Pedro A. Morettin
Wilton de O. Bussab

ESTATÍSTICA BÁSICA

6ª edição
Revista e atualizada

Editora Saraiva

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.

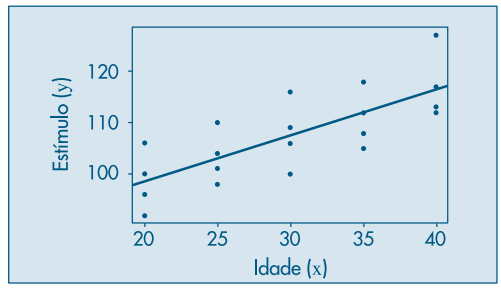
Continuando com o Exemplo 15.1, tanto X (idade) como Y (tempo de resposta ao estímulo) são v.a. contínuas, e podemos pensar em introduzir um modelo alternativo para y_{ij}, dada a relação entre X e Y. Observando as médias de Y, segundo os grupos de idades, ou seja, $E(Y|x)$, percebemos que estas aumentam 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.

Figura 16.1: Gráfico de dispersão de idade e reação ao estímulo, com reta ajustada.



A forma da função $\mu(x)$ deve ser definida pelo pesquisador, em função do grau de conhecimento teórico que ele tem do fenômeno sob estudo. Um modelo alternativo a (16.2) seria, então,

$$y_{ij} = \mu(x_i) + e_{ij}, \tag{16.4}$$

com $E(Y|x_i) = \mu(x_i) = \alpha + \beta x_i$, $i = 1, 2, ..., 5$. Entretanto, a forma usual de escrever o modelo é

$$y_i = \mu(x_i) + e_i, \tag{16.5}$$

onde y_i indica o tempo de reação do i-ésimo indivíduo com x_i anos de idade, $i = 1, 2, ..., n$, e n é o número total de observações. Teremos, então, com essa notação, valores repetidos para X, por exemplo, $x_1 = ... = x_4 = 20$. Convém reforçar a idéia que estamos propondo um modelo de comportamento para as médias das subpopulações, logo teremos de estimar os parâmetros envolvidos na função $\mu(x)$, baseados numa amostra de $n = 20$ observações, no exemplo.

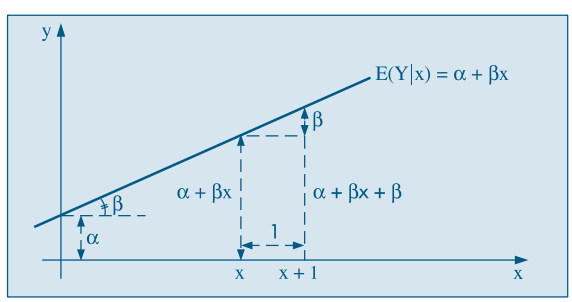
No caso de (16.3) o modelo pode ser escrito como

$$y_i = E(Y|x_i) + e_i = \alpha + \beta x_i + e_i, \quad i = 1, 2, ..., n, \tag{16.6}$$

devendo-se encontrar os valores mais prováveis para α e β, segundo algum critério, a partir de n observações de pares de valores de (X, Y).

Antes de prosseguirmos, seria conveniente interpretar os parâmetros envolvidos no modelo (16.5). Sabemos que α, o intercepto, representa o ponto onde a reta corta o eixo das ordenadas, e β, o coeficiente angular, representa o quanto varia a média de Y para um aumento de uma unidade da variável X. Esses parâmetros estão representados na Figura 16.2.

Figura 16.2: Representação do modelo $E(Y|x) = \alpha + \beta x$.



Voltando ao nosso exemplo, onde X é a idade e Y o tempo de reação, β representa o acréscimo no tempo médio de reação para cada ano de envelhecimento das pessoas. Aqui α representa o tempo de reação para a idade zero (recém-nascido), o que é uma inadequação do modelo.

Observação. Chamamos (16.3) de modelo linear, pois este representa uma reta. Todavia, em casos mais gerais, o termo *linear* refere-se ao modo como os *parâmetros entram no modelo*, ou seja, de forma linear. Por exemplo, o modelo

$$E(Y|x) = \alpha + \beta x + \gamma x^2,$$

embora graficamente represente uma parábola, é *modelo linear em* α, β *e* γ. Por outro lado,

$$E(Y|x) = \alpha e^{\beta x} \tag{16.7}$$

não é um modelo linear em α *e* β.

Determinados modelos não-lineares podem ser transformados em lineares, por meio de transformações das variáveis. Assim, tomando-se o logaritmo (de base e) em (16.7) obtemos

$$\ell nE(Y|x) = \ell n(\alpha) + \beta x = \alpha' + \beta x,$$

que é linear em α' *e* β.

Ao lado de um tratamento formal para estudar o modelo (16.6), devemos usar as técnicas de análise de dados que estudamos na Parte 1 do livro. Em particular, podemos fazer diversos tipos de gráficos *antes* que o modelo seja ajustado, *durante* o processo de ajuste e, finalmente, *depois* que o modelo foi ajustado.

A Figura 16.1 é um exemplo de um gráfico que deve ser feito antes de selecionar o modelo. Ou seja, temos um gráfico de dispersão entre as variáveis X (idade) e Y (tempo de reação ao estímulo). Esse tipo de diagrama permite ver qual o tipo de relação existente entre as variáveis, se há valores atípicos, se há valores repetidos (como no Exemplo 15.1), se a variabilidade de Y está aumentando ou não com X etc. Nesse mesmo exemplo, se decidirmos incluir a variável "acuidade visual" no modelo, teríamos duas variáveis explicativas e poderíamos fazer, por exemplo, gráficos de dispersão entre a resposta e cada variável explicativa e entre as duas variáveis explicativas. Este último nos daria uma idéia do *planejamento* envolvido, ou seja, se os pares de valores das variáveis explicativas estão cobrindo o plano (x_1, x_2), se há grupos de pontos etc.

Exemplos de gráficos depois do ajuste serão vistos na seção 16.5, quando fizermos uma análise dos resíduos, para avaliar a adequação do modelo aos dados. Gráficos durante o ajuste são utilizados quando estudarmos a possibilidade de considerar vários modelos alternativos para o problema em questão. Esse tópico não será explorado com detalhes no livro.

16.2 Estimação dos Parâmetros

Como no capítulo anterior, iremos encontrar os estimadores de mínimos quadrados para os parâmetros do modelo linear (16.6), mas o mesmo desenvolvimento pode ser aplicado em modelos mais complexos. Será necessário ainda introduzir algumas suposições para as v.a. envolvidas. A primeira delas é que a variável X é por hipótese controlada e não está sujeita a variações aleatórias. Dizemos que X é uma variável fixa (ou sem erro ou determinística). Segundo, para dado valor x de X, os erros distribuem-se ao redor da média $\alpha + \beta x$ com média zero, isto é,

$$E(e_i | x) = 0. \tag{16.8}$$

Em terceiro lugar, e pela mesma razão apresentada no capítulo anterior, devemos supor que os erros tenham a mesma variabilidade em torno dos níveis de X, ou seja,

$$\text{Var}(e_i | x) = \sigma_e^2. \tag{16.9}$$

E em quarto lugar, introduziremos a restrição de que os erros sejam não-correlacionados.

Colhida uma amostra de n indivíduos, teremos n pares de valores (x_i, y_i), $i = 1, ..., n$, que devem satisfazer ao modelo (16.6), isto é,

$$y_i = \alpha + \beta x_i + e_i, \quad i = 1, ..., n. \tag{16.10}$$

Temos, então, n equações e $n + 2$ incógnitas (α, β, e_1, e_2, ..., e_n). Precisamos introduzir um critério que permita encontrar α e β. Como no capítulo anterior, vamos adotar o critério que consiste em encontrar os valores de α e β que minimizam a soma dos quadrados dos erros, dados por

$$e_i = y_i - (\alpha + \beta x_i), \quad i = 1, ..., n. \tag{16.11}$$

Obtemos, então, a quantidade de informação perdida pelo modelo ou soma dos quadrados dos erros (ou desvios)

$$\text{SQ}(\alpha, \beta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \{y_i - (\alpha + \beta x_i)\}^2. \tag{16.12}$$

Para cada valor de α e β teremos um resultado para essa soma de quadrados, e a solução de mínimos quadrados (MQ) é aquela que torna essa soma mínima. Temos, então, o problema de encontrar o mínimo de uma função de duas variáveis, α e β, no caso (ver Morettin et al., 2005). Derivando em relação a α e β e igualando a zero, observamos que as soluções $\hat{\alpha}$ e $\hat{\beta}$ devem satisfazer

$$\begin{aligned} n\hat{\alpha} + \hat{\beta} \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i, \\ \hat{\alpha} \sum_{i=1}^{n} x_i + \hat{\beta} \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i, \end{aligned} \tag{16.13}$$

as quais produzem as soluções

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x},$$
$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}. \tag{16.14}$$

Substituindo em (16.3), teremos o estimador para a média $\mu(x)$, dado por

$$\hat{\mu}(x_i) = \hat{\alpha} + \hat{\beta}x_i, \quad i = 1, ..., n, \tag{16.15}$$

que iremos indicar por

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i, \tag{16.16}$$

ou, ainda, por

$$\hat{y}_i = \bar{y} - \hat{\beta}\bar{x} + \hat{\beta}x_i = \bar{y} + \hat{\beta}(x_i - \bar{x}). \tag{16.17}$$

Exemplo 16.1. Voltemos ao Exemplo 15.1 e vamos ajustar o modelo (16.10), com:

y_i: tempo de reação do i-ésimo indivíduo,

x_i: idade do i-ésimo indivíduo,

e_i: desvio, $i = 1, 2, ..., 20$.

Da Tabela 16.1 obtemos as informações:

$$n = 20, \quad \sum y_i = 2.150, \quad \sum x_i = 600, \quad \sum x_i y_i = 65.400,$$
$$\bar{y} = 107{,}50, \quad \bar{x} = 30, \quad \sum x_i^2 = 19.000.$$

Substituindo em (16.14) obtemos

$$\hat{\beta} = \frac{65.400 - (20)(30)(107{,}50)}{19.000 - (20)(30)^2} = 0{,}90,$$
$$\hat{\alpha} = 107{,}50 - (0{,}90)(30) = 80{,}50,$$

o que nos dá o *modelo ajustado*

$$\hat{y}_i = 80{,}50 + 0{,}90x_i, \quad i = 1, 2, ..., 20. \tag{16.18}$$

Com esse modelo podemos prever, por exemplo, o tempo médio de reação para pessoas de 20 anos, que será indicado por $\hat{y}$ (20) e determinado por

$$\hat{y}(20) = 80{,}50 + (0{,}90)(20) = 98{,}50.$$

De modo análogo, os tempos médios para as idades 25, 30, 35 e 40 serão, respectivamente, estimados por: 103,00, 107,50, 112,00, e 116,50. Esses valores são muito próximos daqueles encontrados na seção 15.3, e a vantagem desse modelo sobre aquele é a possibilidade de estimar o tempo de reação médio para um grupo de idades não observado. Suponhamos, por exemplo, que se deseja estimar o tempo médio para o grupo de pessoas com 33 anos; este será dado por

$$\hat{y}(33) = 80{,}50 + (0{,}90)(33) = 110{,}20.$$

Na Figura 16.1 aparecem representados os dados observados, bem como a reta ajustada. Podemos observar que o modelo parece ser adequado, não apresentando nenhum ponto com desvio exagerado.

Problemas

1. Usando os dados do Exemplo 15.1:
   (a) Encontre a reta de mínimos quadrados $\hat{z}_i = \alpha + \beta x_i$, onde z mede a acuidade visual e x, a idade.
   (b) Interprete o significado de α e β nesse problema.
   (c) Para cada indivíduo, encontre o desvio $\hat{e}_i = z_i - \hat{z}_i$; existe algum com valor muito exagerado?

2. A tabela abaixo indica o valor y do aluguel e a idade x de cinco casas.
   (a) Encontre a reta de MQ, supondo a relação $E(y|x) = \alpha + \beta x$.
   (b) Faça o gráfico dos pontos e da reta ajustada. Você acha que o modelo adotado é razoável?
   (c) Qual o significado do coeficiente angular nesse caso?
   (d) E do coeficiente linear?

x	10	13	5	7	20
y	4	3	6	5	2

3. Um laboratório está interessado em medir o efeito da temperatura sobre a potência de um antibiótico. Dez amostras de 50 gramas cada foram guardadas a diferentes temperaturas, e após 15 dias mediu-se a potência. Os resultados estão no quadro abaixo.
   (a) Faça a representação gráfica dos dados.
   (b) Ajuste a reta de MQ, da potência como função da temperatura.
   (c) O que você acha desse modelo?
   (d) A que temperatura a potência média seria nula?

Temperatura	30°		50°			70°			90°	
Potência	38	43	32	26	33	19	27	23	14	21

4. Ainda usando os dados do exemplo numérico 15.1, investigue o ajuste da reta de MQ na variável tempo de reação como função da acuidade visual.

16.3 Avaliação do Modelo

Nesta seção e nas seguintes estudaremos várias formas de avaliar se o modelo linear postulado é adequado ou não, dadas as suposições que fizemos sobre ele.

16.3.1 Estimador de σ_e^2

Como no capítulo anterior, para julgar a vantagem da adoção de um modelo mais complexo (linear ou outro qualquer), vamos usar a estratégia de compará-lo com o modelo mais simples, que é aquele discutido na seção 15.2, ou seja,

$$y_i = \mu + e_i. \tag{16.19}$$

A vantagem será sempre medida por meio da diminuição dos erros de previsão, ou ainda, da variância residual S_e^2. Para o modelo ajustado (16.16), cada *resíduo* é dado por

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta} x_i. \tag{16.20}$$

Como vimos na seção 16.1, vários gráficos envolvendo esses resíduos podem ser feitos para avaliar se eles são "bons representantes" dos verdadeiros e_i desconhecidos, no sentido de que as suposições feitas sobre estes estão satisfeitas. Esses gráficos serão estudados na seção 16.5.

Quando estes resíduos forem pequenos, temos uma indicação de que o modelo está produzindo bons resultados. Para julgarmos se o resíduo é pequeno ou não, devemos compará-lo com os resíduos do modelo alternativo, dados por $y_i - \bar{y}$. Da dificuldade de compará-los individualmente, preferimos trabalhar com as respectivas somas de resíduos quadráticos, dadas por

$$\text{SQTot} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{16.21}$$

e

$$\text{SQRes} = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{16.22}$$

Exemplo 16.1. (continuação) Na quinta coluna da Tabela 16.1 aparecem os resíduos

$$\hat{e}_i = y_i - \hat{y}_i = y_i - (80{,}50 + 0{,}90x_i)$$

que elevados ao quadrado e somados produzirão

$$\text{SQRes} = 563{,}00.$$

Sabemos que SQTot = 1.373,00, o que mostra uma sensível redução de 810 unidades. Mais ainda, a comparação da quinta coluna da Tabela 16.1 com a coluna $e(3)$ da Tabela 15.4 mostra o melhor comportamento dos resíduos do modelo de regressão (16.18).

Tabela 16.1: Resíduos para o modelo (16.18).

i	Variáveis			Resíduos
	Tempo de Reação	Sexo	Idade	$y_i - \hat{y}_i$
1	96	H	20	–2,5
2	92	M	20	–6,5
3	106	H	20	7,5
4	100	M	20	1,5
5	98	M	25	–5,0
6	104	H	25	1,0
7	110	H	25	7,0
8	101	M	25	–2,0
9	116	M	30	8,5
10	106	H	30	–1,5
11	109	H	30	1,5
12	100	M	30	–7,5
13	112	M	35	0,0
14	105	M	35	–7,0
15	118	H	35	6,0
16	108	H	35	–4,0
17	113	M	40	–4,5
18	112	M	40	–5,5
19	127	H	40	9,5
20	117	H	40	–0,5
SQRes				563
S_e^2				31,28
S_e				5,59
$2S_e$				11,18

No entanto, a comparação direta dessas somas de quadrados não nos parece justa, pois o modelo (16.18) tem mais parâmetros do que o modelo (16.19). Vejamos, então, como comparar as variâncias residuais. Para o modelo simples (16.19) o estimador não-viesado de σ_e^2 é

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{\text{SQTot}}{n-1}. \tag{16.23}$$

Também vimos que para o modelo (16.2), com I níveis ou subpopulações, o estimado da variância residual era

$$S_e^2 = \frac{\text{SQDen}}{n-I} = \frac{\text{SQRes}}{n-I}, \tag{16.24}$$

e I também denota o número de parâmetros desconhecidos do modelo (as médias μ_i). Portanto, de modo geral, perde-se um grau de liberdade para cada parâmetro envolvido no modelo e é natural definir o estimador de σ_e^2 num modelo de regressão como sendo

$$S_e^2 = \frac{\text{SQRes}}{n-p}, \tag{16.25}$$

onde p é o número de parâmetros do modelo. No caso particular da regressão linear simples, $p = 2$ e

$$S_e^2 = \frac{\text{SQRes}}{n-2}, \tag{16.26}$$

será um estimador não-viesado de σ_e^2, isto é, $E(S_e^2) = \sigma_e^2$. Veja o Problema 32.

Exemplo 16.2. Continuando o exemplo anterior, obteremos

$$S^2 = 1.373/19 = 72{,}26, \quad S = 8{,}50$$

e

$$S_e^2 = 563/18 = 31{,}28, \quad S_e = 5{,}59,$$

números que sugerem uma diminuição significativa nos resíduos. Observe que, passando de um modelo com um parâmetro para outro com dois, há uma redução de 813 unidades na soma de quadrados residuais. Ou seja, perdendo um grau de liberdade, reduziu-se a soma dos resíduos quadráticos em 810 unidades, o que é mais uma evidência da vantagem de adoção do segundo modelo.

16.3.2 Decomposição da Soma de Quadrados

Ao passarmos do modelo simples para o modelo de regressão linear, vimos que a redução da soma de quadrados é dada por SQTot – SQRes. Esse lucro é devido à adoção do segundo modelo e será indicado por SQReg, significando a *soma dos quadrados devida à regressão*. Segue-se que

$$\text{SQReg} = \text{SQTot} - \text{SQRes}, \tag{16.27}$$

ou seja,

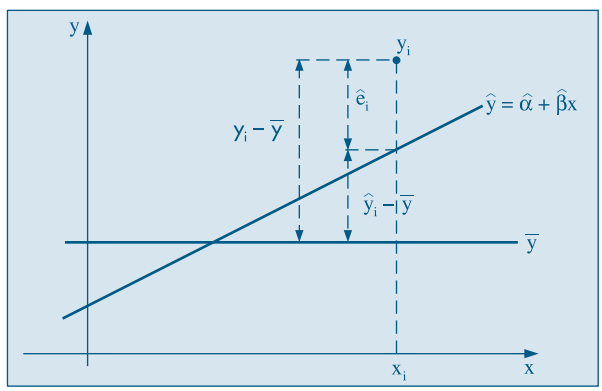
$$\text{SQTot} = \text{SQReg} + \text{SQRes}. \tag{16.28}$$

Observando a Figura 16.3, notamos que vale a seguinte relação:

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) = \hat{e}_i + (\hat{y}_i - \bar{y}). \tag{16.29}$$

Em palavras, o desvio de uma observação em relação à média pode ser decomposto como o desvio da observação em relação ao valor ajustado pela regressão, mais o desvio do valor ajustado em relação à média.

Figura 16.3: Representação gráfica dos diversos desvios.



Elevando-se ao quadrado ambos os membros da igualdade (16.29), tomando-se a soma e observando-se que a soma do duplo produto se anula (veja o Problema 31), obtemos

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} \hat{e}_i^2, \tag{16.30}$$

ou

$$\text{SQTot} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \text{SQRes}, \tag{16.31}$$

do que deduzimos que

$$\text{SQReg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2. \tag{16.32}$$

De (16.17) obtemos que

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

portanto, podemos escrever

$$\text{SQReg} = \hat{\beta}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{16.33}$$

Daqui se pode observar que, quanto maior o valor de $\hat{\beta}$, maior será a redução da soma dos quadrados dos resíduos.

16.3.3 Tabela de Análise de Variância

Do mesmo modo como foi feito na seção 15.2, podemos resumir as informações anteriores numa única tabela ANOVA, ilustrada na Tabela 16.2.

Tabela 16.2: Tabela ANOVA para modelo de regressão.

F.V.	g.l.	SQ	QM	F
Regressão	1	SQReg	SQReg = QMReg	$QMReg/S_e^2$
Resíduo	$n-2$	SQRes	$SQRes/(n-2) = S_e^2$	
Total	$n-1$	SQTot	$SQTot/(n-1) = S^2$	

Também podemos medir o lucro relativo que se ganha ao introduzir o modelo, usando a estatística

$$R^2 = \frac{\text{SQReg}}{\text{SQTot}}, \tag{16.34}$$

definida anteriormente. A estatística F será discutida na seção 16.4.

Exemplo 16.3. Dos cálculos que nos levaram ao modelo (16.18), podemos construir a Tabela 16.3. Temos que

$$R^2 = \frac{810}{1.373} = 59\%.$$

Tabela 16.3: Tabela ANOVA para o modelo (16.18).

F.V.	g.l.	SQ	QM	F
Regressão	1	810	810	25,90
Resíduo	18	563	31,28	
Total	19	1.373	72,26	

O modelo proposto diminui a variância residual em mais da metade e explica 59% da variabilidade total. Verificamos, então, que é vantajosa a adoção do modelo linear (16.18) para explicar o tempo médio de reação ao estímulo, em função da idade.

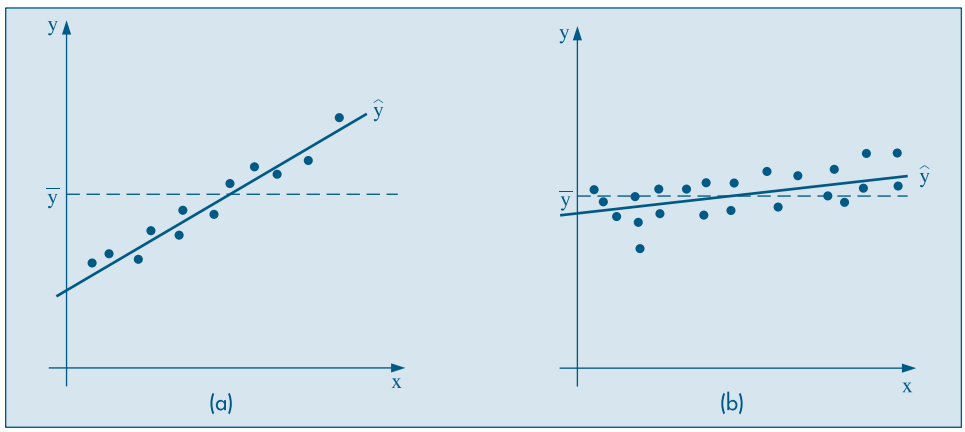
A estratégia adotada para verificar se compensa ou não utilizar o modelo $y = \alpha + \beta x + e$ é observar a redução no resíduo quando comparado com o modelo $y = \mu + e$. Se a redução for muito pequena, os dois modelos serão praticamente equivalentes, e isso ocorre quando a inclinação β for zero ou muito pequena, não compensando usar um modelo mais complexo. Estaremos, pois, interessados em testar a hipótese

$$H_0: \beta = 0, \tag{16.35}$$

o que irá exigir que se coloque uma estrutura de probabilidades sobre os erros. Esse assunto será objeto da próxima seção. A Figura 16.4 ilustra as duas situações que podem ocorrer.

Na Figura 16.4 (a) temos o caso em que claramente a variável auxiliar ajuda a prever a variável resposta. Na situação da Figura 16.4 (b) teremos dúvidas se vale a pena ou não introduzir um modelo mais complexo, ganhando muito pouco em termos de explicação.

Figura 16.4: Retas ajustadas a dois conjuntos de dados. (a) x explica y; (b) x não explica y.



Para a avaliação final do modelo devemos investigar com mais cuidado o comportamento dos resíduos, o que será feito na seção 16.5.

Problemas

5. Usando os resultados do Problema 1, construa a tabela ANOVA para o modelo $\hat{z} = \hat{\alpha} + \hat{\beta}x$, encontrado naquele problema.
   (a) Qual a estimativa S^2? E S_e^2?
   (b) Você acha que a redução nos resíduos foi grande?
   (c) Qual o valor de R^2? Interprete esse número.

6. Um estudo sobre duração de certas operações está investigando o tempo requerido (em segundos) para acondicionar objetos e o volume (em dm^3) que eles ocupam. Uma amostra foi observada e obtiveram-se os seguintes resultados:

Tempo	10,8	14,4	19,6	18,0	8,4	15,2	11,0	13,3	23,1
Volume	20,39	24,92	34,84	31,72	13,59	30,87	17,84	23,22	39,65

   (a) Faça o diagrama de dispersão dos dados.
   (b) Estime a reta de regressão do tempo de operação em função do volume.
   (c) Construa a tabela ANOVA para o modelo.
   (d) Qual o valor de S^2? É pequeno quando comparado com S_e^2?
   (e) Você acha que conhecer o volume do pacote ajuda a prever o tempo de empacotamento?

7. Construa a tabela ANOVA para o Problema 2 e interprete os resultados.

8. Construa a tabela ANOVA com os dados do Problema 3.

9. Idem para o Problema 4.

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;

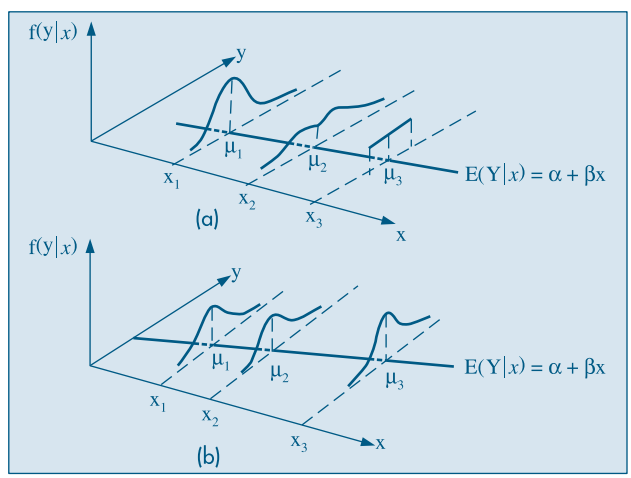
(ii) Se $i \neq j$, $\text{Cov}(e_i, e_j) = 0$, isto é, para duas observações distintas, os erros são não-correlacionados.

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:

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

    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

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

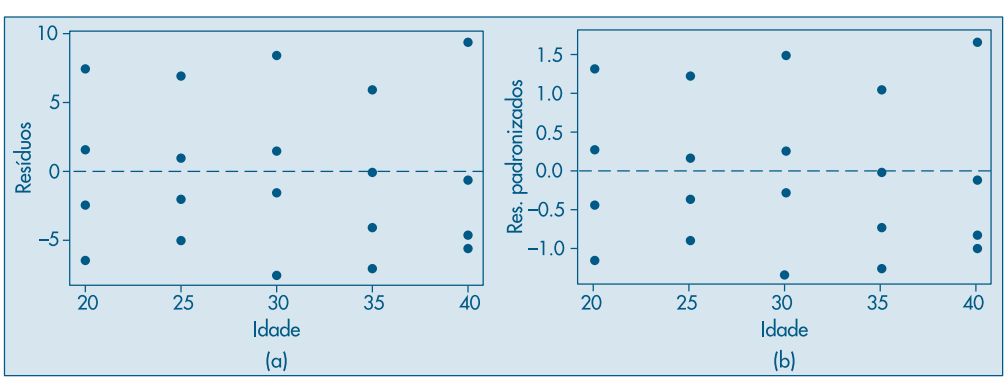
onde $\nu_{ii} = 1/n + (x_i - \bar{x})^2/\sum(x_i - \bar{x})^2$. O denominador de (16.62) é o desvio padrão de $\hat{e}_i$. Não iremos explorar aqui a análise feita com esse tipo de resíduo.

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).

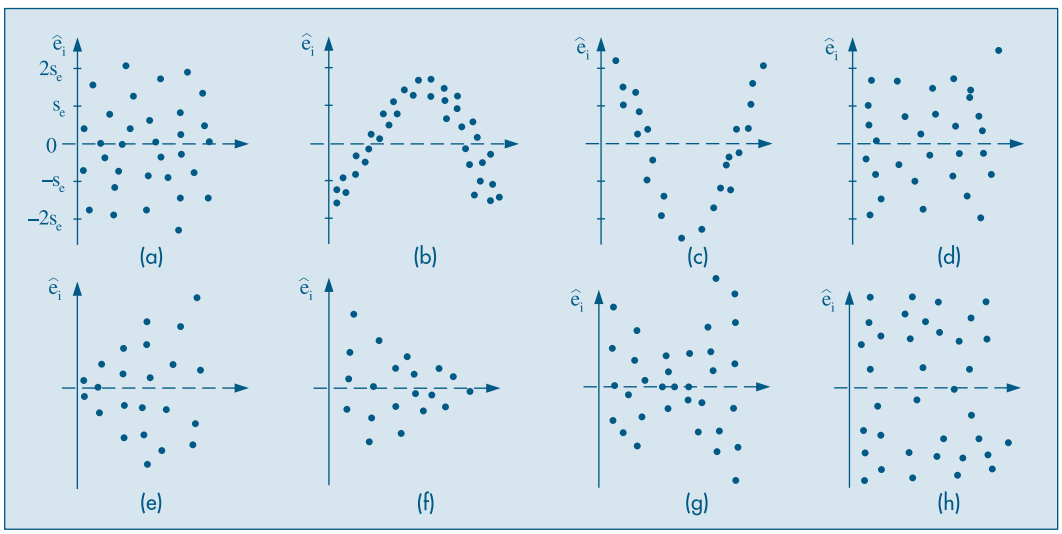
Idade	$\hat{e}_i$	$\hat{z}_i$	$\hat{r}_i$	Idade	$\hat{e}_i$	$\hat{z}_i$	$\hat{r}_i$
20	–2,5	–0,45	–0,49	30	1,5	0,27	0,28
20	–6,5	–1,16	–1,26	30	–7,5	–1,34	–1,37
20	7,5	1,34	1,45	35	0,0	0,0	0,0
20	1,5	0,27	0,29	35	–7,0	–1,25	–1,30
25	–5,0	–0,89	–0,92	35	6,0	1,07	1,11
25	1,0	0,18	0,19	35	–4,0	–0,72	–0,75
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).

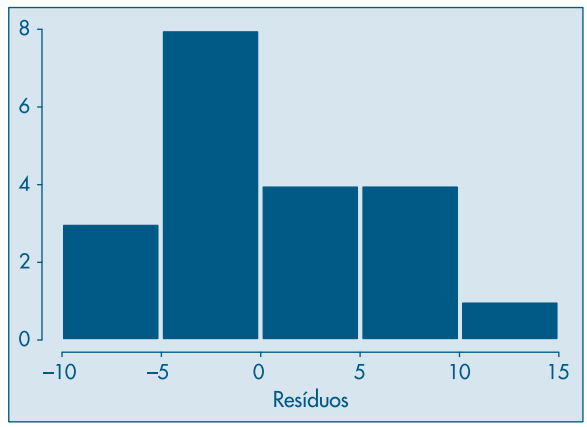
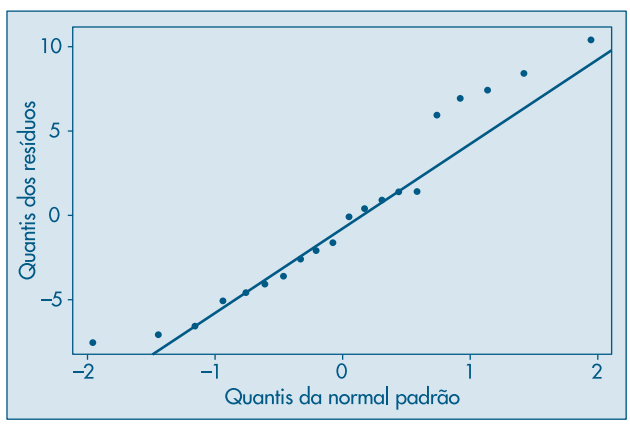


Figura 16.9: Gráfico $q \times q$ (normalidade) para os resíduos do modelo (16.18).



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

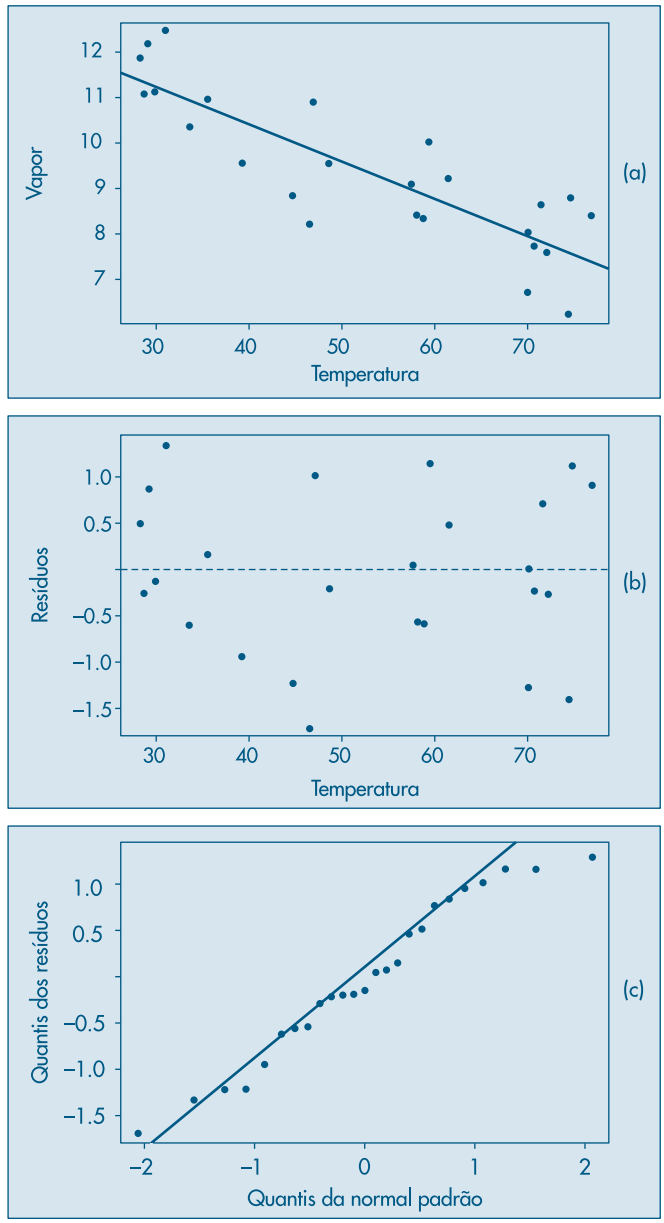
$$\hat{y}_i = 9{,}424 - 0{,}0798(x_i - 52{,}6), \tag{16.63}$$

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.

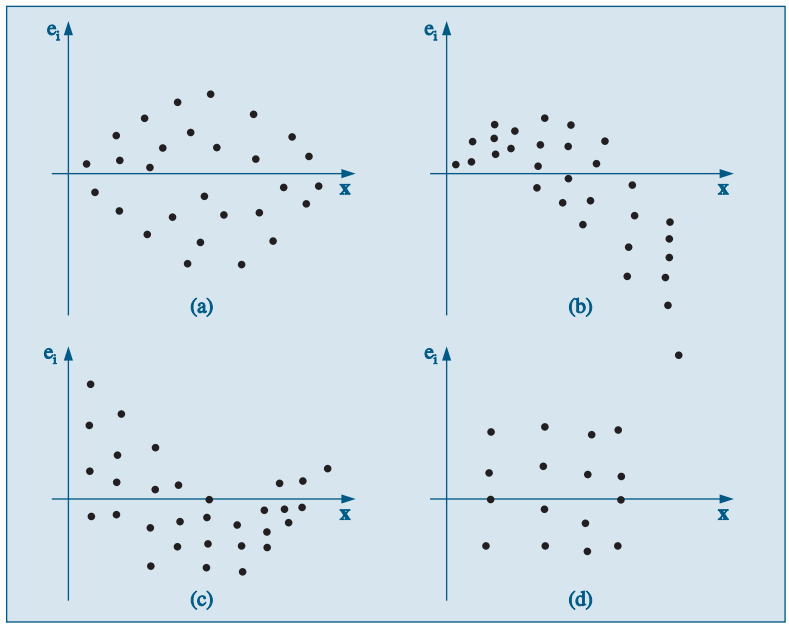
Figura 16.10: (a) gráfico de dispersão com reta ajustada;
(b) resíduos vs temperatura;
(c) gráfico $q \times q$ (normalidade).



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?

14. Para cada gráfico de resíduo abaixo, indique qual a possível transgressão observada.



15. Abaixo estão os valores da variável preditora (x), os resíduos observados depois do ajuste do modelo e a ordem em que os dados foram obtidos.

Preditor	11	20	14	22	12	25	15
Resíduo	–1	–2	3	–3	–1	5	0
Ordem	9	6	13	1	7	14	8

Preditor	14	19	21	18	22	16	21
Resíduo	0	3	–2	2	–5	0	1
Ordem	3	12	4	11	2	10	5

(a) Verifique se existe alguma possível transgressão das suposições, analisando o gráfico $(x_i, \hat{e}_i)$.

(b) Faça o gráfico do resíduo contra a ordem do experimento. Você observa alguma inconveniência?

16.6 Alguns Modelos Especiais

Nesta seção introduziremos alguns modelos particulares simples e que são de interesse prático. Iniciamos com o modelo que teoricamente passa pela origem. Depois, consideramos modelos não-lineares, mas que podem ser linearizados por meio de alguma transformação.

16.6.1 Reta Passando pela Origem

Em algumas situações temos razões teóricas (ou ditadas pelas peculiaridades do problema a analisar) para supor que o modelo deva ser do tipo

$$y_i = \beta x_i + e_i, \quad i = 1, ..., n. \tag{16.65}$$

Com as mesmas suposições anteriores e observada uma amostra (x_i, y_i), $i = 1, ..., n$, é fácil ver que o EMQ de β é

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}. \tag{16.66}$$

Deixamos a cargo do leitor verificar como ficam os resultados obtidos anteriormente para o modelo completo nesse caso particular. Por exemplo,

$$E(\hat{\beta}) = \beta,$$
$$\text{Var}(\hat{\beta}) = \frac{\sigma_e^2}{\sum_{i=1}^{n} x_i^2}.$$

Exemplo 16.9. A mensuração exata (Y) de uma substância do sangue, por meio de uma análise química, é muito cara. Um novo método mais barato resulta na medida X, que supostamente pode ser usada para prever o valor de Y. Nove amostras de sangue foram obtidas e avaliadas pelos dois métodos, obtendo-se as medidas abaixo.

X	119	155	174	190	196	233	272	253	276
Y	112	152	172	183	192	228	263	239	263

Algumas estatísticas obtidas são:

$$n = 9, \qquad \sum_i x_i = 1.868, \qquad \sum_i y_i = 1.804,$$
$$\sum_i x_i y_i = 396.933, \qquad \sum_i x_i^2 = 411.436, \qquad \sum_i y_i^2 = 383.028.$$

Vamos ajustar o modelo (16.65) a esses dados. Obtemos

$$\hat{\beta} = 396.933/411.436 = 0{,}9648,$$

resultando no modelo ajustado

$$\hat{y}_i = 0{,}9648 x_i, \qquad i = 1, 2, ..., 9.$$

É fácil ver que $S_e^2 = 5{,}9136$ e $S_e = 2{,}4318$. Para testar a hipótese H_0: $\beta = 0$, usamos a estatística

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

que resulta ser igual a $t(\hat{\beta}) = (0{,}9648/2{,}4318)\sqrt{411.436} = 254{,}48$, o que claramente leva à rejeição de H_0. Um intervalo de confiança para β, com coeficiente de confiança 95% é

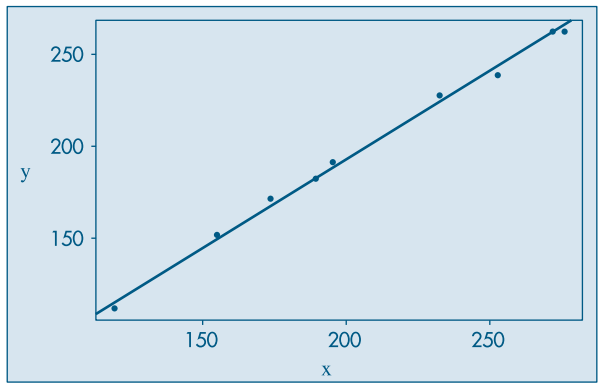
$$0{,}9648 \pm (2{,}306)\ \frac{2{,}4318}{\sqrt{411.436}} = 0{,}9648 \pm 0{,}0087,$$

ou seja,

$$\text{IC}(\beta;\ 0{,}95) = [0{,}9561;\ 0{,}9735].$$

Os dados e a reta ajustada estão na Figura 16.11.

Figura 16.11: Dados e reta ajustada para o Exemplo 16.8.



16.6.2 Modelos Não-Lineares

Quando usamos modelos de regressão, ou qualquer outro tipo de modelo, a situação ideal é aquela em que o pesquisador, por razões teóricas inerentes ao problema real sob estudo, pode sugerir a forma funcional da relação entre duas ou mais variáveis. Na prática, isso nem sempre acontece. Muitas vezes o pesquisador está interessado em usar técnicas de regressão para explorar modelos convenientes sugeridos pelos dados observados.

Como vimos, o primeiro passo para investigar o tipo de modelo a ser adotado é a representação gráfica dos dados, a qual pode sugerir a forma da curva relacionando as variáveis, além de fornecer outras informações (veja o final da seção 16.1). Por exemplo, com os dados da Tabela 16.6 obtemos o diagrama de dispersão da Figura 16.12. Notamos claramente a inadequação da reta como modelo, sendo que provavelmente uma relação exponencial do tipo

$$f(x) = \alpha e^{\beta x} \tag{16.67}$$

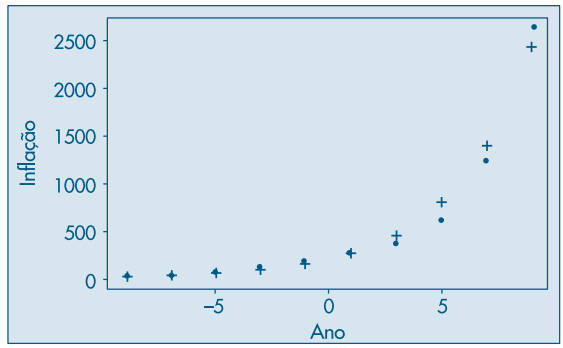
seja mais adequada. Um modelo que pode, então, ser sugerido, é

$$y_i = \alpha e^{\beta x_i} + \varepsilon_i, \quad i = 1, ..., n. \tag{16.68}$$

Tabela 16.6: Taxa de Inflação no Brasil de 1961 a 1979.

Ano	t	Inflação (Y)	$Y^* = \log Y$
1961	–9	9	2,2
1963	–7	24	3,2
1965	–5	72	4,3
1967	–3	128	4,8
1969	–1	192	5,2
1971	1	277	5,6
1973	3	373	5,9
1975	5	613	6,4
1977	7	1.236	7,1
1979	9	2.639	7,9

Figura 16.12: Dados de inflação no Brasil (pontos) e modelo exponencial ajustado (+).



Suponha que queiramos estimar os parâmetros α e β pelo método de mínimos quadrados. Devemos minimizar

$$S(\alpha, \beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha e^{\beta x_i})^2. \tag{16.69}$$

Derivando S em relação a α e β e igualando a zero, obtemos as duas equações

$$\hat{\alpha} \sum_{i=1}^{n} e^{2\hat{\beta} x_i} = \sum_{i=1}^{n} y_i e^{\hat{\beta} x_i},$$

$$\hat{\alpha}^2 \sum_{i=1}^{n} x_i e^{2\hat{\beta} x_i} = \hat{\alpha} \sum_{i=1}^{n} x_i \, y_i e^{\hat{\beta} x_i}. \tag{16.70}$$

A solução desse sistema de equações não-lineares exige o uso de procedimentos de otimização não-lineares, como Newton-Raphson, Gauss-Newton, "scoring" e outros. Ou seja, os pontos de máximo da função S são obtidos numericamente, dada a impossibilidade de termos soluções analíticas para as equações (16.70). Mas devemos dizer que essa é a regra, mais do que a exceção, em problemas encontrados na prática. Portanto, a utilização desses procedimentos de otimização é um requisito importante para estudantes de áreas como estatística, economia, engenharia etc.

Neste livro, vamos nos limitar a tratar de alguns casos onde transformações das variáveis sob estudo permitirão o uso de um modelo linear simples.

Suponha que a função (16.67) seja apropriada para os dados da Tabela 16.6. Considere o modelo

$$y_i = \alpha e^{\beta x_i} \varepsilon_i, \quad i = 1, \ldots, n. \tag{16.71}$$

Observe que nesse modelo os erros ε_i entram de forma *multiplicativa* e não aditiva, como no caso do modelo (16.6). Considerando, agora, o logaritmo (na base e) de ambos os lados de (16.71) e chamando

$$y_i^* = \log y_i, \quad \alpha^* = \log \alpha, \quad \varepsilon_i^* = \log \varepsilon_i, \tag{16.72}$$

podemos escrever o modelo na forma

$$y_i^* = \alpha^* + \beta x_i + \varepsilon_i^*, \quad i = 1, \ldots, n. \tag{16.73}$$

Note que esse modelo é *linear* em α^* e β, e temos que supor que os erros ε_i sejam *positivos*; do contrário, não podemos tomar logaritmos deles. Por outro lado, os erros

ε_i^* podem ser negativos, positivos ou nulos. Portanto, para o modelo linear (16.73) podemos fazer as suposições usuais das seções anteriores.

Exemplo 16.10. Utilizando os dados da Tabela 16.6, devemos, inicialmente, calcular os logaritmos naturais da variável *Y*. Note que nesse exemplo a variável explicativa é o tempo, convenientemente codificado. Na Figura 16.13 temos o diagrama de dispersão dos dados transformados e da reta ajustada, a saber

$$\hat{y}_i^* = 5{,}27 + 0{,}28t, \quad t = -9, \ldots, 9. \tag{16.74}$$

A análise de tal modelo pode ser conduzida como antes. Veja o Problema 35.

Observe que o modelo original ajustado é

$$\hat{y}_i = 194{,}42 \cdot e^{0{,}28t}, \quad i = 1, \ldots, 10, \tag{16.75}$$

pois $\alpha = e^{5{,}27}$. Essa curva está representada na Figura 16.12. Os resíduos do modelo (16.74), transformado, e do modelo (16.75), original, são dados na Tabela 16.7 e nas Figuras 16.14 e 16.15, respectivamente. Note que em ambos os casos os resíduos não parecem ser aleatórios, havendo curvaturas, sugerindo a possibilidade de um modelo com termos quadráticos ou cúbicos, por exemplo.

Figura 16.13: Diagrama de dispersão para o logaritmo da inflação com reta ajustada.

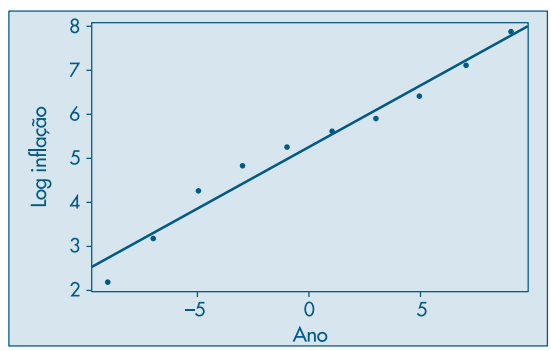


Tabela 16.7: Resíduos para os modelos linear e exponencial.

t	Resíduos Reta	Resíduos Exponencial
–9	–0,55	–6,643
–7	–0,11	–3,386
–5	0,43	24,057
–3	0,37	44,067
–1	0,21	45,061
1	0,05	19,757
3	–0,21	–77,348
5	–0,27	–175,412
7	–0,13	–145,251
9	0,11	222,632

Figura 16.14: Resíduos da reta ajustada ao logaritmo da inflação *versus* ano.

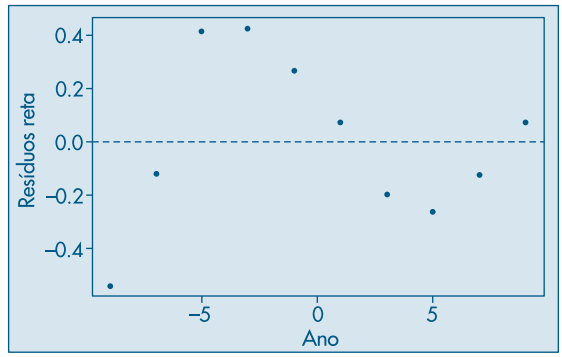
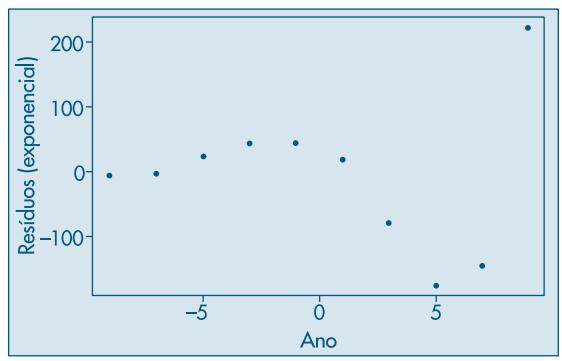


Figura 16.15: Resíduos do modelo exponencial ajustado aos dados originais *versus* ano.



Os histogramas e gráficos $q \times q$ para normalidade dos resíduos estão nas Figuras 16.16 e 16.17. Notamos que o histograma é assimétrico, mostrando claramente o valor correspondente a $t = 9$. Como há poucos pontos, a análise de resíduos fica prejudicada; o gráfico $q \times q$ mostra os pontos não muito próximos de retas.

Figura 16.16: Histogramas: (a) resíduos reta ajustada ao log (inflação); (b) resíduos modelo exponencial.

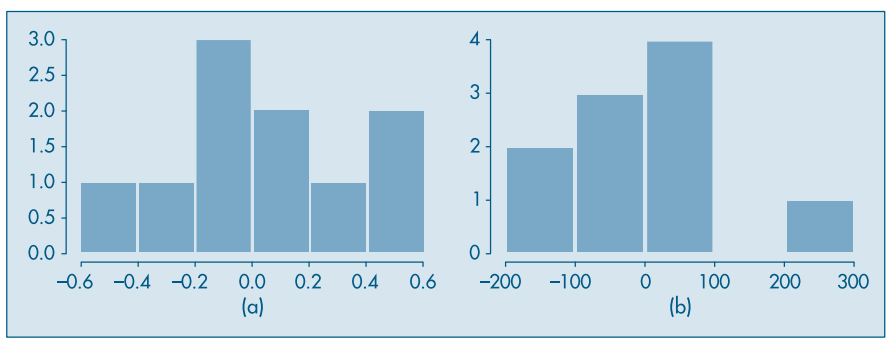
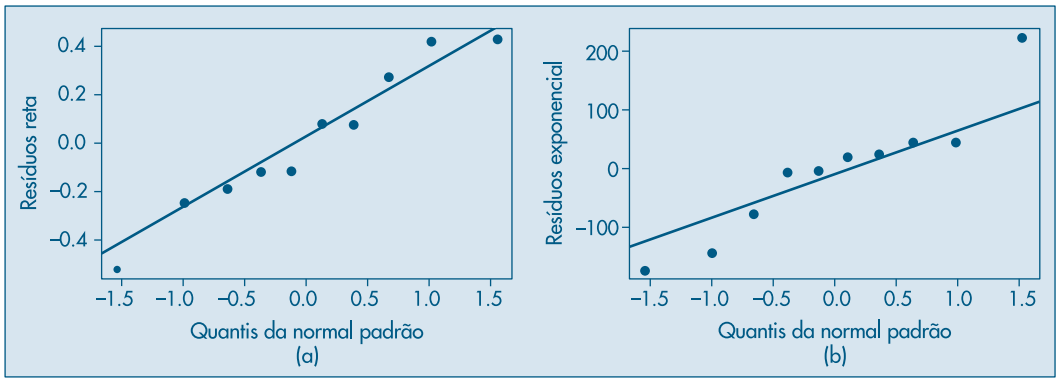


Figura 16.17: Gráficos $q \times q$ dos resíduos: (a) reta; (b) exponencial.



16.7 Regressão Resistente

Nesta seção vamos considerar apenas o caso de regressão linear simples. Ou seja, temos os valores observados (x_i, y_i), $i = 1, ..., n$ e queremos ajustar o modelo (16.6).

Notamos que os estimadores $\hat{\alpha}$ e $\hat{\beta}$ em (16.14) são baseados em $\bar{x}, \bar{y}$ e desvios em relação a essas médias.

A regressão resistente baseia-se em medianas, em vez de médias. Inicialmente, dividimos o conjunto dos n pontos em três grupos, de tamanhos aproximadamente iguais, baseados principalmente na ordenação da variável x e no gráfico de dispersão. Chamemos esses grupos de E (de esquerda), C (de centro) e D (de direita). Se $n = 3k$, cada grupo terá k pontos. Se $n = 3k + 1$, colocamos k pontos nos grupos E e D e k +1 pontos no grupo C. Finalmente, se $n = 3k + 2$, colocamos $k + 1$ pontos nos grupos E e D e k pontos no grupo C.

Para cada grupo obtemos um *ponto resumo*, formado pela mediana dos x_i e a mediana dos y_i naquele grupo. Denominemos esses pontos por

$$(x_E, y_E), (x_C, y_C), (x_D, y_D).$$

Na Figura 16.18 temos um exemplo com três grupos com $k = 3$ em cada grupo.

Figura 16.18: Reta resistente com três grupos.