10 INTRODUÇÃO AOS RELACIONAMENTOS MULTIVARIADOS

Os Capítulos 7 a 9 introduziram métodos para analisar a associação entre duas variáveis. Na maior parte das pesquisas em Ciências Sociais, estas análises são apenas a primeira etapa. Etapas subsequentes usam métodos multivariados para incluir na análise outras variáveis que possam influenciar aquela associação.

Os Exemplos 8.1 e 8.3 mostraram que a identificação partidária nos Estados Unidos está associada ao gênero, com os homens tendo maior probabilidade do que as mulheres de serem Republicanos. Para analisar por que é assim, poderíamos analisar se as diferenças entre homens e mulheres na ideologia partidária (mensurada em uma escala conservadora-liberal) poderia explicar a associação. Por exemplo, talvez os homens tendam a ser mais conservadores do que as mulheres e, sendo conservadores, tendem a estar associados ao partido republicano. Se compararmos homens e mulheres apenas para aqueles classificados como liberais e, então, novamente apenas para aqueles classificados como conservadores, ainda será verdadeiro que os homens têm maior probabilidade de serem Republicanos do que as mulheres? Ou a diferença entre homens e mulheres, na identificação da política partidária, poderia ser explicada por algum outro fator, como renda, nível educacional ou religião?

Vários tipos de questões de pesquisa requerem a inclusão de variáveis às análises. Estas questões geralmente envolvem noções de conexões *causais* entre as variáveis. A Seção 10.1 discute métodos de causação e valores atípicos para testar suposições causais. A Seção 10.2 introduz o *controle estatístico*, uma ferramenta fundamental para estudar como uma associação muda ou possivelmente até mesmo desaparece após a remoção da influência das outras variáveis. A Seção 10.3 mostra tipos de relacionamentos multivariados que o controle estatístico pode revelar.

10.1 ASSOCIAÇÃO E CAUSALIDADE

A causalidade é fundamental para o esforço científico. A maioria das pessoas está acostumada com este conceito, pelo menos no sentido informal. Sabemos, por exemplo, que estar exposto a um vírus pode causar um resfriado e que o fumo pode causar câncer nos pulmões. Mas como podemos julgar se existe um relacionamento causal entre duas variáveis nas Ciências Sociais? Por exemplo, o que causa a delinquência juvenil? Ser pobre? Vir de uma família de pai ou mãe solteiro(a)? Falta de educação moral ou religiosa? Fatores genéticos? Uma combinação destes ou outros fatores? Iremos, agora, olhar algumas diretrizes que nos ajudam a avaliar a hipótese da forma "X causa Y". (Neste capítulo, para enfatizar que estamos considerando propriedades probabilísticas das variáveis em vez dos

seus valores particulares, representaremos as variáveis com letra maiúscula.)

Relacionamentos causais geralmente têm uma assimetria, com uma variável tendo uma influência na outra, mas não vice-versa. Uma flecha entre duas variáveis X e Y, apontando para a variável resposta, representa uma associação causal entre as variáveis. Portanto:

$$X \rightarrow Y$$

especifica que X é uma variável explicativa que tem uma influência causal em Y. Por exemplo, suponha que suspeitamos que ser um escoteiro tem um efeito causal em ser um delinquente juvenil, com escoteiros tendo menor probabilidade de serem delinquentes. Estamos supondo que $E \rightarrow D$, onde E (para Escoteiro) e D (para Delinquente) representam as variáveis binárias "se escoteiro (sim, não)" e "se um delinquente juvenil (sim, não)".

Se suspeitamos que uma variável esteja relacionada a outra de forma causal, como analisamos se ela realmente está? Um relacionamento deve satisfazer três critérios para ser considerado causal. Estes critérios, que iremos discutir, são:

- Uma associação entre as variáveis.
- Uma ordem apropriada no tempo.
- A eliminação de explicações alternativas.

Se todos os três critérios forem satisfeitos, então a evidência sustenta um relacionamento causal. Se um ou mais critérios não são satisfeitos, então concluímos que não existe um relacionamento causal.

Associação

O primeiro critério para a causalidade é a **associação**. Devemos mostrar que X e Y estão associados para sustentar a hipótese de que X causa Y. Se $X \rightarrow Y$, então à medida que X muda, a distribuição de Y deve mudar de alguma forma. Se o escotismo causa taxas baixas de delinquência, por exemplo, então a proporção da população de delinquentes deve ser maior para não escoteiros do que para escoteiros. Para dados amostrais, um teste como o qui-quadrado para dados categóricos ou um teste t para a inclinação da regressão ou para a comparação das médias para dados quantitativos analisa se este critério é satisfeito.

A associação sozinha não pode estabelecer causalidade.

Associação não implica causalidade.

O restante desta seção explica por quê.

Ordem no tempo

O segundo critério para a causalidade é que duas variáveis tenham uma **ordem no tempo** adequada, com a causa precedendo o efeito. Algumas vezes isto é apenas uma questão de lógica. Por exemplo, raça, idade e gênero existem antes de atitudes ou realizações, assim qualquer associação causal deve tratá-las como causas em vez de efeitos.

Em outros casos, a direção causal não é tão óbvia. Considere escotismo e delinquência. É logicamente possível que o escotismo reduza a tendência à delinquência. Por outro lado, é também possível que meninos delinquentes evitem o escotismo e meninos não delinquentes, não. Portanto, a ordem no tempo não está clara e ambas as possibilidades $E \rightarrow D$ e $D \rightarrow E$ são plausíveis. Apenas mostrando que uma associação existe não resolve o dilema, porque uma proporção mais baixa de delinquentes entre os membros dos escoteiros é consistente com ambas as explicações.

Quando um estudo é experimental em vez de observacional, a ordem no tempo pode ser ajustada. Por exemplo, um novo medicamento tem um efeito benéfico no tratamento de uma doença? Poderíamos atribuir aleatoriamente sujeitos que sofrem da doença para receber o medicamento ou o placebo. Então, para analisar se o medicamento tem uma influência causal na doença, observaríamos se a propor-

ção que foi tratada com sucesso é significativamente maior para o grupo tratado com o medicamento. O resultado para um sujeito (sucesso ou não) é observado *após* o tratamento, assim a ordem no tempo está correta.

É difícil estudar causa e efeito quando duas variáveis não apresentam uma ordem temporal mas são mensuradas juntas no tempo. As variáveis podem estar associadas meramente porque ambas têm uma tendência ao longo do tempo. Suponha que a taxa de divórcios e a taxa de crimes têm uma tendência de aumento por um período de 10 anos. Elas terão, então, uma correlação positiva: taxas mais altas de crimes ocorrem em anos que apresentam taxas mais altas de divórcios. Isto não implica que uma taxa crescente de divórcios cause um aumento na taxa de crimes. Elas também poderiam estar correlacionadas positivamente com outras variáveis que têm uma tendência crescente no tempo, como o preço médio de vendas de casas, percentual de pessoas que usam telefone celulares e número de buscas na internet utilizando o Google.

Eliminação de uma explicação alternativa

Suponha que duas variáveis estão associadas e têm a sua própria ordem no tempo para satisfazer a relação causal. Isto ainda é insuficiente para indicar causalidade. Pode haver uma **explicação alternativa** para a associação.

Por exemplo, os pilotos de aeronaves ligam o aviso de "colocar cintos de segurança" antes de se deparar com a turbulência. Observamos uma associação, com uma maior turbulência ocorrendo quando o aviso de colocar o cinto está ligado do que quando ele está desligado. Existe, também, uma ordem no tempo adequada com o aviso aparecendo antes da ocorrência da turbulência. Mas isso não implica que ligar o aviso causa turbulência.

Uma explicação alternativa para uma associação é responsável pela rejeição de muitas hipóteses de relacionamentos causais. Muitas explicações alternativas envolvem uma variável adicional Z ou um conjunto de variáveis. Por exemplo, pode haver uma variável Z que causa tanto X quanto Y. O relacionamento pode ser **espúrio**, como será definido na Seção 10.3, com tanto X quanto Y sendo dependentes de Z.

Com dados observacionais, é fácil encontrar associações, mas estas associações são geralmente explicadas por outras variáveis que podem não ter sido mensuradas em um estudo. Por exemplo, alguns estudos médicos encontraram associações entre o consumo do café e várias respostas, tais como como a possibilidade de um ataque cardíaco. Mas, após levar em consideração outras variáveis associadas à quantidade consumida de café, tais como o país de residência, ocupação e níveis de estresse, tais associações desapareceram ou enfraqueceram consideravelmente.

O critério de eliminar uma explicação alternativa, para a causalidade é o mais difícil de ser obtido. Podemos achar que encontramos um relacionamento causal, mas podemos meramente não ter pensado em uma razão em particular que possa explicar a associação. Por essa razão, nunca podemos *provar* que uma variável é a causa de outra. Podemos, contudo, refutar a hipótese de causalidade mostrando que a evidência empírica contradiz pelo menos um destes três critérios.

Associação, causalidade e evidência incidental

A associação entre o fumo e câncer de pulmão é uma que, agora, é considerada como tendo uma ligação causal. A associação é moderadamente forte, existe uma ordem própria no tempo (câncer do pulmão precedido por um período de consumo de cigarros) e nenhuma explicação alternativa foi encontrada para explicar o relacionamento. Além disto, a ligação causal tem sido sustentada por teorias biológicas que explicam como o fumo poderia causar o câncer de pulmão.

Algumas vezes você escuta pessoas dando evidências incidentais para tentar refutar relacionamentos causais. "Meu tio Paulo tem 85 anos de idade e ainda fuma duas carteiras de cigarro por dia e está forte como um cavalo." Uma associação não precisa ser perfeita, entretanto, para ser causal. Nem todas as pessoas que fumam duas carteiras por dia terão câncer de pulmão, mas o percentual será mais alto entre os fumantes do que entre os não fumantes. Talvez o tio Paulo esteja ainda com boa saúde, mas isto não nos deveria encorajar a desafiar o destino fumando duas carteiras por dia. A evidência incidental não é o suficiente para refutar a causalidade a não ser que ela possa reduzir um dos três critérios para a causalidade.

10.2 CONTROLE DE OUTRAS VARIÁVEIS

Um componente fundamental para avaliar se X pode causar Y é procurar por uma explicação alternativa. Fazemos isto estudando se a associação entre X e Y continua quando removemos os efeitos de outras variáveis dessa associação. Em uma análise multivariada, uma variável está ou é **controlada** quando sua influência é removida.

Um experimento de laboratório controla variáveis que poderiam afetar os resultados mantendo constantes os seus valores. Por exemplo, um experimento em química ou física pode controlar a temperatura e a pressão atmosférica mantendo-as constante em um ambiente laboratorial durante o curso do experimento. Um experimento de laboratório investigando o efeito de diferentes doses de um carcinógeno em ratos pode controlar a idade e dieta dos ratos.

Controle estatístico na pesquisa social

Diferente das ciências de laboratórios, a ciência social é geralmente observacional em vez de experimental. Não podemos fixar valores das variáveis que gostaríamos de controlar, como inteligência ou instrução, antes de obter dados sobre as variáveis de interesse. Mas podemos aproximar um tipo de controle experimental agrupando observações com valores iguais (ou similares) nas variáveis controle. Classe social ou variáveis relacionadas como o grau de instrução ou a renda são geralmente as principais candidatas para serem controladas na pesquisa social. Para controlar o grau de instrução, por exemplo, poderíamos agrupar os resultados amostrais em sujeitos que não completaram o ensino médio, com ensino médio completo, mas sem graduação universitária, com ensino superior incompleto e aqueles com pelo menos uma graduação completa. Isto é **controle estatístico** e não controle experimental.

O exemplo seguinte ilustra o controle estatístico em um cenário da ciência social mantendo uma variável chave constante.

EXEMPLO 10.1 Efeito casual da altura no desempenho em matemática

Os estudantes altos tendem a ter mais habilidade matemática do que estudantes baixos? Podemos pensar assim observando uma amostra aleatória de estudantes do distrito escolar de Lake Wobegon que fizeram um teste para avaliar o desempenho em matemática. A correlação foi de 0,81 entre a altura e o escore em matemática. Os estudantes mais altos tendem a ter escores mais altos.

Ser alto é uma influência causal no conhecimento matemático? Talvez uma explicação alternativa para esta associação é que a amostra tinha estudantes de várias idades. À medida que a idade aumenta, tanto a altura quanto o desempenho em matemática tenderiam a aumentar. Os estudantes mais velhos tendem a ser mais altos e estudantes

mais velhos tendem a ter conhecimentos mais sólidos em matemática.

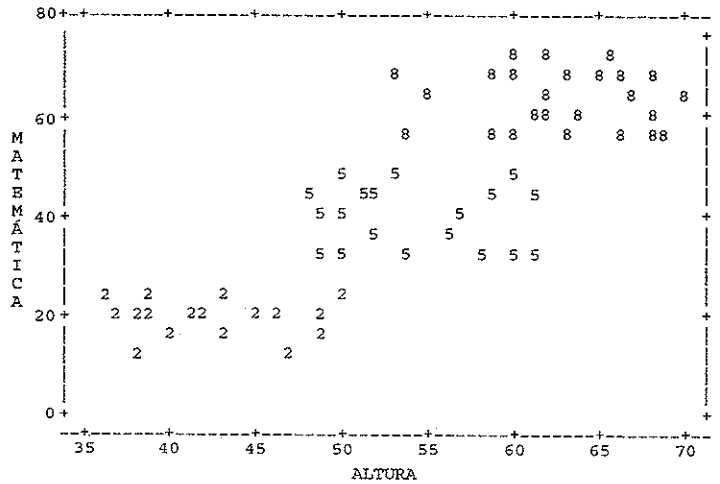
Podemos remover o efeito da idade da associação pelo *controle estatístico*, estudando a associação entre a altura e o escore no teste de matemática para estudantes com a mesma idade. Isto é, controlamos o efeito idade analisando a associação separadamente para cada idade. Então, a variação na idade não pode causar conjuntamente a variação na altura e no escore do teste.

Na verdade, o teste de desempenho foi administrado a estudantes da 2ª, 5ª e 8ª séries de Lake Wobegon, assim a amostra continha considerável variabilidade nas idades dos estudantes. A Figura 10.1 mostra um diagrama de dispersão das observações, com rótulos indicando a série de cada estudante. O padrão geral dos pontos mostra uma correlação positiva forte, com escores de matemática mais altos associados aos estudantes mais altos. Observe os pontos dentro de apenas uma série (para a qual a idade é aproximadamente constante) e você verá uma variação aleatória, sem um padrão, em particular, de aumento ou diminuição. A correlação entre a altura e o escore do teste de matemática está próxima a zero para estudantes com aproximadamente a mesma idade. A altura não tem um efeito causal no escore do teste porque a associação desaparece quando a idade é mantida constante. ■

Em resumo, controlamos a variabilidade mantendo o seu valor constante. Podemos, então, estudar o relacionamento entre X e Y para os casos com valores iguais (ou similares) daquela variável. A variável controlada é chamada de **variável controle**. Mantendo a variável controle constante, removemos a influência daquela variável na associação entre X e Y.

Tipos de associação para o controle estatístico

O diagrama de dispersão na Figura 10.1 descreve a associação entre duas variáveis quantitativas controladas por uma terceira variável. Podemos descrever a associação



☑ **Figura 10.1** Saída computacional mostrando o relacionamento entre "altura" e o "escore em um teste de matemática", com observações identificadas por série. Os estudantes em cada uma das séries têm aproximadamente a mesma idade.

entre uma variável quantitativa e uma variável categórica comparando as médias. Por exemplo, suponha que na sua faculdade a renda média para o corpo docente masculino é mais alta do que a renda média do corpo docente feminino. Suponha que os professores homens tendem a ser mais velhos e com mais experiência em relação às poucas professoras que têm sido contratadas até recentemente. Então, esta diferença iria diminuir e poderia desaparecer se controlarmos a posição acadêmica ou o número de anos desde a última graduação. Veja o Exercício 10.18 para um exemplo desse tipo. Se relativamente mais docentes mulheres estão em faculdades com salários mais baixos (Artes e Ciências, Educação) e relativamente mais docentes masculinos estão em faculdades com salários mais altos (Medicina, Direito, Engenharia), esta diferença iria diminuir e possivelmente desaparecer se controlarmos a faculdade em que o professor trabalha.

Para estudar a associação entre duas variáveis categóricas, controlada uma terceira variável, formamos tabelas de contingência colocando essas variáveis separadamente para os sujeitos em cada nível da variável controle. As tabelas separadas que exibem os relacionamentos dentro dos níveis fixos da variável controle são chamadas de **tabelas parciais**.

EXEMPLO 10.2 Controle estatístico para escotismo e delinquência

A Tabela 10.1 é uma tabela hipotética relacionando escotismo e delinquência. O percentual de delinquentes entre os escoteiros é mais baixo do que entre os não escoteiros. Esta tabela é **bivariada**, significando que ela contém dados de apenas *duas* variáveis. Todas as outras variáveis são ignoradas. Nenhuma é controlada.

Na procura de uma explicação possível para a associação, vamos controlar a frequência à igreja. Talvez os meninos que frequentam a igreja tenham uma probabilidade maior de serem escoteiros, e frequentadores de igreja tenham uma probabilidade menor de serem delinquentes. Então, a diferença nas taxas de delinquência entre escoteiros e não escoteiros pode ser devido à variação da frequência à igreja.

Para controlar a frequência à igreja examinamos a associação entre escotismo e delinquência dentro das tabelas parciais formadas para os vários níveis de frequência à igreja. A Tabela 10.2 mostra tabelas parciais para três níveis: baixa = não mais do que uma vez por ano, média = mais do que uma vez por ano, mas menos do que uma vez por semana e alta = pelo menos uma vez por semana. Adicionando estas três tabelas parciais, é produzida a tabela bivariada (Tabela 10.1) que ignora a frequência à igreja. Por exemplo, o número de escoteiros que são delinquentes é 36 = 10 + 18 + 8.

Em cada tabela parcial, o percentual de delinquentes é o mesmo para escoteiros como para não escoteiros. Controlando a frequência à igreja, não aparece nenhuma associação entre escotismo e delinquência. Esses dados fornecem uma explicação alternativa para a associação entre escotismo e delinquência, nos tornando céticos para qualquer ligação causal. A explicação

☑ **Tabela 10.1** Tabela de contingência relacionando escotismo e delinquência

		Delinquência		
		Sim	Não	Total
Escoteiros	Sim	36 (9%)	364(91%)	400
	Não	60(15%)	340(85%)	400

alternativa é que estas variáveis dependem da frequência à igreja. Jovens que frequentam a igreja são menos prováveis de serem delinquentes e mais prováveis de serem escoteiros. Para um nível fixo de frequência à igreja, o escotismo não tem associação com delinquência. Visto que a associação pode ser explicada pela dependência de frequência à igreja, não existe uma ligação causal entre escotismo e delinquência. ■

Alguns dos exemplos neste capítulo são dados artificiais para esclarecer os relacionamentos e tornar mais simples a explicação dos conceitos. Na prática, alguma distorção ocorre por causa da variação amostral. Mesmo se uma associação entre duas variáveis realmente desaparece sob um controle, as tabelas *amostrais* parciais não iriam ficar exatamente iguais àquelas na Tabela 10.2. Por causa da variação amostral, elas não iriam mostrar uma *completa* ausência de associação. Além disso, poucas associações desaparecem *completamente* sob um controle. Pode haver *alguma* conexão causal entre duas variáveis, mesmo dentro de cada tabela parcial, mas não tão forte como a tabela bivariada sugere.

Tenha cuidado com as variáveis ocultas

Nem sempre é óbvio qual variável requer controle em um estudo. Conhecer a teoria e pesquisar previamente o campo do estudo ajuda o pesquisador a saber quais variáveis controlar. Uma armadilha potencial em quase toda pesquisa da ciência social é a possibilidade de que uma variável importante não tenha sido incluída no estudo. Se você não conseguir controlar uma variável que influencia fortemente a associação entre as variáveis de maior interesse, você obterá resultados enganadores.

Uma variável que *não* é mensurada em um estudo (ou talvez nem conhecida pelos pesquisadores), mas que influencia a associação de interesse é, algumas vezes, referida como uma **variável oculta**. Na interpretação da correlação positiva entre altura e conhecimento matemático no Exemplo 10.1 (página 341), seríamos negligentes se não percebêssemos que a correlação poderia ser devido a uma variável oculta, como a idade do estudante.

Quando você ler sobre um estudo que relata uma associação, veja se é possível pensar em uma variável oculta que poderia ser responsável pela associação. Por exemplo, suponha que um estudo relate uma correlação positiva entre o GPA obtido na universidade e a renda auferida, mais tarde, na vida profissional. Ter um bom aproveitamento na faculdade é responsável por um salário maior? Uma explicação alternativa é que o GPA alto e o salário alto, ambos, poderiam ser causados por uma variável oculta como a tendência do indivíduo ao trabalho duro.

10.3 TIPOS DE RELACIONAMENTOS MULTIVARIADOS

A Seção 10.2 mostrou que uma associação entre duas variáveis pode mudar dramaticamente quando controlamos uma terceira variável. Esta seção descreve tipos de relacionamentos multivariados que geralmente ocorrem na pesquisa social. Representamos a variável resposta por Y. Na prática, podem existir muitas variáveis explicativas e de controle e nós as representamos por $X_1, X_2, \ldots$

☑ **Tabela 10.2** Tabela de contingência relacionando escotismo e delinquência, controlada pela frequência à igreja

		Frequência à igreja					
		Baixa		Média		Alta	
Delinquência		Sim	Não	Sim	Não	Sim	Não
Escoteiro	Sim	10 (20%)	40 (80%)	18 (12%)	132 (88%)	8 (4%)	192 (96%)
	Não	40 (20%)	160 (80%)	18 (12%)	132 (88%)	2 (4%)	48 (96%)

Associações espúrias

Uma associação entre X_1 e Y é dita ser **espúria** se essas duas variáveis são dependentes de uma terceira variável X_2 e a associação desaparece quando X_2 é controlada. Tal associação resulta do relacionamento de X_1 e Y com a variável controle X_2, em vez de indicar uma conexão causal. A associação entre X_1 e Y desaparece quando removemos o efeito de X_2 mantendo-a constante. Mostrar que a associação entre duas variáveis é espúria refuta a hipótese de uma conexão causal entre elas.

EXEMPLO 10.3 Exemplos de associações espúrias

A associação entre altura e o escore no teste de conhecimento de matemática do Exemplo 10.1 desaparece nos níveis fixos de idade. Aquela associação é espúria, com a idade sendo a causa comum tanto da altura quanto do escore em matemática.

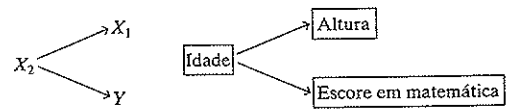
A Tabela 10.1 (página 343) exibiu uma associação entre escotismo e delinquência. Controlando a frequência à igreja, as tabelas parciais da Tabela 10.2 (página 344) não mostraram associação. Isto também é consistente com a espuriedade. A Tabela 10.2 mostra que, à medida que a frequência à igreja aumenta, o percentual de delinquentes diminui (compare os percentuais nas tabelas parciais) e o percentual de escoteiros aumenta. Pela natureza dessas associações, não é surpreendente que a Tabela 10.1 exiba taxas gerais de baixa delinquência para escoteiros do que para não escoteiros. ■

A Figura 10.2 descreve graficamente uma associação espúria, usando X_1 = altura e Y = escore no teste de matemática. Elas estão associadas somente porque ambas dependem de uma causa comum, X_2 = idade. À medida que X_2 muda, ela produz mudanças simultaneamente em X_1 e Y, tal que X_1 e Y estão associados. Na verdade, elas estão associadas somente por causa da sua dependência comum da terceira variável (idade).

EXEMPLO 10.4 Poucas férias causam o aumento no risco de morte?

Quando uma associação é observada entre duas variáveis, estudos posteriores geralmente tentam determinar se aquela associação pode ser espúria controlando as variáveis que poderiam ser uma causa comum. Por exemplo, alguns estudos têm observado uma associação entre a frequência de férias e a qualidade da saúde. Em particular, um estudo de 20 anos de acompanhamento usando participantes mulheres do Framingham Heart Study (Estudo Coronariano de Framingham) descobriu[1] que uma menor frequência de férias estava associada a uma maior frequência de mortes ocasionadas por ataques cardíacos.

Um estudo posterior[2] questionou se isto poderia ser uma associação artificial,



☑ **Figura 10.2** Descrição gráfica de uma associação espúria entre X_1 e Y. A associação desaparece quando controlamos X_2, que causalmente afeta X_1 e Y.

explicada pelos efeitos do *status* socioeconômico (SES). Por exemplo, talvez SES mais alto seja responsável tanto por mortalidade mais baixa e por férias mais frequentes. Mas, após controlar a educação, a renda familiar e outras variáveis potencialmente importantes com conjunto de dados muito maiores, este estudo também observou um risco mais alto de ataque cardíaco e mortes relacionadas para aqueles que gozavam de menos férias. Talvez a associação não seja artificial, a não ser que os pesquisadores encontrem outra variável para controlar tal que a associação desapareça. ■

Relacionamentos encadeados

As associações artificiais não são as únicas para as quais a associação desaparece quando controlamos uma terceira variável. Outra forma é com uma *cadeia* de causação, na qual X_1 afeta X_2, que em ordem afeta Y. A Figura 10.3 descreve a cadeia. Aqui, X_1 é uma causa *indireta* em vez de direta de Y. A variável X_2 é chamada de uma **variável interveniente** (ou algumas vezes de uma variável **mediadora**).

EXEMPLO 10.5 A educação é responsável por uma vida longa?

Um artigo do *New York Times* (por G. Kolata, 3 de janeiro de 2007) resumiu estudos de pesquisas tratando da longevidade de humana. Ele observou que, consistentemente, por meio de estudos em muitos países, a longevidade estava positivamente associada ao nível educacional. Muitos pesquisadores acreditam que a educação é a variável mais importante na explicação de quanto tempo uma pessoa vive. Ter mais instrução é responsável por ter uma vida mais longa?

Estabelecer conexões causais é difícil. Em algumas sociedades, talvez a causação pudesse ir em outra direção, com crianças doentes não indo à escola ou abandonando os estudos cedo porque estavam doentes. Muitos pesquisadores acreditam que poderia haver uma cadeia de causação, talvez com a renda como uma variável interveniente. Por exemplo, talvez ter mais instrução leve a uma maior riqueza, o que, então, (possivelmente por uma variedade de razões, como acesso a um melhor plano de saúde) leve a viver mais. A Figura 10.4 descreve um modelo causal encadeado.

O suporte para este modelo ocorre se a associação entre educação e duração da vida desaparece após o controle da renda; isto é, se dentro de níveis fixos de renda (a variável interveniente), nenhuma associação ocorre. Se isto acontece, a instrução não afeta diretamente a duração da vida, mas é uma causa indireta intermediada pela renda.

Para os relacionamentos espúrios e relacionamentos encadeados, uma associação entre Y e X_1 desaparece quando controlamos uma terceira variável, X_2. A diferença entre as duas é a ordem entre as variáveis. Para uma associação espúria, X_2 é causada antes de X_1 e Y (como na Figura 10.2), enquanto que em uma associação de cadeia X_2 intervém entre as duas (Figura 10.3).

Para ilustrar, um estudo[3] de taxas de mortalidade nos Estados Unidos descobriu que os estados que apresentam maior desigualdade na renda tendiam a ter taxas maiores de mortalidade ajustadas à idade. Entretanto, esta associação desaparece após controlar o percentual dos residentes do estado que tinham pelo menos o ensino médio completo. Isto pode refletir um relacionamento encadeado ou um relacionamento espúrio? Maiores graus de instrução tenderiam a um resultado de menor desigualdade na renda, que resultaria em menores taxas de mortalidade. Portanto, o relacionamento encadeado

$$X_1 \longrightarrow X_2 \longrightarrow Y$$

☑ **Figura 10.3** Um relacionamento em cadeia, no qual X_1 afeta indiretamente Y por intermédio de uma variável interveniente X_2.

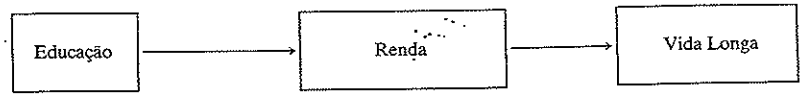
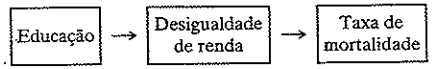


Figura 10.4 Exemplo de um relacionamento encadeado. A renda é uma variável interveniente e a associação entre educação e vida longa desaparece quando ela é controlada.



é plausível. Para o relacionamento ser espúrio, a educação precisaria ter um efeito causal tanto na desigualdade de rendimento quanto na mortalidade. Isso também é plausível. Apenas visualizando os padrões de associações, não sabemos qual deles fornece uma melhor explicação.

Múltiplas causas

As variáveis respostas na pesquisa em Ciências Sociais quase sempre têm mais do que uma causa. Por exemplo, uma variedade de fatores provavelmente tem influências causais nas respostas como Y = delinquência juvenil ou Y = longevidade. A Figura 10.5 descreve X_1 e X_2 como causas separadas de Y. Dizemos que Y tem **múltiplas causas**.

Algumas vezes as variáveis que são causas separadas de Y são, elas próprias, estatisticamente independentes. Isto é, elas são *causas independentes*. Por exemplo, X_1 = gênero e X_2 = raça são essencial e estatisticamente independentes. Se ambas têm efeitos na delinquência juvenil, com taxas da delinquência variando de acordo tanto com gênero quanto com a raça, elas são provavelmente causas independentes.

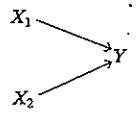
Nas Ciências Sociais, muitas das variáveis explicativas estão associadas. Ser pobre e vir de uma família de pai/mãe solteiro(a) pode causar delinquência, mas estes fatos estão, eles próprios, provavelmente associados. Em virtude de ligações complexas de associações, quando controlamos uma variável X_2 ou um conjunto de variáveis $X_2, X_3, \ldots$, a associação X_1Y geralmente muda de alguma forma. Geralmente a associação diminui um pouco, embora normalmente ela não desapareça completamente como no relacionamento espúrio ou encadeado. Algumas vezes é porque X_1 tem efeitos diretos em Y e também efeitos indiretos por intermédio de outras variáveis. A Figura 10.6 ilustra esse fato. Por exemplo, talvez vir de uma família de pai/mãe solteiro(a) tenha efeito direto na delinquência, mas também efeitos indiretos pela probabilidade maior de ser pobre. A maioria das variáveis respostas tem muitas causas tanto diretas quanto indiretas.

Variáveis supressoras

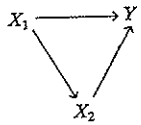
Até agora, discutimos exemplos nos quais uma associação desaparece ou muda quando controlamos outra variável. Em contrapartida, ocasionalmente duas variáveis não mostram associação até que a terceira variável seja controlada. Esta variável controle é chamada de uma **variável supressora**.

EXEMPLO 10.6 A idade suprime a associação entre educação e renda

O nível educacional está relacionado positivamente com a renda? A Tabela 10.3 mostra tal relacionamento, mensurado como variáveis binárias, controlando a



☑ **Figura 10.5** Descrição gráfica das múltiplas causas de Y.



☑ **Figura 10.6** Descrição gráfica dos efeitos diretos e indiretos de X_1 em Y.

idade. Em cada tabela parcial, o percentual de sujeitos no nível de alta renda é maior quando o nível educacional é maior.

Suponha que, agora, a idade seja ignorada, juntando essas duas tabelas parciais. A tabela bivariada para educação e renda é o primeiro painel da Tabela 10.4. Cada freqüência é igual a 250. Agora tanto com uma educação alta quanto uma baixa, o percentual dos que têm alta renda é de 50%. Para a tabela bivariada, nenhuma associação existe entre educação e renda.

Uma olhada nas outras duas tabelas bivariadas na Tabela 10.4 revela como isto poderia acontecer. A idade está possivelmente associada à renda, mas negativamente associadas à educação. Sujeitos mais velhos tendem a ter renda mais alta, mas eles tendem a ter uma educação mais baixa. Portanto, quando a idade é ignorada em vez de controlada, damos um aumento inadvertido ao número relativo de pessoas com níveis altos de renda mas com baixo nível educacional e com níveis baixos de renda mas com níveis altos de educação. ■

Em virtude do potencial para suprimir variáveis, pode ser informativo controlar outras variáveis mesmo quando a análise bivariada não mostrar uma associação. Entretanto, veja Pedhazur (1997, p. 186-88) para uma discussão da importância de existir alguma razão teórica para se esperar efeitos de supressão.

Interação estatística

Geralmente o efeito de uma variável explicativa em uma variável resposta muda de acordo com o nível de outra variável explicativa ou variável controle. Quando o efeito verdadeiro de X_1 em Y muda em diferentes níveis de X_2, o relacionamento é dito exibir uma *interação estatística*.

☑ **Interação estatística**

A **interação estatística** existe entre X_1 e X_2 nos seus efeitos em Y quando o efeito verdadeiro de um previsor em Y muda à medida que o valor do outro previsor muda.

☑ **Tabela 10.3** Tabelas parciais relacionando educação e renda, controlando a idade

		Idade = jovem			Idade = mais velho		
	Renda	**Alta**	**Baixa**	**% Alta**	**Alta**	**Baixa**	**% Alta**
Educação	Alta	125	225	35,7%	125	25	83,3%
	Baixa	25	125	16,7%	225	125	64,3%

☑ **Tabela 10.4** Tabelas parciais relacionando educação, renda e idade

	Renda			Renda			Educação	
Educação	**Alta**	**Baixa**	**Idade**	**Alta**	**Baixa**	**Idade**	**Alta**	**Baixa**
Alta	250	250	Jovem	350	150	Jovem	150	350
Baixa	250	250	Mais velho	150	350	Mais velho	350	150

EXEMPLO 10.7 Interação entre educação e gênero na modelagem da renda

Considere o relacionamento entre Y = renda anual (em milhares de dólares) e X_1 = número de anos de educação, por X_2 = gênero. Muitos estudos nos Estados Unidos descobriram que a inclinação da equação de regressão relacionando Y a X_1 é maior para homens do que para mulheres. Suponha que, na população, as equações de regressão são $E(y) = -10 + 4x_1$ para homens e $E(y) = -5 + 2x_1$ para mulheres. Na média, a renda para os homens aumenta em \$4000 para cada ano de educação, enquanto para as mulheres aumenta em \$2000 para cada ano de educação. Isto é, o efeito da educação na renda varia de acordo com o gênero, com o efeito sendo maior para os homens do que para as mulheres. Assim, existe uma interação entre educação e gênero nos seus efeitos na renda. ■

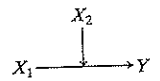
EXEMPLO 10.8 Interação entre SES e idade na qualidade da saúde

Um artigo[4] usando uma amostra de canadenses do *National Population Health Survey* (Levantamento Nacional de Dados da Saúde da População) observou que a qualidade da saúde (mensurada por autoavaliação e índices de saúde) tende a estar positivamente correlacionada com o SES (mensurado por anos de educação e renda familiar anual). Além disso, a associação fica mais forte com a idade. Por exemplo, o intervalo entre os níveis de baixo SES e alto SES tende a ser maior em idades mais avançadas. Portanto, existe uma interação entre SES e idade nos seus efeitos na saúde. ■

Para avaliar se uma amostra exibe evidência de interação, podemos comparar o efeito de X_1 em Y em diferentes níveis de X_2. Quando o efeito amostral é similar em cada nível de X_2, é mais simples usar análise estatística que assuma uma ausência de interação. A interação é digna de nota quando a variabilidade nos efeitos é grande. Por exemplo, talvez a associação seja positiva em um nível de X_2 e negativa em outro, ou forte em um nível e fraca ou não existente em outro. Nos próximos capítulos iremos aprender como analisar inferencialmente se os dados amostrais fornecem forte evidência de interação populacional.

A Figura 10.7 descreve um relacionamento com três variáveis tendo interação estatística. Aqui, X_2 afeta o relacionamento entre X_1 e Y. Quando isto acontece, então, da mesma forma, X_1 afeta o relacionamento entre X_2 e Y.

Suponha que *não* exista interação entre X_1 e X_2 nos seus efeitos em Y. Isto



☑ **Figura 10.7** Descrição gráfica da interação estatística. O efeito de um previsor em Y depende do nível do outro previsor.

não significa que X_1 e X_2 não tenham associação. Pode haver uma ausência de interação estatística mesmo quando todas as variáveis estão associadas. Por exemplo, as Tabelas 10.2 e 10.3 não mostraram interação – em cada caso a associação era similar em cada tabela parcial. Entretanto, em cada caso os previsores estavam associados, entre si e com a resposta. Na Tabela 10.3, por exemplo, a idade estava negativamente associada à educação e positivamente associada à renda.

Resumo dos relacionamentos multivariados

Em resumo,

- Para relacionamentos espúrios (isto é, X_2 afeta X_1 e Y) e relacionamentos encadeados (X_2 intervém entre X_1 e Y), a associação X_1Y desaparece quando controlamos X_2.
- Para múltiplas causas, uma associação pode mudar sob um controle, mas não desaparecer.
- Quando existe uma variável supressora, uma associação aparece somente sob controle.
- Quando existe interação estatística, uma associação tem forças diferentes e/ou direções em valores diferentes de uma variável controle.

Esta lista não esgota as possíveis estruturas de associação. É até mesmo possível que, após controlar uma variável, cada associação em uma tabela parcial tenha a direção oposta como a da associação bivariada. Isto é chamado de **paradoxo de Simpson** e está ilustrado nos Exercícios 10.14, 10.29 e 10.30.

A confusão torna difícil a avaliação dos efeitos

Quando duas variáveis explicativas têm efeitos em uma variável resposta, mas estão também associadas entre si, diz-se que existe um efeito de **confusão**. É difícil determinar se qualquer uma delas realmente causa a resposta, porque o efeito da variável poderia ser, pelo menos, devido à sua associação com a outra variável. Geralmente, observamos um efeito diferente em Y para uma variável quando controlamos a outra variável do que quando a ignoramos.

A confusão é inevitável na pesquisa em Ciências Sociais. É a razão principal da dificuldade de estudar vários assuntos importantes, como o que causa o crime, o que causa o crescimento da economia ou o que causa o sucesso de um estudante na escola.

10.4 QUESTÕES INFERENCIAIS NO CONTROLE ESTATÍSTICO

Para conduzir bem uma pesquisa, você deve selecionar as variáveis chave, determinar quais variáveis controlar, escolher um modelo apropriado, analisar os dados e interpretar os resultados de maneira adequada. As três primeiras seções deste capítulo ignoraram as questões inferenciais para evitar confundi-las com os novos conceitos apresentados. Agora, discutiremos algumas questões inferenciais no estudo das associações enquanto controlamos outras variáveis.

O efeito de pequenas amostras nas análises parciais

Suponha que controlamos X_2 estudando a associação X_1Y. O tamanho da amostra em um nível fixo de X_2 pode ser bem menor do que em todo o conjunto de dados. Mesmo se nenhuma redução ocorrer na associação relativo à totalidade dos dados, os erros padrão das estimativas dos parâmetros tendem a ser maiores. Portanto, os intervalos de confiança para estes parâmetros em níveis fixos de X_2 tendem a ser maiores e os valores da estatística-teste tendem a ser menores.

Para dados categóricos, por exemplo, poderíamos calcular a estatística χ^2 dentro de uma tabela parcial em particular, para testar se as variáveis são independentes naquele nível de X_2. Este valor do χ^2 pode ser menor em relação ao valor do χ^2 para a tabela bivariada X_1Y. Isto pode ser devido em parte a uma associação mais fraca, mas poderia também refletir o tamanho reduzido da amostra. A Seção 8.4 (página 264) mostrou que tamanhos amostrais grandes tendem a produzir valores do χ^2 maiores, para um particular grau de associação.

Os efeitos da categorização no controle de uma variável

Para maior clareza, as variáveis controle nos exemplos deste capítulo tinham somente poucas categorias. Na prática, evite categorizações excessivas de variáveis controle quantitativas. Quanto maior o número de níveis de controle mais aproximadamente constante a variável controle está dentro de cada tabela parcial. A não ser que a variável controle *naturalmente* tenha somente dois níveis (por exemplo, gênero), geralmente é melhor usar pelo menos três ou quatro tabelas parciais.

Por outro lado, é preferível não usar tantas tabelas parciais porque, nesse caso, cada uma pode ter um tamanho amostral pequeno. Estimativas separadas podem ter erros padrão grandes, resultando em inferências imprecisas dentro das tabelas parciais e comparações de associações entre tabelas. Felizmente, os métodos avançados de construção de modelos apresentados no restante do livro nos permitem conduzir um controle estatístico e avaliar padrões de associação e interação sem necessariamente executar análises separadas nas várias combinações dos níveis das variáveis controle.

Comparando e agrupando medidas

Geralmente, é útil comparar os valores dos parâmetros descrevendo o efeito de um previsor em uma resposta nos diferentes níveis de uma variável controle. Pode-se construir um intervalo de confiança para uma diferença entre os valores de dois parâmetros da mesma forma que o Capítulo 7 mostrou para a diferença de proporções ou uma diferença de médias. Suponha que a estimativa das duas amostras é baseada em amostras aleatórias independentes, com erros padrão ep_1 e ep_2. Então a Seção 7.1 (página 212) observou que o erro padrão para a diferença entre as estimativas é $\sqrt{(ep_1)^2 + (ep_2)^2}$. Para amostras aleatórias grandes, a maioria das estimativas tem aproximadamente distribuições amostrais normais. Então, um intervalo de confiança para a diferença entre os parâmetros é:

$$(\text{Estimativa}_2 - \text{Estimativa}_1) \pm z\sqrt{(ep_1)^2 + (ep_2)^2}.$$

Se o intervalo não inclui 0, a evidência sugere que os valores dos parâmetros diferem.

EXEMPLO 10.9 Comparando associações de felicidade para homens e mulheres

Existe uma diferença entre homens e mulheres na associação entre felicidade e felicidade no casamento? Para os dados da PSG de 2004 nas variáveis "HAPPY" e "HAPMAR", o valor amostral de gama para a tabela 3×3 relacionando estas duas variáveis ordinais é 0,674 ($ep = 0{,}0614$, $n = 326$) para homens e 0,689 ($ep = 0{,}0599$, $n = 350$) para mulheres.

Um intervalo de 95% de confiança para a diferença entre os valores populacionais dos gamas é:

$$(0{,}689 - 0{,}674) \pm 1{,}96\sqrt{(0{,}0614)^2 + (0{,}0599)^2}, \quad \text{ou} \quad 0{,}015 \pm 0{,}168,$$

que é (−0,153; 0,183). É plausível que os valores dos gamas populacionais sejam idênticos. Se eles não forem, eles não devem ser muito diferentes. ■

Quando a associação entre duas variáveis é similar na análise parcial, podemos formar uma medida que resuma a força da associação, condicional na variável controle. Isto é referido como uma medida de **associação parcial**. O restante do livro mostra como fazer isto em várias situações usando modelos que lidam com todas as variáveis ao mesmo tempo.

10.5 RESUMO DO CAPÍTULO

É necessário usar a análise multivariada para estudar bem os efeitos em uma variável resposta. Para demonstrar um relacionamento causal, devemos mostrar a **associação** entre as variáveis, assegurar a **ordem apropriada no tempo** e **eliminar explicações alternativas** para a associação.

Para considerar as explicações alternativas introduzimos as **variáveis controle**. Realizamos o controle estatístico analisando as associações enquanto mantemos os valores das variáveis controle essencialmente constantes. Isto nos ajuda a detectar:

- **Dados espúrios**, nos quais X_2 afeta tanto Y quanto X_1.
- **Relacionamentos encadeados** nos quais X_2 é uma **variável interveniente**, tal que X_1 afeta Y indiretamente por meio de seus efeitos em X_2.
- **Variáveis supressoras**, nas quais a associação X_1Y aparece somente depois de controlar X_2.
- **Interação estatística**, na qual o efeito de X_1 em Y varia de acordo com o valor de X_2.

A Tabela 10.5 resume alguns relacionamentos possíveis. O restante do livro apresenta métodos estatísticos para relacionamentos multivariados. À medida que você aprender sobre estes métodos, tenha cuidado de não estender de forma excessiva as suas conclusões: perceba as limitações em fazer inferências causais

☑ **Tabela 10.5** Alguns relacionamentos entre três variáveis

Gráfico	Nome do relacionamento	O que acontece após controlar X_2
$X_2 \nearrow X_1$, $X_2 \searrow Y$	Associação espúria X_1Y	A associação entre X_1 e Y desaparece
$X_1 \longrightarrow X_2 \longrightarrow Y$	Relacionamento encadeado; X_2 intervém; X_1 indiretamente causa Y	A associação entre X_1 e Y desaparece
$X_2 \downarrow$, $X_1 \longrightarrow Y$	Interação	A associação entre X_1 e Y varia de acordo com o nível do X_2
$X_2 \searrow Y \nearrow X_1$	Múltiplas causas	Associação entre X_1 e Y não muda
$X_1 \longrightarrow Y$, $X_1 \searrow X_2 \nearrow Y$	Tanto efeitos diretos quanto indiretos de X_1 em Y	Associação entre X_1 e Y muda, mas não desaparece

com dados inferenciais e tenha em mente que qualquer inferência que você fizer deve geralmente ser temporária por causa das suposições que possam ter sido violadas ou variáveis ocultas que você não incluiu no contesto da modelagem da regressão, veja Berk (2004), Freedman (2005) e Pedhazur (1997).

EXERCÍCIOS

Praticando o básico

10.1 Declare os três critérios para um relacionamento causal. Para cada um, descreva um relacionamento entre duas variáveis que não seja causal porque aquele critério seria violado.

10.2 Uma criança imagina a causa de as mulheres terem filhos. Para cada mulher que vive na sua quadra, ela observa se o seu cabelo é cinza e se tem filhos jovens. As quatro mulheres com cabelo cinza não tem filhos jovens, enquanto todas as cinco mulheres que não têm o cabelo cinza têm filhos jovens. Observando esta associação, a criança conclui que não ter cabelo cinza é a causa das mulheres terem filhos.
- **(a)** Forme uma tabela de contingência apresentando esses dados.
- **(b)** Use este exemplo para explicar por que a associação não implica causação.

10.3 Para todos os incêndios em Chicago no ano passado, os dados estão disponíveis em X = número de bombeiros presentes no incêndio e Y = custo dos danos causados pelo fogo. A correlação é positiva.
- **(a)** Isto significa que ter mais bombeiros em um incêndio causa a piora dos danos? Explique.
- **(b)** Identifique uma terceira variável que poderia ser uma causa comum de X e Y. Construa um diagrama de dispersão hipotético (como o da Figura 10.1) identificando os pontos de acordo com o seu valor na terceira variável para ilustrar o seu argumento.

10.4 As cidades norte-americanas têm uma correlação positiva entre Y = taxa de crimes e X = tamanho da força policial. Isto implica que X causa Y? Explique.

10.5 Existe uma associação entre o GPA universitário e se o estudante universitário já usou maconha. Explique como:
- **(a)** A direção de uma flecha causal pode ir em ambas as direções.
- **(b)** Uma terceira variável pode ser responsável pela associação.

10.6 Explique o que significa *controlar* uma variável; use um exemplo para ilustrar.

10.7 Explique o que é pretendido por uma associação *espúria*; trace um diagrama de dispersão para ilustrar.
- **(a)** Ilustre usando X_1 = tamanho do sapato, X_2 = idade e Y = número de livros lidos, para crianças das escolas de Winnipeg, Canadá.
- **(b)** Ilustre usando X_1 = altura, X_2 = gênero e Y = renda anual, para uma amostra aleatória de adultos. Suponha que, no geral, os homens tendem a ter um renda maior, em média, do que as mulheres.

10.8 A Figura 9.17, na página 317, mostrou uma correlação negativa entre a taxa de nascimentos e a posse de televisores. Identifique uma variável para ajudar a explicar como esta associação poderia ser espúria.

10.9 Uma matéria da *Associated Press* (15 de fevereiro de 2002) citou um estudo da Universidade da Califórnia, em São Diego, que relatava, baseado em um levantamento de dados nacional, que aqueles que tinham em média 8 horas de sono por noite tinham uma probabilidade 12% maior de morrer dentro dos próximos seis anos do que aqueles que tinham em média 6,5 a 7,5 horas de sono por noite.
- **(a)** Explique como a idade do sujeito poderia estar positivamente associada tanto com o tempo gasto dormindo quanto com o índice de mortalidade crescente e, portanto, poderia

explicar a associação entre o tempo dormindo e o índice de mortalidade.

(b) Se a associação desaparece quando controlamos a idade do sujeito, você acha que a idade é mais provável de ser a causa comum ou uma variável interveniente?

10.10 Um estudo descobriu que as crianças que tomam o café da manhã têm melhores notas em matemática do que aqueles que não tomam o café da manhã. Este resultado foi baseado na associação entre X = toma café da manhã (sim, não) e Y = nota do último curso realizado. Como esse resultado poderia ser espúrio e como você poderia verificar esta possibilidade?

10.11 Suponha que a raça esteja relacionada à frequência de detenção juvenil, com adolescentes negros tendo maior probabilidade de serem detidos do que adolescentes brancos. Uma possível explicação de um relacionamento encadeado é que (1) a raça afeta a renda familiar e os negros têm renda familiar mais baixa do que os brancos e (2) ser pobre aumenta a chance de uma detenção juvenil. Mostre uma figura que exiba um relacionamento encadeado. Para sustentar esta explicação, o que seria necessário acontecer à diferença entre as taxas de detenção entre brancos e negros após o controle da renda familiar?

10.12 Um estudo feito na sua universidade descobre que dos inscritos na pós-graduação no ano passado o percentual de homens admitidos foi maior do que o de mulheres. Entretanto, para cada departamento que recebeu as inscrições, o percentual admitido de homens foi *menor* do que o de mulheres. Como isto poderia ter acontecido? Na sua resposta, explique quem faz o papel das variáveis resposta, explicativa, controle, bivariada e das tabelas parciais. (O Exercício 15.12, na página 566, mostra dados que têm comportamento semelhante.)

10.13 A Tabela 10.6 relaciona o nível ocupacional (administração, produção) e a escolha do partido político controlados pela renda.

(a) Construa a tabela bivariada entre o nível ocupacional e o partido político, ignorando a renda. Existe uma associação? Se existir, descreva-a.

(b) As tabelas parciais descrevem uma associação? Interprete-as.

(c) Usando a natureza da associação entre a renda e cada uma das outras variáveis, explique por que a tabela bivariada tem uma associação tão diferente das tabelas parciais.

(d) Construa um diagrama encadeado que poderá explicar os relacionamentos, identificando a variável interveniente.

(e) Mostre que os dados são, também, consistentes com uma associação espúria e construa o diagrama correspondente. Qual diagrama parece ser mais apropriado? Por quê?

10.14 Nos julgamentos de assassinatos[5] em 20 condados da Flórida em um período de dois anos, foram sentenciados à morte 19 dos 151 assassinos nos quais um branco matou um branco, 0 dos 9 em que um branco matou um negro, 11 dos 63 em que um negro matou um branco e 6 dos 103 em que um negro matou um negro.

(a) Construa tabelas parciais relacionando a raça do réu e a condenação à morte, controladas pela raça da vítima. Nessas tabelas, compare as proporções de réus brancos e negros que receberam a pena de morte.

(b) Construa a tabela bivariada ignorando a raça da vítima. Descreva a associação e compare com (a).

(c) O *paradoxo de Simpson* declara que as associações em todas as tabelas parciais podem ter uma direção diferente do que a da associação na tabela bivariada. Mostre que estes dados satisfazem o paradoxo de Simpson, com réus brancos tendo uma chance menor ou maior de serem condenados à morte do que os réus negros se a raça da vítima for controlada.

(d) Descrevendo como a raça da vítima está associada a cada uma dessas variáveis, explique porque a associação parcial difere da associação bivariada e como isso acontece.

(e) Para essas variáveis, indique se cada um dos diagramas seguintes parece fornecer um modelo razoável. Informe o seu raciocínio. (Aqui, P = pena de morte, R = raça do réu, V = raça da vítima.)

(i.) $V \nearrow P$, $V \searrow R$ (ii.) $R \longrightarrow V \longrightarrow P$

(iii.) $R \searrow P$, $V \nearrow P$ (iv.) $R \longrightarrow P$, $R \searrow V \nearrow P$

10.15 Para os dados de vendas de casas mostrados parcialmente na Tabela 9.4, na página 310, o número de quartos têm uma correlação positiva moderadamente forte com o preço de venda. Controlando o tamanho da casa, entretanto, esta associação diminui consideravelmente.

(a) Explique como isto poderia acontecer, ilustrando com um diagrama e mostrando os efeitos potenciais diretos e indiretos do número de quartos no preço de venda.

(b) Explique o que significa dizer que existe confusão nos efeitos do tamanho da casa e o número de quartos nos seus efeitos no preço de venda.

☑ **Tabela 10.6**

		Renda							
		Alta		Média-Alta		Média-Baixa		Baixa	
		Adm.	Prod.	Adm.	Prod.	Adm.	Prod.	Adm.	Prod.
Partido	Republicano	45	5	100	25	75	300	45	405
	Democrata	405	45	300	75	25	100	5	45

Obs. Adm.: Funcionário Administrativo; Prod.: Operário da produção.

10.16 Para a Tabela 9.16, na página 333, que fornece dados de todos os condados da Flórida para muitas variáveis, existe uma correlação positiva moderada (r = 0,47) entre a taxa de crimes o percentual dos que têm ensino médio completo. O percentual morando em áreas urbanas está também fortemente correlacionado com a taxa de crimes (r = 0,68) e com a taxa de concluintes do ensino médio (r = 0,79).

(a) Explique por que a associação entre a taxa de crimes e a taxa de concluintes do ensino médio poderia desaparecer ou até mesmo mudar de direção, quando controlamos o percentual morando em áreas urbanas.

(b) Sob o controle em (a), se a associação desaparece, que tipo de relacionamento é mais plausível – um relacionamento espúrio ou um relacionamento encadeado? Explique.

10.17 A oposição à disponibilidade legal do aborto é mais forte entre os religiosos do que entre os não religiosos e é também mais forte entre aqueles com atitudes sexuais conservadoras do que entre aqueles com atitudes mais permissivas. Faça um diagrama de como as três variáveis podem estar relacionadas, tratando a atitude em relação ao aborto como a variável resposta. (*Nota*: mais do que um diagrama é plausível.)

10.18 A Tabela 10.17 lista o salário médio, em milhares de dólares, do corpo docente com contrato de nove meses nas instituições de ensino superior em 2003-2004, por gênero e posição acadêmica.

(a) Suponha que o gênero é a variável explicativa. Identifique a variável resposta e a variável controle.

(b) Descreva o relacionamento bivariado entre gênero e salário.

(c) Descreva o relacionamento entre gênero e salário, controlando a posição acadêmica.

(d) A hipótese de interesse para estas variáveis é "controlada a posição acadêmica, o salário anual e o gênero são independentes". Faça um diagrama causal que seja consistente

☑ **Tabela 10.7**

	Posição acadêmica				
Gênero	**Professor**	**Associado**	**Adjunto**	**Aux. de Ensino**	**Geral**
Homens	88,3	63,5	53,7	51,0	67,5
Mulheres	76,7	59,1	49,7	47,4	55,4

Fonte: National Center for Education Statistics, *Digest of Education Statistics*. 2004, Tabela 236.

com essa hipótese. Considere a sua interpretação da parte (c) e comente se a hipótese parece plausível.

(e) A diferença geral de 12,1 mil dólares entre a renda média de homens e mulheres é maior do que a diferença para cada posição acadêmica. Explique como isto pode acontecer.

10.19 A Tabela 10.8 relaciona Y = escore no exame (1 = abaixo da mediana, 2 = acima da mediana) para gênero, controlado pelo assunto do exame (Matemática, Língua). Mostre que o assunto do exame é uma variável supressora.

☑ **Tabela 10.8**

	Matemática		Língua	
Gênero	**y = 1**	**y = 2**	**y = 1**	**y = 2**
Homens	100	50	50	100
Mulheres	50	100	100	50

10.20 Quando analisamos os dados dos distritos censitários na área da grande Los Angeles, não encontramos nenhuma correlação significativa entre os impostos e o tamanho mediano do terreno. Todavia, uma correlação positiva considerável ocorre quando controlamos o percentual dos distritos comerciais. Explique como o percentual dos distritos comerciais poderia ser uma variável supressora se ela estiver correlacionada positivamente com o tamanho mediano do terreno.

10.21 De acordo com a U.S. Census Bureau (Agência do Censo dos Estados Unidos) em 2000, a renda mediana da população era estimada em $29661 para mulheres brancas, $25736 para mulheres negras, $40350 para homens brancos, $30886 para homens negros. Compare a diferença das rendas medianas entre homens e mulheres para (a) sujeitos brancos e (b) sujeitos negros. Se estas são estimativas próximas das medianas da população, explique por que existe interação e descreva a sua natureza.

10.22 Para empregados do nível administrativo mais baixo de uma cadeia de *fast food*, a equação de previsão relacionando Y = renda anual (em milhares de dólares) a X_1 = número de anos de experiência no emprego é igual a $\hat{y} = 14{,}2 + 1{,}1x_1$ para homens e $\hat{y} = 14{,}2 + 0{,}4x_1$ para mulheres. Explique como estas equações mostram evidência de interação estatística.

10.23 Um estudo da associação entre se fumante (sim, não) e se tem algum tipo de câncer (sim, não) tem uma razão de chances de 1,1 para sujeitos com menos do que 30 anos, 2,4 para sujeitos de 30 a 50 anos e 4,3 para sujeitos com mais de 50 anos.

(a) Identifique a variável resposta, a explicativa e a controle.

(b) O estudo mostra evidência de interação? Explique.

10.24 Um estudo com os estudantes da Oregon State University encontrou uma associação entre a frequência de comparecimento à igreja e ser favorável à legalização da maconha. Ambas as variáveis foram mensuradas em categorias ordenadas. Quando o gênero do estudante foi controlado, as medidas do gama para as duas tabelas parciais foram:

Homens: gama = −0,287, erro padrão = 0,081

Mulheres: gama = −0,581, erro padrão = 0,091

(a) Interprete os valores amostrais do gama.

(b) Este resultado mostra um leve grau de _____, visto que a associação é um pouco mais forte para mulheres do que para homens.

(c) Construa e interprete um intervalo de 95% de confiança para a diferença entre os valores dos gamas populacionais.

Conceitos e aplicações

10.25 Considere o arquivo de dados *Student Survey* (Exercício 1.11, na página 25). Construa tabelas parciais relacionando as opiniões sobre o aborto com as opiniões sobre a vida após a morte, controlando a variável frequência aos serviços religiosos mensurada utilizando as duas seguintes categorias (nunca ou ocasionalmente e quase todas as semanas ou toda a semana). Prepare um relatório (a) propondo e interpretando um possível diagrama de flechas, antes de analisar os dados, para os relacionamentos entre as variáveis; (b) interpretando as associações amostrais na tabela bivariada e nas parciais; (c) revisando, se necessário, o diagrama de flechas baseado na evidência dos dados amostrais.

10.26 Para os valores do levantamento de dados dos estudantes (Exercício 1.11) existem pares de variáveis para os quais você espera que a associação desapareça sob o controle de uma terceira variável? Explique.

10.27 Usando a PSG mais recente, construa uma tabela de contingência relacionando o gênero (variável "SEX" da PSG) e identificação partidária ("PARTYID"). Existe ainda uma lacuna no gênero? Controle a ideologia política ("POLVIEWS") formando tabelas parciais para os sujeitos mais conservadores e para os mais liberais. A associação parece persistir para estes sujeitos?

10.28 Suponha que X_1 = educação do pai está positivamente associada a Y = renda do filho aos 40 anos. Entretanto, para a análise de regressão conduzida separadamente para níveis fixos de X_2 = educação do filho, a correlação não difere significativamente de zero. Você acha que isso reflete mais provavelmente um relacionamento encadeado ou um espúrio? Explique.

10.29 A Tabela 10.9 mostra o número médio de filhos de famílias canadenses, classificados por se a família fala inglês ou francês e se residia em Quebec ou em outra província. Considere Y = número de filhos da família, X_1 = principal língua da família e X_2 = província (Quebec, Outra).

(a) Descreva a associação entre Y e X_1, baseado nas médias gerais nessa tabela.

(b) Descreva a associação entre Y e X_1, controlados por X_2.

☑ **Tabela 10.9**

Província	**Inglês**	**Francês**
Quebec	1,64	1,80
Outra	1,97	2,14
Total	1,95	1,85

(c) Explique como é possível que para cada nível da variável província a média seja mais alta para famílias que falam francês, mas a média total seja maior para famílias que falam inglês. (Isto ilustra o *paradoxo de Simpson*. Veja Exercício 10.14.)

10.30 Os escores de matemática da 8ª série da *National Assessment of Educational Progress* (Avaliação Nacional do Progresso Educacional) tinham médias de 277 em Nebraska e de 271 em Nova Jersey. Para estudantes brancos, as médias foram de 281 em Nebraska e de 283 em Nova Jersey. Para estudantes negros, as médias foram de 236 em Nebraska e de 242 em Nova Jersey. Para outros estudantes não brancos, as médias foram de 259 em Nebraska e e 260 em Nova Jersey.[6]

(a) Identifique a variável de grupo especificando os dois estados como uma variável explicativa. Qual é a variável resposta e a controle?

(b) Explique como é possível que Nova Jersey tenha a média mais alta para cada raça, mas Nebraska tenha a média mais alta quando os dados são combinados. (Isto ilustra o *paradoxo de Simpson*.)

10.31 O Exemplo 7.1 (página 216) discutiu um estudo que descobriu que rezar não reduz a incidência de complicações para pacientes de cirurgia coronariana.
(a) Assim como uma associação não implica causalidade, a ausência de associação não implica ausência de causalidade, porque pode haver uma explicação alternativa. Ilustre isso usando esse estudo.
(b) Um resumo desse estudo publicado na *Time Magazine* (4 de dezembro de 2006, p. 87) observou que "as orações feitas por estranhos foram fornecidas pelo clero e eram todas idênticas. Talvez isto tenha impedido que elas fossem verdadeiramente sinceras. Em resumo, os possíveis fatores de confusão do estudo o tornaram bastante limitado". Explique o que significa "possível confusão", no contexto desse estudo.

10.32 Um estudo, que observa sujeitos que dizem que se exercitam regularmente, relatou que somente metade tem doenças graves por ano, em média, do que aqueles que dizem que não se exercitam regularmente. A seção dos resultados do artigo declara: "Analisaremos, a seguir, se a idade era a variável de confusão afetando esta associação". Explique o que esta frase significa e como a idade poderia explicar potencialmente a associação entre exercício e doença.

10.33 Um estudo financiado pela Wobegon Springs Mineral Water, Inc., descobriu que a probabilidade de que um recém-nascido tenha um defeito de nascença é menor para as famílias que compram regularmente água engarrafada do que para as que não compram. Esta associação reflete uma ligação causal entre comprar água engarrafada e a redução de defeitos de nascença? Por que sim ou por que não?

10.34 O percentual de mulheres que tem câncer de mama é maior agora do que no início deste século. Suponha que a incidência do câncer tende a aumentar com a idade e suponha que as mulheres tendem a viver mais agora do que no início do século. Como pode a comparação das taxas de câncer de agora com as do início do século mostrar resultados diferentes destes se controlarmos a idade da mulher?

10.35 O índice de mortalidade bruto é o número de mortes em um ano, pelo tamanho da população, multiplicado por 1000. De acordo com a U.S. Bureau of the Census (Agência do Censo dos Estados Unidos), recentemente o México tinha um índice de mortalidade bruto de 4,6 (isto é, 4,6 mortes por 1000 habitantes) enquanto que os Estados Unidos tinham o índice de mortalidade bruto de 8,4. O índice de mortalidade geral poderia ser maior nos Estados Unidos mesmo se os Estados Unidos tivessem um índice de mortalidade mais baixo do que o México para pessoas de idades específicas? Explique.

10.36 Nos Estados Unidos, a idade mediana dos residentes é mais baixa em Utah. Em cada nível de idade, o índice de mortalidade por doença cardíaca é maior em Utah do que no Colorado; mas, no geral, o índice de mortalidade por doença cardíaca é mais baixo em Utah do que no Colorado. Existem contradições aqui ou isto é possível? Explique.

10.37 Um estudo do relacionamento entre o GPA do ensino médio e o trabalho da mãe (sim, não) levanta a hipótese de que existe uma interação com o gênero do estudante. Controlando o gênero, a Tabela 10.10 mostra os resultados.
(a) Descreva o relacionamento entre o trabalho da mãe e o GPA para as mulheres e para os homens. Esta amostra mostra evidência de interação estatística? Explique.
(b) Um artigo em um periódico escrito sobre o estudo declara: "Tem uma mãe que trabalha fora de casa parece ter efeitos positivos no desempenho da filha no ensino médio, mas nenhum efeito substantivo no desempenho do filho". Explique como a Tabela 10.10 sugere esta interpretação.

☑ **Tabela 10.10** GPA médio *versus* mães com e sem emprego, controlado pelo gênero

Gênero	Mãe com emprego	Mãe sem emprego
Mulheres	2,94	2,71
Homens	2,72	2,74

10.38 Dê um exemplo de três variáveis para as quais o efeito de X_1 em Y seria:
(a) Espúrio, desaparecendo quando X_2 é controlado.
(b) Parte de um relacionamento encadeado, desaparecendo quando uma variável interveniente X_2 é controlada.
(c) Enfraquecido, mas não eliminado, quando X_2 é controlado.
(d) Não afetado quando X_2 é controlado.
(e) Diferente em níveis diferentes de X_2 (isto é, mostrando interação).
(f) De confusão com o efeito de X_2.

10.39 Um estudo sobre o comportamento compulsivo de comprar realizou um levantamento de dados nacional, por telefone, em 2004, com adultos entre 18 anos ou mais[7]. O estudo descobriu que sujeitos com rendas mais baixas tinham maior probabilidade de serem compradores compulsivos. Eles relataram: "Compradores compulsivos não diferem significativamente dos outros respondentes na média do total da fatura do cartão de crédito, mas a renda mais baixa dos compradores compulsivos foi um fator de confusão". Explique o que significa dizer que a renda foi um fator de confusão e explique por que uma comparação da média do total da fatura do cartão de crédito entre compradores compulsivos e não compulsivos poderia mudar dependendo do controle da renda.

10.40 Um estudo recente (na *Behavior Modification*, v. 29, p. 677, 2005) relatou uma correlação de 0,68 entre os escores de um índice de depressão com os escores de um índice que mensurou o consumo de gordura saturada. Verdadeiro ou falso: pode-se concluir que, se você aumentar o consumo de gordura saturada em um desvio padrão, seu grau de depressão aumentará por mais do que a metade de um desvio padrão.

10.41 Um estudo (em *Adolescence*, v. 335, p. 445, 2000) observou uma correlação amostral de 0,45 entre depressão e solidão e −0,74 entre solidão e autoestima. Verdadeiro ou falso: pela regra da cadeia, a correlação amostral entre depressão e autoestima era negativa.

Selecione a(s) melhor(es) resposta(s) nos Exercícios 10.42 a 10.45.

10.42 Para todos os julgamentos sobre homicídios ocorridos na Flórida entre 1976 e 1987, a diferença das proporções de brancos e negros que receberam a pena de morte foi de 0,026 quando a vítima era negra e −0,077 quando a vítima era branca.[8] Isto mostra evidência de:
(a) uma associação espúria.
(b) uma interação estatística.
(c) um relacionamento em cadeia.
(d) Todas as respostas acima.

10.43 A interação estatística se refere a qual das seguintes afirmações?
(a) Existe uma associação entre duas variáveis.
(b) O efeito de uma variável explicativa em uma variável resposta muda consideravelmente sob os níveis de uma variável controle.
(c) A associação parcial é a mesma em cada nível da variável controle, mas é diferente da associação bivariada geral, ignorando a variável controle.
(d) Para um conjunto de três variáveis, cada par de variáveis está associado.
(e) Todas as respostas acima.

10.44 O Exemplo 9.10 (página 310) utilizou um conjunto de dados sobre as vendas de casas para determinar a regressão

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

(a) de uma interação de X_1 e X_2 nos seus efeitos em Y.

(b) de uma associação espúria entre preço de venda e tamanho.

(c) de um relacionamento encadeado, pelo qual "nova" afeta "tamanho" que afeta o "preço de venda".

(d) de que o tamanho da casa não tem um efeito causal no preço.

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

(a) X_1 e X_2 são provavelmente causas independentes de Y.

(b) A associação entre Y e X_1 é provavelmente espúria quando X_2 é controlado.

(c) Visto que ambas as variáveis afetam Y, provavelmente existe interação.

(d) As variáveis provavelmente satisfazem o relacionamento encadeado.

(e) A raça é provavelmente a variável supressora.

(f) Nenhuma das respostas acima.

NOTAS

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

11 REGRESSÃO MÚLTIPLA E CORRELAÇÃO

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

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

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

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

11.1 O MODELO DE REGRESSÃO MÚLTIPLA

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

A função de regressão múltipla

Suponha que existam duas variáveis explicativas, representadas por x_1 e x_2. Como nos capítulos anteriores, usamos a letra minúscula para representar observações ou valores particulares das variáveis. A