Capítulo 16

Regressão Linear Simples

16.1 Introdução

No Capítulo 8 introduzimos o conceito de regressão para duas v.a. quantitativas, X e Y. Vimos que a esperança condicional de Y, dado que $X = x$, por exemplo, denotada por $E(Y|x)$, é uma função de x, ou seja,

$$E(Y|x) = \mu(x). \tag{16.1}$$

Em (8.27) definimos precisamente essa função. Uma definição similar vale para $E(X|y)$, que será uma função de y. Estamos considerando aqui o caso em que X e Y são definidas sobre uma mesma população P. Por exemplo, X pode ser a idade e Y o tempo de reação ao estímulo, no Exemplo 15.1. Nesse exemplo, a análise sugeriu a existência de uma relação mais forte entre as duas variáveis, e a modelamos por

$$y_{ij} = \mu_i + e_{ij},\ i = 1, ..., 5, \quad j = 1, ..., 4, \tag{16.2}$$

onde μ_i é a média do grupo de idade i. Podemos pensar que o fator idade determina cinco subpopulações (ou estratos) em P e de lá escolhemos cinco amostras aleatórias de tamanhos $n_i = 4$, $i = 1, ..., 5$.

Em (16.1), $\mu(x)$ pode ser qualquer função de x; veja o Exemplo 8.21. Um caso simples de interesse é aquele em que X e Y têm distribuição conjunta normal bidimensional. Nesse caso, $\mu(x)$ e $\mu(y)$ são, de fato, funções lineares. Veja a seção 8.8.

Continuando com o Exemplo 15.1, tanto X (idade) como Y (tempo de resposta ao estímulo) são v.a. contínuas, e podemos pensar em introduzir um modelo alternativo para y_{ij}, dada a relação entre X e Y. Observando as médias de Y, segundo os grupos de idades, ou seja, $E(Y|x)$, percebemos que estas aumentam conforme as pessoas envelhecem. A Figura 16.1 mostra os dados observados, onde notamos uma tendência crescente, bem como os valores repetidos de Y para cada nível de idade x.

Um modelo razoável para $E(Y|x)$ pode ser

$$E(Y|x) = \mu(x) = \alpha + \beta x, \tag{16.3}$$

ou seja, o tempo médio de reação é uma função linear da idade.

16.4 Propriedades dos Estimadores

Iremos agora estudar as propriedades amostrais dos estimadores $\hat{\alpha}$ e $\hat{\beta}$, e para isso é conveniente voltar ao modelo e às suposições adotadas para a variável aleatória Y sob investigação. Lembremos que a variável X é suposta controlada, fixa, e para cada valor x de X teremos associada uma distribuição de probabilidades para Y, como ilustra a Figura 16.5 (a), onde supomos que a dispersão é a mesma para cada nível da variável X. A Figura 16.5 (b) ilustra o caso que será considerado aqui, em que estas distribuições condicionais são normais, com a mesma variância. Note que $E(Y|x)$ é linear, como estamos considerando neste capítulo.

Formalmente, o modelo

$$Y_i = E(Y|x_i) + e_i = \alpha + \beta x_i + e_i, \quad i = 1, ..., n$$

deve satisfazer as seguintes suposições:

(i) Para cada valor de x_i, o erro e_i tem média zero e variância constante σ_e^2;

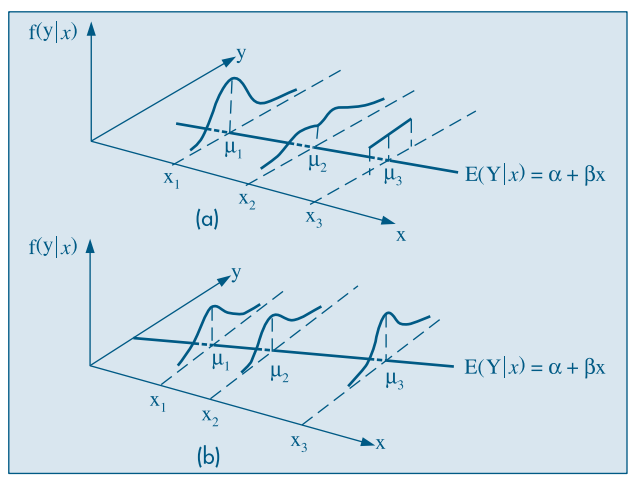
(ii) Se $i \neq j$, $\text{Cov}(e_i, e_j) = 0$, isto é, para duas observações distintas, os erros são não-correlacionados.

Segue-se que

$$E(Y_i|x_i) = \alpha + \beta x_i \quad \text{e} \quad \text{Var}(Y_i|x_i) = \sigma_e^2,$$

e ainda que Y_i e Y_j são não-correlacionados, para $i \neq j$.

Figura 16.5: (a) médias alinhadas, distribuições com a mesma variância;
(b) médias alinhadas, distribuições normais com a mesma variância.



16.4.1 Média e Variância dos Estimadores

Nesta seção vamos obter a média e a variância dos estimadores $\hat{\alpha}$ e $\hat{\beta}$, dados em (16.14).

Proposição 16.1. Para o estimador $\hat{\beta}$ temos

$$E(\hat{\beta}) = \beta, \tag{16.36}$$

$$\text{Var}(\hat{\beta}) = \frac{\sigma_e^2}{\sum_{i=1}^n (x_i - \bar{x})^2}. \tag{16.37}$$

Prova. Inicialmente, vamos escrever β de um modo mais conveniente (veja o Problema 30):

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i - \bar{Y}\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n \frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} Y_i = \sum_{i=1}^n w_i Y_i,$$

onde estamos usando a notação Y (maiúscula) e x (minúscula) para diferenciar o fato de que a primeira está sendo considerada aleatória e a segunda, fixa; e

$$w_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \sum_{i=1}^n w_i = 0.$$

Observe que estamos usando o fato de $\sum_{i=1}^n (x_i - \bar{x}) = 0$ e que

$$\sum_{i=1}^n w_i x_i = \sum_{i=1}^n w_i x_i - \bar{x}\sum_{i=1}^n w_i = \sum_{i=1}^n w_i (x_i - \bar{x})$$

$$= \sum_{i=1}^n \frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} (x_i - \bar{x}) = 1.$$

Usando propriedades da esperança e variância de somas de v.a. (veja o Capítulo 8), podemos escrever

$$E(\hat{\beta}) = E(\sum_{i=1}^n w_i Y_i) = \sum_{i=1}^n w_i E(Y_i)$$

$$= \sum_{i=1}^n w_i(\alpha + \beta x_i) = \alpha\sum_{i=1}^n w_i + \beta\sum_{i=1}^n w_i x_i = \beta,$$

o que mostra que o estimador é não-viesado. Para a variância,

$$\text{Var}(\hat{\beta}) = \text{Var}(\sum_{i=1}^n w_i Y_i) = \sum_{i=1}^n w_i^2 \,\text{Var}(Y_i),$$

pois as observações são não-correlacionadas, e, portanto,

$$\text{Var}(\hat{\beta}) = \sum_{i=1}^n w_i^2 \sigma_e^2 = \sigma_e^2 \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)^2 = \sigma_e^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2},$$

e o resultado segue.

Proposição 16.2. Para o estimador $\hat{\alpha}$ temos:

$$E(\hat{\alpha}) = \alpha, \tag{16.38}$$

$$\text{Var}(\hat{\alpha}) = \sigma_e^2 \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{16.39}$$

Prova. Precisaremos dos seguintes resultados (Problema 33):

$$\text{Cov}(\bar{y}, \hat{\beta}) = 0, \tag{16.40}$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2. \tag{16.41}$$

Como

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(\alpha + \beta x_i + e_i)$$

$$= \alpha + \beta\bar{x} + \frac{1}{n}\sum_{i=1}^{n} e_i,$$

temos que

$$E(\bar{y}) = \alpha + \beta\bar{x} + \frac{1}{n}\sum_{i=1}^{n} E(e_i) = \alpha + \beta\bar{x},$$

dado que x é supostamente fixa e não uma v.a. Também,

$$\text{Var}(\bar{y}) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(e_i) = \frac{\sigma_e^2}{n}.$$

Temos, então, que

$$E(\hat{\alpha}) = E(\bar{y} - \hat{\beta}\,\bar{x}) = \alpha + \beta\bar{x} - \beta\bar{x} = \alpha,$$

e

$$\text{Var}(\hat{\alpha}) = \text{Var}(\bar{y} - \hat{\beta}\,\bar{x}) = \text{Var}(\bar{y}) + \text{Var}(\hat{\beta}\,\bar{x}) - 2\text{Cov}(\bar{y}, \hat{\beta}\,\bar{x})$$

$$= \text{Var}(\bar{y}) + \bar{x}^2\text{Var}(\hat{\beta}) - 2\bar{x}\text{Cov}(\bar{y}, \hat{\beta})$$

e usando os diversos resultados obtidos acima, obtemos (16.39).

16.4.2 Distribuições Amostrais dos Estimadores dos Parâmetros

Para completar o estudo das propriedades dos estimadores, vamos introduzir uma terceira suposição:

(iii) Os erros e_i são v.a. com distribuição normal, isto é,

$$e_i \sim N(0;\ \sigma_e^2), \tag{16.42}$$

o que implica

$$y_i \sim N(\alpha + \beta x_i;\ \sigma_e^2). \tag{16.43}$$

Como $\hat{\beta}$ e $\hat{\alpha}$ são combinações lineares de v.a. normais e independentes, temos o seguinte resultado:

Proposição 16.3. Os estimadores $\hat{\alpha}$ e $\hat{\beta}$ têm ambos distribuição normal, com médias e variâncias dadas pelas Proposições 16.1 e 16.2, isto é,

$$\hat{\alpha} \sim N\left(\alpha;\ \frac{\sigma_e^2 \sum x_i^2}{n\sum(x_i - \bar{x})^2}\right), \tag{16.44}$$

$$\hat{\beta} \sim N\left(\beta;\ \frac{\sigma_e^2}{\sum(x_i - \bar{x})^2}\right). \tag{16.45}$$

Os resultados acima permitem concluir que

$$\frac{\hat{\beta} - \beta}{\sigma_e}\ \sqrt{\sum(x_i - \bar{x})^2} \sim N(0,\ 1), \tag{16.46}$$

$$\frac{\hat{\alpha} - \alpha}{\sigma_e}\ \sqrt{\frac{n\sum(x_i - \bar{x})^2}{\sum x_i^2}} \sim N(0,\ 1). \tag{16.47}$$

16.4.3 Intervalos de Confiança para α e β

Substituindo σ_e por seu estimador S_e em (16.46) e (16.47), sabemos que as estatísticas resultantes terão distribuição t de Student, com $(n - 2)$ graus de liberdade, o que permitirá construir intervalos de confiança para os parâmetros.

Proposição 16.4. As estatísticas

$$t(\hat{\beta}) = \frac{\hat{\beta} - \beta}{S_e}\ \sqrt{\sum(x_i - \bar{x})^2} \tag{16.48}$$

e

$$t(\hat{\alpha}) = \frac{\hat{\alpha} - \alpha}{S_e}\ \sqrt{\frac{n\sum(x_i - \bar{x})^2}{\sum x_i^2}} \tag{16.49}$$

têm distribuição t de Student com $(n - 2)$ graus de liberdade.

Esse resultado, combinado com os procedimentos de construção de intervalos de confiança já estudados, nos leva aos seguintes intervalos para α e β, com γ denotando o coeficiente de confiança e $t_\gamma(n - 2)$ denotando o valor obtido da Tabela V, com $(n - 2)$ graus de liberdade:

$$\text{IC}(\alpha;\ \gamma) = \hat{\alpha} \pm t_\gamma(n - 2)S_e\sqrt{\frac{\sum x_i^2}{n\sum(x_i - \bar{x})^2}}, \tag{16.50}$$

$$\text{IC}(\beta;\ \gamma) = \hat{\beta} \pm t_\gamma(n - 2)S_e\sqrt{\frac{1}{\sum(x_i - \bar{x})^2}}. \tag{16.51}$$

Exemplo 16.4. Da tabela ANOVA do Exemplo 16.3 podemos retirar as informações necessárias para construir intervalos de confiança para α e β. Temos que $\sum x_i^2 = 19.000$, $\sum (x_i - \bar{x})^2 = 1.000$, e $\bar{x} = 30$.

Temos, também, $S_e^2 = 31{,}28$ e, portanto, $S_e = 5{,}59$. Se $\gamma = 0{,}95$, obtemos $t_{0{,}95}(18) = 2{,}101$. Os intervalos são dados por:

$$\text{IC}(\alpha;\ 0{,}95) = 80{,}50 \pm (2{,}101)(5{,}59)\sqrt{\frac{19.000}{(1.000)(20)}} = 80{,}50 \pm 11{,}45,$$

$$\text{IC}(\beta;\ 0{,}95) = 0{,}90 \pm (2{,}101)(5{,}59)\ \sqrt{1/1.000}$$
$$= 0{,}90 \pm 0{,}30.$$

Ou seja,

$$\text{IC}(\alpha;\ 0{,}95) = [69{,}05;\ 91{,}95],$$
$$\text{IC}[\beta;\ 0{,}95] = [0{,}60;\ 1{,}20].$$

Este último resultado é mais uma evidência de que $\beta \neq 0$, o que reforça conclusões anteriores.

Os intervalos de confiança (16.50) e (16.51) podem ser utilizados para testar hipóteses do tipo

$$H_0\colon \alpha = \alpha_0,$$
$$H_0\colon \beta = \beta_0.$$

Em particular, temos o resultado:

Proposição 16.5. A estatística para testar $H_0\colon \alpha = 0$ é

$$t(\hat{\alpha}) = \frac{\hat{\alpha}}{S_e}\sqrt{\frac{n\sum (x_i - \bar{x})^2}{\sum x_i^2}}, \tag{16.52}$$

e a estatística para testar $H_0\colon \beta = 0$ é

$$t(\hat{\beta}) = \frac{\hat{\beta}}{S_e}\sqrt{\sum (x_i - \bar{x})^2}, \tag{16.53}$$

cada uma tendo distribuição t de Student com $(n - 2)$ graus de liberdade.

Observe que

$$[t(\hat{\beta})]^2 = \frac{\hat{\beta}^2 \sum (x_i - \bar{x})^2}{S_e^2},$$

e usando o resultado (16.33) podemos escrever

$$[t(\hat{\beta})]^2 = \frac{\text{SQReg}}{S_e^2}, \tag{16.54}$$

que é a estatística F que aparece na tabela ANOVA. Assim, para testar a hipótese $H_0\colon \beta = 0$, pode-se usar a estatística (16.54), que segue uma distribuição $F(1,\ n - 2)$.

Exemplo 16.5. Para testar separadamente as hipóteses acima, os valores das estatísticas correspondentes serão:

$$t(\hat{\alpha}) = (80{,}5/5{,}59)\sqrt{\frac{(20)(1.000)}{19.000}} = 14{,}77,$$

$$t(\hat{\beta}) = (0{,}90/5{,}59)\sqrt{1.000} = 5{,}09,$$

os quais devem ser comparados com 2,101, que é o valor crítico de $t(18)$, no nível de significância 5%. Vemos que em ambos os casos rejeitamos as hipóteses de que os parâmetros sejam iguais a zero. Comparando o resultado de $t(\hat{\beta})$ com o valor F da tabela ANOVA, constatamos que $t^2(\hat{\beta}) = 25{,}90 = F$, de acordo com o apresentado acima. Algumas vezes, para indicar a significância das estatísticas, a reta ajustada é escrita do seguinte modo:

$$\begin{array}{rccc} \hat{y} = & 80{,}50 & + & 0{,}90x\,, \\ & (14{,}77) & & (5{,}09) \end{array}$$

onde entre parênteses aparece o valor de t, para indicar com que intensidade o parâmetro pode ser considerado distinto de zero.

16.4.4 Intervalo de Confiança para $\mu(z)$ e Intervalo de Predição

O modelo linear (16.6), estudado até agora, será utilizado freqüentemente para fazer previsões da variável resposta (y) para algum nível da variável de controle (x). Usando o enunciado do Exemplo 16.1, poderíamos estar interessados em saber qual o tempo de reação aos 28 anos. É importante estabelecer se queremos estimar o tempo médio para o grupo etário de 28 anos ou o tempo de reação provável para uma pessoa de 28 anos. Veremos que a estimação pontual é a mesma nos dois casos, porém os intervalos de "confiança" serão distintos. Para entender bem as diferenças sugerimos recordar as soluções aos exercícios 23, 24 e 25 do Capítulo 15.

Do modelo (16.3) e do exposto até agora, temos o seguinte resultado.

Proposição 16.6. A distribuição amostral do estimador (16.15) é dada por

$$\widehat{\mu(x_i)} = \hat{y}_i = \hat{\alpha} + \hat{\beta}x_i \sim N(\alpha + \beta x_i, \text{Var}(\hat{y}_i)) \tag{16.55}$$

onde

$$\text{Var}(\widehat{\mu(x_i)}) = \text{Var}(\hat{y}_i) = \sigma_e^2\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right] \tag{16.56}$$

Prova. Das proposições 16.1 e 16.2 vem:

$$E(\widehat{\mu(x_i)}) = E(\hat{\alpha}) + E(\hat{\beta})x_i = \alpha + \beta x_i = \mu(x_i)$$

o que demonstra a primeira parte da proposição. De (16.17) temos

$$\hat{y}_i = \bar{y} + \hat{\beta}(x_i - \bar{x}),$$

portanto

$$\text{Var}(\hat{y}_i) = \text{Var}(\bar{y}) + (x_i - \bar{x})^2 \text{Var}(\hat{\beta}) + 2(x_i - \bar{x}) \text{Cov}(\bar{y}, \hat{\beta}),$$

mas de (16.40), $\text{Cov}(\bar{y}, \hat{\beta}) = 0$, e de (16.37) vem

$$\text{Var}(\hat{y}_i) = \frac{\sigma_e^2}{n} + (x_i - \bar{x})^2 \frac{\sigma_e^2}{\sum(x_i - \bar{x})^2} = \sigma_e^2 \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right],$$

o que conclui a prova.

Com a proposição acima e substituindo σ_e^2 por seu estimador S_e^2 é fácil verificar que o Intervalo de Confiança para $\mu(x)$ será dado por:

$$\text{IC}(\mu(x);\ \gamma) = \hat{y}_i \pm t_\gamma(n-2)\ S_e \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \tag{16.57}$$

Vejamos agora como construir um intervalo de predição para uma futura observação. Imitando a proposta do Problema 15.24, uma futura observação para um dado nível x_f é dada por

$$Y_f(x) = \mu(x_f) + \varepsilon_f$$

e o estimador será

$$\hat{Y}_f = \hat{y}_f + \varepsilon_f = \hat{y}_f,$$

onde substituímos o valor desconhecido ε_f pelo seu valor esperado que é zero.

Da expressão anterior, calculamos:

$$\text{Var}(\hat{Y}_f) = \text{Var}(\hat{y}_f) + \text{Var}(\varepsilon_f) = \sigma_e^2 \left[\frac{1}{n} + \frac{(x_1 - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right] + \sigma_e^2,$$

ou seja,

$$\text{Var}(\hat{Y}_f) = \sigma_e^2 \left[1 + \frac{1}{n} + \frac{(x_1 - \bar{x})^2}{\sum(x_1 - \bar{x})^2} \right]. \tag{16.58}$$

Substituindo σ_e^2 pelo seu estimador S_e^2, teremos um estimador da variância, e analogamente o intervalo de predição abaixo:

$$\text{IP}(Y_f;\ \gamma) = \hat{y}_f \pm t_\gamma S_e \sqrt{1 + \frac{1}{n} + \frac{(x_f - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \tag{16.59}$$

Exemplo 16.6. Qual o tempo de reação aos 28 anos?

A estimativa pontual é dada por:

$$\hat{y}(28) = 80{,}5 + 0{,}9(28) = 105{,}7.$$

Considerando como resposta adequada o tempo de reação médio do grupo de 28 anos, podemos escrever o Intervalo de Confiança para a média, ou seja:

$$\text{IC}(\mu(28);\ 0{,}95) = 105{,}7 \pm (2{,}101)(5{,}59)\ \sqrt{\frac{1}{20} + \frac{(28-30)^2}{1000}} =$$

$$= 105{,}7 \pm 2{,}7 = \,]103{,}0;\ 108{,}4[\,.$$

Se quiséssemos saber dentro de que intervalo 95% das futuras observações iriam estar, construiríamos o Intervalo de Predição:

$$\text{IP}(Y_f;\ 0{,}95) = 105{,}7 \pm (2{,}101)(5{,}59)\ \sqrt{1 + \frac{1}{20} + \frac{(28-30)^2}{1000}} =$$

$$= 105{,}7 \pm 12{,}1 = \,]93{,}6;\ 117{,}8[.$$

Problemas

10. Usando a tabela ANOVA, construída no Problema 5:
    (a) Construa o IC(β; 95%).
    (b) Construa o IC(α; 90%).
    (c) Use a estatística F para testar a hipótese H_0: $\beta = 0$.
    (d) Construa o IC para a acuidade visual média do grupo etário de 28 anos.
    (e) E qual seria o Intervalo de Predição da acuidade visual das pessoas de 28 anos?

11. Com as informações do Exemplo 15.1, e a ANOVA construída no Problema 9, você diria que a acuidade visual ajuda a prever o tempo de reação dos indivíduos? Que estatística você usou para justificar seu argumento e por quê?

12. Investigando a relação entre a quantidade de fertilizante usado (x) e a produção de soja (y) numa estação experimental com 20 canteiros, obteve-se a equação de MQ:

$$\hat{y} = \underset{(3{,}22)}{15{,}00} + \underset{(1{,}65)}{2{,}83x}.$$

    Com esses resultados você diria que a quantidade de fertilizante influi na produção? Por quê?

16.5 Análise de Resíduos

Para verificar se um modelo é adequado, temos que investigar se as suposições feitas para o desenvolvimento do modelo estão satisfeitas. Para tanto, estudamos o comportamento do modelo usando o conjunto de dados observados, notadamente as discrepâncias entre os valores observados e os valores ajustados pelo modelo, ou seja, fazemos uma *análise dos resíduos*.

O i-ésimo resíduo é dado por

$$\hat{e}_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n. \tag{16.60}$$

Lembremos que já utilizamos estes resíduos para obter medidas da qualidade e dos estimadores dos parâmetros do modelo. Agora iremos estudar o comportamento individual e conjunto destes resíduos, comparando com as suposições feitas sobre os verdadeiros erros e_i. Existem várias técnicas formais para conduzir essa análise, mas aqui iremos ressaltar basicamente métodos gráficos. Para mais detalhes, ver Draper e Smith (1998).

Uma representação gráfica bastante útil é obtida plotando-se pares $(x_i, \hat{e}_i)$, $i = 1, ..., n$. Outras vezes, é de maior utilidade fazer a representação gráfica dos chamados resíduos padronizados,

$$\hat{z}_i = \frac{y_i - \hat{y}_i}{S_e} = \frac{\hat{e}_i}{S_e}, \tag{16.61}$$

plotando-se os pares $(x_i, \hat{z}_i)$. Observe que a forma dos dois gráficos será semelhante, havendo apenas uma mudança de escala das ordenadas nos dois casos. Por isso, iremos usar a primeira representação, indicando no gráfico a posição do valor S_e.

Outro resíduo usado é o chamado *resíduo estudentizado*, definido por

$$\hat{r}_i = \frac{\hat{e}_i}{S_e\sqrt{1 - v_{ii}}}, \tag{16.62}$$

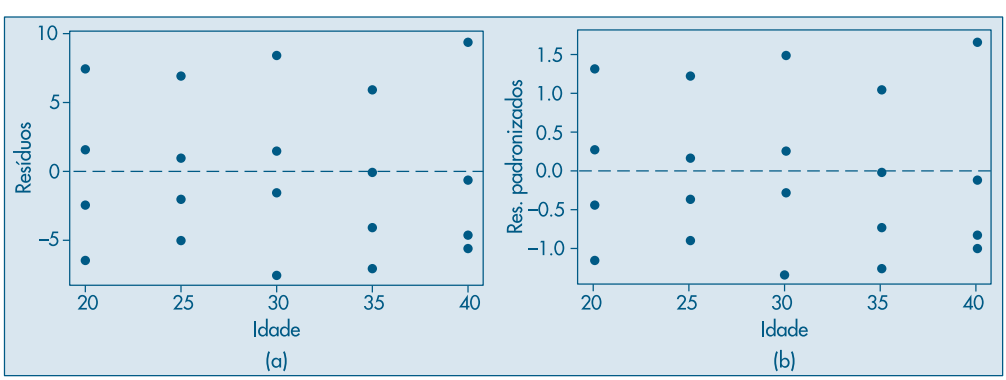
onde $v_{ii} = 1/n + (x_i - \bar{x})^2/\sum(x_i - \bar{x})^2$. O denominador de (16.62) é o desvio padrão de $\hat{e}_i$. Não iremos explorar aqui a análise feita com esse tipo de resíduo.

Exemplo 16.7. Voltemos ao Exemplo 15.1. Os resíduos do modelo (16.18) estão reproduzidos na Tabela 16.4, dos quais foram obtidos os demais. Os dois primeiros resíduos estão representados na Figura 16.6. Note que os dois gráficos são parecidos e levarão ao mesmo tipo de diagnóstico. Comentários adicionais sobre esse exemplo serão feitos abaixo.

Tabela 16.4: Resíduos para o modelo (16.18).

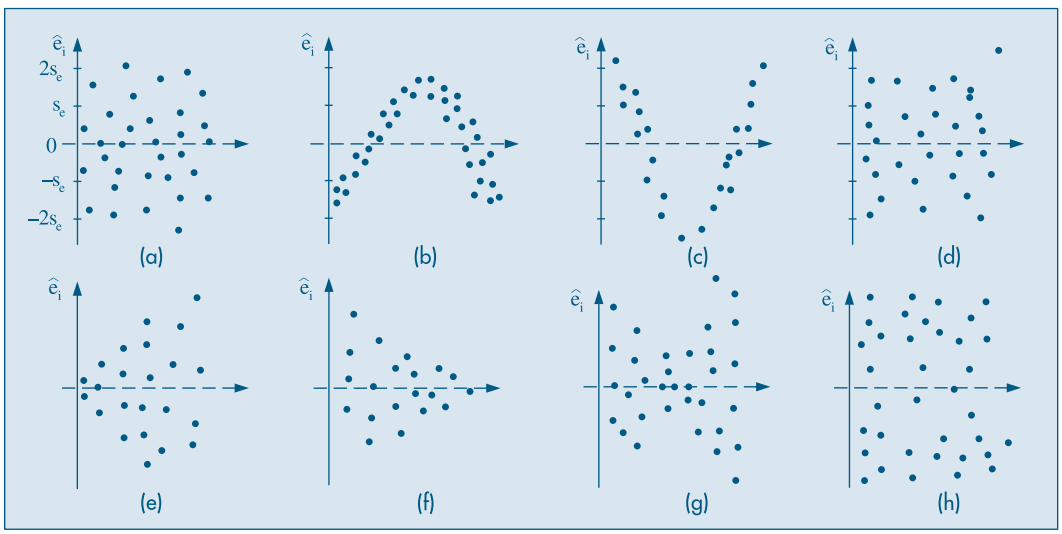
Idade	$\hat{e}_i$	$\hat{z}_i$	$\hat{r}_i$	Idade	$\hat{e}_i$	$\hat{z}_i$	$\hat{r}_i$
20	−2,5	−0,45	−0,49	30	1,5	0,27	0,28
20	−6,5	−1,16	−1,26	30	−7,5	−1,34	−1,37
20	7,5	1,34	1,45	35	0,0	0,0	0,0
20	1,5	0,27	0,29	35	−7,0	−1,25	−1,30
25	−5,0	−0,89	−0,92	35	6,0	1,07	1,11
25	1,0	0,18	0,19	35	−4,0	−0,72	−0,75
25	7,0	1,25	1,30	40	−4,5	−0,80	−0,86
25	−2,0	−0,36	0,37	40	−5,5	−0,98	−1,06
30	8,5	1,52	1,56	40	9,5	1,70	1,84
30	−1,5	−0,27	−0,28	40	−0,5	−0,09	−0,10

Figura 16.6: Resíduos para o Exemplo 16.1. (a) $\hat{e}_i = y_i - \hat{y}_i$; (b) resíduos padronizados.



Obtido o gráfico dos resíduos, precisamos saber como identificar possíveis inadequações. Apresentamos na Figura 16.7 alguns tipos usuais de gráficos de resíduos. A Figura 16.7 (a) é a situação ideal para os resíduos, distribuídos aleatoriamente em torno do zero, sem nenhuma observação muito discrepante.

Figura 16.7: Gráficos de resíduos. (a) situação ideal; (b), (c) modelo não-linear; (d) elemento atípico; (e), (f), (g) heterocedasticidade; (h) não-normalidade.



Nas situações (b) e (c) temos possíveis inadequações do modelo adotado, e as curvaturas sugerem que devemos procurar outras funções matemáticas que expliquem melhor o fenômeno.

A Figura 16.7 (d) mostra a existência de um elemento discrepante, e deve ser investigada a razão desse desvio tão marcante. Pode ser um erro de medida, ou a discrepância pode ser real. Em situações como essa, em que há observações muito diferentes das demais, métodos chamados robustos têm de ser utilizados.

Os casos (e), (f) e (g) indicam claramente que a suposição de homoscedasticidade (mesma variância) não está satisfeita. Em (h), parece haver maior incidência de observações nos extremos, mostrando que a suposição de normalidade não está satisfeita.

Analisados os resíduos e diagnosticada uma possível transgressão das suposições, devemos propor alterações que tornem o modelo mais adequado aos dados e às suposições feitas.

A verificação da hipótese de normalidade pode ser realizada fazendo-se um histograma dos resíduos ou um gráfico de $q \times q$, como explicado no Capítulo 3.

Exemplo 16.7. (continuação) A análise dos resíduos do modelo (16.18) mostra que esses não violam as suposições de média zero e variância comum. A Figura 16.8 mostra

o histograma dos resíduos, e a Figura 16.9 mostra um gráfico $q \times q$. Esse gráfico, feito com o SPlus, coloca nos eixos das ordenadas os valores crescentes dos $\hat{e}_i$ e no eixo das abscissas os quantis de uma normal padrão. Se os valores fossem de uma normal, eles deveriam se dispor ao longo de uma reta. Notamos que tanto o histograma quanto o gráfico de quantis mostram que os resíduos não são normalmente distribuídos.

Figura 16.8: Histograma dos resíduos do modelo (16.18).

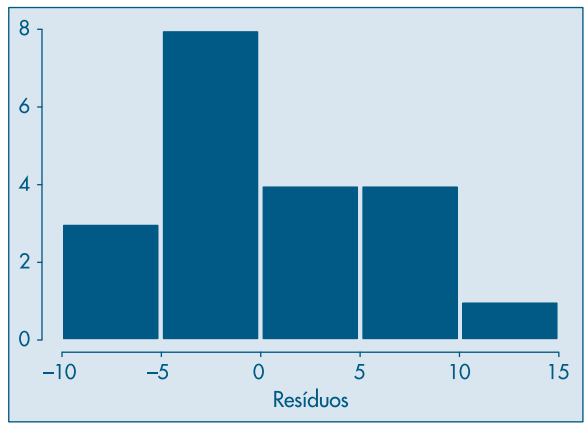
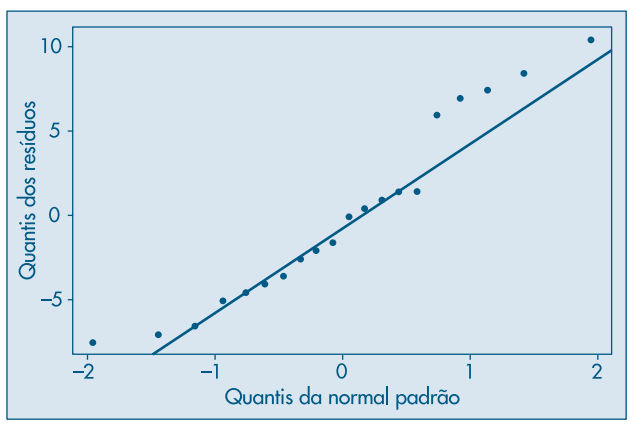


Figura 16.9: Gráfico $q \times q$ (normalidade) para os resíduos do modelo (16.18).



Quando a suposição de variância comum não estiver satisfeita, usualmente faz-se uma transformação da variável resposta y, ou da preditora x, ou de ambas. Para detalhes, ver Bussab (1986) e a seção 16.6.

Exemplo 16.8. Num processo industrial, além de outras variáveis, foram medidas: X = temperatura média (°F) e Y = quantidade de vapor. Os dados estão na Tabela 16.5 (Draper & Smith, 1998, Appendix A).

Tabela 16.5: Temperatura e quantidade de vapor de um processo industrial.

Nº	x_i	y_i	$\hat{e}_i$
1	35,3	10,98	0,174
2	29,7	11,13	−0,123
3	30,8	12,51	1,345
4	58,8	8,40	−0,531
5	61,4	9,27	0,547
6	71,3	8,73	0,797
7	74,4	6,36	−1,326
8	76,7	8,50	0,998
9	70,7	7,82	−0,161
10	57,5	9,14	0,106
11	46,4	8,24	−1,680
12	28,9	12,19	0,873
13	28,1	11,88	0,499
14	39,1	9,57	−0,933
15	46,8	10,94	1,052
16	48,5	9,58	−0,173
17	59,3	10,09	1,199
18	70,0	8,11	0,073
19	70,0	6,83	−1,207
20	74,5	8,88	1,202
21	72,1	7,68	−0,189
22	58,1	8,47	−0,517
23	44,6	8,86	−1,204
24	33,4	10,36	−0,598
25	28,6	11,08	−0,261

Fonte: Draper e Smith (1998).

O gráfico de dispersão e a reta de MQ estão na Figura 16.10 (a). A reta estimada de MQ é dada por

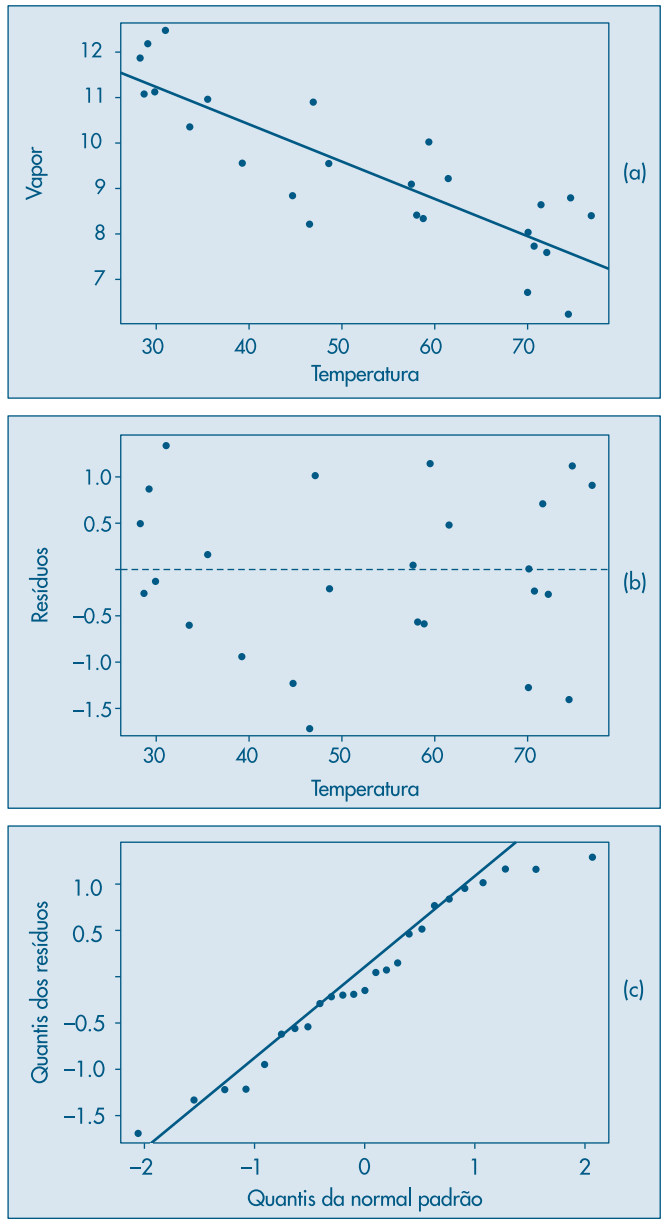
$$\hat{y}_i = 9{,}424 - 0{,}0798(x_i - 52{,}6), \tag{16.63}$$

ou ainda

$$\hat{y}_i = 13{,}623 - 0{,}0798x_i, \tag{16.64}$$

de modo que $\hat{\alpha} = 13{,}623$ e $\hat{\beta} = -0{,}0798$. Os resíduos $\hat{e}_i = y_i - \hat{y}_i$ estão na quarta coluna da Tabela 16.5 e seu gráfico contra x_i na Figura 16.10 (b). O gráfico $q \times q$ para verificar a suposição de normalidade está na Figura 16.10 (c). Observamos que há vários pontos afastados da reta.

Figura 16.10: (a) gráfico de dispersão com reta ajustada;
(b) resíduos vs temperatura;
(c) gráfico $q \times q$ (normalidade).



Problemas

13. Com o modelo linear já obtido para a acuidade visual como função da idade, construa os tipos de resíduos apresentados no Exemplo 16.6. Represente-os graficamente. Você observa alguma transgressão das suposições básicas?