6.23)$$
$$(0{,}006)\ (0{,}0021)$$

De acordo com essa função de regressão, estima-se que um aumento de 1 por cento na renda corresponda a um aumento de 0,0554 por cento na pontuação nos exames.

A Figura 6.6 mostra a função de regressão log-log estimada da Equação (6.23). Como Y está em logaritmos, o eixo vertical na Figura 6.6 é o logaritmo da pontuação nos exames e o gráfico de dispersão é o logaritmo da pontuação nos exames *versus* renda na diretoria. Para fins de comparação, a Figura 6.6 também mostra a função de regressão estimada para uma especificação log-linear, que é

$$\widehat{PontExame} = 6{,}439 + 0{,}00284 Renda, \quad \overline{R}^2 = 0{,}497. \qquad (6.24)$$
$$(0{,}003)\ (0{,}00018)$$

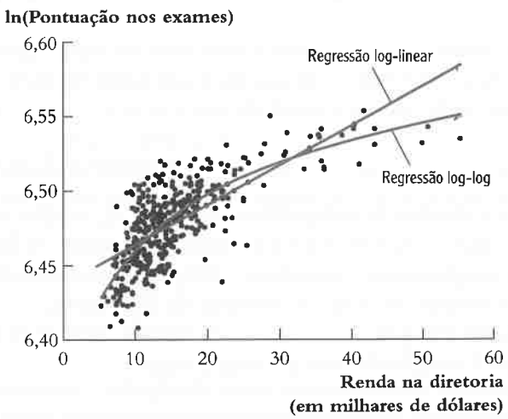
Como o eixo vertical está em logaritmos, a função de regressão na Equação (6.24) é a linha reta da Figura 6.6. Na Figura 6.6 você pode ver que a especificação log-log se ajusta ligeiramente melhor do que a especificação log-linear. Isso é consistente com o $\overline{R}^2$ maior para a regressão log-log (0,557) do que para a regressão log-linear (0,497). Mesmo assim, a especificação log-log não se ajusta tão bem aos dados: nos valores mais baixos da renda, a maioria das observações situa-se abaixo da curva log-log, ao passo que no nível médio da renda a maioria das observações situa-se acima da função de regressão estimada.

Os três modelos de regressão logarítmica estão resumidos no Conceito-Chave 6.2.

A dificuldade de comparar regressões logarítmicas. Qual dos modelos de regressão logarítmica se ajusta melhor aos dados? Como vimos na discussão das equações (6.23) e (6.24), podemos utilizar o $\overline{R}^2$ para

FIGURA 6.6 Funções de Regressão Log-Linear e Log-Log

Na função de regressão log-linear, $\ln(Y)$ é uma função linear de X. Na função de regressão log-log, $\ln(Y)$ é uma função linear de $\ln(X)$.



Conceito-Chave 6.2

Logaritmos na Regressão: Três Casos

Logaritmos podem ser utilizados para transformar a variável dependente Y, a variável independente X, ou ambas (desde que elas sejam positivas). A tabela a seguir resume esses três casos e a interpretação do coeficiente de regressão β_1. Em cada caso, é possível estimar β_1 aplicando MQO após a tomada do logaritmo da variável dependente e/ou independente.

Caso	Especificações da Regressão	Interpretação de β_1
I	$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$	A variação de 1 por cento em X está associada a uma variação de $0{,}01\beta_1$ em Y.
II	$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$	A variação de uma unidade em X ($\Delta X = 1$) está associada a uma variação de $100\beta_1$ por cento em Y.
III	$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$	A variação de 1 por cento em X está associada a uma variação de β_1 por cento em Y, de modo que β_1 é a elasticidade de Y com relação a X.

comparar os modelos log-linear e log-log; constatamos que o modelo log-log apresentava o $\overline{R}^2$ maior. Da mesma forma, podemos utilizar o $\overline{R}^2$ para comparar a regressão linear-log da Equação (6.18) com a regressão linear de Y contra X. Na regressão entre pontuação nos exames e renda, a regressão linear-log possui um $\overline{R}^2$ de 0,561, ao passo que a regressão linear possui um $\overline{R}^2$ de 0,508, de modo que o modelo linear-log se ajusta melhor aos dados.

Como podemos comparar o modelo linear-log e o modelo log-log? Infelizmente, o $\overline{R}^2$ *não pode* ser utilizado para comparar essas duas regressões, uma vez que suas variáveis dependentes são diferentes (uma é Y_i e a outra é $\ln(Y_i)$). Lembre-se de que o $\overline{R}^2$ mede a fração da variância da variável dependente explicada pelos regressores. Como as variáveis dependentes dos modelos log-log e linear-log são diferentes, não faz sentido comparar seus $\overline{R}^2$s.

Em virtude desse problema, a melhor coisa a fazer em uma aplicação em particular é decidir se faz sentido especificar Y em logaritmos, utilizando a teoria econômica e o conhecimento que você e os outros têm do problema. Por exemplo, pesquisadores de economia do trabalho geralmente modelam o salário utilizando logaritmos, uma vez que as comparações de salários, os reajustes salariais e assim por diante freqüentemente são discutidos de modo mais natural em termos percentuais. Na modelagem da pontuação nos exames, parece natural (ao menos para nós) discutir os resultados dos exames em termos dos pontos feitos, e não dos aumentos percentuais na pontuação, de modo que nos concentramos em modelos em que a variável dependente é a pontuação no exame, e não seu logaritmo.

Calculando valores previstos de Y quando este se encontra em logaritmos. [5] Se a variável dependente Y foi transformada em logaritmos, a regressão estimada pode ser utilizada para calcular diretamente o valor previsto de $\ln(Y)$. Contudo, é um pouco complicado calcular o próprio valor previsto de Y.

Para visualizar isso, considere o modelo de regressão log-linear da Equação (6.19) e reescreva-o de modo que ele esteja especificado em termos de Y, e não de $\ln(Y)$. Para fazer isso, tome a função exponencial dos dois lados da Equação (6.19); o resultado é

$$Y_i = \exp(\beta_0 + \beta_1 X_i + u_i) = e^{\beta_0 + \beta_1 X_i} e^{u_i}. \tag{6.25}$$

Se u_i é distribuído independentemente de X_i, o valor esperado de Y_i dado X_i é $E(Y_i|X_i) = E(e^{\beta_0+\beta_1 X_i} e^{u_i}|X_i) = e^{\beta_0+\beta_1 X_i} E(e^{u_i})$. O problema é que, mesmo que $E(u_i) = 0$, $E(e^{u_i}) \neq 1$. Assim, o valor previsto apropriado de Y_i não é obtido simplesmente tomando-se a função exponencial de $\hat{\beta}_0 + \hat{\beta}_1 X_i$, isto é, fazendo $\hat{Y}_i = e^{\hat{\beta}_0+\hat{\beta}_1 X_i}$; esse valor previsto é viesado em razão da ausência do fator $E(e^{u_i})$.

Uma solução para esse problema é estimar o fator $E(e^{u_i})$ e utilizá-lo no cálculo do valor previsto de Y, porém, como isso é complicado, não prosseguiremos.

[5] Este material é mais avançado e por isso pode ser pulado sem perda de continuidade.

Outra "solução", que é o enfoque utilizado neste livro, é calcular os valores previstos do logaritmo de Y, mas não transformá-los em suas unidades originais. Na prática, isso em geral é aceitável, uma vez que, quando a variável dependente é especificada como um logaritmo, freqüentemente é mais natural utilizar somente a especificação logarítmica (e as interpretações percentuais associadas) ao longo da análise.

Modelos Polinomiais e Logarítmicos de Pontuação nos Exames e na Renda na Diretoria

Na prática, a teoria econômica e o julgamento cuidadoso podem sugerir uma forma funcional a ser utilizada, mas no final a forma verdadeira da função de regressão da população é desconhecida. Na prática, ajustar uma função não-linear envolve, portanto, a decisão sobre o método ou a combinação de métodos que funciona melhor. Para fins de ilustração, compararemos os modelos logarítmico e polinomial da relação entre renda na diretoria e pontuação nos exames.

Especificações polinomiais. Consideremos duas especificações polinomiais, a quadrática (veja a Equação (6.2)) e a cúbica (veja a Equação (6.11)), especificadas utilizando-se potências de *Renda*. Como o coeficiente de $Renda^3$ na Equação (6.11) era significante ao nível de 5 por cento, a especificação cúbica ofereceu uma melhora em relação à quadrática, de modo que selecionamos o modelo cúbico como a especificação polinomial preferida.

Especificações logarítmicas. A especificação logarítmica na Equação (6.18) parecia fornecer um bom ajuste a esses dados, mas não a testamos formalmente. Uma maneira de fazer isso é ampliá-la com potências mais altas do logaritmo da renda. Se esses termos adicionais não são estatisticamente diferentes de zero, podemos concluir que a especificação na Equação (6.18) é adequada, uma vez que não pode ser rejeitada contra uma função polinomial do logaritmo. Desse modo, a regressão cúbica estimada (especificada em potências do logaritmo da renda) é

$$\begin{aligned}\widehat{PontExame} &= \underset{(79,4)}{486,1} + \underset{(87,9)}{113,4}\ln(Renda) - \underset{(31,7)}{26,9}[\ln(Renda)]^2 \\ &\quad + \underset{(3,74)}{3,06}[\ln(Renda)]^3,\ \overline{R}^2 = 0,560.\end{aligned} \tag{6.26}$$

A estatística t do coeficiente do termo cúbico é 0,818, de modo que a hipótese nula de que o coeficiente verdadeiro é zero não é rejeitada ao nível de 10 por cento. A estatística F que testa a hipótese conjunta de que os coeficientes verdadeiros dos termos quadrático e cúbico são iguais a zero é 0,44, com um valor p de 0,64, de modo que essa hipótese nula conjunta não é rejeitada ao nível de 10 por cento. Desse modo, o modelo logarítmico cúbico na Equação (6.26) não fornece uma melhora estatisticamente significante em relação ao modelo da Equação (6.18), que é linear no logaritmo da renda.

Comparando as especificações cúbica e linear-log. A Figura 6.7 mostra as funções de regressão estimadas da especificação cúbica da Equação (6.11) e a especificação linear-log da Equação (6.18). As duas funções de regressão estimadas são muito semelhantes. Uma ferramenta estatística que compara essas especificações é o $\overline{R}^2$. O $\overline{R}^2$ da regressão logarítmica é 0,561 e o da regressão cúbica é 0,555. Como a especificação logarítmica possui uma pequena vantagem em termos do $\overline{R}^2$, e como essa especificação não precisa de polinômios de ordens mais altas no logaritmo da renda para o ajuste aos dados, adotamos a especificação logarítmica da Equação (6.18).

6.3 Interações entre Variáveis Independentes

Na introdução deste capítulo, nós nos perguntamos se a redução da razão aluno-professor poderia ter um efeito maior sobre a pontuação nos exames em diretorias em que muitos alunos ainda estão aprendendo inglês

do que naquelas em que poucos ainda estão aprendendo inglês. Isso poderia ocorrer, por exemplo, se os alunos que ainda estão aprendendo inglês se beneficiassem de forma diferenciada de uma instrução individualizada ou em grupos pequenos. Se for esse o caso, a presença de muitos alunos que estão aprendendo inglês em uma diretoria interagiria com a razão aluno-professor de tal forma que o efeito de uma variação na razão aluno-professor sobre a pontuação nos exames dependeria da fração de alunos aprendendo inglês.

Nesta seção, explicamos como incluir essas interações entre duas variáveis independentes no modelo de regressão múltipla. A interação possível entre a razão aluno-professor e a fração de alunos que está aprendendo inglês é um exemplo da situação mais geral em que o efeito de uma variação em uma variável independente sobre Y depende do valor de outra variável independente. Consideremos três casos: quando ambas as variáveis independentes são binárias, quando uma é binária e a outra é contínua, e quando ambas são contínuas.

Interações entre Duas Variáveis Binárias

Considere a regressão da população do logaritmo do salário (Y_i, onde $Y_i = \ln(Salário_i)$) contra duas variáveis binárias, o sexo do indivíduo (D_{1i}, que é igual a 1 se a i-ésima pessoa é mulher) e se esse indivíduo tem curso superior (D_{2i}, onde $D_{2i} = 1$ se a i-ésima pessoa tem curso superior). A regressão linear da população de Y_i sobre essas duas variáveis binárias é

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + u_i. \tag{6.27}$$

Nesse modelo de regressão, β_1 é o efeito de ser mulher sobre o logaritmo do salário, mantendo constante o nível de instrução, e β_2 é o efeito de ter um curso superior, mantendo constante o sexo.

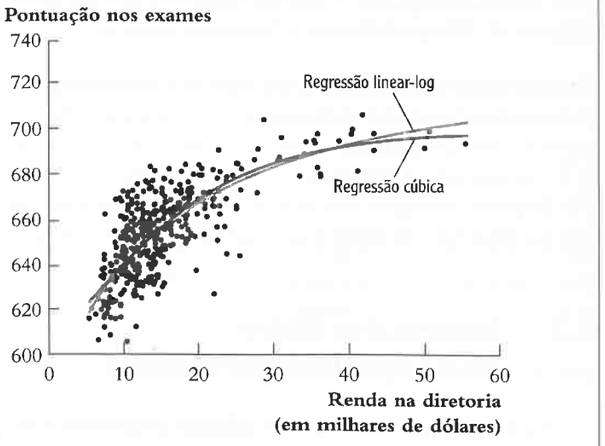
A especificação na Equação (6.27) tem uma limitação importante: o efeito de ter um curso superior nessa especificação, mantendo constante o sexo, é o mesmo para homens e mulheres. Entretanto, não há motivos para que isso seja assim. Expresso matematicamente, o efeito de D_{2i} sobre Y_i, mantendo constante D_{1i}, poderia depender do valor de D_{1i}. Em outras palavras, poderia haver uma interação entre sexo e ter um curso superior, de modo que o valor no mercado de trabalho de um curso superior seria diferente para homens e mulheres.

Embora a especificação na Equação (6.27) não permita essa interação entre sexo e ter um curso superior, é fácil modificar a especificação para que ela permita a interação por meio da introdução de outro regressor, o produto de duas variáveis binárias, $D_{1i} \times D_{2i}$. A regressão resultante é

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} \times D_{2i}) + u_i. \tag{6.28}$$

FIGURA 6.7 Funções de Regressão Linear-Log e Cúbica

A função de regressão cúbica estimada (veja a Equação (6.11)) e a função de regressão linear-log estimada (veja a Equação (6.18)) são praticamente idênticas nesta amostra.



O novo regressor, o produto $D_{1i} \times D_{2i}$, é chamado de **termo de interação** ou **regressor interado**, e o modelo de regressão da população na Equação (6.28) é chamado de **modelo de regressão com interação** entre variáveis binárias.

O termo de interação na Equação (6.28) permite que o efeito da população de ter um curso superior (variando D_{2i} de $D_{2i} = 0$ para $D_{2i} = 1$) sobre o logaritmo do salário (Y_i) dependa do sexo (D_{1i}). Para mostrar isso matematicamente, calcule o efeito da população de uma variação em D_{2i} utilizando o método geral exposto no Conceito-Chave 6.1. O primeiro passo é calcular a expectativa condicional de Y_i para $D_{2i} = 0$, dado um valor de D_{1i}; esta é $E(Y_i|D_{1i} = d_1, D_{2i} = 0) = \beta_0 + \beta_1 \times d_1 + \beta_2 \times 0 + \beta_3 \times (d_1 \times 0) = \beta_0 + \beta_1 d_1$. O próximo passo é calcular a expectativa condicional de Y_i após a variação, isto é, para $D_{2i} = 1$, dado o mesmo valor de D_{1i}; esta é $E(Y_i|D_{1i} = d_1, D_{2i} = 1) = \beta_0 + \beta_1 \times d_1 + \beta_2 \times 1 + \beta_3 \times (d_1 \times 1) = \beta_0 + \beta_1 d_1 + \beta_2 + \beta_3 d_1$. O efeito dessa variação é a diferença entre os valores esperados (isto é, a diferença na Equação (6.6)), que é

$$E(Y_i \mid D_{1i} = d_1, D_{2i} = 1) - E(Y_i \mid D_{1i} = d_1, D_{2i} = 0) = \beta_2 + \beta_3 d_1. \tag{6.29}$$

Assim, na especificação da interação entre variáveis binárias da Equação (6.28), o efeito de ter um curso superior (uma variação unitária em D_{2i}) depende do sexo da pessoa (o valor de D_{1i}, que é d_1 na Equação (6.29)). Se a pessoa é do sexo masculino ($d_1 = 0$), o efeito de ter um curso superior é β_2, mas, se a pessoa é do sexo feminino ($d_1 = 1$), o efeito é $\beta_2 + \beta_3$. O coeficiente β_3 sobre o termo de interação é a diferença entre o efeito de ter um curso superior para mulheres *versus* homens.

Embora esse exemplo tenha sido exposto utilizando o logaritmo do salário, sexo e ter um curso superior, o ponto é geral. A regressão com interação entre variáveis binárias permite que o efeito da variação em uma das variáveis independentes binárias dependa do valor de outra variável binária.

Para interpretarmos os coeficientes, consideramos cada combinação possível das variáveis binárias. Esse método, que se aplica a todas as regressões com variáveis binárias, está resumido no Conceito-Chave 6.3.

Aplicação para a razão aluno-professor e porcentagem de alunos que está aprendendo inglês. Seja $RAPAlta_i$ uma variável binária que é igual a um se a razão aluno-professor é de 20 ou mais e igual a zero nos demais casos; seja $AIAlta_i$ uma variável binária que é igual a um se a porcentagem de alunos que está aprendendo inglês é de 10 por cento ou mais e igual a zero nos demais casos. A regressão interada da pontuação nos exames contra $RAPAlta_i$ e $AIAlta_i$ é

$$\begin{aligned} \widehat{PontExame} &= 664{,}1 - 18{,}2 AIAlta - 1{,}9 RAPAlta - 3{,}5 (RAPAlta \times AIAlta), \\ &\quad (1{,}4) \quad (2{,}3) \qquad (1{,}9) \qquad (3{,}1) \\ \overline{R}^2 &= 0{,}290. \end{aligned} \tag{6.30}$$

O efeito previsto da mudança de uma diretoria com razão aluno-professor baixa para outra com razão aluno-professor alta, mantendo constante a porcentagem alta ou baixa de alunos que está aprendendo inglês, é dado pela Equação (6.29), com coeficientes estimados substituindo os coeficientes da população. De acordo com as estimativas na Equação (6.30), esse efeito é $-1{,}9 - 3{,}5 AIAlta$. Isto é, se a fração de alunos que está aprendendo inglês é baixa ($AIAlta_i = 0$), o efeito de passar de $RAPAlta_i = 0$ para $RAPAlta_i = 1$ sobre a pontuação nos exames equivale a uma queda de 1,9 ponto. Se a fração de alunos que está aprendendo inglês é alta, estima-se que a pontuação nos exames caia em 1,9 + 3,5 = 5,4 pontos.

A regressão estimada na Equação (6.30) também pode ser utilizada para estimar a pontuação média nos exames para cada uma das quatro combinações possíveis das variáveis binárias. Isso é feito utilizando-se o procedimento do Conceito-Chave 6.3. Assim, a pontuação média nos exames da amostra para diretorias com razão aluno-professor baixa ($RAPAlta_i = 0$) e fração baixa de alunos que está aprendendo inglês ($AIAlta_i = 0$) é de 664,1. Para diretorias com $RAPAlta_i = 1$ (razão aluno-professor alta) e $AIAlta_i = 0$ (fração baixa de alunos que está aprendendo inglês), a média da amostra é de 662,2 (= 664,1 – 1,9). Quando $RAPAlta_i = 0$ e $AIAlta_i = 1$, a média da amostra é de 645,9 (= 664,1 – 18,2); e quando $RAPAlta_i = 1$ e $AIAlta_i = 1$, a média da amostra é de 640,5 (= 664,1 – 18,2 – 1,9 – 3,5).

Conceito-Chave 6.3

Método para Interpretar os Coeficientes em Regressões com Variáveis Binárias

Em primeiro lugar, calcule os valores esperados de Y para cada caso possível descrito pelo conjunto de variáveis binárias. A seguir, compare esses valores esperados. Cada coeficiente pode então ser expresso tanto como um valor esperado quanto como a diferença entre dois ou mais valores esperados.

Interações entre uma Variável Contínua e uma Variável Binária

A seguir, considere a regressão da população do logaritmo do salário ($Y_i = \ln(Salário_i)$) contra uma variável contínua, os anos de experiência profissional de um indivíduo (X_i) e uma variável binária, se o trabalhador tem curso superior (D_i, onde $D_i = 1$, se a i-ésima pessoa tem curso superior). Conforme a Figura 6.8 mostra, existem três formas para que a reta de regressão da população que relaciona Y e a variável contínua X possa depender da variável binária D.

Na Figura 6.8a, as duas retas de regressão diferem somente em seu intercepto. O modelo de regressão da população correspondente é

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i. \tag{6.31}$$

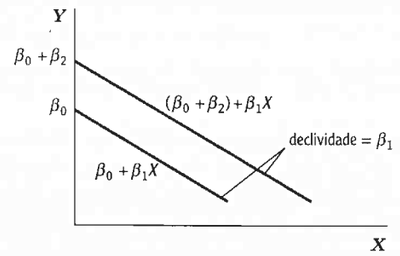
Esse é o modelo de regressão múltipla familiar com uma função de regressão da população que é linear em X_i e D_i. Quando $D_i = 0$, a função de regressão da população é $\beta_0 + \beta_1 X_i$, de modo que o intercepto é β_0 e a declividade é β_1. Quando $D_i = 1$, a função de regressão da população é $\beta_0 + \beta_1 X_i + \beta_2$, de modo que a declividade permanece β_1, mas o intercepto passa a ser $\beta_0 + \beta_2$. Desse modo, β_2 é a diferença entre os interceptos das duas retas de regressão, como mostra a Figura 6.8a. Expresso em termos do exemplo do salário, β_1 é o efeito de um ano adicional de experiência profissional sobre o logaritmo do salário, mantendo constante o fato de ter ou não um curso superior, e β_2 é o efeito de ter um curso superior sobre o logaritmo do salário, mantendo constantes os anos de experiência. Nessa especificação, o efeito de um ano adicional de experiência profissional é o mesmo para indivíduos com ou sem curso superior, isto é, as duas retas na Figura 6.8a possuem a mesma declividade.

Na Figura 6.8b, as duas retas têm declividade e intercepto diferentes. Declividades distintas permitem que o efeito de um ano adicional de trabalho seja diferente para indivíduos com e sem curso superior. Para permitir inclinações diferentes, adicione um termo de interação à Equação (6.31):

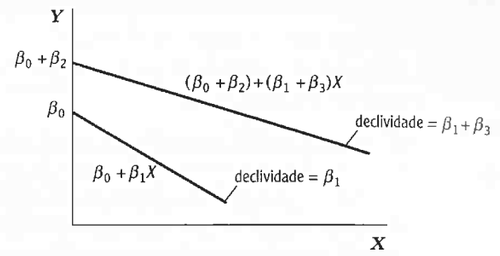
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i, \tag{6.32}$$

onde $X_i \times D_i$ é uma nova variável, o produto de X_i por D_i. Para interpretar os coeficientes dessa regressão, aplique o procedimento do Conceito-Chave 6.3. Sua aplicação mostra que, se $D_i = 0$, a função de regressão da população é $\beta_0 + \beta_1 X_i$, ao passo que, se $D_i = 1$, a função de regressão da população é $(\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_i$. Desse modo, essa especificação permite duas funções de regressão da população que relacionam Y_i e X_i, dependendo do valor de D_i, como é mostrado na Figura 6.8b. A diferença entre os dois interceptos é β_2 e a diferença entre as duas declividades é β_3. No exemplo do salário, β_1 é o efeito de um ano adicional de experiência profissional para indivíduos sem curso superior ($D_i = 0$) e $\beta_1 + \beta_3$ é o efeito para indivíduos com curso superior, de modo que β_3 é a *diferença* entre o efeito de um ano adicional de experiência profissional para indivíduos com curso superior *versus* indivíduos sem curso superior.

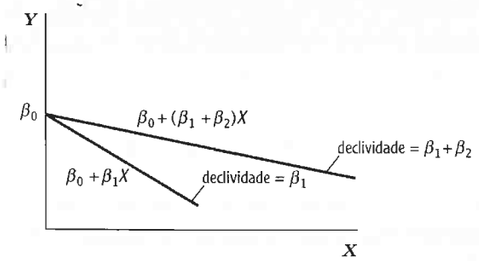
FIGURA 6.8 Funções de Regressão Utilizando Variáveis Binárias e Variáveis Contínuas



(a) Interceptos diferentes, mesma declividade



(b) Interceptos diferentes, declividades diferentes



(c) Mesmo intercepto, declividades diferentes

Interações entre variáveis binárias e variáveis contínuas podem produzir três funções de regressão da população: (a) $\beta_0 + \beta_1 X + \beta_2 D$ permite interceptos diferentes, mas tem a mesma declividade; (b) $\beta_0 + \beta_1 X + \beta_2 D + \beta_3 (X \times D)$ permite interceptos diferentes e declividades diferentes; e (c) $\beta_0 + \beta_1 X + \beta_2 (X \times D)$ tem o mesmo intercepto, mas permite declividades diferentes.

Uma terceira possibilidade, mostrada na Figura 6.8c, é a de que as duas retas têm declividades diferentes, mas o mesmo intercepto. O modelo de regressão com interação para esse caso é

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i \times D_i) + u_i. \tag{6.33}$$

Os coeficientes dessa especificação também podem ser interpretados utilizando o Conceito-Chave 6.3. Em termos do exemplo do salário, essa especificação permite efeitos diferentes da experiência sobre o logaritmo do salário entre indivíduos com e sem curso superior, mas requer que o logaritmo do salário esperado seja o mesmo para ambos os grupos quando eles não têm experiência anterior. Dito de outra forma, essa especificação equivale à igualdade do salário inicial médio da população para indivíduos com e sem curso superior. Isso não faz muito sentido para essa aplicação; na prática, essa especificação é utilizada com menor freqüência do que a Equação (6.32), que permite interceptos e declividades diferentes.

As três especificações, equações (6.31), (6.32) e (6.33), são versões do modelo de regressão múltipla do Capítulo 5 e, uma vez que uma nova variável $X_i \times D_i$ seja criada, os coeficientes das três podem ser estimados por MQO.

Os três modelos de regressão com uma variável independente binária e uma variável independente contínua estão resumidos no Conceito-Chave 6.4.

Aplicação para a razão aluno-professor e a porcentagem de alunos que está aprendendo inglês. O efeito do corte da razão aluno-professor sobre a pontuação nos exames depende de a porcentagem de alunos que está aprendendo inglês ser alta ou baixa? Uma forma de responder a essa pergunta é utilizar uma especificação que permite duas retas de regressão diferentes, dependendo de a porcentagem de alunos que está apren-

Conceito-Chave 6.4

Interações entre Variáveis Binárias e Variáveis Contínuas

Utilizando-se o termo de interação $X_i \times D_i$, a reta de regressão da população que relaciona Y_i e a variável contínua X_i pode ter uma declividade que dependa da variável binária D_i. Existem três possibilidades:

1. Intercepto diferente, mesma declividade (veja a Figura 6.8a):
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i.$$
2. Intercepto diferente e declividade diferente (veja a Figura 6.8b):
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i.$$
3. Mesmo intercepto, declividade diferente (veja a Figura 6.8c):
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i \times D_i) + u_i.$$

dendo inglês ser alta ou baixa. Isso é obtido utilizando-se a especificação intercepto diferente/declividade diferente:

$$\widehat{PontExame} = \underset{(11,9)}{682,2} - \underset{(0,59)}{0,97} RAP + \underset{(19,5)}{5,6} AIAlta - \underset{(0,97)}{1,28}(RAP \times AIAlta) \quad (6.34)$$
$$\overline{R}^2 = 0,305,$$

onde a variável binária $AIAlta_i$ é igual a um se a porcentagem de alunos que está aprendendo inglês na diretoria é maior do que 10 por cento e igual a zero nos demais casos.

Para diretorias com uma fração baixa de alunos que está aprendendo inglês ($AIAlta_i = 0$), a reta de regressão estimada é $682,2 - 0,97RAP_i$. Para diretorias com uma fração alta de alunos que está aprendendo inglês ($AIAlta_i = 1$), a reta de regressão estimada é $682,2 + 5,6 - 0,97RAP_i - 1,28RAP_i = 687,8 - 2,25RAP_i$. De acordo com essas estimativas, prevê-se que a redução da razão aluno-professor em um aumentará a pontuação nos exames em 0,97 pontos nas diretorias com frações baixas de alunos aprendendo inglês e em 2,25 pontos nas diretorias com frações altas de alunos aprendendo inglês. A diferença entre esses dois efeitos, 1,28 pontos, é o coeficiente do termo de interação na Equação (6.34).

A regressão de MQO na Equação (6.34) pode ser utilizada para testar várias hipóteses sobre a reta de regressão da população. Em primeiro lugar, a hipótese de que as duas retas são na verdade a mesma pode ser testada calculando-se a estatística F que testa a hipótese conjunta de que o coeficiente de $AIAlta$ e o coeficiente do termo de interação $RAP_i \times AIAlta_i$ são iguais a zero. A estatística F é 89,9, que é significante ao nível de 1 por cento.

Em segundo lugar, o teste da hipótese de que as duas retas têm a mesma declividade pode ser feito ao testar se o coeficiente do termo de interação é igual a zero. A estatística $-1,28/0,97 = -1,32$ é menor do que 1,645 em valor absoluto, de modo que a hipótese nula de que as duas retas têm a mesma declividade não pode ser rejeitada utilizando-se um teste bicaudal ao nível de significância de 10 por cento.

Em terceiro lugar, o teste da hipótese de que as duas retas têm o mesmo intercepto pode ser feito ao testar se o coeficiente da população de $AIAlta$ é igual a zero. A estatística $t = 5,6/19,5 = 0,29$, de modo que a hipótese de que as retas têm o mesmo intercepto não pode ser rejeitada ao nível de 5 por cento.

Esses três testes geram resultados aparentemente contraditórios: o teste conjunto com a utilização da estatística F rejeita a hipótese conjunta de que a declividade e o intercepto são os mesmos, mas os testes das hipóteses individuais que utilizam a estatística t não as rejeitam. Isso ocorre porque os regressores $AIAlta$ e $RAP \times AIAlta$ são altamente correlacionados, o que resulta em erros padrão grandes dos coeficientes individuais. Mesmo que seja impossível dizer qual dos coeficientes é diferente de zero, existe uma forte evidência contra a hipótese de que *ambos* são iguais a zero.

Finalmente, a hipótese de que a razão aluno-professor não entra nessa especificação pode ser testada pelo cálculo da estatística F para a hipótese conjunta de que os coeficientes da RAP e do termo de interação são iguais a zero. A estatística F é 5,64, com um valor p de 0,004. Desse modo, os coeficientes da razão aluno-professor são estatisticamente significantes ao nível de significância de 1 por cento.

Interações entre Duas Variáveis Contínuas

Agora suponha que ambas as variáveis independentes (X_{1i} e X_{2i}) sejam contínuas. Por exemplo, seja Y_i o logaritmo do salário do i-ésimo trabalhador, seja X_{1i} os anos de experiência profissional e seja X_{2i} o número de anos que o indivíduo freqüentou a escola. Se a função de regressão da população é linear, o efeito de um ano adicional de experiência sobre o salário não depende do número de anos de estudo ou, de forma equivalente, o efeito de um ano adicional de instrução não depende do número de anos de experiência profissional. Na realidade, entretanto, pode haver uma interação entre essas duas variáveis de modo que o efeito de um ano adicional de experiência sobre o salário dependa do número de anos de estudo. Essa interação pode ser modelada ampliando-se o modelo de regressão linear com um termo de interação que é o produto de X_{1i} por X_{2i}:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i. \quad (6.35)$$

O termo de interação permite que o efeito de uma variação unitária em X_1 dependa de X_2. Para visualizar isso, aplique o método geral para o cálculo dos efeitos em modelos de regressão não-linear do Conceito-Chave 6.1. A diferença na Equação (6.6), calculada para a função de regressão interada na Equação (6.35), é $\Delta Y = (\beta_1 + \beta_3 X_2)\Delta X_1$ (veja o Exercício 6.5(a)). Portanto, o efeito de uma variação em X_1 sobre Y, mantendo X_2 constante, é

$$\frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2, \quad (6.36)$$

que depende de X_2. Por exemplo, no caso do salário, se β_3 for positivo, o efeito de um ano adicional de experiência sobre o logaritmo do salário é maior, no montante β_3, para cada ano adicional de instrução do trabalhador.

Um cálculo semelhante mostra que o efeito de uma variação ΔX_2 em X_2, sobre Y, mantendo X_1 constante, é $\frac{\Delta Y}{\Delta X_2} = (\beta_2 + \beta_3 X_1)$.

Colocando esses dois efeitos juntos vemos que o coeficiente β_3 do termo de interação é o efeito de um aumento unitário em X_1 *e* X_2, maior do que a soma dos efeitos de um aumento unitário somente em X_1 e de um aumento unitário somente em X_2. Isto é, se X_1 varia em ΔX_1 *e* X_2 varia em ΔX_2, a variação esperada em Y é $\Delta Y = (\beta_1 + \beta_3 X_2)\Delta X_1 + (\beta_2 + \beta_3 X_1)\Delta X_2 + \beta_3 \Delta X_1 \Delta X_2$ (veja o Exercício 6.5(c)). O primeiro termo é o efeito de variar X_1 mantendo X_2 constante; o segundo termo é o efeito de variar X_2 mantendo X_1 constante; e o último termo, $\beta_3 \Delta X_1 \Delta X_2$, é o efeito extra de variar tanto X_1 quanto X_2.

O Conceito-Chave 6.5 resume as interações entre duas variáveis.

Quando as interações são combinadas com transformações logarítmicas, elas podem ser utilizadas para estimar a elasticidade-preço quando esta depende das características do bem (veja o quadro para um exemplo).

Aplicação para a razão aluno-professor e porcentagem de alunos aprendendo inglês. Os exemplos anteriores consideraram interações entre a razão aluno-professor e uma variável binária indicando se a porcentagem dos alunos aprendendo inglês era grande ou pequena. Uma forma diferente de estudar essa interação é examinar a interação entre a razão aluno-professor e a variável contínua porcentagem de alunos aprendendo inglês ($\%AI$). A regressão de interação estimada é

$$\widehat{PontExame} = \underset{(11,8)}{686,3} - \underset{(0,59)}{1,12} RAP - \underset{(0,37)}{0,67}\%AI + \underset{(0,019)}{0,0012}(RAP \times \%AI), \quad (6.37)$$
$$\overline{R}^2 = 0,422.$$

Demanda por Periódicos de Economia

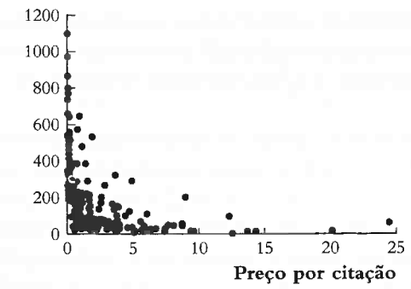
Os economistas acompanham as pesquisas mais recentes em suas áreas de especialização. A maioria das pesquisas em economia aparece primeiro em periódicos de economia, de modo que os economistas — ou suas bibliotecas — assinam esses periódicos.

Em que medida a demanda por periódicos de economia pelas bibliotecas é elástica? Para descobrir isso, analisamos a relação entre o número de assinaturas de um periódico nas bibliotecas dos Estados Unidos (Y_i) e o preço de sua assinatura para a biblioteca, utilizando dados de 180 periódicos de economia do ano 2000. Como o produto de um periódico não é o papel em que ele é impresso, mas as idéias que contém, seu preço logicamente não é medido em dólares por ano ou dólares por página, mas em dólares por idéia. Embora não possamos medir diretamente "idéias", uma boa medida indireta é o número de vezes que os artigos de um periódico são citados posteriormente por outros pesquisadores. Portanto, medimos o preço como "preço por citação" no periódico. A gama de preços é enorme, de 0,5 centavo de dólar por citação (*American Economic Review*) a 20 centavos de dólar por citação ou mais. Alguns periódicos têm preços por citação altos porque possuem poucas citações, outros porque o preço da assinatura anual para biblioteca é muito alto: em 2000, uma assinatura para biblioteca do *Journal of Econometrics* custou cerca de US$ 1.900, quarenta vezes o preço de uma assinatura da *American Economic Review*!

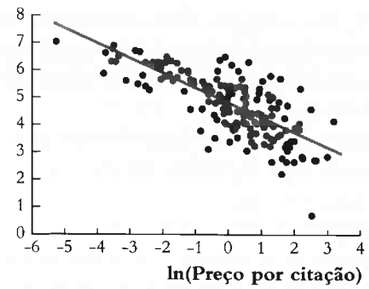
Como estamos interessados em estimar elasticidades, utilizamos uma especificação log-log (veja o Conceito-Chave 6.2). Os gráficos de dispersão das figuras 6.9a e 6.9b fornecem suporte empírico para essa transformação. Como alguns dos periódicos mais antigos e respeitados têm preços por citação mais baixos, uma regressão do logaritmo da quantidade contra o logaritmo do preço poderia ter viés de omissão de variáveis. Nossas regressões, portanto, incluem duas variáveis de controle, o logaritmo da idade e o logaritmo do número de caracteres por ano no periódico.

(Continua)

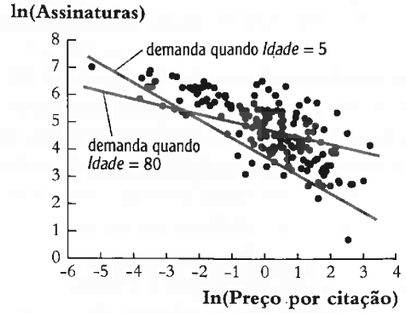
FIGURA 6.9 Assinaturas de Periódicos de Economia por Bibliotecas e Seus Preços



(a) Assinaturas e preço por citação



(b) ln(Assinaturas) e ln(Preço por citação)



(c) ln(Assinaturas) e ln(Preço por citação)

Existe uma relação inversa não-linear entre o número de assinaturas feitas por bibliotecas dos Estados Unidos (quantidade) e o preço por citação (preço), como mostra a Figura 6.9a para 180 periódicos de economia em 2000. Mas, como podemos ver na Figura 6.9b, a relação entre logaritmo da quantidade e logaritmo do preço parece ser aproximadamente linear. A Figura 6.9c mostra que a demanda é mais elástica para periódicos novos (Idade = 5 anos) do que para periódicos antigos (Idade = 80 anos).

Os resultados da regressão estão resumidos na Tabela 6.1 e produzem as seguintes conclusões (veja se você consegue encontrar a base para essas conclusões na tabela):

1. A demanda é menos elástica para os periódicos mais antigos do que para os mais novos.
2. A evidência sustenta uma função linear do logaritmo do preço em vez de uma cúbica.
3. A demanda é maior para periódicos com mais caracteres, mantendo constantes o preço e a idade.

Logo, qual é a elasticidade da demanda por periódicos de economia? Ela depende da idade do periódico. Curvas de demanda para um periódico de 80 anos para uma novata de 5 anos estão superpostas no gráfico de dispersão da Figura 6.9c; a elasticidade da demanda do periódico mais antigo é −0,28 (*EP* = 0,06) ao passo que a do periódico mais novo é −0,67 (*EP* = 0,08).

Essa demanda é muito inelástica: a demanda é pouco sensível ao preço, especialmente para periódicos mais antigos. Para as bibliotecas, ter as pesquisas mais recentes à mão é uma necessidade, e não um luxo. Para fins de comparação, os especialistas estimam que a elasticidade da demanda por cigarros esteja na faixa de −0,3 a −0,5. Os periódicos de economia viciam, aparentemente, tanto quanto os cigarros — porém, são muito mais saudáveis![6]

[6] Esses dados foram fornecidos por cortesia do professor Theodore Bergstrom do Departamento de Economia da Universidade da Califórnia, Santa Bárbara, Estados Unidos. Se você estiver interessado em conhecer mais sobre a economia dos periódicos de economia, veja Bergstrom (2001).

TABELA 6.1 Estimativa da Demanda por Períodos de Economia

Variável Dependente: Logaritmo de Assinaturas em Bibliotecas dos Estados Unidos em 2000; 180 Observações

Regressor	(1)	(2)	(3)	(4)
ln(*Preço por citação*)	−0,533** (0,034)	−0,408** (0,044)	−0,961** (0,160)	−0,899** (0,145)
[ln(*Preço por citação*)]^2			0,017 (0,025)	
[ln(*Preço por citação*)]^3			0,0037 (0,0055)	
ln(*Idade*)		0,424** (0,119)	0,373** (0,118)	0,374** (0,118)
ln(*Idade*) × ln(*Preço por citação*)			0,156** (0,052)	0,141** (0,040)
ln(*Caracteres* ÷ 1.000.000)		0,206* (0,098)	0,235* (0,098)	0,229* (0,096)
Intercepto	4,77** (0,055)	3,21** (0,38)	3,41** (0,38)	3,43** (0,38)
Estatística *F* e Estatística-Resumo				
Estatística *F* testando os coeficientes dos termos quadrático e cúbico (valor *p*)			(0,779)	0,25
EPR	0,750	0,705	0,691	0,688
$\overline{R}^2$	0,555	0,607	0,622	0,626

A estatística *F* testa a hipótese de que os coeficientes de [ln(*Preço por citação*)]^2 e [ln(*Preço por citação*)]^3 são iguais a zero. Os erros padrão estão entre parênteses abaixo dos coeficientes, e os valores *p* estão entre parênteses abaixo da estatística *F*. Os coeficientes individuais são estatisticamente significantes ao nível de *5 por cento ou **1 por cento.

Interações na Regressão Múltipla

Conceito-Chave 6.5

O termo de interação entre as duas variáveis independentes X_1 e X_2 é seu produto, $X_1 \times X_2$. A inclusão desse termo de interação permite que o efeito de uma variação em X_1 sobre Y dependa do valor de X_2 e, inversamente, permite que o efeito de uma variação em X_2 sobre Y dependa do valor de X_1.

O coeficiente de $X_1 \times X_2$ é o efeito de um aumento unitário em X_1 e X_2, maior do que a soma dos efeitos individuais de um aumento unitário somente em X_1 e somente em X_2. Isso é verdadeiro se X_1 e/ou X_2 são contínuas ou binárias.

Quando a porcentagem de alunos que está aprendendo inglês é igual à mediana ($\%AI = 8{,}85$), estima-se que a declividade da reta que relaciona a pontuação nos exames e a razão aluno-professor seja $-1{,}11$ ($= -1{,}12 + 0{,}0012 \times 8{,}85$). Quando a porcentagem de alunos aprendendo inglês está no 75º percentil ($\%AI = 23{,}0$), estima-se que a reta seja mais plana, com uma declividade de $-1{,}09$ ($= -1{,}12 + 0{,}0012 \times 23{,}0$). Isto é, para uma diretoria com 8,85 por cento de alunos aprendendo inglês, o efeito estimado de uma redução unitária na razão aluno-professor é um aumento da pontuação nos exames de 1,11 pontos; porém, para uma diretoria com 23,0 por cento de alunos aprendendo inglês, estima-se que a redução da razão aluno-professor em uma unidade aumente a pontuação nos exames em apenas 1,09 pontos. A diferença entre esses efeitos estimados não é estatisticamente significante, entretanto a estatística t que testa se o coeficiente do termo de interação é zero é $t = 0{,}0012/0{,}019 = 0{,}06$, que não é significante ao nível de 10 por cento.

Para manter a discussão centrada em modelos não-lineares, as especificações nas seções 6.1-6.3 excluem as variáveis de controle adicionais, tal como a situação econômica dos alunos. Conseqüentemente, esses resultados provavelmente estão sujeitos a um viés de omissão de variáveis. Para obter conclusões importantes quanto ao efeito da diminuição da razão aluno-professor sobre a pontuação nos exames, essas especificações não-lineares devem ser ampliadas com variáveis de controle; é para tal exercício que nos voltamos agora.

6.4 Efeitos Não-Lineares da Razão Aluno-Professor sobre a Pontuação nos Exames

Nesta seção, apontamos três questões específicas sobre a pontuação nos exames e a razão aluno-professor. A primeira seria: após o controle das diferenças entre as características econômicas nas diversas diretorias, o efeito da redução da razão aluno-professor sobre a pontuação nos exames depende da fração de alunos que está aprendendo inglês? A segunda seria: esse efeito depende do valor da razão aluno-professor? A terceira e mais importante seria: após levar em consideração os fatores econômicos e as não-linearidades, qual é o efeito estimado da redução da razão aluno-professor em dois alunos por professor sobre a pontuação nos exames, conforme nossa superintendente do Capítulo 4 se propõe a fazer?

Respondemos a essas perguntas quando consideramos as especificações de regressão não-lineares do tipo discutido nas seções 6.2 e 6.3, estendidas para incluir duas medidas da situação econômica dos alunos: a porcentagem de alunos com direito a um almoço subsidiado e o logaritmo da renda média na diretoria. O logaritmo da renda é utilizado porque a análise empírica da Seção 6.2 sugere que essa especificação capta a relação não-linear entre pontuação nos exames e na renda. Assim como na Seção 5.12, não incluímos o gasto por aluno como um regressor e, dessa forma, consideramos o efeito da redução da razão aluno-professor, o que permite que o gasto por aluno aumente (isto é, não estamos mantendo o gasto por aluno constante).

Discussão dos Resultados da Regressão

Os resultados da regressão de MQO estão resumidos na Tabela 6.2. As colunas (1)-(7) mostram regressões separadas. As entradas da tabela são coeficientes, erros padrão, determinadas estatísticas F e seus valores p e estatísticas-resumo, conforme indicado na descrição de cada linha.

A primeira coluna de resultados de regressão, chamada regressão (1) na tabela, é a regressão (4) da Tabela 5.2 — repetida aqui por conveniência. Essa regressão não controla a renda, de modo que a primeira coisa que fazemos é verificar se os resultados variam substancialmente quando o logaritmo da renda é incluído como uma variável econômica de controle adicional. Os resultados estão na regressão (2) da Tabela 6.2. O logaritmo da renda é estatisticamente significante ao nível de 1 por cento, e o coeficiente da razão aluno-professor torna-se um tanto mais próximo de zero, diminuindo de $-1{,}00$ para $-0{,}73$, embora permaneça estatisticamente significante ao nível de 1 por cento. A variação no coeficiente de *RAP* entre as regressões (1) e (2) é grande o suficiente para garantir a inclusão do logaritmo da renda nas regressões restantes como um impedimento ao viés de omissão de variáveis.

TABELA 6.2 Modelos de Regressão Não-Lineares para a Pontuação nos Exames

Variável Dependente: Pontuação Média nos Exames na Diretoria; 420 Observações

Regressor	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Razão aluno-professor (*RAP*)	−1,00** (0,27)	−0,73** (0,26)	−0,97 (0,59)	−0,53 (0,34)	64,33** (24,86)	83,70** (28,50)	65,29** (25,26)
RAP^2					−3,42** (1,25)	−4,38** (1,44)	−3,47** (1,27)
RAP^3					0,059** (0,021)	0,075** (0,024)	0,060** (0,021)
% de alunos aprendendo inglês	−0,122** (0,033)	−0,176** (0,034)					−0,166** (0,034)
% de alunos aprendendo inglês ≥10%? (Binária, *AIAlta*)			5,64 (19,51)	5,50 (9,80)	−5,47** (1,03)	816,1* (327,7)	
AIAlta × *RAP*			−1,28 (0,97)	−0,58 (0,50)		−123,3* (50,2)	
AIAlta × RAP^2						6,12* (2,54)	
AIAlta × RAP^3						−0,101* (0,043)	
% com direito a almoço subsidiado	−0,547** (0,024)	−0,398** (0,033)		−0,411** (0,029)	−0,420** (0,029)	−0,418** (0,029)	−0,402** (0,033)
Renda média na diretoria (logaritmo)		11,57** (1,81)		12,12** (1,80)	11,75** (1,78)	11,80** (1,78)	11,51** (1,81)
Intercepto	700,1** (5,6)	658,6** (8,6)	682,2** (11,9)	653,6** (9,9)	252,0 (163,6)	122,3 (185,5)	244,8 (165,7)
Estatística *F* e Valores *p* para Hipóteses Conjuntas							
(a) Todas as variáveis *RAP* e interações = 0			5,64 (0,004)	5,92 (0,003)	6,31 (<0,001)	4,96 (<0,001)	5,91 (0,001)
(b) RAP^2, $RAP^3 = 0$					6,17 (<0,001)	5,81 (0,003)	5,96 (0,003)
(c) *AIAlta* × *RAP*, *AIAlta* × RAP^2, *AIAlta* × $RAP^3 = 0$						2,69 (0,046)	
EPR	9,08	8,64	15,88	8,63	8,56	8,55	8,57
$\overline{R}^2$	0,773	0,794	0,305	0,795	0,798	0,799	0,798

Essas regressões foram estimadas utilizando dados sobre diretorias regionais de ensino K-8 na Califórnia, descritos no Apêndice 4.1. Os erros padrão estão entre parênteses abaixo dos coeficientes e os valores p estão entre parênteses abaixo da estatística F. Os coeficientes individuais são estatisticamente significantes ao nível de significância de *5 por cento ou de **1 por cento.

A regressão (3) da Tabela 6.2 é a regressão interada da Equação (6.34) com a variável binária para uma porcentagem alta ou baixa de alunos aprendendo inglês, mas sem variáveis econômicas de controle. Quando as variáveis econômicas de controle (logaritmo da renda e porcentagem com direito a almoço subsidiado) são adicionadas (veja a regressão (4) da tabela), os coeficientes mudam, porém em nenhum dos casos o coeficiente do termo de interação é significante ao nível de 5 por cento. Com base na evidência da regressão (4), a hipótese de que o efeito da *RAP* é o mesmo para diretorias com porcentagens baixas e altas de alunos aprendendo inglês não pode ser rejeitada ao nível de 5 por cento (estatística t é $t = -0,58/0,50 = -1,16$).

A regressão (5) examina se o efeito da variação da razão aluno-professor depende do valor da razão aluno-professor ao incluir uma especificação cúbica em *RAP* além de outras variáveis de controle na regressão (4) (o termo de interação, *AIAlta* × *RAP*, foi excluído porque não era significante na regressão (4) ao nível de 10 por cento). As estimativas da regressão (5) são consistentes com um efeito não-linear da razão aluno-professor. A hipótese nula de que a relação é linear é rejeitada ao nível de significância de 1 por cento contra a alternativa de que ela é cúbica (a estatística F que testa a hipótese de que os coeficientes verdadeiros de RAP^2 e RAP^3 são zero é 6,17, com um valor $p < 0,001$).

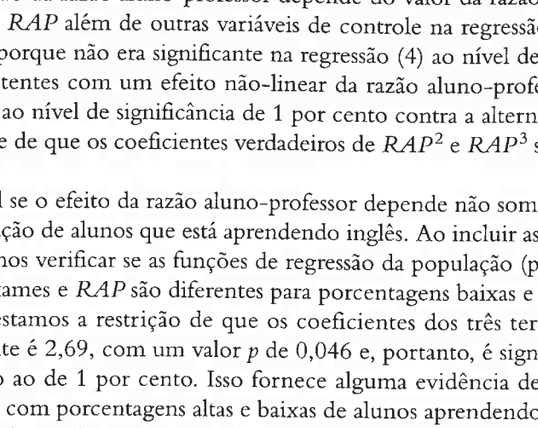
A regressão (6) examina de maneira adicional se o efeito da razão aluno-professor depende não somente do valor da razão aluno-professor, mas também da fração de alunos que está aprendendo inglês. Ao incluir as interações entre *AIAlta* e *RAP*, RAP^2 e RAP^3, podemos verificar se as funções de regressão da população (possivelmente cúbicas) que relacionam pontuação nos exames e *RAP* são diferentes para porcentagens baixas e altas de alunos aprendendo inglês. Para fazermos isso, testamos a restrição de que os coeficientes dos três termos de interação são iguais a zero. A estatística F resultante é 2,69, com um valor p de 0,046 e, portanto, é significante ao nível de significância de 5 por cento, mas não ao de 1 por cento. Isso fornece alguma evidência de que as funções de regressão são diferentes para diretorias com porcentagens altas e baixas de alunos aprendendo inglês; contudo, a comparação das regressões (6) e (4) deixa claro que essas diferenças estão associadas aos termos quadrático e cúbico.

A regressão (7) é uma modificação da regressão (5), em que a variável contínua *%AI* é utilizada em vez da variável binária *AIAlta* para controlar a porcentagem de alunos que está aprendendo inglês na diretoria. Os coeficientes dos outros regressores não variam substancialmente com essa modificação, o que indica que os resultados na regressão (5) não são sensíveis à medida da porcentagem de alunos que está aprendendo inglês efetivamente utilizada na regressão.

Em todas as especificações, a hipótese de que a razão aluno-professor não entra nas regressões é rejeitada ao nível de 1 por cento.

FIGURA 6.10 Três Funções de Regressão Relacionando Pontuação nos Exames e Razão Aluno-Professor

As regressões cúbicas das colunas (5) e (7) da Tabela 6.2 são praticamente idênticas. Elas indicam um pouco de não-linearidade na relação entre pontuação nos exames e razão aluno-professor.

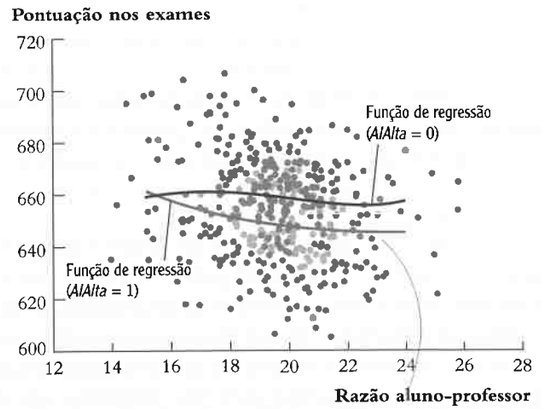


As especificações não-lineares na Tabela 6.2 são interpretadas mais facilmente na forma gráfica. A Figura 6.10 mostra o gráfico das funções de regressão estimadas relacionando a pontuação nos exames e a razão aluno-professor para a especificação linear (2) e as especificações cúbicas (5) e (7), juntamente com um gráfico de dispersão dos dados.[7] Essas funções de regressão estimadas mostram o valor previsto da pontuação nos exames como uma função da razão aluno-professor, mantendo fixos os outros valores das variáveis independentes na regressão. As funções de regressão estimadas são próximas umas das outras, embora as regressões cúbicas sejam mais planas para valores grandes da razão aluno-professor.

A regressão (6) indica uma diferença estatisticamente significante nas funções cúbicas de regressão relacionando pontuação nos exames e *RAP*, dependendo de a porcentagem de alunos aprendendo inglês na diretoria ser grande ou pequena. A Figura 6.11 mostra o gráfico dessas duas funções de regressão estimadas de modo que possamos ver se essa diferença, além de ser estatisticamente significante, é de relevância prática. Como a Figura 6.11 mostra, para razões aluno-professor entre 17 e 23 — uma gama que inclui 88 por cento das observações —, as duas funções são separadas por aproximadamente 10 pontos, mas sob outros aspectos são muito semelhantes; isto é, para *RAP* entre 17 e 23, diretorias com uma porcentagem menor de alunos aprendendo inglês se saem melhor, mantendo constante a razão aluno-professor; porém, o efeito de uma variação na razão aluno-professor é essencialmente o mesmo nos dois grupos. As duas funções de regressão são diferentes para razões aluno-professor abaixo de 16,5, mas devemos tomar cuidado para não exagerar em nossa interpretação. As diretorias com $RAP < 16,5$ constituem apenas 6 por cento das observações, de modo que as diferenças entre as funções de regressão não-lineares estão refletindo as diferenças nessas poucas diretorias com razões aluno-professor muito baixas. Portanto, com base na Figura 6.11, concluímos que o efeito de uma variação na razão aluno-professor sobre a pontuação nos exames não depende da porcentagem de alunos aprendendo inglês para a gama de razões aluno-professor na qual temos a maioria dos dados.

FIGURA 6.11 Funções de Regressão para Diretorias com Porcentagens Altas e Baixas de Alunos Aprendendo Inglês

Diretorias com porcentagens baixas de alunos aprendendo inglês (*AIAlta* = 0) são mostradas por pontos cinza-escuro e diretorias com *AIAlta* = 1 são mostradas por pontos cinza-claro. A função de regressão cúbica para *AIAlta* = 1 da regressão (6) na Tabela 6.2 está aproximadamente 10 pontos abaixo da função de regressão cúbica para *AIAlta* = 0 para $17 \leq RAP \leq 23$, mas sob outros aspectos as duas funções têm formas e inclinações semelhantes nessa gama de valores.
As declividades das funções de regressão diferem mais para valores muito grandes e muito pequenos de *RAP*, nos quais há poucas observações.



[7] Para cada curva, o valor previsto foi calculado fixando-se para cada variável independente — exceto *RAP* — o valor médio da amostra e calculando-se o valor previsto por meio da multiplicação desses valores fixos das variáveis independentes pelos respectivos coeficientes estimados da Tabela 6.2. Isso foi feito para diversos valores da *RAP* e o gráfico dos valores previstos ajustados resultantes é a reta de regressão estimada relacionando a pontuação nos exames e a *RAP*, mantendo constantes as outras variáveis em suas médias da amostra.

Resumo dos Resultados

Esses resultados nos permitem responder às três questões levantadas no início desta seção.

Após o controle da situação econômica, a existência de muitos ou poucos alunos aprendendo inglês na diretoria não tem uma influência substancial sobre o efeito de uma variação na razão aluno-professor sobre a pontuação nos exames. Nas especificações lineares, não existe evidência estatisticamente significante dessa diferença. A especificação cúbica na regressão (6) fornece uma evidência estatisticamente significante (ao nível de 5 por cento) de que as funções de regressão são diferentes para diretorias com porcentagens altas e baixas de alunos aprendendo inglês; como mostra a Figura 6.11, contudo, as funções de regressão estimadas têm declividades semelhantes na gama de razões aluno-professor que contém a maioria de nossos dados.

Após o controle da situação econômica, há evidência de um efeito não-linear da razão aluno-professor sobre a pontuação nos exames. Esse efeito é estatisticamente significante ao nível de 1 por cento (os coeficientes de RAP^2 e RAP^3 são sempre significantes ao nível de 1 por cento).

Agora podemos voltar ao problema da superintendente que iniciou o Capítulo 4. Ela quer saber o efeito da redução da razão aluno-professor em dois alunos por professor sobre a pontuação nos exames. Na especificação linear (2), esse efeito não depende da razão aluno-professor em si; o efeito estimado dessa redução é uma melhora da pontuação nos exames em 1,46 (= $-0,73 \times -2$) ponto. Nas especificações não-lineares, esse efeito depende do valor da razão aluno-professor. Se a sua diretoria atualmente possui uma razão aluno-professor de 20 e ela considera o corte para 18, então, baseado na regressão (5), o efeito estimado dessa redução é uma melhora da pontuação nos exames de 3,00 pontos, ao passo que, baseado na regressão (7), essa estimativa é de 2,93. Se a sua diretoria atualmente possui uma razão aluno-professor de 22 e considera o corte para 20, então, baseado na regressão (5), o efeito estimado dessa redução é uma melhora da pontuação nos exames de 1,93 ponto, ao passo que, baseado na regressão (7), essa estimativa é de 1,90. As estimativas das especificações não-lineares sugerem que o corte da razão aluno-professor terá um efeito um pouco maior se essa razão já for pequena.

6.5 Conclusão

Neste capítulo, apresentamos várias maneiras de modelar funções de regressão não-lineares. Como esses modelos são variantes do modelo de regressão múltipla, os coeficientes desconhecidos podem ser estimados por MQO e as hipóteses sobre seus valores podem ser testadas pelo uso das estatísticas t e F, como descrito no Capítulo 5. Nesses modelos, o efeito esperado de uma variação em uma das variáveis independentes, X_1, sobre Y, mantendo constantes as outras variáveis independentes X_2, ..., X_k, em geral depende dos valores de X_1, X_2, ..., X_k.

Existem muitos modelos neste capítulo, e por isso é normal você sentir-se um pouco confuso sobre qual deles utilizar em determinada aplicação. Como você poderia analisar possíveis não-linearidades na prática? Na Seção 6.1, expusemos um enfoque geral para tal análise, mas esse enfoque requer que você tome decisões e exercite seu julgamento ao longo do processo. Seria conveniente ter uma receita única que você pudesse seguir e que funcionasse para todas as aplicações, porém, na prática, a análise de dados raramente é simples.

O único passo mais importante na especificação de funções de regressão não-lineares é "usar a cabeça". Antes de examinar os dados, você poderia pensar em um motivo, com base na teoria econômica ou no julgamento cuidadoso, pelo qual a declividade da função de regressão da população possa depender do valor daquela, ou de outra variável independente. Se for esse o caso, que tipo de dependência você esperaria? E, ainda mais importante, quais não-linearidades (se houver) teriam implicações importantes para as questões importantes tratadas em seu estudo? Respostas cuidadosas para essas perguntas concentrarão sua análise. Na aplicação da pontuação nos exames, por exemplo, tal raciocínio nos levou a investigar se a contratação de mais professores poderia ter um efeito maior em diretorias com uma porcentagem alta de alunos aprendendo inglês, talvez porque esses alunos se beneficiariam de forma diferenciada de uma atenção mais individualizada. Formulando a pergunta de forma precisa, fomos capazes de encontrar uma resposta precisa: após controlarmos a situação econômica dos alunos, não encontramos evidências estatisticamente significantes de tal interação.

Resumo

1. Em uma regressão não-linear, a declividade da função de regressão da população depende do valor de uma ou mais das variáveis independentes.
2. O efeito de uma variação da(s) variável(is) independente(s) sobre Y pode ser calculado avaliando-se a função de regressão para dois valores da(s) variável(is) independente(s). O procedimento está resumido no Conceito-Chave 6.1.
3. Uma regressão polinomial inclui potências de X como regressores. Uma regressão quadrática inclui X e X^2; uma regressão cúbica inclui X, X^2 e X^3.
4. Pequenas variações em logaritmos podem ser interpretadas como variações proporcionais ou variações percentuais de uma variável. Regressões que envolvem logaritmos são utilizadas para estimar variações proporcionais e elasticidades.
5. O produto de duas variáveis é chamado termo de interação. Quando os termos de interação são incluídos como regressores, permitem que a declividade da regressão para uma variável dependa do valor de outra variável.

Termos-chave

modelo de regressão quadrática (135)
função de regressão não-linear (137)
modelo de regressão polinomial (140)
modelo de regressão cúbica (140)
elasticidade-preço (141)
função exponencial (142)
logaritmo natural (142)
modelo linear-log (143)
modelo log-linear (143)
modelo log-log (144)
termo de interação (149)
regressor interado (149)
modelo de regressão com interação (149)

Revisão dos Conceitos

6.1 Faça um esboço de uma função de regressão que seja crescente (que tenha declividade positiva) e mais inclinada para valores pequenos de X e menos inclinada para valores grandes. Explique como você especificaria uma regressão não-linear para modelar essa forma de curva. Você pode pensar em uma relação econômica com essa forma?

6.2 Uma função de produção "Cobb-Douglas" relaciona produção (Q) a fatores de produção, capital (K), mão-de-obra (L) e matéria-prima (M), e um termo de erro u utilizando a equação $Q = \lambda K^{\beta_1} L^{\beta_2} M^{\beta_3} e^u$, onde λ, β_1, β_2 e β_3 são parâmetros de produção. Suponha que você tenha dados sobre a produção e os fatores de produção de uma amostra aleatória de empresas com a mesma função de produção Cobb-Douglas. Como você utilizaria a análise de regressão para estimar os parâmetros de produção?

6.3 Uma função "demanda por moeda" padrão utilizada por macroeconomistas tem a forma $\ln(m) = \beta_0 + \beta_1 \ln(PIB) + \beta_2 R$, onde m é a quantidade (real) de moeda, PIB é o valor (real) do produto interno bruto e R é o valor da taxa nominal de juros medida em porcentagem ao ano. Suponha que $\beta_1 = 1,0$ e $\beta_2 = -0,02$. O que acontecerá com o valor de m se o PIB aumentar em 2 por cento? O que acontecerá com m se a taxa de juros aumentar de 4 para 5 por cento?

6.4 Você estimou um modelo de regressão linear relacionando Y e X. Seu professor diz: "Acho que a relação entre Y e X é não-linear". Explique como você testaria a adequação de sua regressão linear.

6.5 Suponha que no problema 6.2 você tenha pensado que o valor de β_2 não era constante, mas que aumentava juntamente com K. Como você poderia utilizar um termo de interação para captar esse efeito?

Exercícios

6.1 As vendas em uma companhia totalizaram US$ 196 milhões em 2001 e subiram para US$ 198 milhões em 2002.

a. Calcule o aumento percentual nas vendas utilizando a fórmula usual $100 \times \frac{Vendas_{2002} - Vendas_{2001}}{Vendas_{2001}}$. Compare esse valor à aproximação $100\,(\ln(Vendas_{2002}) - \ln(Vendas_{2001}))$.

b. Repita o item (a) supondo $Vendas_{2002} = 205$; $Vendas_{2002} = 250$; $Vendas_{2002} = 500$.

c. Em que medida a aproximação é boa quando a variação é pequena? A qualidade da aproximação se deteriora à medida que a variação percentual aumenta?

6.2 Suponha que um pesquisador colete dados sobre as casas vendidas em determinado bairro no ano passado e obtenha os resultados da regressão mostrados na tabela abaixo.

***a.** Utilizando os resultados da coluna (1), responda: qual é a variação esperada no preço de se construir uma ampliação de 500 pés quadrados (46 m^2) em uma casa? Calcule um intervalo de confiança de 95 por cento para a variação percentual no preço.

Variável Dependente: ln(*Preço*)					
Regressor	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
Tamanho	0,00042 (0,000038)				
ln(*Tamanho*)		0,69 (0,054)	0,68 (0,087)	0,57 (2,03)	0,69 (0,055)
ln(*Tamanho*)^2				0,0078 (0,14)	
Dormitórios			0,0036 (0,037)		
Piscina	0,082 (0,032)	0,071 (0,034)	0,071 (0,034)	0,071 (0,036)	0,071 (0,035)
Vista	0,037 (0,029)	0,027 (0,028)	0,026 (0,026)	0,027 (0,029)	0,027 (0,030)
Piscina × *vista*					0,0022 (0,10)
Condição	0,13 (0,045)	0,12 (0,035)	0,12 (0,035)	0,12 (0,036)	0,12 (0,035)
Intercepto	10,97 (0,069)	6,60 (0,39)	6,63 (0,53)	7,02 (7,50)	6,60 (0,40)
Estatísticas-resumo					
EPR	0,102	0,098	0,099	0,099	0,099
$\overline{R}^2$	0,72	0,74	0,73	0,73	0,73

Definições das variáveis: *Preço* = preço de venda (US$); *Tamanho* = tamanho da casa (em pés quadrados); *Dormitórios* = número de dormitórios; *Piscina* = variável binária (1 se a casa tem uma piscina, 0 se não for o caso); *Vista* = variável binária (1 se a casa tem uma linda vista, 0 se não for o caso); *Condição* = variável binária (1 se o corretor de imóveis relata que a casa se encontra em uma condição excelente, 0 se não for o caso).

b. Comparando as colunas (1) e (2), é melhor utilizar *Tamanho* ou ln(*Tamanho*) para explicar os preços das casas?

***c.** Utilizando a coluna (2), responda: qual é o efeito estimado da piscina sobre o preço?

d. A regressão na coluna (3) acrescenta o número de dormitórios à regressão. Qual é o tamanho do efeito estimado de um dormitório adicional? O efeito é estatisticamente significante? Por que você considera o efeito estimado tão pequeno? (*Dica*: Que outras variáveis estão sendo mantidas constantes?)

***e.** O termo quadrático ln(*Tamanho*)^2 é importante?

f. Utilize a regressão da coluna (5) para calcular a variação esperada no preço quando uma piscina é acrescentada a uma casa sem uma vista. Repita o exercício para uma casa com uma vista. Há uma diferença grande? A diferença é estatisticamente significante?

6.3 Após ler neste capítulo a análise da pontuação nos exames e do tamanho da turma, um educador comenta: "De acordo com a minha experiência, o desempenho do aluno depende do tamanho da turma, mas de uma forma diferente da mostrada por sua regressão. Em vez disso, os alunos vão bem quando o tamanho da turma é menor do que 20 alunos e vão muito mal quando o tamanho da turma é maior do que 25. Não há ganhos na redução do tamanho da turma para menos de 20 alunos, a relação é constante na região intermediária entre 20 e 25 alunos e não há perdas no aumento do tamanho da turma quando ele já é maior do que 25". O educador está descrevendo um "efeito limiar" em que o desempenho é constante para tamanhos de turma inferiores a 20, então há um salto e é constante para tamanhos de turma entre 20 e 25 e há outro salto para tamanhos de turma superiores a 25. Para modelar esses efeitos limiares, defina as variáveis binárias:

$RAPpequeno = 1$ se $RAP < 20$ e $RAPpequeno = 0$ nos demais casos
$RAPmoderado = 1$ se $20 \leq RAP \leq 25$ e $RAPmoderado = 0$ nos demais casos
$RAPgrande = 1$ se $RAP > 25$ e $RAPgrande = 0$ nos demais casos.

a. Considere a regressão $PontExame_i = \beta_0 + \beta_1 RAPpequeno_i + \beta_2 RAPgrande_i + u_i$. Esboce a função de regressão que relaciona *PontExame* e *RAP* para valores hipotéticos dos coeficientes da regressão que são consistentes com a declaração do educador.

b. Uma pesquisadora tenta estimar a regressão $PontExame_i = \beta_0 + \beta_1 RAPpequeno_i + \beta_2 RAPmoderado_i + \beta_3 RAPgrande_i + u_i$ e descobre que seu computador trava. Por quê?

***6.4** Explique como você utilizaria o "Enfoque nº 2"da Seção 5.8 para calcular o intervalo de confiança discutido abaixo da Equação (6.8). (*Dica*: Isso requer a estimação de uma nova regressão utilizando uma definição diferente dos regressores e da variável dependente. Veja o Exercício (5.8).)

6.5 Considere o modelo de regressão $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i$. Utilize o Conceito-Chave 6.1 para mostrar que:

a. $\frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2$(efeito da variação em X_1, mantendo X_2 constante).

b. $\frac{\Delta Y}{\Delta X_2} = \beta_2 + \beta_3 X_1$ (efeito da variação em X_2, mantendo X_1 constante).

c. Se X_1 varia em ΔX_1 *e* X_2 varia em ΔX_2, então $\Delta Y = (\beta_1 + \beta_3 X_2)\Delta X_1 + (\beta_2 + \beta_3 X_1)\Delta X_2 + \beta_3 \Delta X_1 \Delta X_2$.

CAPÍTULO 7

Avaliando Estudos Baseados na Regressão Múltipla

Os três últimos capítulos explicaram como utilizar a regressão múltipla para analisar a relação entre variáveis em uma base de dados. Neste capítulo, damos um passo para trás e perguntamos: o que torna um estudo que utiliza regressão múltipla confiável ou não? Nós nos concentramos em estudos estatísticos cujo objetivo é estimar o efeito causal de uma variação em alguma variável independente, tal como o tamanho da turma de alunos, sobre uma variável dependente, tal como a pontuação nos exames. Quando a regressão múltipla fornece uma estimativa útil do efeito causal para tais estudos e, igualmente importante, quando ela falha em fazê-lo?

Para responder a essa pergunta, este capítulo apresenta uma estrutura para avaliar estudos estatísticos em geral, independentemente de eles utilizarem regressão múltipla ou não. Essa estrutura baseia-se nos conceitos de validade interna e externa. Um estudo é válido internamente se as suas inferências estatísticas sobre os efeitos causais são válidas para a população e o cenário estudados; um estudo é válido externamente se as suas inferências podem ser generalizadas para outras populações e cenários. Nas seções 7.1 e 7.2, discutimos validade interna e externa, enumeramos um conjunto de ameaças possíveis a essas validades e discutimos como identificar aquelas ameaças na prática. Algumas daquelas ameaças não podem ser tratadas utilizando as ferramentas econométricas apresentadas até aqui; este capítulo oferece uma visão geral dos métodos, estudados nos capítulos restantes deste livro, para tratar as ameaças.

Como uma ilustração da estrutura de validade interna e externa, na Seção 7.3 avaliamos a validade interna e externa do estudo do efeito do corte da razão aluno-professor sobre a pontuação nos exames apresentada nos capítulos 4-6.

7.1 Validade Interna e Validade Externa

Os conceitos de validade interna e validade externa, definidos no Conceito-Chave 7.1, fornecem uma estrutura para avaliar se um estudo estatístico ou econométrico é útil para responder a uma questão específica de interesse.

Validade interna e validade externa distinguem entre população e cenário estudados e população e cenário para os quais os resultados são generalizados. A **população estudada** é a população de entidades — pessoas, empresas, diretorias regionais de ensino e assim por diante — da qual a amostra foi selecionada. A população para a qual os resultados são generalizados, ou a **população de interesse**, é a população de entidades para a qual as inferências causais do estudo serão aplicadas. Por exemplo, o diretor de uma escola de ensino médio pode querer generalizar nossos resultados sobre tamanhos de turma e pontuação nos exames do ensino fundamental das diretorias regionais de ensino da Califórnia (a população estudada) para a população de escolas de ensino médio (a população de interesse).

Por "cenário" entende-se o ambiente institucional, legal, social e econômico. Por exemplo, seria importante saber se os resultados de um experimento de laboratório que avalia métodos para o crescimento de tomates orgânicos poderiam ser generalizados para o campo, isto é, se os métodos orgânicos que funcionam no cenário de um laboratório também funcionam no cenário do mundo real. Fornecemos outros exemplos de diferenças em populações e cenários mais adiante nesta seção.

Ameaças à Validade Interna

A validade interna possui dois componentes. Em primeiro lugar, o estimador do efeito causal deveria ser não viesado e consistente. Por exemplo, suponha que $\hat{\beta}_{RAP}$ seja o estimador de MQO do efeito de uma variação unitária na razão aluno-professor sobre a pontuação nos exames em uma dada regressão; então, $\hat{\beta}_{RAP}$ deve ser um estimador não viesado e consistente do verdadeiro efeito causal da população resultante de uma variação na razão aluno-professor, β_{RAP}.

Validade Interna e Validade Externa

Conceito-Chave 7.1

Uma análise estatística é **válida internamente** se as inferências estatísticas sobre os efeitos causais são válidas para a população estudada. A análise é **válida externamente** se as suas inferências e conclusões puderem ser generalizadas com base na população e no cenário estudados para outras populações e cenários.

Em segundo lugar, os testes de hipótese deveriam ter o nível de significância desejado (a taxa de rejeição efetiva do teste sob a hipótese nula deveria ser igual ao nível de significância desejado) e os intervalos de confiança deveriam ter o nível de confiança desejado. Por exemplo, se um intervalo de confiança é construído como $\hat{\beta}_{RAP} \pm 1{,}96EP(\hat{\beta}_{RAP})$, deveria conter o verdadeiro efeito causal da população, β_{RAP}, com probabilidade de 95 por cento entre as amostras repetidas.

Na análise de regressão, os efeitos causais são estimados utilizando a função de regressão estimada e os testes de hipótese são conduzidos utilizando os coeficientes da regressão estimada e seus erros padrão. Portanto, os requisitos para a validade interna em um estudo baseado em regressão de MQO são os seguintes: que o estimador de MQO seja não viesado e consistente e que os erros padrão sejam calculados de maneira que os intervalos de confiança tenham o nível de confiança desejado. Há vários motivos para que isso não aconteça, os quais constituem ameaças à validade interna. Essas ameaças levam a violações de uma ou mais das hipóteses de mínimos quadrados do Conceito-Chave 5.4. Por exemplo, uma ameaça que discutimos em detalhe é o viés de omissão de variáveis; ele leva a uma correlação entre um ou mais regressores e o termo de erro, o que viola a primeira hipótese de mínimos quadrados. Se os dados sobre a variável omitida estiverem disponíveis, então essa ameaça poderá ser evitada pela inclusão daquela variável como um regressor adicional.

Na Seção 7.2 há uma discussão detalhada das diversas ameaças à validade interna na análise de regressão múltipla e da forma de eliminá-las.

Ameaças à Validade Externa

Ameaças potenciais à validade externa surgem das diferenças entre a população e o cenário estudados e a população e o cenário de interesse.

Diferenças em populações. Diferenças entre a população estudada e a população de interesse podem representar uma ameaça à validade externa. Por exemplo, estudos laboratoriais sobre os efeitos tóxicos de produtos químicos normalmente utilizam populações de animais como ratos (a população estudada), mas os resultados são utilizados para a elaboração de normas de saúde e segurança para populações humanas (a população de interesse). O fato de ratos e homens serem suficientemente diferentes para ameaçar a validade externa de tais estudos é uma questão polêmica.

De forma mais geral, o verdadeiro efeito causal pode não ser o mesmo na população estudada e na população de interesse. Isso porque a população pode ter sido escolhida de um modo que a torna diferente da população de interesse em virtude de diferenças nas características das populações, de diferenças geográficas ou ainda porque o estudo está obsoleto.

Diferenças em cenários. Ainda que a população estudada e a população de interesse sejam idênticas, generalizar os resultados do estudo pode não ser possível se os **cenários** forem diferentes. Por exemplo, um estudo do efeito de uma campanha publicitária contra o consumo abusivo de álcool sobre a embriaguez na universidade não pode ser generalizado para outro grupo idêntico de universitários se a idade permitida por lei para o

consumo de bebidas alcoólicas nas duas universidades é diferente. Nesse caso, o cenário legal em que o estudo foi conduzido difere daquele em que seus resultados são aplicados.

De modo mais geral, exemplos de diferenças em cenários incluem diferenças no ambiente institucional (universidades públicas *versus* universidades religiosas), diferenças na legislação (diferenças na idade permitida por lei) ou diferenças no ambiente físico (embriaguez em festa no sul da Califórnia *versus* Fairbanks, Alasca).

Aplicação ao caso de pontuação nos exames e razão aluno-professor. Os capítulos 5 e 6 relataram melhorias estimadas estatisticamente significantes, mas bastante pequenas, da pontuação nos exames como resultado da redução na razão aluno-professor. Essa análise se baseou nos resultados de exames para as diretorias regionais de ensino da Califórnia. Suponha por ora que esses resultados sejam válidos internamente. Para quais outras populações e cenários de interesse eles poderiam ser generalizados?

Quanto mais próximos a população e o cenário estudados estiverem da população e do cenário de interesse, mais fortes serão os argumentos para a validade externa. Por exemplo, alunos universitários e seu curso são muito diferentes de alunos de escolas de ensino fundamental e seu curso, de modo que é implausível que o efeito da redução no tamanho das turmas, estimado utilizando os dados do ensino fundamental das diretorias regionais de ensino da Califórnia, seja generalizado para as universidades. Por outro lado, alunos, currículo e organização do ensino fundamental são muito semelhantes por todos os Estados Unidos, de modo que é plausível que os resultados da Califórnia possam ser generalizados para o desempenho em exames padronizados do ensino fundamental de outras diretorias regionais de ensino nesse país.

Como avaliar a validade externa de um estudo. A validade externa deve ser considerada utilizando o conhecimento específico de populações e cenários estudados e de populações e cenários de interesse. Diferenças importantes entre ambos lançarão dúvidas sobre a validade externa do estudo.

Às vezes há dois ou mais estudos sobre populações diferentes, mas relacionadas. Se for esse o caso, a validade externa desses estudos pode ser verificada pela comparação de seus resultados. Por exemplo, na Seção 7.3, analisamos os dados sobre pontuação nos exames e tamanho da turma para o ensino fundamental nas diretorias regionais de ensino de Massachusetts e comparamos esses resultados com os da Califórnia. Em geral, resultados semelhantes em dois ou mais estudos sustentam o direito à validade externa, ao passo que diferenças nos resultados lançam dúvidas sobre sua validade externa.[1]

Como desenhar um estudo válido externamente. Como as ameaças à validade externa originam-se de uma falta de comparabilidade de populações e cenários, essas ameaças são minimizadas da melhor forma nos estágios iniciais de um estudo, antes de os dados serem coletados. O desenho de um estudo foge ao escopo deste livro; o leitor interessado pode consultar Shadish, Cook e Campbell (2002).

7.2 Ameaças à Validade Interna na Análise de Regressão Múltipla

Estudos baseados na análise de regressão são válidos internamente se os coeficientes da regressão estimada são não viesados e consistentes e se os seus erros padrão produzem intervalos de confiança ao nível de confiança desejado. Nesta seção, pesquisamos cinco motivos pelos quais o estimador de MQO dos coeficientes da regressão múltipla podem ser viesados, mesmo em amostras grandes: variáveis omitidas, erro de especificação da forma funcional da função de regressão, medida imprecisa das variáveis independentes ("erros nas variáveis"), seleção da amostra e causalidade simultânea. Todas as fontes de viés surgem porque o regressor está correlacionado com o

[1] Uma comparação de diversos estudos relacionados sobre o mesmo tópico é chamada de metanálise. A discussão do quadro sobre o "efeito Mozart" no Capítulo 5, por exemplo, baseia-se em uma metanálise. A realização de uma metanálise com base em vários estudos apresenta seus próprios desafios. Como você separa os estudos bons dos ruins? Como você compara estudos quando as variáveis dependentes diferem? Você deveria dar mais importância a um estudo grande em relação a um estudo pequeno? Uma discussão da metanálise e seus desafios foge ao escopo deste livro. O leitor interessado pode consultar Hedges e Olkin (1985) e Cooper e Hedges (1994).

termo de erro na regressão da população, violando a primeira hipótese de mínimos quadrados do Conceito-Chave 5.4. Para cada uma, discutimos o que pode ser feito para reduzir esse viés. A seção termina com uma discussão das circunstâncias que levam a erros padrão inconsistentes e o que pode ser feito com relação a isso.

Viés de Omissão de Variáveis

Lembre-se de que o viés de omissão de variáveis surge quando uma variável que tanto determina Y quanto é correlacionada com um ou mais dos regressores incluídos é omitida da regressão. Esse viés persiste mesmo em amostras grandes, de modo que o estimador de MQO é inconsistente. A melhor forma de minimizar o viés de omissão de variáveis depende da disponibilidade de dados para a variável omitida potencial.

Soluções para o viés de omissão de variáveis quando a variável omitida é observada. Se você dispõe de dados para a variável omitida, pode incluí-la em uma regressão múltipla e, desse modo, atacar o problema. Contudo, a adição de uma nova variável tem custos e benefícios. Por um lado, a omissão da variável poderia resultar em um viés de omissão de variáveis. Por outro, a inclusão da variável quando ela não pertence à regressão (isto é, quando seu coeficiente de regressão da população é igual a zero) reduz a precisão dos estimadores dos outros coeficientes da regressão. Em outras palavras, a decisão de incluir ou não uma variável envolve um dilema entre viés e variância dos coeficientes de interesse. Na prática, há quatro passos que podem ajudá-lo a decidir se você deve ou não incluir uma variável ou um conjunto de variáveis em uma regressão.

O primeiro passo é identificar os principais coeficientes de interesse em sua regressão. Nas regressões de pontuação nos exames, trata-se do coeficiente da razão aluno-professor, uma vez que a questão colocada originalmente refere-se ao efeito de uma redução nessa razão sobre a pontuação nos exames.

O segundo passo é perguntar-se: Quais são as fontes mais prováveis de um importante viés de omissão de variáveis nessa regressão? A resposta requer a aplicação da teoria econômica e um conhecimento profundo, e deveria ocorrer antes de você estimar quaisquer regressões; como isso é feito antes da análise dos dados, é chamado de raciocínio *a priori* ("antes do fato"). No exemplo da pontuação nos exames, esse passo envolve a identificação dos determinantes da pontuação nos exames que, se ignorados, poderiam tornar viesado nosso estimador do efeito do tamanho da turma. O resultado desse passo é uma especificação de regressão base, o ponto de partida para sua análise de regressão empírica, e uma lista com variáveis "questionáveis" adicionais que podem ajudar a diminuir o possível viés de omissão de variáveis.

O terceiro passo é ampliar sua especificação de base com as variáveis questionáveis adicionais identificadas no segundo passo e testar as hipóteses de que seus coeficientes são iguais a zero. Se os coeficientes das variáveis adicionais forem estatisticamente significantes ou se os coeficientes de interesse estimados mudarem consideravelmente quando as variáveis adicionais forem incluídas, então elas deverão permanecer na especificação e você deverá modificar sua regressão básica. Caso contrário, essas variáveis poderão ser excluídas da regressão.

O quarto passo é apresentar um resumo preciso de seus resultados na forma tabular. Isso oferece "total transparência" a um cético potencial, que pode então tirar suas próprias conclusões. As tabelas 5.2 e 6.2 são exemplos dessa estratégia. Por exemplo, na Tabela 6.2 poderíamos ter apresentado apenas a regressão na coluna (7), uma vez que ela resume os efeitos e as não-linearidades relevantes das outras regressões da tabela. A apresentação das outras regressões, contudo, permite ao leitor cético tirar suas próprias conclusões.

Esses passos estão resumidos no Conceito-Chave 7.2.

Soluções para o viés de omissão de variáveis quando a variável omitida não é observada. A adição de uma variável omitida a uma regressão não é uma opção se você não dispõe de dados sobre aquela variável. Ainda assim, há três outros modos de resolver o problema do viés de omissão de variáveis. Cada uma dessas três soluções contorna esse viés por meio da utilização de tipos diferentes de dados.

A primeira solução é utilizar dados em que a mesma unidade de observação é analisada em pontos diferentes no tempo. Por exemplo, a pontuação nos exames e os dados a ela relacionados podem ser coletados para as mesmas diretorias em 1995 e novamente em 2000. Os dados nessa forma são chamados de dados de painel. Conforme explicado no Capítulo 8, os dados de painel tornam possível o controle de variáveis omitidas não observadas, desde que elas não variem ao longo do tempo.

A segunda solução é utilizar a regressão de variáveis instrumentais. Esse método se apóia em uma nova variável, chamada de variável instrumental. A regressão de variáveis instrumentais será discutida no Capítulo 10.

Conceito-Chave 7.2

Devo Incluir Mais Variáveis em Minha Regressão?

Se você inclui outra variável em sua regressão múltipla, elimina a possibilidade de viés de omissão de variáveis resultante da exclusão daquela variável, porém a variância do estimador dos coeficientes de interesse pode aumentar. Seguem-se algumas diretrizes que podem ajudá-lo a decidir se deve incluir uma variável adicional:

1. Seja específico com relação aos(s) coeficiente(s) de interesse.
2. Use um raciocínio *a priori* para identificar as fontes potenciais mais importantes de viés de omissão de variáveis, que leve a uma especificação de base e a algumas variáveis "questionáveis".
3. Teste se as variáveis questionáveis adicionais têm coeficientes diferentes de zero.
4. Forneça tabulações representativas "totalmente transparentes" de seus resultados de modo que outros vejam o efeito da inclusão das variáveis questionáveis sobre o(s) coeficiente(s) de interesse. Seus resultados mudarão se você incluir uma variável questionável?

A terceira solução é utilizar um projeto de estudo no qual o efeito de interesse (por exemplo, o efeito da redução do tamanho da turma sobre os resultados do aluno) é estudado por meio da utilização de um experimento controlado aleatório. Esses experimentos serão discutidos no Capítulo 11.

Erro de Especificação da Forma Funcional da Função de Regressão

Se a verdadeira função de regressão da população for não-linear, mas a regressão estimada for linear, então esse **erro de especificação da forma funcional** torna o estimador de MQO viesado. Esse viés é um tipo de viés de omissão de variáveis, em que as variáveis omitidas são os termos que refletem os aspectos não-lineares ausentes da função de regressão. Por exemplo, se a função de regressão da população for um polinômio quadrático, então uma regressão que omita o quadrado da variável independente terá viés de omissão de variáveis.

Soluções para o erro de especificação da forma funcional. Quando a variável dependente é contínua (como a pontuação nos exames), esse problema da não-linearidade potencial pode ser resolvido com a utilização dos métodos do Capítulo 6. Se, contudo, essa variável é discreta ou binária (por exemplo, Y_i é igual a um se a i-ésima pessoa tem curso superior e igual a zero nos demais casos), as coisas tornam-se mais complicadas. A regressão com uma variável dependente discreta será discutida no Capítulo 9.

Erros nas Variáveis

Suponha que em nossa regressão de pontuação nos exames contra a razão aluno-professor tenhamos confundido sem querer nossos dados, de modo que acabamos regredindo a pontuação nos exames para alunos da 5ª série sobre a razão aluno-professor para alunos da 8ª série naquela diretoria. Embora as razões aluno-professor para alunos do ensino fundamental e para alunos da 8ª série possam ser correlacionadas, elas não são iguais, de modo que essa confusão levaria a um viés no coeficiente estimado. Esse é um exemplo de **viés de erros nas variáveis** porque sua fonte é um erro na medida da variável independente. Esse viés persiste mesmo em amostras muito grandes, de modo que o estimador de MQO é inconsistente se há erro de medida.

Há muitas fontes possíveis de erro de medida. Se os dados são coletados por meio de uma pesquisa, um entrevistado pode dar a resposta errada. Por exemplo, uma pergunta do Current Population Survey envolve o salário do ano anterior. Um entrevistado pode não saber o seu salário exato ou pode informar um valor errado por qualquer outro motivo. Se, por outro lado, os dados são obtidos de registros administrativos computadorizados, pode ter havido erros de digitação quando eles entraram no sistema pela primeira vez.

Para verificar que o viés de erros nas variáveis resulta em uma correlação entre o regressor e o termo de erro, suponha que haja um único regressor X_i (por exemplo, a renda efetiva), mas que seja medido de forma imprecisa por $\tilde{X}_i$ (a estimativa da renda do entrevistado). Como a variável observada é $\tilde{X}_i$, e não X_i, a equação da regressão efetivamente estimada é aquela baseada em $\tilde{X}_i$. A equação de regressão da população $Y_i = \beta_0 + \beta_1 X_i + u_i$, escrita em termos da variável medida de forma imprecisa $\tilde{X}_i$, é

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \tilde{X}_i + [\beta_1 (X_i - \tilde{X}_i) + u_i] \\ &= \beta_0 + \beta_1 \tilde{X}_i + v_i, \end{aligned} \tag{7.1}$$

onde $v_i = \beta_1 (X_i - \tilde{X}_i) + u_i$. Portanto, a equação de regressão da população escrita em termos de $\tilde{X}_i$ possui um termo de erro que contém a diferença entre X_i e $\tilde{X}_i$. Se essa diferença for correlacionada com o valor medido $\tilde{X}_i$, então o regressor $\tilde{X}_i$ será correlacionado com o termo de erro e $\hat{\beta}_1$ será viesado e inconsistente.

O tamanho preciso e a direção do viés em $\hat{\beta}_1$ dependem da correlação entre $\tilde{X}_i$ e $(X_i - \tilde{X}_i)$. Essa correlação, por sua vez, depende da natureza específica do erro de medida.

Por exemplo, suponha que o entrevistado na pesquisa forneça seu melhor palpite ou lembrança do valor efetivo da variável independente X_i. Uma forma conveniente de representar isso matematicamente é supor que o valor medido de X_i seja igual ao valor efetivo, não medido, somado a um componente puramente aleatório, w_i. Portanto, o valor medido da variável, representado por $\tilde{X}_i$, é $\tilde{X}_i = X_i + w_i$. Como o erro é puramente aleatório, podemos supor que w_i tem média zero e variância σ_w^2 e é não-correlacionado com X_i e com o erro da regressão u_i. Sob essa hipótese, um pouco de álgebra[2] mostra que $\hat{\beta}_1$ tem o limite de probabilidade

$$\hat{\beta}_1 \xrightarrow{p} \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2} \beta_1. \tag{7.2}$$

Isto é, se o efeito da imprecisão de medida consiste simplesmente na adição de um elemento aleatório ao valor efetivo da variável independente, então $\hat{\beta}_1$ é inconsistente. Como a razão $\frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}$ é menor do que um, $\hat{\beta}_1$ será viesado em direção a zero, mesmo em amostras grandes. No caso extremo em que o erro de medida é tão grande que essencialmente nenhuma informação sobre X_i permanece, a razão entre as variâncias na expressão final da Equação (7.2) é zero e $\hat{\beta}_1$ converge em probabilidade para zero. No outro extremo, quando não há erro de medida, $\sigma_w^2 = 0$, logo $\hat{\beta}_1 \xrightarrow{p} \beta_1$.

Embora o resultado na Equação (7.2) seja específico para esse tipo particular de erro de medida, ele ilustra a proposição mais geral de que, se a variável independente é medida de forma imprecisa, então o estimador de MQO é viesado, mesmo em amostras grandes. O Conceito-Chave 7.3 resume o viés de erros nas variáveis.

Soluções para o viés de erros nas variáveis. A melhor forma de resolver o problema de erros nas variáveis é obter uma medida precisa de X. Se for impossível, contudo, há métodos econométricos que podem ser utilizados para diminuir o viés de erros nas variáveis.

Um desses métodos é a regressão de variáveis instrumentais. Isso depende de haver outra variável (a variável "instrumental") correlacionada ao valor efetivo X_i, mas não-correlacionada ao erro de medida. Esse método será estudado no Capítulo 10.

Um segundo método é o desenvolvimento de um modelo matemático do erro de medida e, se possível, a utilização das fórmulas resultantes para ajustar as estimativas. Por exemplo, se uma pesquisadora acredita que a variável medida é, na verdade, a soma do valor efetivo e de um termo de erro de medida aleatório e se ela conhece ou pode estimar a razão σ_w^2 / σ_X^2, então pode utilizar a Equação (7.2) para calcular um estimador de β_1 que corrija o viés para baixo. Como esse enfoque requer conhecimento especializado sobre a natureza do erro de medida, os detalhes normalmente são específicos para dada base de dados e seus problemas de medida e não prosseguiremos com esse enfoque.

[2] Sob essa hipótese de erro de medida, $v_i = \beta_1 (X_i - \tilde{X}_i) + u_i = -\beta_1 w_i + u_i$, $\text{cov}(\tilde{X}_i, u_i) = 0$ e $\text{cov}(\tilde{X}_i, w_i) = \text{cov}(X_i + w_i, w_i) = \sigma_w^2$, logo $\text{cov}(\tilde{X}_i, v_i) = -\beta_1 \text{cov}(\tilde{X}_i, w_i) + \text{cov}(\tilde{X}_i, u_i) = -\beta_1 \sigma_w^2$. Portanto, a partir da Equação (5.1), $\hat{\beta}_1 \xrightarrow{p} \beta_1 - \beta_1 \sigma_w^2 / \sigma_{\tilde{X}}^2$. Agora $\sigma_{\tilde{X}}^2 = \sigma_X^2 + \sigma_w^2$, então $\hat{\beta}_1 \xrightarrow{p} \beta_1 - \beta_1 \sigma_w^2 / (\sigma_X^2 + \sigma_w^2) = [\sigma_X^2 / (\sigma_X^2 + \sigma_w^2)] \beta_1$.

Viés de Erros nas Variáveis

Conceito-Chave 7.3

O viés de erros nas variáveis no estimador de MQO surge quando uma variável independente é medida de forma imprecisa. Esse viés depende da natureza do erro de medida e persiste mesmo que o tamanho da amostra seja grande. Se a variável medida é igual à variável efetiva mais um termo de erro de medida independentemente distribuído com média zero, então o estimador de MQO em uma regressão com uma única variável do lado direito é viesado em direção a zero e seu limite de probabilidade é dado na Equação (7.2).

Seleção da Amostra

O **viés de seleção da amostra** ocorre quando a disponibilidade dos dados é influenciada por um processo de seleção relacionado ao valor da variável dependente. Esse processo pode introduzir uma correlação entre o termo de erro e o regressor, o que leva a um viés no estimador de MQO.

A seleção da amostra que não está relacionada ao valor da variável dependente não introduz viés. Por exemplo, se os dados são coletados de uma população por amostragem aleatória simples, o método de amostragem (a população ao acaso) não tem relação nenhuma com o valor da variável dependente. Tal amostragem não introduz viés.

O viés pode ser introduzido quando o método de amostragem está relacionado ao valor da variável dependente. Um exemplo de viés de seleção da amostra em votações foi dado no quadro do Capítulo 2. Naquele exemplo, o método de seleção da amostra (números de telefone de proprietários de automóveis selecionados aleatoriamente) estava relacionado com a variável dependente (quem o indivíduo apoiava na eleição para presidente dos Estados Unidos em 1936), uma vez que em 1936 os proprietários de automóveis que possuíam telefone eram muito provavelmente republicanos.

Um exemplo de seleção da amostra em economia surge da utilização de uma regressão de salários sobre instrução para estimar o efeito de um ano adicional de instrução sobre os salários. Por definição, somente os indivíduos que possuem um emprego têm salário. Os fatores (observáveis e não observáveis) que determinam o fato de uma pessoa ter um emprego — instrução, experiência, domicílio, capacidade, sorte e assim por diante — são semelhantes aos fatores que determinam o quanto essa pessoa recebe quando está empregada. Assim, o fato de alguém ter um emprego sugere, mantendo tudo o mais constante, que o termo de erro na equação de salário para aquela pessoa é positivo. Dito de outra forma, o fato de alguém ter um emprego ou não é em parte determinado pelas variáveis omitidas no termo de erro da regressão de salário. Portanto, o simples fato de alguém ter um emprego — e assim aparece na base de dados — fornece informações de que o termo de erro na regressão é positivo, ao menos na média, e que poderia ser correlacionado com os regressores. Isso também pode levar a um viés no estimador de MQO.

O Conceito-Chave 7.4 resume o viés de seleção da amostra.

Soluções para o viés de seleção. Os métodos que discutimos até o momento não eliminam o viés de seleção da amostra. Os métodos para a estimação de modelos com seleção da amostra fogem ao escopo deste livro. Esses métodos baseiam-se nas técnicas que serão apresentadas no Capítulo 9, ocasião em que serão fornecidas referências adicionais.

Causalidade Simultânea

Até agora, supusemos que a causalidade vai dos regressores para a variável dependente (X causa Y). E se a causalidade também vai da variável dependente para um ou mais regressores (Y causa X)? Se for esse o caso, a causalidade vai para trás e para a frente, isto é, há **causalidade simultânea**. Se ela existe, uma regressão de MQO capta ambos os efeitos, de modo que o estimador de MQO é viesado e inconsistente.

Viés de Seleção da Amostra

Conceito-Chave 7.4

O viés de seleção da amostra surge quando um processo de seleção influencia a disponibilidade de dados e tal processo está relacionado com a variável dependente. A seleção da amostra induz a uma correlação entre um ou mais regressores e o termo de erro, o que leva a um viés e à inconsistência do estimador de MQO.

Por exemplo, nosso estudo sobre pontuação nos exames se concentrou no efeito da redução da razão aluno-professor sobre a pontuação, de modo que se presume que a causalidade deva ir da razão aluno-professor para a pontuação. Suponha, contudo, que uma iniciativa governamental tenha subsidiado a contratação de professores em diretorias regionais de ensino com baixa pontuação nos exames. Se fosse esse o caso, a causalidade iria para ambos os sentidos: pelos motivos pedagógicos usuais, as razões aluno-professor baixas provavelmente levam a uma alta pontuação nos exames; porém, em razão do programa do governo, a baixa pontuação nos exames levaria a razões aluno-professor baixas.

A causalidade simultânea leva a uma correlação entre o regressor e o termo de erro. No exemplo da pontuação nos exames, suponha que haja um fator omitido que leve a uma baixa pontuação; em virtude do programa do governo, esse fator que gera baixa pontuação resulta, por sua vez, em uma razão aluno-professor baixa. Portanto, um termo de erro negativo na regressão da população da pontuação nos exames sobre a razão aluno-professor diminui a pontuação, mas, em virtude do programa do governo, leva também a uma redução da razão aluno-professor. Em outras palavras, a razão é positivamente correlacionada com o termo de erro na regressão populacional. Isso, por sua vez, leva a um viés de causalidade simultânea e à inconsistência do estimador de MQO.

Essa correlação entre o termo de erro e o regressor pode ser expressa matematicamente pela introdução de uma equação adicional que descreva a ligação causal inversa. Por conveniência, considere apenas as duas variáveis X e Y e ignore outros possíveis regressores. Dessa forma, há duas equações, uma em que X causa Y e outra em que Y causa X:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \text{ e} \tag{7.3}$$

$$X_i = \gamma_0 + \gamma_1 Y_i + \nu_i. \tag{7.4}$$

A Equação (7.3) é aquela familiar em que β_1 é o efeito de uma variação em X sobre Y, onde u representa outros fatores. A Equação (7.4) representa o efeito causal inverso de Y sobre X. No problema da pontuação nos exames, a Equação (7.3) representa o efeito pedagógico do tamanho da turma sobre a pontuação nos exames, ao passo que a Equação (7.4) representa o efeito causal inverso da pontuação nos exames sobre o tamanho da turma induzido pelo programa do governo.

A causalidade simultânea leva a uma correlação entre X_i e o termo de erro u_i na Equação (7.3). Para visualizar isso, imagine que u_i seja negativo, o que diminui Y_i. Contudo, o valor menor de Y_i afeta o valor de X_i por meio da segunda dessas equações e, se γ_1 for positivo, um valor baixo de Y_i levará a um valor baixo de X_i. Portanto, se γ_1 for positivo, X_i e u_i serão positivamente correlacionados.[3]

Como isso pode ser expresso matematicamente utilizando-se um sistema de duas equações simultâneas, o viés de causalidade simultânea às vezes é chamado de **viés de equações simultâneas**. O viés de causalidade simultânea está resumido no Conceito-Chave 7.5.

[3] Para mostrar isso matematicamente, observe que a Equação (7.4) implica que $\text{cov}(X_i, u_i) = \text{cov}(\gamma_0 + \gamma_1 Y_i + \nu_i, u_i) = \gamma_1 \text{cov}(Y_i, u_i) + \text{cov}(\nu_i, u_i)$. Supondo que $\text{cov}(\nu_i, u_i) = 0$, pela Equação (7.3), isso implica que $\text{cov}(X_i, u_i) = \gamma_1 \text{cov}(Y_i, u_i) = \gamma_1 \text{cov}(\beta_0 + \beta_1 X_i + u_i, u_i) = \gamma_1 \beta_1 \text{cov}(X_i, u_i) + \gamma_1 \sigma_u^2$. Resolvendo $\text{cov}(X_i, u_i)$, chegamos ao resultado $\text{cov}(X_i, u_i) = \gamma_1 \sigma_u^2 / (1 - \gamma_1 \beta_1)$.

Conceito-Chave 7.5

Viés de Causalidade Simultânea

O viés de causalidade simultânea, também chamado de viés de equações simultâneas, surge em uma regressão de Y sobre X quando, além da ligação causal de interesse de X para Y, há uma ligação causal de Y para X. Essa causalidade inversa faz com que X esteja correlacionado com o termo de erro na regressão da população de interesse.

Soluções para o viés de causalidade simultânea. Há duas maneiras de diminuir o viés de causalidade simultânea. Uma é utilizar a regressão de variáveis instrumentais, o tópico do Capítulo 10. A segunda é projetar e implementar um experimento controlado aleatório em que o canal de causalidade inversa é anulado; tais experimentos serão discutidos no Capítulo 11.

Fontes de Inconsistência dos Erros Padrão de MQO

Erros padrão inconsistentes representam uma ameaça diferente à validade interna. Mesmo que o estimador de MQO seja consistente e a amostra seja grande, erros padrão inconsistentes produzem testes de hipótese com tamanho que difere do nível de significância desejado e intervalos de confiança de "95 por cento" que deixam de incluir o valor verdadeiro em 95 por cento das amostras repetidas.

Há dois motivos principais pelos quais os erros padrão são inconsistentes: tratamento inadequado da heteroscedasticidade e correlação do termo de erro entre observações.

Heteroscedasticidade. Conforme discutido na Seção 4.9, por motivos históricos, alguns pacotes de regressão relatam erros padrão somente homoscedásticos. Se, contudo, o erro da regressão é heteroscedástico, aqueles erros padrão não constituem uma base confiável para testes de hipótese e intervalos de confiança. A solução para esse problema é utilizar erros padrão robustos quanto à heteroscedasticidade e construir estatísticas F utilizando um estimador de variância robusto quanto à heteroscedasticidade. Erros padrão desse tipo são fornecidos como uma opção em pacotes modernos.

Correlação do termo de erro entre observações. Em alguns cenários, o termo de erro da população pode estar correlacionado ao longo das observações. Isso não acontecerá se os dados forem obtidos por amostragem ao acaso da população porque a aleatoriedade do processo de amostragem assegura que os erros sejam independentemente distribuídos de uma observação para a seguinte. Às vezes, contudo, a amostragem é apenas parcialmente aleatória. A circunstância mais comum ocorre quando os dados são observações repetidas da mesma entidade ao longo do tempo, por exemplo, a mesma diretoria regional de ensino para diversos anos. Se as variáveis omitidas que formam o erro da regressão são persistentes (como a demografia da diretoria), isso induz a uma correlação "serial" desse erro ao longo do tempo. Outro exemplo é aquele em que uma amostragem baseia-se em uma unidade geográfica. Se há variáveis omitidas que reflitam influências geográficas, essas variáveis podem resultar na correlação dos erros de regressão para observações adjacentes.

A correlação do erro de regressão entre observações não torna o estimador de MQO viesado ou inconsistente, mas viola a segunda hipótese de mínimos quadrados do Conceito-Chave 5.4. A partir disso inferimos que os erros padrão de MQO — somente os homoscedásticos *e* os robustos quanto à heteroscedasticidade — são incorretos no sentido de que não produzem intervalos de confiança com o nível de confiança desejado.

Em muitos casos, esse problema pode ser consertado pela utilização de uma fórmula alternativa para erros padrão. Fornecemos tal fórmula para o cálculo de erros padrão robustos quanto à heteroscedasticidade e quanto à correlação serial na discussão sobre regressão com dados de séries temporais no Capítulo 12.

7.3 Exemplo: Pontuação nos Exames e Tamanho da Turma

A estrutura de validade interna e validade externa nos ajuda a fazer um exame crítico do que aprendemos — e do que não aprendemos — em nossa análise dos dados da Califórnia sobre pontuação nos exames.

Validade Externa

O fato de a análise da Califórnia poder ser generalizada — isto é, se ela é válida externamente — depende da população e do cenário para o qual a generalização é feita. Aqui, consideramos se os resultados podem ser generalizados para o desempenho em outros exames padronizados e para o ensino público fundamental em outras diretorias regionais de ensino dos Estados Unidos.

Na Seção 7.1, você viu que a existência de mais de um estudo sobre o mesmo tópico fornece uma oportunidade para avaliar a validade externa de todos os estudos pela comparação de seus resultados. No caso da pontuação nos exames e do tamanho da turma, outras bases de dados comparáveis estão, de fato, disponíveis. Nesta seção, examinamos uma base de dados diferente, baseada nos resultados de exames padronizados para alunos da rede pública na 4ª série em 220 diretorias regionais de ensino no Estado de Massachusetts em 1998. Tanto os exames de Massachusetts quanto os da Califórnia são medidas amplas do conhecimento do aluno e da aptidão acadêmica, ainda que cada um tenha a sua particularidade. Da mesma forma, a organização das aulas no ensino fundamental é muito semelhante nos dois estados (assim como na maior parte do ensino fundamental nas diretorias regionais de ensino dos Estados Unidos), embora aspectos do financiamento do ensino fundamental e do currículo sejam diferentes. Portanto, a obtenção de resultados semelhantes com relação ao efeito da razão aluno-professor sobre o desempenho nos exames com dados de Massachusetts e Califórnia evidenciaria a validade externa dos resultados da Califórnia. Inversamente, a obtenção de resultados diferentes nos dois estados levantaria dúvidas sobre a validade interna ou externa de pelo menos um dos estudos.

Comparação dos dados da Califórnia e de Massachusetts. Assim como os dados da Califórnia, os dados de Massachusetts estão ao nível da diretoria regional de ensino. As definições das variáveis na base de dados de Massachusetts são iguais, ou praticamente iguais, àquelas na base de dados da Califórnia. O Apêndice 7.1 fornece mais informações sobre a base de dados de Massachusetts, incluindo as definições das variáveis.

A Tabela 7.1 apresenta estatísticas-resumo para as amostras da Califórnia e de Massachusetts. A pontuação média nos exames é maior em Massachusetts, porém o exame é diferente; logo, uma comparação direta das pontuações não é apropriada. A razão aluno-professor média é maior na Califórnia (19,6 *versus* 17,3). A renda média na diretoria é 20 por cento maior em Massachusetts, porém o desvio padrão da renda é maior na Califórnia, isto é, há uma dispersão maior nas rendas médias das diretorias da Califórnia em relação a Massachusetts. A porcentagem média de alunos que ainda está aprendendo inglês e a porcentagem média de alunos com direito a almoço subsidiado são muito maiores na Califórnia do que em Massachusetts.

TABELA 7.1 Estatísticas-Resumo para as Bases de Dados da Pontuação nos Exames da Califórnia e de Massachusetts

	Califórnia		Massachusetts	
	Média	Desvio padrão	Média	Desvio padrão
Pontuação nos exames	654,1	19,1	709,8	15,1
Razão aluno-professor	19,6	1,9	17,3	2,3
% aprendendo inglês	15,8%	18,3%	1,1%	2,9%
% com direito a almoço subsidiado	44,7%	27,1%	15,3%	15,1%
Renda média na diretoria (US$)	15.317	7.226	18.747	5.808
Número de observações	420		220	
Ano	1999		1998	

Pontuação nos exames e renda média na diretoria. Para economizar espaço, não apresentamos gráficos de dispersão para todos os dados de Massachusetts. Como foi observado no Capítulo 6, contudo, é interessante examinar a relação entre pontuação nos exames e renda média na diretoria em Massachusetts. A Figura 7.1 mostra o gráfico de dispersão dessa relação. O padrão geral desse gráfico é semelhante àquele da Figura 6.2 para os dados da Califórnia: a relação entre renda e pontuação nos exames parece muito inclinada para valores baixos de renda e pouco inclinada para valores altos. Evidentemente, a regressão linear mostrada na figura não revela essa aparente não-linearidade. A Figura 7.1 também mostra as funções de regressão cúbica e logarítmica. A função de regressão cúbica possui um $\overline{R}^2$ ligeiramente maior do que a especificação logarítmica (0,486 *versus* 0,455). A comparação das figuras 6.7 e 7.1 mostra que o padrão geral de não-linearidade encontrado nos dados de renda e pontuação nos exames da Califórnia também está presente nos dados de Massachusetts. Contudo, as formas funcionais precisas que melhor descrevem essa não-linearidade diferem uma da outra, com a especificação cúbica ajustando-se melhor para Massachusetts e a especificação linear-log ajustando-se melhor para a Califórnia.

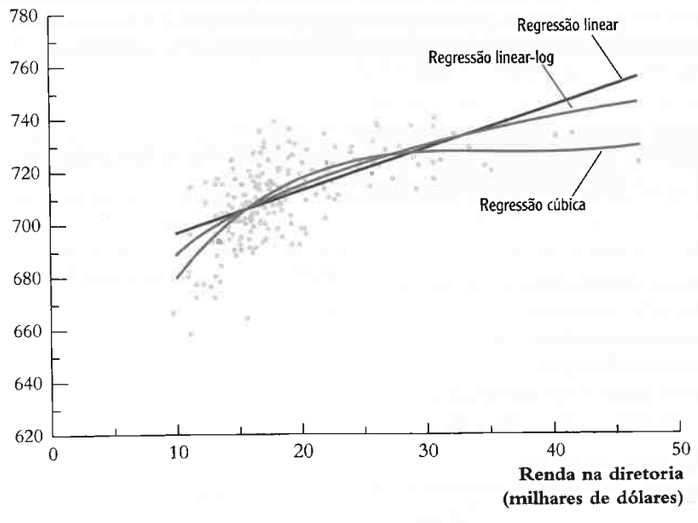
Resultados da regressão múltipla. A Tabela 7.2 apresenta os resultados da regressão para os dados de Massachusetts. A primeira regressão, presente na coluna (1) da tabela, possui apenas a razão aluno-professor como regressor. A declividade é negativa (–1,72), e a hipótese de que o coeficiente é zero pode ser rejeitada ao nível de significância de 1 por cento ($t = -1,72/0,50 = -3,44$).

As colunas restantes apresentam os resultados da inclusão de variáveis adicionais que controlam características dos alunos e da introdução de não-linearidades na função de regressão estimada. O controle da porcentagem de alunos que está aprendendo inglês, da porcentagem de alunos com direito a almoço subsidiado e da renda média na diretoria reduz o coeficiente estimado sobre a razão aluno-professor em 60 por cento, de –1,72 na regressão (1) para –0,69 na regressão (2) e –0,64 na regressão (3).

A comparação dos $\overline{R}^2$ nas regressões (2) e (3) indica que a especificação cúbica (3) fornece um modelo melhor para a relação entre pontuação nos exames e renda do que a especificação logarítmica (2), mesmo que a razão aluno-professor seja mantida constante. Não há evidência estatisticamente significante de uma relação não-linear entre pontuação nos exames e razão aluno-professor: a estatística F na regressão (4) que testa se os coeficientes da população de RAP^2 e RAP^3 são iguais a zero tem um valor p de 0,641. Da mesma forma, não há evidência de que uma redução na razão aluno-professor tenha um efeito diferente em diretorias que têm muitos alunos

FIGURA 7.1 Estatísticas-Resumo para as Bases de Dados da Pontuação nos Exames da Califórnia e de Massachusetts

A função de regressão linear estimada não capta a relação não-linear entre renda e pontuação nos exames para os dados de Massachusetts. As funções estimadas linear-log e cúbica são semelhantes para rendas na diretoria entre US$ 13.000 e US$ 30.000, a região que contém a maior parte das observações.



aprendendo inglês em relação àquelas que têm poucos (a estatística t de $AIAlta \times RAP$ na regressão (5) é 0,80/0,56 = 1,43). Finalmente, a regressão (6) mostra que o coeficiente estimado da razão aluno-professor não muda substancialmente quando a porcentagem de alunos que está aprendendo inglês (que é insignificante na regressão (3)) é excluída. Em suma, os resultados da regressão (3) não são sensíveis a mudanças na forma funcional e na especi-

TABELA 7.2 Estimativas de Regressões Múltiplas da Razão Aluno-Professor e Pontuação nos Exames: Dados de Massachusetts

Variável Dependente: Média da Pontuação Combinada nos Exames de Inglês, Matemática e Ciências na Diretoria Regional de Ensino; Quarta Série; 220 Observações.

Regressor	(1)	(2)	(3)	(4)	(5)	(6)
Razão aluno-professor (*RAP*)	–1,72** (0,50)	–0,69* (0,27)	–0,64* (0,27)	12,4 (14,0)	–1,02** (0,37)	–0,67* (0,27)
RAP^2				-0,680 (0,737)		
RAP^3				0,011 (0,013)		
% aprendendo inglês		–0,411 (0,306)	–0,437 (0,303)	–0,434 (0,300)		
% aprendendo inglês > mediana (Binária, *AIAlta*)					–12,6 (9,8)	
$AIAlta \times RAP$					0,80 (0,56)	
% com direito a almoço subsidiado		–0,521** (0,077)	–0,582** (0,097)	–0,587** (0,104)	–0,709** (0,091)	–0,653** (0,72)
Renda na diretoria (logaritmo)		16,53** (3,15)				
Renda na diretoria			–3,07 (2,35)	–3,38 (2,49)	–3,87* (2,49)	–3,22 (2,31)
Renda na diretoria^2			0,164 (0,085)	0,174 (0,089)	0,184* (0,090)	0,165 (0,085)
Renda na diretoria^3			–0,0022* (0,0010)	–0,0023* (0,0010)	–0,0023* (0,0010)	–0,0022* (0,0010)
Intercepto	739,6** (8,6)	682,4** (11,5)	744,0** (21,3)	665,5** (81,3)	759,9** (23,2)	747,4** (20,3)
Estatísticas *F* e Valores *p* Testando a Exclusão de Grupos de Variáveis						
Todas as variáveis *RAP* e as interações = 0				2,86 (0,038)	4,01 (0,020)	
RAP^2, $RAP^3 = 0$				0,45 (0,641)		
Renda^2, *Renda*^{3}			7,74 (< 0,001)	7,75 (< 0,001)	5,85 (0,003)	6,55 (0,002)
AIAlta, $AIAlta \times RAP$					1,58 (0,208)	
EPR	14,64	8,69	8,61	8,63	8,62	8,64
$\overline{R}^2$	0,063	0,670	0,676	0,675	0,675	0,674

Essas regressões foram estimadas utilizando dados sobre o ensino fundamental em diretorias regionais de ensino de Massachusetts descritos no Apêndice 7.1. Os erros padrão estão entre parênteses abaixo dos coeficientes; os valores p estão entre parênteses abaixo da estatística F. Os coeficientes individuais são estatisticamente significantes ao nível de *5 por cento ou ao nível de **1 por cento.

ficação consideradas nas regressões (4)-(6) da Tabela 7.2. Portanto, adotamos a regressão (3) como nossa estimativa de base do efeito de uma variação na razão aluno-professor sobre a pontuação nos exames com base nos dados de Massachusetts.

Comparação entre os resultados de Massachusetts e da Califórnia. Com relação aos dados da Califórnia, constatamos que:

a. A adição de variáveis que controlam características da situação do aluno reduziu o coeficiente da razão aluno-professor de –2,28 (veja a Tabela 5.2, regressão (1)) para –0,73 (veja a Tabela 6.2, regressão (2)), uma redução de 68 por cento.
b. A hipótese de que o verdadeiro coeficiente sobre a razão aluno-professor seja igual a zero foi rejeitada ao nível de significância de 1 por cento, mesmo após a adição de variáveis que controlam a situação do aluno e as características econômicas da diretoria.
c. O efeito de um corte na razão aluno-professor não dependeu de maneira significativa da porcentagem de alunos que está aprendendo inglês na diretoria.
d. Há alguma evidência de que a relação entre pontuação nos exames e razão aluno-professor seja não-linear.

Constatamos o mesmo para Massachusetts? Para os itens (a), (b) e (c), a resposta é sim. A inclusão de variáveis adicionais de controle reduziu o coeficiente da razão aluno-professor de –1,72 (Tabela 7.2, regressão (1)) para –0,69 (Tabela 7.2, regressão (2)), uma redução de 60 por cento. Os coeficientes da razão aluno-professor permanecem significantes após a adição de variáveis de controle. Nos dados de Massachusetts, esses coeficientes são significantes apenas ao nível de 5 por cento, ao passo que, nos dados da Califórnia, eles são significantes ao nível de 1 por cento. Contudo, o número de observações para os dados da Califórnia é praticamente o dobro, de modo que não é surpreendente que as estimativas para esse Estado sejam mais precisas. Assim como nos dados da Califórnia, não há evidência estatisticamente significante nos dados de Massachusetts de uma interação entre a razão aluno-professor e a variável binária que indica uma alta porcentagem de alunos aprendendo inglês na diretoria.

A constatação (d), entretanto, não é válida para os dados de Massachusetts: a hipótese de que a relação entre a razão aluno-professor e a pontuação nos exames é linear não pode ser rejeitada ao nível de significância de 5 por cento quando testada contra uma especificação cúbica.

Como os dois exames padronizados são diferentes, não é possível comparar os coeficientes diretamente: um ponto no exame de Massachusetts não é igual a um ponto no exame da Califórnia. Se, contudo, as pontuações nos exames forem expressas na mesma unidade, será possível comparar os efeitos estimados do tamanho da turma. Uma maneira de fazer isso é transformar as pontuações nos exames por meio de uma padronização: subtrair a média da amostra e dividir pelo desvio padrão de modo que elas tenham uma média igual a zero e uma variância igual a um. Os coeficientes de declividade na regressão com a pontuação nos exames transformada são iguais aos coeficientes de declividade na regressão original, divididos pelo desvio padrão da pontuação. Portanto, o coeficiente da razão aluno-professor, dividido pelo desvio padrão da pontuação nos exames, pode ser comparado nos dois conjuntos de dados.

A Tabela 7.3 apresenta essa comparação. A primeira coluna apresenta a estimativa de MQO do coeficiente da razão aluno-professor em uma regressão com a porcentagem de alunos que está aprendendo inglês, a porcentagem de alunos que tem direito a almoço subsidiado e a renda média na diretoria incluídas como variáveis de controle. A segunda coluna apresenta o desvio padrão da pontuação nos exames entre as diretorias. As duas últimas colunas apresentam o efeito estimado de uma redução da razão aluno-professor em dois alunos por professor (a proposta de nossa superintendente) sobre a pontuação nos exames — a primeira em unidades de pontuação e a segunda em unidades de desvio padrão. Na especificação linear, o coeficiente de MQO estimado utilizando os dados da Califórnia é –0,73, de modo que se estima que o corte da razão aluno-professor em dois aumente a pontuação nos exames na diretoria em $-0,73 \times (-2) = 1,46$ ponto. Como o desvio padrão da pontuação nos exames é 19,1 pontos, isso corresponde a $1,46/19,1 = 0,076$ desvios padrão da distribuição da pontuação nos exames entre as diretorias. O erro padrão dessa estimativa é $0,26 \times 2/19,1 = 0,027$. Os efeitos estimados para os modelos não-lineares e seus erros padrão foram calculados pelo método descrito na Seção 6.1.

TABELA 7.3 Razões Aluno-Professor e Pontuação nos Exames: Comparação das Estimativas para Califórnia e Massachusetts

			Efeito Estimado da Redução de 2 Alunos por Professor em Unidades de:	
	Estimativa de MQO $\hat{\beta}_{RAP}$	Desvio Padrão da Pontuação nos Exames entre Diretorias	Pontuação dos Exames	Desvio Padrão
Califórnia				
Linear: Tabela 6.2(2)	–0,73 (0,26)	19,1	1,46 (0,52)	0,076 (0,027)
Cúbica: Tabela 6.2(7) *Reduzir RAP de 20 para 18*	—	19,1	2,93 (0,70)	0,153 (0,037)
Cúbica: Tabela 6.2(7) *Reduzir RAP de 22 para 20*	—	19,1	1,90 (0,69)	0,099 (0,036)
Massachusetts				
Linear: Tabela 7.2(3)	–0,64 (0,27)	15,1	1,28 (0,54)	0,085 (0,036)

Os erros padrão estão entre parênteses.

Com base no modelo linear utilizando dados da Califórnia, estima-se que uma redução de dois alunos por professor aumente a pontuação nos exames em 0,076 unidade de desvio padrão, com um erro padrão de 0,027. Os modelos não-lineares para os dados da Califórnia sugerem um efeito um pouco maior; o efeito específico depende da razão aluno-professor inicial. Com base nos dados de Massachusetts, esse efeito estimado é de 0,085 unidade de desvio padrão, com um erro padrão de 0,036.

Essas estimativas são essencialmente as mesmas. É previsto que o corte da razão aluno-professor aumente a pontuação nos exames, porém a melhoria prevista é pequena. Nos dados da Califórnia, por exemplo, a diferença da pontuação nos exames entre a diretoria mediana e a diretoria no 75º percentil é de 12,2 pontos no exame (veja a Tabela 4.1), ou 0,64 (= 12,2/19,1) desvios padrão. O efeito estimado pelo modelo linear é pouco mais de um décimo disso; em outras palavras, de acordo com essa estimativa, o corte da razão aluno-professor em dois faria com que uma diretoria movesse somente um décimo do caminho da mediana para o 75º percentil da distribuição da pontuação nos exames entre as diretorias. A redução da razão aluno-professor em dois é uma mudança grande para uma diretoria, mas os benefícios estimados mostrados na Tabela 7.3, apesar de não serem nulos, são pequenos.

A análise dos dados de Massachusetts sugere que os resultados da Califórnia são válidos externamente, pelo menos quando generalizados para o ensino fundamental em diretorias regionais de ensino de outras partes dos Estados Unidos.

Validade Interna

A semelhança entre os resultados para a Califórnia e para Massachusetts não garante sua validade *interna*. Na Seção 7.2, enumeramos cinco ameaças possíveis para a validade interna que poderiam induzir um viés no efeito estimado do tamanho da turma sobre a pontuação nos exames. Consideramos agora cada uma dessas ameaças.

Variáveis omitidas. As regressões múltiplas apresentadas neste capítulo e nos anteriores controlam uma característica do aluno (porcentagem aprendendo inglês), uma característica econômica familiar (porcentagem de alunos com direito a almoço subsidiado) e uma medida mais ampla da riqueza na diretoria (renda média na diretoria).

Variáveis possíveis omitidas, tais como outras características da escola e do aluno, continuam omitidas e isso pode provocar um viés de variável omitida. Por exemplo, se a razão aluno-professor estiver correlacionada com a qualificação do professor (talvez porque os melhores professores são atraídos para escolas com razões aluno-professor menores) e a qualidade do professor afetar a pontuação nos exames, a omissão da qualidade do professor poderá tornar o coeficiente da razão aluno-professor viesado. Da mesma forma, diretorias com uma razão aluno-professor baixa também podem oferecer várias oportunidades de aprendizado extracurricular. Além disso, as diretorias com uma razão aluno-professor baixa podem atrair famílias mais comprometidas com a melhora do aprendizado de seus filhos em casa. Esses fatores omitidos podem levar a um viés de variável omitida.

Uma forma de eliminar o viés de omissão de variáveis — pelo menos na teoria — é conduzir um experimento. Por exemplo, alunos poderiam ser designados aleatoriamente para turmas de tamanhos diferentes e seu desempenho posterior em exames padronizados poderia ser comparado. Tal estudo foi na verdade conduzido no Estado do Tennessee; vamos examiná-lo no Capítulo 11.

Forma funcional. A análise feita aqui e no Capítulo 6 explorou diversas formas funcionais. Observamos que algumas das possíveis não-linearidades investigadas não eram estatisticamente significantes, ao passo que aquelas que eram não alteravam substancialmente o efeito estimado de uma redução na razão aluno-professor. Embora seja possível conduzir análises adicionais de forma funcional, isso sugere que os principais resultados desses estudos provavelmente não são sensíveis a especificações diferentes de regressão não-linear.

Erros nas variáveis. A razão aluno-professor média na diretoria é uma medida ampla e potencialmente imprecisa do tamanho da turma. Por exemplo, como os alunos trocam continuamente de diretoria, pode ser que a razão aluno-professor não represente com precisão os tamanhos efetivos de turma experimentados pelos alunos que se submetem ao exame, o que por sua vez poderia levar o efeito estimado do tamanho da turma a um viés em direção a zero. Outra variável com erro de medida potencial é a renda média na diretoria. Esses dados foram obtidos do censo de 1990, ao passo que os outros dados são de 1998 (Massachusetts) ou 1999 (Califórnia). Se a composição econômica da diretoria tivesse mudado substancialmente ao longo da década de 1990, essa seria uma medida imprecisa da verdadeira renda média na diretoria.

Seleção. Os dados da Califórnia e de Massachusetts cobrem todo o ensino público fundamental nas diretorias regionais de ensino do Estado que satisfazem a restrições de tamanho mínimo, de modo que não há motivo para acreditar que a seleção da amostra seja um problema nesse caso.

Casualidade simultânea. A casualidade simultânea surgiria se o desempenho nos exames padronizados afetasse a razão aluno-professor. Isso poderia ocorrer, por exemplo, se houvesse um mecanismo burocrático ou político para aumentar a alocação de fundos para diretorias ou escolas com desempenho fraco, que por sua vez resultaria na contratação de mais professores. Em Massachusetts não havia tal mecanismo para equalização da alocação de fundos a escolas na época desses exames. Na Califórnia, uma série de ações legais levou a alguma equalização da alocação de fundos, porém essa redistribuição de fundos não se baseou no desempenho dos alunos. Portanto, a casualidade simultânea não parece ser um problema nem em Massachusetts nem na Califórnia.

Heteroscedasticidade e correlação do termo de erro entre observações. Todos os resultados relatados aqui e nos capítulos anteriores utilizam erros padrão robustos quanto à heteroscedasticidade, de modo que a heteroscedasticidade não ameaça a validade interna. A correlação do termo de erro entre as observações, contudo, poderia ameaçar a consistência dos erros padrão, uma vez que não foi utilizada uma amostragem aleatória simples (a amostra consiste de todo o ensino fundamental nas diretorias regionais de ensino do Estado). Embora haja fórmulas alternativas de erro padrão que poderiam ser aplicadas a essa situação, os detalhes são complicados e especializados e por isso sua discussão será deixada para textos mais avançados.

Discussão e Implicações

A semelhança entre os resultados para a Califórnia e para Massachusetts sugere que esses estudos são válidos externamente, no sentido de que os principais resultados podem ser generalizados para o desempenho em exames padronizados no ensino fundamental em outras diretorias regionais de ensino dos Estados Unidos.

Algumas das principais ameaças potenciais à validade interna foram atacadas pelo controle da situação do aluno, da situação econômica familiar e da riqueza na diretoria por meio da procura por não-linearidades na função de regressão. Mesmo assim, algumas ameaças potenciais à validade interna persistem. O principal candidato é o viés de omissão de variáveis, que talvez surja pelo fato de as variáveis de controle não captarem outras características das diretorias regionais de ensino ou as oportunidades de aprendizado extracurricular.

Tomando como base os dados da Califórnia e de Massachusetts, estamos capacitados para responder à pergunta feita pela superintendente no Capítulo 4.1: após controlar a situação econômica da família, as características do aluno e a riqueza na diretoria e modelar as não-linearidades na função de regressão, prevê-se que o corte da razão aluno-professor em dois alunos por professor aumente a pontuação nos exames em aproximadamente 0,08 desvios padrão da distribuição da pontuação nos exames entre as diretorias. Esse efeito, embora estatisticamente significante, é muito pequeno. Esse pequeno efeito estimado é consistente com os resultados dos vários estudos que investigaram os efeitos de reduções no tamanho da turma sobre a pontuação nos exames.[4]

A superintendente pode agora utilizar essa estimativa para ajudá-la na decisão de reduzir ou não o tamanho das turmas. Ao tomar essa decisão, ela deverá pesar os custos e os benefícios da redução proposta. Os custos incluem salários dos professores e despesas com salas de aula adicionais. Os benefícios incluem melhor desempenho acadêmico, que medimos pelo desempenho em exames padronizados, mas há outros benefícios potenciais que não estudamos, como menor taxa de evasão escolar e melhores salários no futuro. O efeito estimado da proposta sobre o desempenho nos exames padronizados é um insumo importante para o cálculo de custos e benefícios.

7.4 Conclusão

Os conceitos de validade interna e validade externa fornecem uma estrutura para avaliar o que aprendemos em um estudo econométrico.

Um estudo baseado em regressão múltipla é válido internamente se os coeficientes estimados são não viesados e consistentes e se os erros padrão são consistentes. Ameaças à validade interna de tal estudo incluem variáveis omitidas, erro de especificação da forma funcional (não-linearidades), medida imprecisa das variáveis independentes (erros nas variáveis), seleção da amostra e causalidade simultânea. Cada uma delas introduz uma correlação entre o regressor e o termo de erro, que por sua vez torna o estimador de MQO viesado e inconsistente. Se os erros são correlacionados ao longo das observações — como podem ser para dados de séries temporais — ou são heteroscedásticos, mas os desvios padrão são calculados utilizando a fórmula somente homoscedástica, então a validade interna fica comprometida em virtude da inconsistência dos desvios padrão. Esse último grupo de problemas pode ser resolvido pelo cálculo apropriado dos desvios padrão.

Um estudo que utiliza a análise de regressão, assim como qualquer estudo estatístico, é válido externamente se os seus resultados podem ser generalizados além da população e do cenário estudados. Algumas vezes a comparação de dois ou mais estudos sobre o mesmo tópico pode ajudar. Independentemente da existência de dois ou mais desses estudos, contudo, a avaliação da validade externa requer um julgamento sobre as semelhanças da população e do cenário estudados com a população e o cenário para os quais os resultados estão sendo generalizados.

As próximas duas partes deste livro desenvolvem formas de eliminar as ameaças que comprometem a validade interna que não podem ser diminuídas somente pela análise de regressão múltipla. A Parte 3 estende o modelo de regressão múltipla em caminhos projetados para diminuir as cinco fontes potenciais de viés no estimador de MQO; a Parte 3 também discute os experimentos controlados aleatórios, um enfoque diferente para obter validade interna. A Parte 4 desenvolve métodos para a análise de dados de séries temporais e para a utilização desses dados na estimativa dos chamados efeitos causais dinâmicos, que variam ao longo do tempo.

[4] Se você estiver interessado em aprender mais sobre a relação entre tamanho da turma e pontuação nos exames, veja as resenhas de Ehrenberg, Brewer, Gamoran e Willms (2001a, 2001b).

Resumo

1. Estudos estatísticos são avaliados verificando-se se a análise é válida interna e externamente. Um estudo é válido internamente se as inferências estatísticas sobre os efeitos causais são válidas para a população estudada. Um estudo é válido externamente se as suas inferências e conclusões podem ser generalizadas com base na população e no cenário estudados para outras populações e cenários.
2. Na análise de regressão, há duas ameaças principais à validade interna. Em primeiro lugar, os estimadores de MQO são inconsistentes se os regressores e os termos de erro são correlacionados. Em segundo lugar, intervalos de confiança e testes de hipótese não são válidos quando os erros padrão são incorretos.
3. Os regressores e os termos de erro podem ser correlacionados quando há variáveis omitidas, uma forma funcional incorreta, um ou mais regressores medidos com erro, a amostra escolhida de forma não aleatória com base na população ou causalidade simultânea entre regressores e variáveis dependentes.
4. Os erros padrão são incorretos quando os erros são heteroscedásticos e o pacote econométrico utiliza os erros padrão somente homoscedásticos, ou quando o termo de erro é correlacionado ao longo de diversas observações.

Termos-chave

população estudada (164)
população de interesse (164)
validade interna (165)
validade externa (165)
erro de especificação da forma funcional (168)
viés de erros nas variáveis (168)
viés de seleção da amostra (170)
viés de causalidade simultânea (170)
viés de equações simultâneas (171)

Revisão dos Conceitos

7.1 Qual é a diferença entre validade interna e externa? E entre população estudada e população de interesse?

7.2 O Conceito-Chave 7.2 descreve o problema de seleção de variáveis em termos de um dilema entre viés e variância. Em que consiste esse dilema? Por que a inclusão de um regressor adicional poderia diminuir o viés? E aumentar a variância?

7.3 Variáveis econômicas freqüentemente são medidas com erro. Isso significa que a análise de regressão não é confiável? Explique.

7.4 Suponha que um Estado ofereça exames padronizados voluntários a todos os alunos da terceira série e que os dados sejam utilizados em um estudo do efeito do tamanho da turma sobre o desempenho dos alunos. Explique como o viés de seleção da amostra pode invalidar os resultados.

7.5 Um pesquisador estima o efeito de gastos com polícia sobre a taxa de criminalidade utilizando dados a nível municipal. Explique como a causalidade simultânea pode invalidar os resultados.

7.6 Um pesquisador estima uma regressão utilizando dois pacotes econométricos diferentes. O primeiro utiliza a fórmula para erros padrão somente homoscedásticos. O segundo utiliza a fórmula robusta quanto à heteroscedasticidade. Os erros padrão são muito diferentes. Qual deles você deve utilizar? Por quê?

Exercícios

***7.1** Suponha que você tenha acabado de ler um estudo estatístico minucioso do efeito da publicidade sobre a demanda por cigarros. Utilizando dados de Nova York durante a década de 1970, o estudo concluiu que a publicidade nos ônibus e no metrô era mais eficiente do que a publicidade nos veículos impressos. Utilize o conceito de validade externa para determinar se esses resultados podem ser aplicados a Boston na década de 1970, a Los Angeles na década de 1970 e a Nova York em 2002.

7.2 Considere o modelo de regressão com uma variável: $Y_i = \beta_0 + \beta_1 X_i + u_i$, e suponha que ele satisfaça a hipótese no Conceito-Chave 4.3. Suponha que Y_i seja medido com erro, de modo que os dados sejam $\tilde{Y}_i = Y_i + w_i$, onde w_i é o erro de medida que é i.i.d. e independente de Y_i e X_i. Considere a regressão da população $\tilde{Y}_i = \beta_0 + \beta_1 X_i + v_i$, onde v_i é o erro da regressão utilizando a variável dependente com erro de medida $\tilde{Y}_i$.

a. Mostre que $v_i = u_i + w_i$.

b. Mostre que a regressão $\tilde{Y}_i = \beta_0 + \beta_1 X_i + v_i$ satisfaz as hipóteses do Conceito-Chave 4.3. (Suponha que w_i seja independente de Y_j e X_j para todos os valores de i e j e possua um quarto momento finito.)

c. Os estimadores de MQO são consistentes?

d. É possível construir intervalos de confiança da forma habitual?

e. Avalie a afirmação: "Erro de medida em X é um problema sério. Erro de medida em Y, não".

7.3 Pesquisadores da área de economia do trabalho estudaram os determinantes do salário das mulheres e descobriram um quebra-cabeça empírico intrigante. Utilizando mulheres empregadas selecionadas aleatoriamente, eles regrediram o salário sobre o número de filhos das mulheres e um conjunto de variáveis de controle (idade, instrução, ocupação etc.). Eles descobriram que mulheres com mais filhos tinham salários maiores, mantendo o controle dos outros fatores. Explique como a seleção da amostra pode ser a causa desse resultado (*Dica*: Observe que a amostra inclui apenas mulheres que estão trabalhando.) (Esse quebra-cabeça empírico motivou a pesquisa de James Heckman sobre seleção da amostra que lhe conferiu o Prêmio Nobel de Economia em 2000.)

APÊNDICE 7.1 Dados de Exames no Ensino Fundamental de Massachusetts

Os dados de Massachusetts para o ensino público fundamental em diretorias regionais de ensino são médias de diretorias em 1998. A pontuação nos exames é extraída do Massachusetts Comprehensive Assessment System (MCAS), um exame aplicado a todos os alunos da quarta série das escolas públicas de Massachusetts no segundo bimestre de 1988. O exame é patrocinado pela Secretaria de Educação de Massachusetts e é obrigatório para todas as escolas públicas. Os dados analisados aqui são a pontuação total global, que é a soma das pontuações nas disciplinas de inglês, matemática e ciências que compõem o exame.

Os dados sobre a razão aluno-professor, a porcentagem de alunos com direito a almoço subsidiado e a porcentagem de alunos que ainda está aprendendo inglês são médias para o ensino fundamental em cada diretoria regional de ensino durante o ano escolar 1997-1998 e foram obtidos da Secretaria de Educação de Massachusetts. Os dados sobre a renda média na diretoria foram obtidos do censo dos Estados Unidos de 1990.