Ocasionalmente, queremos prever y quando $\log(y)$ é usado como a variável dependente em um modelo de regressão. A Seção 6.4 explica esse método simples. Finalmente, algumas vezes estamos interessados em conhecer o sinal e a magnitude dos resíduos de observações específicas. A análise de resíduos pode ser usada para determinarmos se elementos específicos da amostra possuem valores previstos que estejam bem acima, ou bem abaixo, dos verdadeiros resultados.

PROBLEMAS

6.1 No Exemplo 4.2, no qual a percentagem de alunos aprovados em um exame de matemática do 10º ano (*mate10*) é a variável dependente, faz sentido incluir *cien11* — a percentagem de alunos do 11º ano aprovados em um exame de ciências — como uma variável explicativa adicional?

6.2 Se iniciarmos com a (6.38) sob as hipóteses MLC, admitirmos n grande e ignorarmos o erro de estimação na $\hat{\beta}_j$, um intervalo de predição de 95% da y^0 será $[\exp(-1{,}96\hat{\sigma})\exp(\widehat{\log y^0}), \exp(1{,}96\hat{\sigma})\exp(\widehat{\log y^0})]$. O ponto de predição de y^0 é $\hat{y}^0 = \exp(\hat{\sigma}^2/2)\exp(\widehat{\log y^0})$.

(i) Para quais valores de $\hat{\sigma}$ o ponto de predição ficará no intervalo de predição de 95%?

Esta condição parece provável de se sustentar na maioria das aplicações?

(ii) Confirme se a condição da parte (i) é cumprida no exemplo dos salários dos CEOs.

6.3 O seguinte modelo permite que o retorno da educação dependa da educação total dos pais, chamada *edupais*:

$$\log(salário) = \beta_0 + \beta_1 educ + \beta_2 educ \cdot edupais + \beta_3 exper + \beta_4 perm + u.$$

(i) Mostre que, em forma decimal, o retorno de mais um ano de educação nesse modelo é

$$\Delta\log(salário)/\Delta educ = \beta_1 + \beta_2 edupais.$$

Que sinal você espera para β_2? Por quê?

(ii) Utilizando os dados contidos no arquivo WAGE2.RAW, a equação estimada é

$$\widehat{\log(salário)} = 5{,}65 + 0{,}047\ educ + 0{,}00078\ educ \cdot edupais + 0{,}019\ exper + 0{,}010\ perm$$
$$(0{,}13)\ (0{,}010)\ (0{,}00021)\ (0{,}004)\ (0{,}003)$$
$$n = 722,\ R^2 = 0{,}169.$$

(Somente 722 observações contêm todas as informações sobre a educação dos pais.) Interprete o coeficiente do termo de interação. Pode ser interessante escolher dois valores específicos para *edupais* — por exemplo, *edupais* = 32 se ambos tiverem educação superior, ou *edupais* = 24 se ambos tiverem educação de nível médio — e comparar o retorno estimado de *educ*.

(iii) Quando *edupais* é adicionada como uma variável separada na equação, obtemos:

$$\widehat{\log(salário)} = 4{,}94 + 0{,}097\ educ + 0{,}033\ edupais + 0{,}0016\ educ \cdot edupais + 0{,}020\ exper + 0{,}010\ perm$$
$$(0{,}38)\ (0{,}027)\ (0{,}017)\ (0{,}0012)\ (0{,}004)\ (0{,}003)$$
$$n = 722,\ R^2 = 0{,}174.$$

O retorno da educação agora depende positivamente da educação dos pais? Teste a hipótese nula de que o retorno da educação não depende da educação dos pais.

6.4 Suponha que queiramos estimar os efeitos do consumo de bebida alcoólica (*álcool*) na nota média de graduação (*GPAgrad*). Além de coletarmos informações sobre as médias das notas de graduação e do uso de bebida alcoólica, também obtenhamos dados sobre a frequência (digamos, porcentagem de aulas frequentadas, chamada *freq*). Uma pontuação de teste padronizado (digamos, *TAA*) e nota média do ensino médio (*GPAem*) também estão disponíveis.

(i) Devemos incluir *freq* juntamente com *álcool* como variáveis explicativas num modelo de regressão múltipla? (Pense em como você interpretaria $\beta_{álcool}$).

(ii) As variáveis *GPAem* e *TAA* devem ser incluídas como variáveis explicativas? Explique.

6.5 Utilizando os dados contidos no arquivo RDCHEM.RAW, a seguinte equação foi obtida por MQO:

$$\widehat{pdintens} = 2{,}613 + 0{,}00030\ vendas - 0{,}0000000070\ vendas^2$$
$$(0{,}429)\ (0{,}00014)\ (0{,}0000000037)$$
$$n = 32,\ R^2 = 0{,}1484.$$

(i) Em que ponto o efeito marginal de *vendas* sobre *pdintens* se torna negativo?

(ii) Você manteria o termo quadrático no modelo? Explique.

(iii) Defina *vendasbil* como vendas expressas em bilhões de dólares: *vendasbil* = *vendas*/1.000. Reescreva a equação com *vendasbil* e *vendasbil*^{2} como as variáveis independentes. Certifique-se de descrever os erros-padrão e o R-quadrado. [*Sugestão*: Observe que *vendasbil*^{2} = *vendas*^{2}/(1.000)^2.]

(iv) Com o propósito de descrever os resultados, qual equação você prefere?

6.6 As três seguintes equações foram estimadas utilizando-se as 1.534 observações contidas no arquivo 401K.RAW.

$$\widehat{taxap} = 80{,}29 + 5{,}44\ taxcomp + 0{,}269\ idade - 0{,}00013\ totemp$$
$$(0{,}78)\ (0{,}52)\ (0{,}045)\ (0{,}00004)$$
$$R^2 = 0{,}100,\ \bar{R}^2 = 0{,}098.$$

$$\widehat{taxap} = 97{,}32 + 5{,}02\ taxcomp + 0{,}314\ idade - 2{,}66\log(totemp)$$
$$(1{,}95)\ (0{,}51)\ (0{,}044)\ (0{,}28)$$
$$R^2 = 0{,}144,\ \bar{R}^2 = 0{,}142.$$

$$\widehat{taxap} = 80{,}62 + 5{,}34\ taxcomp + 0{,}290\ idade - 0{,}00043\ totemp$$
$$(0{,}78)\ (0{,}52)\qquad (0{,}045)\qquad (0{,}00009)$$
$$+ 0{,}0000000039\ totemp^2$$
$$(0{,}0000000010)$$
$$R^2 = 0{,}108,\ \bar{R}^2 = 0{,}106.$$

Qual desses três modelos você prefere? Por quê?

6.7 Sejam $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ as estimativas MQO da regressão de y_i sobre $x_{i1}, \ldots, x_{ik}$, $i = 1, 2, \ldots, n$. Para constantes diferentes de zero $c_1, \ldots, c_k$, argumente que o intercepto e as inclinações MQO da regressão de $c_0 y_i$ sobre $c_1 x_{i1}, \ldots, c_k x_{ik}$, $i = 1, 2, \ldots, n$ são dados por $\tilde{\beta}_0 = c_0\hat{\beta}_0, \tilde{\beta}_1 = (c_0/c_1)\hat{\beta}_1, \ldots, \tilde{\beta}_k = (c_0/c_k)\hat{\beta}_k$. (*Sugestão*: Use o fato de que $\hat{\beta}_j$ soluciona as condições de primeira ordem em (3.13), e que $\tilde{\beta}_j$ deve solucionar as condições de primeira ordem envolvendo as variáveis dependente e independentes redimensionadas.)

6.8 Quando $taxafreq^2$ e $ACT \cdot taxafreq$ são adicionadas à equação (6.19), o R-quadrado passa a ser 0,232. Esses termos adicionais são conjuntamente significantes no nível de 10%? Você os incluiria no modelo?

6.9 A seguinte equação foi estimada utilizando os dados contidos no arquivo CEOSAL1.RAW:

$$\widehat{\log(salário)} = 4{,}322 + 0{,}276\log(vendas) + 0{,}0215\ roe - 0{,}00008\ roe^2$$
$$(0{,}324)\ (0{,}33)\qquad (0{,}129)\qquad (0{,}00026)$$
$$n = 209,\ R^2 = 0{,}282.$$

Esta equação permite que *roe* tenha um efeito decrescente sobre log(*salário*). Essa generalidade é necessária? Justifique.

APÊNDICE 6A

6A. Uma Breve Introdução à Reamostragem

Em muitos casos em que fórmulas de erros-padrão são difíceis de serem obtidas matematicamente, ou em que se acha que eles não são uma aproximação muito boa da verdadeira variação amostral de um estimador, podemos nos valer de um **método de reamostragem**. A ideia geral é tratar os dados observados como uma população da qual podemos extrair amostras. O método de reamostragem mais comum é o de **reamostragem**. (Existem, na verdade, várias versões da reamostragem, mas a mais geral, e aplicada com maior rapidez é chamada *reamostragem não paramétrica*, e é esta que descrevemos aqui.)

Suponha que temos uma estimativa, $\hat{\theta}$, de um parâmetro populacional, θ. Obtivemos esta estimativa, que pode ter sido uma função das estimativas MQO (ou estimativas que trataremos em capítulos posteriores), de uma amostra aleatória de tamanho n. Gostaríamos de obter um erro-padrão de $\hat{\theta}$ que possa ser usado na construção de estatística t ou intervalos de confiança. Notadamente, podemos

obter um erro-padrão válido calculando a estimativa de diferentes amostras aleatórias extraídas dos dados originais.

A implementação é fácil. Se listarmos nossas observações de 1 a n extrairemos n números aleatoriamente, com substituição, desta lista. Isto produzirá um novo conjunto de dados (de tamanho n) que consistirá dos dados originais, mas com muitas observações aparecendo várias vezes (exceto no caso bastante raro em que reamostramos os dados originais). Cada vez que amostramos aleatoriamente dos dados originais, podemos estimar θ usando o mesmo procedimento que usamos nos dados originais. Que $\hat{\theta}^{(b)}$ denote a estimativa da reamostra b. Agora, se repetirmos a reamostragem e a estimação m vezes, teremos m novas estimativas, $\{\hat{\theta}^{(b)}: b = 1, 2, \ldots, m\}$. O **erro-padrão de reamostragem** de $\hat{\theta}$ é simplesmente o desvio-padrão amostral da $\hat{\theta}^{(b)}$, ou seja,

$$\text{bse}(\hat{\theta}) = \left[(m-1)^{-1}\sum_{b=1}^{m}(\hat{\theta}^{(b)} - \bar{\hat{\theta}})^2\right]^{1/2}. \qquad (6.48)$$

em que $\bar{\hat{\theta}}$ é a média das reamostragens.

Se a obtenção de uma estimativa de θ numa amostra de tamanho n exige pouco tempo computacional, como no caso dos MQO e todos os outros estimadores que encontramos neste livro, podemos nos permitir preferir que m — o número de replicações de reamostragens — seja grande. Um valor típico é $m = 1.000$, mas mesmo $m = 500$ ou um valor menor pode produzir um erro-padrão confiável. Observe que o tamanho de m — o número de vezes que reamostramos os dados originais — não tem nada a ver com o tamanho da amostra, n, (Para certos problemas de estimação além do escopo deste livro, um n grande pode forçar que se faça um número menor de replicações de reamostragem). Muitos pacotes estatísticos e econométricos possuem comandos de reamostragem integrados, e isso torna simples o cálculo dos erros-padrão de reamostragem, especialmente quando comparado com o trabalho frequentemente requerido para se obter uma fórmula analítica de um erro-padrão assimptótico.

Pode-se de fato fazer melhor na maioria dos casos usando-se o exemplo de reamostragem para calcular-se p-valores de estatísticas t (e estatísticas F), ou para se obter intervalos de confiança, em vez de obter-se um erro-padrão de reamostragem para ser usado na construção de estatísticas t ou intervalos de confiança. Veja Horowitz (2001) para um tratamento abrangente.