Capítulo 12 | Regressão linear múltipla e alguns modelos de regressão não-linear

12.2 Estimação dos coeficientes

Modelo de regressão linear múltipla

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i$$

ou

$$y_i = \hat{y}_i + e_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + e_i,$$

onde ϵ_i e e_i são o erro aleatório e o resíduo, respectivamente, associados com a resposta y_i e com valor ajustado $\hat{y}_i$.

Tabela 12.1 Dados para o Exemplo 12.1

Óxido Nitroso, y	Umidade, x_1	Temperatura, x_2	Pressão, x_3	Óxido Nitroso, y	Umidade, x_1	Temperatura, x_2	Pressão, x_3
0,90	72,4	76,3	29,18	1,07	23,2	76,8	29,38
0,91	41,6	70,3	29,35	0,94	47,4	86,6	29,35
0,96	34,3	77,1	29,24	1,10	31,5	76,9	29,63
0,89	35,1	68,0	29,27	1,10	10,6	86,3	29,56
1,00	10,7	79,0	29,78	1,10	11,2	86,0	29,48
1,10	12,9	67,4	29,39	0,91	73,3	76,3	29,40
1,15	8,3	66,8	29,69	0,87	75,4	77,9	29,28
1,03	20,1	76,9	29,48	0,78	96,6	78,7	29,29
0,77	72,2	77,7	29,09	0,82	107,4	86,8	29,03
1,07	24,0	67,7	29,60	0,95	54,9	70,9	29,37

Fonte: Charles T. Hare, "Light-Duty Diesel Emission Correction Factors for Ambient Conditions" (Fatores de correlação da emissão de diesel de cargas leves com as condições ambientais), EPA-600/2-77-116, Agência de Proteção Ambiental Norte-Americana.

12.3 Modelo de regressão linear usando matrizes (opcional)

Modelo linear geral

$$y = \mathbf{X}\beta + \epsilon,$$

onde

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix},$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Tabela 12.2 Dados para o Exemplo 12.3

y (% de sobrevivência)	x_1 (% de peso)	x_2 (% de peso)	x_3 (% de peso)
25,5	1,74	5,30	10,80
31,2	6,32	5,42	9,40
25,9	6,22	8,41	7,20
38,4	10,52	4,63	8,50
18,4	1,19	11,60	9,40
26,7	1,22	5,85	9,90
26,4	4,10	6,62	8,00
25,9	6,32	8,72	9,10
32,0	4,08	4,42	8,70
25,2	4,15	7,60	9,20
39,7	10,15	4,83	9,40
35,7	1,72	3,12	7,60
26,5	1,70	5,30	8,20

Tabela 12.3 Dados para o Exemplo 12.4

Tempo de esterilização, x_2 (min)	Temperatura, x_1 (°C) 75	100	125
15	14,05	10,55	7,55
	14,93	9,48	6,59
20	16,56	13,63	9,23
	15,85	11,75	8,78
25	22,41	18,55	15,93
	21,66	17,98	16,44

12.4 Propriedades dos estimadores de quadrados mínimos

Teorema 12.1

Para a equação de regressão linear

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

uma estimativa não viciada de σ^2 é dada pelo quadrado médio residual ou do erro

$$s^2 = \frac{SQE}{n - k - 1}, \quad \text{onde}$$

$$SQE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrados médios	F
Regressão	SQR	k	$QMR = \frac{SQR}{k}$	$f = \frac{QMR}{QME}$
Erro	SQE	$n - (k + 1)$	$QME = \frac{SQE}{n - (k+1)}$	
Total	SQT	$n - 1$		

12.5 Inferências na regressão linear múltipla

Intervalo de confiança para $\mu_{Y \mid x_{10}, x_{20}, \ldots, x_{k0}}$

Um intervalo de confiança de $100(1-\alpha)\%$ para a *resposta média* $\mu_{Y \mid x_{10}, x_{20}, \ldots, x_{k0}}$ é

$$\hat{y}_0 - t_{\alpha/2} s \sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} < \mu_{Y \mid x_{10}, x_{20}, \ldots, x_{k0}} < \hat{y}_0 + t_{\alpha/2} s \sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0},$$

onde $t_{\alpha/2}$ é um valor da distribuição t com $n-k-1$ graus de liberdade.

Intervalo de predição para y_0

Um intervalo de predição $100(1-\alpha)\%$ para uma *resposta única* y_0 é dado por

$$\hat{y}_0 - t_{\alpha/2}\, s \sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} < y_0 <$$

$$\hat{y}_0 + t_{\alpha/2}\, s \sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0},$$

onde $t_{\alpha/2}$ é o valor de uma distribuição t com $n-k-1$ graus de liberdade.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	399,45437	133,15146	30,98	<,0001
Error	9	38,67640	4,29738		
Corrected Total	12	438,13077			

Root MSE	2,07301	R-Square	0,9117
Dependent Mean	29,03846	Adj R-Sq	0,8823
Coeff Var	7,13885		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	39,15735	5,88706	6,65	<,0001
x1	1	1,01610	0,19090	5,32	0,0005
x2	1	-1,86165	0,26733	-6,96	<,0001
x3	1	-0,34326	0,61705	-0,56	0,5916

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	25,5000	27,3514	1,4152	24,1500	30,5528	21,6734	33,0294	-1,8514
2	31,2000	32,2623	0,7846	30,4875	34,0371	27,2482	37,2764	-1,0623
3	25,9000	27,3495	1,3588	24,2757	30,4234	21,7425	32,9566	-1,4495
4	38,4000	38,3096	1,2818	35,4099	41,2093	32,7960	43,8232	0,0904
5	18,4000	15,5447	1,5789	11,9730	19,1165	9,6499	21,4395	2,8553
6	26,7000	26,1081	1,0358	23,7649	28,4512	20,8658	31,3503	0,5919
7	26,4000	28,2532	0,8094	26,4222	30,0841	23,2189	33,2874	-1,8532
8	25,9000	26,2219	0,9732	24,0204	28,4233	21,0414	31,4023	-0,3219
9	32,0000	32,0882	0,7828	30,3175	33,8589	27,0755	37,1008	-0,0882
10	25,2000	26,0676	0,6919	24,5024	27,6329	21,1238	31,0114	-0,8676
11	39,7000	37,2524	1,3070	34,2957	40,2090	31,7086	42,7961	2,4476
12	35,7000	32,4879	1,4648	29,1743	35,8015	26,7459	38,2300	3,2121
13	26,5000	28,2032	0,9841	25,9771	30,4294	23,0122	33,3943	-1,7032

Figura 12.1 Impressão *SAS* para os dados do Exemplo 12.3.

Fonte de variação	Soma dos quadrados	g.l.
Regressão	$\sum_{i=1}^{n} \hat{y}_i^2 = \mathbf{y'X(X'X)^{-1}X'y}$	$k + 1$
Erro	$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ $= \mathbf{y'[I_n - X(X'X)^{-1}X']y}$	$n - (k + 1)$
Total	$\sum_{i=1}^{n} y_i^2 = \mathbf{y'y}$	n

Fonte de variação	Soma dos quadrados	g.l.
Regressão	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \mathbf{y}'[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}]\mathbf{y}$	k
Erro	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \mathbf{y}'[\mathbf{I_n} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}$	$n-(k+1)$
Total	$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \mathbf{y}'[\mathbf{I_n} - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}']\mathbf{y}$	$n-1$

12.6 Escolha de um modelo ajustado por meio de testes de hipóteses

R^2 ajustado

$$R^2_{\text{ajt}} = 1 - \frac{SQE/(n - k - 1)}{SQT/(n - 1)}.$$

12.7 Caso especial de ortogonalidade (opcional)

Tabela 12.4 Análise de variância para variáveis ortogonais

Fonte da variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	f calculado
β_1	$R(\beta_1) = b_1^2 \sum_{i=1}^{n} x_{1i}^2$	1	$R(\beta_1)$	$\frac{R(\beta_1)}{s^2}$
β_2	$R(\beta_2) = b_2^2 \sum_{i=1}^{n} x_{2i}^2$	1	$R(\beta_2)$	$\frac{R(\beta_2)}{s^2}$
⋮	⋮	⋮	⋮	⋮
β_k	$R(\beta_k) = b_k^2 \sum_{i=1}^{n} x_{ki}^2$	1	$R(\beta_k)$	$\frac{R(\beta_k)}{s^2}$
Erro	SQE	$n-k-1$	$s^2 = \frac{SQE}{n-k-1}$	
Total	$SQT = Syy$	$n-1$		

Tabela 12.5 Dados para o Exemplo 12.8

Raio da partícula	Temperatura do pó	Taxa de extrusão	Temperatura de evaporação
82	150 (–1)	12 (–1)	220 (–1)
93	190 (+1)	12 (–1)	220 (–1)
114	150 (–1)	24 (+1)	220 (–1)
124	150 (–1)	12 (–1)	250 (+1)
111	190 (+1)	24 (+1)	220 (–1)
129	190 (+1)	12 (–1)	250 (+1)
157	150 (–1)	24 (+1)	250 (+1)
164	190 (+1)	24 (+1)	250 (+1)

Tabela 12.6 Análise de variância para os dados do raio da partícula

Fonte da variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	*f* calculado	Valor *P*
β_1	$(2{,}5)^2\,(8) = 50$	1	50	2,16	0,2156
β_2	$(14{,}75)^2\,(8) = 1.740{,}50$	1	1.740,50	75,26	0,0010
β_k	$(21{,}75)^2\,(8) = 3.784{,}50$	1	3.784,50	163,65	0,0002
Erro	92,5	4	23,1250		
Total	5.667,50	7			

12.8 Variáveis categóricas ou indicadoras

Três categorias

A estimação dos coeficientes pelo método dos mínimos quadrados continua a ser aplicada. No caso de três níveis ou categorias de uma única variável indicadora, o modelo incluirá *dois* regressores, digamos z_1 e z_2, onde a atribuição (0, 1) é como se segue:

$$\begin{array}{cc} z_1 & z_2 \\ \left[\begin{array}{cc} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ \cdots & \cdots \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \\ \cdots & \cdots \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{array}\right] \end{array}$$

Em outras palavras, se há ℓ categorias, o modelo inclui $\ell - 1$ termos.

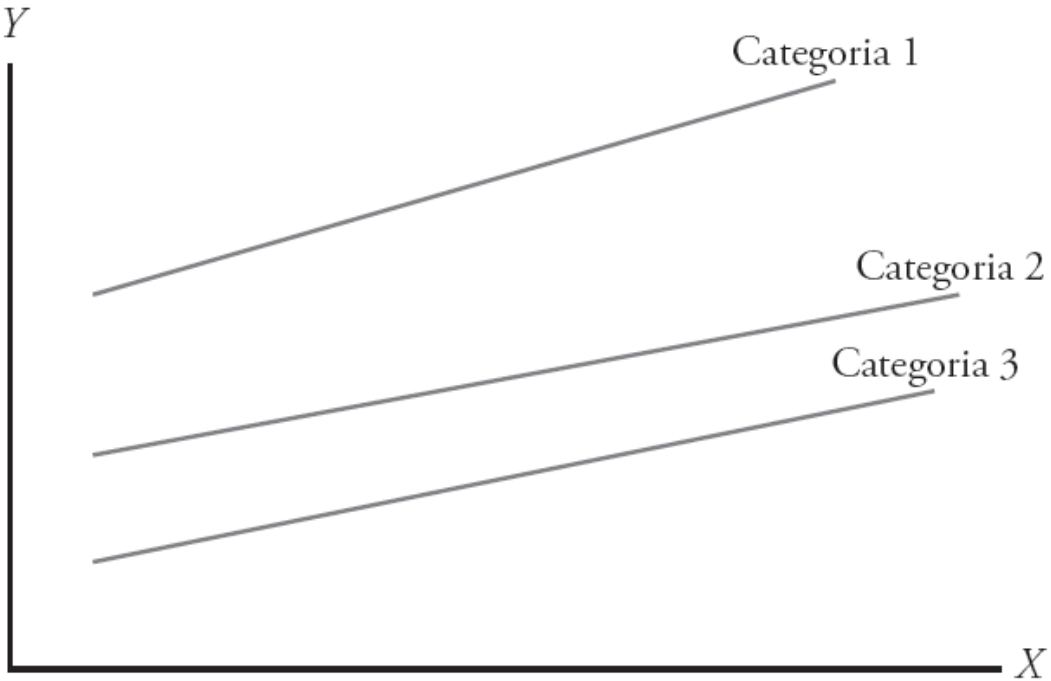

Figura 12.2 Caso de três categorias.

Tabela 12.7 Dados para o Exemplo 12.9

x (pH)	y (Quantidade de sólidos suspensos)	Polímero
6,5	292	1
6,9	329	1
7,8	352	1
8,4	378	1
8,8	392	1
9,2	410	1
6,7	198	2
6,9	227	2
7,5	277	2
7,9	297	2
8,7	364	2
9,2	375	2
6,5	167	3
7,0	225	3
7,2	247	3
7,6	268	3
8,7	288	3
9,2	342	3

$$y = \alpha + \beta_1 x + \beta_2 z_1 + \beta_3 z_2 + \varepsilon$$

$$z_1 = \begin{cases} 1\ se\ polimero\ 1 \\ 0\ caso\ contrario \end{cases}$$

$$z_2 = \begin{cases} 1\ se\ polimero\ 2 \\ 0\ caso\ contrario \end{cases}$$

Se polímero =1

$$y = \alpha + \beta_1 x + \beta_2 z_1 = (\alpha + \beta_2) + \beta_1 x + \varepsilon$$

Se polímero =2

$$y = \alpha + \beta_1 x + \beta_3 z_2 = (\alpha + \beta_3) + \beta_1 x + \varepsilon$$

Se polímero =3

$$y = \alpha + \beta_1 x = \alpha + \beta_1 x + \varepsilon$$

Com interações

$$y = \alpha + \beta_1 x + \beta_2 z_1 + \beta_3 z_2 + \beta_4 x z_1 + \beta_5 z_2 + \varepsilon$$

$$z_1 = \begin{cases} 1\ se\ polimero\ 1 \\ 0\ caso\ contrario \end{cases}$$

$$z_2 = \begin{cases} 1\ se\ polimero\ 2 \\ 0\ caso\ contrario \end{cases}$$

Se polímero =1

$$y = \alpha + \beta_1 x + \beta_2 z_1 + \beta_4 x z_1 = (\alpha + \beta_2) + (\beta_1 + \beta_4) x + \varepsilon$$

Se polímero =2

$$y = \alpha + \beta_1 x + \beta_3 z_2 + \beta_5 x z_2 = (\alpha + \beta_3) + (\beta_1 + \beta_5) x + \varepsilon$$

Se polímero =3

$$y = \alpha + \beta_1 x = \alpha + \beta_1 x + \varepsilon$$

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	80181,73127	26727,24376	73,68	<,0001
Error	14	5078,71318	362,76523		
Corrected Total	17	85260,44444			

R-Square	Coeff Var	Root MSE	y Mean
0,940433	6,316049	19,04640	301,5556

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-161,8973333	37,43315576	-4,32	0,0007
x	54,2940260	4,75541126	11,42	<,0001
z1	89,9980606	11,05228237	8,14	<,0001
z2	27,1656970	11,01042883	2,47	0,0271

Figura 12.3 Impressão *SAS* para o Exemplo 12.9.

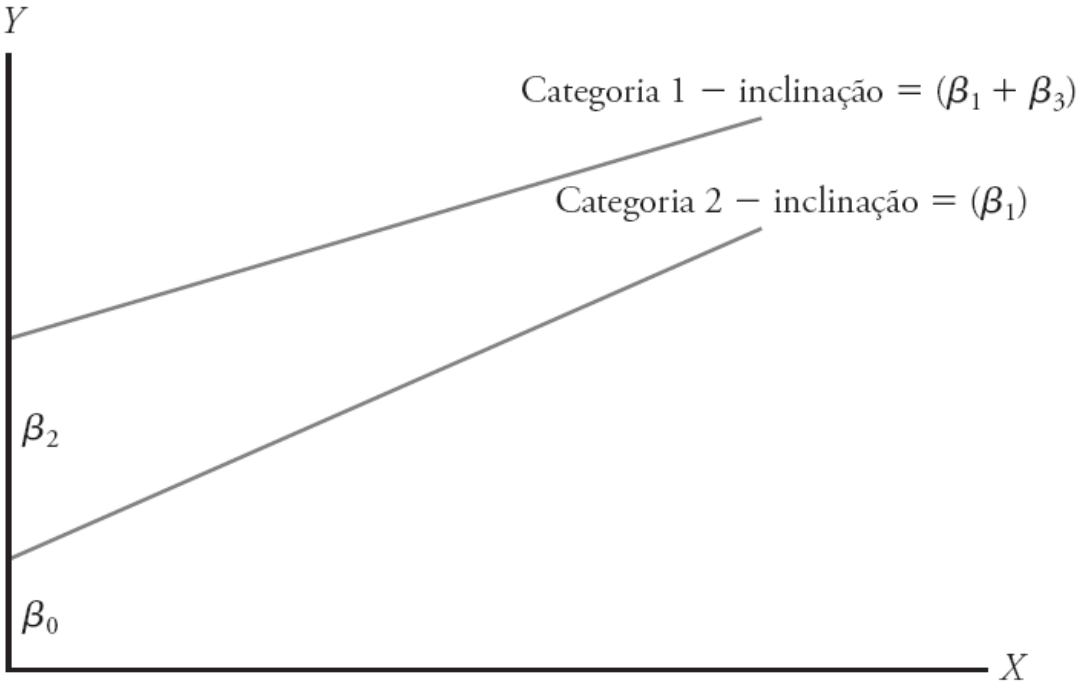

Figura 12.4 Não-paralelismo em variáveis categóricas.

12.9 Métodos seqüenciais para seleção de modelos

Tabela 12.8 Dados relacionados ao comprimento das crianças*

Comprimento da criança, y (cm)	Idade, x_1 (dias)	Comprimento ao nascimento, x_2 (cm)	Peso no nascimento, x_3 (kg)	Tamanho do peito no nascimento, x_4 (cm)
57,5	78	48,2	2,75	29,5
52,8	69	45,5	2,15	26,3
61,3	77	46,3	4,41	32,2
67,0	88	49,0	5,52	36,5
53,5	67	43,0	3,21	27,2
62,7	80	48,0	4,32	27,7
56,2	74	48,0	2,31	28,3
68,5	94	53,0	4,30	30,3
69,2	102	58,0	3,71	28,7

*Dados analisados pelo Centro de Consultoria Estatística, do Instituto Politécnico e Universidade Estadual da Virgínia, em Blacksburg, Virgínia.

Tabela 12.9 Valores t para os dados de regressão da Tabela 12.8

Variável x_1	Variável x_2	Variável x_3	Variável x_4
$R(\beta_1\|\beta_2, \beta_3, \beta_4)$	$R(\beta_2\|\beta_1, \beta_3, \beta_4)$	$R(\beta_3\|\beta_1, \beta_2, \beta_4)$	$R(\beta_4\|\beta_1, \beta_2, \beta_3)$
= 0,0644	= 0,6334	= 6,2523	= 0,0241
t = 0,2947	t = 0,9243	t = 2,9040	t = 0,1805

Tabela 12.10 Conjunto de dados para o Exemplo 12.12

Observação	Captura com rede quadrada, y	Captura com rede de varredura, x_1	Altura das plantas, x_2 (cm)
1	18,0000	4,15476	52,705
2	8,8750	2,02381	42,069
3	2,0000	0,15909	34,766
4	20,0000	2,32812	27,622
5	2,3750	0,25521	45,879
6	2,7500	0,57292	97,472
7	3,3333	0,70139	102,062
8	1,0000	0,13542	97,790
9	1,3333	0,12121	88,265
10	1,7500	0,10937	58,737
11	4,1250	0,56250	42,386
12	12,8750	2,45312	31,274
13	5,3750	0,45312	31,750
14	28,0000	6,68750	35,401
15	4,7500	0,86979	64,516
16	1,7500	0,14583	25,241
17	0,1333	0,01562	36,354