CAPÍTULO 8
MODELO DE REGRESSÃO BIVARIADO

RESUMO:

Modelos de regressão fazem o trabalho pesado dos analistas de dados em vários campos das ciências sociais. Iniciamos o capítulo com uma discussão do ajuste de uma reta em um gráfico de dispersão, então passamos para a discussão das inferências adicionais que podem ser feitas quando passamos do coeficiente de correlação para o modelo de regressão com duas variáveis. Também discutimos medidas de ajuste e a natureza do teste de hipótese e da significância estatística nos modelos de regressão. Ao longo deste capítulo, apresentamos importantes conceitos em texto, fórmulas matemáticas e ilustrações gráficas. Este capítulo conclui com uma discussão sobre os pressupostos do modelo de regressão e os requisitos matemáticos mínimos para sua estimação.

8.1 REGRESSÃO BIVARIADA

No capítulo 7, introduzimos três diferentes testes de hipótese bivariados. Neste capítulo adicionamos um quarto, a regressão bivariada. Esse é um importante passo inicial para o modelo de regressão múltiplo – tópico do capítulo 9 –, que consiste em "controlar por" outra variável (Z) quando examinamos a relação entre nossa variável independente de interesse (X) e nossa variável dependente (Y). É crucial desenvolver um entendimento aprofundado da regressão bivariada antes de passarmos para a regressão múltipla. Nas seções que seguem, começamos com uma visão geral do modelo de regressão bivariado, no qual ajustamos uma reta a partir dos nossos dados em um gráfico de dispersão. Então discutimos as incertezas associadas a essa reta e como utilizamos várias medidas dessa incerteza para fazer inferências sobre a população subjacente.

8.2 AJUSTANDO UMA LINHA: POPULAÇÃO – AMOSTRA

A ideia básica da regressão bivariada é que ajustamos a "melhor" reta para o gráfico de dispersão dos nossos dados. Essa reta, que é definida pela sua inclinação e seu intercepto-y, serve como um **modelo estatístico** da realidade. Nesse sentido, a regressão bivariada é bastante diferente das três técnicas de testes de hipótese que introduzimos no capítulo 7; embora essas técnicas permitam o teste de hipótese, elas não produzem um modelo estatístico. Você talvez se lembre, do seu curso de geometria no colégio, que a fórmula para uma reta é expressa por

$$Y = mX + b,$$

em que b é o intercepto-y e m é a inclinação – frequentemente referido como o componente de "aumento e crescimento" da fórmula da linha. Para um aumento de uma unidade em X, m é a quantidade correspondente de aumento em Y (ou redução em Y, se m é negativo). Juntos, esses dois elementos (m e b) são descritos como **parâmetros**[1] da reta. Você talvez se lembre de exercícios de matemática do colegial no qual lhe eram fornecidos os valores de m e b e pediam para que você desenhasse a reta resultante em um gráfico. Isso exemplifica que uma vez que conhecemos esses dois parâmetros da reta, podemos desenhar a linha para qualquer valor que assuma X assuma[2].

Em um modelo de regressão bivariado, representamos o parâmetro do intercepto-y pela letra grega alpha (α) e o da inclinação pela letra grega beta (β)[3]. Como prenunciado por todas as outras discussões sobre variáveis, Y é a variável dependente e X é a variável independente. Nossa teoria sobre a população subjacente na qual estamos interessados é expressa no **modelo de regressão populacional**:

$$Y_i = \alpha + \beta X_i + u_i .$$

Note que nesse modelo existe um componente adicional, u_i, que não corresponde ao que estamos acostumados a observar na fórmula da reta de nossas aulas de geometria. Esse é o termo **estocástico** ou componente "randômico" da nossa variável dependente. Temos esse termo porque não esperamos que todos os pontos dos nossos dados se alinhem perfeitamente em uma reta. Isso se relaciona diretamente a nossa discussão nos capítulos anteriores sobre a natureza probabilística (em oposição a determinística) das teorias causais sobre fenômenos políticos. Estamos, afinal, tentando

[1] Na descrição de uma reta, os parâmetros (m e b, neste caso) são fixos, enquanto as variáveis (X e Y) variam.

[2] Se isso não é familiar para você, ou se deseja meramente refrescar sua memória, você pode fazer o exercício 1 no final deste capítulo antes de continuar a ler.

[3] Diferentes livros sobre regressão utilizam notações ligeiramente diferentes, portanto é importante não assumir que todos os livros utilizam a mesma notação.

explicar processos que envolvem o comportamento humano. Como o ser humano é complexo, certamente haverá uma grande quantidade de ruído aleatório em nossas medidas sobre seu comportamento. Assim, pensamos sobre os valores da nossa variável dependente Y_i como a combinação de um componente sistemático, $\alpha + \beta X_i$ e de um componente estocástico, u_i.

Como temos discutido, raramente trabalhamos com dados populacionais. Em vez disso, utilizamos dados amostrais para fazer inferências sobre uma população subjacente de nosso interesse. Em nosso modelo de regressão bivariado, utilizamos informações sobre o **modelo de regressão amostral** para fazer inferências sobre o modelo de regressão populacional não observado. Para distinguir entre esses dois modelos, colocamos acentos circunflexos (ˇ) acima dos termos do modelo de regressão amostral que estimam os termos para um modelo de regressão populacional não observado. Como eles possuem acentos, definimos $\check{\alpha}$ e $\check{\beta}$ como **parâmetros estimados**. Esses termos são nossos melhores palpites dos parâmetros populacionais não observados α e β.

$$\text{Modelo de regressão amostral: } Y_i = \check{\alpha} + \check{\beta} X_i + \check{u}_i .$$

Note que, em nosso modelo de regressão amostral, α, β e u_i possuem acentos, mas Y_i e X_i não. Isso ocorre uma vez que Y_i e X_i são valores de casos da população que estão na nossa amostra. Desse modo, Y_i e X_i são valores mensurados, em vez de estimados, e os utilizamos para estimar os valores de α, β e u_i. Os valores que definem a linha são os componentes estimados sistemáticos de Y. Para cada valor Y_i, utilizamos $\check{\alpha}$ e $\check{\beta}$ para calcular o valor predito de Y_i, que chamamos de $\check{Y}_i$, em que:

$$\check{Y}_i = \check{\alpha} + \check{\beta} X_i .$$

Essa fórmula também pode ser reescrita em termos de expectativas

$$E(Y \mid X_i) = \check{Y}_i + \check{\alpha} + \check{\beta} X_i ,$$

que significa que o valor esperado de Y para dado Y_i (ou $\check{Y}_i$) é igual a nossa fórmula para a reta da regressão bivariada. Podemos, portanto, dizer que cada Y_i tem um componente sistemático estimado, $\check{Y}_i$, e um componente estocástico estimado, $\check{u}_i$. Assim, podemos escrever nosso modelo como

$$Y_i = \check{Y}_i + \check{u}_i ,$$

e podemos reescrever essa fórmula em termos de $\check{u}_i$ para melhor entender o componente estocástico estimado:

$$\breve{u}_i = Y_i - \breve{Y}_i.$$

Dessa fórmula, conseguimos observar que o componente estocástico estimado $(\breve{u}_i)$ é igual à diferença entre o valor real da variável dependente (Y_i) e o valor da variável dependente predito pelo nosso modelo de regressão bivariado. O componente estocástico estimado também é conhecido como **resíduo**. "Resíduo" é outra palavra para "sobra", e é um termo apropriado, porque $\breve{u}_i$ é o resto de Y_i após termos desenhado a reta definida por $\breve{Y}_i = \breve{\alpha} + \beta X_i$. Outro modo de se referir a $\breve{u}_i$ é pela fórmula $\breve{u}_i = Y_i - \breve{Y}_i$, que é chamada de **termo de erro da amostra**. Porque $\breve{u}_i$ é uma estimativa de u_i, um modo correspondente de se referir a u_i é chamá-lo de **termo de erro da população**.

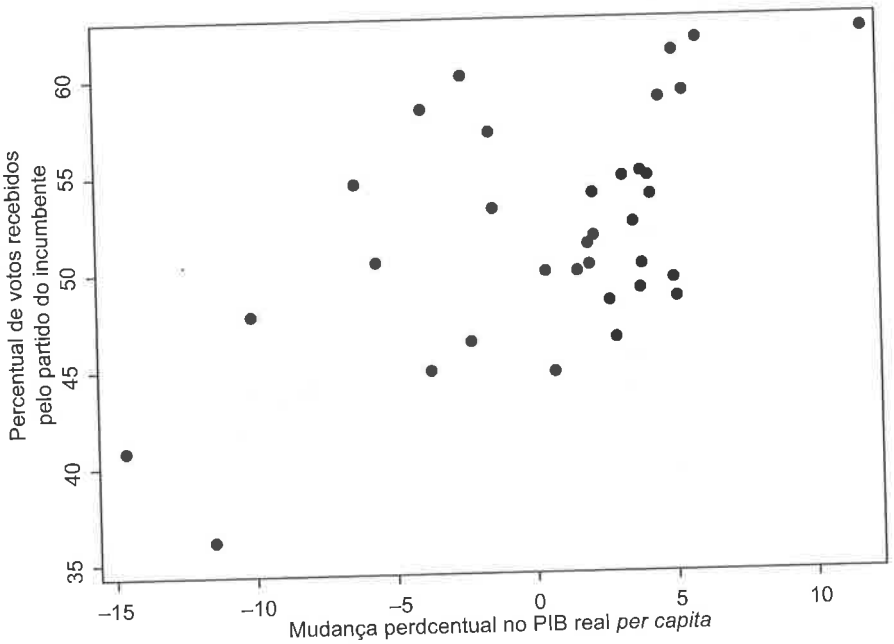

Figura 8.1 – Gráfico de dispersão da mudança no PIB e voto no partido do incumbente.

8.3 QUAL LINHA SE AJUSTA MELHOR? ESTIMANDO A RETA DE REGRESSÃO

Considere o gráfico de dispersão da Figura 8.1. Nossa tarefa é traçar uma linha reta[4] que descreva a relação entre nossa variável independente *X* e nossa variável dependente *Y*. Como desenhamos essa reta? Claramente, queremos desenhar uma reta que passe o mais próximo possível dos nossos casos no gráfico de dispersão. Como

[4] Por "linha reta" queremos dizer uma linha com um uma única inclinação que não mude quando nos movemos da esquerda para a direita no nosso gráfico.

os dados possuem um padrão geral do canto inferior esquerdo para o canto superior direito, sabemos que a inclinação da nossa reta será positiva.

Na Figura 8.2, desenhamos três retas com inclinações positivas – nomeadas A, B e C – no gráfico de dispersão para os dados de crescimento e voto e escrevemos os parâmetros da fórmula correspondente a cada reta no lado direito do gráfico. Então, como decidimos qual reta "melhor" se ajusta aos dados que estamos observando no nosso gráfico de dispersão dos valores de X_i e Y_i? Como estamos interessados em explicar nossa variável dependente, queremos que os valores residuais $\breve{u}_i$, os quais são distâncias verticais entre cada Y_i e o correspondente $\breve{Y}_i$, sejam os menores possíveis. Mas como essas distâncias verticais apresentam valores positivos e negativos, não podemos somente somá-los para cada linha e termos um bom ajuste entre cada linha e nossos dados[5].

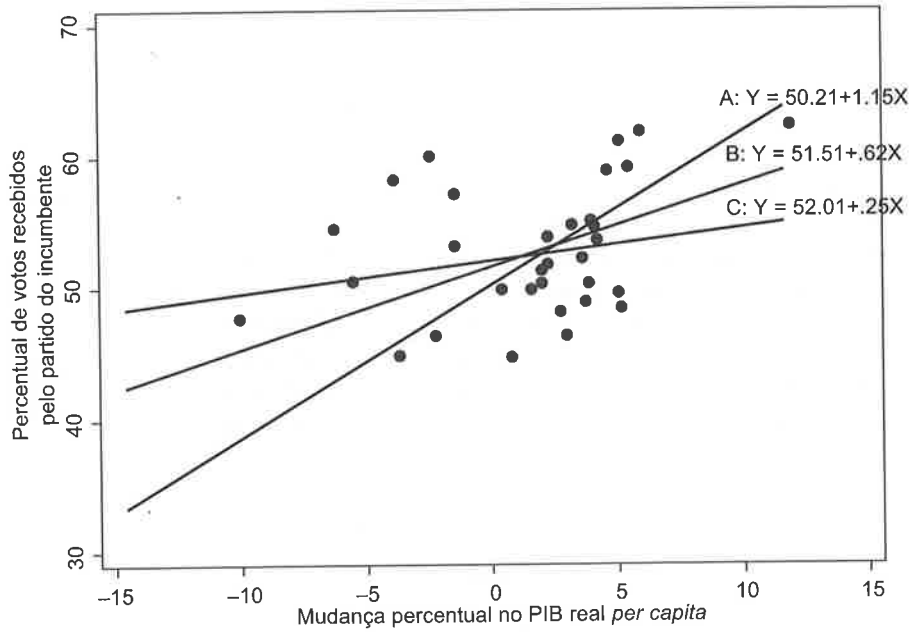

Figura 8.2 – Três retas possíveis.

Então precisamos de um método que avalie o ajuste de cada uma das retas em que os resíduos positivos e negativos não cancelem uns aos outros. Uma possibilidade é somar o valor absoluto total dos resíduos para cada uma das linhas:

$$\sum_{i=1}^{n} |\breve{u}_i| .$$

[5] Inicialmente, poderíamos pensar que gostaríamos de minimizar a soma dos resíduos. Mas a linha que minimiza a soma dos resíduos é, na verdade, uma linha reta paralela ao eixo x. Tal linha não nos ajuda a explicar a relação entre X e Y.

Outra possibilidade é somar todos os resíduos elevados ao quadrado para cada uma das linhas:

$$\sum_{i=1}^{n} \breve{u}_i \dagger.$$

Seja qual for a nossa escolha, queremos escolher a reta que tenha o menor valor total. A Tabela 8.1 apresenta esses cálculos para as três retas da Figura 8.2.

Tabela 8.1 – Medidas dos resíduos totais para três retas diferentes.

Linha	Fórmula paramétrica	$\sum_{i=1}^{n} [\breve{u}_i]$	$\sum_{i=1}^{n} \breve{u}_i^2$
A	$Y = 50,21 + 1,15X_i$	149,91	1086,95
B	$Y = 51,51 + 0,62X_i$	137,60	785,56
C	$Y = 52,01 + 0,21X_i$	146,50	926,16

Nos dois cálculos, podemos observar que a reta B é a que melhor se ajusta aos dados. Embora o cálculo utilizando o valor absoluto seja tão válido quanto o que emprega o valor da soma dos resíduos levado ao quadrado, estatísticos tendem a preferir o último (ambos os métodos identificam a mesma reta como a "melhor"). Assim, traçamos uma reta que minimiza a soma dos resíduos ao quadrado $\sum_{i=1}^{n} \breve{u}_i^2$. Esse método de estimação de parâmetros da regressão é conhecido por regressão dos **mínimos quadrados ordinários (MQO)**. Para uma regressão MQO bivariada, as fórmulas para a estimação dos parâmetros de uma reta que satisfazem esse critério são[6]

$$\breve{\beta} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2},$$

$$\alpha = \bar{Y} - \breve{\beta}\bar{X}.$$

Se examinarmos a fórmula para o $\breve{\beta}$, podemos observar que o numerador é o mesmo que o numerador para o cálculo da covariância entre X e Y. Assim, a lógica de como cada um dos casos contribui nessa fórmula, como exposta na Figura 8.2, é a mesma. O denominador na fórmula para o $\breve{\beta}$ é a soma dos desvios dos valores de X_i em relação à média de $X(\bar{X})$ elevada ao quadrado. Assim, para uma dada covariância entre X e Y, quanto maior (menor) for a dispersão de X, menor (maior) será o parâmetro de inclinação da reta de regressão.

Uma das propriedades matemáticas da regressão dos MQO é que a reta produzida pelos parâmetros estimados perpassa os valores das médias de X e Y. Isso torna a es-

[6] As fórmulas para estimação dos parâmetros do MQO são obtidas a partir da definição que o valor da soma dos resíduos ao quadrado é igual a zero e do uso de cálculo diferencial para a solução dos valores de $\breve{\beta}$ e $\breve{\alpha}$.

timação de $\breve{\alpha}$ bastante simples. Se começarmos no ponto definido pelo valor médio de X e o valor médio de Y e então utilizarmos o parâmetro de inclinação estimado ($\breve{\beta}$) para desenhar uma reta, o valor de X em que Y é igual a zero é o $\breve{\alpha}$. A Figura 8.3 mostra a reta da regressão MQO em um gráfico de dispersão. Podemos observar a partir desse gráfico que a reta da regressão MQO passa pelo ponto em que a linha que descreve o valor médio de X encontra a linha que descreve o valor médio de Y.

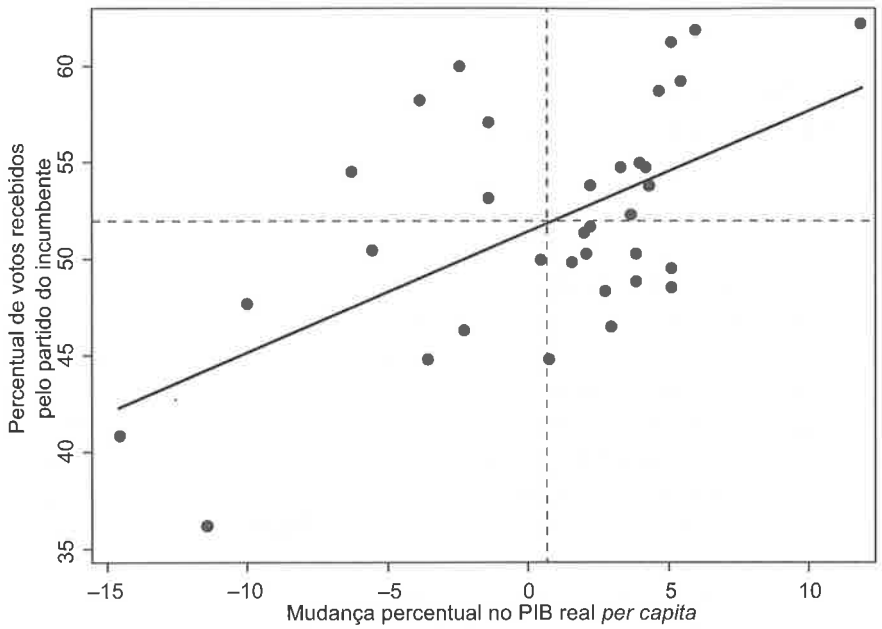

Figura 8.3 – Linha de regressão MQO em gráfico de dispersão com quadrantes de média delimitados.

Utilizando os dados apresentados na Tabela 7.10 na nossa fórmula, calculamos $\breve{\alpha} = 51{,}51$ e $\breve{\beta} = 0{,}62$, fazendo nossa fórmula da reta de regressão igual a $Y = 51{,}51 + 0{,}62X$. Para pensar o que isso nos diz sobre a política, primeiro é preciso lembrar que Y é o percentual de votos recebido pelo candidato do partido do presidente em exercício e que X é o crescimento real *per capita* do PIB. Então, se nossa medida de crescimento é igual a zero, esperaríamos que o partido do presidente em exercício obtivesse 51,51% dos votos. Se o crescimento não for igual a zero, multiplicamos o valor do crescimento por 0,62 e adicionamos (ou subtraímos, se o crescimento for negativo) o resultado de 51,51 para obter nosso melhor palpite para o valor da variável "voto". Caminhar para a direita ou para a esquerda ao longo da reta de regressão para nossa amostra na Figura 8.3 significa que estamos aumentando ou diminuindo o valor da variável "crescimento". Para cada movimento para direita-esquerda, observamos um aumento ou decréscimo no valor esperado para o percentual de votos do partido do incumbente. Se retornarmos a lógica de conceber a inclinação como um movimento de aumento e crescimento, nosso parâmetro estimado de inclinação responde à pergunta de qual é a mudança esperada em Y quando observamos o aumento de

uma unidade em X. Em outras palavras, espera-se que o aumento de uma unidade em nossa variável independente, "crescimento", leve a um aumento de 0,62 na nossa variável dependente, "voto no partido do incumbente"[7].

Podemos afirmar a partir da Figura 8.3 que existem pontos que ficam acima e abaixo da nossa reta de regressão. Portanto, sabemos que nosso modelo não se ajusta perfeitamente ao mundo real. Na próxima seção, discutiremos uma série de inferências que podemos fazer sobre a incerteza associada ao modelo de regressão da nossa amostra.

8.4 MENSURANDO NOSSA INCERTEZA SOBRE A RETA DA REGRESSÃO DE MQO

Como vimos nos capítulos 6 e 7, inferências sobre a população subjacente feitas a partir de dados de *survey* possuem diferentes graus de incerteza. No capítulo 7, discutimos o papel do valor-*p* para expressar essa incerteza. Em um modelo de regressão de MQO, temos diferentes modos para mensurar nossa incerteza. Discutimos essas medidas, primeiramente, em termos do ajuste geral entre X e Y e, então, discutimos a incerteza sobre os parâmetros individuais (nossa incerteza sobre os parâmetros individuais é utilizada para testar nossas hipóteses). Ao longo desta discussão, utilizaremos nosso exemplo da reta de regressão ajustada para os dados das eleições presidenciais americanas que empregamos para testar a teoria do voto econômico. Os resultados numéricos desse modelo produzidos pelo *software* estatístico Stata estão expostos na Figura 8.4. Esses resultados numéricos podem ser particionados em três áreas separadas. A tabela no canto superior esquerdo da Figura 8.4 nos fornece as medidas de variação do nosso modelo. O conjunto de estatísticas listadas no canto superior direito da Figura 8.4 nos dá um conjunto de estatísticas de resumo sobre o modelo como um todo. Ao longo da parte inferior da Figura 8.4 temos uma tabela de estatísticas sobre os parâmetros estimados do modelo. O nome da variável dependente, "VOTE", é exposto no topo dessa tabela. Abaixo, temos os nomes das nossas variáveis independentes, "GROWTH" e "_cons", que é a abreviatura de *constant* [constante] (outro nome para o intercepto-*y*), também conhecido como $\breve{\alpha}$. À direita na tabela da parte inferior da Figura 8.4, observamos que a próxima coluna tem o nome "Coef.", que é a abreviatura para *coefficient* [coeficiente], outro nome para o parâmetro estimado de inclinação. Nessa coluna, observamos os valores de $\breve{\beta}$ e $\breve{\alpha}$, que são 0,62 e 51,51 quando arredondamos os resultados para a segunda casa decimal[8].

---

[7] Tome cuidado para não inverter as variáveis independente e dependente na descrição dos resultados. Não é correto interpretar os resultados afirmando que "para cada 0,62 ponto de mudança na taxa de crescimento da economia dos Estados Unidos, devemos esperar observar, na média, o aumento de 1% no percentual de votos de um candidato do partido do presidente em exercício". Certifique-se de que você consegue observar a diferença entre essas descrições.

[8] A escolha de quantas casas decimais deve ser feita com base no valor da variável dependente. Neste caso, como nossa variável dependente é o percentual de votos, escolhemos a segunda cada decimal. Cientistas políticos usualmente não consideram resultados eleitorais além das duas primeiras casas decimais.

```
. reg VOTE GROWTH

      Source |       SS       df       MS              Number of obs =      34
-------------+------------------------------           F(  1,    32) =   15.70
       Model |  385.31241      1  385.312461            Prob > F      =  0.0004
    Residual |  785.539343     32  24.5481045           R-squared     =  0.3291
-------------+------------------------------           Adj R-squared =  0.3081
       Total |  1170.8518      33  35.4803577           Root MSE      =  4.9546

------------------------------------------------------------------------------
        VOTE |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      GROWTH |   .6249078   .1577315     3.96   0.000     .3036193    .941963
       _cons |   51.50816   .8569026    60.11   0.000     49.76271    53.25361
------------------------------------------------------------------------------
```

Figura 8.4 – Resultados do Stata para o modelo de regressão bivariada VOTE = $\alpha + \beta \times$ GROWTH.

8.4.1 QUALIDADE DO AJUSTE – RAIZ DO ERRO QUADRÁTICO MÉDIO (*ROOT MEAN-SQUARED ERROR*)

Medidas da qualidade geral do ajuste entre um modelo de regressão e a variável dependente são chamadas de medidas de qualidade de ajuste. Uma das medidas mais intuitivas (apesar do nome) é a raiz do erro quadrático médio (*root mean-squared error – root MSE*). Essa estatística é algumas vezes referida como o erro-padrão do modelo de regressão. Ela proporciona uma medida da precisão média do modelo utilizando a mesma métrica da variável dependente. Essa estatística ("*Root MSE*" na Figura 8.4) é calculada por:

$$root\ MSE = \sqrt{\frac{\sum_{i=1}^{n} \breve{u}_i^2}{n}}.$$

Na fórmula, os valores são elevados ao quadrado e depois é feita a raiz quadrada da quantidade para lidar com o fato de que alguns dos resíduos serão positivos (pontos para os quais Y_i está acima da reta de regressão) e alguns serão negativos (pontos para os quais Y_i está abaixo da reta de regressão). Uma vez que tenhamos feito isso, podemos observar que essa estatística é basicamente a distância média entre os pontos que representam nossos dados e a reta de regressão.

Para os resultados numéricos descritos na Figura 8.4, podemos observar que o root MSE para nosso modelo bivariado do voto no partido do incumbente é de 4,95. Esse valor é encontrado na sexta linha da coluna de resultados do lado direito da Figura 8.4. Ele indica que, na média, a variação dos valores preditos de nosso modelo é de 4,95 pontos percentuais em relação aos votos ganhos pelo partido do incumbente. Vale enfatizar que o *root MSE* sempre é expresso na métrica em que a variável dependente está mensurada. A única razão para esse valor corresponder a um percentual é porque a nossa variável dependente é o percentual de votos.

8.4.2 QUALIDADE DO AJUSTE: R^2

Outro indicador de qualidade de ajuste é a **estatística *R*-quadrado** (tipicamente escrita como R^2). A estatística R^2 varia entre 0 e 1 e indica a proporção da variação da variável dependente que é explicada pelo modelo. A ideia básica da estatística R^2 é mostrada na Figura 8.5, que consiste em um diagrama de Venn que descreve a variação em X e Y, assim como a covariação entre X e Y. A ideia por trás desse diagrama é que descrevemos a variação de cada variável com um círculo. Quanto maior o círculo, maior a variação. Nesse diagrama, a variação de Y é composta por duas áreas, a e b, e a variação de X consiste nas áreas b e c. A área a representa a variação em Y que não é relacionada à variação em X, e a área b representa a covariação entre X e Y.

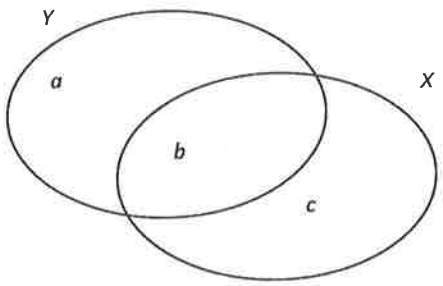

Figura 8.5 – Diagrama de Venn da variância e da covariância para X e Y.

Em um modelo bivariado, a área a é o resíduo ou a variação estocástica de Y. A estatística R^2 é igual à área b sobre a variação total em Y, que é igual à soma das áreas a e b. Assim, quanto menor o valor da área a e maior o valor da área b, maior é a estatística R^2. A fórmula para a variação total em Y (áreas a e b na Figura 8.5), também conhecida como a soma total dos quadrados, é dada por:

$$\text{TSS}=\sum_{i=1}^{n}(Y_i-\bar{Y})^2 .$$

A fórmula para a variação residual de Y, a área que não é explicada por X, chamada de soma dos resíduos quadrados (RSS), é dada por:

$$\text{RSS}=\sum_{i=1}^{n}\check{u}_i^2 .$$

Uma vez que tenhamos calculado essas duas quantidades, podemos calcular a estatística R^2:

$$\text{R}^2=1=\frac{\text{RSS}}{\text{TSS}} .$$

A fórmula para a contraparte de TSS que não é a RSS, chamada de soma dos quadrados do modelo (MSS), é dada por:

$$\text{MSS}=\sum_{i=1}^{n}(\check{Y}_i-\bar{Y})^2 .$$

Isso também pode ser usado para calcular o R^2 por:

$$\text{R}^2=\frac{\text{MSS}}{\text{TSS}} .$$

A partir dos resultados numéricos descritos na Figura 8.4, podemos observar que a estatística R^2 para nosso modelo bivariado do voto no partido do incumbente é 0,329. Esse número aparece na quarta linha da coluna de resultados do lado direito da Figura 8.4. Ele indica que nosso modelo explica quase 33% da variação da variável dependente. Também podemos observar na Figura 8.4 os valores para MSS, RSS e TSS sob a coluna chamada "SS", no canto superior esquerdo da tabela.

8.4.3 ESTA QUALIDADE DO AJUSTE É "BOA"?

Uma pergunta lógica a se fazer quando observamos uma medida de qualidade do ajuste é "Qual seria um valor bom ou ruim para o *root MSE* e/ou R^2?" Essa não é uma pergunta fácil de ser respondida. Em parte, a resposta depende do que você pretende com seu modelo. Se você está tentando prever resultados eleitorais, digamos que prever o resultado com um erro médio de 4,95 pode não ser muito bom. Afinal, a maioria das eleições presidenciais é bastante apertada, portanto, 4,95% são muitos votos. De fato, podemos observar que em treze das 34 eleições que estamos examinando o candidato eleito venceu por uma margem menor que 4,95%, o que faz com que nosso modelo não seja apropriado para mais de um terço da nossa amostra de eleições. Por outro lado, podemos dizer que somos capazes de chegar perto do resultado e que, em termos do R^2, explicamos quase 33% da variação do voto no partido do presidente em exercício para as eleições entre 1876 e 2008 utilizando apenas uma medida econômica. Quando começamos a pensar sobre todas as diferenças da estratégia de campanha, personalidades, escândalos, guerras e todo o resto que não está nesse modelo simples, o nível de precisão é bastante expressivo. De fato, esse modelo nos informa algo bastante notável sobre a política nos Estados Unidos: que a economia é muito importante para explicar os resultados eleitorais.

8.4.4 INCERTEZA SOBRE OS COMPONENTES INDIVIDUAIS DO MODELO DE REGRESSÃO AMOSTRAL

Antes de iniciarmos o conteúdo desta seção, queremos alertá-lo sobre a existência de uma grande quantidade de fórmulas matemáticas nela. Para utilizar uma metáfora familiar, à medida que você for se confrontando com as fórmulas nesta seção, é im-

portante que foque os contornos da floresta e não fique preso aos detalhes das muitas árvores com as quais deparará ao longo do caminho. Em vez de memorizar cada fórmula, concentre-se em entender o que faz com que cada um dos resultados gerados pelas fórmulas seja maior ou menor.

Uma parte crucial da incerteza no modelo MQO de regressão é o grau de incerteza sobre a estimação dos valores dos parâmetros individuais da população a partir do modelo de regressão amostral. Podemos utilizar a mesma lógica que discutimos no capítulo 6 sobre fazer inferências a partir da amostra dos valores da população para cada um dos parâmetros individuais de um modelo de regressão amostral.

Uma medida que é utilizada para estimação sobre a incerteza de cada um dos parâmetros populacionais é a variância estimada do componente estocástico populacional, u_i. Essa variância não observada, σ^2, é estimada a partir dos resíduos $(\breve{u}_i)$, utilizando a fórmula abaixo, após os parâmetros da regressão com os dados da amostra terem sido estimados:

$$\breve{\sigma}^2 = \frac{\sum_{i=1}^{n} \breve{u}_i^2}{n-2}.$$

Analisando a fórmula, podemos observar dois componentes que desempenham um papel na determinação da magnitude do valor estimado. O primeiro componente é o valor individual dos resíduos $(\breve{u}_i)$. Lembre-se que esses valores (calculados por $u_i = Y_i - \breve{Y}_i$) são a distância vertical entre cada valor Y_i observado e a reta de regressão. Quanto maiores forem esses valores, mais distantes estão os casos individuais da reta de regressão. O segundo componente dessa fórmula é o n, o tamanho da amostra. A essa altura, você deve estar familiarizado com a ideia de que quanto maior o tamanho da amostra, menor é a variância estimada. Esse é o caso da nossa fórmula para $\breve{\sigma}^2$.

Uma vez que tenhamos estimado $\breve{\sigma}^2$, a variância e os erros-padrão para o parâmetro de inclinação estimados $(\breve{\beta})$ são estimados utilizando as seguintes fórmulas:

$$\text{var}(\breve{\beta}) = \frac{\breve{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})\dagger},$$

$$\text{se}(\breve{\beta}) = \sqrt{\text{var}(\breve{\beta})} = \frac{\breve{\sigma}^2}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})\dagger}}.$$

Ambas as fórmulas podem ser divididas em dois componentes que determinam a magnitude dos valores estimados. No numerador, temos o valor de $\breve{\sigma}^2$. Então, quanto maior for esse valor, maior será a variância e o erro-padrão do parâmetro de inclinação. Isso faz sentido, porque quanto mais distantes estiverem os pontos que representam nossos dados da reta de regressão, menos confiança teremos no valor da inclinação. Se observarmos o denominador nessa equação, temos o termo $\sum_{i=1}^{n}(X_i - \bar{X})^2$, que é uma medida da variação dos valores de X_i em torno da média de $X(\bar{X})$. Quanto maior essa variação, menor será a variância e o erro-padrão do parâmetro de inclinação estimado. Essa é uma propriedade importante; em termos do mundo real, isso

significa que quanto maior a variação de X, mais precisamente conseguiremos estimar a relação entre X e Y.

A variância e o erro-padrão do parâmetro do intercepto ($\check{\alpha}$) são estimados pelas seguintes fórmulas:

$$\text{var}(\check{\alpha}) = \frac{\check{\sigma}^2 \sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} (X_i - \bar{X})^2},$$

$$\text{se}(\check{\alpha}) = \sqrt{\text{var}(\check{\alpha})} = \sqrt{\frac{\check{\sigma}^2 \sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} (X_i - \bar{X})^2}}.$$

A lógica para os componentes dessas fórmulas é ligeiramente mais complicada, porque podemos observar que a soma dos valores de X_i ao quadrado aparece no numerador. Observamos, contudo, que o denominador contém a medida da variação dos valores X_i ao redor da média ($\bar{X}$) multiplicada por n, o número de casos. Assim, a mesma lógica básica se mantém para estes termos: quanto maiores os valores de $\check{u}_i$ forem, maiores serão a variância e o erro-padrão do parâmetro do intercepto estimado; e quanto maior a variação dos valores de X_i em torno da sua média, menores serão a variância e o erro-padrão do parâmetro de intercepto estimado.

Menos óbvio – mas não menos verdade – é o fato de que, nas fórmulas anteriores, quanto maior o tamanho das amostras, menores serão os erros-padrão estimados[9]. Então, assim como aprendemos sobre os efeitos do tamanho da amostra quando calculamos o erro-padrão da média no capítulo 6, existe um efeito idêntico aqui. Amostras de maior tamanho produzirão, mantendo os demais fatores iguais, menores erros-padrão para os coeficientes estimados da nossa regressão.

8.4.5 INTERVALO DE CONFIANÇA PARA OS PARÂMETROS ESTIMADOS

No capítulo 6 discutimos como utilizamos a distribuição normal (apoiada pelo teorema do limite central) para estimar intervalos de confiança para a média populacional não observada a partir de dados amostrais. Empregamos os mesmos passos lógicos para estimar os intervalos de confiança para os parâmetros não observados do modelo de regressão populacional por meio dos resultados da nossa regressão com dados amostrais. As fórmulas para estimação de intervalos de confiança são:

$$\check{\beta} \pm [t \times \text{se}(\check{\beta})],$$
$$\check{\alpha} \pm [t \times \text{se}(\check{\alpha})],$$

[9] Isso é verdade porque o numerador da expressão contém $\check{\sigma}$, que, como previamente observado, tem o tamanho da amostra, n, como seu numerador.

em que o valor para t é determinado a partir da tabela-t disponível no Apêndice B. Então, por exemplo, se queremos calcular um intervalo de confiança de 95%, significa que devemos buscar na coluna pelo valor de 0,025[10]. Uma vez que tenhamos determinado a coluna apropriada, selecionamos o valor que corresponde ao grau de liberdade. O grau de liberdade para o teste-t é igual ao número de observações (n) menos o número de parâmetros estimados (k). No caso de nosso modelo de regressão apresentado na Figura 8.4, temos $n = 34$ e $k = 2$, então nosso grau de liberdade é igual a 32. Observando a coluna do valor 0,025, temos que na linha para $\text{gl} = 30$ o valor de $t = 2{,}042$. Contudo, como nosso grau de liberdade é igual a 32, o valor-t que deixa 0,025 em cada uma das caudas é igual a 2,037[11]. Assim, nossos intervalos de confiança de 95% são:

$$\breve{\beta} \pm [t \times \text{se}(\breve{\beta})] = 0{,}6249078 \pm (2{,}037 \times 0{,}1577315) = 0{,}30 \text{ a } 0{,}94,$$
$$\breve{\alpha} \pm [t \times \text{se}(\breve{\alpha})] = 51{,}50816 \pm (2{,}037 \times 0{,}8569026) = 49{,}76 \text{ a } 53{,}25.$$

Esses valores estão expostos no canto inferior direito da tabela apresentada na parte inferior da Figura 8.4.

A abordagem tradicional do teste de hipótese utilizando a regressão de MQO é que especificamos uma hipótese nula e uma **hipótese alternativa** e, então, comparamos as duas. Embora possamos testar hipóteses sobre os parâmetros de inclinação ou do intercepto, usualmente estamos mais preocupados com testes sobre o parâmetro de inclinação. Em particular, normalmente estamos preocupados com o teste de que o parâmetro de inclinação para a população é igual a zero. A lógica desse teste de hipótese é bastante próxima à lógica dos testes bivariados introduzidos no capítulo 7. Primeiro, observamos o parâmetro de inclinação da amostra, que consiste em uma estimativa do parâmetro de inclinação da população. A partir do valor desse parâmetro, do intervalo de confiança e do tamanho da nossa amostra, avaliamos quão provável é observarmos o valor da inclinação para a amostra se o valor verdadeiro, mas não observado na população, for igual a zero. Se a resposta for "bastante provável", então concluímos que o parâmetro de inclinação na amostra é igual a zero.

Para entender por que frequentemente estabelecemos para nossa comparação que o valor da inclinação é zero, pense sobre o que isso representa na fórmula da reta. Lembre-se que a inclinação é a mudança em Y para o aumento em uma unidade de X. Se a mudança é igual a zero, então não existe covariação entre X e Y e, portanto, falhamos em superar o terceiro obstáculo causal.

Esses tipos de teste podem ser realizados utilizando apenas uma ou as duas caudas da distribuição normal. A maioria dos programas estatísticos reportam os resultados para o teste bicaudal de que o parâmetro em questão não é igual a zero. Apesar disso,

---

[10] Para entender isso, pense novamente no capítulo 6, quando introduzimos os intervalos de confiança. Um intervalo de confiança de 95% significaria que deixaríamos um total de 5% nas caudas da curva normal. Como temos duas caudas, utilizamos a coluna com valor de 0,025.

[11] O valor exato de t é calculado automaticamente pelos *softwares* estatísticos. Para uma ferramenta on-line que fornece os valores exatos, conferir: <http://www.stat.tamu.edu/~west/applets/tcal.html>.

muitas teorias em ciência política são mais apropriadamente traduzidas em testes unicaudais de hipóteses, que algumas vezes são referidos como testes de hipótese "direcionais". Revisamos esses dois tipos de testes de hipótese nas próximas seções utilizando nossa regressão da Figura 8.4.

8.4.6 TESTES DE HIPÓTESE BICAUDAIS

A forma mais comum de teste estatístico de hipótese sobre os parâmetros de uma regressão MQO é o teste bicaudal para a hipótese de que o parâmetro de inclinação é igual a zero. Ele é expresso como:

$$H_0 : \beta = 0,$$
$$H_1 : \beta \neq 0,$$

em que H_0 é a hipótese nula e H_1 é a hipótese alternativa. Note que essas duas hipóteses rivais são expressas em termos do parâmetro de inclinação para o modelo de regressão populacional. Para examinar qual dessas hipóteses tem apoio dos dados, calculamos a **razão-*t*** na qual β é definido a partir do valor especificado na hipótese nula (neste caso, zero, porque $H_0 : \beta = 0$), que é representado como β^*:

$$t_{n-k} = \frac{\breve{\beta} - \beta^*}{\text{se}(\breve{\beta})}.$$

Para o parâmetro de inclinação do modelo bivariado de regressão apresentado na Figura 8.4, podemos calcular isso por:

$$t_{32} = \frac{\breve{\beta} - \beta^*}{\text{se}(\breve{\beta})} = \frac{0,6249078 - 0}{0,1577315} = 3,96.$$

A partir do que temos visto nos capítulos anteriores, podemos dizer que a razão-*t* calculada é bastante grande. Lembre-se que o padrão comum de significância estatística nas ciências sociais é que o valor-*p* seja menor que 0,05. Se observarmos o valor para grau de liberdade igual a 30 no Apêndice B, temos que, para ter um valor-*p* menor que 0,05, precisamos de uma razão-*t* de 2,042 ou maior (2,037 se utilizarmos o valor exato do grau de liberdade). Claramente superamos esse padrão[12]. De fato, se observarmos na coluna mais à direita do Apêndice B quando o grau de liberdade é igual a 30, vemos que a estatística-*t* calculada excede o valor *t* necessário para *p* menor que 0,002 (na coluna nomeada "0,001" temos o valor de 3,385 para graus de liberdade igual a 30). Isso significa que é extremamente improvável que H_0 tenha apoio, o que, por sua vez, aumenta nossa confiança em H_1. Se observamos a tabela na parte inferior da Figura 8.4, podemos observar que a estatística-*t* e o valor-*p* correspondente para esse

[12] Como esse é um teste de hipótese bicaudal, para o padrão $p < 0,05$ precisamos olhar a coluna com o nome "0,025". Isso ocorre porque, nesse caso, teremos 0,025 em cada uma das caudas.

teste de hipótese são apresentados na quarta e na quinta coluna da linha GROWTH. Vale notar que, embora o valor-p reportado seja 0,000, isso não significa que a probabilidade da hipótese nula ser verdade seja realmente igual a zero. Em vez disso, significa que este é um número extremamente pequeno que é arredondado como zero quando é reportado com três casas decimais.

Utilizamos exatamente a mesma lógica para testar hipóteses sobre o parâmetro do intercepto-y. A fórmula para a estatística-t é

$$t_{n-k} = \frac{\breve{\alpha} - \alpha^*}{\text{se}(\breve{\alpha})}.$$

Na Figura 8.4, observamos os cálculos para as seguintes hipóteses nula e alternativa:

$$H_0 : \alpha = 0,$$
$$H_1 : \alpha \neq 0.$$

A razão-t resultante é um gritante 60,11! Isso faz sentido quando pensamos sobre essa quantidade em termos do mundo real. Lembre-se que o intercepto-y é o valor esperado da variável dependente Y quando a variável independente X é igual a zero. Em nosso modelo, isso significa que queremos saber o valor esperado do voto no partido do presidente em exercício quando o crescimento é igual a zero. Isso porque, mesmo quando a economia está encolhendo, existem alguns partidários obstinados que votarão no partido do incumbente. Assim, faz sentido que a hipótese nula $H_0 : \alpha = 0$ seja bastante fácil de ser rejeitada.

Talvez uma hipótese nula mais interessante seja que os incumbentes obterão 50% dos votos se o crescimento for igual a zero. Nesse caso,

$$H_0 : \alpha = 50,$$
$$H_1 : \alpha \neq 50.$$

A estatística-t calculada correspondente é

$$t_{32} = \frac{\breve{\alpha} - \alpha^*}{\text{se}(\breve{\alpha})} = \frac{51,50816 - 50}{0,8569026} = 1,76.$$

Examinando a linha em que temos os valores da estatística-t quando o grau de liberdade é igual a 30, temos que a estatística-t é menor que 2,042, que equivale ao valor-p < 0,05 (valor na coluna nomeada como "0,025"), mas é maior do que 1,697, que equivale ao valor-p < 0,10 (valor na coluna nomeada como "0,05"). Com uma tabela-t mais detalhada ou um computador, poderíamos calcular o valor-p exato para esse teste de hipótese, que é 0,09. Com esses resultados, estamos em uma área um pouco nebulosa. Podemos ter bastante confiança de que o intercepto não é igual a 50, mas podemos apenas rejeitar a hipótese nula ($H_0 : \alpha = 50$) ao nível de confiança de 0,10 em vez do padrão amplamente aceito para a significância estatística de 0,05. Pensemos por um segundo sobre nosso interesse no intercepto com valor igual a 50.

Mesmo que o teste para a hipótese alternativa que acabamos de realizar ($H_0 : \alpha \neq 50$) seja do nosso interesse, poderíamos nos interessar ainda em saber se o partido do presidente em exercício venceria as eleições se o crescimento fosse igual a zero? Antes de abordarmos essa questão, é preciso explicar a relação entre intervalos de confiança e testes de hipótese bicaudais.

8.4.7 A RELAÇÃO ENTRE INTERVALOS DE CONFIANÇA E TESTES DE HIPÓTESE BICAUDAIS

Nas últimas três seções, introduzimos o intervalo de confiança e o teste de hipótese como duas das formas de inferir os parâmetros da população a partir do modelo de regressão para a amostra. Esses dois métodos para fazer inferências estão matematicamente relacionados. Isso ocorre pois ambos são baseados na tabela-*t*. A relação entre eles é tal que, se o intervalo de confiança de 95% não incluir um determinado valor, então a hipótese nula de que o parâmetro da população é igual a tal valor (em um teste de hipótese bicaudal) terá um valor-*p* menor que 0,05. Podemos observar isso para cada um dos três testes que discutimos na seção sobre testes de hipótese bicaudais:

- Porque o intervalo de confiança de 95% para o nosso parâmetro de inclinação não inclui o valor 0, o valor-*p* para o teste de hipótese de que $\beta = 0$ é menor que 0,05.
- Porque o intervalo de confiança de 95% para o nosso parâmetro do intercepto não inclui o valor 0, o valor-*p* para o teste de hipótese de que $\alpha = 0$ é menor que 0,05.
- Porque o intervalo de confiança de 95% para o nosso parâmetro do intercepto inclui o valor 50, o valor-*p* para o teste de hipótese de que $\alpha = 50$ é maior que 0,05.

8.4.8 TESTE DE HIPÓTESE UNICAUDAL

Como pontuamos nas seções anteriores, a forma mais comum de teste estatístico de hipótese para os parâmetros de um modelo de regressão de MQO é o teste bicaudal para a hipótese nula de que o parâmetro de inclinação é igual a zero. Não é por acaso que esse é o tipo mais comum. Por padrão, a maioria dos *softwares* estatísticos reporta os resultados utilizando esse tipo de teste. Na realidade, porém, a maioria das hipóteses em ciência política são de que um parâmetro é positivo ou negativo, e não que o parâmetro é diferente de zero. Chamamos esse tipo de hipótese de **hipótese direcional**. Considere, por exemplo, no nosso exemplo da teoria do voto econômico, como a traduziríamos em uma hipótese sobre o parâmetro do intercepto. Nossa teoria é que um melhor desempenho da economia levará a um aumento do percentual de votos recebidos pelo candidato do partido do incumbente. Em outras palavras, esperamos observar uma relação positiva entre crescimento econômico e o percentual de votos do partido do incumbente, o que significa que esperamos que β seja maior que zero.

Expressamos uma hipótese direcional por:

$$H_0 : \beta \leq 0,$$
$$H_1 : \beta > 0,$$

em que H_0 é a hipótese nula e H_1 é a hipótese alternativa. Como ocorre com o teste bicaudal, essas duas hipóteses rivais são expressas em termos do parâmetro de inclinação para o modelo de regressão da população. Para testar qual dessas hipóteses é respaldada pelos dados, calculamos a estatística-*t* em que β é definido como igual ao valor especificado na hipótese nula[13] (neste caso, zero, porque $H_0 : \beta \leq 0$), que é representado na fórmula abaixo por:

$$t_{n-k} = \frac{\breve{\beta} - \beta^*}{\text{se}(\breve{\beta})}.$$

Para o parâmetro de inclinação no modelo de regressão bivariado apresentado na Figura 8.4, calculamos a estatística-*t* por:

$$t_{32} = \frac{\breve{\beta} - \beta^*}{\text{se}(\breve{\beta})} = \frac{0{,}6249078 - 0}{0{,}1577315} = 3{,}96.$$

Esses cálculos parecem familiares para você? Deveriam, porque essa estatística-*t* é calculada exatamente do mesmo modo que a estatística-*t* para o teste bicaudal desse parâmetro foi calculada. As diferenças advêm de como utilizamos a tabela-*t* do Apêndice B para chegar ao valor-*p* apropriado para o teste de hipótese. Como este é um teste de hipótese unicaudal, utilizamos a coluna nomeada como "0,05" em vez da coluna nomeada como "0,025" para avaliarmos se nosso valor-*p* é tal que temos $p < 0{,}05$. Em outras palavras, necessitaríamos de uma estatística-*t* com o valor de 1,697 para grau de liberdade igual a 30 (1,694 para grau de liberdade igual a 32) para alcançar esse nível de significância em um teste de hipótese unicaudal. Para um teste bicaudal de hipótese, necessitamos de uma estatística-*t* de 2,047 para grau de liberdade igual a 30 (e 2,042 para grau de liberdade igual a 32).

Retornando ao nosso teste de hipótese sobre o intercepto e o valor igual a 50, se mudarmos de um teste bicaudal para um teste unicaudal, temos:

$$H_0 : \alpha \leq 50,$$
$$H_1 : \alpha > 50,$$

e o valor da nossa estatística-*t* ainda é:

$$t_{32} = \frac{\breve{\alpha} - \alpha^*}{\text{se}(\breve{\alpha})} = \frac{51{,}50816 - 50}{0{,}8569026} = 1{,}76.$$

Mas, para grau de liberdade igual 32, esse teste de hipótese unicaudal leva a um valor-*p* de 0,04. Em outras palavras, este é um caso em que a formulação do nosso teste como um teste unicaudal faz uma diferença bastante grande, especialmente porque muitos estudiosos consideram 0,05 como o padrão para a significância estatística.

[13] Escolhemos 0 quando a hipótese nula é $H_0 : \beta \leq 0$, porque esse é o valor crítico para a hipótese nula. Sob essa hipótese nula, zero é o limiar, e qualquer evidência de que β é igual a um valor menor ou igual a zero é favorável à hipótese nula.

Podemos observar a partir desses exemplos e da tabela-*t* que, quando temos uma hipótese direcional, podemos rejeitar mais facilmente a hipótese nula. Uma das peculiaridades da pesquisa em ciência política é que, mesmo quando as hipóteses são direcionais, pesquisadores frequentemente reportam os resultados para o teste bicaudal. Discutiremos a questão de como apresentar nossos resultados do modelo de regressão em detalhes no capítulo 12.

8.5 PRESSUPOSTOS, MAIS PRESSUPOSTOS E OS REQUISITOS MATEMÁTICOS MÍNIMOS

Se pressupostos fossem água, você precisaria de um guarda-chuva agora. Sempre que você estima um modelo de regressão, está implicitamente fazendo um amplo conjunto de pressupostos sobre o modelo populacional não observado. Nesta seção, dividimos estes pressupostos em pressupostos sobre o componente estocástico do modelo populacional e sobre a especificação do modelo. Adicionalmente, apresentamos alguns requisitos matemáticos mínimos que devem ser observados antes que se possa estimar um modelo de regressão. Na última seção, listamos esses pressupostos e requisitos e discutimos brevemente como eles se aplicam ao nosso exemplo de modelo bivariado para o impacto do crescimento econômico sobre o voto do partido do incumbente.

8.5.1 PRESSUPOSTOS SOBRE O COMPONENTE ESTOCÁSTICO DO MODELO POPULACIONAL

Os pressupostos mais importantes sobre o componente estocástico do modelo populacional u_i são sobre sua distribuição. Eles podem ser sumarizados por:

$$u_i \sim N(0, \sigma^2),$$

que significa que assumimos[14] que u_i é distribuído normalmente $(\sim N)$ com a média igual a zero e a variância igual a σ^2. Essa sintética afirmação matemática contém três dos cinco pressupostos que fazemos sobre o componente estocástico do modelo populacional ao estimar um modelo de regressão. Passaremos por cada um deles.

u_i é normalmente distribuído

O pressuposto de que u_i é normalmente distribuído permite que utilizemos a tabela-*t* para fazer inferências probabilísticas sobre o modelo de regressão da população

[14] Estritamente falando, não precisamos adotar todos estes pressupostos para estimar os parâmetros de um modelo de MQO. Mas precisamos adotá-los para interpretar da maneira padrão os resultados de um modelo de MQO.

a partir do nosso modelo de regressão da amostra. A principal justificativa para esse pressuposto é o teorema do limite central que discutimos no capítulo 6.

$E(u_i) = 0$: sem viés

O pressuposto de que u_i tem média ou valor esperado igual a zero é também conhecido como pressuposto do viés zero. Considere o que aconteceria se $E(u_i) \neq 0$. Em outras palavras, nesse caso *esperaríamos* que nosso modelo de regressão não fosse preciso. Quando isso ocorre, essencialmente estamos ignorando alguns dos *insights* teóricos sobre as causas subjacentes de *Y*. Lembre-se de que o termo estocástico é supostamente randômico. Se $E(u_i) \neq 0$, então deve existir algum componente não randômico nesse termo. É importante notar que não esperamos que todos os nossos valores u_i sejam iguais a zero, porque sabemos que alguns terão valores acima e abaixo da reta de regressão. Porém, esse pressuposto significa que nosso melhor palpite ou valor esperado para cada valor individual u_i é zero.

Se pensarmos sobre o exemplo utilizado neste capítulo, este pressuposto significa que não há razão para esperarmos que qualquer um dos valores preditos para o percentual de voto do incumbente esteja subestimado ou sobrestimado em nosso modelo. Se, por outro lado, tivermos alguma expectativa nesse sentido, não poderíamos assumir tal pressuposto. Digamos, por exemplo, que esperássemos que, durante períodos de guerra, o partido do incumbente se saia melhor do que esperaríamos que se saísse caso apenas a economia fosse considerada. Sob essa circunstância não poderíamos assumir esse pressuposto. A solução para esse problema seria incluir outra variável independente em nosso modelo para medir se o país estava em guerra no ano da eleição. Uma vez que tenhamos controlado todas as fontes potenciais de viés, podemos nos sentir confortáveis para fazer esse pressuposto. A inclusão de variáveis independentes adicionais é o assunto principal do capítulo 9.

u_i tem variância igual a σ^2: homocedasticidade

O pressuposto de que u_i tem variância igual σ^2 parece bastante simples. Mas, como essa noção de variância não contém o subscrito *i*, significa que assumimos que a variância para cada caso em uma população subjacente é a mesma. A palavra para descrever essa situação é "homocedasticidade", que significa "a variância do erro é constante". Se esse pressuposto não foi assumido, temos uma situação em que a variância de u_i é σ^2, situação conhecida como "heterocedasticidade", que significa "variância do erro não é constante". Quando temos heterocedasticidade, nosso modelo de regressão ajusta alguns casos dos casos da população melhor do que outros. Essa pode ser uma causa potencial de problemas quando estamos estimando intervalos de confiança e testando hipóteses.

Em nosso exemplo, esse pressuposto seria violado se, por alguma razão, algumas eleições fossem mais difíceis do que outras de serem preditas pelo modelo. Nesse caso, nosso modelo seria heterocedástico. Isso poderia acontecer, por exemplo, se eleições que ocorreram após

debates políticos serem transmitidos pela televisão fossem mais difíceis de serem preditas pelo nosso modelo utilizando apenas a variável independente de desempenho econômico. Sob essas circunstâncias, o pressuposto da homocedasticidade não seria razoável.

Sem autocorrelação

Também assumimos que não existe autocorrelação. A autocorrelação ocorre quando o termo estocástico para dois ou mais casos estão relacionados sistematicamente uns com os outros. Isso claramente viola o cerne da ideia de que esses termos são estocásticos ou randômicos. Formalmente, expressamos esse pressuposto por

$$\text{cov}_{u_i, u_j} = 0 \forall i \neq j;$$

isso significa que a covariância entre os termos de erro da população u_i e u_j é igual a zero para qualquer i não igual a j (para quaisquer dois casos únicos).

A forma mais comum de autocorrelação ocorre em modelos com dados de séries temporais. Como discutimos no capítulo 4, dados de séries temporais envolvem a mensuração das variáveis relevantes ao longo do tempo para uma única unidade espacial. No exemplo que estamos utilizando, empregamos medidas para crescimento econômico e para o percentual de votos recebido pelo partido do presidente em exercício mensuradas de quatro em quatro anos para os EUA. Se, por alguma razão, os termos de erro para pares de eleições adjacentes fossem correlacionados, teríamos autocorrelação.

Os valores de *X* são mensurados sem erro

A princípio, o pressuposto de que os valores de *X* foram mensurados sem erro pode parecer fora de lugar em uma lista de pressupostos sobre o componente estocástico da população. Mas esse pressuposto é assumido para simplificar as inferências que fazemos sobre o nosso modelo de regressão da população a partir do nosso modelo de regressão da amostra. Assumindo que *X* é mensurado sem erro, aceitamos que qualquer variabilidade da nossa reta de regressão é devida ao componente estocástico u_i, e não a modelos de mensuração em *X*. Colocando de outro modo, se *X* também tem um componente estocástico, necessitamos modelar *X* antes de modelar *Y*, e isto complicaria substancialmente o processo de estimação de *Y*.

Estaremos provavelmente bastante desconfortáveis com esse pressuposto em praticamente qualquer modelo de regressão que estimamos com dados do mundo real. No exemplo que estamos utilizando, estamos assumindo que temos as medidas corretas para a mudança percentual no PIB real *per capita* de 1876 a 2008. Se refletirmos um pouco mais sobre essa medida, podemos pensar sobre todos os tipos de potenciais erros na mensuração. E quanto às atividades econômicas ilegais que são de difícil mensuração para o governo? Como a medida é *per capita*, quão confiante estamos de que o denominador desse cálculo, a população, é mensurada corretamente?

Apesar dos problemas óbvios com esse pressuposto, o assumimos todas as vezes que estimamos um modelo de MQO. A menos que utilizemos técnicas estatísticas consideravelmente mais complicadas, esse é um pressuposto com o qual temos que conviver e manter em nossa mente quando avaliamos a confiança geral que temos no que nosso modelo está nos dizendo.

Lembre-se do capítulo 5, em que, durante a discussão sobre a mensuração de nossos conceitos de interesse, argumentamos que a mensuração é importante porque se mensurarmos de maneira inadequada podemos fazer inferências causais incorretas sobre o mundo real. Esse pressuposto deve tornar as importantes lições deste capítulo claras.

8.5.2 PRESSUPOSTOS SOBRE AS ESPECIFICAÇÕES DO NOSSO MODELO

Os pressupostos sobre as especificações do nosso modelo podem ser sumarizados em um único pressuposto: que o modelo possui *a* especificação correta. Dividimos esse pressuposto em dois para lançar luz sobre um leque de modos pelos quais esse pressuposto pode ser violado.

Nenhuma variável causal foi deixada de fora; nenhuma variável não causal foi incluída

Este pressuposto significa que, ao especificarmos nosso modelo de regressão bivariado da relação entre X e Y, não pode haver nenhuma outra variável Z que também causa Y[15]. Também significa que X deve causar Y. Em outras palavras, este é apenas outro modo de dizer que o modelo de regressão da amostra que especificamos é o verdadeiro modelo de regressão para a população subjacente.

À medida que fomos utilizando o exemplo deste capítulo, já começamos a sugerir variáveis independentes adicionais que teorizamos ser causalmente relacionadas a nossa variável dependente. Para aceitar este pressuposto, precisamos incluir todas essas variáveis em nosso modelo. A adição de outras variáveis independentes em nosso modelo é o assunto do capítulo 9.

Linearidade dos parâmetros

O pressuposto da linearidade dos parâmetros é uma maneira sofisticada de dizer que nosso parâmetro β da população para a relação entre X e Y não varia. Em outras palavras, a relação entre X e Y é a mesma para todos os valores de X.

---

[15] Uma exceção a isso é o caso especial em que existe uma variável Z que é causalmente relacionada a Y, mas não correlacionada com X e u_i. Nesse caso, ainda seriamos capazes de produzir uma estimação razoável da relação entre X e Y apesar de deixar Z fora do modelo. Discutimos mais este caso no capítulo 9.

No contexto do nosso exemplo atual, isso significa que estamos assumindo que o impacto do aumento de uma unidade na mudança real do PIB *per capita* é sempre a mesmo. Então, quando nos movemos do valor de –10 para –9, temos o mesmo efeito que ao nos mover do valor 1 para 2. No capítulo 10, discutimos algumas técnicas para relaxar esse pressuposto.

8.5.3 REQUISITOS MATEMÁTICOS MÍNIMOS

Para uma regressão bivariada, temos dois requisitos mínimos que devem ser satisfeitos pelos nossos dados amostrais antes de podermos estimar nossos parâmetros. Adicionaremos outros requisitos quando expandirmos o modelo bivariado para o modelo de regressão multivariado.

X deve variar

Pense sobre como seria o gráfico de dispersão dos dados da nossa amostra se X não variasse. Basicamente, teríamos uma pilha de valores de Y no mesmo ponto do eixo-x. A únicà linha razoável que poderia ser traçada através desse conjunto de pontos seria uma linha reta paralela ao eixo-y. Lembre-se que nosso objetivo é explicar nossa variável dependente Y. Sob essas circunstâncias nós falharíamos de modo miserável, porque qualquer valor Y seria tão bom quanto qualquer outro, dado que X possui apenas um valor. Assim, precisamos de alguma variação em X para conseguir estimar um modelo de regressão de MQO.

$n > k$

Para estimar um modelo de regressão, o número de casos (n) deve exceder o número de parâmetros a ser estimados (k). Assim, quando estimamos um modelo de regressão bivariado com dois parâmetros, devemos ter *no mínimo* três casos.

8.5.4 COMO PODEMOS SATISFAZER TODOS ESSES PRESSUPOSTOS?

Os requisitos matemáticos para estimar um modelo de regressão não são tão severos, mas uma questão sensível a se fazer neste ponto é: "Como podemos razoavelmente satisfazer todos os pressupostos listados todas as vezes que estamos rodando um modelo de regressão?". Para responder a essa pergunta, nos referimos à discussão feita no capítulo 1 sobre a analogia entre mapas e modelos. *Sabemos* que todos os nossos pressupostos podem não ser passíveis de serem satisfeitos. Também sabemos que estamos tentando simplificar a realidade. O único modo pelo qual podemos fazer isso é por meio de um conjunto grande de pressupostos não realistas sobre o mundo. É crucial,

porém, nunca perdermos de vista o fato de que estamos aceitando esses pressupostos. No próximo capítulo relaxamos um dos pressupostos mais irrealistas que fazemos em um modelo de regressão bivariado para controlarmos por uma segunda variável, Z.

CONCEITOS INTRODUZIDOS NESTE CAPÍTULO

- Estatística R^2 (R-quadrado) – uma medida de qualidade de ajuste do modelo que varia entre 0 e 1 e representa a proporção da variável dependente que é explicada pelo modelo.
- Estocástico – aleatório.
- Hipótese alternativa – uma expectativa baseada na teoria oposta à hipótese nula.
- Hipótese direcional – uma hipótese alternativa em que esperamos que a relação seja positiva ou negativa.
- Mínimos quadrados ordinários – também conhecidos como "MQO", são o método mais popular para estimar um modelo de regressão com dados amostrais.
- Modelo de regressão amostral – uma estimativa baseada na amostra do modelo de regressão para a população.
- Modelo de regressão populacional – uma formulação teórica da relação linear proposta entre, ao menos, uma variável independente e a variável dependente.
- Modelo estatístico – uma representação numérica da relação entre, ao menos, uma variável independente e a variável dependente.
- Parâmetro – um sinônimo para "fronteira" com uma conotação mais matemática. No contexto da estatística, o valor de uma característica da população desconhecida.
- Parâmetro estimado – cálculo para uma característica da população feito a partir de uma amostra.
- Razão-t – a razão entre um parâmetro estimado e seu erro-padrão estimado.
- Resíduo – o mesmo que o termo de erro da população.
- Raiz do erro quadrático médio *(Root mean-squared error)* – cálculo da qualidade do ajuste do modelo feito a partir da raiz quadrada da soma do quadrado de cada um dos valores de termo de erro do modelo da amostra divididos pelo número de casos. Também conhecido como "erro-padrão do modelo".
- Termo de erro da amostra – em um modelo de regressão para uma amostra, uma estimação do resíduo baseada na amostra.
- Termo de erro da população – em um modelo de regressão para a população, a diferença entre o valor predito pelo modelo para a variável dependente e o verdadeiro valor da variável dependente.

EXERCÍCIOS

1. Desenhe os eixos X e Y no meio de um quadrado com dimensões de 10 cm × 10 cm. O ponto de intersecção entre as linhas X e Y é conhecido por "origem" e é definido como o ponto em que X e Y são iguais a zero. Desenhe cada uma das retas entre os valores −5 e 5 de X e escreva as equações de regressão correspondentes:

   a) Intercepto-y = 2, inclinação = 2

   b) Intercepto-y = −2, inclinação = 2

   c) Intercepto-y = 0, inclinação = 1

   d) Intercepto-y = 2, inclinação = −2

2. Resolva cada uma das seguintes expressões matemáticas de modo que o resultado seja um componente do modelo de regressão bivariado para a amostra:

   a) $\breve{\alpha} + \breve{\beta} X_i + \breve{u}_i$

   b) $Y_i - E(Y \mid X_i)$

   c) $\breve{\beta} X_i + \breve{u}_i - Y_i$

3. Utilizando o banco de dados "state_data.dta", estimamos um modelo de regressão bivariado utilizando os dados de renda *per capita* ("pcinc" no nosso banco de dados) para cada um dos estados americanos e do distrito de Columbia como variável dependente e o percentual de residentes no estado com nível superior completo ("pctba" no nosso banco de dados) como variável independente. A equação estimada foi:

$$pcinc_i = 11519{,}78 + 1028{,}96 pctba_i$$

   Interprete os parâmetros estimados para o efeito do nível de educação no estado no valor da renda média.

4. No banco de dados descrito no exercício 3, o valor de pctba para Illinois é igual a 29,9. Qual é o valor predito pelo modelo para a renda *per capita* em Illinois?

5. O erro-padrão estimado para o parâmetro de inclinação do modelo descrito no exercício 3 foi igual a 95,7. Construa um intervalo de confiança de 95% para esse parâmetro estimado. Mostre todos os seus cálculos. O que isso nos diz sobre a relação estimada?

6. Teste a hipótese de que o parâmetro para pctba não é igual a zero. Mostre todos os cálculos. O que isso nos diz sobre a relação estimada?

7. Teste a hipótese de que o parâmetro para pctba é maior do que zero. Mostre todos os cálculos. O que isso nos diz sobre a relação estimada?

8. A estatística R^2 para o modelo descrito no exercício 3 é igual a 0,70 e o *root MSE* = 3773,8. O que esses números nos dizem sobre o nosso modelo?

9. Estime e interprete os resultados para um modelo de regressão bivariado diferente do modelo utilizado no exercício 3 utilizando o banco de dados "state_data.dta".

10. Pense em cada um dos pressupostos que você assumiu quando estava respondendo ao exercício 9. Qual deles você se sente mais confortável em fazer e qual você se sente menos confortável? Explique suas respostas.

11. No exercício 10 do capítulo 7, você calculou um coeficiente de correlação para a relação entre duas variáveis contínuas. Agora, estime um modelo de regressão bivariado utilizando as mesmas duas variáveis. Produza uma tabela de resultados e escreva sobre o que essa tabela diz sobre a política no Reino Unido em 2005.