Modelos de Regressão

Clarice Garcia Borges Demétrio
Departamento de Ciências Exatas, ESALQ, USP
Caixa Postal 9
13418-900 Piracicaba, SP
Email: Clarice@carpa.ciagri.usp.br
Fax: 019 34294346

Sílvio Sandoval Zocchi
Departamento de Ciências Exatas, ESALQ, USP
Caixa Postal 9
13418-900 Piracicaba, SP
Email: sszocchi@carpa.ciagri.usp.br
Fax: 019 34294346

29 de março de 2011

Prefácio

Estas notas são resultantes de vários anos de lecionamento da disciplina LCE Regressão e Covariância,

Agradecimentos

Oa autores agradecem a todos que direta ou indiretamente contribuíram para a realização desse texto.

Sumário

Capítulo 1

Conceitos gerais

1.1 Natureza das variáveis

Um problema comum em Estatística é o estudo da relação entre duas variáveis X e Y, isto é, procura-se uma função de X que explique Y

$$X, Y \to Y \simeq f(X).$$

Em geral, a relação não é perfeita. Os pontos não se situam perfeitamente sobre a função que relaciona as duas variáveis. Mesmo se existe uma relação exata entre as variáveis como temperatura e pressão, flutuações em torno da curva aparecerão devido a erros de medidas.

Freqüentemente, o tipo de curva a ser ajustada é sugerido por evidência empírica ou por argumentos teóricos. O modelo a ser adotado depende de vários fatores, por exemplo, natureza das variáveis, relação linear ou não, homogeneidade de variâncias ou não, tipos de erros, independência dos erros etc.

A natureza das variáveis X e Y pode variar, isto é, elas podem ser fixas (ou controladas) ou aleatórias. Além disso, ambas podem ser medidas com ou sem erro (de mensuração). De forma esquemática, tem-se:

$$X \begin{cases} \text{fixa} & \begin{cases} \text{com erro} \\ \text{sem erro} \end{cases} \\ \text{aleatória} & \begin{cases} \text{com erro} \\ \text{sem erro} \end{cases} \end{cases}$$

$$Y \begin{cases} \text{fixa} & \begin{cases} \text{com erro} \\ \text{sem erro} \end{cases} \\ \text{aleatória} & \begin{cases} \text{com erro} \\ \text{sem erro} \end{cases} \end{cases}$$

o que sugere 16 combinações possíveis entre X e Y.

Assim, por exemplo, se

- X representa a variável sexo, ela é uma variável de classificação, fixa, medida sem erro, que pode assumir o valor 0, se feminino, ou 1 se masculino ou vice-versa;
- X representa um número (fixado) de frutos (2, 3, 4) por ramo em um determinado ano e Y, o número de gemas floríferas nos mesmos ramos no ano seguinte, tem-se que X é fixa, sem erro e Y é aleatória, sem erro de mensuração;
- X representa as quantidades 30, 60 e 90kg de nitrogênio/ha colocadas no solo, ela é fixa, possivelmente, medida com erro;
- X representa quantidades de nitrogênio no solo e Y quantidades de nitrogênio na planta, ambas são aleatórias, possivelmente, medidas com erro. Pode-se, porém, controlar X por meio da especificação de determinadas características do solo.

1.1.1 Relações entre tipos de variáveis e tipos de erros

(i) Considerando-se X fixa (ou controlada), tem-se:

$$X_{CE} = X_{CS} + e_X$$

sendo

X_{CE} : X controlada, medida com erro

X_{CS} : X controlada, medida sem erro

e_X : erro de medida em X.

Como exemplos, têm-se doses de pesticidas, de adubos etc.

(ii) Considerando-se Y fixa (ou controlada), tem-se

$$Y_{CE} = Y_{CS} + e_Y$$

sendo

Y_{CE} : Y controlada, medida com erro

Y_{CS} : Y controlada, medida sem erro

e_Y : erro de medida em Y.

(iii) Considerando-se que X é uma variável aleatória com distribuição de média μ_X, tem-se:

$$X_{AS} = \mu_X + \varepsilon_X$$

e

$$X_{AE} = \mu_X + \varepsilon_X + e_X = X_{AS} + e_X$$

sendo

X_{AE} : X aleatória, medida com erro

X_{AS} : X aleatória, medida sem erro

ε_X é erro aleatório

e_X é erro de mensuração.

Como exemplos, têm-se quantidades de nutrientes encontradas no solo.

(iv) Considerando-se que Y é uma variável aleatória com distribuição de média μ_Y, tem-se:

$$Y_{AS} = \mu_Y + \varepsilon_Y$$

e

$$Y_{AE} = \mu_Y + \varepsilon_Y + e_Y = Y_{AS} + e_Y$$

sendo

Y_{AE} : Y aleatória, medida com erro

Y_{AS} : Y aleatória, medida sem erro

ε_Y é erro aleatório

e_Y é erro de mensuração.

Como exemplos, têm-se quantidades de nutrientes encontradas na planta, medidas de comprimento, peso, volume etc.

Na maior parte dos casos, tanto X como Y são medidas com erros e o que se procura fazer é tornar esses erros desprezíveis. Apenas como exemplos, sejam alguns casos das 16 combinações possíveis entre X e Y.

Caso 1: Y_{CS} vs X_{CS} (Y controlado sem erro versus X controlado sem erro).

Esse é um problema matemático (modelo determinístico) em que $Y = f(X)$. Como exemplo, tem-se a lei física:

$$E = rJ$$

sendo E, tensão, J, intensidade da corrente e r, resistência.

Se, porém, forem observados n pares de valores *E, J*, as medidas observadas dependerão da precisão dos equipamentos, estando, portanto, sujeitas a erros, e pode-se estimar r por meio de uma equação de regressão que passa pela origem.

Caso 2: Y_{CE} vs X_{CS} (Y controlada com erro versus X controlada sem erro).

Nesse caso, a variável Y está afetada por apenas um tipo de erro, isto é,

$$Y_{CE} = f(X_{CS}) + e_Y.$$

Em geral, considera-se que $E(e_Y) = 0$, e portanto,

$$E(Y_{CE}) = f(X_{CS}).$$

Caso 3: Y_{AS} vs X_{CS} (Y aleatória sem erro versus X controlada sem erro).

Nesse caso, também, a variável Y está afetada por apenas um tipo de erro, isto é,

$$Y_{AS} = f(X_{CS}) + \varepsilon_Y = \mu_Y + \varepsilon_Y.$$

Caso 4: Y_{AE} vs X_{CS} (Y aleatória com erro versus X controlada sem erro).

Nesse caso, a variável Y está afetada por dois tipos de erros, isto é,

$$Y_{AE} = f(X_{CS}) + \varepsilon_Y + e_Y = \mu_Y + \varepsilon_Y + e_Y$$

se a função $f(.)$ for conhecida. Se $f(.)$ não é conhecida, ou quando Y é afetada por k variáveis, isto é,

$$Y = g(X, X_1, X_2, \cdots, X_k) + \varepsilon_Y + e_Y$$

sendo $g(X, X_1, X_2, \cdots, X_k) = f(X) + h(X_1, X_2, \cdots, X_k)$, pode-se ter

$$Y = f(X_{CS}) + \xi_Y + \varepsilon_Y + e_Y = \mu_Y + \xi_Y + \varepsilon_Y + e_Y$$

em que ξ_Y é o erro devido à não consideração de todas as variáveis que afetam Y, isto é, tem-se, também, um erro de especificação do modelo.

1.1.2 Funções lineares e não lineares (especificação da função $f(.)$)

Nos estudos de regressão busca-se relacionar uma variável aleatória Y com uma ou mais variáveis X's, especificando-se a função $f(.)$. Quando Y depende apenas de uma variável X, isto é,

$$Y = f(X, \beta_0, \beta_1, \cdots, \beta_k) + \varepsilon_Y$$

tem-se que $f(.)$ é linear nos parâmetros $\beta_0, \beta_1, \cdots, \beta_k$ se

$$\frac{\partial f}{\partial \beta_i} = h(X), i = 0, 1, \cdots, k,$$

sendo $h(X)$ dependente apenas de X.

Outro caso comum é considerar

$$Y = f(X_1, X_2, \cdots, X_k, \beta_0, \beta_1, \cdots, \beta_k) + \varepsilon_Y$$

que é linear nos parâmetros se

$$\frac{\partial f}{\partial \beta_i} = h(X_1, X_2, \cdots, X_k),$$

isto é, h(.) depende apenas de $X_1, X_2, \cdots, X_k$. Se pelo menos uma das derivadas parciais $\frac{\partial f}{\partial \beta_i}$ depende de pelo menos um dos parâmetros, então, $f(.)$ é uma função não linear dos parâmetros.

Como exemplos de funções lineares, têm-se:
(i) $f(X, \beta_0) = \beta_0$, pois, $\frac{\partial f}{\partial \beta_0} = 1$,

(ii) $f(X, \beta_0, \beta_1) = \beta_0 + \beta_1 X$, pois, $\frac{\partial f}{\partial \beta_0} = 1$ e $\frac{\partial f}{\partial \beta_1} = X$,

(iii) $f(X, \beta_0, \beta_1) = \beta_0 + \beta_1 \frac{1}{X}$, pois, $\frac{\partial f}{\partial \beta_0} = 1$ e $\frac{\partial f}{\partial \beta_1} = \frac{1}{X}$,

(iv) $f(X_1, X_2, X_3, \beta_0, \beta_1, \beta_2, \beta_3) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$,
pois, $\frac{\partial f}{\partial \beta_0} = 1$, $\frac{\partial f}{\partial \beta_1} = X_1$, $\frac{\partial f}{\partial \beta_2} = X_2$ e $\frac{\partial f}{\partial \beta_3} = X_3$,

(v) $f(X, \beta_0, \beta_1, \beta_2, \beta_3) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$,
pois, $\frac{\partial f}{\partial \beta_0} = 1$, $\frac{\partial f}{\partial \beta_1} = X$, $\frac{\partial f}{\partial \beta_2} = X^2$ e $\frac{\partial f}{\partial \beta_3} = X^3$

(vi) $f(X, \beta_0, \beta_1) = \beta_0 + \beta_1 \log(X)$, pois, $\frac{\partial f}{\partial \beta_0} = 1$ e $\frac{\partial f}{\partial \beta_1} = \log(X)$.

Como exemplos de funções não lineares, podem ser citadas:
(i) $f(X, \beta_0, \beta_1, \beta_2) = \beta_0 \text{sen}(\beta_1 X + \beta_2)$,
pois, $\frac{\partial f}{\partial \beta_0} = \text{sen}(\beta_1 X + \beta_2)$, $\frac{\partial f}{\partial \beta_1} = \beta_0 X \cos(\beta_1 X + \beta_2)$ e $\frac{\partial f}{\partial \beta_2} = \beta_0 \cos(\beta_1 X + \beta_2)$,

(ii) $f(X, \beta_0, \beta_1, \beta_2) = \beta_0 + \beta_1 e^{\beta_2 X}$,
pois, $\frac{\partial f}{\partial \beta_0} = 1$, $\frac{\partial f}{\partial \beta_1} = e^{\beta_2 X}$ e $\frac{\partial f}{\partial \beta_2} = \beta_1 X e^{\beta_2 X}$

(iii) $f(X, \beta_0, \beta_1, \beta_2) = \frac{\beta_0 + \beta_1 X}{1 + \beta_2 X}$,
pois, $\frac{\partial f}{\partial \beta_0} = \frac{1}{1 + \beta_2 X}$, $\frac{\partial f}{\partial \beta_1} = \frac{X}{1 + \beta_2 X}$ e $\frac{\partial f}{\partial \beta_2} = -\frac{(\beta_0 + \beta_1 X) X}{(1 + \beta_2 X)^2}$.

1.1.3 Tipos de modelos

Em função da natureza das variáveis X e Y, diferentes tipos de modelos podem ser considerados. Se X e Y são fixos, tem-se um **modelo determinístico**. Se Y é aleatório, três tipos de modelos podem ser considerados

- **Modelo tipo I**, em que os X's são fixos.
- **Modelo tipo II**, em que os X's são aleatórios.
- **Modelo Misto**, em que parte dos X's são fixos e parte, aleatórios.

Observação: Será considerado, aqui, apenas o caso em que os Y são aleatórios.

Para o **Modelo tipo I**, os valores da variável X são selecionados pelo pesquisador, não havendo variação aleatória associada a eles. A seleção dos X's pode envolver um conjunto específico de valores ou valores que estão simplesmente dentro de uma amplitude de variação. Assim, por exemplo, a resposta a um inseticida pode ser medida para uma série específica de doses, enquanto que peso do corpo humano pode ser obtido para uma amplitude de alturas restritas por uma descrição (faixa etárea, raça etc). Quando valores esperados estão sendo considerados, os mesmos X's são usados ao definir uma amostragem repetida que é a sua base. Estes X's devem ser medidos sem erro.

Valores da variável X, por exemplo, horas de luz artificial, níveis de temperatura, quantias de produtos e espaçamentos entre plantios podem ser igual ou convenientemente espaçados para o aumento da eficiência do tratamento.

Medida de Y sem erro não é um requisito teórico, desde que o erro de medida tenha uma distribuição com média conhecida, geralmente, considerada igual a zero. A variância de Y é, então, a soma de uma variância biológica (ou outra) em Y e a variância de erro de medida. É importante, naturalmente, manter os erros de medidas em um mínimo.

Suponha que o **Modelo tipo I** seja apropriado e que o problema seja especificado de uma das formas que se segue.

1. Assume-se que existe uma relação funcional ou matemática entre Y e X mas que são possíveis erros observacionais em Y. O problema é estimar essa relação. Se os X's são medidos sem erros (na realidade, X possui erros pequenos, porém, para estudos teóricos considera-se que não os tem) como na Figura 1.1, então, há uma única linha de regressão dada por $\mathrm{E}(Y \mid X) = \mathrm{E}(Y) = \alpha + \beta X$.

2. Se os X's são, também, medidos com erro, então, deve-se visualizar uma distribuição bivariada para cada ponto da reta verdadeira (Figura 1.2). Para estimar a relação funcional devem ser adotados procedimentos específicos (modelo funcional dentro do estudo de *Modelos de regressão com erros de medidas*).

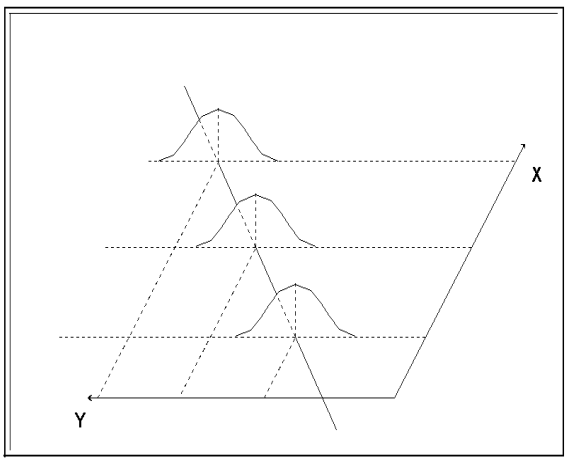


Figura 1.1: Erros de medida em Y

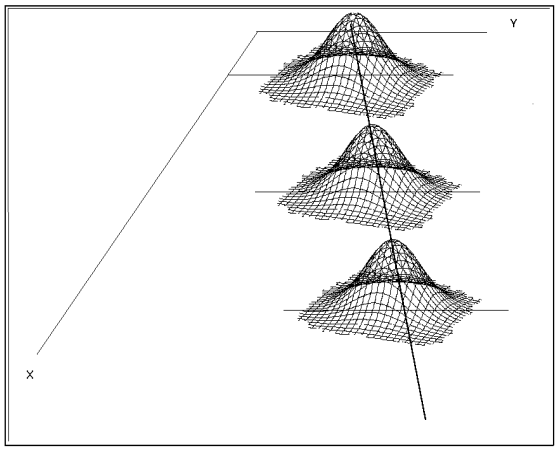


Figura 1.2: Erros de medida em X e Y

3. Existe uma relação estatística ou associação entre X e Y. Inicialmente, uma distribuição bivariada sobre o plano X, Y é apropriada. Entretanto, X é restrita em lugar de aleatória como na Figura 1.3. Consequentemente, só há uma regressão significativa a ser estimada, aquela de Y em relação a X. Erros de medidas em X ou Y são provavelmente desprezíveis em relação à amplitude escolhida dos X's ou à variação aleatória dos Y's.

Para o **Modelo tipo II**, ambos X e Y são aleatórios. Este é o caso clássico de regressão bivariada, assumindo-se normalidade (Figura 1.4). Nesse caso a amostragem aleatória é de indivíduos, em que são feitos pares de medidas. A escolha de qual variável é dependente é determinada pelo problema. As duas linhas de regressão são possíveis, isto é, $Y|X$ e $X|Y$. Se X e Y são variáveis aleatórias com erros de medidas tem-se o modelo estrutural da teoria de *Modelos de regressão com erros de medidas.*

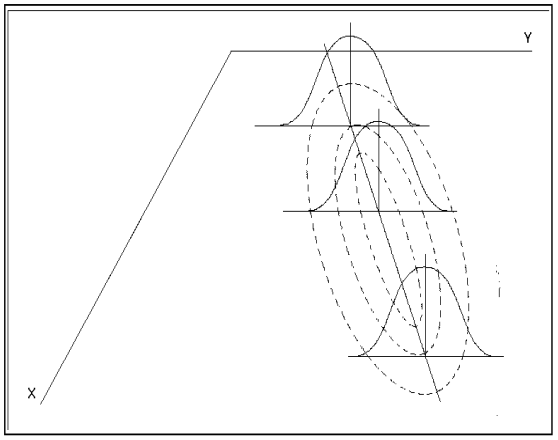


Figura 1.3: Restrições em X

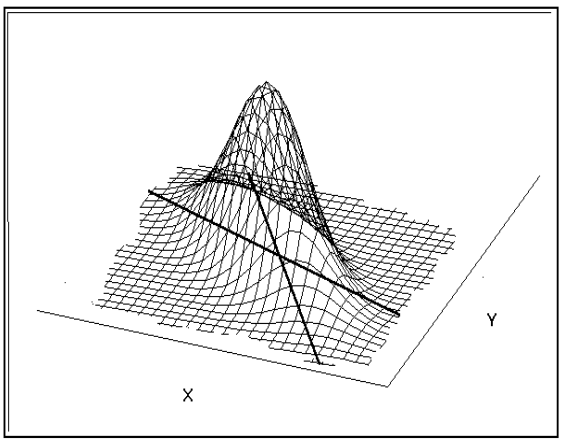


Figura 1.4: Superfície normal bivariada

1.2 Diagramas de dispersão

Antes de se iniciar qualquer análise de regressão de um conjunto de dados, é importante que se plotem os pares de dados em diagramas de dispersão, para que se tenha idéia a respeito do tipo de relação existente entre as variáveis, da variabilidade associada a elas e da presença de pontos atípicos. Entretanto, esses gráficos devem ser olhados com cuidado quando existem duas ou mais variáveis explanatórias, pois eles não levam em consideração a correlação existente entre elas. Assim, por exemplo, a Figura 1.5 mostra que existe uma relação linear entre as variáveis Y e X, existem dois pontos discrepantes e uma aparente heterogeneidade de variâncias.

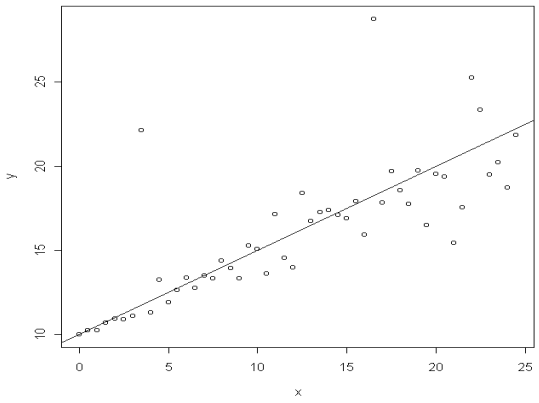


Figura 1.5: Gráfico de dispersão

1.3 Exemplos

1. Os dados da Tabela 1.1 (Snedecor e Cochran, 1967) referem-se a um experimento, em que 9 amostras de solos foram preparadas, variando-se os níveis de fósforo orgânico (X). Nessas amostras foi plantado milho e, após 38 dias, as plantas foram colhidas e o conteúdo de fósforo foi determinado. Em seguida, determinou-se, por uma expressão, o fósforo disponível (Y) para a planta no solo.

Tabela 1.1: Valores de fósforo orgânico X e de fósforo disponível (Y)

X (ppm)	1	4	5	9	13	11	23	23	28
Y (ppm)	64	71	54	81	93	76	77	95	109

Nesse caso, a variável X é fixa. A Figura 1.6 mostra que existe uma relação linear entre as variáveis Y e X. O número de observações é relativamente pequeno para que se possam fazer considerações sobre pontos discrepantes e variabilidade.

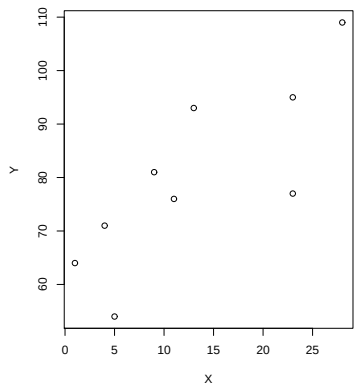


Figura 1.6: Gráficos de dispersão de Y em relaçao a X, Tabela 1.1.

2. Os dados da Tabela 1.2 (Duarte, 1989) referem-se a um experimento de irrigação em batata plantada em terra roxa estruturada (solo argiloso) em que foram medidas as lâminas (L, mm) de água a diferentes distâncias do aspersor e as correspondentes produtividades (P, t/ha). Em geral, para esse tipo de solo, o excesso de água causa diminuição de produtividade.

Tabela 1.2: Valores de lâminas (L, mm) de água a diferentes distâncias do aspersor e as correspondentes produtividades (P, t/ha)

L	285	380	400	425	455	490	520	550	575	615	680	785
P	14,94	15,98	21,21	22,71	22,38	24,83	24,42	30,59	29,96	31,07	29,80	22,61

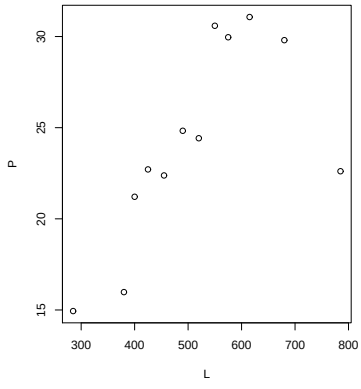


Figura 1.7: Gráficos de dispersão de P em relaçao a L, Tabela 1.2.

Nesse caso, a variável X é aleatória, mas pode ser considerada *controlada* se for de interesse

do pesquisador. A Figura 1.7 mostra que existe uma relação linear entre as variáveis P e L, e, embora o número de observações seja pequeno, parece que existe um ponto discrepante ou que a relação não é linear.

3. Paes de Camargo *et al* (1982), estudando a construção de um tensiômetro de leitura direta, obtiveram os resultados que aparecem na Tabela 1.3 para valores de alturas da câmara no tensiômetro (X), em mm, e tensão da água no solo (Y), em mb. Ver Pereira & Arruda (1987).

Tabela 1.3: Valores de alturas da câmara no tensiômetro (X), em mm, e tensão da água no solo (Y), em mb

X	9	12	30	42	57	102	147	210	290
Y	217	291	439	515	603	681	716	746	755

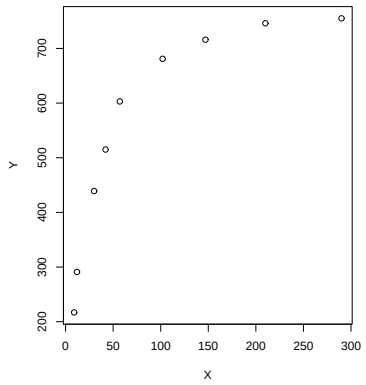


Figura 1.8: Gráficos de dispersão de Y em relaçao a X, Tabela 1.3.

Nesse caso, a variável X é fixa. A Figura 1.8 mostra que existe uma relação não linear entre as variáveis Y e X e nenhum ponto discrepante.

4. Os dados da Tabela 1.4 (Snedecor e Cochran, 1967) referem-se a medidas de concentrações de fósforo inorgânico (X_1) e fósforo orgânico (X_2) no solo e de conteúdo de fósforo (Y) nas plantas crescidas naquele solo. O objetivo desse tipo de experimento é estudar a relação existente entre o conteúdo de fósforo na planta e duas fontes de fósforo no solo.

   Nesse caso, as variáveis X_1 e X_2 são aleatórias, mas podem ser consideradas *controladas* se for de interesse do pesquisador. A Figura 1.9 mostra os gráficos de dispersão para as variáveis duas a duas. Pode-se ver que, aparentemente não existe relação linear entre as

Tabela 1.4: Valores de concentrações de fósforo inorgânico (X_1) e fósforo orgânico (X_2) no solo e de conteúdo de fósforo (Y)

Amostra	X_1	X_2	Y	Amostra	X_1	X_2	Y
1	0,4	53	64	10	12,6	58	51
2	0,4	23	60	11	10,9	37	76
3	3,1	19	71	12	23,1	46	96
4	0,6	34	61	13	23,1	50	77
5	4,7	24	54	14	21,6	44	93
6	1,7	65	77	15	23,1	56	95
7	9,4	44	81	16	1,9	36	54
8	10,1	31	93	17	26,8	58	168
9	11,6	29	93	18	29,9	51	99

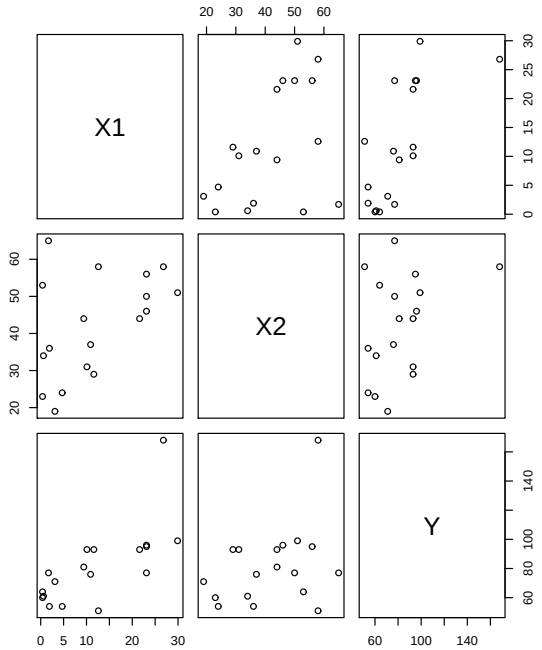


Figura 1.9: Gráficos de dispersão para as variáveis duas a duas, Tabela 1.4.

variáveis Y e X_1 e Y e X_2 e, em ambos os casos, aparece um ponto discrepante. Já entre X_1 e X_2, existe uma relação linear com uma aparente heterogeneidade de variâncias.

5. Os dados da Tabela 1.5 (Zambrosi e Alleoni, 2002) referem-se a resultados de um experimento casualizado em blocos, planejado para estudar o efeito da calagem sobre a CTC

do solo medida por dois métodos diferentes.

Tabela 1.5: Valores de CTC direta e indireta, em $mmol_c/kg$, na profundidade de 5 a 10 cm, 18 meses após a calagem incorporada ao solo, segundo a dose de calcário, em t/ha

	bloco 1		bloco 2		bloco 3		bloco 4	
Dose	direta	indireta	direta	indireta	direta	indireta	direta	indireta
0,00	38,80	83,00	38,80	90,70	45,60	85,80	50,20	85,50
2,00	59,20	87,60	53,00	84,60	57,20	97,50	62,80	80,80
4,90	60,60	106,60	73,30	111,40	79,30	102,40	77,90	112,40
7,80	68,80	177,00	90,70	112,20	84,50	125,60	73,80	106,40

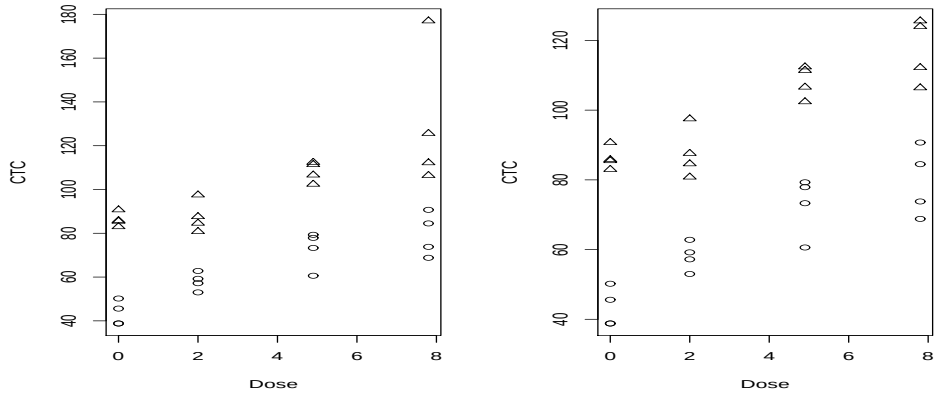


Figura 1.10: Gráficos de dispersão de CTC direta e indireta em relação à dose, com a observação $177,00$ e corrigida, respectivamente, Tabela 1.5.

Nesse caso, a variávek X é fixa. A Figura 1.10 mostra que existe uma relação linear entre as medidas de CTC e as doses de calcário, em t/ha, para ambos os métodos e que, aparentemente, há um paralelismo entre as retas a serem ajustadas. Nessa análise inicial foi detectada a presença de um dados discrepante $(177,00)$ correspondente ao bloco 1, dose $7,80$ e CTC indireta. Em conversa com o pesquisador responsável foi verificado que se tratava de um erro grosseiro de transcrição de dados e que o valor correto era $(124,00)$.

6. Os dados da Tabela 5.1 (Steel e Torrie, 1980) referem-se a um estudo sobre a resposta da cultura do milho como função da quantidade de fosfato, porcentagem de saturação de bases (X_2) e sílica (X_3) em solos ácidos. A resposta (Y), em porcentagem, foi medida como a diferença entre as produções (em lb/acre) nas parcelas recebendo fosfato e aquelas não recebendo fosfato (X_1), dividida pelas produções das parcelas que não receberam fosfato,

e multiplicadas por 100. Considerando-se esses dados, foi obtida a variável produtividade Y_1 das parcelas recebendo fosfato, dada por $Y_1 = X_1(1 + \frac{Y}{100})$.

Tabela 1.6: Dados de resposta da cultura do milho (Y) ao fosfato, em porcentagem, produtividade na testemunha (X_1), em lb/acre, porcentagem de saturação de bases (X_2) e pH do solo (X_3)

Y	X_1	X_2	X_3	Y	X_1	X_2	X_3
88	844	67	5,75	18	1262	74	6,10
80	1678	57	6,05	18	4624	69	6,05
42	1573	39	5,45	4	5249	76	6,15
37	3025	54	5,70	2	4258	80	5,55
37	653	46	5,55	2	2943	79	6,40
20	1991	62	5,00	-2	5092	82	6,55
20	2187	69	6,40	-7	4496	85	6,50

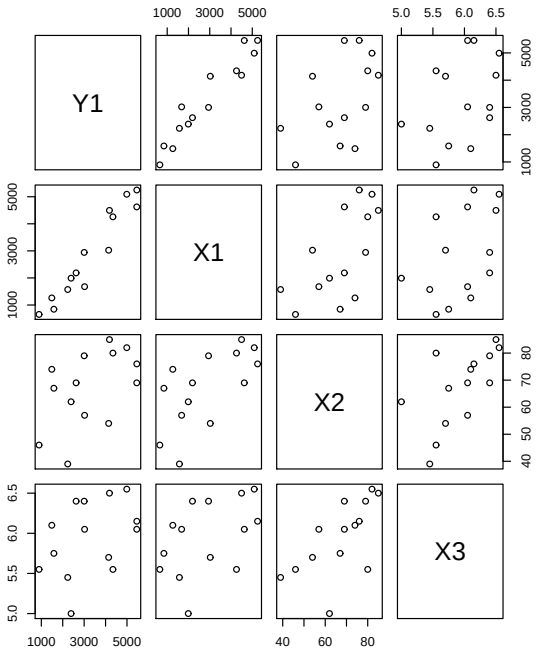


Figura 1.11: Gráficos de dispersão para as variáveis duas a duas, Tabela 5.1.

Nesse caso, as variáveis X_1, X_2 e X_3 são aleatórias, e o interesse do pesquisador está, principalmente no estudo de correlações entre as variáveis.. Na Figura 5.1 podem ser vistos os gráficos de dispersão para as variáveis duas a duas. Observa-se que existe uma correlação linear grande e positiva entre as variáveis X_1 e X_2.

1.4 Exercícios

1.4.1 Para cada um dos conjuntos de dados apresentados a seguir, discuta a natureza das variáveis, faça os possíveis diagramas de dispersão e discuta a relação entre as variáveis, tendência, dispersão e pontos atípicos.

1. Os dados que se seguem (Snedecor e Cochran, 1967) referem-se a medidas de alturas de feijão (Y), durante 7 semanas (amostras aleatórias independentes)

Idade em semanas (X)	1	2	3	4	5	6	7
Alturas em cm (Y)	5	13	16	23	33	38	40

2. Os dados que se seguem (Steel e Torrie, 1980) referem-se a peso médio (X) de 50 galinhas e consumo de alimentos (Y), para 10 linhagens White Leghorn.

Amostra	1	2	3	4	5	6	7	8	9	10
X	4,6	5,1	4,8	4,4	5,9	4,7	5,1	5,2	4,9	5,1
Y	87,1	93,1	89,8	91,4	99,5	92,1	95,5	99,3	93,4	94,4

3. Os dados que se seguem (Mead e Curnow, 1980) referem-se a concentrações de CO_2 (X) aplicadas sobre folhas de trigo a uma temperatura de 35^0C e a quantias de CO_2 ($Y, cm^3/dm^2$/hora) absorvido pelas folhas.

Amostra	1	2	3	4	5	6	7	8	9	10	11
X	75	100	100	120	130	130	160	190	200	240	250
Y	0,00	0,65	0,50	1,00	0,95	1,30	1,80	2,80	2,50	4,30	4,50

4. Os dados que se seguem (Ryan, Joiner e Ryan Jr., 1976) referem-se a medidas de diâmetro a 4,5 pés acima do solo (D, polegadas) e altura (H, pés) de 31 cerejeiras ("black cherry") em pé e de volume (V, pés cúbicos) de árvores derrubadas. O objetivo desse tipo de experimento é verificar de que forma essas variáveis estão relacionadas para, por meio de medidas nas árvores em pé, poder predizer o volume de madeira em uma área de floresta (*Allegheny National Forest*).

Amostra	X_1	X_2	Y	Amostra	X_1	X_2	Y
1	8,3	70	10,3	17	12,9	85	33,8
2	8,6	65	10,3	18	13,3	86	27,4
3	8,8	63	10,2	19	13,7	71	25,7
4	10,5	72	16,4	20	13,8	64	24,9
5	10,7	81	18,8	21	14,0	78	34,5
6	10,8	83	19,7	22	14,2	80	31,7
7	11,0	66	15,6	23	14,5	74	36,3
8	11,0	75	18,2	24	16,0	72	38,3
9	11,1	80	22,6	25	16,3	77	42,6
10	11,2	75	19,9	26	17,3	81	55,4
11	11,3	79	24,2	27	17,5	82	55,7
12	11,4	76	21,0	28	17,9	80	58,3
13	11,4	76	21,4	29	18,0	80	51,5
14	11,7	69	21,3	30	18,0	80	51,0
15	12,0	75	19,1	31	20,6	87	77,0
16	12,9	74	22,2				

5. Os dados que se seguem referem-se a números de ovos postos por 14 galinhas e números de folículos ovulados.

nº. de ovos	39	29	46	28	31	25	49	57	51	21	42	38	34	47
nº. de folículos	37	34	52	26	32	25	55	65	44	25	45	26	29	30

1.4.2 O manejo de irrigação é uma preocupação constante para aqueles que fazem uso dela, pois é anti-econômico irrigar a uma velocidade superior àquela da infiltração (a água irá escorrer e não infiltrar). Em função disso, são conduzidos ensaios que têm como finalidade estimar as equações de infiltração acumulada em relação ao tempo acumulado e de velocidade de infiltração em relação ao tempo acumulado e à velocidade básica de infiltração para um solo. Essas equações são importantes para a determinação do tempo de irrigação para atingir uma determinada lâmina de água, no caso de irrigação superficial e para a escolha do tipo de aspersor que deve ter intensidade de aplicação menor do que a velocidade de infiltração básica.

Os dados que se seguem referem-se a tempos acumulados (T, minutos) de observação e correspondentes medidas de infiltração acumulada (I, cm) da água no solo, usando o método do infiltômetro de anel.

T	I	T	I	T	I
1	0,8	16	3,9	96	13,8
2	1,3	26	4,7	126	16,9
4	1,8	36	6,9	156	20,0
6	2,1	51	8,6	186	23,5
11	3,1	66	10,1	216	26,4

Baseando-se nos dados apresentados,

a) calcule a velocidade de infiltração V (cm/min), dada por $V = 1/T$;

b) discuta a natureza das variáveis: tempo acumulado, infiltração acumulada e velocidade de infiltração;

c) faça diagramas de dispersão para infiltração acumulada versus tempo acumulado, velocidade de infiltração versus tempo acumulado e discuta a relação entre as variáveis, tendência, dispersão e pontos atípicos;

d) calcule a velocidade de infiltração básica aproximada (média dos últimos cinco valores)

Observação Em geral, na literatura (Bernardo, S. 1989, Manual de Irrigação), são propostos os modelos não lineares para estimar as equações de infiltração acumulada em relação a tempo acumulado e de velocidade de infiltração em relação a tempo acumulado:

$$I = aT^b + cT \quad \text{e} \quad V = dT^{b-1} + c$$

ou

$$I = aT^b \quad \text{e} \quad V = dT^{b-1}$$

em que a, b, c e d são parâmetros a serem estimados e c refere-se à velocidade de infiltração básica.

1.4.3 Mostre quais funções das que se seguem são lineares nos parâmetros e quais são não lineares.

a) $f(X, \beta_0, \beta_1) = \beta_0 + \beta_1 X^{-2}$

b) $f(X, \beta_0, \beta_1) = \beta_0 + \beta_1 X^3$

c) $f(X, \beta_0, \beta_1) = \dfrac{\beta_0}{\beta_0 + \beta_1 X}$

d) $f(X, \beta_0, \beta_1, \beta_2) = \beta_2 \exp\{-\exp(\beta_0 + \beta_1 X)\}$

e) $f(X, \beta_0, \beta_1, \beta_2) = \beta_0 + \beta_1 \beta_2^X$

f) $f(X, \beta_0, \beta_1, \beta_2) = \beta_0 + \beta_1 X I_{\{X \leq 0\}} + \beta_2 X I_{\{X > 0\}}$

g) $f(X_1, X_2, \beta_1, \beta_2) = \beta_1 X_1 + \beta_2 X_2$

h) $f(X_1, X_2, \beta_0, \beta_1, \beta_2, \beta_{12}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2$

i) $f(X_1, X_2, \beta_0, \beta_1, \beta_2) = \dfrac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}$

Capítulo 2

Regressão linear simples

2.1 Introdução

A teoria de *Regressão* teve origem no século XIX com Galton. Em um de seus trabalhos ele estudou a relação entre a altura dos pais e dos filhos (X_i e Y_i), procurando saber como a altura do pai influenciava a altura do filho. Notou que se o pai fosse muito alto ou muito baixo, o filho teria uma altura tendendo à média. Por isso, ele chamou de regressão, ou seja, existe uma tendência de os dados regredirem à média.

A utilização de modelos de regressão, pode ter por objetivos:

i) **Predição**. Uma vez que se espera que uma parte (que se deseja que seja a maior) da variação de Y é explicada pelas variáveis X, então, pode-se utilizar o modelo para obter valores de Y correspondentes a valores de X que não estavam entre os dados. Esse processo denomina-se **predição** e, em geral, são usados valores de X que estão dentro do intervalo de variação estudado. A utilização de valores fora desse intervalo recebe o nome de **extrapolação** e, deve ser usada com muito cuidado, pois o modelo adotado pode não ser correto fora do intervalo estudado. Este, talvez, seja o uso mais comum dos modelos de regressão.

ii) **Seleção de variáveis**. Freqüentemente, não se tem idéia de quais são as variáveis que afetam significativamente a variação de Y. Para responder a esse tipo de questão, conduzem-se estudos para os quais um grande número de variáveis está presente. A análise de regressão pode auxiliar no processo de seleção de variáveis, eliminando aquelas cuja contribuição não seja importante.

iii) **Estimação de parâmetros**. Dado um modelo e um conjunto de dados (amostra) referente às variáveis resposta e preditoras, estimar parâmetros, ou ainda, ajustar o modelo aos dados, significa obter valores (estimativas) para os parâmetros, por algum processo, tendo por base o modelo e os dados observados. Em alguns casos, o valor do coeficiente tem valor por si só. Como exemplo, pode-se citar o estudo de estabilidade de variedades.

Em outros casos, o interesse está em uma função dos parâmetros. Como exemplo, pode-se citar o cálculo de doses letais.

iv) **Inferência**. O ajuste de um modelo de regressão tem, em geral, por objetivos básicos, além de estimar os parâmetros, realizar inferências sobre eles, tais como testes de hipóteses e intervalos de confiança.

Em geral, as variáveis X's são chamadas **variáveis independentes** ou **explanatórias** ou **"carriers"**, enquanto que a variável Y é chamada **variável dependente** ou **resposta**.

2.2 Modelo estatístico

Suponha que a relação verdadeira entre X e Y é uma linha reta, e que cada observação Y, em cada nível de X, é uma variável aleatória (Figura 2.1).

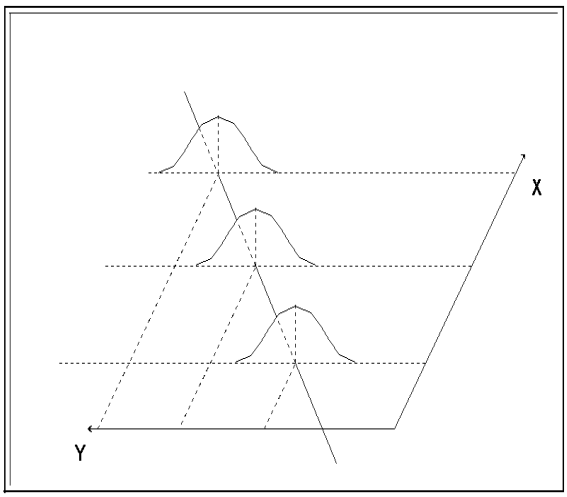


Figura 2.1: Erros em Y

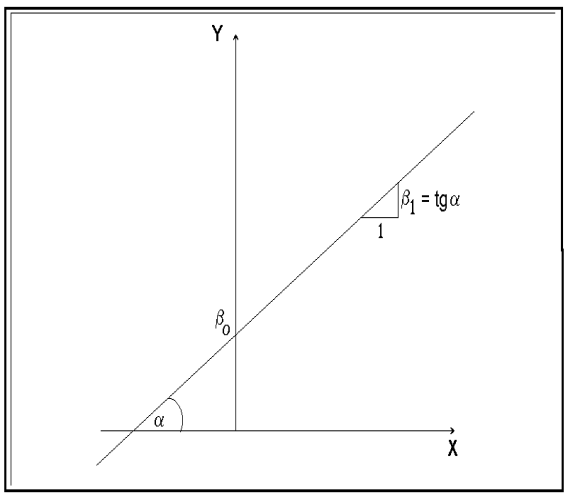


Figura 2.2: Interpretação dos coeficientes

Então, o valor esperado de Y para cada valor de X é

$$\mathrm{E}(Y|X) = \beta_0 + \beta_1 X$$

sendo que os parâmetros da equação da reta, β_0 e β_1, são constantes desconhecidas.

Verifica-se que para $X = 0$, β_0 representa o ponto onde a reta corta o eixo dos Y's e por isso é chamado **intercepto** (ou **coeficiente linear**). Já β_1 é chamado **coeficiente de regressão** ou **coeficiente angular** da reta, pois, da interpretação geométrica da derivada tem-se

$$\beta_1 = tg\alpha$$

sendo α o ângulo que a reta forma com o eixo dos X's. Além disso, tem-se que para um aumento de 1 unidade de X há um aumento de β_1 unidades na $\mathrm{E}(Y|X)$ (Figura 2.2).

Assim, dados n pares de valores, $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$, se for admitido que Y é função linear de X, pode-se estabelecer uma regressão linear simples, cujo modelo estatístico é

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \cdots, n$$

sendo β_0 e β_1 os parâmetros a serem estimados.

Ao se estabelecer esse modelo, pressupõe-se que:

(i) A relação entre Y e X é linear.

(ii) Os valores de X são fixos (ou controlados).

(iii) A média do erro é nula, isto é, $\mathrm{E}(\varepsilon_i) = 0$.

(iv) Para um dado valor de X, a variância do erro ε_i é sempre σ^2, isto é,

$$\mathrm{Var}(\varepsilon_i) = \mathrm{E}(\varepsilon_i^2) - [\mathrm{E}(\varepsilon_i)]^2 = \mathrm{E}(\varepsilon_i^2) = \sigma^2$$

o que implica em

$$\mathrm{Var}(Y_i) = \mathrm{E}[Y_i - \mathrm{E}(Y_i)]^2 = \mathrm{E}(\varepsilon_i^2) = \sigma^2.$$

Diz-se, então, que o erro é homocedástico, ou que se tem homocedasticia (do erro ou da variável dependente).

(v) O erro de uma observação é independente do erro de outra observação, isto é,

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_{i'}) = \mathrm{E}(\varepsilon_i \varepsilon_{i'}) - \mathrm{E}(\varepsilon_i)\mathrm{E}(\varepsilon_{i'}) = \mathrm{E}(\varepsilon_i \varepsilon_{i'}) = 0, \quad \text{para} \quad i \neq i'.$$

(vi) Os erros têm distribuição normal.

Logo, combinando (iii), (iv) e (v) tem-se $\varepsilon_i \sim N(0, \sigma^2)$ e, portanto, $Y_i \sim \mathrm{N}(\beta_0 + \beta_1 X_i, \sigma^2)$. A suposição de normalidade é necessária para a elaboração dos testes de hipóteses e obtenção de intervalos de confiança.

2.3 Estimação dos parâmetros

O problema agora é estimar os parâmetros β_0 e β_1 de tal forma que os desvios dos valores observados em relação aos estimados sejam mínimos (Figura 2.3).

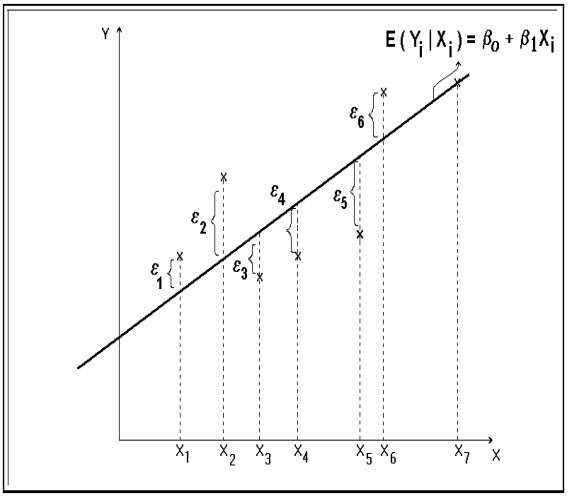


Figura 2.3: Regressão linear

Isso equivale a minimizar o comprimento do vetor $\boldsymbol{\varepsilon}= (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)'$. Usando a norma euclideana para avaliar o comprimento de $\boldsymbol{\varepsilon}$, tem-se:

$$Z =|| \boldsymbol{\varepsilon} ||^2= \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [Y_i - \mathrm{E}(Y_i)]^2 = \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 X_i]^2$$

Deseja-se, portanto, estimar β_0 e β_1 tais que Z seja mínima. Esse método é chamado **método dos mínimos quadrados**. Para isso, obtêm-se as derivadas parciais:

$$\begin{cases} \dfrac{\partial Z}{\partial \beta_0} &= 2\sum_{i=1}^{n}[Y_i - \beta_0 - \beta_1 X_i](-1) \\ \dfrac{\partial Z}{\partial \beta_1} &= 2\sum_{i=1}^{n}[Y_i - \beta_0 - \beta_1 X_i](-X_i) \end{cases}$$

e fazendo-se $\dfrac{\partial Z}{\partial \beta_0} = 0$ e $\dfrac{\partial Z}{\partial \beta_1} = 0$, obtêm-se as equações normais:

$$\sum_{i=1}^{n}[Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i] = 0 \Leftrightarrow n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \tag{2.1}$$

$$\sum_{i=1}^{n}[Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i]X_i = 0 \Leftrightarrow \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i. \tag{2.2}$$

De (2.1) tem-se

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{\hat{\beta}_1}{n}\sum_{i=1}^{n} X_i \tag{2.3}$$

ou

$$\boxed{\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.} \tag{2.4}$$

Substituindo-se (2.3) em (2.2) tem-se

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - \frac{\sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n}}{\sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}} = \frac{n\sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

ou, ainda, considerando-se $x_i = X_i - \bar{X}$ e $y_i = Y_i - \bar{Y}$, e como $\sum_{i=1}^n x_i = \sum_{i=1}^n (X_i - \bar{X}) = 0$ e $\sum_{i=1}^n y_i = \sum_{i=1}^n (Y_i - \bar{Y}) = 0$, têm-se as expressões equivalentes:

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2} = \frac{\sum_{i=1}^n X_i y_i}{\sum_{i=1}^n x_i^2} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.} \tag{2.5}$$

Obtendo-se as derivadas parciais de segunda ordem de Z em relação a β_0 e a β_1, tem-se:

$$\frac{\partial^2 Z}{\partial \beta_0^2} = 2\sum_{i=1}^n 1 = 2n > 0,$$

$$\frac{\partial^2 Z}{\partial \beta_0 \partial \beta_1} = 2\sum_{i=1}^n X_i$$

e

$$\frac{\partial^2 Z}{\partial \beta_1^2} = 2\sum_{i=1}^n X_i^2.$$

Portanto,

$$\begin{vmatrix} \frac{\partial^2 Z}{\partial \beta_0^2} & \frac{\partial^2 Z}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 Z}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 Z}{\partial \beta_1^2} \end{vmatrix} = \begin{vmatrix} 2n & 2\sum_{i=1}^n X_i \\ 2\sum_{i=1}^n X_i & 2\sum_{i=1}^n X_i^2 \end{vmatrix} = 4\left[n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2\right] = 4n\sum_{i=1}^n (X_i - \overline{X})^2 \geq 0,$$

e tem os elementos da diagonal positivos, o que mostra que Z é mínima para $\hat{\beta}_0$ e $\hat{\beta}_1$. Logo, a reta estimada pelo método dos mínimos quadrados é dada por:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i.$$

A solução do sistema de equações normais possui as seguintes propriedades:

a) O ponto $(\bar{X}, \bar{Y})$ é um ponto da reta estimada $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. (Verifique!)

b) Usando-se (2.1), tem-se:

$$\sum_{i=1}^n \hat{\varepsilon}_i = \sum_{i=1}^n (Y_i - \hat{Y}_i) = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

decorrendo que $\sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{Y}_i$.

c) Usando-se (2.2), tem-se:

$$\sum_{i=1}^{n} X_i\hat{\varepsilon}_i = \sum_{i=1}^{n} X_i(Y_i - \hat{Y}_i) = \sum_{i=1}^{n} X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0,$$

decorrendo que $\sum_{i=1}^{n} X_i Y_i = \sum_{i=1}^{n} X_i\hat{Y}_i$.

d) Usando-se (b) e (c), tem-se $\sum_{i=1}^{n} \hat{Y}_i\hat{\varepsilon}_i = 0$

$$\sum_{i=1}^{n} \hat{Y}_i\hat{\varepsilon}_i = \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 X_i)\hat{\varepsilon}_i = \hat{\beta}_0 \sum_{i=1}^{n} \hat{\varepsilon}_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i\hat{\varepsilon}_i = 0.$$

e) Os estimadores de quadrados mínimos $\hat{\beta}_0$ e $\hat{\beta}_1$ são funções lineares das observações Y_i's, isto é,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} Y_i$$

$$\boxed{\hat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i} \tag{2.6}$$

sendo

$$c_i = \frac{(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{x_i}{\sum_{i=1}^{n} x_i^2}, \tag{2.7}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} = \frac{\sum_{i=1}^{n} Y_i}{n} - \sum_{i=1}^{n} c_i Y_i \bar{X} = \sum_{i=1}^{n}(\frac{1}{n} - c_i\bar{X})Y_i,$$

$$\boxed{\hat{\beta}_0 = \sum_{i=1}^{n} d_i Y_i,} \tag{2.8}$$

sendo

$$d_i = \frac{1}{n} - c_i\bar{X}. \tag{2.9}$$

Note que

e.1) $\sum_{i=1}^{n} c_i = 0$

$$\sum_{i=1}^{n} \frac{(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \sum_{i=1}^{n}(X_i - \bar{X}) = 0.$$

e.2) $\sum_{i=1}^{n} c_i X_i = 1$

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

e.3) $\sum_{i=1}^{n} d_i = 1$ (Prove!)

e.4) $\sum_{i=1}^{n} d_i X_i = 0$ (Prove!)

f) Os estimadores de mínimos quadrados de β_0 e de β_1 são não viesados, isto é,

$$\mathrm{E}(\hat{\beta}_0) = \beta_0 \quad \text{e} \quad \mathrm{E}(\hat{\beta}_1) = \beta_1$$

A partir de (2.6), tem-se

$$\mathrm{E}(\hat{\beta}_1) = \mathrm{E}\left(\sum_{i=1}^{n} c_i Y_i\right) = \sum_{i=1}^{n} \mathrm{E}(c_i Y_i) = \sum_{i=1}^{n} c_i \mathrm{E}(\beta_0 + \beta_1 X_i + \varepsilon_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i X_i$$

e usando-se (e.1) e (e.2) tem-se:

$$\boxed{\mathrm{E}(\hat{\beta}_1) = \beta_1.}$$

A partir de (2.4), tem-se:

$$\mathrm{E}(\hat{\beta}_0) = \mathrm{E}(\bar{Y} - \hat{\beta}_1 \bar{X}) = \frac{\sum_{i=1}^{n} \mathrm{E}(Y_i)}{n} - \beta_1 \bar{X} = \frac{1}{n} \sum_{i=1}^{n}(\beta_0 + \beta_1 X_i) - \beta_1 \bar{X} = \beta_0 + \beta_1 \bar{X} - \beta_1 \bar{X}.$$

Portanto,

$$\boxed{\mathrm{E}(\hat{\beta}_0) = \beta_0.}$$

Faça o mesmo, usando (e.3) e (e.4).

g) A variância dos estimadores de mínimos quadrados de β_0 e β_1 é mínima entre as variâncias de quaisquer outros estimadores lineares (em Y) de β_0 e β_1 (**Teorema de Gauss**).

Dado que $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ e $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}$ e lembrando-se que os Y_i's são independentes, tem-se:

g.1) $\text{Var}(\hat{\beta}_1) = \text{Var}\left[\frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}\right] = \frac{1}{(\sum_{i=1}^n x_i^2)^2}\sum_{i=1}^n \text{Var}(x_i Y_i) = \frac{1}{(\sum_{i=1}^n x_i^2)^2}\sum_{i=1}^n x_i^2\sigma^2$

Portanto,

$$\boxed{\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}.} \tag{2.10}$$

g.2) $\text{Var}(\hat{\beta}_0) = \text{Var}(\bar{Y} - \hat{\beta}_1\bar{X}) = \text{Var}(\bar{Y}) + \bar{X}^2\text{Var}(\hat{\beta}_1) - 2\bar{X}\text{Cov}(\bar{Y}, \hat{\beta}_1)$ mas

$$\text{Var}(\bar{Y}) = \frac{1}{n^2}\sum_{i=1}^n \text{Var}(Y_i) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

e

$$\begin{aligned}\text{Cov}(\bar{Y}, \hat{\beta}_1) &= \text{Cov}\left(\frac{\sum_{i=1}^n Y_i}{n}, \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}\right) = \frac{1}{n\sum_{i=1}^n x_i^2}\text{Cov}\left(\sum_{i=1}^n Y_i, \sum_{i=1}^n x_i Y_i\right)\\ &= \frac{1}{n\sum_{i=1}^n x_i^2}\sum_{i=1}^n x_i\text{Var}(Y_i) = \frac{1}{n\sum_{i=1}^n x_i^2}\sum_{i=1}^n x_i\sigma^2\end{aligned}$$

$$\boxed{\text{Cov}(\bar{Y}, \hat{\beta}_1) = 0.} \tag{2.11}$$

Logo, $\text{Var}(\hat{\beta}_0) = \frac{\sigma^2}{n} + \bar{X}^2\frac{\sigma^2}{\sum_{i=1}^n x_i^2} - 0$

$$\boxed{\text{Var}(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n x_i^2}\right)\sigma^2.} \tag{2.12}$$

g.3) $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \text{Cov}(\bar{Y} - \bar{X}\hat{\beta}_1, \hat{\beta}_1) = \text{Cov}(\bar{Y}, \hat{\beta}_1) - \bar{X}\text{Var}(\hat{\beta}_1)$ o que implica em:

$$\boxed{\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{X}}{\sum_{i=1}^n x_i^2}\sigma^2.} \tag{2.13}$$

g.4) $\text{Var}(\hat{Y}_i) = \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 X_i) = \text{Var}(\hat{\beta}_0) + X_i^2\text{Var}(\hat{\beta}_1) + 2X_i\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$

$$\begin{aligned}\text{Var}(\hat{Y}_i) &= \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n x_i^2} + X_i^2\frac{1}{\sum_{i=1}^n x_i^2} - 2X_i\frac{\bar{X}}{\sum_{i=1}^n x_i^2}\right)\sigma^2\\ &= \left[\frac{1}{n} + \frac{1}{\sum_{i=1}^n x_i^2}(X_i^2 - 2X_i\bar{X} + \bar{X}^2)\right]\sigma^2 = \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^n x_i^2}\right]\sigma^2\end{aligned}$$

$$\boxed{\text{Var}(\hat{Y}_i) = \left[\frac{1}{n} + \frac{x_i^2}{\sum_{i=1}^n x_i^2}\right]\sigma^2.} \tag{2.14}$$

Teorema de Gauss

Considere o **Modelo I** estabelecido e suas pressuposições. Sejam $\hat{\beta}_0$ e $\hat{\beta}_1$ os estimadores não viesados de mínimos quadrados de β_0 e β_1 e $\tau = a_1\beta_0 + a_2\beta_1$ uma combinação linear de β_0 e β_1. Então, dentre todos os estimadores imparciais de τ, lineares em Y, o estimador

$$\hat{\tau} = a_1\hat{\beta}_0 + a_2\hat{\beta}_1$$

tem variância mínima, isto é, se $T = \sum_{i=1}^n l_i Y_i$, em que l_i são constantes arbitrárias e $\text{E}(T) = \tau$, então,

$$\text{Var}(\hat{\tau}) \leq \text{Var}(T).$$

Demonstração:

i) O estimador $\hat{\tau}$ de τ é não-viesado.

$$\text{E}(\hat{\tau}) = \text{E}(a_1\hat{\beta}_0 + a_2\hat{\beta}_1) = a_1\beta_0 + a_2\beta_1 = \tau.$$

ii) O estimador $\hat{\tau}$ de τ é também linear em Y.

Usando-se (2.6) e (2.8), tem-se:

$$\hat{\tau} = a_1\hat{\beta}_0 + a_2\hat{\beta}_1 = a_1\sum_{i=1}^n d_i Y_i + a_2\sum_{i=1}^n c_i Y_i = \sum_{i=1}^n (a_1 d_i + a_2 c_i)Y_i = \sum_{i=1}^n \kappa_i Y_i$$

sendo

$$\kappa_i = a_1 d_i + a_2 c_i, \tag{2.15}$$

$$c_i = \frac{(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2},$$

$$d_i = \frac{1}{n} - c_i\bar{X}.$$

Portanto, $\hat{\tau}$ é linear em Y.

iii) A variância de $\hat{\tau}$ é dada por:

$$\text{Var}(\hat{\tau}) = \text{Var}(a_1\hat{\beta}_0 + a_2\hat{\beta}_1) = a_1^2\text{Var}(\hat{\beta}_0) + a_2^2\text{Var}(\hat{\beta}_1) + 2a_1a_2\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$$

e usando-se (2.10), (2.12) e (2.13), tem-se:

$$\boxed{\text{Var}(\hat{\tau}) = \left[\frac{a_1^2}{n} + \frac{(a_2 - a_1\bar{X})^2}{\sum_{i=1}^n x_i^2}\right]\sigma^2.}$$

iv) Por imposição o estimador $T = \sum_{i=1}^n l_i Y_i$ é não viesado, isto é, $\text{E}(T) = \tau$, o que implica em:

$$\begin{aligned}\text{E}(T) &= \text{E}(\sum_{i=1}^n l_i Y_i) = \sum_{i=1}^n l_i \text{E}(\beta_0 + \beta_1 X_i + \varepsilon_i) \\ &= \beta_0 \sum_{i=1}^n l_i + \beta_1 \sum_{i=1}^n l_i X_i = a_1\beta_0 + a_2\beta_1.\end{aligned}$$

Portanto,

$$a_1 = \sum_{i=1}^n l_i \tag{2.16}$$

e

$$a_2 = \sum_{i=1}^n l_i X_i. \tag{2.17}$$

v) $\text{Var}(T) = \text{Var}(\sum_{i=1}^n l_i Y_i) = \sum_{i=1}^n l_i^2 \text{Var}(Y_i)$

Logo,

$$\boxed{\text{Var}(T) = \textstyle\sum_{i=1}^n l_i^2 \sigma^2.}$$

vi) $\text{Cov}(T, \hat{\tau}) = \text{Cov}(\sum_{i=1}^n l_i Y_i, \sum_{i=1}^n \kappa_i Y_i) = \sum_{i=1}^n l_i \kappa_i \text{Var}(Y_i) = \sum_{i=1}^n l_i \kappa_i \sigma^2$ e, usando-se (2.15) e (2.9), tem-se

$$\begin{aligned}\text{Cov}(T, \hat{\tau}) &= \sum_{i=1}^n l_i(a_1 d_i + a_2 c_i)\sigma^2 = \sum_{i=1}^n l_i \left[\frac{a_1}{n} - c_i \bar{X} a_1 + a_2 c_i\right]\sigma^2 \\ &= \sum_{i=1}^n l_i \left[\frac{a_1}{n} + (a_2 - \bar{X}a_1)c_i\right]\sigma^2\end{aligned}$$

e ainda, usando-se (2.7), (2.16) e (2.17), tem-se

$$\begin{aligned}\text{Cov}(T, \hat{\tau}) &= \left[\frac{a_1 \sum_{i=1}^n l_i}{n} + (a_2 - \bar{X}a_1)\frac{\sum_{i=1}^n l_i (X_i - \bar{X})}{\sum_{i=1}^n x_i^2}\right]\sigma^2 \\ &= \left[\frac{a_1^2}{n} + (a_2 - \bar{X}a_1)\frac{(a_2 - \bar{X}a_1)}{\sum_{i=1}^n x_i^2}\right]\sigma^2.\end{aligned}$$

Portanto,

$$\boxed{\text{Cov}(T, \hat{\tau}) = \left[\frac{a_1^2}{n} + \frac{(a_2 - \bar{X}a_1)^2}{\sum_{i=1}^n x_i^2}\right]\sigma^2 = \text{Var}(\hat{\tau}).}$$

vii) $\text{Var}(T - \hat{\tau})$

$$0 \leq \text{Var}(T - \hat{\tau}) = \text{Var}(T) + \text{Var}(\hat{\tau}) - 2\text{Cov}(T, \hat{\tau}) = \text{Var}(T) - \text{Var}(\hat{\tau}).$$

Portanto,

$$\boxed{\text{Var}(\hat{\tau}) \leq \text{Var}(T).}$$

Assim:

1) Se $T = \hat{\tau}$, isto é, se $\kappa_i = l_i = \dfrac{a_1}{n} + (a_2 - \bar{X}a_1)c_i$, então,

$$\text{Var}(\hat{\tau}) = \text{Var}(T).$$

2) Caso contrário, isto é, se $\kappa_i \neq l_i$, então,

$$\text{Var}(\hat{\tau}) < \text{Var}(T).$$

Casos especiais

1) Se $a_1 = 0$ e $a_2 = 1$, então, $\hat{\tau} = \hat{\beta}_1$. Logo, $\hat{\beta}_1$ é o estimador não viesado, de variância mínima de β_1.

2) Se $a_1 = 1$ e $a_2 = 0$, então, $\hat{\tau} = \hat{\beta}_0$. Logo, $\hat{\beta}_0$ é o estimador não viesado, de variância mínima de β_0.

3) Se $a_1 = 1$ e $a_2 = X_0$, então, $\hat{\tau} = \hat{Y}_{X_0} = \hat{\beta}_0 + \hat{\beta}_1 X_0$. Logo, $\hat{Y}_{X_0}$ é o estimador não viesado, de variância mínima de $\text{E}(Y_{X_0})$.

g) Como $Y_i \sim \text{N}(\beta_0 + \beta_1 X_i, \sigma^2)$, e, além disso, $\hat{\beta}_0$ e $\hat{\beta}_1$ são combinações lineares dos Y_i's, então,

$$\boxed{\hat{\beta}_0 \sim \text{N}(\beta_0, \text{Var}(\hat{\beta}_0))} \tag{2.18}$$

pois, $\text{E}(\hat{\beta}_0) = \beta_0$ e $\text{Var}(\hat{\beta}_0) = (\dfrac{1}{n} + \dfrac{\bar{X}^2}{\sum_{i=1}^n x_i^2})\sigma^2$ e

$$\boxed{\hat{\beta}_1 \sim \text{N}(\beta_1, \text{Var}(\hat{\beta}_1))} \tag{2.19}$$

pois, $\mathrm{E}(\hat{\beta}_1) = \beta_1$ e $\mathrm{Var}(\hat{\beta}_1) = \dfrac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$.

Além disso,

$$\boxed{\hat{Y}_i \sim N(\beta_0 + \beta_1 X_i, \mathrm{Var}(\hat{Y}_i))} \tag{2.20}$$

pois, $\mathrm{E}(\hat{Y}_i) = \beta_0 + \beta_1 X_i$ e $\mathrm{Var}(\hat{Y}_i) = \left(\dfrac{1}{n} + \dfrac{x_i^2}{\sum_{i=1}^{n} x_i^2}\right)\sigma^2$.

Observação: O problema aqui é que σ^2 é desconhecido e precisa ser estimado (ver expressão (2.28)).

2.4 Uma forma alternativa para o modelo de regressão linear simples – Variável X centrada

Uma forma reparametrizada com que se apresenta o modelo de regressão linear simples é obtida pela utilização da variável preditora centrada, isto é, pela utilização de $x_i = X_i - \bar{X}$ como variável preditora. Assim, tem-se:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i = (\beta_0 + \beta_1 \bar{X}) + \beta_1 (X_i - \bar{X}) + \varepsilon_i = \alpha + \beta_1 x_i + \varepsilon_i \tag{2.21}$$

De forma semelhante ao que foi feito no item (2.3), na página 22, tem-se:

$$Z = || \boldsymbol{\varepsilon} ||^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [Y_i - \mathrm{E}(Y_i|X_i)]^2 = \sum_{i=1}^{n} [Yi - \alpha - \beta_1 x_i]^2$$

que minimizado leva à estimativa de quadrados mínimos de α dada por:

$$\boxed{\hat{\alpha} = \bar{Y}} \tag{2.22}$$

e à estimativa para o β_1 dada pela expressão (2.5) na página 23, com variância dada pela expressão (2.10) na página 26. Mostra-se, ainda que,

$$\boxed{\mathrm{E}(\hat{\alpha}) = \alpha,}$$

$$\boxed{\mathrm{Var}(\hat{\alpha}) = \frac{1}{n}\sigma^2} \tag{2.23}$$

e

$$\boxed{\mathrm{Cov}(\hat{\alpha}, \hat{\beta}_1) = 0.} \tag{2.24}$$

Vê-se, portanto, que os estimadores de quadrados mínimos, $\hat{\alpha}$ e $\hat{\beta}_1$, não são correlacionados, pois $\text{Cov}(\hat{\alpha}, \hat{\beta}_1) = 0$.

2.5 Análise de variância e teste F

Obtenção das somas de quadrados

Pela Figura 2.4, vê-se que o desvio de uma determinada observação em relação ao valor estimado correspondente pode ser decomposto da seguinte forma:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = (Y_i - \bar{Y}) - (\hat{Y}_i - \bar{Y})$$

isto é,

desvio não explicado pelo modelo = desvio total - desvio devido ao modelo.

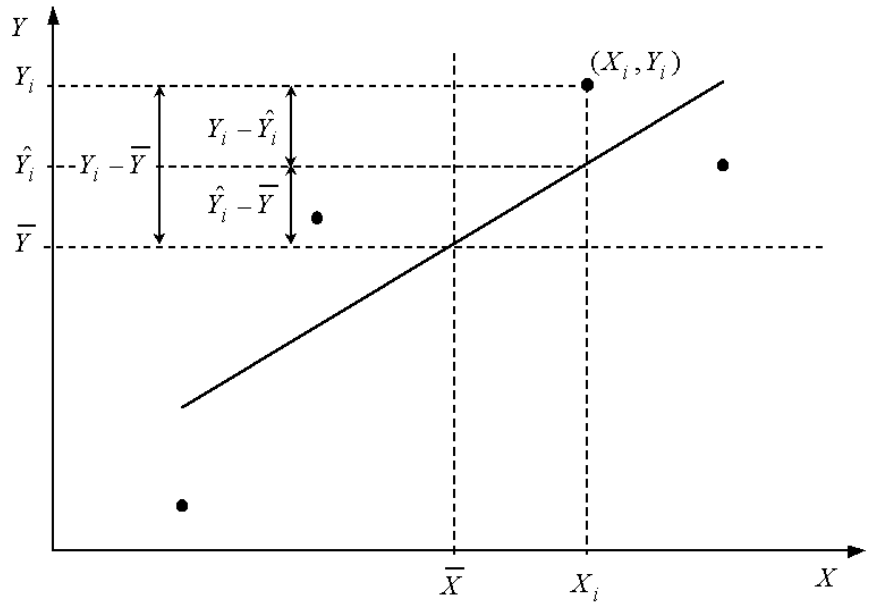


Figura 2.4: Decomposição dos desvios $\hat{\varepsilon}_i = Y_i - \hat{Y}_i = (Y_i - \bar{Y}) - (\hat{Y}_i - \bar{Y})$

Tem-se, então, que a soma de quadrados dos desvios (parte não explicada pelo

modelo) é dada por:

$$\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i-\bar{Y}-\hat{Y}_i+\bar{Y})^2$$

$$= \sum_{i=1}^{n}[(Y_i-\bar{Y})-(\hat{Y}_i-\bar{Y})]^2 = \sum_{i=1}^{n}(Y_i-\bar{Y})^2 - 2\sum_{i=1}^{n}(Y_i-\bar{Y})(\hat{Y}_i-\bar{Y}) + \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2.$$

Mas, já foi visto em (b), na página 24, que

$$\sum_{i=1}^{n}\hat{\varepsilon}_i = 0 \Rightarrow \sum_{i=1}^{n}Y_i = \sum_{i=1}^{n}\hat{Y}_i$$

e, em (d), na página 24, que

$$\sum_{i=1}^{n}\hat{Y}_i\hat{\varepsilon}_i = \sum_{i=1}^{n}\hat{Y}_i(Y_i-\hat{Y}_i) = 0 \Rightarrow \sum_{i=1}^{n}\hat{Y}_i^2 = \sum_{i=1}^{n}\hat{Y}_iY_i.$$

Então, $\sum_{i=1}^{n}(Y_i-\bar{Y})(\hat{Y}_i-\bar{Y}) = \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2$ e, portanto,

$$\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \sum_{i=1}^{n}(Y_i-\bar{Y})^2 - \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2.$$

Mas,

$$\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2 = \sum_{i=1}^{n}(\hat{\beta}_0+\hat{\beta}_1X_i-\bar{Y})^2 = \sum_{i=1}^{n}(\bar{Y}-\hat{\beta}_1\bar{X}+\hat{\beta}_1X_i-\bar{Y})^2$$

$$= \hat{\beta}_1^2\sum_{i=1}^{n}(X_i-\bar{X})^2 = \hat{\beta}_1^2\sum_{i=1}^{n}x_i^2 = \frac{(\sum_{i=1}^{n}x_iY_i)^2}{\sum_{i=1}^{n}x_i^2}$$

que por depender do coeficiente $\hat{\beta}_1$ é chamada soma de quadrados de regressão. Tem-se, portanto,

$$SQRes = SQTotal - SQReg$$

ou, ainda

$$\boxed{SQTotal = SQReg + SQRes}$$

isto é, a variabilidade total dos dados (medida pela $SQTotal$) pode ser subdividida em duas partes:

- uma parte que depende da magnitude do coeficiente $\hat{\beta}_1$, isto é, depende de quanto o modelo explica (medida pela $SQReg$);

- outra que depende da falta de ajuste do modelo ou de quanto o modelo não explica (medida pela $SQRes$).

Note-se que a $SQReg$, além de depender da magnitude do coeficiente de regressão, depende, também, da soma de quadrados de desvios dos X's. Portanto, é importante que os valores de X sejam bem escolhidos, de forma que a variação fique representada adequadamente e que a magnitude da $SQReg$ possa ser atribuída basicamente ao coeficiente de regressão.

Valor esperado das Somas de Quadrados

a) **SQTotal**

Dado que $SQTotal = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$, em que $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ e $\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{\varepsilon}$, então,

$$Y_i - \bar{Y} = \beta_1(X_i - \bar{X}) + \varepsilon_i - \bar{\varepsilon} = \beta_1 x_i + \varepsilon_i - \bar{\varepsilon}$$

e

$$SQTotal = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\beta_1 x_i + \varepsilon_i - \bar{\varepsilon})^2 = \beta_1^2 \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2 + 2\beta_1 \sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})x_i.$$

Portanto,

$$\mathrm{E}(SQTotal) = \beta_1^2 \sum_{i=1}^{n} x_i^2 + \mathrm{E}\left[\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2\right] + 2\beta_1 \mathrm{E}\left[\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})x_i.\right]$$

Mas, lembrando que $\mathrm{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \mathrm{E}(\varepsilon_i^2) = \sigma^2$ e que os ε_i's são independentes, isto é, para $i \neq i'$ $\mathrm{Cov}(\varepsilon_i, \varepsilon_{i'}) = \mathrm{E}(\varepsilon_i \varepsilon_{i'}) = 0$, tem-se

$$\mathrm{E}\left[\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})x_i\right] = \sum_{i=1}^{n}\mathrm{E}(\varepsilon_i - \bar{\varepsilon})x_i = 0$$

e

$$\begin{aligned}
\mathrm{E}\left[\sum_{i=1}^{n}(\varepsilon_i-\bar{\varepsilon})^2\right] &= \sum_{i=1}^{n}\mathrm{E}(\varepsilon_i-\bar{\varepsilon})^2=\sum_{i=1}^{n}\mathrm{E}(\varepsilon_i^2-2\varepsilon_i\bar{\varepsilon}+\bar{\varepsilon}^2)\\
&= \sum_{i=1}^{n}\left[\mathrm{E}(\varepsilon_i^2)-2\mathrm{E}(\varepsilon_i\bar{\varepsilon})+\mathrm{E}(\bar{\varepsilon}^2)\right]\\
&= \sum_{i=1}^{n}\left\{\sigma^2-2\mathrm{E}\left(\varepsilon_i\frac{\varepsilon_1+\cdots+\varepsilon_n}{n}\right)+\mathrm{E}\left[\left(\sum_{i=1}^{n}\frac{\varepsilon_1+\cdots+\varepsilon_n}{n}\right)^2\right]\right\}\\
&= \sum_{i=1}^{n}\left[\sigma^2-2\frac{\sigma^2}{n}+\frac{\sigma^2}{n}\right]=(n-1)\sigma^2.
\end{aligned}$$

Então,

$$\boxed{\mathrm{E}(SQTotal)=\beta_1^2\textstyle\sum_{i=1}^{n}x_i^2+(n-1)\sigma^2.} \tag{2.25}$$

b) **SQReg**

Dado que $SQReg=\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2=\hat{\beta}_1^2\sum_{i=1}^{n}x_i^2$ e que $\sum_{i=1}^{n}x_iX_i=\sum_{i=1}^{n}x_i^2$ tem-se:

$$
\begin{aligned}
\mathrm{E}(SQReg) &= \mathrm{E}\left(\hat{\beta}_1^2 \sum_{i=1}^{n} x_i^2\right) = \sum_{i=1}^{n} x_i^2 \mathrm{E}(\hat{\beta}_1^2) = \sum_{i=1}^{n} x_i^2 \frac{1}{(\sum_{i=1}^{n} x_i^2)^2} \mathrm{E}\left(\sum_{i=1}^{n} x_i Y_i\right)^2 \\
&= \frac{1}{\sum_{i=1}^{n} x_i^2} \mathrm{E}\left[\sum_{i=1}^{n} x_i(\beta_0 + \beta_1 X_i + \varepsilon_i)\right]^2 \\
&= \frac{1}{\sum_{i=1}^{n} x_i^2} \mathrm{E}\left[\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i X_i + \sum_{i=1}^{n} x_i \varepsilon_i\right]^2 \\
&= \frac{1}{\sum_{i=1}^{n} x_i^2} \mathrm{E}\left[\beta_1 \sum_{i=1}^{n} x_i X_i + \sum_{i=1}^{n} x_i \varepsilon_i\right]^2 \\
&= \frac{1}{\sum_{i=1}^{n} x_i^2} \mathrm{E}\left[\beta_1^2 \left(\sum_{i=1}^{n} x_i^2\right)^2 + 2\beta_1 \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} x_i \varepsilon_i + \left(\sum_{i=1}^{n} x_i \varepsilon_i\right)^2\right] \\
&= \beta_1^2 \sum_{i=1}^{n} x_i^2 + 2\beta_1 \sum_{i=1}^{n} x_i \mathrm{E}(\varepsilon_i) + \frac{1}{\sum_{i=1}^{n} x_i^2} \mathrm{E}\left(\sum_{i=1}^{n} x_i \varepsilon_i\right)^2 \\
&= \beta_1^2 \sum_{i=1}^{n} x_i^2 + \frac{1}{\sum_{i=1}^{n} x_i^2} \sum_{i=1}^{n} x_i^2 \sigma^2 = \beta_1^2 \sum_{i=1}^{n} x_i^2 + \sigma^2
\end{aligned}
$$

$$
\boxed{\mathrm{E}(SQReg) = \beta_1^2 \textstyle\sum_{i=1}^{n} x_i^2 + \sigma^2.} \tag{2.26}
$$

c) **SQRes**

Como *SQRes = SQTotal - SQReg* , então, usando-se (2.25) e (2.26), tem-se:

$$
\mathrm{E}(SQRes) = \mathrm{E}(SQTotal) - \mathrm{E}(SQReg) = \beta_1^2 \sum_{i=1}^{n} x_i^2 + (n-1)\sigma^2 - \beta_1^2 \sum_{i=1}^{n} x_i^2 - \sigma^2 = (n-2)\sigma^2
$$

$$
\boxed{\mathrm{E}(SQRes) = (n-2)\sigma^2.} \tag{2.27}
$$

Estimador da variância residual

Dado que

$$\mathrm{E}(SQRes) = (n-2)\sigma^2,$$

como consequência, tem-se que:

$$\mathrm{E}\left(\frac{SQRes}{n-2}\right) = \sigma^2,$$

e, portanto, um estimador não viesado para σ^2 é dado por

$$\hat{\sigma}^2 = \frac{SQRes}{n-2} = QMRes. \tag{2.28}$$

Tem-se, então, a partir de (2.10), (2.12) e (2.13), as variâncias e covariância estimadas, substituindo-se σ^2 por $QMRes$.

Independência entre parâmetros estimados e $SQRes$

Conforme será visto, matricialmente, no item (3.4) tem-se que $SQRes$ é independente de $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\alpha}$.

Distribuição das Somas de Quadrados

Conforme será visto no item (3.4) tem-se:

$$\frac{1}{\sigma^2}SQTotal = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n} y_i^2 \sim \chi^2\left(n-1, \frac{1}{2\sigma^2}\beta_1^2\sum_{i=1}^{n} x_i^2\right),$$

$$\frac{1}{\sigma^2}SQReg = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \sim \chi^2\left(1, \frac{1}{2\sigma^2}\beta_1^2\sum_{i=1}^{n} x_i^2\right)$$

e

$$\frac{1}{\sigma^2}SQRes = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \sim \chi^2(n-2).$$

Independência das $SQReg$ e $SQRes$

Dado que

$$SQReg = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

e

$$SQRes = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,$$

e ainda, $\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 x_i = \bar{Y} + \hat{\beta}_1 x_i$ e $\hat{Y}_i - \bar{Y} = \hat{\beta}_1 x_i$, então, usando-se (2.10) e (2.11), tem-se:

$$\begin{aligned}
\text{Cov}(\hat{Y}_i - \bar{Y}, Y_i - \hat{Y}_i) &= \text{Cov}(\hat{\beta}_1 x_i, Y_i - \bar{Y} - \hat{\beta}_1 x_i) \\
&= \text{Cov}(\hat{\beta}_1 x_i, Y_i) - \text{Cov}(\hat{\beta}_1 x_i, \bar{Y}) - \text{Var}(\hat{\beta}_1 x_i) \\
&= x_i \text{Cov}(\frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}, Y_i) - x_i \text{Cov}(\hat{\beta}_1, \bar{Y}) - x_i^2 \text{Var}(\hat{\beta}_1) \\
&= x_i^2 \frac{\sigma^2}{\sum_{i=1}^n x_i^2} - x_i^2 \frac{\sigma^2}{\sum_{i=1}^n x_i^2} = 0
\end{aligned}$$

pois, $\text{Cov}(\hat{\beta}_1, \bar{Y}) = 0$ (página 26), e, como os Y_i's têm distribuição normal, isso implica na independência das $SQReg$ e $SQRes$.

Quadro da análise da variância e teste F

O interesse agora é testar a hipótese $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$, isto é, se realmente existe uma relação linear entre Y e X. Já foi visto que:

$$\frac{1}{\sigma^2} SQRes \sim \chi^2_{n-2} \quad \text{e} \quad \frac{1}{\sigma^2} SQReg \sim \chi^2_{1,\delta}$$

sendo $\delta = \frac{1}{\sigma^2}\beta_1^2 \sum_{i=1}^n x_i^2$ o parâmetro de não centralidade, e, além disso, são independentes. Logo, sob $H_0 : \beta_1 = 0$, $\delta = 0$,

$$\frac{1}{\sigma^2} SQReg \sim \chi^2_1 \ \ (central)$$

e

$$F = \frac{\frac{SQReg}{\sigma^2}}{\frac{SQRes}{(n-2)\sigma^2}} \sim F_{1,n-2}.$$

Portanto, rejeita-se a hipótese $H_0 : \beta_1 = 0$, a um nível de $100\gamma\%$ de significância, se:

$$F_{calc} > F_{1,n-2;\gamma}$$

ou se

$$P(F_{1,n-2} > F_{calc}) < \gamma$$

sendo, em geral, $\gamma = 0,05$ ou $\gamma = 0,01$.

A partir dos resultados obtidos, pode-se obter o esquema do quadro da análise da variância e teste F mostrados na Tabela 2.1.

Tabela 2.1: Esquema de análise de variância e teste F

Causas de variação	G.L.	S.Q.	Q.M.	E(Q.M.)	F
Regressão linear	1	$\frac{(\sum_{i=1}^{n} x_i Y_i)^2}{\sum_{i=1}^{n} x_i^2}$	$\frac{SQReg}{1}$	$\sigma^2 + \beta_1^2 \sum_{i=1}^{n} x_i^2$	$\frac{QMReg}{QMRes}$
Resíduo	$n-2$	*por diferença*	$\frac{SQRes}{n-2}$	σ^2	
Total	$n-1$	$\sum_{i=1}^{n} Y_i^2 - C$			

sendo $C = \frac{(\sum_{i=1}^{n} Y_i)^2}{n}$.

2.6 Estimação por intervalo

O método utilizado aqui para a construção de um intervalo de confiança será o método da quantidade pivotal. Se $Q = q(Y_1, Y_2, \ldots, Y_n; \theta)$, isto é, uma função da amostra aleatória $Y_1, Y_2, \ldots, Y_n$ e de θ, o parâmetro de interesse e tem uma distribuição que independe de θ, então Q é uma quantidade pivotal. Logo, para qualquer γ fixo, tal que $0 < \gamma < 1$, existem q_1 e q_2, dependendo de γ, tais que

$$P[q_1 < Q < q_2] = 1 - \gamma$$

e a partir dessa expressão, pode-se obter um intervalo de confiança para θ com um coeficiente de confiança $1 - \gamma$.

Dado o modelo definido por (2.21), já foi visto que

$$\hat{\alpha} \sim N(\alpha, \frac{\sigma^2}{n}),$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2}\right]\sigma^2\right)$$

e

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}\right).$$

Por outro lado, tem-se que

$$\frac{1}{\sigma^2} SQRes \sim \chi^2_{n-2} \Leftrightarrow W = (n-2)\frac{QMRes}{\sigma^2} \sim \chi^2_{n-2}$$

e dada uma variável aleatória $Z \sim N(0,1)$ e, além disso, sendo Z e $QMRes$ independentes,

$$Q = \frac{Z}{\sqrt{\frac{W}{n-2}}} \sim t_{n-2}$$

que é o fundamento para a construção dos intervalos de confiança que se seguem.

Intervalo de confiança para α

Dado que

$$Z = \frac{\hat{\alpha} - \alpha}{\sqrt{V(\hat{\alpha})}} = \frac{\hat{\alpha} - \alpha}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

então,

$$\frac{\hat{\alpha} - \alpha}{\sqrt{\frac{\sigma^2}{n}}} \sqrt{\frac{(n-2)\sigma^2}{(n-2)QMRes}} = \frac{\hat{\alpha} - \alpha}{\sqrt{\hat{V}(\hat{\alpha})}} \sim t_{n-2}$$

e um intervalo de confiança para α, com um coeficiente de confiança $1-\gamma$ é obtido a partir de:

$$P\left[-t_{\frac{\gamma}{2}} \leq \frac{\hat{\alpha} - \alpha}{\sqrt{\hat{V}(\hat{\alpha})}} \leq t_{\frac{\gamma}{2}}\right] = 1 - \gamma$$

obtendo-se

$$P\left[\hat{\alpha} - t_{\frac{\gamma}{2}}\sqrt{\frac{QMRes}{n}} \leq \alpha \leq \hat{\alpha} + t_{\frac{\gamma}{2}}\sqrt{\frac{QMRes}{n}}\right] = 1 - \gamma$$

ou ainda, dada a simetria da distribuição t pode-se escrever:

$$\boxed{IC[\alpha]_{1-\gamma} : \hat{\alpha} \pm t_{n-2;\frac{\gamma}{2}}\sqrt{\frac{QMRes}{n}}.}$$

Intervalo de confiança para β_0

De forma semelhante, tem-se:

$$Z = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{V(\hat{\beta}_0)}} = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n x_i^2}\right]\sigma^2}} \sim N(0,1) \quad \text{e} \quad \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n x_i^2}\right]QMRes}} \sim t_{n-2}.$$

Logo,

$$\boxed{IC[\beta_0]_{1-\gamma} : \hat{\beta}_0 \pm t_{n-2;\frac{\gamma}{2}}\sqrt{\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n x_i^2}\right]QMRes}.}$$

Intervalo de confiança para β_1

De forma semelhante, tem-se:

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{V(\hat{\beta}_1)}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{1}{\sum_{i=1}^n x_i^2}\sigma^2}} \sim N(0,1) \quad \text{e} \quad \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{1}{\sum_{i=1}^n x_i^2} QMRes}} \sim t_{n-2}.$$

Logo,

$$\boxed{IC[\beta_1]_{1-\gamma} : \hat{\beta}_1 \pm t_{n-2;\frac{\gamma}{2}} \sqrt{\frac{QMRes}{\sum_{i=1}^n x_i^2}}.}$$

Intervalo de confiança para $\mathbf{E}(Y_i) = \beta_0 + \beta_1 X_i = \alpha + \beta_1 x_i$

Já foi visto que a aproximação de mínimos quadrados para Y_i é dada por

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = \hat{\alpha} + \hat{\beta}_1 x_i$$

com

$$\mathrm{E}(\hat{Y}_i) = \mathrm{E}(Y_i) = \beta_0 + \beta_1 X_i = \alpha + \beta_1 x_i$$

e

$$\mathrm{Var}(\hat{Y}_i) = \left[\frac{1}{n} + \frac{x_i^2}{\sum_{i=1}^n x_i^2}\right] \sigma^2.$$

Além disso,

$$\hat{Y}_i \sim N\left(\mathrm{E}(Y_i), \left[\frac{1}{n} + \frac{x_i^2}{\sum_{i=1}^n x_i^2}\right] \sigma^2\right).$$

Logo,

$$Z_i = \frac{\hat{Y}_i - \mathrm{E}(Y_i)}{\sqrt{\mathrm{Var}(\hat{Y}_i)}} \quad \text{e} \quad \frac{\hat{Y}_i - \mathrm{E}(Y_i)}{\sqrt{\left[\frac{1}{n} + \frac{x_i^2}{\sum_{i=1}^n x_i^2}\right] QMRes}} \sim t_{n-2}.$$

Portanto,

$$\boxed{IC[\mathrm{E}(Y_i)]_{1-\gamma} : \hat{Y}_i \pm t_{n-2;\frac{\gamma}{2}} \sqrt{\left[\frac{1}{n} + \frac{x_i^2}{\sum_{i=1}^n x_i^2}\right] QMRes}.}$$

Intervalo de previsão para $Y_h = \beta_0 + \beta_1 X_h + \varepsilon_h = \alpha + \beta_1 x_h + \varepsilon_h$ (Intervalo de previsão)

Frequentemente, há interesse em se estimar o valor de uma nova observação Y_h relativa ao valor X_h da variável preditora, isto é, deseja-se prever o valor da variável resposta para uma nova observação $X = X_h$.

O estimador de

$$Y_h = \beta_0 + \beta_1 X_h + \varepsilon_h = \alpha + \beta_1 x_h + \varepsilon_h$$

é dado por:

$$\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h = \hat{\alpha} + \hat{\beta}_1 x_h$$

e o erro de previsão é

$$(\hat{Y}_h - Y_h) = (\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)X_h - \varepsilon_h = (\hat{\alpha} - \alpha) + (\hat{\beta}_1 - \beta_1)x_h - \varepsilon_h$$

obtendo-se:

$$\mathrm{E}(\hat{Y}_h - Y_h) = 0 \Rightarrow \mathrm{E}(\hat{Y}_h) = \mathrm{E}(Y_h) \Rightarrow \mathrm{E}(\hat{Y}_h) = \beta_0 + \beta_1 X_h = \alpha + \beta_1 x_h = Y_h - \varepsilon_h \neq Y_h$$

e

$$\mathrm{Var}(\hat{Y}_h - Y_h) = \mathrm{Var}(\hat{Y}_h) + \mathrm{Var}(Y_h) = \left(\frac{1}{n} + \frac{x_h^2}{\sum_{i=1}^n x_i^2} + 1\right)\sigma^2$$

pois, $\hat{Y}_h$ e Y_h são variáveis aleatórias independentes, pela pressuposição (v) da página 21.

Para avaliar a precisão de $\hat{Y}_h$ como previsão do valor da nova observação, determina-se o intervalo de previsão para Y_h. Uma vez que, para determinado valor (X_h) da variável preditora, os valores de Y variam em torno de sua verdadeira média, isto é, em torno de $\mathrm{E}(Y_h)$ com variância σ^2, a variância que interessa é $\sigma^2 + \mathrm{Var}(\hat{Y}_h)$. Logo,

$$\boxed{IC[Y_h]_{1-\gamma} : \hat{Y}_h \pm t_{n-2;\frac{\gamma}{2}} \sqrt{\left(\frac{1}{n} + \frac{x_h^2}{\sum_{i=1}^n x_i^2} + 1\right) QMRes.}}$$

A Figura 2.5 mostra o aspecto que, em geral, assumem o intervalo de confiança para $\mathrm{E}(Y_h)$ e o intervalo de previsão para Y_h.

O conceito de intervalo de previsão é análogo ao de intervalo de confiança, com a diferença de que, enquanto o intervalo de confiança refere-se a uma constante (o parâmetro β_1, por exemplo), o intervalo de previsão refere-se a uma variável aleatória (Y_h, no caso).

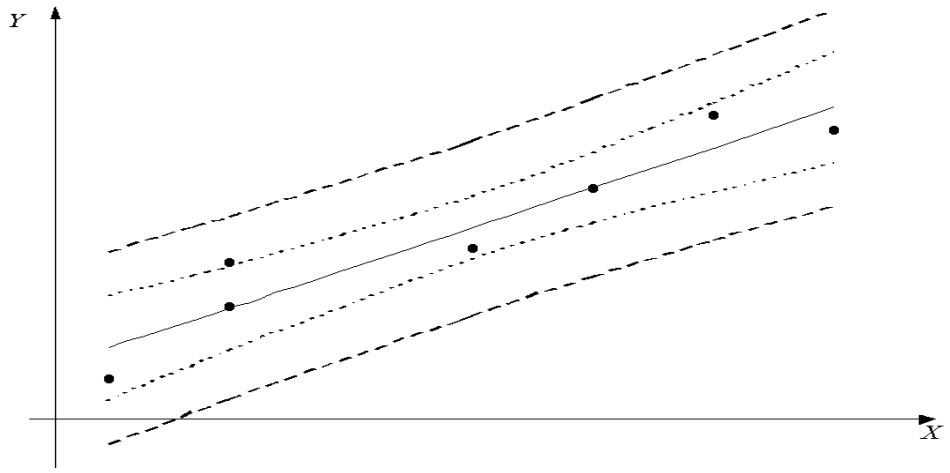


Figura 2.5: Intervalo de confiança (....) para $E(Y_h)$ e intervalo de previsão (- - -) para Y_h

2.7 Testes de hipóteses para os parâmetros

Teste de hipóteses para α

Em função do que já foi visto tem-se que o teste da hipótese:

$$H_0 : \alpha = \alpha_0 \quad versus \quad \begin{cases} H_{a_1} : \alpha < \alpha_0 \\ H_{a_2} : \alpha > \alpha_0 \\ H_{a_3} : \alpha \neq \alpha_0 \end{cases}$$

é obtido a partir de:

$$\frac{\hat{\alpha} - \alpha_0}{\sqrt{\hat{V}(\hat{\alpha})}} \sim t_{n-2}.$$

Assim, obtém-se:

$$\boxed{t_{calc} = \frac{\hat{\alpha} - \alpha_0}{\sqrt{\frac{QMRes}{n}}}}$$

e, a um nível de $100\gamma\%$ de significância, rejeita-se H_0, em favor de:

$H_{a_1} : \alpha < \alpha_0$ se $t_{calc} < -t_{n-2;\gamma}$ ou se $P(t_{n-2} < t_{calc}) < \gamma$;

$H_{a_2} : \alpha > \alpha_0$ se $t_{calc} > t_{n-2;\gamma}$ ou se $P(t_{n-2} > t_{calc}) < \gamma$;

$H_{a_3} : \alpha \neq \alpha_0$ se $|t_{calc}| > t_{n-2;\frac{\gamma}{2}}$ ou se $P(|t_{n-2}| > |t_{calc}|) < \gamma$;

isto é, as regiões de rejeição de H_0 são dadas pelos intervalos de t correspondentes às áreas hachuradas nas Figuras 2.6, 2.7 e 2.8, respectivamente.

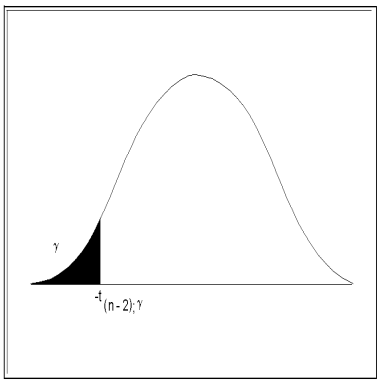


Figura 2.6: H_0 vs H_{a_1}

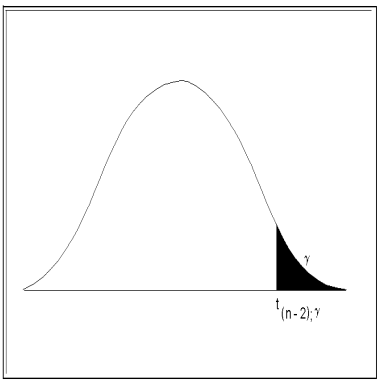


Figura 2.7: H_0 vs H_{a_2}

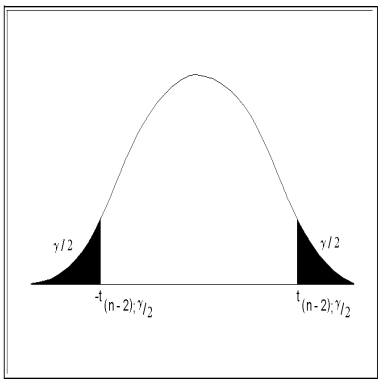


Figura 2.8: H_0 vs H_{a_3}

Teste de hipóteses para β_0

De forma semelhante, obtém-se o teste de hipóteses para β_0, isto é, o teste de:

$$H_0 : \beta_0 = \beta_{00} \quad versus \quad \begin{cases} H_{a_1} : \beta_0 < \beta_{00} \\ H_{a_2} : \beta_0 > \beta_{00} \\ H_{a_3} : \beta_0 \neq \beta_{00} \end{cases}$$

é obtido a partir de:

$$\boxed{t_{calc} = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2}\right) QMRes}}}$$

com regiões de rejeição de H_0 dadas pelos intervalos de t correspondentes às áreas hachuradas nas Figuras 2.6, 2.7 e 2.8, respectivamente.

Observação: Um caso particular importante é aquele em que $\beta_{00} = 0$, isto é, a reta passa pela origem.

Teste de hipóteses para β_1

De forma semelhante, obtém-se o teste de hipóteses para β_1, isto é, o teste de:

$$H_0 : \beta_1 = \beta_{10} \quad versus \quad \begin{cases} H_{a_1} : \beta_1 < \beta_{10} \\ H_{a_2} : \beta_1 > \beta_{10} \\ H_{a_3} : \beta_1 \neq \beta_{10} \end{cases}$$

é obtido a partir de:

$$\boxed{t_{calc} = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{QMRes}{\sum_{i=1}^{n} x_i^2}}}}$$

com regiões de rejeição de H_0 dadas pelos intervalos de t correspondentes às áreas hachuradas nas Figuras 2.6, 2.7 e 2.8, respectivamente.

Observação: No caso particular em que $\beta_{10} = 0$ (teste bilateral), tem-se que $t^2_{calc} = F_{calc}$.

2.8 Exemplo de aplicação

Considere o Exercício número 1 do item 1.4.1 da página 14. Usando-se, por exemplo, o SAS, obtêm-se os resultados da Tabela 2.2

Tabela 2.2: Esquema de análise de variância e teste F

Causas de variação	G.L.	S.Q.	Q.M.	F
Regressão linear	1	1.056,57	1.056,57	225,49 **
Resíduo	5	23,43	4,68	
Total	6	1.080,00		

$F_{1,5;0,05} = 6,61$, $F_{1,5;0,01} = 16,26$ e $P(F_{1;5} > 225,49) = 0,0000237$

Como $F_{calc} = 225,49 > F_{1,5;0,01} = 16,26$ ou, ainda, $P(F_{1;5} > 225,49) < 0,01$, rejeita-se $H_0 : \beta_1 = 0$, ao nível de 1% de significância. As estimativas e desvios padrões obtidos para os parâmetros foram:

$$\hat{\beta}_0 = -0,57, \quad s(\hat{\beta}_0) = 1,83,$$

$$\hat{\beta}_1 = 6,14, \quad s(\hat{\beta}_1) = 0,41,$$

ficando a reta estimada

$$\hat{Y}_i = -0,57 + 6,14X_i.$$

A estatística para o teste da hipótese $H_0 : \beta_0 = 0$ versus $H_a : \beta_0 \neq 0$ é :

$$t_{calc} = -0,31 < t_{5;0,025} = 2,571 \quad \text{ou} \quad P(|t_5| > 0,31) = 0,767$$

isto é, não se rejeita H_0 ao nível de 5% de significância, o que indicaria a possiblidade do ajuste de uma reta passando pela origem, e o que nesse caso é perfeitamente explicado na prática, pois no dia 0 a planta terá altura 0.

A estatística t para o teste da hipótese $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$, como esperado, é :

$$t_{calc} = 15,01 = \sqrt{225,49} = \sqrt{F_{calc}}.$$

Intervalos de confiança, com coeficientes de confiança de 95%, para β_0 e para β_1 são dados por:

$$IC(\beta_0)_{0,95} : (-5,275; 4,132)$$

e

$$IC(\beta_1)_{0,95} : (5,091; 7,195),$$

mostrando que existem evidências de que β_0 não é significativamente diferente de zero (o intervalo para β_0 inclui o zero) ao nível de 5% de significância, enquanto que β_1 o é (o intervalo não inclui o zero), confirmando o resultado obtido pelo teste F.

São obtidos, ainda, os resultados apresentados a seguir.

X	Y	$\hat{Y}$	$s(\hat{Y})$	LI_{IC}	LS_{IC}	LI_{IP}	LS_{IP}
1	5	5,57	1,48	1,78	9,36	-1,16	12,30
2	13	11,71	1,16	8,74	14,69	5,40	18,02
3	16	17,86	0,92	15,50	20,21	11,82	23,90
4	23	24,00	0,82	21,90	26,10	18,05	29,95
5	33	30,14	0,92	27,79	32,49	24,10	36,18
6	38	36,28	1,16	33,31	39,26	29,98	42,60
7	40	42,43	1,48	38,64	46,22	35,70	49,16

em que LI_{IC} e LS_{IC} são os limites do intervalo de confiança para $E(Y_h)$, com um coeficiente de 95% de confiança, e LI_{IP} e LS_{IP} são os limites do intervalo de previsão para Y_h, com um coeficiente de 95% de confiança. A Figura 2.9 mostra os intervalos de confiança para $E(Y_h)$ e de previsão para Y_h, bem como a reta estimada e os valores observados.

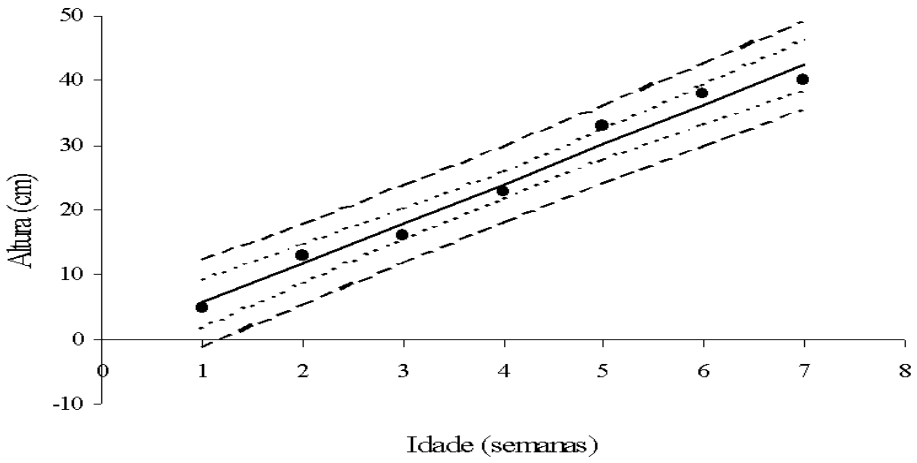


Figura 2.9: Intervalo de confiança para $E(Y_h)$ e intervalo de previsão para Y_h

2.9 Regressão linear por anamorfose

Existem determinados tipos de modelos não lineares que através de uma transformação tornam-se lineares e os parâmetros do modelo inicial podem, então, ser estimados através de funções deles. Geralmente, essas estimativas são usadas como valores iniciais para um processo iterativo. Como exemplos podem ser citados:

- **Modelo de Cobb-Douglas**, muito usado na área de Economia, e dado por:

$$R = \alpha Z^{\beta}$$

sendo R a renda bruta e Z, a área plantada.

Para linearizar esse modelo basta usar a função logarítmica e tem-se:

$$\log R = \log \alpha + \beta \log Z \Rightarrow Y = \beta_0 + \beta_1 X$$

sendo $Y = \log R$ a nova variável resposta, $X = \log Z$, a nova variável explicativa e por uma regressão linear simples estimam-se os parâmetros β_0 e β_1, e consequentemente, $\hat{\alpha} = e^{\hat{\beta}_0}$ e $\hat{\beta} = \hat{\beta}_1$.

- **Polinômios inversos**, cujas curvas são hiperbólicas, muito usados para descrever a relação existente entre peso e densidade de plantas, crescimento de plantas e balanço de íons, produtividade e doses de adubo, velocidade de reação e concentração de substrato em reações químicas de enzimas (**Equação de Michaelis-Menten**). A vantagem dos polinômios inversos em relação aos polinômios ordinários, é que, em geral, são funções não negativas, limitadas (por assíntotas) e não simétricas, o que pode muitas vezes explicar melhor fenômenos que ocorrem na prática (Nelder, 1966). Podem ser escritos, por exemplo, dentre outras, na forma linear

$$\frac{Z}{W} = \alpha Z + \beta \Rightarrow W = \frac{Z}{\alpha Z + \beta}.$$

em que W é a variável resposta (peso, altura, produtividade, velocidade de reação) e Z é a variável explicativa (densidade de plantas, balanço de íons, dose de adubo, concentração de substrato). Verifica-se que, à medida que Z aumenta, W tende para uma assíntota superior α^{-1}, isto é,

$$\lim_{Z \to \infty} \frac{Z}{\alpha Z + \beta} = \frac{1}{\alpha},$$

e que para valores de Z suficientemente pequenos, W é aproximadamente proporcional a $\beta^{-1}Z$. Tem como casos limites, uma reta quando $\alpha = 0$ e uma constante quando $\beta = 0$.

Na forma quadrática, tem-se:

$$\frac{Z}{W} = \alpha Z + \beta + \gamma Z^2 \Rightarrow W = \frac{Z}{\alpha Z + \beta + \gamma Z^2}$$

em que W é a variável resposta e Z é a variável explicativa. Para valores de Z suficientemente pequenos, W é aproximadamente proporcional a $\beta^{-1}Z$ e para valores grandes de Z

é aproximadamente proporcional a $(\gamma Z)^{-1}$. O valor máximo de W ocorre para $Z = \sqrt{\frac{\beta}{\gamma}}$ e é dado por $\frac{1}{2\sqrt{\beta\gamma}+\alpha}$, tal que α não afeta a posição do máximo, mas somente o valor que W assume.

A obtenção de estimativas iniciais para α, β e γ podem ser obtidas linearizando-se esses modelos da seguinte forma:

$$\frac{1}{W} = \alpha + \beta\frac{1}{Z} \Rightarrow Y = \beta_0 + \beta_1 X$$

e

$$\frac{1}{W} = \alpha + \beta\frac{1}{Z} + \gamma Z \Rightarrow Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

sendo que $Y = \frac{1}{W}$ é nova variável resposta, $X = \frac{1}{Z}$, $X_1 = \frac{1}{Z}$ e $X_2 = Z$ são as novas variáveis explicativas e por uma regressão linear simples estimam-se os parâmetros β_0, β_1 e β_2, e consequentemente, $\hat{\alpha} = \hat{\beta}_0$, $\hat{\beta} = \hat{\beta}_1$ e $\hat{\gamma} = \hat{\beta}_2$.

2.10 Teste para falta de ajuste (ou teste de linearidade)

Já foi visto que o

$$QMRes = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

da análise de variância da regressão é uma estimativa não tendenciosa da variância do erro ou da variância residual (σ^2), sob a suposição de que o modelo ajustado é correto. Suponha que o modelo proposto é

$$\boxed{\mathrm{E}(Y_i) = \mu(X_i) \Rightarrow Y_i = \mu(X_i) + \varepsilon_i} \tag{2.29}$$

e que o modelo correto seria

$$\boxed{\mathrm{E}(Y_i) = \gamma(X_i) \Rightarrow Y_i = \gamma(X_i) + \varepsilon_i^*.} \tag{2.30}$$

Figura 2.10: Modelos linear e quadrático

com $\mathrm{E}(\varepsilon_i^*) = 0$ e $\mathrm{Var}(\varepsilon_i^*) = \mathrm{E}[(\varepsilon_i^*)^2] = \sigma^2$.

Comparando-se os dois modelos, tem-se que o termo $B_i = \gamma(X_i) - \mu(X_i)$ estará incluído em ε_i de (2.29). Logo,

$$\mathrm{E}(\varepsilon_i) = B_i \quad \text{e} \quad \mathrm{E}(\varepsilon_i^2) = \mathrm{E}[(\varepsilon_i^* + B_i)^2] = \sigma^2 + B_i^2,$$

sendo que $B_i = \gamma(X_i) - \mu(X_i)$ é o viés, como mostra a Figura 2.10, no caso em que $\mu(X_i) = \beta_0 + \beta_1 X_i$ e $\gamma(X_i) = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$. Isso mostra que ao se usar o modelo (2.29), se ele for correto $B_i = 0$ e o $QMRes$ será uma estimativa não tendenciosa para a variância residual, isto é, $\mathrm{E}(QMRes) = \sigma^2$; se, por outro lado, não for correto, então, $\mathrm{E}(QMRes) = \sigma^2 + \dfrac{1}{n-2} B_i^2$.

Nesse caso em que (2.29) é o modelo de regressão linear simples, um gráfico pode mostrar essa falta de ajuste. Já, quando se têm modelos mais complicados, ou então, mais de uma variável explanatória, fica mais difícil. Necessário se torna, portanto, a obtenção de uma estimativa da variância residual σ^2 que independa do modelo. Isso é possível, usando-se o planejamento de coleta de observações repetidas de Y para cada X distinto, como mostra a Figura 2.11, para um determinado X_i. Considere k níveis de X_i para os quais são observados n_i valores de Y (Tabela 2.3).

Tabela 2.3: Valores de Y correspondentes a k níveis de X_i

X	Y				Totais	Médias
X_1	Y_{11}	Y_{12}	$\cdots$	Y_{1n_1}	$T_1 = Y_{1.}$	$\bar{Y}_1$
X_2	Y_{21}	Y_{22}	$\cdots$	Y_{2n_2}	$T_2 = Y_{2.}$	$\bar{Y}_2$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
X_k	Y_{k1}	Y_{k2}	$\cdots$	Y_{kn_k}	$T_k = Y_{k.}$	$\bar{Y}_k$

Essa outra estimativa de σ^2 é dada pelo **Quadrado Médio do Resíduo** de uma

análise de variância em que cada valor distinto de X é considerado como se fosse um *tratamento* a que está submetida a variável Y. Têm-se, então, dois resíduos: aquele a que se chama **desvios de regressão** (ou **resíduo da regressão**) e o **resíduo** propriamente dito (ou **erro puro**).

Figura 2.11: Valores repetidos de X_i

Figura 2.12: Decomposição de desvios totais

Figura 2.13: Decomposição de desvios de tratamentos

Tem-se, então, que a média das observações para o nível i é dada por

$$\bar{Y}_i = \frac{Y_{i1} + Y_{i2} + \cdots + Y_{in_i}}{n_i}$$

e, pode-se ter

$$\mathrm{E}(\bar{Y}_i) = \mu(X_i) \ \text{ (modelo proposto)} \quad \text{ou} \quad \mathrm{E}(\bar{Y}_i) = \gamma(X_i) \ \text{ (modelo correto).}$$

Logo,

$$d_{ij} = Y_{ij} - \bar{Y}_i \quad \text{e} \quad \frac{1}{n-k}\sum_{i=1}^{n} d_{ij}^2 = \hat{\sigma}^2 \Rightarrow \text{erro puro.}$$

Pela Figura 2.12, tem-se:

$$(Y_{ij} - \bar{Y}) = (Y_{ij} - \bar{Y}_i) + (\bar{Y}_i - \bar{Y}),$$

e, portanto,

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)(\bar{Y}_i - \bar{Y}) \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_i - \bar{Y})^2\end{aligned}$$

isto é,

$$SQTotal = SQErroPuro + SQTrat$$

em que

$$SQTotal = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - C$$

$$C = \frac{(\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij})^2}{N}, \text{ sendo } N = \sum_{i=1}^{k} n_i$$

$$SQTrat = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_i - \bar{Y})^2 = \sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2 = \sum_{i=1}^{k}\frac{T_i^2}{n_i} - C$$

$$SQErroPuro = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2 = SQTotal - SQTrat$$

pois,

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)(\bar{Y}_i - \bar{Y}) = \sum_{i=1}^{k}(\bar{Y}_i - \bar{Y})\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i) = \sum_{i=1}^{k}(\bar{Y}_i - \bar{Y})(Y_{i.} - n_i\frac{Y_{i.}}{n_i}) = 0.$$

Na realidade isso é equivalente ao modelo estatístico correspondente a um ensaio inteiramente casualizado (em que os tratamentos são os níveis de X) dado por:

$$Y_{ij} = \alpha + \gamma_i + \varepsilon_{ij}$$

sendo que γ_i é o efeito do i-ésimo tratamento, e dando origem ao esquema de **Análise de Variância** apresentado na Tabela 2.4.

Tabela 2.4: Esquema de análise de variância

Causas de variação	G.L.	S.Q.
Entre níveis de X	$k-1$	$SQTrat$
Resíduo	$N-k$	$SQRes$
Total	$N-1$	$SQTotal$

O interesse, agora, está em verificar se existe uma relação linear entre as médias de *tratamentos* (os diferentes valores de X considerados como níveis de uma variável qualitativa) e os valores de X_i's, isto é, desdobrar os $(k-1)$ graus de liberdade de *tratamentos* em 1 grau de liberdade para Regressão linear e $(k-2)$ graus de liberdade para desvios de regressão. Assim, tem-se o modelo para médias de tratamentos, dado por:

$$\mathrm{E}(\bar{Y}_i) = \beta_0 + \beta_1 X_i = \alpha + \beta_1 x_i$$

sendo $\mathrm{E}(\bar{Y}_i)$ estimado por:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = \hat{\alpha} + \hat{\beta}_1 x_i.$$

Tem-se, então, para um dado X_i (Figura 2.13)

$$\bar{Y}_i - \bar{Y} = (\bar{Y}_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

ou seja,

Entre níveis de X = falta de ajuste + efeito do modelo.

Portanto,

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_i - \bar{Y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_i - \hat{Y}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\hat{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$$

sendo

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 0 \quad (Prove!).$$

Então,

$$SQTrat = SQDesvios\ de\ Reg + SQReg$$

em que

$$SQReg = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{k} n_i(\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{k} n_i(\hat{\beta}_0 + \hat{\beta}_1 X_i - \bar{Y})^2 = \hat{\beta}_1^2 \sum_{i=1}^{k} n_i x_i^2.$$

Mas, como

$$\mathrm{E}(\bar{Y}_i) = \beta_0 + \beta_1 X_i = \alpha + \beta_1 x_i$$

tem-se que

$$Y_{ij} = \beta_0 + \beta_1 X_i + \varepsilon_{ij} = \alpha + \beta_1 x_i + \varepsilon_{ij}$$

e, portanto,

$$\varepsilon_{ij} = Y_{ij} - \beta_0 - \beta_1 X_i = Y_{ij} - \alpha - \beta_1 x_i.$$

Logo,

$$Z(\beta_0, \beta_1) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\varepsilon_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \beta_0 - \beta_1 X_i)^2$$

e pelo pelo método dos mínimos quadrados,

$$\begin{cases} \dfrac{\partial Z}{\partial \beta_0} &= 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \beta_0 - \beta_1 X_i)(-1) \\ \dfrac{\partial Z}{\partial \beta_1} &= 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \beta_0 - \beta_1 X_i)(-X_i) \end{cases}$$

$$\begin{cases} \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij} - \hat{\beta}_0 \sum_{i=1}^{k} n_i - \hat{\beta}_1 \sum_{i=1}^{k} n_i X_i = 0 \\ \\ \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_i Y_{ij} - \hat{\beta}_0 \sum_{i=1}^{k} n_i X_i - \hat{\beta}_1 \sum_{i=1}^{k} n_i X_i^2 = 0 \end{cases}$$

$$\begin{cases} N\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{k} n_i X_i = \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij} = \sum_{i=1}^{k} n_i \bar{Y}_i \\ \\ \hat{\beta}_0 \sum_{i=1}^{k} n_i X_i + \hat{\beta}_1 \sum_{i=1}^{k} n_i X_i^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_i Y_{ij} = \sum_{i=1}^{k} n_i X_i \bar{Y}_i. \end{cases}$$

Logo,

$$\boxed{\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}}$$

e

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{k} n_i X_i \bar{Y}_i - \frac{\sum_{i=1}^{k} n_i X_i \sum_{i=1}^{k} n_i \bar{Y}_i}{N}}{\sum_{i=1}^{k} n_i X_i^2 - \frac{(\sum_{i=1}^{k} n_i X_i)^2}{N}} \\ \\ &= \frac{\sum_{i=1}^{k} n_i (X_i - \bar{X})(\bar{Y}_i - \bar{Y})}{\sum_{i=1}^{k} n_i (X_i - \bar{X})^2} \end{aligned}$$

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{k} n_i x_i \bar{Y}_i}{\sum_{i=1}^{k} n_i x_i^2}.}$$

Portanto,

$$SQReg = \hat{\beta}_1^2 \sum_{i=1}^{k} n_i x_i^2 = \frac{(\sum_{i=1}^{k} n_i x_i \bar{Y}_i)^2}{\sum_{i=1}^{k} n_i x_i^2}$$

e

$$SQD = SQDesvios\ de\ Reg = SQTrat\ -\ SQReg$$

ficando o novo quadro da análise de variância dado pela Tabela 2.5.

Verifica-se que

$$\mathrm{E}(QMD) = \mathrm{E}\left[\frac{SQD}{k-2}\right] = \sigma^2 + \frac{\sum_{i=1}^{k} n_i [\mu(X_i) - (\beta_0 + \beta_1 X_i)]^2}{k-2}.$$

Interessa, inicialmente, testar a falta de ajuste (ou linearidade) do modelo, isto é, testar a hipótese:

$$H_0 : \mu(X) = \beta_0 + \beta_1 X \quad \Rightarrow \quad \mu(X) - \beta_0 - \beta_1 X = 0.$$

Sob essa hipótese

$$\mathrm{E}(QMD) = \sigma^2 \ \ e \ \ \frac{1}{\sigma^2} SQD \sim \chi^2_{k-2}.$$

Tabela 2.5: Esquema de análise de variância

Causas de variação	G.L.	S.Q.	Q.M.	F
Regressão linear	1	$SQReg$	$QMReg$	F_{Reg}
Desvios de regressão	$k-2$	SQD	QMD	F_D
Entre níveis de X	$k-1$	$SQTrat$	$QMTrat$	F_{Trat}
Resíduo	$N-k$	$SQRes$	$QMRes$	
Total	$N-1$	$SQTotal$		

Além disso,

$$\frac{1}{\sigma^2}SQRes \sim \chi^2_{N-k}.$$

Logo a estatística

$$F_D = \frac{QMD}{QMRes} \sim F_{k-2,N-k}.$$

Portanto, rejeita-se H_0 , a um nível de $100\gamma\%$ de significância, se $F_D > F_{k-2,N-k;\gamma}$ ou se $Pr(F_{k-2,N-k} > F_D) < \gamma$. Isso significa que existem evidências de que o modelo linear não satisfaz, havendo necessidade de se procurar outro modelo. Além disso, faz-se, também, o teste para a regressão linear, isto é, o teste da hipótese:

$$H_0 : \beta_1 = 0 \;\text{ versus }\; H_a : \beta_1 \neq 0.$$

Como resultados desses dois testes podem ocorrer as situações:

$$\textbf{Caso 1}: \begin{cases} Teste\ de\ falta\ de\ ajuste : não\ significativo \\ Teste\ da\ regressão\ \ (H_0 : \beta_1 = 0) : não\ significativo \\ Modelo\ estimado : \hat{Y}_{ij} = \hat{\beta}_0 = \bar{Y} \end{cases}$$

$$\textbf{Caso 2}: \begin{cases} Teste\ de\ falta\ de\ ajuste : não\ significativo \\ Teste\ da\ regressão\ \ (H_0 : \beta_1 = 0) : significativo \\ Modelo\ estimado : \hat{Y}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 X_i \end{cases}$$

$$\textbf{Caso 3}: \begin{cases} Teste\ de\ falta\ de\ ajuste : significativo \\ Teste\ da\ regressão\ \ (H_0 : \beta_1 = 0) : não\ significativo \\ Modelo\ sugerido : Y_{ij} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_{ij}\ ou\ de\ grau\ superior \end{cases}$$

$$\textbf{Caso 4}: \begin{cases} Teste\ de\ falta\ de\ ajuste : significativo \\ Teste\ da\ regressão\ \ (H_0 : \beta_1 = 0) : significativo \\ Modelo\ sugerido : Y_{ij} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_{ij}\ ou\ de\ grau\ superior \end{cases}$$

Esses 4 casos são mostrados, respectivamente, nas Figuras 2.14, 2.15, 2.16 e 2.17.

Exemplo: Considere os dados do Exercício 2, item 1.4.1, página 14.

Figura 2.14: Caso 1 Figura 2.15: Caso 2 Figura 2.16: Caso 3 Figura 2.17: Caso 4

a) A partir do modelo: $Y_{ij} = \beta_0 + \beta_1 X_i + \varepsilon_{ij}$, tem-se a Tabela 2.6.

Tabela 2.6: Análise de regressão

Causas de variação	G.L.	S.Q.	Q.M.	F
Regressão linear	1	90,83	90,83	
Resíduo 1	8	44,77	5,60	
Total	9	135,60		

b) A partir do modelo: $Y_{ij} = \alpha + \gamma_i + \varepsilon_{ij}$, tem-se a Tabela 2.7.

Tabela 2.7: Análise de variância

Causas de variação	G.L.	S.Q.	Q.M.	F
Entre níveis de X	7	132,71	18,96	
Resíduo	2	2,89	1,443	
Total	9	135,60		

c) Combinando-se os dois quadros, tem-se a Tabela 2.8.

d) **Conclusões**: Como para *falta de ajuste*, $F_{calc} = 4,84 < F_{6;2;0,05}$ ou se $P(F_{6;2} > 4,84) = 0,1812 > 0,05$, não se rejeita H_0, ao nível de 5% de significância. Vê-se, ainda, que o teste para a hipótese $H_0 : \beta_1 = 0$ é significativo ao nível de 5% de significância, indicando a evidência da tendência linear, isto é, a relação existente entre consumo de alimentos e peso médio das galinhas. A Figura 2.18 mostra a reta ajustada e os valores observados. Convém observar que esse exemplo tem um número pequeno de observações e, além disso, apenas um dos pesos $(5, 1)$ está repetido três vezes.

Tabela 2.8: Análise de variância

Causas de variação	G.L.	S.Q.	Q.M.	F
Regressão linear	1	90,83	90,83	62,93 *
Desvios de regressão	6	41,88	6,98	4,84 ns
Entre níveis de X	7	132,71		
Resíduo	2	2,89	1,443	
Total	9	135,60		

$F_{6;2;0,05} = 19,33$, $F_{6;2;0,01} = 99,33$ e $P(F_{6;2} > 4,84) = 0,1812$

$F_{1,2;0,05} = 18,51$, $F_{1,2;0,01} = 98,50$ e $P(F_{1;2} > 62,33) = 0,0155$

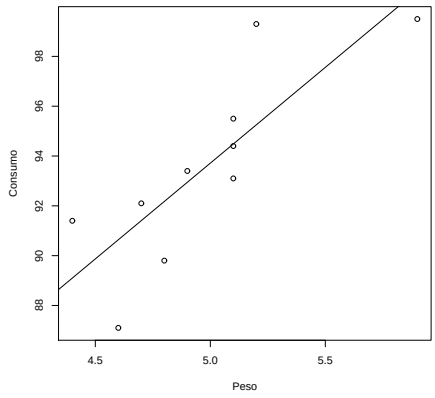


Figura 2.18: Reta ajustada e valores observados

Se a *falta de ajuste* fosse significativa, concluir-se-ia que o modelo linear utilizado não era o adequado, havendo necessidade de se utilizar um modelo de grau maior. O quadrado médio residual não estimaria corretamente a variância residual (σ^2), pois estaria incluindo um erro sistemático devido ao uso de um modelo inadequado.

2.11 Coeficiente de determinação

É definido por

$$R^2 = \frac{SQReg}{SQTotal} = 1 - \frac{SQRes}{SQTotal}$$

e indica a proporção da variação de Y que é explicada pela regressão. Note que $0 \leq R^2 \leq 1$.

É, portanto, uma medida descritiva da qualidade do ajuste obtido. Entretanto, o valor do coeficiente de determinação depende do número de observações da amostra, tendendo

a crescer quando n diminui; no limite para $n = 2$, tem-se sempre $R^2 = 1$, pois dois pontos determinam uma reta e os desvios são, portanto, nulos. Numa tentativa de correção desse problema, foi definido o **coeficiente de determinação ajustado** para graus de liberdade, indicado por $\bar{R}^2$. Tem-se que:

$$1 - R^2 = 1 - \frac{SQReg}{SQTotal} = \frac{SQRes}{SQTotal}$$

O **coeficiente de determinação ajustado** é definido por:

$$1 - \bar{R}^2 = \frac{\frac{1}{n-2}SQRes}{\frac{1}{n-1}SQTotal} = \frac{n-1}{n-2}(1 - R^2)$$

ou ainda,

$$\bar{R}^2 = R^2 - \frac{1}{n-2}(1 - R^2)$$

Excluindo-se o caso em que $R^2 = 1$, tem-se que $\bar{R}^2 < R^2$. Note que $\bar{R}^2$ pode ser negativo.

A estatística R^2 deve ser usada com precaução, pois é sempre possível torná-la maior pela adição de um número suficiente de termos. Assim, se, por exemplo, não há pontos repetidos (mais do que um valor Y para um mesmo X) um polinômio de grau $n - 1$ dará um ajuste perfeito ($R^2 = 1$) para n dados. Quando há valores repetidos, R^2 não será nunca igual a 1, pois o modelo não poderá explicar a variabilidade devido ao erro puro.

Embora R^2 aumente se uma nova variável é adicionada ao modelo, isso não significa necessariamente que o novo modelo é superior ao anterior. A menos que a soma de quadrados residual do novo modelo seja reduzida de uma quantia igual ao quadrado médio residual original, o novo modelo terá um quadrado médio residual maior do que o original, devido à perda de 1 grau de liberdade. Na realidade esse novo modelo poderá ser pior do que o anterior.

A magnitude de R^2, também, depende da amplitude de variação da variável regressora. Geralmente, R^2 aumentará com maior amplitude de variação dos X's e diminuirá em caso contrário. Pode-se mostrar que:

$$E(R^2) \approx \frac{\beta_1^2 \sum_{i=1}^{n} x_i^2}{\beta_1^2 \sum_{i=1}^{n} x_i^2 + \sigma^2}$$

Assim, um valor grande de R^2 poderá ser grande simplesmente porque X variou em uma amplitude muito grande. Por outro lado R^2 poderá ser pequeno porque a amplitude dos X's foi muito pequena para permitir que uma relação com Y fosse detectada.

Em geral, também, R^2 não mede a magnitude da inclinação da linha reta. Um valor grande de R^2 não significa uma reta mais inclinada. Além do mais, ele não leva em consideração a falta de ajuste do modelo; ele poderá ser grande, mesmo que Y e X estejam não linearmente relacionados (ver Figura 22).

Dessa forma, vê-se que R^2 não deve ser considerado sozinho, mas sempre aliado a outros diagnósticos do modelo.

No caso em que existem repetições para as doses de X tem-se:

$$R^2 = \frac{SQReg}{SQTrat},$$

$$1 - R^2 = 1 - \frac{SQReg}{SQTrat} = \frac{SQ \quad Falta \quad de \quad Ajuste}{SQTrat}$$

e o **coeficiente de determinação ajustado** definido por:

$$1 - \bar{R}^2 = \frac{\frac{1}{t-2} SQ \quad Falta \quad de \quad Ajuste}{\frac{1}{t-1} SQTrat} = \frac{t-1}{t-2}(1 - R^2)$$

ou ainda,

$$\bar{R}^2 = R^2 - \frac{1}{t-2}(1 - R^2)$$

2.12 Exercícios

1. Considere o modelo de regressão linear passando pela origem

$$Y_i = \beta X_i + \varepsilon_i, \qquad (i = 1, \ldots, n). \tag{2.31}$$

Pede-se:

a) Mostre que a estimativa de quadrados mínimos de β é dada por:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} {X_i}^2}.$$

b) Obtenha Var($\hat{\beta}$).

2. Seja

$$\begin{aligned} Y_1 &= \theta + \varepsilon_1 \\ Y_2 &= 2\theta - \phi + \varepsilon_2 \\ Y_3 &= \theta + 2\phi + \varepsilon_3 \end{aligned}$$

em que $\mathrm{E}(\varepsilon_i) = 0$ $(i = 1, 2, 3)$. Encontre as estimativas de quadrados mínimos de θ e ϕ.

3. Encontre as estimativas de mínimos quadrados dos parâmetros dos modelos que se seguem. Obter as variâncias e covariâncias das estimativas dos parâmetros, supondo que $\mathrm{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$, que o erro de uma observação é independente do erro de outra observação e que X é controlada sem erro ou com erro desprezível.

   (a) $Y_i = i + \theta + \varepsilon_i, (i = 1, 2, 3)$.

   (b) $Y_i = i\theta + \varepsilon_i, (i = 1, \ldots, 4)$.

   (c) $Y_1 = \theta + \varepsilon_1$
   $Y_2 = 2\theta - \phi + \varepsilon_2$
   $Y_3 = \theta + 2\phi + \varepsilon_3$.

   (d) $Y_i = \beta_0 + \beta_1 X_i + \beta_2(3X_i^2 - 2) + \varepsilon_i, (i = 1, 2, 3)$, sendo $X_1 = -1$, $X_2 = 0$ e $X_3 = 1$. Mostre que as estimativas de mínimos quadrados de β_0 e β_1 não se alteram se $\beta_2 = 0$.

   (e) Modelo de regressão linear reparametrizado

$$Y_i = \alpha + \beta_1(X_i - \bar{X}) + \varepsilon_i = \alpha + \beta_1 x_i + \varepsilon_i, (i = 1, \ldots, n).$$

   sendo $x_i = X_i - \bar{X}$ chamada variável centrada.

   (f) Modelo de regressão linear segmentada

$$Y_i = \begin{cases} \alpha + \varepsilon_i & (i = 1, 2, 3) \\ \alpha + \beta(X_i - X_3) + \varepsilon_i & (i = 4, 5) \end{cases}$$

   sendo $X_1 = 0$, $X_2 = 2$, $X_3 = 4$, $X_4 = 6$ e $X_5 = 8$.

   (g) Modelo de regressão linear segmentada

$$Y_i = \begin{cases} \alpha + \beta_1(X_i - X_3) + \varepsilon_i & (i = 1, 2) \\ \alpha + \beta_2(X_i - X_3) + \varepsilon_i & (i = 3, 4, 5) \end{cases} \tag{2.32}$$

   sendo $X_1 = 0$, $X_2 = 2$, $X_3 = 4$, $X_4 = 6$ e $X_5 = 8$.

(h) Modelo de regressão linear segmentada

$$Y_i = \begin{cases} \alpha + \beta_1(X_i - X_k) + \varepsilon_i & (i = 1, \ldots, k) \\ \alpha + \beta_2(X_i - X_k) + \varepsilon_i & (i = k+1, \ldots, n) \end{cases} \tag{2.33}$$

sendo $1 < k < n$.

4. Considere os conjuntos de dados apresentados nos Exercícios 1 a 5 do item 1.4.1 (pág. 14 e 15) e o modelo de regressão

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

Pede-se:

(a) Obtenha as estimativas de quadrados mínimos de β_0 e β_1.

(b) Obtenha $\text{Var}(\hat{\beta}_0)$, $\text{Var}(\hat{\beta}_1)$ e $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$.

(c) Onde couber, considere o modelo de regressão

$$X_i = \beta_0' + \beta_1' Y_i + \varepsilon_i'$$

e obtenha as estimativas de quadrados mínimos de β_0' e β_1'.

(d) Obtenha $\text{Var}(\hat{\beta}_0')$, $\text{Var}(\hat{\beta}_1')$ e $\text{Cov}(\hat{\beta}_0', \hat{\beta}_1')$.

(e) Complete os gráficos de dispersão com as retas de regressão.

(f) Comente sobre o ajuste, apenas olhando os gráficos.

5. Obtenha as estimativas de quadrados mínimos dos parâmetros do modelo (2.32), considerando o conjunto de dados a seguir

i	1	2	3	4	5
X_i	0	2	4	6	8
Y_i	4	6	10	9	6

6. Considere o conjunto de dados apresentado no Exercício 8 do item 1.4.1 (pág. 18) e o modelo de regressão

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i.$$

Pede-se:

(a) Obtenha as estimativas de quadrados mínimos de β_0, β_1 e β_2.

(b) Complete o gráfico de dispersão com a curva de regressão.

(c) Comente sobre o ajuste, apenas olhando o gráfico.

b) Obtenha $\text{Var}(\hat{\beta}_0)$, $\text{Var}(\hat{\beta}_1)$, $\text{Var}(\hat{\beta}_2)$, $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$, $\text{Cov}(\hat{\beta}_0, \hat{\beta}_2)$ e $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$.

7. Considere o modelo de regressão linear passando pela origem (2.31). Obtenha as somas de quadrados e o quadro da análise de variância.

8. Considere o modelo de regressão linear múltipla reparametrizado

$$Y_i = \alpha + \beta_1(X_{1i} - \bar{X}_1) + \beta_2(X_{2i} - \bar{X}_2) + \varepsilon_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \qquad (i = 1, \ldots, n)$$

onde $x_{1i} = X_{1i} - \bar{X}_1$ e $x_{2i} = X_{2i} - \bar{X}_2$ são variáveis centradas. Obtenha as somas de quadrados e o quadro da análise de variância.

9. Considere o modelo de regressão linear segmentada

$$Y_i = \alpha + \beta_1(X_i - X_u)I_{\{i<u\}} + \beta_2(X_i - X_u)I_{\{i\geq u\}} + \varepsilon_i, \qquad (i = 1, \ldots, u, \ldots, n)$$

onde $I_{\{.\}}$ são variáveis indicadoras, que assumem o valor 1 quando a condição entre chaves estiver satisfeita ou o valor 0, caso contrário. Obtenha as somas de quadrados e o quadro da análise de variância.

10. Considere os conjuntos de dados apresentados nos Exercícios 1 a 5 do item 1.4.1 (pág. 14 a 15), os resultados do Exercício 4 do item 2.12 (pág. 61) e o modelo de regressão

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

Pede-se:

(a) Obter o quadro da análise de regressão e fazer o teste F. Tirar conclusões a um nível de de 5% de significância.

(b) Testar a hipótese $H_0 : \beta_1 = 0$ com um nível de significância de 5% de probabilidade e verificar que $F_{\text{calc}} = t^2_{\text{calc}}$.

(c) Testar a hipótese $H_0 : \beta_0 = 0$ com um nível de significância de 5%.

(d) Construir os intervalos de confiança para β_0 e β_1, considerando um coeficiente de de 95% de confiança.

11. Considere o conjunto de dados apresentado no Exercício do item 1.4.2 (pág. 16). Obtenha as estimativas iniciais de mínimos quadrados de a, b e d, linearizando os modelos:

$$I = aT^b \quad \text{e} \quad V = dT^{b-1}$$

12. Linearize as funções relacionadas a seguir, ou seja, determine as transformações $f_1(\cdot), \ldots, f_4(\cdot)$ tais que $Y^* = \beta_0 + \beta_1 X^*$, onde $X^* = f_1(X)$, $Y^* = f_2(Y)$, $\beta_0 = f_3(\theta_0)$ e $\beta_1 = f_4(\theta_1)$.

(a) Função potência: $Y = \theta_0 X^{\theta_1}$, $Y > 0$, $\theta_0 > 0$ e $X > 0$.

(b) Função exponencial: $Y = \theta_0 \exp(\theta_1 X)$, $Y > 0$ e $\theta_0 > 0$.

(c) Função logarítmica: $Y = \theta_0 + \theta_1 \ln X$, $X > 0$.

(d) Função hiperbólica: $Y = \dfrac{X}{\theta_0 X - \theta_1}$, $X > \theta_1/\theta_0$ e $Y > 0$.

(e) Função logística com dois parâmetros: $Y = \dfrac{\exp(\theta_0 + \theta_1 X)}{1 + \exp(\theta_0 + \theta_1 X)}$, $0 < Y < 1$

(f) Função de Gompertz com dois parâmetros: $Y = \exp[-\exp(\theta_0 + \theta_1 X)]$, $0 < Y < 1$.

13. Faça um estudo completo das funções relacionadas no item anterior, ou seja, para cada uma delas:

    (a) Obtenha os conjuntos domínio e imagem.

    (b) Obtenha os intervalos de X para os quais a função é crescente e aqueles para os quais ela é decrescente.

    (c) Estude a função quanto à concavidade e a existência de pontos de inflexão.

    (d) Verifique se há assíntotas verticais ou horizontais.

    (e) Construa gráficos representativos do seu comportamento.

14. Paes de Camargo *et al* (1982), estudando a construção de um tensiômetro de leitura direta, obtiveram os resultados que se seguem para valores de alturas da câmara no tensiômetro (X), em mm, e tensão da água no solo (Y), em mb.

X	9	12	30	42	57	102	147	210	290
Y	217	291	439	515	603	681	716	746	755

Fonte: Paes de Camargo, A.; Grohmann, F.; Paes de Camargo, M.B. (1982) Tensiômetro simples de leitura direta. *Pesquisa Agropecuária Brasileira*, *17*(12): 1763-1772.

Pede-se:

(a) Ajustar a função hiperbólica (12d) aos dados observados utilizando as transformações necessárias.

(b) Construir o diagrama de dispersão incluindo a função ajustada.

(c) Obter a assíntota horizontal da função ajustada e interpretá-la.

Nota: Este exercício foi apresentado por Pereira, A.R. & Arruda, H.V. (1987). *Ajuste prático de curvas na pesquisa biológica*. Fundação Cargill, Campinas, SP, pág. 14-15.

15. Os dados que se seguem referem-se a valores simulados de X_i e Y_i $(i = 1, \ldots, 6)$.

X	0	2	3	4	6	9
Y	13,464	9,025	7,389	4,953	4,055	2,718

Construa o diagrama de dispersão e, com base nele, escolha uma das funções apresentadas no 2º exercício. Em seguida, ajuste-a utilizando as transformações necessárias.

16. Em um estudo da calagem para a sucessão batata-triticale-milho, Quaggio, J.A. *et al.* (1985) obtiveram os resultados para teor de cálcio no solo X, em meq/100cm^3 e porcentagem de tubérculos maduros Y, apresentados a seguir.

X	0,1	0,2	0,2	0,4	0,5	0,8	1,0	1,6	3,2
Y	63	79	80	86	88	89	93	93	96

Fonte: Quaggio, J.A.; Ramos, V.J.; Bataglia, O.C.; van Raij, B.; Sakai, M. (1985) Calagem para a sucessão batata-triticale-milho usando calários com diferentes teores de magnésio. *Bragantia*, *44*(1): 391-406.

Pede-se ajustar o modelo

$$Y_i = \theta_0 + \frac{\theta_1}{X_i} + \varepsilon_i$$

a esses dados, utilizando as transformações necessárias.

17. Considere a amostra de 10 pares de valores X_i, Y_i apresentados na Tabela 2.9, os dados relativos aos Exercícios 1 e 2 do item 1.4.1 (pág. 14) e o modelo de regressão linear simples

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

Supondo que $\varepsilon_i \sim N(0, \sigma^2)$, $\text{Cov}(\varepsilon_i, \varepsilon_{i'}) = 0$, $(i \neq i')$ e que X é fixa, pede-se:

(a) Determine as estimativas dos parâmetros.

(b) Faça a análise de variância da regressão.

(c) Teste a hipótese $H_0 : \beta_0 = 0$ contra $H_\text{a} : \beta_0 \neq 0$, a um nível de significância $\gamma = 0,05$.

(d) Teste a hipótese $H_0 : \beta_1 = 0$ contra $H_a : \beta_1 \neq 0$, a um nível de significância $\gamma = 0,05$.

(e) Determine o valor do coeficiente de determinação.

(f) Determine o valor do coeficiente de determinação corrigido.

(g) Construa os intervalos de confiança para $E(Y_i)$ $(i = 1, \ldots, n)$, com um coeficiente de 95% confiança.

(h) Determine a estimativa de Y_h para $X_h = (X_3 + X_4)/2$ (média entre o terceiro e o quarto valores de X) e construa o intervalo de previsão para Y_h com um coeficiente de 95% de confiança.

Tabela 2.9: Valores de X_i e Y_i $(i = 1, \ldots, 10)$.

X	0	1	1	2	3	3	4	5	5	6
Y	3	2	3	5	4	4	7	6	7	9

Fonte: HOFFMAN, R. & VIEIRA, S. (1983). *Análise de Regressão. Uma Introdução à Econometria*. 2ª ed. Ed, Hucitec, São Paulo, pág. 124.

18. Considere o conjunto de dados

X	$-2a$	$-a$	0	a	$2a$
Y	0	0	3	6	6

sendo a uma constante positiva não nula, e o modelo de regressão

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

Supondo que $\varepsilon_i \sim N(0, \sigma^2)$, $\text{Cov}(\varepsilon_i, \varepsilon_{i'}) = 0$, $(i \neq i')$ e que X é fixa, pede-se:

(a) Obter as estimativas de mínimos quadrados $\hat{\beta}_0$ e $\hat{\beta}_1$ de β_0 e β_1.

(b) Obter as estimativas das variâncias e covariâncias de $\hat{\beta}_0$ e $\hat{\beta}_1$.

(c) Obter o quadro da análise de variância e tirar conclusões com um nível de significância 5%.

(d) Testar a hipótese $H_0 : \beta_1 = 0$ a um nível de 5% de significância.

(e) Testar a hipótese $H_0 : \beta_0 = 0$ a um nível de 5% de significância.

(f) Obter os intervalos de confiança para β_0 e β_1 com um coeficiente de 95% confiança.

(g) Obter os intervalos de confiança para $E(Y_i)$, $(i = 1, \ldots, n)$ com um coeficiente de 95% confiança.

(h) Obter os intervalos de previsão para Y, para $X = \frac{X_i + X_{i+1}}{2}$, $(i = 1, \ldots, n-1)$ com um coeficiente de 95% confiança.

19. Considere o seguinte modelo de regressão

$$Y_i = \beta_0 + \beta_1(X_i - X_k) + \varepsilon_i, \qquad (i = 1, \ldots, k, \ldots, n).$$

Supondo que $\varepsilon_i \sim N(0, \sigma^2)$, $\text{Cov}(\varepsilon_i, \varepsilon_{i'}) = 0$, $(i \neq i')$ e que X é fixa, obtenha o esquema do quadro da análise de variância.

20. Considere os dados da Tabela 2.9. Pede-se:

    (a) Fazer o teste para falta de ajuste da regressão linear a um nível de 5% significância.

    (b) Obter a equação da reta e representá-la em um gráfico juntamente com os valores observados.

21. Os dados da Tabela 2.10 são provenientes de um ensaio inteiramente casualizado e referem-se a concentrações de CO_2 aplicadas (X) sobre folhas de trigo a uma temperatura de 35ºC e quantidades de CO_2 absorvido (Y) pelas folhas, em $cm^3/dm^2/h$.

Tabela 2.10: Concentrações de CO_2 aplicado (X) sobre folhas de trigo a uma temperatura de 35ºC e quantidades de CO_2 absorvido (Y) pelas folhas, em $cm^3/dm^2/h$.

X	Y			
75	0,00			
100	0,65	0,50	0,40	
120	1,00			
130	0,95	1,30		
160	1,80	1,80	2,10	
190	2,80			
200	2,50	2,90	2,45	3,05
240	4,30			
250	4,50			

Fonte: MEAD, R. & CURNOV, R.N. (1983). *Statistical Methods in Agriculture and Experimental Biology*. Chapman & Hall, pág. 140.

Pede-se:

(a) Fazer o teste para falta de ajuste da regressão linear a um nível de 5% significância.

(b) Obter a equação da reta e representá-la em um gráfico juntamente com as médias (observadas) para doses.

Tabela 2.11: Doses de radiação gama (X) aplicadas sobre explantes de abacaxi e pesos (Y) dos mesmos, em g, 45 dias após a irradiação

.

X	Y									
0	9,45	10,84	10,12	11,14	10,30	11,04	11,45	12,23	9,46	12,75
30	10,14	10,73	9,02	0,91	1,35	6,89	1,14	8,98	9,18	0,82
40	8,61	5,48	8,88	9,23	6,15	8,86	7,32	7,66	9,63	5,70
50	6,46	5,88	7,14	2,49	8,33	6,93	6,18	4,14	6,75	5,50
60	7,22	5,49	0,45	6,00	5,05	0,15	4,97	3,52	7,07	9,93
70	2,46	4,45	5,04	6,19	4,15	5,49	4,65	2,78	5,98	0,70
80	3,75	5,75	2,94	0,23	2,22	2,65	2,61	4,13	2,80	4,95

Fonte: Márcia Scherer (2002). Indução de mutação visando o melhoramento de abacaxi.

22. Os dados da Tabela 2.11 são provenientes de um ensaio inteiramente casualizado e referem-se a pesos (Y), em g, de explantes de abacaxi, 45 dias após terem recebido diferentes doses de radiação gama (X). Pede-se:

    (a) Fazer o teste para falta de ajuste da regressão linear a um nível de 5% significância.

    (b) Obter a equação da reta e e representá-la em um gráfico juntamente com as médias (observadas) para doses.

23. Os dados da Tabela 2.12 referem-se a produções de ruibarbo para enlatamento, por datas de colheita, de um experimento em blocos ao acaso. Pede-se:

    (a) Fazer a análise de variância para ensaios em blocos ao acaso.

    (b) Fazer o teste para falta de ajuste da regressão linear a um nível de 5% significância (desdobrar Tratamentos em Regressão Linear e Desvios de Regressão).

    (c) Obter a equação da reta para as médias de tratamentos e representá-la em um gráfico juntamente com as médias (observadas) de tratamentos.

Tabela 2.12: Produções de ruibarbo para enlatamento.

	Blocos			
Data de colheita	I	II	III	IV
03/5	21,2	21,4	12,0	17,2
07/5	19,3	17,4	24,5	30,2
11/5	22,8	29,0	18,5	24,5
15/5	26,0	34,0	33,0	30,2
19/5	43,5	37,0	25,1	23,5
23/5	32,1	30,5	35,7	32,3
27/5	33,0	32,2	35,4	35,4

Fonte: MEAD, R. & CURNOV, R.N. (1983) Statistical Methods in Agriculture and Experimental Biology. Chapman & Hall, pág. 145.

Capítulo 3

Regressão Linear Múltipla

3.1 Modelo estatístico - Notação matricial

Tem-se uma regressão linear múltipla quando se admite que a variável resposta (Y) é função de duas ou mais variáveis explicativas (regressoras). O modelo estatístico de uma regressão linear múltipla com k variáveis regressoras $(X_1, X_2, \ldots, X_k)$ é:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i$$

ou, na forma reparametrizada com variáveis centradas:

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

em que $\alpha = \beta_0 + \beta_1 \bar{X}_1 + \beta_2 \bar{X}_2 + \ldots + \beta_k \bar{X}_k$ e $x_{ij} = X_{ij} - \bar{X}_j$, sendo $\bar{X}_j = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_{ij}$, $j = 1, \ldots, k$.

Em notação matricial, o modelo de regressão linear múltipla fica:

$$\boxed{\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon} = \boldsymbol{\mu} + \boldsymbol{\varepsilon},} \tag{3.1}$$

em que $\boldsymbol{Y}$ é o vetor, de dimensões $n \times 1$, da variável aleatória Y; $\boldsymbol{X}$ é a matriz, de dimensões $n \times p$, conhecida do delineamento, e como sempre ocorre em modelos de regressão, a menos de multicolinearidade, é de posto completo $p = k + 1$, sendo $p = k + 1$ o número de parâmetros; $\boldsymbol{\theta}$ é o vetor, de dimensões $p \times 1$, de parâmetros desconhecidos; $\boldsymbol{\varepsilon}$ é o vetor, de dimensões $n \times 1$ e de variáveis aleatórias não observáveis, isto é,

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1k} \\ 1 & X_{21} & X_{22} & \ldots & X_{2k} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & X_{n1} & X_{n2} & \ldots & X_{nk} \end{bmatrix} \quad \text{ou} \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix},$$

$$\boldsymbol{\theta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \cdots \\ \beta_k \end{bmatrix} \quad \text{ou} \quad \boldsymbol{\theta} = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \cdots \\ \beta_k \end{bmatrix} \quad \text{e} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{bmatrix}$$

ou ainda,

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1k} \\ 1 & X_{21} & X_{22} & \ldots & X_{2k} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & X_{n1} & X_{n2} & \ldots & X_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \cdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{bmatrix}$$

$$= \begin{bmatrix} \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \ldots + \beta_k X_{1k} + \varepsilon_1 \\ \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \ldots + \beta_k X_{2k} + \varepsilon_2 \\ \cdots \\ \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \ldots + \beta_k X_{nk} + \varepsilon_n \end{bmatrix}.$$

ou, com variáveis centradas,

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \cdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{bmatrix}$$

$$= \begin{bmatrix} \alpha + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_k x_{1k} + \varepsilon_1 \\ \alpha + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_k x_{2k} + \varepsilon_2 \\ \cdots \\ \alpha + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_k x_{nk} + \varepsilon_n \end{bmatrix}.$$

De forma semelhante ao que foi visto em regressão linear simples, têm-se as suposições:

(i) a variável resposta Y é função linear das variáveis explicativas X_j, $j = 1, 2, \ldots, k$;

(ii) as variáveis explicativas X_j são fixas;

(iii) $\mathrm{E}(\varepsilon_i) = 0$, ou seja, $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, sendo $\mathbf{0}$ um vetor de zeros de dimensões $n \times 1$;

(iv) os erros são homocedásticos, isto é, $\text{Var}(\varepsilon_i) = \text{E}(\varepsilon_i^2) = \sigma^2$;

(v) os erros são independentes, isto é, $\text{Cov}(\varepsilon_i, \varepsilon_{i'}) = \text{E}(\varepsilon_i \varepsilon_{i'}) = 0$, $i \neq i'$;

(vi) os erros têm distribuição normal.

Logo, combinando-se (iv) e (v) tem-se $\text{Var}(\boldsymbol{\varepsilon}) = \text{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \boldsymbol{I}\sigma^2$, sendo $\boldsymbol{I}$ uma matriz identidade, de dimensões $n \times n$. Portanto, considerando-se, também, (vi) tem-se $\boldsymbol{\varepsilon} \sim \text{N}(\boldsymbol{0}, \boldsymbol{I}\sigma^2)$ e $\boldsymbol{Y} \sim \text{N}(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{I}\sigma^2)$, pois, $\text{E}(\boldsymbol{Y}) = \boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\theta}$ e $\text{Var}(\boldsymbol{Y}) = \text{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{I}\sigma^2$. A suposição de normalidade é necessária para a elaboração dos testes de hipóteses e obtenção de intervalos de confiança.

Caso particular: No modelo de regressão linear simples,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i = \alpha + \beta_1 x_i + \varepsilon_i, \tag{3.2}$$

tem-se

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \cdots & \cdots \\ 1 & X_n \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \text{e} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{bmatrix}$$

ou ainda,

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \cdots & \cdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 + \varepsilon_1 \\ \beta_0 + \beta_1 X_2 + \varepsilon_2 \\ \cdots \\ \beta_0 + \beta_1 X_n + \varepsilon_n \end{bmatrix} \tag{3.3}$$

$$= \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \cdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} \alpha + \beta_1 x_1 + \varepsilon_1 \\ \alpha + \beta_1 x_2 + \varepsilon_2 \\ \cdots \\ \alpha + \beta_1 x_n + \varepsilon_n \end{bmatrix}.$$

3.2 Estimação dos parâmetros – Método dos quadrados mínimos

O número de parâmetros a serem estimados é $p = k + 1$. Se existirem apenas p observações, a estimação dos parâmetros reduz-se a um problema matemático de resolução de um sistema de p equações a p incógnitas, não sendo possível fazer qualquer análise estatística. Deve-se, portanto, ter $n > p$.

Um método utilizado para a estimação dos p ($p < n$) parâmetros é o método dos quadrados mínimos. Se o modelo adotado é dado por (3.1), então, o vetor de erros, $\boldsymbol{\varepsilon}$, fica:

$$\boldsymbol{\varepsilon} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{Y} - \boldsymbol{\mu}.$$

De uma forma geral, tem-se que tanto melhor será o modelo quanto menor for o comprimento de $\boldsymbol{\varepsilon}$. Usando a norma Euclideana para o comprimento de $\boldsymbol{\varepsilon}$, tem-se:

$$Z(\boldsymbol{\theta}) = \| \boldsymbol{\varepsilon} \|^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = \sum_{i=1}^{n} \varepsilon_i^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}) = \boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta}$$

Logo,

$$\frac{\partial Z}{\partial \boldsymbol{\theta}} = -2\boldsymbol{X}^T\boldsymbol{Y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta}$$

e fazendo-se $\dfrac{\partial Z}{\partial \boldsymbol{\theta}} = 0$, tem-se:

$$\boxed{\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{X}^T\boldsymbol{Y}} \tag{3.4}$$

que é o sistema de equações normais, em que

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} n & \sum_{i=1}^n X_{i1} & \sum_{i=1}^n X_{i2} & \dots & \sum_{i=1}^n X_{ik} \\ \sum_{i=1}^n X_{i1} & \sum_{i=1}^n X_{i1}^2 & \sum_{i=1}^n X_{i1}X_{i2} & \dots & \sum_{i=1}^n X_{i1}X_{ik} \\ \sum_{i=1}^n X_{i2} & \sum_{i=1}^n X_{i1}X_{i2} & \sum_{i=1}^n X_{i2}^2 & \dots & \sum_{i=1}^n X_{i2}X_{ik} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{i=1}^n X_{ik} & \sum_{i=1}^n X_{i1}X_{ik} & \sum_{i=1}^n X_{i2}X_{ik} & \dots & \sum_{i=1}^n X_{ik}^2 \end{bmatrix} \text{ e } \boldsymbol{X}^T\boldsymbol{Y} = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_{i1}Y_i \\ \sum_{i=1}^n X_{i2}Y_i \\ \dots \\ \sum_{i=1}^n X_{ik}Y_i \end{bmatrix}$$

ou, usando variáveis centradas,

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} n & 0 & 0 & \dots & 0 \\ 0 & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \dots & \sum_{i=1}^n x_{i1}x_{ik} \\ 0 & \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 & \dots & \sum_{i=1}^n x_{i2}x_{ik} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \sum_{i=1}^n x_{i1}x_{ik} & \sum_{i=1}^n x_{i2}x_{ik} & \dots & \sum_{i=1}^n x_{ik}^2 \end{bmatrix} \text{ e } \boldsymbol{X}^T\boldsymbol{Y} = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_{i1}Y_i \\ \sum_{i=1}^n x_{i2}Y_i \\ \dots \\ \sum_{i=1}^n x_{ik}Y_i \end{bmatrix}.$$

Como X tem posto coluna completo, $p = r(X)$, então, o sistema de equações normais é consistente e tem solução única dada por:

$$\boxed{\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}.} \tag{3.5}$$

Caso particular: No modelo de regressão linear simples dado por (2.21), tem-se

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_2 & \dots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \cdots & \cdots \\ 1 & X_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix},$$

$$\begin{aligned} (\boldsymbol{X}^T\boldsymbol{X})^{-1} &= \frac{1}{n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} \begin{bmatrix} \sum_{i=1}^n X_i^2 & -\sum_{i=1}^n X_i \\ -\sum_{i=1}^n X_i & n \end{bmatrix} \\ &= \frac{1}{n\sum_{i=1}^n x_i^2} \begin{bmatrix} \sum_{i=1}^n X_i^2 & -\sum_{i=1}^n X_i \\ -\sum_{i=1}^n X_i & n \end{bmatrix} \end{aligned}$$

pois, como já foi visto, $n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2 = n\sum_{i=1}^n x_i^2$. Também,

$$\boldsymbol{X}^T\boldsymbol{Y} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_2 & \dots & X_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i Y_i \end{bmatrix}.$$

Logo, o sistema de equações normais fica

$$\begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix} \begin{bmatrix} \hat\beta_0 \\ \hat\beta_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i Y_i \end{bmatrix}$$

e, portanto, o estimador de quadrados mínimos de $\boldsymbol{\theta}$ é dado por

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \begin{bmatrix} \hat\beta_0 \\ \hat\beta_1 \end{bmatrix} = \frac{1}{n\sum_{i=1}^n x_i^2} \begin{bmatrix} \sum_{i=1}^n X_i^2 & -\sum_{i=1}^n X_i \\ -\sum_{i=1}^n X_i & n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i Y_i \end{bmatrix} \\ &= \begin{bmatrix} \bar Y - \hat\beta_1 \bar X \\ \\ \dfrac{n\sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n\sum_{i=1}^n x_i^2} \end{bmatrix}, \end{aligned}$$

isto é,

$$\boxed{\hat{\boldsymbol{\theta}} = \begin{bmatrix} \hat\beta_0 \\ \hat\beta_1 \end{bmatrix} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \begin{bmatrix} \bar Y - \hat\beta_1 \bar X \\ \\ \dfrac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2} \end{bmatrix},} \tag{3.6}$$

pois, como já foi visto, $n\sum_{i=1}^n X_iY_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i = n\sum_{i=1}^n x_iY_i$. Para a variável centrada, isto é, pela utilização de $x_i = X_i - \bar{X}$ como variável preditora, tem-se:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i = (\beta_0 - \beta_1\bar{X}) + \beta_1(X_i - \bar{X}) + \varepsilon_i = \alpha + \beta_1 x_i + \varepsilon_i$$

e, portanto, em notação matricial

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \cdots \\ 1 & x_n \end{bmatrix}, \quad \boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} n & 0 \\ 0 & \sum_{i=1}^n x_i^2 \end{bmatrix} \quad \text{e} \quad \boldsymbol{X}^T\boldsymbol{Y} = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_iY_i \end{bmatrix}.$$

Logo,

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_1 \end{bmatrix} = \frac{1}{n\sum_{i=1}^n x_i^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & 0 \\ 0 & n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_iY_i \end{bmatrix},$$

isto é,

$$\boxed{\hat{\boldsymbol{\theta}} = \begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_1 \end{bmatrix} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \begin{bmatrix} \bar{Y} \\ \dfrac{\sum_{i=1}^n x_iY_i}{\sum_{i=1}^n x_i^2} \end{bmatrix}.} \tag{3.7}$$

Propriedades:

(*i*) Os elementos de $\hat{\boldsymbol{\theta}}$ são combinações lineares dos Y_i, isto é,

$$\hat{\theta}_j = \sum_{i=1}^n c_{ij}Y_i$$

em que os c_{ij} são os elementos da matriz resultante de $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

(*ii*) A solução de quadrados mínimos $\hat{\boldsymbol{\theta}}$ é um estimador não viesado para $\boldsymbol{\theta}$, isto é,

$$\begin{aligned} \mathrm{E}(\hat{\boldsymbol{\theta}}) &= \mathrm{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathrm{E}(\boldsymbol{Y}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathrm{E}(\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}) \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{I}\boldsymbol{\theta} = \boldsymbol{\theta}. \end{aligned}$$

(*iii*) A matriz de variâncias e covariâncias de $\hat{\boldsymbol{\theta}}$ é dada por:

$$\begin{aligned} \mathrm{Var}(\hat{\boldsymbol{\theta}}) &= \mathrm{Var}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathrm{Var}(\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{I}\sigma^2\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2 = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2, \end{aligned}$$

pois, $\mathrm{Var}(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}\mathrm{Var}(\boldsymbol{Y})\boldsymbol{A}^T$. Portanto,

$$\boxed{V = \text{Var}(\hat{\boldsymbol{\theta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2.} \tag{3.8}$$

Caso particular: No modelo de regressão linear simples, definido em (3.1) fica:

$$V = \frac{1}{n\sum_{i=1}^n x_i^2}\begin{bmatrix} \sum_{i=1}^n X_i^2 & -\sum_{i=1}^n X_i \\ -\sum_{i=1}^n X_i & n \end{bmatrix}\sigma^2.$$

Mas,

$$\sum_{i=1}^n X_i^2 = \sum_{i=1}^n X_i^2 - 2\frac{(\sum_{i=1}^n X_i)^2}{n} + 2\frac{(\sum_{i=1}^n X_i)^2}{n} = \sum_{i=1}^n X_i^2 - 2\bar{X}\sum_{i=1}^n X_i + n\bar{X}^2 + n\bar{X}^2$$

$$= \sum_{i=1}^n (X_i^2 - 2\bar{X}X_i + \bar{X}^2) + n\bar{X}^2 = \sum_{i=1}^n x_i^2 + n\bar{X}^2.$$

Logo,

$$\boxed{V = \begin{bmatrix} \text{Var}(\hat{\beta}_0) & \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \text{Var}(\hat{\beta}_1) \end{bmatrix} = \begin{bmatrix} \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n x_i^2} & -\frac{\bar{X}}{\sum_{i=1}^n x_i^2} \\ -\frac{\bar{X}}{\sum_{i=1}^n x_i^2} & \frac{1}{\sum_{i=1}^n x_i^2} \end{bmatrix}\sigma^2} \tag{3.9}$$

e, com a variável X centrada, tem-se

$$\boxed{V = \begin{bmatrix} \text{Var}(\hat{\alpha}) & \text{Cov}(\hat{\alpha}, \hat{\beta}_1) \\ \text{Cov}(\hat{\alpha}, \hat{\beta}_1) & \text{Var}(\hat{\beta}_1) \end{bmatrix} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2 = \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum_{i=1}^n x_i^2} \end{bmatrix}\sigma^2.} \tag{3.10}$$

Vê-se, portanto, que os estimadores de quadrados mínimos, $\hat{\alpha}$ e $\hat{\beta}_1$, não são correlacionados, pois $\text{Cov}(\hat{\alpha}, \hat{\beta}_1) = 0$.

(*iv*) Como resultado dos itens (i), (ii) e (iii) e o fato que $\boldsymbol{Y} \sim \text{N}(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{I}\sigma^2)$, tem-se que

$$\hat{\boldsymbol{\theta}} \sim \text{N}(\boldsymbol{\theta}, (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2).$$

(*v*) Dada uma combinação linear dos parâmetros

$$\boldsymbol{c}^T\boldsymbol{\theta} = c_0\beta_0 + c_1\beta_1 + \ldots + c_k\beta_k$$

em que $\boldsymbol{c}^T = [c_0, c_1, \ldots, c_k]$, um estimador de quadrados mínimos, não viesado e de variância mínima é dado por $\boldsymbol{c}^T\hat{\boldsymbol{\theta}}$. Portanto,

$$\mathrm{E}(\boldsymbol{c}^T\hat{\boldsymbol{\theta}}) = \boldsymbol{c}^T\boldsymbol{\theta} \quad \text{e} \quad \mathrm{Var}(\boldsymbol{c}^T\hat{\boldsymbol{\theta}}) = \boldsymbol{c}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c}\,\sigma^2.$$

Além disso,

$$\boldsymbol{c}^T\hat{\boldsymbol{\theta}} \sim \mathrm{N}(\boldsymbol{c}^T\boldsymbol{\theta}, \boldsymbol{c}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c}\,\sigma^2).$$

(*vi*) A aproximação de quadrados mínimos para $\boldsymbol{Y}$ é dada por:

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{Y}$$

sendo $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Portanto,

$$\mathrm{E}(\hat{\boldsymbol{Y}}) = \boldsymbol{X}\boldsymbol{\theta} \quad \text{e} \quad \mathrm{Var}(\hat{\boldsymbol{Y}}) = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2.$$

Além disso,

$$\hat{\boldsymbol{Y}} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2).$$

(*vi*) **Interpretação geométrica**

A matriz $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ é o projetor ortogonal de $\boldsymbol{Y}$ no espaço coluna de $\boldsymbol{X}$, $C(\boldsymbol{X})$ (ver Figura 3.1). Ela é chamada, em inglês, *hat matrix*, e representada por $\boldsymbol{H}$, por ser a matriz que "coloca o chapéu" em $\boldsymbol{Y}$.

Figura 3.1: Projeção ortogonal de Y em $C(X)$

Pelo teorema de Pitágoras, a partir da Figura 3.1 tem-se:

$$|| \boldsymbol{Y} ||^2 = || \hat{\boldsymbol{Y}} ||^2 + || \hat{\boldsymbol{\varepsilon}} ||^2 \tag{3.11}$$

isto é, o vetor de observações pode ser decomposto na soma de dois vetores ortogonais: o vetor $\hat{\boldsymbol{Y}}$ do espaço estimação, e o vetor $\hat{\boldsymbol{\varepsilon}}$ do espaço resíduo. Esta decomposição é o fundamento da análise de variância, que será vista no item 3.5.

3.3 Notação matricial alternativa

Considerando-se variáveis centradas, o vetor de parâmetros pode ser escrito como

$$\boldsymbol{\theta} = \begin{bmatrix} \alpha \\ \boldsymbol{\beta} \end{bmatrix}$$

sendo $\boldsymbol{\beta}^T = \begin{bmatrix} \beta_1 & \beta_2 & \cdots & \beta_k \end{bmatrix}$. Além disso, a matriz $\boldsymbol{X}$ pode ser escrita como:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \boldsymbol{X}_1 \end{bmatrix}.$$

Logo,

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} n & \mathbf{0}^T \\ \mathbf{0} & \boldsymbol{X}_1^T\boldsymbol{X}_1 \end{bmatrix}, \quad (\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix} \frac{1}{n} & \mathbf{0}^T \\ \mathbf{0} & (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1} \end{bmatrix} \quad \text{e} \quad \boldsymbol{X}^T\boldsymbol{Y} = \begin{bmatrix} \mathbf{1}^T\boldsymbol{Y} \\ \boldsymbol{X}_1^T\boldsymbol{Y} \end{bmatrix}$$

e, portanto,

$$\boxed{\hat{\boldsymbol{\theta}} = \begin{bmatrix} \hat{\alpha} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \begin{bmatrix} \bar{Y} \\ (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{Y} \end{bmatrix}.} \tag{3.12}$$

3.4 Análise de variância e teste F

Obtenção das Somas de Quadrados

Por meio da decomposição ortogonal dada pela equação (3.11), fica fácil interpretar a decomposição da soma total de quadrados de desvios. Assim, tem-se:

$$|| \boldsymbol{Y} ||^2 = || \hat{\boldsymbol{Y}} ||^2 + || \hat{\boldsymbol{\varepsilon}} ||^2 = || \boldsymbol{X}\hat{\boldsymbol{\theta}} ||^2 + || \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}} ||^2$$

ou

$$\begin{aligned} \boldsymbol{Y}^T\boldsymbol{Y} &= (\boldsymbol{X}\hat{\boldsymbol{\theta}})^T\boldsymbol{X}\hat{\boldsymbol{\theta}} + (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}) \\ &= \hat{\boldsymbol{\theta}}^T\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\theta}} + \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^T\boldsymbol{X}^T\boldsymbol{Y} + \hat{\boldsymbol{\theta}}^T\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\theta}} \end{aligned}$$

e usando-se a expressão do sistema de equações normais dado por (3.4) tem-se que:

$$\boldsymbol{Y}^T\boldsymbol{Y} = \hat{\boldsymbol{\theta}}^T\boldsymbol{X}^T\boldsymbol{Y} + (\boldsymbol{Y}^T\boldsymbol{Y} - \hat{\boldsymbol{\theta}}^T\boldsymbol{X}^T\boldsymbol{Y})$$

e substituindo-se $\hat{\boldsymbol{\theta}}$ pela expressão (3.5), fica

$$\boldsymbol{Y}^T\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{Y}^T[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{H}\boldsymbol{Y} + \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$$

em que

$\boldsymbol{Y}^T\boldsymbol{Y} = \sum_{i=1}^n Y_i^2 = SQTotal$ (não corrigida);
$\boldsymbol{Y}^T\boldsymbol{H}\boldsymbol{Y} = \hat{\boldsymbol{\theta}}^T\boldsymbol{X}^T\boldsymbol{Y} = SQParâmetros = SQP$ e
$\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{Y} - \hat{\boldsymbol{\theta}}^T\boldsymbol{X}^T\boldsymbol{Y} = SQRes.$

Logo,

$$SQTotal = SQP + SQRes.$$

Usando-se a notação alternativa apresentada no item (3.3), tem-se,

$$SQP = \hat{\boldsymbol{\theta}}^T\boldsymbol{X}^T\boldsymbol{Y} = \begin{bmatrix} \hat{\alpha} & \hat{\boldsymbol{\beta}}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{1}^T\boldsymbol{Y} \\ \boldsymbol{X}_1^T\boldsymbol{Y} \end{bmatrix} = \hat{\alpha}\boldsymbol{1}^T\boldsymbol{Y} + \hat{\boldsymbol{\beta}}^T\boldsymbol{X}_1^T\boldsymbol{Y}$$

$$= \frac{1}{n}\boldsymbol{Y}^T\boldsymbol{1}\boldsymbol{1}^T\boldsymbol{Y} + \hat{\boldsymbol{\beta}}^T\boldsymbol{X}_1^T\boldsymbol{Y} = \frac{1}{n}\boldsymbol{Y}^T\boldsymbol{J}\boldsymbol{Y} + \hat{\boldsymbol{\beta}}^T\boldsymbol{X}_1^T\boldsymbol{Y}$$

que na notação $R(.)$ é representada por

$$SQP = R(\alpha, \boldsymbol{\beta}) = R(\alpha) + R(\boldsymbol{\beta}|\alpha) = \text{Correção} + SQReg$$

sendo $R(\alpha, \boldsymbol{\beta})$, a redução na soma de quadrados residual devido à inclusão dos parâmetros $\alpha, \beta_1, \beta_2, \ldots, \beta_k$; $R(\alpha)$, a redução na soma de quadrados residual devido à inclusão do parâmetro α e $R(\boldsymbol{\beta}|\alpha)$, a redução na soma de quadrados residual devido à inclusão dos parâmetros $\beta_1, \beta_2, \ldots, \beta_k$ dado que α já estava no modelo. Logo,

$SQReg = SQP - \text{Correção}$ e
$SQTotal = SQTotal(\text{não corrigida}) - \text{Correção} = SQReg + SQRes.$

Caso particular: No modelo de regressão linear simples em que

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i = \alpha + \beta_1(X_i - \bar{X}) + \varepsilon_i,$$

tem-se:

$$SQP = \begin{bmatrix} \hat{\beta}_0 & \hat{\beta}_1 \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i Y_i \end{bmatrix} = \begin{bmatrix} \hat{\alpha} & \hat{\beta}_1 \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_i Y_i \end{bmatrix}$$

$$= \hat{\alpha}\sum_{i=1}^n Y_i + \hat{\beta}_1\sum_{i=1}^n x_i Y_i = SQ(\alpha) + SQ(\beta_1|\alpha) = \frac{(\sum_{i=1}^n Y_i)^2}{n} + \frac{(\sum_{i=1}^n x_i Y_i)^2}{\sum_{i=1}^n x_i^2}$$

em que

$$SQ(\alpha) = \frac{(\sum_{i=1}^n Y_i)^2}{n} = C \text{ (correção) e}$$

$$SQ(\beta_1|\alpha) = \frac{(\sum_{i=1}^n x_i Y_i)^2}{\sum_{i=1}^n x_i^2} = SQReg.$$

Número de graus de liberdade associado às Somas de Quadrados

O número de graus de liberdade associados à uma soma de quadrados é dado pela característica da matriz idempotente de sua forma quadrática, isto é, o número de graus de liberdade de $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}$ é dado pela característica($\boldsymbol{A}$). Além disso, para $\boldsymbol{A}$ uma matriz idempotente, tem-se que característica($\boldsymbol{A}$) = traço($\boldsymbol{A}$). Vale lembrar, também, que para c uma constante e $\boldsymbol{A}$ e $\boldsymbol{B}$ matrizes, tem-se:

(i) traço($\boldsymbol{A}$) = traço($\boldsymbol{A}^T$);

(ii) traço($c\boldsymbol{A}$) = c traço($\boldsymbol{A}$);

(iii) traço($\boldsymbol{A}+\boldsymbol{B}$) =traço($\boldsymbol{A}$) + traço($\boldsymbol{B}$);

(iv) traço($\boldsymbol{A}\boldsymbol{B}$) =traço($\boldsymbol{B}\boldsymbol{A}$).

Além disso, tem-se que:
$SQTotal = \boldsymbol{Y}^T\boldsymbol{Y} - \frac{1}{n}\boldsymbol{Y}^T\boldsymbol{1}\boldsymbol{1}^T\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{Y} - \frac{1}{n}\boldsymbol{Y}^T\boldsymbol{J}\boldsymbol{Y} = \boldsymbol{Y}^T(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y}$,
$SQP = \hat{\boldsymbol{\theta}}^T\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{H}\boldsymbol{Y}$,
$SQReg = SQP - \frac{1}{n}\boldsymbol{Y}^T\boldsymbol{1}\boldsymbol{1}^T\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{H}\boldsymbol{Y} - \frac{1}{n}\boldsymbol{Y}^T\boldsymbol{J}\boldsymbol{Y} = \boldsymbol{Y}^T(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y}$ e
$SQRes = \boldsymbol{Y}^T\boldsymbol{Y} - \hat{\boldsymbol{\theta}}^T\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$,

sendo $\frac{1}{n}\boldsymbol{J}$, $\boldsymbol{H}$ e $\boldsymbol{I} - \boldsymbol{H}$ matrizes simétricas e idempotentes. Também, $\boldsymbol{H} - \frac{1}{n}\boldsymbol{J}$ é simétrica e para mostrar que é idempotente, considere a forma alternativa apresentada no item (3.3). Logo,

$$\boldsymbol{J}\boldsymbol{H} = \boldsymbol{J}\begin{bmatrix} \boldsymbol{1} & \boldsymbol{X}_1 \end{bmatrix}\begin{bmatrix} \frac{1}{n} & \boldsymbol{0}^T \\ \boldsymbol{0} & (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1} \end{bmatrix}\begin{bmatrix} \boldsymbol{1}^T \\ \boldsymbol{X}_1^T \end{bmatrix}$$

$$= \begin{bmatrix} n & 0 & \dots & 0 \\ n & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ n & 0 & \dots & 0 \end{bmatrix}\begin{bmatrix} \frac{1}{n} & \boldsymbol{0}^T \\ \boldsymbol{0} & (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1} \end{bmatrix}\begin{bmatrix} \boldsymbol{1}^T \\ \boldsymbol{X}_1^T \end{bmatrix} = \begin{bmatrix} 1 & \boldsymbol{0}^T \\ 1 & \boldsymbol{0}^T \\ \dots & \dots \\ 1 & \boldsymbol{0}^T \end{bmatrix}\begin{bmatrix} \boldsymbol{1}^T \\ \boldsymbol{X}_1^T \end{bmatrix} = \boldsymbol{J}$$

e, portanto, $(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J})(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J}) = \boldsymbol{H} - \frac{1}{n}\boldsymbol{H}\boldsymbol{J} - \frac{1}{n}\boldsymbol{J}\boldsymbol{H} + \frac{1}{n}\boldsymbol{H}\boldsymbol{J} = (\boldsymbol{H} - \frac{1}{n}\boldsymbol{J})$, isto é, $\boldsymbol{H} - \frac{1}{n}\boldsymbol{J}$ é idempotente. Pode-se verificar, então, que estão associados, respectivamente, $n-1$, p, $p-1$ e $n-p$ graus de liberdade às $SQTotal$, SQP, $SQReg$ e $SQRes$.

Valor esperado das Somas de Quadrados e Quadrados Médios

É conveniente lembrar que a esperança da forma quadrática $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}$ é dada por:

$$\mathrm{E}(\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}) = \text{traço }(\boldsymbol{A})\sigma^2 + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}$$

em que, $\boldsymbol{A}$ é uma matriz simétrica, conhecida de dimensões $n \times n$ e $\boldsymbol{\mu} = E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\theta}$.

Então,

$$\begin{aligned}
\mathrm{E}(SQTotal) &= \mathrm{E}[\boldsymbol{Y}^T(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y}] \\
&= [\mathrm{tr}(\boldsymbol{I}) - \frac{1}{n}\mathrm{tr}(\boldsymbol{J})]\sigma^2 + \boldsymbol{\theta}^T\boldsymbol{X}^T(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{X}\boldsymbol{\theta} \\
&= (n-1)\sigma^2 + \boldsymbol{\theta}^T\boldsymbol{X}^T(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{X}\boldsymbol{\theta}
\end{aligned}$$

em que

$$\boldsymbol{I} - \frac{1}{n}\boldsymbol{J} = \frac{1}{n}\begin{bmatrix} n-1 & -1 & \cdots & -1 \\ -1 & n-1 & \cdots & -1 \\ \cdots & \cdots & \cdots & \cdots \\ -1 & -1 & \cdots & n-1 \end{bmatrix},$$

$$\begin{aligned}
\boldsymbol{X}^T(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{X} &= \frac{1}{n}\begin{bmatrix} \mathbf{1}^T \\ \boldsymbol{X}_1^T \end{bmatrix}\begin{bmatrix} n-1 & -1 & \cdots & -1 \\ -1 & n-1 & \cdots & -1 \\ \cdots & \cdots & \cdots & \cdots \\ -1 & -1 & \cdots & n-1 \end{bmatrix}\begin{bmatrix} \mathbf{1} & \boldsymbol{X}_1 \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{0}^T \\ \boldsymbol{X}_1^T \end{bmatrix}\begin{bmatrix} \mathbf{1} & \boldsymbol{X}_1 \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \boldsymbol{X}_1^T\boldsymbol{X}_1 \end{bmatrix}
\end{aligned}$$

e

$$\boldsymbol{\theta}^T\boldsymbol{X}^T(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{X}\boldsymbol{\theta} = \begin{bmatrix} \alpha & \boldsymbol{\beta}^T \end{bmatrix}\begin{bmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \boldsymbol{X}_1^T\boldsymbol{X}_1 \end{bmatrix}\begin{bmatrix} \alpha \\ \boldsymbol{\beta} \end{bmatrix} = \boldsymbol{\beta}^T\boldsymbol{X}_1^T\boldsymbol{X}_1\boldsymbol{\beta}.$$

Portanto,

$$\boxed{\mathrm{E}(SQTotal) = (n-1)\sigma^2 + \boldsymbol{\theta}^T\boldsymbol{X}^T(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{X}\boldsymbol{\theta} = (n-1)\sigma^2 + \boldsymbol{\beta}^T\boldsymbol{X}_1^T\boldsymbol{X}_1\boldsymbol{\beta},} \qquad (3.13)$$

$$\mathrm{E}(SQRes) = \mathrm{E}[\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}] = \mathrm{tr}(\boldsymbol{I}-\boldsymbol{H})\sigma^2 + \boldsymbol{\theta}^T\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{X}\boldsymbol{\theta}$$

mas,

$$\mathrm{tr}(\boldsymbol{H}) = \mathrm{tr}[\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T] = r(\boldsymbol{X}) = p$$

e

$$\boldsymbol{\theta}^T\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta} = 0.$$

Logo,

$$\boxed{\mathrm{E}(SQRes) = (n-p)\sigma^2} \tag{3.14}$$

e

$$\begin{aligned}\mathrm{E}(SQReg) &= \mathrm{E}(SQTotal) - \mathrm{E}(SQRes)\\ &= (n-1)\sigma^2 + \boldsymbol{\theta}^T\boldsymbol{X}^T(\boldsymbol{I}-\frac{1}{n}\boldsymbol{J})\boldsymbol{X}\boldsymbol{\theta} - (n-p)\sigma^2 = (p-1)\sigma^2 + \boldsymbol{\theta}^T\boldsymbol{X}^T(\boldsymbol{I}-\frac{1}{n}\boldsymbol{J})\boldsymbol{X}\boldsymbol{\theta}\end{aligned}$$

$$\boxed{\mathrm{E}(SQReg) = (p-1)\sigma^2 + \boldsymbol{\theta}^T\boldsymbol{X}^T(\boldsymbol{I}-\frac{1}{n}\boldsymbol{J})\boldsymbol{X}\boldsymbol{\theta} = (p-1)\sigma^2 + \boldsymbol{\beta}^T\boldsymbol{X}_1^T\boldsymbol{X}_1\boldsymbol{\beta}.} \tag{3.15}$$

Por definição Quadrado Médio é o quociente das Somas de Quadrados pelos respectivos números de graus de liberdade, isto é,

$$QMReg = \frac{SQReg}{p-1} \quad \text{e}$$

$$QMRes = \frac{SQRes}{n-p}.$$

Logo,

$$\mathrm{E}(QMReg) = \sigma^2 + \frac{1}{p-1}\boldsymbol{\beta}^T\boldsymbol{X}_1^T\boldsymbol{X}_1\boldsymbol{\beta} \quad \text{e}$$

$$\mathrm{E}(QMRes) = \sigma^2.$$

Caso particular: No modelo de regressão linear simples dado pela equação (2.21), $p = r(X) = 2$, $\boldsymbol{\beta} = \beta_1$ e $\boldsymbol{X}_1^T\boldsymbol{X}_1 = \sum_{i=1}^n x_i^2$. Portanto, tem-se

$$\mathrm{E}(SQTotal) = (n-1)\sigma^2 + \beta_1^2\sum_{i=1}^n x_i^2$$

$$\mathrm{E}(SQReg) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} x_i^2 \quad \text{e}$$

$$\mathrm{E}(SQRes) = (n-2)\sigma^2.$$

Estimador da variância residual

Dado que

$$\mathrm{E}(QMRes) = \mathrm{E}\left(\frac{SQRes}{n-p}\right) = \sigma^2,$$

então, um estimador não viesado para σ^2

$$\hat{\sigma}^2 = \frac{SQRes}{n-p} = QMRes.$$

Têm-se, então, a partir de (3.9) e (3.10), as matrizes de variâncias e covariâncias estimadas, substituindo-se σ^2 por $QMRes$.

Independência de $\hat{\boldsymbol{\theta}}$ e $\hat{\sigma}^2$

Lembrando que se $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{V})$, então, uma forma quadrática $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}$ e uma forma linear $\boldsymbol{B}\boldsymbol{Y}$ são independentes se e só se $\boldsymbol{B}\boldsymbol{V}\boldsymbol{A} = \boldsymbol{0}$. Dado que $\boldsymbol{V} = \boldsymbol{I}\sigma^2$,

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \Rightarrow \boldsymbol{B} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

e

$$SQRes = \boldsymbol{Y}^T[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{Y} \Rightarrow \boldsymbol{A} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

tem-se

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{I}[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\sigma^2 = \boldsymbol{0}.$$

Portanto, $\hat{\boldsymbol{\theta}}$ e $\hat{\sigma}^2$ são independentes.

Distribuição das Somas de Quadrados

Lembrando que se $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{V})$, então, a forma quadrática $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}$ tem uma distribuição χ^2 não central com $r(A)$ graus de liberdade e parâmetro de não centralidade $\delta = \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}$, isto é,

$$\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \sim \chi^2_{[r(A),\delta]},$$

se e só se $\boldsymbol{AV}$ é idempotente. Se $\delta = \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu} = 0$, então a distribuição é uma χ^2 central com $r(A)$ graus de liberdade, isto é,

$$\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \sim \chi^2_{[r(A)]}.$$

Para o modelo linear dado pela equação (3.1) tem-se $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\theta}$, $\boldsymbol{V} = \boldsymbol{I}\sigma^2$ e que as matrizes $\frac{1}{n}\boldsymbol{J}$, $\boldsymbol{H} = \boldsymbol{J}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, $\boldsymbol{I} - \boldsymbol{H}$ e $\boldsymbol{H} - \frac{1}{n}\boldsymbol{J}$ são simétricas e idempotentes. Então,

$$\frac{1}{\sigma^2}SQTotal = \frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y} \sim \chi^2_{[n-1,\delta]}$$

$$\frac{1}{\sigma^2}SQReg = \frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y} \sim \chi^2_{[p-1,\delta]}$$

e

$$\frac{1}{\sigma^2}SQRes = \frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \sim \chi^2_{(n-p)}$$

para $\delta = \frac{1}{2\sigma^2}\boldsymbol{\beta}^T\boldsymbol{X}_1^T\boldsymbol{X}_1\boldsymbol{\beta}$, pois

$$\begin{aligned}\boldsymbol{\theta}^T\boldsymbol{X}^T(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{X}\boldsymbol{\theta} &= \boldsymbol{\theta}^T\boldsymbol{X}^T(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J})\boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{\theta}^T(\boldsymbol{X}^T\boldsymbol{X} - \frac{1}{n}\boldsymbol{X}^T\boldsymbol{J}\boldsymbol{X})\boldsymbol{\theta} \\ &= \begin{bmatrix} \alpha & \boldsymbol{\beta}^T \end{bmatrix}\begin{bmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \boldsymbol{X}_1^T\boldsymbol{X}_1 \end{bmatrix}\begin{bmatrix} \alpha \\ \boldsymbol{\beta} \end{bmatrix} = \boldsymbol{\beta}^T\boldsymbol{X}_1^T\boldsymbol{X}_1\boldsymbol{\beta}.\end{aligned}$$

Caso particular: No modelo de regressão linear simples dado pela equação (2.21), $p = r(X) = 2$, $\boldsymbol{\beta} = \beta_1$ e $\boldsymbol{X}_1^T\boldsymbol{X}_1 = \sum_{i=1}^n x_i^2$. Portanto, para $\delta = \frac{1}{2\sigma^2}\beta_1^2\sum_{i=1}^n x_i^2$

$$\frac{1}{\sigma^2}SQTotal = \frac{1}{\sigma^2}\sum_{i=1}^n(Y_i - \bar{Y})^2 = \frac{1}{\sigma^2}\sum_{i=1}^n y_i^2 \sim \chi^2_{(n-1,\delta)}$$

$$\frac{1}{\sigma^2}SQReg = \frac{1}{\sigma^2}\sum_{i=1}^n(\hat{Y}_i - \bar{Y})^2 \sim \chi^2_{(1,\delta)}$$

e

$$\frac{1}{\sigma^2}SQRes = \frac{1}{\sigma^2}\sum_{i=1}^n(Y_i - \hat{Y}_i)^2 \sim \chi^2_{(n-2)}.$$

Independência das $SQReg$ e $SQRes$

Lembrando que se $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{V})$, então, as formas quadráticas $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}$ e $\boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{Y}$ são distribuídas independentemente se $\boldsymbol{A}\boldsymbol{V}\boldsymbol{B} = \boldsymbol{0}$ (ou, equivalentemente se $\boldsymbol{B}\boldsymbol{V}\boldsymbol{A} = \boldsymbol{0}$). Dado que

$$SQReg = \boldsymbol{Y}^T(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y} \quad \text{e} \quad SQRes = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$$

então,

$$(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J})\boldsymbol{I}(\boldsymbol{I} - \boldsymbol{H})\sigma^2 = (\boldsymbol{H} - \boldsymbol{H}\boldsymbol{H} - \frac{1}{n}\boldsymbol{J} + \frac{1}{n}\boldsymbol{J}\boldsymbol{H})\sigma^2 = \boldsymbol{0}$$

mostrando que $SQReg$ e $SQRes$ são independentes.

Distribuição do quociente $\dfrac{QMReg}{QMRes}$

Em geral, as hipóteses de interesse são sobre os parâmetros $\boldsymbol{\beta}^T = \begin{bmatrix} \beta_1 & \beta_2 & \cdots & \beta_k \end{bmatrix}$, mantendo-se o parâmetro β_0 (ou α) no modelo, isto é, deseja-se testar a hipótese $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ (correspondendo ao modelo $\boldsymbol{Y} = \alpha + \boldsymbol{\varepsilon}$) versus H_a : não H_0. Isso equivale a verificar se realmente existe uma relação linear entre Y e X_1, X_2, ..., X_k. Já foi visto que:

$$\frac{1}{\sigma^2}SQRes \sim \chi^2_{n-p} \quad \text{e} \quad \frac{1}{\sigma^2}SQReg \sim \chi^2_{p-1,\delta}$$

em que $\delta = \dfrac{1}{2\sigma^2}\boldsymbol{\beta}^T\boldsymbol{X}_1^T\boldsymbol{X}_1\boldsymbol{\beta}$ é o parâmetro de não centralidade, e, além disso, são independentes. Logo, sob $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$, tem-se que $\delta = 0$,

$$\frac{1}{\sigma^2}SQReg \sim \chi^2_{p-1} \quad \text{(central)}$$

e

$$F = \frac{\dfrac{SQReg}{(p-1)\sigma^2}}{\dfrac{SQRes}{(n-p)\sigma^2}} \sim F_{p-1,n-p}.$$

Portanto, rejeita-se a hipótese $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$, a um nível de $100\gamma\%$ de significância, se

$$F_{calc} = \frac{QMReg}{QMRes} > F_{p-1,n-p;\gamma}$$

ou se $P(F_{1,n-2} > F_{calc}) < \gamma$, em que, em geral, $\gamma = 0,05$ ou $\gamma = 0,01$.

Quadro da análise da variância e teste F

A partir dos resultados obtidos, pode-se obter o esquema do quadro da análise da variância e teste F mostrados na Tabela 13. Essa forma de apresentação da análise é chamada, por alguns autores, de *análise parcial*.

A soma de quadrados de regressão é a redução na soma de quadrados residual devido à inclusão dos parâmetros $\beta_1, \beta_2, \ldots, \beta_k$ (ou das variáveis $X_1, X_2, \ldots, X_k$) no modelo $Y = \beta_0 + \varepsilon$, isto é, $SQReg = R(\beta_1, \beta_2, \ldots, \beta_k|\beta_0) = R(\beta_0, \beta_1, \beta_2, \ldots, \beta_k) - R(\beta_0)$, sendo $R(\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$ a soma de quadrados de resíduos para o modelo $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$ e $R(\beta_0)$ a soma de quadrados de resíduos para o modelo $Y = \beta_0 + \varepsilon$.

Tabela 3.1: Esquema de análise de variância e teste F

Causas de variação	G.L.	S.Q.	Q.M.	E(Q.M.)	F
Regressão	$k = p - 1$	$\hat{\boldsymbol{\theta}}^T \boldsymbol{X}^T \boldsymbol{Y} - C$	$\frac{SQReg}{p-1}$	$\sigma^2 + \frac{1}{p-1} \boldsymbol{\beta}^T \boldsymbol{X}_1^T \boldsymbol{X}_1 \boldsymbol{\beta}$	$F = \frac{QMReg}{QMRes}$
Resíduo	$n - p$	por diferença	$\frac{SQRes}{n-p}$	σ^2	
Total	$n - 1$	$\sum_{i=1}^n Y_i^2 - C$			

$C = \frac{(\sum_{i=1}^n Y_i)^2}{n}$.

Outro tipo de análise é a *análise sequencial*, em que os parâmetros vão sendo adicionados ao modelo sequencialmente. Considerando-se, por exemplo, uma variável resposta Y e duas variáveis X_1 e X_2 regressoras, o modelo completo poderia ser

$$\mathrm{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

e o quadro de Análise de Variância sequencial poderia ser o apresentado na Tabela 14, sendo $R(\beta_1|\beta_0)$ a redução na soma de quadrados residual devido à inclusão do parâmetro β_1 (ou da variável X_1) no modelo $Y = \beta_0 + \varepsilon$; $R(\beta_2|\beta_0, \beta_1)$ a redução na soma de quadrados residual devido à inclusão do parâmetro β_2 (ou da variável X_2) no modelo $Y = \beta_0 + \beta_1 X_1 + \varepsilon$

Tabela 3.2: Esquema de análise de variância sequencial e teste F

Causas de variação	G.L.	S.Q.	Q.M.	F
β_0	1	$R(\beta_0)$		
$\beta_1\|\beta_0$	1	$R(\beta_1\|\beta_0) = R(\beta_0, \beta_1) - R(\beta_0)$		F_1
$\beta_2\|\beta_1, \beta_0$	1	$R(\beta_2\|\beta_0, \beta_1) = R(\beta_0, \beta_1, \beta_2) - R(\beta_0, \beta_1)$		F_2
Parâmetros	3	$R(\boldsymbol{\theta})$		
Resíduo	$n - 3$	por diferença	$\frac{SQRes}{n-p}$	
Total	n	$\boldsymbol{Y}^T \boldsymbol{Y}$		

Nesse caso, as hipóteses testadas pelo teste F são de acordo com a ordem estabelecida no quadro de análise de variância, isto é,

(i) F_1 para testar $H_0 : \beta_1 = 0$ ignorando X_2

(i) F_2 para testar $H_0 : \beta_2 = 0$ ajustado para X_1.

Testes de Hipóteses

Seja $\boldsymbol{c}^T\boldsymbol{\theta}$ uma combinação linear dos parâmetros e seu estimador $\boldsymbol{c}^T\hat{\boldsymbol{\theta}}$. Então, do teorema de Gauss-Markov,

$$\boldsymbol{c}^T\hat{\boldsymbol{\theta}} \sim \mathrm{N}(\boldsymbol{c}^T\boldsymbol{\theta}, \boldsymbol{c}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c}\ \sigma^2)$$

e, portanto,

$$Z = \frac{\boldsymbol{c}^T\hat{\boldsymbol{\theta}} - \boldsymbol{c}^T\boldsymbol{\theta}}{\sigma\sqrt{\boldsymbol{c}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c}}} \sim \mathrm{N}(0,1)$$

pois, $\mathrm{E}(Z) = 0$ e $\mathrm{Var}(Z) = 1$. Por outro lado,

$$W = (n-p)\frac{QMRes}{\sigma^2} \sim \chi^2_{n-p}.$$

Logo,

$$\frac{Z}{\sqrt{\frac{W}{n-p}}} = \frac{\boldsymbol{c}^T\hat{\boldsymbol{\theta}} - \boldsymbol{c}^T\boldsymbol{\theta}}{\sqrt{\boldsymbol{c}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c}\ QMRes}} \sim t_{n-p}.$$

Portanto, rejeita-se, a um nível $100\gamma\%$ de significância, a hipótese $H_0 : \boldsymbol{c}^T\boldsymbol{\theta} = \boldsymbol{c}^T\boldsymbol{\theta}_0$ em favor de

$H_{a_1} : \boldsymbol{c}^T\boldsymbol{\theta} < \boldsymbol{c}^T\boldsymbol{\theta}_0$ se $t_{calc} < -t_{n-p;\gamma}$ ou se $P(t_{n-p} < t_{calc}) < \gamma$;

$H_{a_2} : \boldsymbol{c}^T\boldsymbol{\theta} > \boldsymbol{c}^T\boldsymbol{\theta}_0$ se $t_{calc} > t_{n-p;\gamma}$ ou se $P(t_{n-p} > t_{calc}) < \gamma$;

$H_{a_3} : \boldsymbol{c}^T\boldsymbol{\theta} \neq \boldsymbol{c}^T\boldsymbol{\theta}_0$ se $|t_{calc}| > t_{n-p;\frac{\gamma}{2}}$ ou se $P(|t_{n-p}| > |t_{calc}|) < \gamma$;

sendo $t_{calc} = \dfrac{\boldsymbol{c}^T\hat{\boldsymbol{\theta}} - \boldsymbol{c}^T\boldsymbol{\theta}_0}{\sqrt{\boldsymbol{c}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c}\ QMRes}}$.

Assim, por exemplo, no teste da hipótese $H_0 : \beta_1 = 0$ ignorando X_2 tem-se $\boldsymbol{c}^T = (0,1,0)$ e $t_{calc} = \dfrac{\hat{\beta}_1}{s(\hat{\beta}_1)}$, sendo $\hat{\beta}_1$ calculado para o modelo ignorando X_2. No teste da hipótese $H_0 : \beta_1 = \beta_2$ que equivale a $H_0 : \beta_1 - \beta_2 = 0$, tem-se $\boldsymbol{c}^T = (0,1,-1)$.

Princípio do Resíduo Condicional

Testes de hipóteses mais gerais podem ser feitos por meio da comparação de modelos e do uso do *Princípio do Resíduo Condicional*. Seja o *modelo completo* (*modelo c*) dado por

$$\boxed{Y_i = \beta_0 + \beta_1 W_{i1} + \beta_2 W_{i2} + \ldots + \beta_k W_{ik} + \varepsilon_i \quad \text{(modelo c)}}$$

sendo $\mathrm{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \mathrm{E}(\varepsilon_i^2) = \sigma^2$, $\mathrm{E}(\varepsilon_i \varepsilon_{i'}) = 0$, $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ e W_{ij}, funções de X_{ij}. Assim, por exemplo,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

ou

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \beta_4 X_{i2}^2 + \beta_5 X_{i1} X_{i2} + \varepsilon_i.$$

Matricialmente, tem-se:

$$\boldsymbol{Y} = \boldsymbol{W}_c \boldsymbol{\theta}_c + \boldsymbol{\varepsilon}_c$$

em que $\boldsymbol{Y}$ é o vetor das observações, de dimensões $n \times 1$; $\boldsymbol{W}_c$ é a matriz do modelo, de dimensões $n \times p$, sendo $p = k+1$ o número de parâmetros; $\boldsymbol{\theta}_c$ é o vetor, de dimensões $p \times 1$, de parâmetros desconhecidos e $\boldsymbol{\varepsilon}_c$ é o vetor, de dimensões $n \times 1$, de variáveis aleatórias não observáveis.

Pelo método dos quadrados mínimos, tem-se que:

$\hat{\boldsymbol{\theta}}_c = (\boldsymbol{W}_c^T \boldsymbol{W}_c)^{-1} \boldsymbol{W}_c^T \boldsymbol{Y}$,

$\hat{\boldsymbol{Y}}_c = \boldsymbol{W}_c \hat{\boldsymbol{\theta}}_c$,

$SQP_c = SQP(\beta_0, \beta_1, \ldots, \beta_k) = \hat{\boldsymbol{\theta}}_c^T \boldsymbol{W}_c^T \boldsymbol{Y}$, com $p = k+1$ graus de liberdade,

$SQTotal$(não corrigida) $= \boldsymbol{Y}^T \boldsymbol{Y}$, com n graus de liberdade,

$SQRes_c = \boldsymbol{Y}^T \boldsymbol{Y} - \hat{\boldsymbol{\theta}}_c^T \boldsymbol{W}_c^T \boldsymbol{Y}$, com $n - p$ graus de liberdade.

Suponha que se deseja testar a hipótese:

$$H_0 : \beta_{m+1} = \beta_{m+2} = \ldots = \beta_k = 0 \quad (m < k).$$

Sob H_0, tem-se o *modelo reduzido* (*modelo r*)

$$\boxed{Y_i = \beta_0 + \beta_1 W_{i1} + \beta_2 W_{i2} + \ldots + \beta_m W_{im} + \varepsilon_i \quad \text{(modelo r)}}$$

com representação matricial dada por

$$\boldsymbol{Y} = \boldsymbol{W}_r \boldsymbol{\theta}_r + \boldsymbol{\varepsilon}_r$$

em que $\boldsymbol{W}_r$ é a matriz do modelo, de dimensões $n \times r$, sendo $r = m+1$ o número de parâmetros; $\boldsymbol{\theta}_r$ é o vetor, de dimensões $r \times 1$, de parâmetros desconhecidos e $\boldsymbol{\varepsilon}_r$ é o vetor, de dimensões $n \times 1$, dos erros.

Pelo método dos quadrados mínimos, tem-se que:

$\hat{\boldsymbol{\theta}}_r = (\boldsymbol{W}_r^T \boldsymbol{W}_r)^{-1} \boldsymbol{W}_r^T \boldsymbol{Y}$,

$\hat{\boldsymbol{Y}}_r = \boldsymbol{W}_r\hat{\boldsymbol{\theta}}_r$,
$SQP_r = SQP(\beta_0, \beta_1, \ldots, \beta_m) = \hat{\boldsymbol{\theta}}_r^T\boldsymbol{W}_r^T\boldsymbol{Y}$, com $r = m+1$ graus de liberdade,
$SQTotal$(não corrigida) $= \boldsymbol{Y}^T\boldsymbol{Y}$, com n graus de liberdade,
$SQRes_r = \boldsymbol{Y}^T\boldsymbol{Y} - \hat{\boldsymbol{\theta}}_r^T\boldsymbol{W}_r^T\boldsymbol{Y}$, com $n-r$ graus de liberdade.

Logo, a redução da SQResíduo devido à adição dos parâmetros $\beta_{m+1}, \beta_{m+2}, \ldots, \beta_k$ ao *modelo* r é dada por

$$R(\beta_{m+1}, \beta_{m+2}, \ldots, \beta_k|\beta_0, \beta_1, \ldots, \beta_m) = SQRes_r - SQRes_c = SQP_c - SQP_r$$

com $(p-r)$ graus de liberdade, o que é conhecido como *Princípio do Resíduo Condicional*.

Assim satisfeitas as pressuposições estabelecidas para o modelo, tem-se

$$F = \frac{\dfrac{R(\beta_{m+1}, \beta_{m+2}, \ldots, \beta_k|\beta_0, \beta_1, \ldots, \beta_m)}{p-r}}{\dfrac{SQRes_c}{n-p}} = \frac{\dfrac{SQRes_r - SQRes_c}{p-r}}{QMRes_c} \sim F_{p-r,n-p}$$

e rejeita-se H_0, a um nível de $100\gamma\%$ de significância, se $F > F_{p-r,n-p;\gamma}$.

Exemplo 1: Seja o caso particular em que $W_{ij} = X_{ij}$, isto é,

$$\boxed{Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i \quad \text{(modelo c)}} \tag{3.16}$$

e a hipótese a ser testada

$$H_0 : \beta_j = 0 \quad \text{para algum} \;\; 1 \leq j \leq k$$

o que equivale ao modelo reduzido

$$\boxed{Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{j-1} X_{i,j-1} + \beta_{j+1} X_{i,j+i} + \ldots + \beta_k X_{ik} + \varepsilon_i \quad \text{(modelo r)}}$$

com $m = p-1$ parâmetros a serem estimados.

A matriz W_r e o vetor $\boldsymbol{\theta}_r$ ficam, então,

$$\boldsymbol{W}_r = \begin{bmatrix} 1 & X_{11} & \ldots & X_{1,j-1} & X_{1,j+1} & \ldots & X_{1k} \\ 1 & X_{21} & \ldots & X_{2,j-1} & X_{2,j+1} & \ldots & X_{2k} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & X_{n1} & \ldots & X_{n,j-1} & X_{n,j+1} & \ldots & X_{nk} \end{bmatrix} \quad \text{e} \quad \boldsymbol{\theta}_r = \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \ldots \\ \gamma_{j-1} \\ \gamma_j \\ \ldots \\ \gamma_{m-1} \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_{j-1} \\ \beta_{j+1} \\ \ldots \\ \beta_k \end{bmatrix}.$$

Logo,

$SQP_r = \hat{\boldsymbol{\theta}}_r^T \boldsymbol{W}_r^T \boldsymbol{Y}$, com $r = p - 1$ graus de liberdade,
$SQRes_r = \boldsymbol{Y}^T \boldsymbol{Y} - SQP_r$ com $(n - p + 1)$ graus de liberdade

e a redução na soma de quadrados do resíduo devido à hipótese H_0 é dada por

$$R(H_0) = SQRes_r - SQRes_c \text{ com } 1 \text{ grau de liberdade}$$

o que leva à estatística

$$F = \frac{R(\beta_j|\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots \beta_k)}{QMRes_c} = \frac{SQRes_r - SQRes_c}{QMRes_c} \sim F_{1,n-p}.$$

e rejeita-se $H_0 : \beta_j = 0$, a um nível de $100\gamma\%$ de significância, se $F > F_{1,n-p;\gamma}$.

Essa mesma hipótese pode ser testada pelo teste t, isto é,

$$t = \frac{\hat{\beta}_j - 0}{s(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{s(\hat{\beta}_j)} \sim t_{n-p}$$

em que $\hat{\beta}_j$ e $s(\hat{\beta}_j)$ são estimados sob o *modelo completo* (*modelo c*). Verifica-se que $F = t^2$. Essa é a chamada análise parcial, pois os β's são estimados conjuntamente.

Exemplo 2: Usando-se o mesmo *modelo c* dado por (3.16) e a hipótese a ser testada

$$H_0 : \beta_1 = \beta_2 = \beta \quad vs \quad H_a : \beta_1 \neq \beta_2$$

tem-se que o modelo reduzido fica

$$Y_i = \beta_0 + \beta X_{i1} + \beta X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i$$

ou, ainda

$$\boxed{Y_i = \beta_0 + \beta X_i^* + \beta_3 X_{i3} + \ldots + \beta_k X_{ik} + \varepsilon_i \quad \text{(modelo r)}}$$

com $m = p - 1$ parâmetros a serem estimados e $X_i^* = X_{i1} + X_{i2}$.

A matriz W_r e o vetor $\boldsymbol{\theta}_r$ ficam, então,

$$\boldsymbol{W}_r = \begin{bmatrix} 1 & X_{11} + X_{12} & X_{13} & \ldots & X_{1k} \\ 1 & X_{21} + X_{22} & X_{23} & \ldots & X_{2k} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & X_{n1} + X_{n2} & X_{n3} & \ldots & X_{nk} \end{bmatrix} \quad \text{e} \quad \boldsymbol{\theta}_r = \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \ldots \\ \gamma_{m-1} \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta \\ \beta_3 \\ \ldots \\ \beta_k \end{bmatrix}.$$

Logo,

$$SQRes_r = \boldsymbol{Y}^T \boldsymbol{Y} - SQP_r \text{ com } (n - p + 1) \text{ graus de liberdade}$$

e a redução na soma de quadrados do resíduo devido à hipótese H_0 é dada por

$$R(H_0) = SQRes_r - SQRes_c \ \text{ com } \ 1 \ \text{ grau de liberdade}$$

o que leva à estatística

$$F = \frac{R(H_0)}{QMRes_c} \sim F_{1,n-p}$$

e rejeita-se $H_0 : \beta_1 = \beta_2$, a um nível de $100\gamma\%$ de significância, se $F > F_{1,n-p;\gamma}$.

Exemplo 3: Seja o *modelo c* dado por (3.16) e o teste da hipótese:

$$H_0 : \begin{cases} \beta_1 = 2 \\ \beta_2 = 1 \end{cases}$$

que corresponde ao modelo reduzido

$$Y_i = \beta_0 + 2X_{i1} + X_{i2} + \beta_3 X_{i3} + \ldots + \beta_k X_{ik} + \varepsilon_i$$

com $m = p - 2$ parâmetros a serem estimados, ou, ainda

$$\boxed{Z_i = \beta_0 + \beta_3 X_{i3} + \ldots + \beta_k X_{ik} + \varepsilon_i \quad \text{(modelo r)}}$$

sendo que $Z_i = Y_i - 2X_{i1} - X_{i2}$ é a nova variável resposta, sendo o termo $2X_{i1} + X_{i2}$ conhecido como *offset*. Então,

$$\boldsymbol{W}_r = \begin{bmatrix} 1 & X_{13} & X_{14} & \ldots & X_{1k} \\ 1 & X_{23} & X_{24} & \ldots & X_{2k} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & X_{n3} & X_{n4} & \ldots & X_{nk} \end{bmatrix} \quad \text{e} \quad \boldsymbol{\theta}_r = \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \ldots \\ \gamma_{m-2} \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_3 \\ \beta_4 \\ \ldots \\ \beta_k \end{bmatrix}.$$

A redução na soma de quadrados do resíduo devido à hipótese $H_0 : \beta_1 = 2 \ \text{ e } \ \beta_2 = 1$ é dada por:

$$R(H_0) = SQRes_r - SQRes_c \ \text{ com } \ 2 \ \text{ graus de liberdade}$$

o que leva à estatística

$$F = \frac{\dfrac{R(H_0)}{2}}{QMRes_c} \sim F_{2,n-p}$$

e rejeita-se $H_0 : \beta_1 = 2 \ \text{ e } \ \beta_2 = 1$, a um nível de $100\gamma\%$ de significância, se $F > F_{2,n-p;\gamma}$.

Exemplo 4: Seja o *modelo c* dado por (3.16) e o teste da hipótese:

$$H_0 : \begin{cases} \beta_1 - 2\beta_2 = 4\beta_3 \\ \beta_1 + 2\beta_2 = 6 \end{cases}$$

Resolvendo-se o sistema de equações para β_1 e β_2, essa hipótese é equivalente a

$$H_0 : \begin{cases} \beta_1 = 3 + 2\beta_3 \\ \beta_2 = \dfrac{3}{2} - \beta_3 \end{cases}$$

que corresponde ao modelo reduzido

$$Y_i = \beta_0 + (3 + 2\beta_3)X_{i1} + (\frac{3}{2} - \beta_3)X_{i2} + \beta_3 X_{i3} + \ldots + \beta_k X_{ik} + \varepsilon_i$$

com $m = p - 2$ parâmetros a serem estimados, ou, ainda

$$\boxed{Z_i = \beta_0 + \beta_3 X_i^* + \ldots + \beta_k X_{ik} + \varepsilon_i \quad \text{(modelo r)}}$$

em que $Z_i = Y_i - 3X_{i1} - \dfrac{3}{2}X_{i2}$ é a nova variável resposta, $X_i^* = 2X_{i1} - X_{i2} + X_{i3}$, sendo o termo $3X_{i1} + \dfrac{3}{2}X_{i2}$ conhecido como *offset*. Então,

$$\boldsymbol{W}_r = \begin{bmatrix} 1 & X_1^* & X_{14} & \ldots & X_{1k} \\ 1 & X_2^* & X_{24} & \ldots & X_{2k} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & X_n^* & X_{n4} & \ldots & X_{nk} \end{bmatrix} \quad \text{e} \quad \boldsymbol{\theta}_r = \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \ldots \\ \gamma_{m-2} \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_3 \\ \beta_4 \\ \ldots \\ \beta_k \end{bmatrix}.$$

Logo,

$$F = \frac{\dfrac{R(H_0)}{2}}{QMRes_c} \sim F_{2,n-p}$$

e rejeita-se $H_0 : \beta_1 - 2\beta_2 = 4\beta_3$ e $\beta_1 + 2\beta_2 = 6$, a um nível de $100\gamma\%$ de significância, se $F > F_{2,n-p;\gamma}$.

3.5 Coeficiente de Determinação Múltiplo

É usado como uma medida descritiva da qualidade do ajuste obtido, definido por

$$R^2 = \frac{SQReg}{SQTotal} = 1 - \frac{SQRes}{SQTotal}$$

e indica a proporção da variação de Y que é explicada pela regressão. Note que $0 \leq R^2 \leq 1$.

Entretanto, o valor do coeficiente de determinação deve ser usado com precaução, pois depende do número de observações da amostra, tendendo a crescer quando n diminui. Além disso, é sempre possível torná-lo maior, pela adição de um número suficiente de termos. Assim, se, por exemplo, não há pontos repetidos (mais do que um valor Y para mesmos X's)

uma regressão múltipla com $n-1=k$ variáveis regressoras dará um ajuste perfeito ($R^2=1$) para n dados. Quando há valores repetidos, R^2 não será nunca igual a 1, pois o modelo não poderá explicar a variabilidade devido ao erro puro.

Embora R^2 aumente se se adiciona uma nova variável ao modelo, isto não significa necessariamente que o novo modelo é superior ao anterior. A menos que a soma de quadrados residual do novo modelo seja reduzida de uma quantia igual ao quadrado médio residual original, o novo modelo terá um quadrado médio residual maior do que o original, devido à perda de 1 grau de liberdade. Na realidade esse novo modelo poderá ser pior do que o anterior.

A magnitude de R^2, também, depende da amplitude de variação das variáveis regressoras. Geralmente, R^2 aumentará com maior amplitude de variação dos X's e diminuirá em caso contrário. Assim, um valor grande de R^2 poderá ser grande simplesmente porque os X's variaram em uma amplitude muito grande. Por outro lado R^2 poderá ser pequeno porque as amplitudes dos X's foram muito pequenas para permitir que uma relação com Y fosse detectada.

Dessa forma, vê-se que R^2 não deve ser considerado sozinho, mas sempre aliado a outros diagnósticos do modelo. Numa tentativa de correção dos problemas apontados, foi definido o **coeficiente de determinação ajustado** para graus de liberdade, indicado por $\bar{R}^2$. Tem-se que:

$$1-R^2=1-\frac{SQReg}{SQTotal}=\frac{SQRes}{SQTotal}.$$

O **coeficiente de determinação ajustado** é definido por:

$$1-\bar{R}^2=\frac{\frac{1}{n-p}SQRes}{\frac{1}{n-1}SQTotal}=\frac{n-1}{n-p}(1-R^2)\Rightarrow\bar{R}^2=R^2-\frac{1}{n-p}(1-R^2)$$

Excluindo-se o caso em que $R^2=1$, tem-se que $\bar{R}^2<R^2$. Note que $\bar{R}^2$ pode ser negativo.

3.6 Exemplo

Considere os dados do Exercício 5 do item 1.4.1 (página 16) referentes a medidas de diâmetro à altura do peito (D) e altura (H) de árvores (*black cherry*) em pé e de volume (V) de árvores derrubadas. O objetivo desse tipo de experimento é verificar de que forma essas variáveis estão relacionadas para, usando-se medidas nas árvores em pé, poder se predizer o volume de madeira em uma área de floresta.

A Figura 3.2 mostra os gráficos de dispersão das variáveis duas a duas para os dados observados sem transformação e com transformação logarítmica. Pode-se ver que existe uma

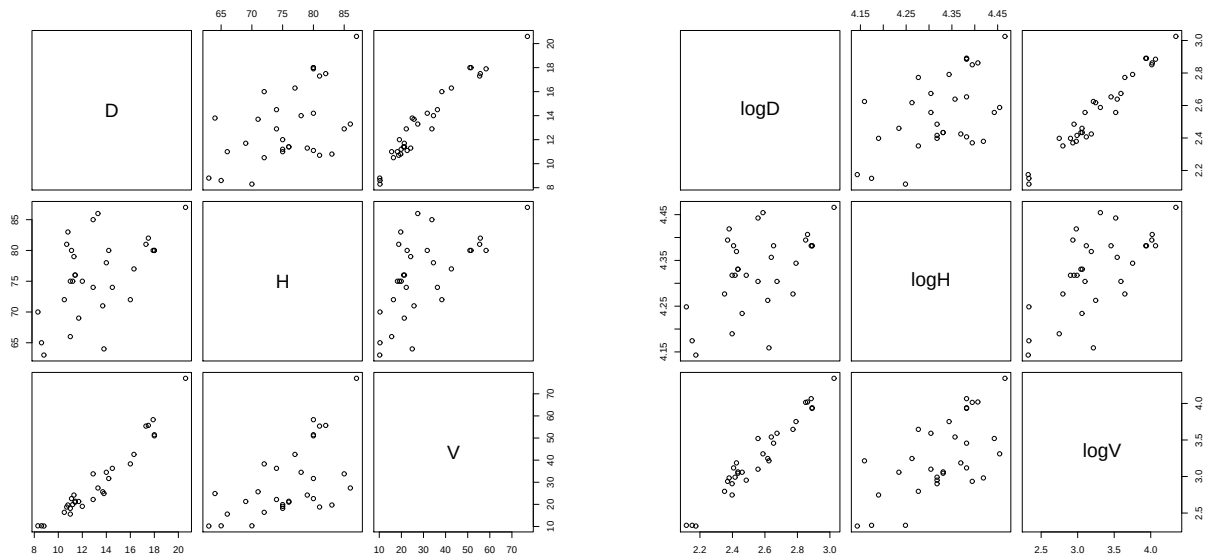


Figura 3.2: Gráfico de dispersão - valores observados e transformados

relação mais forte entre volume e diâmetro à altura do peito, do que entre volume e altura. Além disso, que a variável altura tem variabilidade maior do que a variável diâmetro à altura do peito.

As Tabelas 3.3 e 3.4 mostram os resultados obtidos para a análise dos dados sem transformação e com transformação logarítmica. Verifica-se que existem evidências, ao nível de 1% de significância, que os efeitos tanto de diâmetro à altura do peito como de altura são significativos, sendo que o efeito de diâmetro à altura do peito é maior que o de altura, tanto para o caso de dados não transformados como para transformados. Há necessidade, porém, de um estudo mais detalhado, usando-se análise de resíduos e diagnósticos, para a escolha do modelo final.

É importante, lembrar, também, que o teste para o modelo com ambas as variáveis (regressão parcial) simultaneamente tem um nível de significância conjunto, enquanto que na análise seqüencial não se sabe o nível conjunto de significância dos testes.

A necessidade de transformação pode ser explicada, considerando-se que volume é proporcional ao produto de diâmetro à altura do peito pela altura, isto é,

$$V \approx \gamma_0 D^{\beta_1} H^{\beta_2}$$

logo,

$$log(V) = \beta_0 + \beta_1 log(D) + \beta_2 log(H) + \varepsilon$$

Testes t (equivalentes aos testes F) e intervalos de confiança para os parâmetros e intervalos de previsão para V podem ser obtidos de forma semelhante ao que já foi visto no

capítulo 2.

3.7 Exercícios

1. Reescrever os modelos de regressão a seguir, em notação matricial ($\boldsymbol{Y} = \boldsymbol{X\theta} + \boldsymbol{\varepsilon}$)

   (a) $Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$

   (b) $Y_i = \beta_0 + \beta_1 X_{i1}^2 + \varepsilon_i$

   (c) $Y_i = \beta_0 + \beta_1 \log(X_{i1}) + \varepsilon_i$

   (d) $Y_i = \beta_0 + \beta_1 (X_{i1} - X_{u1}) + \varepsilon_i$

   (e) $Y_i = \beta_0 + \beta_1 (X_{i1} - X_{u1}) I_{\{i<u\}} + \varepsilon_i$

   (f) $Y_i = \beta_1 X_{i1}^3 + \varepsilon_i$

   (g) $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$

   (h) $Y_i = \beta_1 x_{i1} + \beta_2 x_{i2}^2 + \varepsilon_i$

   (i) $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \varepsilon_i$

   (j) $Y_i = \beta_0 + \beta_1 (X_{i1} - X_{u1}) I_{\{i<u\}} + \beta_2 (X_{i1} - X_{u1}) I_{\{i\geq u\}} + \varepsilon_i$

   sendo $x_{i1} = X_{i1} - \bar{X}_1$, $x_{i2} = X_{i2} - \bar{X}_2$ e $i = 1, \ldots, u, \ldots, n$.

2. Considerando-se os modelos do item anterior, $\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ e $\mathrm{E}(\boldsymbol{\varepsilon\varepsilon}^T) = \boldsymbol{I}\sigma^2$, obtenha, usando o método dos quadrados mínimos, $\hat{\boldsymbol{\theta}}$ e $\mathrm{Var}(\hat{\boldsymbol{\theta}})$.

3. Considere o modelo de regressão linear múltipla

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \qquad (i = 1, \ldots, n).$$

   (a) Simule um conjunto de dados considerando-se que $X_{11} = -1$, $X_{21} = -1$, $X_{31} = 1$, $X_{41} = 1$, $X_{12} = -1$, $X_{22} = 1$, $X_{32} = -1$, $X_{42} = 1$, $\alpha = 10$, $\beta_1 = 2$, $\beta_2 = -1$ e $\varepsilon_i \sim \mathrm{N}(0; 9)$.

   (b) Baseando-se nos dados simulados no item (a), obtenha as somas de quadrados e o quadro da análise de variância.

4. Considere o modelo de regressão linear segmentada

$$Y_i = \alpha + \beta_1 (X_i - X_u) I_{\{i<u\}} + \beta_2 (X_i - X_u) I_{\{i\geq u\}} + \varepsilon_i, \qquad (i = 1, \ldots, u, \ldots, n)$$

   em que $I_{\{\cdot\}}$ são variáveis indicadoras, que assumem o valor 1 quando a condição entre chaves estiver satisfeita ou o valor 0, caso contrário.

(a) Simule um conjunto de dados considerando-se que $X_1 = -2$, $X_2 = -1$, $X_3 = 0$, $X_4 = 1$, $X_5 = 2$, $\alpha = 10$, $\beta_1 = 2$, $\beta_2 = -1$, $u = 3$ e $\varepsilon_i \sim \mathrm{N}(0; 36)$.

(b) Baseando-se nos dados simulados no item (a), obtenha as somas de quadrados e o quadro da análise de variância.

5. Considere os dados da Tabela 3.5 e o modelo de regressão linear múltipla

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \qquad (i = 1, \ldots, n).$$

Pede-se:

(a) Determine as estimativas dos parâmetros.

(b) Faça a análise de variância da regressão.

(c) Teste a hipótese $H_0 : \beta_0 = 0$ contra $H_\mathrm{a} : \beta_0 \neq 0$, a um nível de significância $\gamma = 0,05$.

(d) Teste a hipótese $H_0 : \beta_1 = 0$ contra $H_a : \beta_1 \neq 0$, a um nível de significância $\gamma = 0,05$.

(e) Determine os valores dos coeficientes de determinação e de determinação corrigido.

(f) Construa os intervalos de confiança para $E(Y_i)$ $(i = 1, \ldots, n)$, com um coeficiente de confiança de 95%.

(g) Teste a hipótese $H_0 : \beta_2 = 0$ contra $H_\mathrm{a} : \beta_2 \neq 0$, a um nível de significância $\gamma = 0,05$.

(h) Determine a estimativa de Y_h para $X_{h1} = 2$ e $X_{h2} = 4$ e construa o intervalo de previsão para Y_h com um coeficiente de confiança de 95%.

6. Considere os valores apresentados na Tabela 3.6 e o modelo de regressão linear múltipla

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

e suponha que os ε_i's são independentes e $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$. Pede-se:

(a) Faça a análise de regressão parcial e interprete o teste F e os testes t para os parâmetros.

(b) Faça a análise de regressão sequencial, considerando a seqüência X_1, $X_2|X_1$ e $X_3|X_1, X_2$.

(c) Fazer o teste F para as hipóteses:

i. $H_0 : \beta_1 = 0$ e $\beta_2 = 0$

ii. $H_0 : \beta_1 = \beta_2$.

iii. $H_0 : \beta_1 = \beta_3$.

iv. $H_0 : \beta_1 = 1$ e $\beta_1 + \beta_2 + \beta_3 = 1$.

v. $H_0 : \beta_2 = -2\beta_3$ e $3\beta_1 = 2\beta_2$.

7. Considere os dados do do exemplo 4 da Seção 1.3 (pág. 10 e 11) o modelo de regressão linear múltipla

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \quad (i = 1, \ldots, n).$$

Pede-se:

(a) Obter as estimativas de mínimos quadrados dos parâmetros do modelo.

(b) Obter as estimativas das variâncias e covariâncias das estimativas de mínimos quadrados dos parâmetros do modelo.

(c) Fazer a análise de variância da regressão e concluir, considerando o nível de significância 0,05.

Quais pressuposições foram consideradas para resolver este exercício? Existe alguma que não esteja satisfeita? Caso a resposta seja sim, quais são as possíveis consequências sobre a análise e as conclusões obtidas?

8. Usando os dados do exemplo 4 da Seção 1.3 (pág. 10 e 11), pede-se:

(a) Fazer a análise de regressão parcial e a sequencial, considerando o modelo de regressão

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \varepsilon_i$$

sendo $X_i = X_{i1}$ e $Z_i = X_{i1}$. Tirar conclusões.

(b) Dados os modelos

$\mathrm{E}(Y) = a_{0X} + a_{1X}X$, $\mathrm{E}(Y) = b_{0Z} + b_{1Z}Z$, $\mathrm{E}(X) = c_{0Z} + c_{1Z}Z$ e $\mathrm{E}(Z) = d_{0X} + d_{1X}X$, obter

i. as estimativas dos parâmetros

ii. $\hat{Y}_{YX}$, $\hat{Y}_{YZ}$, $\hat{X}_{XZ}$ e $\hat{Z}_{ZX}$ e

iii. os resíduos (chamados resíduos parciais) $\hat{\varepsilon}_{YX} = Y - \hat{Y}_{YX}$, $\hat{\varepsilon}_{YZ} = Y - \hat{Y}_{YZ}$, $\hat{\varepsilon}_{XZ} = X - \hat{X}_{XZ}$ e $\hat{\varepsilon}_{ZX} = Z - \hat{Z}_{ZX}$

(c) Obter as estimativas dos parâmetros β_{YX} e β_{YZ} das retas passando pela origem (pois, em ambos os modelos ambas as variáveis têm soma nula), considerando, respectivamente, $\hat{\varepsilon}_{YZ}$ versus $\hat{\varepsilon}_{XZ}$ e $\hat{\varepsilon}_{YX}$ versus $\hat{\varepsilon}_{ZX}$. Verifique que $\hat{\beta}_{YX} = \hat{\beta}_1$ e $\hat{\beta}_{YZ} = \hat{\beta}_2$.
Nota: Vê-se, portanto, que: β_1 é o coeficiente de regressão entre Y e X ajustado para Z (ou eliminada a influência de Z sobre Y e sobre X) e β_2 é o coeficiente de regressão entre Y e Z ajustado para X (ou eliminada a influência de X sobre Y e sobre Z).

(d) Obter a correlação entre $\hat{\varepsilon}_{YZ}$ e $\hat{\varepsilon}_{XZ}$.

Tabela 3.3: Análise de variância, teste F e estimativas - Dados sem transformação

Modelo $E(Y) = \beta_0 + \beta_1 X_1$				
Causas de variação	G.L.	S.Q.	Q.M.	F
DAP	1	$7.581,8$	$7.581,8$	$419,4**$
Resíduo	29	$524,3$	$18,1$	
Total	30	$8.106,1$		
$\hat{Y} = -36,94 + 5,066X_1$			$R^2 = 0,935$	$\bar{R}^2 = 0,933$
$s(\hat{\beta}_0) = 3,36$ e $s(\hat{\beta}_1) = 0,247$				

Modelo $E(Y) = \beta_0 + \beta_2 X_2$				
Causas de variação	G.L.	S.Q.	Q.M.	F
Altura	1	$2.901,2$	$2.901,2$	$16,2**$
Resíduo	29	$5.204,9$	$179,5$	
Total	30	$8.106,1$		
$\hat{Y} = -87,12 + 1,543X_2$			$R^2 = 0,358$	$\bar{R}^2 = 0,336$
$s(\hat{\beta}_0) = 29,27$ e $s(\hat{\beta}_2) = 0,384$				

Modelo $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ - Parcial				
Causas de variação	G.L.	S.Q.	Q.M.	F
DAP e Altura	2	$7.684,4$	$3.842,2$	$255,0**$
Resíduo	28	$421,9$	$15,1$	
Total	30	$8.106,1$		
$\hat{Y} = -57,99 + 4,708X_1 + 0,339X_2$			$R^2 = 0,948$	$\bar{R}^2 = 0,944$
$s(\hat{\beta}_0) = 8,64$, $s(\hat{\beta}_1) = 0,264$ e $s(\hat{\beta}_2) = 0,130$				

Modelo $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ - Sequencial				
Causas de variação	G.L.	S.Q.	Q.M.	F
DAP	1	$7.581,8$	$7.581,8$	$503,1**$
Altura|DAP	1	$102,4$	$102,4$	$6,8*$
Resíduo	28	$421,9$	$15,1$	
Total	30	$8.106,1$		

Modelo $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ - Sequencial				
Causas de variação	G.L.	S.Q.	Q.M.	F
Altura	1	$2.901,2$	$2.901,2$	$192,5**$
DAP|Altura	1	$4.783,0$	$4.783,0$	$317,4**$
Resíduo	28	$421,9$	$15,1$	
Total	30	$8.106,1$		
$F_{1,29;0,05} = 4,18$, $F_{2,28;0,05} = 3,34$ e $F_{1,28;0,05} = 4,20$				
$F_{1,29;0,01} = 7,60$, $F_{2,28;0,01} = 5,45$ e $F_{1,28;0,01} = 7,64$				

Tabela 3.4: Análise de variância, teste F e estimativas - Dados transformados

Modelo $E[log(Y)] = \beta_0 + \beta_1\ log(X_1)$				
Causas de variação	G.L.	S.Q.	Q.M.	F
DAP	1	$7,9254$	$7,9254$	$599,7**$
Resíduo	29	$0,3832$	$0,0132$	
Total	30	$8,3087$		

$\widehat{log(Y)} = -2,353 + 2,2\ log(X_1) \quad R^2 = 0,954 \quad \bar{R}^2 = 0,952$

$s(\hat{\beta}_0) = 0,231$ e $s(\hat{\beta}_1) = 0,089$

Modelo $E[log(Y)] = \beta_0 + \beta_2\ log(X_2)$				
Causas de variação	G.L.	S.Q.	Q.M.	F
Altura	1	$3,496$	$3,496$	$21,06**$
Resíduo	29	$4,8130$	$0,166$	
Total	30	$8,3087$		

$\widehat{log(Y)} = -13,96 + 3,982\ log(X_2) \quad R^2 = 0,421 \quad \bar{R}^2 = 0,401$

$s(\hat{\beta}_0) = 3,76$ e $s(\hat{\beta}_2) = 0,868$

Modelo $E[log(Y)] = \beta_0 + \beta_1\ log(X_1) + \beta_2\ log(X_2)$ - Parcial				
Causas de variação	G.L.	S.Q.	Q.M.	F
DAP e Altura	2	$8,1228$	$4,0614$	$615,36**$
Resíduo	28	$0,1855$	$0,0066$	
Total	30	$8,3087$		

$\widehat{log(Y)} = -6,632 + 1,983\ log(X_1) + 1,117\ log(X_2) \quad R^2 = 0,978 \quad \bar{R}^2 = 0,976$

$s(\hat{\beta}_0) = 0,799,\ s(\hat{\beta}_1) = 0,0,075$ e $s(\hat{\beta}_2) = 0,204$

Modelo $E[log(Y)] = \beta_0 + \beta_1\ log(X_1) + \beta_2\ log(X_2)$ - Sequencial				
Causas de variação	G.L.	S.Q.	Q.M.	F
DAP	1	$7,9254$	$7,9254$	$1196,5**$
Altura|DAP	1	$0,1978$	$0,1978$	$29,9**$
Resíduo	28	$0,1855$	$0,0066$	
Total	30	$8,3087$		

Modelo $E[log(Y)] = \beta_0 + \beta_1\ log(X_1) + \beta_2\ log(X_2)$ - Sequencial				
Causas de variação	G.L.	S.Q.	Q.M.	F
Altura	1	$3,4957$	$3,4957$	$527,8**$
DAP|Altura	1	$4,6275$	$4,6275$	$698,6**$
Resíduo	28	$0,1855$	$0,0066$	
Total	30	$8,3087$		

$F_{1,29;0,05} = 4,18,\ F_{2,28;0,05} = 3,34$ e $F_{1,28;0,05} = 4,20$

$F_{1,29;0,01} = 7,60,\ F_{2,28;0,01} = 5,45$ e $F_{1,28;0,01} = 7,64$

Tabela 3.5: Valores de X_{i1}, X_{i2} e Y_i, $(i = 1, \ldots, 6)$.

X_1	0	1	1	2	2	3
X_2	0	2	4	2	4	6
Y	1,5	6,5	10,0	11,0	11,5	16,5

Fonte: HOFFMAN, R. & VIEIRA, S. (1983). *Análise de Regressão. Uma Introdução à Econometria.* 2ª ed. Ed. Hucitec, São Paulo, pág. 124.

Tabela 3.6: Valores de X_{i1}, X_{i2}, X_{i3} e Y_i $(i = 1, \ldots, 14)$.

i	X_{i1}	X_{i2}	X_{i3}	Y_i	i	X_{i1}	X_{i2}	X_{i3}	Y_i
1	-2	2	-2	8,5	8	1	0	0	6,0
2	-1	-1	0	1,0	9	1	1	0	7,0
3	-1	0	0	4,0	10	0	0	-1	5,0
4	-1	0	0	4,0	11	0	0	0	5,0
5	-1	1	0	5,0	12	0	0	0	5,0
6	1	-1	0	3,0	13	0	0	1	3,0
7	1	0	0	6,0	14	2	-2	2	0,5

Fonte: HOFFMAN, R. & VIEIRA, S. (1983). *Análise de Regressão. Uma Introdução à Econometria.* 2ª ed. Ed, Hucitec, São Paulo, pág. 147.

O programa em R utilizado para a obtenção dos resultados foi

```
# Libraries needed #
library(MASS)
library(car)

#  Minitab Cherry Tree Data
#  ========================
# Volume of usable wood in 31 black cherry trees from
# Minitab Student Handbook (1985), Ryan, Joiner and Ryan.
#     D = diameter at 4.5 ft from ground (inches)
#     H = height (feet)
#     V = volume (cubic feet)

##trees<-read.table("Tree.dat", header=TRUE)
##require(trees)
data(trees, package='datasets')
data() # lists all datasets
attach(trees)

D<-trees[,1]
H<-trees[,2]
V<-trees[,3]
# first examine the data
par(mfrow=c(2,3))
hist(D, main="Girth")
hist(H, main="Height")
hist(V, main="Volume")

boxplot(D, main="Girth")
boxplot(H, main="Height")
boxplot(V, main="Volume")

#Scatterplot
pairs(trees)

plot(trees)
scatterplot.matrix(trees) # uses library(car)
```

```
## Fitted models ##
mod1<-lm(V~1)
mod2<-lm(V~D)
summary(mod2)
mod3<-lm(V~H)
summary(mod3)
mod4<-lm(V~D+H)
summary(mod4)
anova(mod1, mod4)
anova(mod1, mod2, mod4)
anova(mod1, mod3, mod4)

## Log transformed data ##

#Scatterplots
logD<-log(D)
logH<-log(H)
logV<-log(V)
ltrees<-cbind(logD,logH,logV)
pairs(ltrees)

## Fitted models ##
mod1<-lm(logV~1)
mod2<-lm(logV~logD)
summary(mod2)
mod3<-lm(logV~logH)
summary(mod3)
mod4<-lm(logV~logD+logH)
summary(mod4)
anova(mod1, mod4)
anova(mod1, mod2, mod4)
anova(mod1, mod3, mod4)
```

Capítulo 4

Análise de Resíduos e Diagnósticos

4.1 Introdução

No modelo linear $\boldsymbol{Y} = \boldsymbol{X\theta} + \boldsymbol{\varepsilon} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$, os elementos ε_i do vetor $\boldsymbol{\varepsilon}$ são as diferenças entre os valores observados (Y_i's) e aqueles esperados pelo modelo (μ_i's). São chamados erros, representam a variação natural dos dados e admite-se que os ε_i's são independentes e, além disso, $\boldsymbol{\varepsilon}_i \sim \mathrm{N}(0, \sigma^2)$, isto é, comportam-se como especificado pelas pressuposições das página 70 e 71. Entretanto, nem sempre é o caso e, se as suposições são violadas, têm-se as *falhas sistemáticas* (não linearidade, não-normalidade, heterocedasticidade, não-independência dos erros, efeito cumulativo de fatores que não foram considerados no modelo etc) e a análise resultante pode levar a conclusões duvidosas. Outro fato bastante comum é a presença de pontos atípicos (*falhas isoladas*), que podem influenciar, ou não, o ajuste do modelo. Elas podem surgir devido a:

- erros grosseiros na variável resposta ou nas variáveis explanatórias, por medidas erradas ou registro da observação, ou ainda, erros de transcrição;

- observação proveniente de uma condição distinta das demais;

- modelo mal especificado (falta de uma ou mais variáveis, modelo inadequado etc);

- escala errada, talvez os dados sejam melhor descritos após uma transformação, do tipo logarítmica ou raiz quadrada;

- distribuição da variável resposta errada, por exemplo, tem uma cauda mais longa do que a distribuição normal.

Ajustado um determinado modelo a um conjunto de dados, para a verificação das pressuposições devem ser considerados como material básico: os valores estimados (ou ajustados), $\hat{\mu}_i = \hat{Y}_i$, os resíduos, $r_i = Y_i - \hat{\mu}_i$, a variância residual estimada, $\hat{\sigma}^2 = s^2 = QMRes$, e os

elementos da diagonal (*leverage*) da matriz de projeção

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \begin{bmatrix} h_1 & h_{12} & \dots & h_{1n} \\ h_{12} & h_2 & \dots & h_{2n} \\ \dots & \dots & \dots & \dots \\ h_{1n} & h_{2n} & \dots & h_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \dots \\ \boldsymbol{x}_n^T \end{bmatrix} (\boldsymbol{X}^T\boldsymbol{X})^{-1} \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \dots & \boldsymbol{x}_n \end{bmatrix}$$

isto é, $h_{ii} = h_i = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$, sendo $\boldsymbol{x}_i^T = \begin{bmatrix} X_{i1} & X_{i2} & \dots & X_{ik} \end{bmatrix}$.

Uma idéia importante é a da deleção (*deletion*), isto é, a comparação do ajuste do modelo escolhido, considerando-se todos os pontos, com o ajuste do mesmo modelo sem os pontos atípicos. As estatísticas obtidas pela omissão de um certo ponto i são denotadas com um índice entre parênteses. Assim, por exemplo, $s^2_{(i)}$ representa a variância residual estimada para o modelo ajustado, excluído o ponto i.

As técnicas usadas para a verificação do ajuste de um modelo a um conjunto de dados podem ser formais ou informais. As informais baseiam-se em exames visuais de gráficos para a detecção de padrões, ou então, de pontos discrepantes. As formais envolvem aninhar o modelo sob pesquisa em uma classe maior pela inclusão de um parâmetro (ou vetor de parâmetros) extra. As mais usadas são baseadas nos testes da razão de verossimilhanças e escore. Parâmetros extras podem aparecer devido a:

- inclusão de uma covariável adicional;

- aninhamento de uma covariável X em uma família indexada por um parâmetro γ, sendo um exemplo a família de Box-Cox;

- inclusão de uma variável construída;

- inclusão de uma variável *dummy* tomando o valor 0 (zero) para a(s) unidade(s) discrepante(s) e 1 (um) para as demais. Isso é equivalente a eliminar essa observação do conjunto de dados, fazer a análise com a(s) observação(ões) discrepante(s) e sem ela(s) e verificar se a mudança no valor da SQRes. é significativa, ou não. Depende, porém, de localizar o(s) ponto(s) discrepante(s).

4.2 Tipos de resíduos

Os resíduos têm papel fundamental na verificação do ajuste de um modelo e vários tipos foram propostos na literatura (Cook & Weisberg, 1982; Atkinson, 1985; Miazaki & Stangenhaus, 1994).

a) **Resíduos ordinários** - Os resíduos do processo de ajustamento por quadrados mínimos são dados por:

$$r_i = \hat{\varepsilon}_i = Y_i - \hat{\mu}_i.$$

Entretanto, enquanto que os erros ε_i's são independentes e com a mesma variância, o mesmo não ocorre com os resíduos ordinários, isto é,

$$\text{Var}(\hat{\boldsymbol{\varepsilon}}) = \text{Var}[(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}] = (\boldsymbol{I} - \boldsymbol{H})\sigma^2.$$

Em particular, a variância do i-ésimo resíduo é dada por:

$$\text{Var}(r_i) = \text{Var}(\hat{\varepsilon}_i) = (1 - h_i)\sigma^2$$

e a covariância dos i-ésimo e i'-ésimo resíduos por:

$$\text{Cov}(r_i, r_{i'}) = \text{Cov}(\hat{\varepsilon}_i, \hat{\varepsilon}_{i'}) = -h_{ii'}\sigma^2.$$

Assim, usar $r_i = \hat{\varepsilon}_i$ pode não ser adequado devido à heterogeneidade de variâncias. Foram, então, propostas diferentes padronizações para sanar esse problema.

b) **Resíduos estudentizados internamente (*Studentized residual*)** - Considerando-se $s^2 = QMRes$ como a estimativa de σ^2, tem-se que um estimador não tendencioso para $\text{Var}(\varepsilon_i)$ é dado por:

$$\widehat{\text{Var}}(\hat{\varepsilon}_i) = (1 - h_i)s^2 = (1 - h_i)QMRes$$

e como $\text{E}(r_i) = \text{E}(\hat{\varepsilon}_i) = \text{E}(Y_i - \hat{\mu}_i) = 0$, então, o resíduo estudentizado internamente é:

$$rsi_i = \frac{\hat{\varepsilon}_i}{\sqrt{\widehat{Var}(\hat{\varepsilon}_i)}} = \frac{r_i}{s\sqrt{(1 - h_i)}} = \frac{Y_i - \hat{\mu}_i}{\sqrt{(1 - h_i)QMRes}}.$$

Esses resíduos são mais sensíveis do que os anteriores por considerarem variâncias distintas. Entretanto, um valor discrepante pode alterar profundamente a variância residual dependendo do modo como se afasta do grupo maior das observações. Além disso, numerador e denominador são variáveis dependentes, isto é, $\text{Cov}(\hat{\varepsilon}, QMRes) \neq 0$.

c) **Resíduos estudentizados externamente (*jackknifed residuals*, *deletion residuals*, *externally studentized residual*, *RStudent*)** - Para garantir a independência do numerador e denominador na padronização dos resíduos, define-se o resíduo estudentizado externamente, como:

$$rse_{(i)} = \frac{r_i}{s_{(i)}\sqrt{(1 - h_i)}}$$

sendo que $s^2_{(i)}$ é o quadrado médio residual livre da influência da observação i, ou seja, a estimativa de σ^2, omitindo-se a observação i. Prova-se que:

$$rse_{(i)} = rsi_i \sqrt{\frac{n-p-1}{n-p-rsi_i^2}}$$

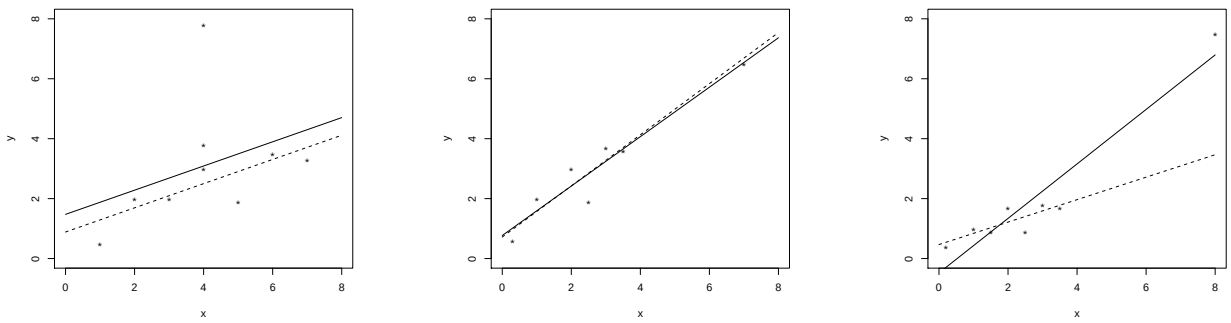
sendo que p é o número de parâmetros independentes.

A vantagem de usar $rse_{(i)}$ é que, sob normalidade, ele tem distribuição t de Student com $(n-p-1)$ graus de liberdade. Embora não seja recomendada a prática de testes de significância na análise de resíduos, sugere-se que a i-ésima observação seja merecedora de atenção especial se $|rse_{(i)}|$ for maior do que o $100(1-\frac{\gamma}{2n})$-ésimo percentil da distribuição t com $(n-p-1)$ graus de liberdade, sendo que γ, o nível de significância, é dividido por n por ser este o número de pontos sob análise.

4.3 Estatísticas para diagnósticos

Um ponto aberrante, ou atípico ou extremo é aquele que tem h_i e/ou resíduo grandes, caracterizando discrepâncias isoladas. No estudo desses pontos são importantes as noções de *leverage*, consistência e influência (ver pág 404, McCullagh & Nelder, 1989), como mostrado na Figura 4.1, estando o ponto extremo indicado por um círculo. Na Figura 4.1(a), o valor de X para o ponto extremo está proximo da média. Sua exclusão não afeta a estimativa do coeficiente angular mas reduz muito o intercepto, melhorando o ajuste do modelo. Na Figura 4.1(b) o ponto extremo é consistente com os demais e sua inclusão não afeta o ajuste grandemente mas melhora a precisão de $\hat{\beta}$. Na Figura 4.1(c) a exclusão do ponto extremo afeta grndemente o ajuste.

Figura 4.1: Diagramas de dispersão



Assim, uma observação pode ser classificada como:

a) *Inconsistente*: quando destoa da tendência geral das demais, isto é, tem resíduo ($rse_{(i)}$) grande, como mostra a Figura 4.1 no confronto de (a) e (c) versus (b). Em geral, é

inconsistente se $|rse_{(i)}| \geq t_{\{\frac{\gamma}{2n};n-p-1\}}$, com nível de significância igual a $100\gamma\%$. Um ponto inconsistente é considerado um *outlier* quando tem *leverage* (h_i) pequeno e resíduo ($rse_{(i)}$) grande.

b) *Ponto de alavanca*: quando tem medida de *leverage* h_i grande, como mostra a Figura 4.1 no confronto de (b) e (c) versus (a). Em geral, é grande se $h_i \geq \frac{2p}{n}$. Pode ser classificado como bom (b), quando consistente, ou ruim (c), quando inconsistente.

c)*Influente*: Quando uma observação está distante das outras em termos das variáveis explanatórias ela pode ser, ou não, influente como mostra a Figura 4.1 no confronto de (a) e (b) versus (c). Uma observação influente é aquela cuja omissão do conjunto de dados resulta em mudanças substanciais em certos aspectos do modelo. Ela pode ser um *outlier*, ou não. Uma observação pode ser influente de diversas maneiras, isto é, no ajuste geral do modelo ($DFFitS_{(i)}$, $C_{(i)}$ ou $D_{(i)}$ grandes), no conjunto de estimativas dos parâmetros ($DFBetaS_{(i)}$ grandes), na estimativa de um determinado parâmetro ($DFBetaS_{(i)}$ grande para um determinado β_j), na escolha de uma transformação da variável resposta ou de uma variável explanatória.

Observação: No pacote estatístico R, a i-ésima observação é considerada influente se $|\text{DFBetaS}_{(i)}| > 1$, se $|\text{DFFitS}_{(i)}| > 3\sqrt{p/(n-p)}$, se $|1 - COVRATIO| > 3p/(n-p)$, se $D_{(i)} > F_{\{0{,}05;p;n-p\}}$, ou se $h_i > 3p/n$.

A seguir são descritas as estatísticas citadas.

a) **Elementos da diagonal da matriz de projeção $\boldsymbol{H}$ (h_i, *leverage*)** - A distância de uma observação em relação às demais é medida pelo h (medida de *leverage*).

No caso particular da regressão linear simples, usando-se variável centrada $x_i = X_i - \bar{X}$, tem-se:

$$\boldsymbol{H} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \cdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum_{i=1}^n x_i^2} \end{bmatrix} \begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \end{bmatrix}$$

e, portanto,

$$h_i = \frac{1}{n} + \frac{x_i^2}{\sum_{i=1}^n x_i^2} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^n x_i^2}, \quad \text{elementos da diagonal de } \boldsymbol{H}$$

e

$$h_{ii'} = \frac{1}{n} + \frac{x_i x_i'}{\sum_{i=1}^n x_i^2} = \frac{1}{n} + \frac{(X_i - \bar{X})(X_{i'} - \bar{X})}{\sum_{i=1}^n x_i^2}, \quad \text{elementos fora da diagonal de } \boldsymbol{H}$$

o que mostra que à medida que X se afasta de $\bar{X}$ o valor de h_i aumenta e que o valor mínimo de h_i é $\frac{1}{n}$. Esse valor mínimo ocorre para todos os modelos que incluem uma constante. No caso em que o modelo de regressão passa pela origem, o valor mínimo de

h_i é 0 para uma observação $X_i = 0$. O valor máximo de h_i é 1, ocorrendo quando o modelo ajustado é irrelevante para a predição em X_i e o resíduo é igual a 0. Sendo $\boldsymbol{H}$ uma matriz de projeção, tem-se

$$\boldsymbol{H} = \boldsymbol{H}^2 = \begin{bmatrix} h_1 & h_{12} & \dots & h_{1n} \\ h_{12} & h_2 & \dots & h_{2n} \\ \dots & \dots & \dots & \dots \\ h_{1n} & h_{2n} & \dots & h_n \end{bmatrix} = \begin{bmatrix} h_1 & h_{12} & \dots & h_{1n} \\ h_{12} & h_2 & \dots & h_{2n} \\ \dots & \dots & \dots & \dots \\ h_{1n} & h_{2n} & \dots & h_n \end{bmatrix} \begin{bmatrix} h_1 & h_{12} & \dots & h_{1n} \\ h_{12} & h_2 & \dots & h_{2n} \\ \dots & \dots & \dots & \dots \\ h_{1n} & h_{2n} & \dots & h_n \end{bmatrix} =$$

e, portanto,

$$h_i = \sum_{i'=1}^{n} h_{ii'}^2 = h_i^2 + \sum_{i' \neq i} h_{ii'}^2 \Rightarrow h_i(1 - h_i) = \sum_{i' \neq i} h_{ii'}^2$$

concluindo-se que $0 \leq h_i \leq 1$ e $\sum_{i'=1}^{n} h_{ii'} = 1$. Além disso,

$$r(\boldsymbol{H}) = \text{tr}(\boldsymbol{H}) = \text{tr}[\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T] = \text{tr}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}] = \text{tr}(\boldsymbol{I}_p) = \sum_{i=1}^{n} h_i = p,$$

e, então, o valor médio de h_i é $\dfrac{p}{n}$.

No processo de ajuste, como

$$\hat{\boldsymbol{\mu}} = \boldsymbol{HY} = \begin{bmatrix} h_1 & h_{12} & \dots & h_{1n} \\ h_{12} & h_2 & \dots & h_{2n} \\ \dots & \dots & \dots & \dots \\ h_{1n} & h_{2n} & \dots & h_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix}$$

tem-se

$$\hat{\mu}_i = \sum_{i'=1}^{n} h_{ii'} Y_{i'} = h_{i1}Y_1 + h_{i2}Y_2 + \ldots + h_iY_i + \ldots + h_{in}Y_n \;\text{ com }\; 1 \leq i \leq n.$$

Vê-se, portanto, que o valor ajustado $\hat{\mu}_i$ é a média ponderada dos valores observados e que o peso de ponderação é o valor de $h_{ii'}$. Assim, o elemento da diagonal de $\boldsymbol{H}$ é o peso com que a observação Y_i participa do processo de obtenção do valor ajustado $\hat{\mu}_i$. Valores de $h_i \geq \dfrac{2p}{n}$, segundo Belsley, Kuh & Welsch (1980, p.17) indicam observações que merecem uma análise mais apurada.

b) **DFBeta** - É importante quando o coeficiente de regressão tem um significado prático. Mede a alteração no vetor estimado $\hat{\boldsymbol{\theta}}$ ao se retirar a i-ésima observação da análise. É dado por

$$DFBeta_{(i)} = \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(i)} = \frac{1}{(1 - h_i)}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i^T r_i$$

Não tem interpretação simples. Cook & Weisberg (1982) propuseram curvas empíricas para o estudo dessa medida.

c) **DFFitS** - Mede a alteração provocada no valor ajustado pela retirada da observação i. É dada por

$$DFFitS_{(i)} = \frac{DFFit_{(i)}}{\sqrt{h_i s^2_{(i)}}} = \frac{\boldsymbol{x}_i(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(i)})}{\sqrt{h_i s^2_{(i)}}} = \frac{\hat{\mu}_i - \hat{\mu}_{(i)}}{\sqrt{h_i s^2_{(i)}}} \frac{1}{(1-h_i)} \boldsymbol{x}_i^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_i r_i$$

ou, ainda,

$$DFFitS_{(i)} = \left(\frac{h_i}{1-h_i}\right)^{\frac{1}{2}} \frac{r_i}{s_{(i)}(1-h_i)^{\frac{1}{2}}} = \left(\frac{h_i}{1-h_i}\right)^{\frac{1}{2}} rse_i$$

sendo que o quociente $\dfrac{h_i}{1-h_i}$ é chamado potencial de influência e é uma medida da distância do ponto X em relação às demais observações. Belsley, Kuh & Welsch (1980, pág. 28) sugerem que valores absolutos excedendo $2\sqrt{\dfrac{p}{n}}$ podem identificar observações influentes.

d) **Distância de Cook**

É também uma medida de afastamento do vetor de estimativas provocado pela retirada da observação i. É uma expressão muito semelhante ao $DFFitS$ mas que usa como estimativa da variância residual aquela obtida com todas as n observações, ou ainda, usa o resíduo estudentizado internamente. É dada por:

$$D_{(i)} = \frac{(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(i)})^T (\boldsymbol{X}^T \boldsymbol{X})(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(i)})}{ps^2} = \frac{h_i}{(1-h_i)^2} \frac{r_i^2}{ps^2} = \left(\frac{r_i}{(1-h_i)^{\frac{1}{2}} s}\right)^2 \frac{h_i}{(1-h_i)} \frac{1}{p}$$

ou, ainda,

$$D_{(i)} = \frac{h_i}{(1-h_i)} \frac{1}{p} rsi_i^2.$$

e) **Distância de Cook modificada**

Atkinson (1981, pág. 25) sugere uma modificação para a distância de Cook

$$C_i = \left(\frac{n-p}{p} \frac{h_i}{1-h_i}\right)^{\frac{1}{2}} |rse_{(i)}| = \left(\frac{n-p}{p}\right)^{\frac{1}{2}} DFFitS_{(i)}.$$

4.4 Tipos de gráficos

a) **Valores observados (Y) vs variáveis explanatórias (X_j)** - Esse tipo de gráfico indica a estrutura que pode existir entre a variável dependente e as diversas covariáveis. Pode indicar, também, a presença de heterocedasticidade. Pode, porém, levar a uma idéia falsa no caso de muitas covariáveis (a não ser que haja ortogonalidade entre todas).

b) **Variável explanatória X_j vs variável explanatória $X_{j'}$** - Esse tipo de gráfico pode indicar a estrutura que pode existir entre duas variáveis explanatórias. Pode indicar, também, a presença de heterocedasticidade. Pode, porém, levar a uma idéia falsa no caso de muitas covariáveis (a não ser que haja ortogonalidade entre todas).

c) **Resíduos vs variáveis explanatórias não incluídas (X_{fora})** - Pode mostrar se existe uma relação entre os resíduos do modelo ajustado e uma variável ainda não incluída no modelo. Pode mostrar, também, a evidência de heterocedasticidade. Pode levar, porém, ao mesmo tipo de problema apontado nos itens (a) e (b). Uma alternativa melhor para esse tipo de gráfico é o gráfico da variável adicionada (*Added variable plot*).

d) **Resíduos vs variáveis explanatórias incluídas (X_{dentro})** - Pode mostrar se ainda existe uma relação sistemática entre os resíduos e a variável X_j já incluída no modelo, isto é, por exemplo se X^2_{dentro} deve ser incluída. Esse tipo de gráfico apresenta o mesmo tipo de problema que o citado nos itens (a), (b) e (c). Alternativa melhor para isso é o gráfico de resíduos parciais (*Partial residual plot*). O padrão para esse tipo de gráfico é uma distribuição aleatória de média 0 e amplitude constante. Desvios sistemáticos podem indicar :

   - termo quadrático (ou ordem superior) faltando,
   - escala errada da variável explanatória.

e) **Resíduos vs valores ajustados** - O padrão para esse tipo de gráfico é uma distribuição aleatória de média 0 e amplitude constante. Pode mostrar heterogeneidade de variâncias.

f) **Gráficos de índices** - servem para localizar observações com resíduo, h (*leverage*), distância de Cook modificada etc, grandes.

g) **Gráfico da variável adicionada ou da regressão parcial (*Added variable plot*)** Embora os gráficos de resíduos vs variáveis não incluídas no modelo possam indicar a necessidade de variáveis extras no modelo, a interpretação exata deles não é clara. A dificuldade está em que, a menos que a variável explanatória, considerada para inclusão, seja ortogonal a todas as variáveis que já estão no modelo, o coeficiente angular do gráfico de resíduos não é o mesmo que o coeficiente angular no modelo ajustado, incluindo a variável em questão. Esse tipo de gráfico pode ser usado para detectar a relação com uma variável explanatória e como isto é influenciado por observações individuais. No caso do modelo linear geral, tem-se

$$\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{X\theta} + \gamma U$$

sendo U a variável a ser adicionada (pode ser uma variável construída). O interesse está em se saber se $\gamma = 0$, isto é, se não há necessidade de se incluir a variável U no modelo.

A partir do sistema de equações normais

$$\begin{bmatrix} \boldsymbol{X}^T \\ U^T \end{bmatrix} \begin{bmatrix} \boldsymbol{X} & U \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\theta}} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{Y} \\ U^T\boldsymbol{Y} \end{bmatrix} \Rightarrow \begin{cases} \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\theta}} + \boldsymbol{X}^T U\hat{\gamma} &= \boldsymbol{X}^T\boldsymbol{Y} \\ U^T\boldsymbol{X}\hat{\boldsymbol{\theta}} + U^T U\hat{\gamma} &= U^T\boldsymbol{Y} \end{cases}$$

tem-se

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T U\hat{\gamma}$$

e

$$\hat{\gamma} = \frac{U^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}}{U^T(\boldsymbol{I}-\boldsymbol{H})U} = \frac{U^T(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}}{U^T(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{I}-\boldsymbol{H})U} = \frac{\boldsymbol{u}^{*T}\boldsymbol{r}}{\boldsymbol{u}^{*T}\boldsymbol{u}^*}$$

que é o coeficiente angular de uma reta que passa pela origem e em que $\boldsymbol{r} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\theta}} = (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}$ são os resíduos de $\boldsymbol{Y}$ ajustado para $\boldsymbol{X}$ e $\boldsymbol{u}^* = (\boldsymbol{I}-\boldsymbol{H})U$ são os resíduos de U ajustado para $\boldsymbol{X}$.

O gráfico da variável adicionada (*added variable plot*) de $\boldsymbol{r}$ versus $\boldsymbol{u}^*$ tem coeficiente angular $\hat{\gamma}$ (diferente do gráfico de $\boldsymbol{r}$ vs U). Ele pode mostrar, também, como a evidência para a inclusão de U depende de observações individuais.

Esse gráfico, portanto, é obtido a partir dos resíduos ordinários da regressão de $\boldsymbol{Y}$ como função de todas as covariáveis, exceto $U = X_j$, versus os resíduos ordinários da regressão de $U = X_j$ como função das mesmas covariáveis usadas para modelar $\boldsymbol{Y}$. Assim, por exemplo, para um modelo com 3 covariáveis, o gráfico da variável adicionada para X_3 é obtido a partir de

$$\hat{\boldsymbol{\mu}} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 \Rightarrow \boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{\mu}}$$

e

$$\hat{X}_3 = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 \Rightarrow \boldsymbol{u}^* = X_3 - \hat{X}_3$$

h) **Gráfico de resíduos parciais ou gráfico de resíduos mais componente (*Partial residual plot*)** - Se o interesse está em se detectar uma estrutura omitida, tal como uma forma diferente de dependência em U, um gráfico usando U pode servir melhor. Esse gráfico também, tem coeficiente angular $\hat{\gamma}$. Consiste em se plotarem os resíduos do modelo $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\theta} + \gamma U$ mais $\hat{\gamma}U$ versus U, isto é, no gráfico de $\tilde{\boldsymbol{r}} = \boldsymbol{r} + \hat{\gamma}U$ versus U. Por isso ele, também, é chamado de gráfico do resíduo mais componente.

i) **Gráficos normal e semi-normal de probabilidades (*Normal Plots* e *Half Normal Plots*)** - Segundo Weisberg (1985) o gráfico normal de probabilidades destaca-se por dois aspectos:

- identificação da distribuição originária dos dados e
- identificação de valores que se destacam no conjunto.

Seja uma amostra aleatória de tamanho n. As estatísticas de ordem correspondentes aos resíduos obtidos a partir do ajuste de um determinado modelo a essa amostra são $d_{(1)}, d_{(2)}, \ldots, d_{(i)}, \ldots, d_{(n)}$.

O fundamento geral para a construção do gráfico normal de probabilidades é que se os valores de uma dada amostra provêm de uma distribuição normal, então os valores das estatísticas de ordem e os Z_i correspondentes, obtidos da distribuição normal padrão são linearmente relacionados. Portanto, o gráfico dos valores, $d_{(i)}$ versus Z_i deve ser uma reta, aproximadamente. Formatos aproximados comuns que indicam ausência de normalidade são:

- **S (Esse)** - indica distribuições com caudas muito curtas, isto é, distribuições cujos valores estão muito próximos da média,
- **S invertido (Esse invertido)** - indica distribuições com caudas muito longas e, portanto, presença de muitos valores extremos,
- **J e J invertido** - indicam distribuições assimétricas, positivas e negativas, respectivamente.

Esses gráficos, na realidade são muito dependentes do número de observações, atingindo a estabilidade quando o número de observações é grande (em torno de 300). Para a construção desse gráfico seguem-se os passos:

i) ajuste um determinado modelo a um conjunto de dados e obtenha $d_{(i)}$, os valores ordenados de uma certa estatística de diagnóstico (resíduos, distância de Cook, h etc);

ii) dada a estatística de ordem na posição (i), calcule a respectiva probabilidade acumulada p_i e o respectivo quantil, ou seja, o inverso da função de distribuição normal $\Phi(.)$, no ponto p_i. Essa probabilidade p_i é, em geral, aproximada por

$$p_i = \frac{i - c}{n - 2c + 1}$$

sendo $0 < c < 1$. Diversos valores têm sido propostos para a constante c. Vários autores recomendam a utilização de $c = \frac{3}{8}$, ficando, então,

$$Z_i = \Phi^{-1}\left(\frac{i - 0,375}{n + 0,25}\right), i = 1, 2, \ldots, n.$$

iii) coloque, em um gráfico, $d_{(i)}$ versus Z_i.

Esse gráfico tem, também, o nome de *Q-Q plot*, por relacionar os valores de um quantil amostral (d_i) versus os valores do quantil correspondente da distribuição normal (Z_i).

A construção do gráfico semi-normal de probabilidades (*half normal plot*) é o resultado do conjunto de pontos obtidos por valores $|d|_{(i)}$ versus Z_i onde $Z_i = \Phi^{-1}\left(\frac{i+n-0,125}{2n+0,5}\right)$.

McCullagh & Nelder (1989) sugerem o uso do gráfico normal de probabilidades para resíduos e o gráfico semi-normal de probabilidades (*half normal plot*) para medidas positivas como é o caso de h (medida de *leverage*) e da distância de Cook modificada. No caso do gráfico normal de probabilidades para resíduos, espera-se que na ausência de pontos discrepantes, o aspecto seja linear, mas não há razão para se esperar que o mesmo aconteça quando são usados h ou a distância de Cook modificada. Os valores extremos aparecerão nos extremos do gráfico, possivelmente com valores que desviam da tendência indicada pelos demais.

Para auxiliar na interpretação do gráfico semi-normal de probabilidades (*half normal plot*), Atkinson (1985) propôs a adição de um envelope simulado. Este envelope é tal que sob o modelo correto as quantias (resíduos, *leverage*, distância de Cook etc) obtidas a partir dos dados observados caem dentro do envelope. Esse gráfico é obtido, seguindo-se os passos:

i) ajuste um determinado modelo a um conjunto de dados e obtenha $d_{(i)}$, os valores absolutos ordenados de uma certa estatística de diagnóstico (resíduos, distância de Cook, h (*leverage*) etc);

ii) simule 19 amostras da variável resposta, usando as estimativas obtidas após um determinado modelo ser ajustado aos dados e os mesmos valores para as variáveis explanatórias;

iii) ajuste o mesmo modelo a cada uma das 19 amostras e calcule os valores absolutos ordenados da estatística de diagnóstico de interesse, $d^*_{j(i)}$, $j = 1, \ldots, 19$, $i = 1, \ldots, n$;

iv) para cada i, calcule a média, o mínimo e o máximo dos $d^*_{j(i)}$;

v) coloque em um gráfico as quantidades obtidas no item anterior e $d_{(i)}$ versus Z_i.

Demétrio & Hinde (1997) apresentam um conjunto de macros que permitem fazer esses gráficos para uma grande variedade de modelos, usando o GLIM.

j) **Valores observados (Y) ou Resíduos versus tempo** - Mesmo que tempo não seja uma variável incluída no modelo, gráficos de respostas (Y) ou de resíduos versus tempo devem ser feitos sempre que possível. Esse tipo de gráfico pode levar à detecção de padrões não suspeitados, devido ao tempo ou, então, a alguma variável altamente correlacionada com tempo.

4.5 Exemplo - Regressão linear simples

Continuando a análise dos dados do Exemplo 2 do item 1.3.1 referentes a doses de de fósforo orgânico X e medidas de fósforo disponível (Y) no solo, pode-se ver na Figura 4.2 um conjunto padrão de gráficos de resíduos e diagnósticos, obtido pelo uso do pacote estatístico R, após o ajuste do modelo, com os comandos

```
par(mfrow=c(2,2))
plot(mod2)
```

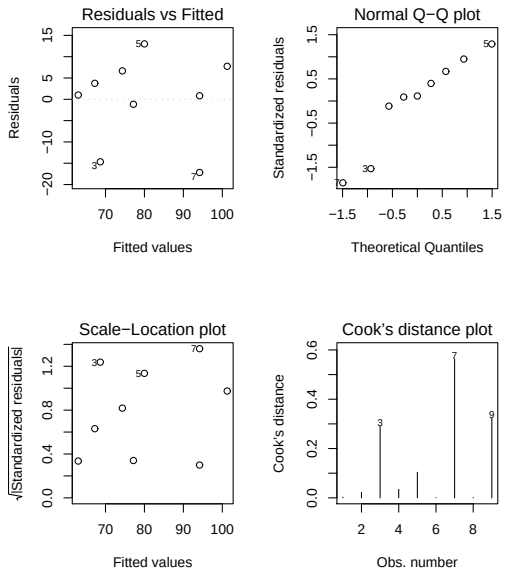


Figura 4.2: Conjunto padrão de gráficos do R para diagnósticos em regressão.

Observa-se que em todos os gráficos há destaque para as observações 3, 5 e 7. Entretanto, nenhuma delas teve valor de distância de Cook D_i maior do que $F_{\{50\%;2;7\}} = 0,7665478$, ou seja, nenhuma observação é considerada influente segundo esse critério.

A Figura 4.3 mostra alguns gráficos, usados para o estudo informal da normalidade dos resíduos, e que podem ser obtidos com os comandos

```
par(mfrow=c(1,3))
qqnorm(rstudent(mod2),ylab="Residuos padr. ext.", main="")
qqline(rstudent(mod2))

hist(rstudent(mod1), probability="TRUE", main="")
```

```
lines(density(rstudent(mod2)))
rug(rstudent(mod2))
boxplot(rstudent(mod2))
```

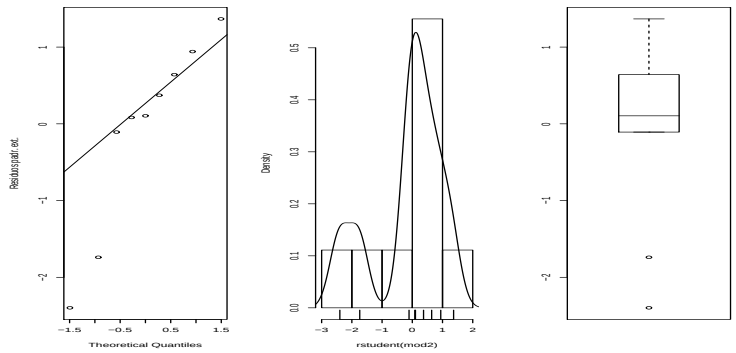


Figura 4.3: Gráfico normal de probabilidades, histograma e *boxplot* dos resíduos estudentizados externamente (rse).

Pela análise da Figura 4.3 suspeita-se da presença de dois valores discrepantes e da bimodalidade da distribuição dos resíduos. Convém ressaltar, no entanto, que os número de observações é pequeno e assim, mesmo supondo que a distribuição dos erros seja normal, a probabilidade de aparecerem valores aparentemente discrepantes é grande. A partir dos resultados do teste de Shapiro-Wilk apresentados a seguir, conclui-se que não há evidências para dizer que a distribuição residual não seja normal, considerando o nível de significância de 5%, pois `p-value = 0.1683 > 0.05`.

```
shapiro.test(rstudent(mod2)) #  Teste de normalidade de Shapiro-Wilk

        Shapiro-Wilk normality test

data:  rstudent(mod2)
W = 0.8829, p-value = 0.1683
```

Analisando o gráfico do perfil de verossimilhança para o modelo de transformação de Box-Cox apresentado na Figura 4.7, observa-se que o intervalo de confiança para λ inclui o valor 1, o que sugere a não necessidade de transformação da variável Y.

Apesar de se ter notado a presença de duas observações aparentemente atípicas no gráfico normal de probabilidades (Figura 4.3) o gráfico de probabilidades dos resíduos estudentizados externamente com envelope simulado, apresentado na Figura 4.5 à esquerda, mostra que tais valores não possuem valores grandes de resíduo pois estão dentro do envelope. Tal

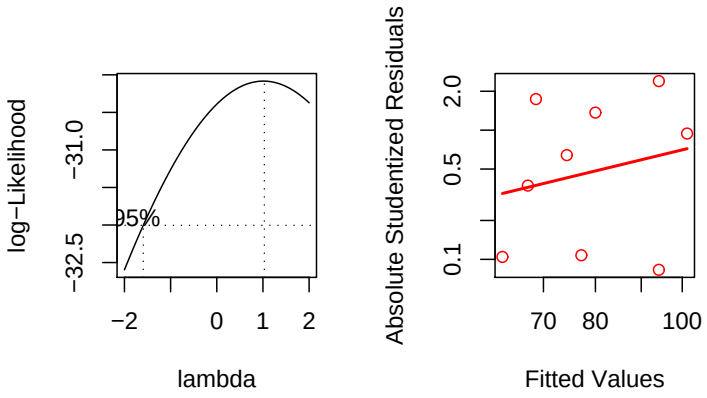


Figura 4.4: Gráfico do perfil de verossimilhança para o modelo de transformação de Box-Cox e gráfico dos valores absolutos de rse *vs* valores ajustados em escala log-log.

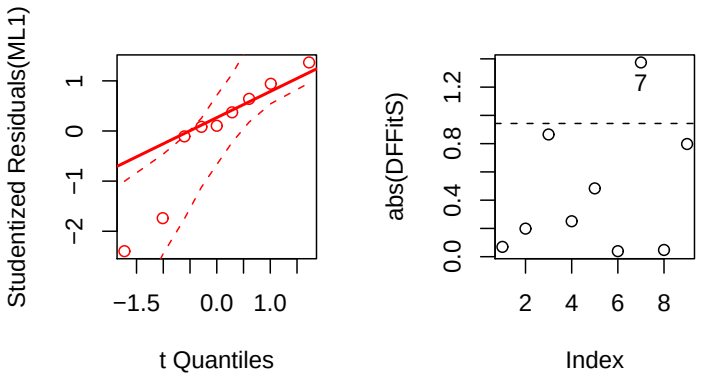


Figura 4.5: Gráfico de probabilidades dos resíduos estudentizados externamente com envelope simulado e gráfico dos valores absolutos de DFFitS *vs* índice, com limite dado por $2\sqrt{p/n} = 0,943$.

informação pode ser confirmada se se realizar o teste para o valor de rse, apresentado a seguir, considerando o valor p de Bonferroni.

```
> outlier.test(mod2)
 max|rstudent| df unadjusted p Bonferroni p
      2.397595  6   0.05346982    0.4812284

Observation: 7  # Não é significativo pois Bonferroni p = 0.4812284 >0.05
```

Por outro lado, analisando o gráfico dos valores absolutos de DFFitS *vs* índice, nota-se que a observação índice 7 é potencialmente influente sobre o ajuste do modelo. Analisando as medidas padrões de influência do R, apresentadas a seguir, duas observações são destacadas (1 e 7).

```
> IM1<-influence.measures(mod2) #Obs.: as medidas dffit e dfb estão estandartizadas
> IM1
Influence measures of
         lm(formula = Y ~ X) :

   dfb.1.     dfb.X   dffit cov.r   cook.d   hat inf
1  0.0695 -5.56e-02  0.0695 1.958 0.002816 0.307   *
2  0.1954 -1.40e-01  0.1986 1.670 0.022492 0.221
3 -0.8404  5.74e-01 -0.8649 0.750 0.290144 0.198
4  0.2143 -1.02e-01  0.2509 1.375 0.034364 0.133
5  0.2758  2.52e-17  0.4834 0.890 0.103919 0.111
6 -0.0287  8.45e-03 -0.0391 1.535 0.000888 0.117
7  0.3122 -1.02e+00 -1.3745 0.472 0.562807 0.247   *
8 -0.0108  3.53e-02  0.0476 1.804 0.001319 0.247
9 -0.3264  6.83e-01  0.7974 1.774 0.323156 0.418
> summary(IM1)
Potentially influential observations of
         lm(formula = Y ~ X) :

  dfb.1_ dfb.X   dffit cov.r   cook.d hat
1  0.07  -0.06    0.07  1.96_*  0.00   0.31
7  0.31  -1.02_* -1.37  0.47    0.56   0.25
```

Convém lembrar que o R considera a i-ésima observação influente se $|\text{DFBetaS}_i| > 1$, se $|\text{DFFitS}_i| > 3\sqrt{p/(n-p)}$, se $|1 - cov.r_i| > 3p/(n-p)$, se $D_i > F_{\{p;n-p\}}$, ou se $h_i > 3p/n$. Para o exemplo em questão, o R considera a i-ésima observação influente se $|\text{DFBetaS}_i| > 1$, se $|\text{DFFitS}_i| > 1,603$, se $|1 - COVRATIO| > 0,857$ se $D_i > F_{\{50\%;2;7\}} = 0,7665478$ ou se $h_i > 0,667$.

Segundo os resultados do teste a seguir, como o valor de $p = 0,5942836 > 0,05$, conclui-se que não há evidências de que haja heterogeneidade de variâncias a um nível de significância de 5% de probabilidade.

```
> # Homogeneidade de variâncias? #
> ncv.test(mod2) # Non-constant Variance Score Test
Non-constant Variance Score Test
```

```
Variance formula: ~ fitted.values
Chisquare = 0.2837042    Df = 1     p = 0.5942836
```

4.6 Exemplo - Regressão linear múltipla

Continuando a análise dos dados do Exemplo 5 do item 1.3.1 referentes a medidas de concentrações de fósforo inorgânico (X_1) e fósforo orgânico (X_2) no solo e de conteúdo de fósforo (Y) nas plantas crescidas naquele solo, pode-se ver na Figura 4.6 um conjunto padrão de gráficos de resíduos e diagnósticos, obtido pelo uso do pacote estatístico R, após o ajuste do modelo. Observa-se uma possível inadequação do modelo, destacando a 17ª observação como influente e com resíduo elevado. Sugere, por outro lado, que uma transformação da variável resposta Y talvez seja necessária.

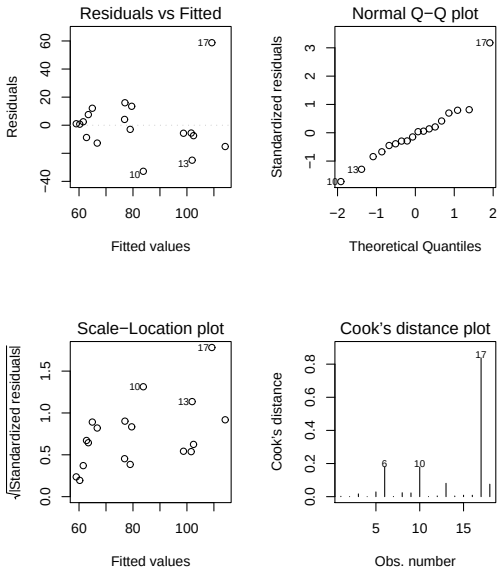


Figura 4.6: Conjunto padrão de gráficos para diagnósticos do R - Sem transformação.

4.7 Família Box-Cox de tranformações

Conforme já foi visto, o modelo linear clássico é válido sob as pressuposições:

(i) simplicidade de estrutura para o valor esperado da variável resposta (aditividade do modelo);

(ii) independência dos erros;

(iii) homogeneidade de variâncias e

(iv) normalidade aproximada de erros aditivos.

Se não for possível satisfazer a esses requisitos na escala original dos dados, pode ser que uma transformação não linear dos dados possa produzir homogeneidade de variâncias e distribuição aproximadamente normal.

Indicações para a necessidade de uma transformação de dados podem ser úteis. Assim, no caso de dados não-negativos, como, por exemplo, volumes, medidas de tempo até que um evento ocorra, pode ser que uma transformação logarítmica leve a uma distribuição aproximadamente normal. Naturalmente, se todos os dados estiverem longe do zero com uma dispersão pequena, a transformação não terá muito efeito. Se, porém, a razão entre o maior e o menor valores for em termos de potências de 10, uma transformação será desejável. Uma outra alternativa é o uso da teoria de Modelos Lineares Generalizados em que se usam outras distribuições, diferentes da normal, para a modelagem dos dados. Assim, o uso de uma transformação logarítmica é diferente do uso de uma distribuição normal com função de ligação logarítmica. No primeiro caso tem-se que os dados na escala original têm uma distribuição log-normal enquanto que no segundo têm distribuição normal (Aitkin *et al*, 1989).

Box & Cox (1964) propuseram uma família de transformações dada por

$$Y(\lambda) = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(Y) & \lambda = 0 \end{cases} \tag{4.1}$$

sendo λ o parâmetro da transformação e Y a variável resposta. Na ausência de uma transformação, $\lambda = 1$.

Para observações $(Y_i, \boldsymbol{x}_i^T)$, $i = 1,\ 2, \ldots,\ n$ e $\boldsymbol{x}_i^T = (X_{i1}, X_{i2}, \ldots, X_{ik})$ assume-se que existe algum λ tal que

$$Y_i(\lambda) \sim N(\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2), \quad i = 1, \ldots, n$$

ou seja, o objetivo é determinar λ (ou uma escala para Y), tal que sejam verdadeiras as pressuposições:

(i) os resíduos sejam normais;

(ii) a variância seja homogênea (constante) e

(iii) o modelo seja aditivo.

O método mais usado para a estimação do parâmetro λ é o método do perfil de verossimilhança. Consiste em considerar um modelo com $p+2$ parâmetros, isto é, os p parâmetros de $\boldsymbol{\beta}$, σ^2 e λ, e estimá-los em dois estágios:

(a) Para λ fixado, obter as estimativas de máxima verossimilhança para $\boldsymbol{\beta}(\lambda)$ e $\sigma^2(\lambda)$. Considerando-se que $Y_i(\lambda) \sim N(\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma^2)$, $\quad i = 1, \ldots, n$, então,

$$f(Y_i(\lambda); \boldsymbol{\beta}, \sigma^2, \lambda) = (2\pi\sigma^2)^{-1/2} \exp\{-[Y_i(\lambda) - \boldsymbol{x}_i^T\boldsymbol{\beta}]^T[Y_i(\lambda) - \boldsymbol{x}_i^T\boldsymbol{\beta}]/(2\sigma^2)\},$$

e a função de verossimilhança para o vetor $\boldsymbol{Y}(\lambda)$ é dada por:

$$L(\boldsymbol{\beta}, \sigma^2, \lambda; \boldsymbol{Y}) = (2\pi\sigma^2)^{-n/2} \exp\{-[\boldsymbol{Y}(\lambda) - \boldsymbol{X}\boldsymbol{\beta}]^T[\boldsymbol{Y}(\lambda) - \boldsymbol{X}\boldsymbol{\beta}]/(2\sigma^2)\}$$

ou, exprimindo-se em termos da variável original $\boldsymbol{Y}$, usando (4.1), fica

$$L(\boldsymbol{\beta}, \sigma^2, \lambda; \boldsymbol{Y}) = (2\pi\sigma^2)^{-n/2} \exp\{-[\boldsymbol{Y}(\lambda) - \boldsymbol{X}\boldsymbol{\beta}]^T[\boldsymbol{Y}(\lambda) - \boldsymbol{X}\boldsymbol{\beta}]/(2\sigma^2)\}\ J$$

sendo J, o jacobiano da transformação definido por:

$$J = \prod_{i=1}^{n} \frac{dY_i(\lambda)}{dY_i} = \prod_{i=1}^{n} Y_i^{\lambda-1}.$$

O logaritmo da função de verossimilhança fica, então,

$$\ell(\boldsymbol{\beta}, \sigma^2, \lambda; \boldsymbol{Y}) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}[\boldsymbol{Y}(\lambda) - \boldsymbol{X}\boldsymbol{\beta}]^T[\boldsymbol{Y}(\lambda) - \boldsymbol{X}\boldsymbol{\beta}] + (\lambda - 1)\sum_{i=1}^{n}\log(Y_i) + \textit{constante}.$$

Derivando-se $\ell(\boldsymbol{\beta}, \sigma^2, \lambda; \boldsymbol{Y})$ em relação a $\boldsymbol{\beta}$ e a σ^2, obtêm-se as derivadas parciais:

$$\begin{cases} \dfrac{\partial \ell}{\partial \boldsymbol{\beta}} &= \dfrac{1}{\sigma^2}\boldsymbol{X}^T[\boldsymbol{Y}(\lambda) - \boldsymbol{X}\boldsymbol{\beta}] \\ \dfrac{\partial \ell}{\partial \sigma^2} &= -\dfrac{n}{2\sigma^2} + \dfrac{1}{2(\sigma^2)^2}[\boldsymbol{Y}(\lambda) - \boldsymbol{X}\boldsymbol{\beta}]^T[\boldsymbol{Y}(\lambda) - \boldsymbol{X}\boldsymbol{\beta}] \end{cases}$$

que igualadas a 0 resultam em:

$$\hat{\boldsymbol{\beta}}(\lambda) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}(\lambda) \ \text{ e } \ \hat{\sigma}^2(\lambda) = \frac{S(\lambda)}{n} = \frac{1}{n}[\boldsymbol{Y}(\lambda) - \boldsymbol{X}\hat{\boldsymbol{\beta}}]^T[\boldsymbol{Y}(\lambda) - \boldsymbol{X}\hat{\boldsymbol{\beta}}]$$

A substituição da estimativa de máxima verossimilhança $\hat{\sigma}^2(\lambda)$ pela de quadrados mínimos, isto é, por $\hat{\sigma}^2(\lambda) = \dfrac{S(\lambda)}{n-p}$ não afeta o próximo passo.

(b) Substituindo-se os resultados obtidos em $\ell(\boldsymbol{\beta}, \sigma^2, \lambda; \boldsymbol{Y})$ obtém-se

$$l_{max}(\lambda) = -\frac{n}{2}\log(\hat{\sigma}^2) + (\lambda - 1)\sum_{i=1}^{n}\log(Y_i) + \textit{constante}$$

Essa expressão parcialmente maximizada, ou também chamada de perfil de verossimilhança (na realidade perfil do logaritmo da função de verossimilhança) é, portanto, uma função de λ que depende da soma de quadrados residual, $S(\lambda)$ e do jacobiano. Ocorre, porém, que as somas de quadrados residuais obtidas para diferentes valores de λ não são comparáveis, pois depende da grandeza das observações.

Uma forma mais simples para $l_{max}(\lambda)$, porém equivalente, é obtida, usando-se a forma normalizada da transformação, isto é, usando-se

$$Z(\lambda) = \frac{Y(\lambda)}{J^{1/n}}$$

e, sendo $J = \prod_{i=1}^{n} Y_i^{\lambda-1}$, tem-se que $J^{1/n} = \dot{Y}^{\lambda-1}$, em que $\dot{Y}$ é a média geométrica das observações. Logo,

$$Z(\lambda) = \frac{Y(\lambda)}{\dot{Y}^{\lambda-1}} = \begin{cases} \dfrac{Y^{\lambda}-1}{\lambda \dot{Y}^{\lambda-1}} & \lambda \neq 0 \\ \dot{Y} log(Y) & \lambda = 0 \end{cases} . \tag{4.2}$$

Verifica-se que o jacobiano dessa transformação, em relação à variável original, passa a ser igual a 1 e, portanto, o logaritmo da função de verossimilhança parcialmente maximizada pode ser escrito como

$$\ell_{max}(\lambda) = -\frac{n}{2}\log(\hat{\sigma}^2) + constante$$

sendo $\hat{\sigma}^2 = \dfrac{R(\lambda)}{n} = \dfrac{1}{n}[\boldsymbol{Z}(\lambda) - \boldsymbol{X}\hat{\boldsymbol{\beta}}]^T[\boldsymbol{Z}(\lambda) - \boldsymbol{X}\hat{\boldsymbol{\beta}}] = \dfrac{1}{n}\boldsymbol{Z}(\lambda)[\boldsymbol{I}(\lambda) - \boldsymbol{H}]\boldsymbol{Z}(\lambda)$ e $\hat{\boldsymbol{\beta}}(\lambda) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z}(\lambda)$. Tem-se, portanto, que $R(\lambda)$ é a soma de quadrados residual de $\boldsymbol{Z}(\lambda)$.

A estimativa de máxima verossimilhança, $\hat{\lambda}$, é o valor de λ que maximiza $l_{max}(\lambda)$, ou, equivalentemente, minimiza $R(\lambda)$. Tem-se, então, que um método prático para se obter $\hat{\lambda}$ segue os passos:

(i) escolhe-se uma grade de valores para λ, de -2 a 2;

(ii) para cada valor fixado de λ, obtêm-se $\hat{\boldsymbol{\beta}}(\lambda)$ e $R(\lambda)$ (ou $S(\lambda)$ se não for utilizada a transformação normalizada);

(iii) faz-se um gráfico dos pares $(\lambda, R(\lambda))$ ou $(\lambda, S(\lambda))$ se não for utilizada a transformação normalizada);

(iv) escolhe-se o valor de λ para o qual o gráfico passa pelo mínimo.

No *R* isso é feito, utilizando-se os comandos:

```
library(MASS)
boxcox(modelo, data= nome)
```

depois de ajustado o modelo desejado aos dados. O gráfico resultante mostra, também, um intervalo de confiança para λ, com um coeficiente de confiança igual a 95% de probabilidade. Há interesse, ainda em se obter um intervalo de confiança para λ por dois motivos. O primeiro deles é para verificar se o intervalo contém o valor $\lambda = 1$, o que indicaria não haver necessidade

de transformação. O segundo, para identificar se o intervalo cobre algum valor de λ cuja interpretação seja mais simples. Um intervalo de confiança, com um coeficiente de confiança de $100(1-\alpha)\%$ para λ é obtido a partir dos valores para os quais

$$\{\lambda : 2[\ell_{max}(\hat{\lambda}) - \ell_{max}(\lambda)] \leq \chi^2_{1,\alpha}\}.$$

Então, um intervalo de confiança, com um coeficiente de confiança aproximadamente igual a 95% de probabilidade, incluirá todos os valores de λ para os quais

$$\{\lambda : 2[\ell_{max}(\hat{\lambda}) - \ell_{max}(\lambda)] \leq 3,84\}.$$

Um procedimento prático seria traçar uma reta no gráfico de $\ell_{max}(\hat{\lambda})$ contra λ, passando pelo ponto

$$\ell_{max}(\hat{\lambda}) - \frac{1}{2}\, 3,84$$

A reta corta a curva em dois pontos e os λ's correspondentes a esses pontos formam o intervalo de confiança aproximado para λ.

Variável construída

Uma maneira de se verificar a necessidade de transformação de variáveis (resposta ou explanatórias) é por meio do uso de uma variável construída como uma variável adicional no modelo. Essa variável construída poderá ser usada com o teste da razão de verossimilhança ou, então, com o *added variable plot* ou o *partial residual plot*.

(i) Transformação para a variável resposta

Considerando-se a família de transformações dada pela expressão (4.2), tem-se que a expansão de $Z(\lambda)$ em série de Taylor em relação a λ_0 conhecido é:

$$Z(\lambda) \approx Z(\lambda_0) + (\lambda - \lambda_0)u(\lambda_0)$$

sendo $u(\lambda_0) = Z'(\lambda)|_{\lambda=\lambda_0}$. Então,

$$\boldsymbol{Z}(\lambda_0) = \boldsymbol{Z}(\lambda) - (\lambda - \lambda_0)\boldsymbol{u}(\lambda_0) = \boldsymbol{X}\boldsymbol{\beta} + \gamma\boldsymbol{u} + \boldsymbol{\varepsilon}$$

Mas, a partir de (4.2) tem-se:

$$u(\lambda) = Z'(\lambda) = \frac{\lambda\dot{Y}^{\lambda-1}Y^{\lambda}\log Y - (Y^{\lambda}-1)(\dot{Y}^{\lambda-1} + \lambda\dot{Y}^{\lambda-1}\log\dot{Y})}{(\lambda\dot{Y}^{\lambda-1})^2}$$

$$= \frac{Y^{\lambda}\log Y - (Y^{\lambda}-1)(\frac{1}{\lambda} + \log\dot{Y})}{\lambda\dot{Y}^{\lambda-1}}$$

Assim para o teste de $H_0 : \lambda = 1$, isto é, a não necessidade de transformação da variável resposta, a variável construída fica:

$$u(1) = Y(\log \frac{Y}{\dot{Y}} - 1)$$

enquanto que para o teste de $H_0 : \lambda = 0$, isto é, a necessidade de transformação logarítmica para a variável resposta, a variável construída fica:

$$u(0) = \dot{Y} \log Y(\frac{\log Y}{2} - \log \dot{Y})$$

(ii) Transformação para uma variável explanatória

Para a verificação da necessidade de transformação da variável explanatória X_k, faz-se:

$$\mathrm{E}(Y) = \sum_{j \neq k} \beta_j X_j + \beta_k X_k^{\lambda} = Z(\lambda)$$

e, usando-se expansão de $Z(\lambda)$ em série de Taylor em relação a λ_0, tem-se

$$Z(\lambda) \approx Z(\lambda_0) + (\lambda - \lambda_0) Z'(\lambda_0) = \sum_{j \neq k} \beta_j X_j + \beta_k X_k^{\lambda_0} + \beta_k (\lambda - \lambda_0) X_k^{\lambda_0} \log X_k$$

pois, $Z'(\lambda) = \beta_k X_k^{\lambda_0} \log X_k$. Portanto, testar $H_0 : \lambda = \lambda_0$ é equivalente a testar $H_0 : \gamma = 0$ para a regressão com a variável construída $u = X^{\lambda_0} \log X_k$ com X^{λ_0}, já incluída no modelo. Assim, para o teste de $H_0 : \lambda = 1$, tem-se:

$$\mathrm{E}(Y) = \sum_{j=1}^{p} \beta_j X_j + (\lambda - 1) \beta_k X_k \log X_k = \boldsymbol{X\beta} + \gamma \boldsymbol{u}$$

em que $u = X_k \log X_k$ é a variável construída. A influência de observações individuais na evidência de uma transformação pode ser pesquisada, usando o gráfico da variável adicionada.

(iii) Transformação simultânea das variáveis resposta e explanatórias

Para a transformação simultânea da variável resposta e de todas as variáveis explanatórias (exceto a constante $1^{\lambda} = 1$) à mesma potência, a variável construída $u^{+}(\lambda_0)$ para $\lambda_0 = 1$ é dada por:

$$u = \sum_{j=2}^{p} \beta_j X_j \log X_j - u(1)$$

em que $u(1) = y(\log \frac{Y}{\dot{Y}} - 1)$ e X_1 é referente ao termo constante.

4.8 Exemplos

(i) Continuando a análise dos dados do Exercício 2 item 1.3.1 referentes a doses de de fósforo orgânico X e medidas de fósforo disponível (Y) no solo, pode-se ver pelo gráfico do perfil de verossimilhança para o modelo de transformação de Box-Cox apresentado na Figura 4.7 que o intervalo de confiança para λ, com um coeficiente de confiança de 95% de probabilidade, inclui o valor 1, o que sugere a não necessidade de transformação da variável Y.

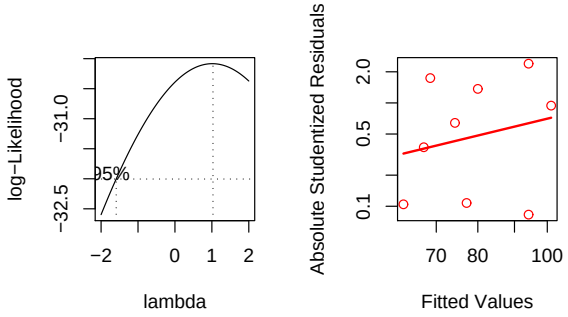


Figura 4.7: Gráfico do perfil de verossimilhança para o modelo de transformação de Box-Cox e gráfico dos valores absolutos de rse *vs* valores ajustados em escala log-log.

(ii) Continuando a análise dos dados do Exemplo 5 do item 1.4 referentes a medidas de concentrações de fósforo inorgânico (X_1) e fósforo orgânico (X_2) no solo e de conteúdo de fósforo (Y) nas plantas crescidas naquele solo, o gráfico do perfil de verossimilhança para a transformação de Box-Cox (ver Figura 4.8), sugere a transformação $Y^{-\frac{1}{2}} = 1/\sqrt{Y}$. Adotando essa transformação, o novo modelo fica:

$$\mathrm{E}\left(\frac{1}{\sqrt{Y}}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2. \tag{4.3}$$

Com a transformação sugerida, o novo conjunto de gráficos para diagnósticos do R apresentado na Figura 4.9 não revela maiores problemas além da presença de uma observação potencialmente influente, a 6ª.

A análise de variância e os testes para os parâmetros do modelo podem, então, ser obtidas, usando-se o R.

```
> Anova(ML2) ######## Atenção!! Não utilizar, neste caso, o comando anova(),
que faz o ajuste seqüencial
Anova Table (Type II tests)

Response: 1/sqrt(Y)
```

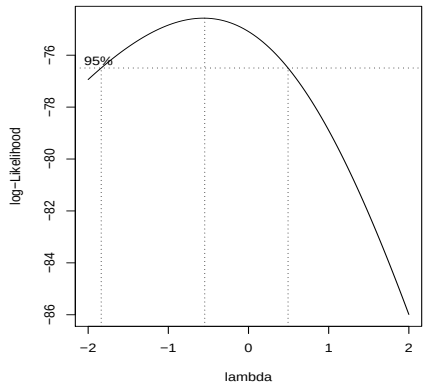


Figura 4.8: Gráfico do perfil de verossimilhança para a transformação de Box-Cox .

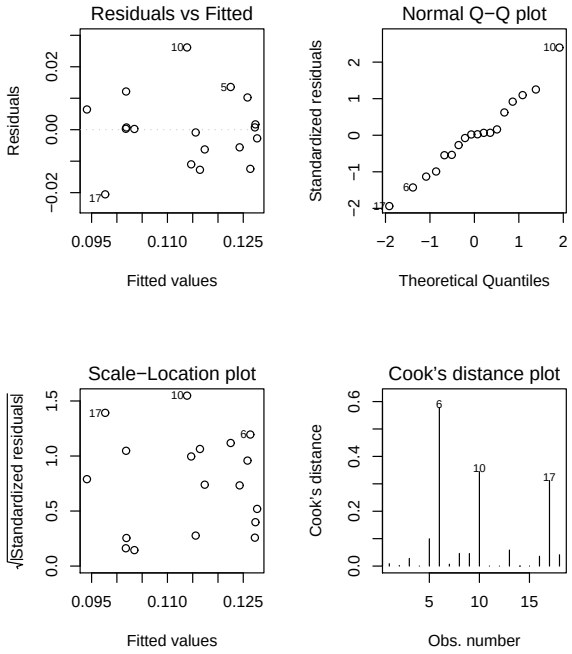


Figura 4.9: Conjunto padrão de gráficos para diagnósticos do R - Dados transformados ($Y^{\frac{1}{2}}$)

```
             Sum Sq Df F value   Pr(>F)
X1        0.00179467  1 12.7947 0.002753 **
X2        0.00000032  1  0.0023 0.962599
Residuals 0.00210400 15
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> summary(ML2)

Call:
lm(formula = 1/sqrt(Y) ~ X1 + X2)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0205370 -0.0060990  0.0002617  0.0052467  0.0261430

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.276e-01  9.342e-03  13.659 7.23e-10 ***
X1          -1.141e-03  3.189e-04  -3.577  0.00275 **
X2           1.133e-05  2.377e-04   0.048  0.96260
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01184 on 15 degrees of freedom
Multiple R-Squared: 0.5171,     Adjusted R-squared: 0.4527
F-statistic: 8.031 on 2 and 15 DF,  p-value: 0.004256
```

Pode-se concluir, a partir dessas análises, que a variável X_2 não é significativa, dado que a variável X_1 está no modelo. Assim, o novo modelo sugerido passa a ser o de regressão linear simples:

$$E\left(\frac{1}{\sqrt{Y}}\right) = \beta_0 + \beta_1 X_1 \tag{4.4}$$

que pode ser ajustado, no R, por meio dos seguintes comandos:

```
> ML3<-lm(1/sqrt(Y)~X1)
> par(mfrow=c(2,2))
> plot(ML3)
> summary(ML3)

Call:
lm(formula = 1/sqrt(Y) ~ X1)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0204612 -0.0061195  0.0002214  0.0051938  0.0263184

Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1279927  0.0042439  30.159 1.58e-15 ***
X1          -0.0011336  0.0002739  -4.138 0.000772 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01147 on 16 degrees of freedom
Multiple R-Squared: 0.517,      Adjusted R-squared: 0.4868
F-statistic: 17.13 on 1 and 16 DF,  p-value: 0.0007717

> anova(ML3)
Analysis of Variance Table

Response: 1/sqrt(Y)
          Df     Sum Sq    Mean Sq F value    Pr(>F)
X1         1 0.00225255 0.00225255  17.127 0.0007717 ***
Residuals 16 0.00210432 0.00013152
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

que revelam que este modelo é um modelo adequado.

Podem-se construir, ainda, outros gráficos para verificar se o modelo 4.4 está adequado, através dos seguintes comandos:

```
par(mfrow=c(2,2))
# Gráfico dos valores observados contra ajustados #
plot(fitted(ML3),1/sqrt(Y))
identify(fitted(ML3),1/sqrt(Y))

# Gráfico abs(DFFitS) vs índice #
plot(abs(dffits(ML3)))
abline(3*sqrt(ML3$rank/ML3$df.residual),0,lty=2)
identify(1:n,abs(dffits(ML3)))

# Gráfico quantil quantil com envelope simulado #
qq.plot(ML3,simulate=TRUE,reps=1000)

# Gráfico do perfil de verossimilhança para a transformação de Box-Cox #
boxcox(ML3)
```

que produzem os gráficos da Figura 4.10 que confirmam a adequação do modelo final (4.4).

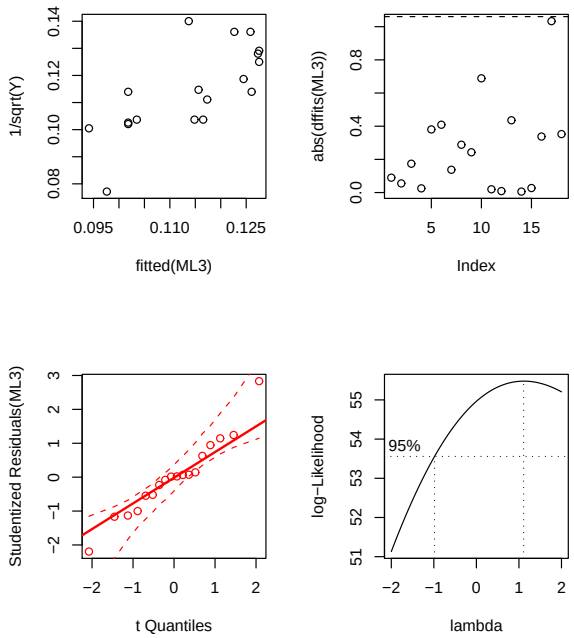


Figura 4.10: Conjunto extra de gráficos para diagnósticos do R - Modelo final (4.4)

(iii) Considere os dados do exercício 8 do item 1.3.1 referentes a um experimento de irrigação em batata plantada em terra roxa estruturada (solo argiloso) em que foram medidas as lâminas (L, mm) de água a diferentes distâncias do aspersor e as correspondentes produtividades (P, t/ha).

Ajustando-se o modelo de regressão linear simples aos dados de irrigação, observa-se, claramente, a influência da 12ª observação sobre o ajuste (ver Figura 4.11), confirmada pelo seu valor grande de distância de Cook (valor maior do que $F_{\{50\%;2;10\}} = 0,7434918$) no último gráfico da figura 4.12. Note que em todos os quatro gráficos dessa figura, tal observação é destacada, sendo que o primeiro gráfico sugere, ainda, que talvez o modelo não esteja adequado, o que pode ser observado pela nuvem de pontos em formato de $\bigcap$ e o terceiro não sugere a necessidade de uma eventual transformação da variável resposta P.

De modo a solucionar os problemas levantados por esta análise inicial, pode-se continuar a análise por dois caminhos diferentes: procurar um novo modelo ou analisar os dados sem a observação destacada. Convém ressaltar que esta última alternativa deverá somente ser escolhida se o pesquisador que realizou o experimento comprovar que houve algum tipo de falha relativa a essa observação. Partindo para o ajuste do modelo de regressão quadrática aos dados tem-se que os gráficos de diagnóstico (não apresentados) revelam problemas semelhantes aos anteriores. Quando se ajusta o modelo de regressão cúbica, no entanto, os gráficos de diagnóstico (figura 4.13), aparentemente, não revelam nenhum problema.

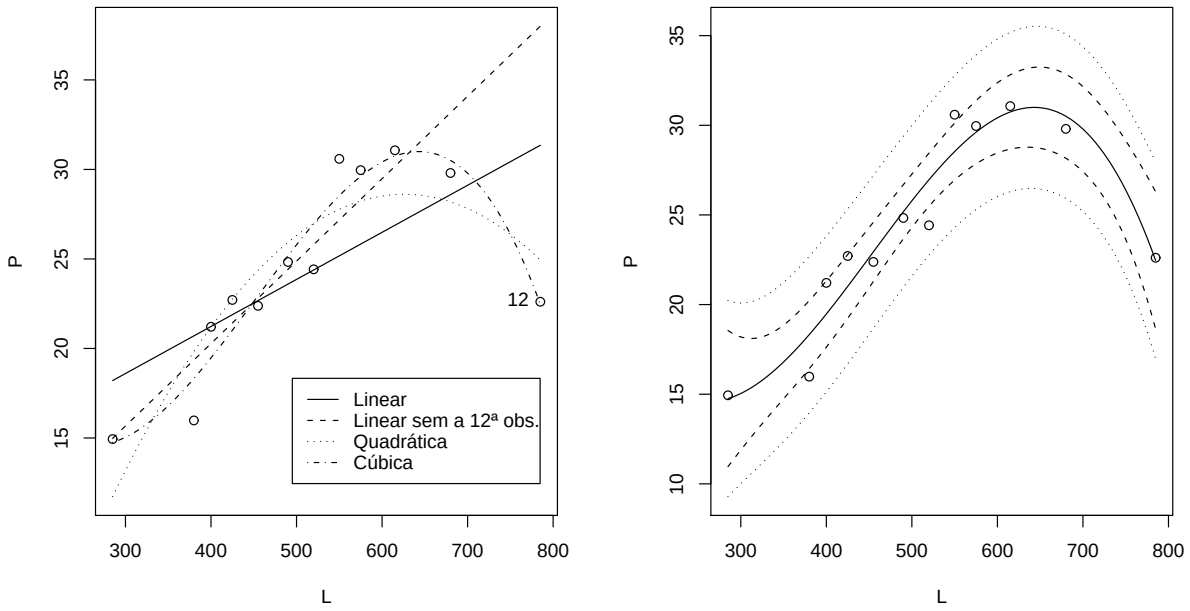


Figura 4.11: Gráfico de dispersão e modelos ajustados e gráfico de dispersão com intervalos de 95% de confiança para a resposta média $E(P)$ $(- - -)$ e para a resposta predita $\hat{P}$ $(\cdots)$, supondo o modelo de regressão cúbica.

Tabela 4.1: Estimativas dos parâmetros dos modelos de regressão adotados, com respectivos erros padrões entre parênteses, r^2,$r^2_{aj.}$ e AIC.

	Estimativas dos parâmetros						
Regressão	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	r^2	$r^2_{aj.}$	AIC
Linear	10,71349	0,02629			0,4523	0,3975	72,35493
	(4,850392)	(0,009148)					
Linear sem a 12ª obs.	1,85402	0,04605			0,8672	0,8524	52,10291
	(3,007421)	(0,006007)					
Quadrática	-28,29	0,1818	-0,0001452		0,8086	0,7661	61,73754
	(9,996)	(0,03841)	(3,547e-05)				
Cúbica	48,69	-0,3009	0,0008074	-5,945e-07	0,928	0,901	52,00372
	(22,11)	(0,1348)	(0,0002625)	(1,632e-07)			

Resumo do ajuste do modelo de regressão cúbica (ver programa em capítulo sobre o R)

```
ML3<-lm(P~L+I(L^2)+I(L^3))
par(mfrow=c(2,2))
plot(ML3)
```

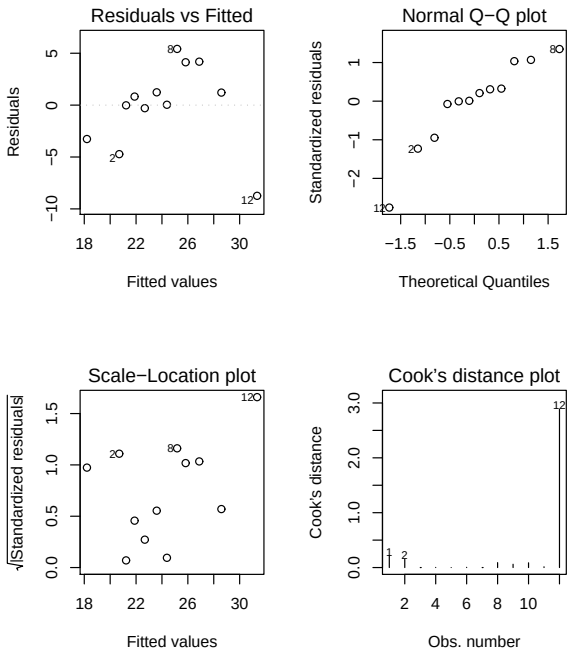


Figura 4.12: Conjunto padrão de gráficos do R para diagnósticos em regressão - Regressão linear.

```
> summary(ML3)

Call:
lm(formula = P ~ L + I(L^2) + I(L^3))

Residuals:
    Min      1Q  Median      3Q     Max
-2.5419 -0.5974  0.1654  0.6854  2.0551

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.869e+01  2.211e+01   2.202  0.05880 .
L           -3.009e-01  1.348e-01  -2.231  0.05617 .
I(L^2)       8.074e-04  2.625e-04   3.076  0.01522 *
I(L^3)      -5.945e-07  1.632e-07  -3.643  0.00656 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

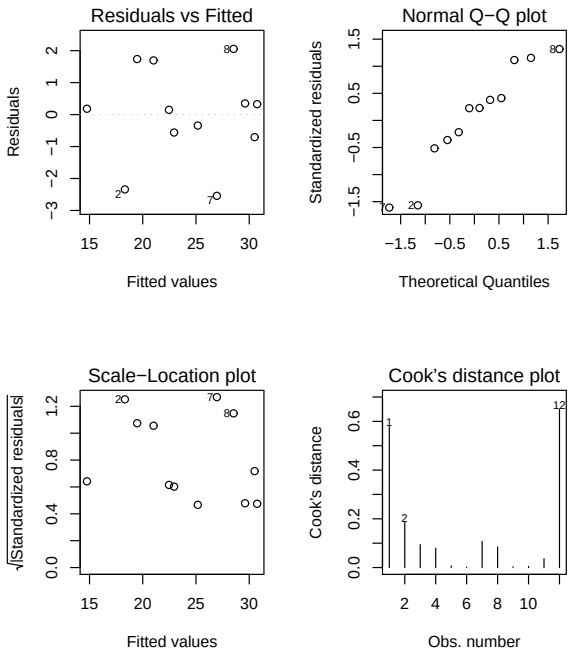


Figura 4.13: Conjunto padrão de gráficos do R para diagnósticos em regressão - Regressão cúbica.

```
Residual standard error: 1.706 on 8 degrees of freedom
Multiple R-Squared: 0.928,      Adjusted R-squared: 0.901
F-statistic: 34.38 on 3 and 8 DF,  p-value: 6.414e-05
```

Análise de variância para o modelo de regressão cúbica.

```
> anova(ML3)
Analysis of Variance Table

Response: P
          Df  Sum Sq Mean Sq F value    Pr(>F)
L          1 146.247 146.247  50.270 0.0001030 ***
I(L^2)     1 115.203 115.203  39.599 0.0002346 ***
I(L^3)     1  38.604  38.604  13.269 0.0065634 **
Residuals  8  23.274   2.909
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Medidas de influência

```
> summary(influence.measures(ML3))
```

Tabela 4.2: Tempos de sobrevivência de ratos após envenenamento.

Tempo	Tipo	Trat.	Tempo	Tipo	Trat.	Tempo	Tipo	Trat.	Tempo	Tipo	Trat.
0,31	1	1	0,45	1	1	0,46	1	1	0,43	1	1
0,82	1	2	1,10	1	2	0,88	1	2	0,72	1	2
0,43	1	3	0,45	1	3	0,63	1	3	0,76	1	3
0,45	1	4	0,71	1	4	0,66	1	4	0,62	1	4
0,36	2	1	0,29	2	1	0,4	2	1	0,23	2	1
0,92	2	2	0,61	2	2	0,49	2	2	1,24	2	2
0,44	2	3	0,35	2	3	0,31	2	3	0,40	2	3
0,56	2	4	1,02	2	4	0,71	2	4	0,38	2	4
0,22	3	1	0,21	3	1	0,18	3	1	0,23	3	1
0,30	3	2	0,37	3	2	0,38	3	2	0,29	3	2
0,23	3	3	0,25	3	3	0,24	3	3	0,22	3	3
0,30	3	4	0,36	3	4	0,31	3	4	0,33	3	4

```
Potentially influential observations of
        lm(formula = P ~ L + I(L^2) + I(L^3)) :

   dfb.1_  dfb.L dfb.I(L^2 dfb.I(L^3 dffit cov.r   cook.d
1   1.02_* -0.90  0.80     -0.73      1.44 22.98_*  0.58
12 -0.40    0.47 -0.54      0.62      1.52 30.42_*  0.65
   hat
1   0.93
12  0.95
```

Exemplo 1:

Os dados da Tabela 4.2, referem-se a tempos de sobrevivência de ratos após envenenamento com 4 tipos de venenos e 3 diferentes tratamentos (Box e Cox, 1964). Como pode ser constatado na Figura 4.14, os dados sem transformação apresentam heterogeneidade de variâncias que é amenizada quando se usam os inversos dos valores observados ou os valores observados elevados a -3/4.

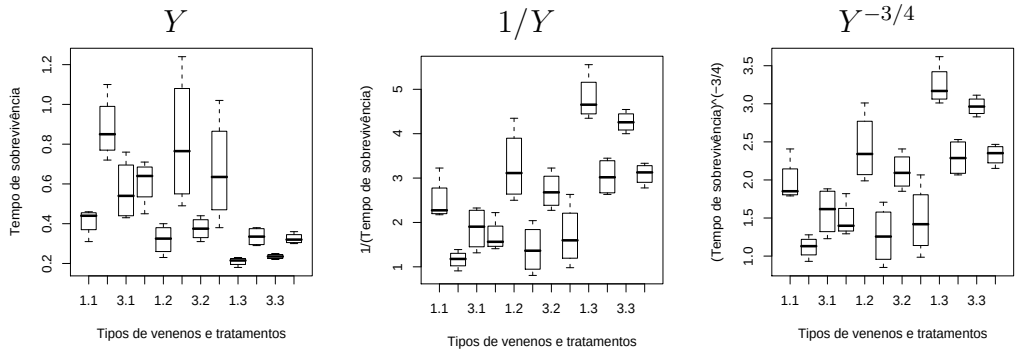


Figura 4.14: Box-plots para as observações da Tabela 4.2, com e sem transformação.

Usando-se o modelo

$$Tempo_{ij} = Tipo_i + Trat_j + Tipo.Trat_{ij} + \varepsilon_{ij}$$

em que $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$, e o gráfico para verificar a necessidade de uma transformação na família Box-Cox, obtém-se $\hat{\lambda} = -0,75$ conforme a Figura 4.15. Entretanto, o valor $\hat{\lambda} = -1$ está no intervalo de confiança e $1/Y$ tem uma interpretação melhor nesse caso, isto é, representa a taxa de mortalidade.

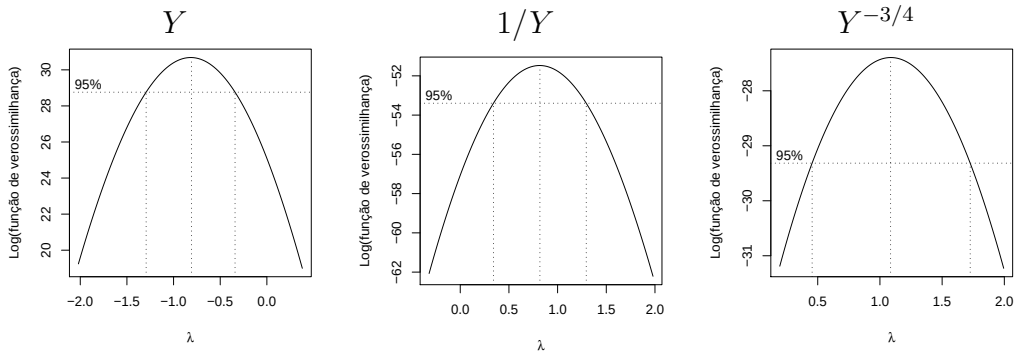


Figura 4.15: Gráficos para escolha de transformação na família Box-Cox, Tabela 4.2.

Ajustando-se, também, os modelos

$$\frac{1}{Tempo_{ij}} = Tipo_i + Trat_j + Tipo.Trat_{ij} + \epsilon_{ij}$$

e

$$Tempo_{ij}^{-3/4} = Tipo_i + Trat_j + Tipo.Trat_{ij} + \delta_{ij}$$

em que $\epsilon_{ij} \sim \mathrm{N}(0, \tau^2)$ e $\delta_{ij} \sim \mathrm{N}(0, \zeta^2)$, obtêm-se os outros dois gráficos da Figura 4.15, mostrando que o valor $\hat{\lambda} = 1$ está incluído no intervalo de confiança e que, portanto, ambas as transformações tornaram a escala da variável da variável resposta adequada. A Figura 4.16, com os gráficos dos valores ajustados *versus* valores observados sem e com transformação, dos valores ajustados *versus* resíduos e gráficos normais de probabilidades, confirma esses resultados, mostrando claramente a falta de ajuste para o caso do modelo normal para a variável sem transformação e que ambas as transformações resolveram o problema de heterogeneidade de variâncias e da falta de normalidade da variável resposta. Outros modelos, supondo distribuição normal com função de ligação inversa, distribuições gama e normal inversa, foram usados e deram resultados piores.

Os resultados da Tabela 4.3 mostram que em ambos os casos existem evidências de efeito significativo de tipo de veneno e de tratamentos mas não de interação entre eles. Entretanto, a evidência é muito mais forte para o caso em que foram feitas as transformações 1/tempo e tempo$^{-3/4}$.

Exemplo 2: Considere os dados do Exercício 5 do item 1.4.1 (página 16) referentes a medidas de diâmetro à altura do peito (D) e altura (H) de árvores (*black cherry*) em pé e de volume (V) de árvores derrubadas. O objetivo desse tipo de experimento é verificar de que forma essas variáveis estão relacionadas para, através de medidas nas árvores em pé, poder se predizer o volume de madeira em uma área de floresta.

4.9 Transformação e função de ligação

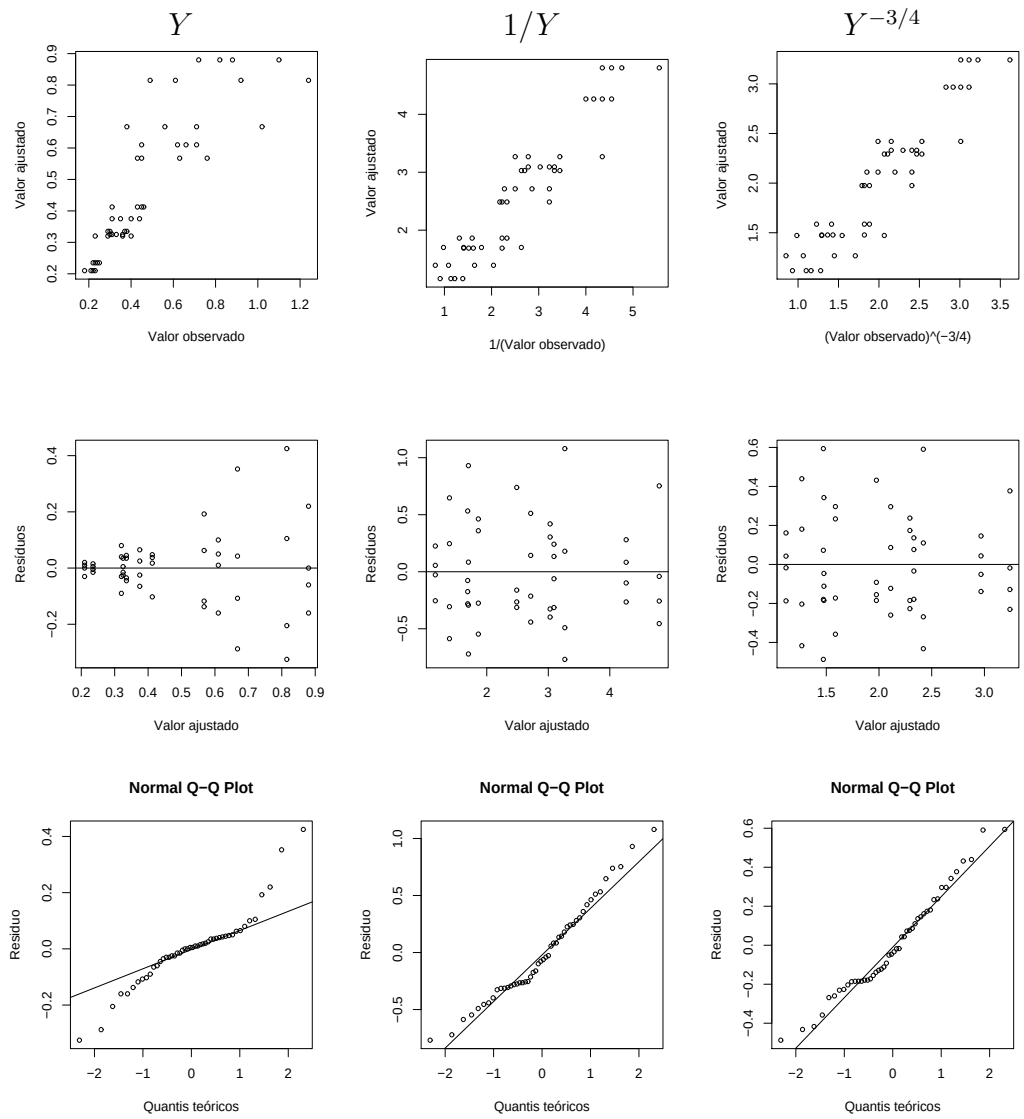


Figura 4.16: Gráficos dos valores ajustados *versus* valores observados sem e com transformação, dos valores ajustados *versus* resíduos e gráficos normais de probabilidades, Tabela 4.2.

A Tabela 4.4 refere-se a produções médias diárias de gordura (kg/dia) no leite de uma única vaca durante 35 semanas (McCulloch, 2001). É comum supor que a produção média de gordura Y_i tem distribuição com média

$$\mu_i = \alpha t_i^{\beta} e^{\gamma t_i},$$

em que t representa a semana, α, β e γ são parâmetros desconhecidos.

Portanto,

$$\log \mu_i = \log \alpha + \beta \log(t_i) + \gamma_i t_i,$$

o que mostra a necessidade do uso de função de ligação logarítmica. Pode-se supor ainda que $Y_i \sim$

Tabela 4.3: Análise da variância para os tempos de sobrevivência de ratos após envenenamento, sem e com transformação inversa, Tabela 4.2.

		Tempo			1/Tempo			$\text{Tempo}^{-3/4}$		
Fonte	GL	SQ	QM	F	SQ	QM	F	SQ	QM	F
Tipo	2	1,0330	0,5165	23,27**	34,877	17,439	72,46**	11,9261	5,9630	68,45**
Tratamento	3	0,9212	0,3071	16,71**	20,414	6,805	28,35**	7,1579	2,3860	27,39**
Interação	6	0,2501	0,0417	1,88	1,571	0,262	1,09	0,4859	0,0810	0,93
Resíduo	36	0,8007	0,0222		8,643	0,240		3,1361	0,0871	

Tabela 4.4: Produções médias diárias de gordura (kg/dia) do leite de uma vaca.

0.31	0.39	0.50	0.58	0.59	0.64	0.68
0.66	0.67	0.70	0.72	0.68	0.65	0.64
0.57	0.48	0.46	0.45	0.31	0.33	0.36
0.30	0.26	0.34	0.29	0.31	0.29	0.20
0.15	0.18	0.11	0.07	0.06	0.01	0.01

$\text{N}(\mu_i, \tau^2)$, isto é,

$$Y_i = \mu_i + \delta_i = \alpha t_i^{\beta} e^{\gamma t_i} + \delta_i,$$

em que $\delta_i \sim \text{N}(0, \tau^2)$. Isso equivale, portanto, ao MLG em que a variável resposta Y tem distribuição normal com função de ligação logarítmica e preditor linear que é igual a $\log \alpha + \beta \log(t_i) + \gamma_i t_i$.

Entretanto, na prática é comum supor que $\log(Y_i) \sim \text{N}(\log \mu_i, \sigma^2)$, isto é,

$$\log(Y_i) = \log \mu_i + \epsilon_i = \log \alpha + \beta \log(t_i) + \gamma t_i + \epsilon_i,$$

em que $\epsilon_i \sim \text{N}(0, \sigma^2)$. Isso equivale, portanto, ao MLG em que a variável resposta $\log(Y)$ tem distribuição normal com função de ligação identidade e mesmo preditor linear $\log \alpha + \beta \log(t_i) + \gamma_i t_i$.

A Figura 4.17 mostra que a distribuição normal com função de ligação logarítmica produz um melhor ajuste do que adotar uma transformação logarítmica dos dados e supor uma distribuição normal com função de ligação identidade. Isso é confirmado nos gráficos de valores ajustados *versus* valores observados apresentado na Figura 4.18. O programa em R encontra-se no Apêndice.

4.10 Exercícios

1. Para cada variável Y apresentada na Tabela 20, ajuste o modelo de regressão linear simples e:

   (a) Obtenha, para $i = 1, \ldots, n$:

      i. os resíduos ordinários (r_i)

      ii. os resíduos estudentizados internamente (rsi_i)

      iii. os resíduos padronizados externamente (rse_i)

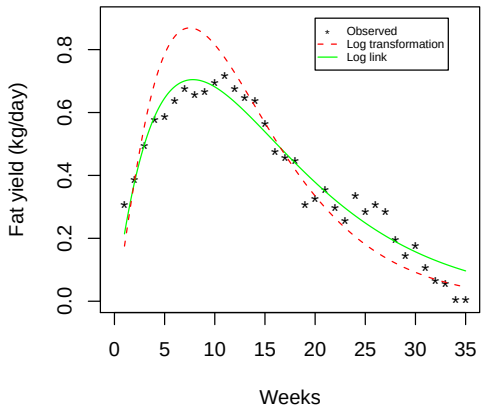


Figura 4.17: Valores observados e curvas ajustadas

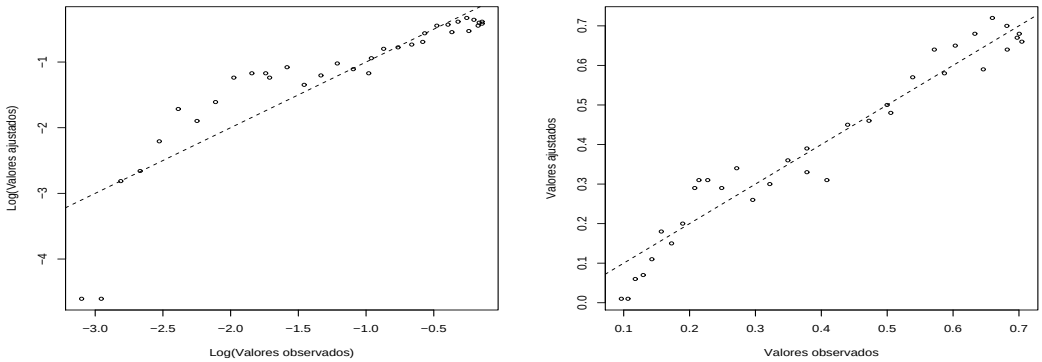


Figura 4.18: Gráficos de valores ajustados *versus* valores observados obtidos para o modelo normal para $\log(Y)$ com função de ligação identidade e para o modelo normal para Y com função de ligação logarítmica.

iv. os elementos h_{ii} da diagonal da matriz de projeção H

v. os valores $DFBetaS_{(i)}$, para β_0 e β_1, e $DFFitS_{(i)}$

vi. os valores da distância de Cook, $D_{(i)}$, e da distância de Cook modificada, $C_{(i)}$

(b) Construa os gráficos:

i. de Y *vs* X

ii. de rse *vs* valores ajustados

iii. de índices para h_{ii}, DFFitS$_{(i)}$ e $D_{(i)}$

(c) Responda:

i. Há pontos de alavanca? São bons ou ruins?

ii. Há pontos inconsistentes?

iii. Há pontos influentes?

iv. Há *outliers*?

Tabela 4.5: Valores de X_i e Y_{ui}, $(u = 1, \ldots, 4)$, $(i = 1, \ldots, 9)$.

X	Y_1	Y_2	Y_3	Y_4
0	9,6	10,4	10,4	18,8
1	8,4	11,2	11,4	15,4
2	12,6	12,8	15,0	13,7
3	15,1	13,9	15,9	11,9
4	22,8	17,2	18,6	9,1
5	19,1	21,6	21,5	11,6
6	21,6	22,4	21,7	8,9
7	24,0	22,9	25,3	12,2
12	33,6	35,4	30,6	16,8

Programa SAS:

```
options nodate ps=25; data aula7a; input X Y1 Y2 Y3 Y4; cards;
 0    9.6   10.4   10.4   18.8
 1    8.4   11.2   11.4   15.4
 2   12.6   12.8   15.0   13.7
 3   15.1   13.9   15.9   11.9
 4   22.8   17.2   18.6    9.1
 5   19.1   21.6   21.5   11.6
 6   21.6   22.4   21.7    8.9
 7   24.0   22.9   25.3   12.2
12   33.6   35.4   30.6   16.8 ; proc reg;
     model Y1 Y2 Y3 Y4=X/p influence r;
     plot (Y1 Y2 Y3 Y4)*X /collect hplots=2 vplots=2;
     plot (STUDENT. RSTUDENT.)*P./collect hplots=2 vplots=1;
     plot (H. DFFITS. COOKD.)*obs./collect hplots=3 vplots=1;
run;
```

2. Prove que

$$C_i = \left(\frac{n-p}{p}\right)^{\frac{1}{2}} |DFFitS_{(i)}|,$$

ou seja, a distância de Cook modificada, proposta por Atkinson (1981) é proporcional ao valor absoluto da medida $DFFitS$, proposta por Belsey *et al* (1980, p.15).

3. Sabe-se que a região de $100(1-\gamma)\%$ de confiança para o vetor de parâmetros $\boldsymbol{\theta}$ é dada pelos valores de $\boldsymbol{\theta}$ tais que

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T X^T X (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \leq p s^2 F_{\{p,\nu,\gamma\}}$$

sendo $s^2 = QMRes$ uma estimativa de σ^2 com $\nu = GLRes$ graus de liberdade e F, o percentil $100\gamma\%$ da distribuição F com p e ν graus de liberdade. Essa região é delimitada por um elipsóide cujo volume, dependente do determinante de $X^T X$ e de s^2, é dado por

$$E \propto \{s^{2p}/|X^T X|\}^{1/2}.$$

Quando, no entanto, a i-ésima observação é deletada, o volume passa a ser

$$E_{(i)} \propto \{s_{(i)}^{2p}/|X_{(i)}^T X_{(i)}|\}^{1/2}.$$

Como uma forma de se determinar a influência da i-ésima observação sobre o volume do elipsóide de confiança, Belsey *et al* (1980, p.22) propuseram a medida chamada $COVRATIO$, dada por

$$COVRATIO = \left\{ \frac{E_{(i)}}{E_i} \right\}^2.$$

Alternativamente, Cook e Weisberg (1982, p.159) usam o logaritmo da razão entre os volumes, porém ajustado pela razão dos valores F da definição das regiões de confiança. Segundo os autores e Atkinson (1987,p.227), essas medidas são confiáveis como medidas de influência geral? Justifique e ilustre por meio de um exemplo.

Gráficos quantil-quantil (*QQ-plot*)

(a) Dados os valores de X: -3, -1, 1, 3, 5, 7,..., 23, 25, pede-se:

i. Simular valores de Y através da função

$$Y = 100 - 3X + e,$$

onde $e \sim N(0, 1)$. Pede-se:

ii. Ajustar, aos dados gerados, o modelo de regessão linear

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

e obter os resíduos padronizados externamente rse_i.

iii. Verificar se os resíduos calculados em (b) têm distribuição normal através do gráfico normal de probabilidades (*normal plot* ou *normal quantile-quantile plot* ou ainda, *QQ-plot*).

iv. Verificar se os resíduos calculados em (b) têm distribuição normal através do *normal probability-probability plot* ou *PP-plot*.

(b) Dado o conjunto de observações {9, 7, 3, 2, 8, 4} da variável aleatória X, construa o gráfico quantil-quantil supondo cada uma das seguintes distribuições:

i. $f(x) = 0,02xI_{[0;10]}(x)$

ii. $f(x) = 0,1I_{[0;10]}(x)$

iii. $f(x) = 0,04xI_{[0;5]}(x) + (0,4 - 0,04x)I_{[5;10]}(x)$

Observação: Considerar, como função de distribuição empírica,

$$\hat{F}(x_{(i)}) = \frac{i - \frac{3}{8}}{n + \frac{1}{4}}, \tag{4.5}$$

sendo $x_{(i)}$ a i-ésima observação do conjunto de n observações colocadas em ordem crescente.

(c) Dado o conjunto de observações {9, 7, 3, 2, 8, 4} da variável aleatória X, construa o gráfico quantil-quantil supondo distribuição uniforme discreta no intervalo [2;9]. Observação: Considere (4.5) como função de distribuição empírica.

(d) Gere uma amostra aleatória de tamanho $n = 20$ de uma variável aleatória X com distribuição normal com parâmetros $\mu = 10$ e $\sigma = 2$.

i. Obtenha os valores da variável padronizada $Z = \frac{X-\bar{x}}{s}$ para essa amostra, sendo $\bar{x}$ a média e s o desvio padrão amostrais.

ii. Construa o gráfico quantil-quantil para Z supondo a distribuição t de Student com $n - 1 = 19$ graus de liberdade e considerando (4.5) como função de distribuição empírica.

iii. Repita o item anterior considerando a normal padronizada e compare o gráfico obtido com o anterior.

iv. Construa o gráfico formado pelos pontos de coordenadas $\left(F^{-1}\left(\frac{1}{2} + \frac{1}{2}\frac{i-\frac{3}{8}}{n+\frac{1}{4}}\right), |z|_{(i)}\right)$, $i = 1, \ldots, n$, sendo $F^{-1}(.)$ a função inversa da função de distribuição acumulada t de Student com $n - 1 = 19$ graus de liberdade.

v. Repita o item anterior supondo que $F^{-1}(.)$ a função inversa da função normal padronizada e compare o gráfico obtido com o anterior. Observação: Este gráfico é chamado *half-normal plot*.

(e) Apresente um revisão bibliográfica sobre o emprego da constante c na função de distribuição empírica

$$\hat{F}(x_{(i)}) = \frac{i - c}{n - 2c + 1},$$

sendo $x_{(i)}$ a i-ésima observação do conjunto de n observações colocadas em ordem crescente. Qual valor de c é utilizado para a construção do gráfico quantil-quantil para a distribuição normal no R? e no SAS? Observações repetidas podem causar problemas? Discuta.

(f) Construir um envelope simulado para cada uma das distribuições apresentadas no item 3b, considerando o coeficiente de confiança 95%.

Gráfico da variável adicionada (*added-variable plot*) ou da regressão parcial (*partial-regression plot*) e gráfico de resíduos parciais (*partial-residual plot*) ou gráfico de resíduos mais componente (*component + residual plot*)

i. Considerando os valores de (X_1, X_2, X_3) apresentados na Tabela 16 (página 99), simule valores de Y_2 através da função

$$Y_2 = 10 + 3X_1 - 2X_2 + e$$

onde $e \sim N(0, 1)$. Pede-se:

A. Ajustar, aos dados gerados, o modelo

$$Y_{2i} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

B. Construir e interpretar o gráfico da variável adicionada (*added variable plot* ou, equivalentemente, *partial regression leverage plot*), para as variáveis X_1, X_2 e X_3.

C. Construir e interpretar o gráfico dos resíduos parciais (*partial residual plot*) para as variáveis X_1, X_2 e X_3.

D. Ajustar, aos dados gerados, o modelo

$$Y_{2i} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i.$$

ii. Considerando os valores de (X_1, X_2, X_3) apresentados na Tabela 16 (página 99), simule valores de Y_3 através da função

$$Y_3 = 5 + 3X_1 - 2X_2 + X_3 + 0,5X_2^2 + e$$

onde $e \sim N(0, 1)$. Pede-se:

A. Ajustar, aos dados gerados, o modelo

$$Y_{3i} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

B. Construir e interpretar o gráfico da variável adicionada (*added variable plot* ou, equivalentemente, *partial regression leverage plot*), para as variáveis X_1, X_2 e X_3.

C. Construa e interprete o gráfico dos resíduos parciais (*partial residual plot* para as variáveis X_1, X_2 e X_3.

D. Baseando-se nas conclusões obtidas a partir dos gráficos construídos nos itens (b) e (c), proponha um modelo aos valores de Y_3 gerados no item (a).

Tabela 8.1.

i	X_{i1}	X_{i2}	X_{i3}
1	-2	2	-2
2	-1	-1	0
3	-1	0	0
4	-1	0	0
5	-1	1	0
6	1	-1	0
7	1	0	0
8	1	0	0
9	1	1	0
10	0	0	-1
11	0	0	0
12	0	0	0
13	0	0	1
14	2	-2	2

(g) Considerando os valores de (X_1, X_2, X_3) apresentados na tabela 8.1, simule os $n = 14$ valores de Y por meio da função $Y_i = 10 + 3X_{i1}^2 - 2X_{i2} + 0X_{i3} + \varepsilon_i$ (i=1,...,n), onde $\varepsilon_i \sim N(0, 1)$, e $Cov(\varepsilon_i, \varepsilon_{i'}) = 0$, $i \neq i'$. Pede-se:

i. Ajustar, aos dados gerados, o modelo

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

(i=1,...,n), considerando-se $\varepsilon_i \sim N(0, \sigma^2)$, e $Cov(\varepsilon_i, \varepsilon_{i'}) = 0$, $i \neq i'$.

ii. Construir e interpretar os gráficos da variável adicionada (ou gráfico da regressão parcial) para as variáveis X_1, X_2 e X_3.

iii. Construir e interpretar os gráficos de resíduos parciais (ou gráfico de resíduos mais componente), para as variáveis X_1, X_2 e X_3.

iv. Com base nos gráficos dos itens anteriores, propor um novo modelo e refazer os gráficos. Repetir o processo até que não ocorram problemas.

4. O arquivo de dados *trees*, disponível no pacote R, contém os dados de 31 cerejeiras (*Black cherry*) da Floresta Nacional de Allegheny, relativos a três variáveis: volume de madeira útil ($Volume$), em pés cúbicos; altura ($Height$), em pés, e circunferência ($Girth$) a $4,5$ pés ($1,37$ metros) de altura. Pede-se:

   (a) Ajustar o modelo $Volume_i = \beta_0 + \beta_1 Girth_i + \beta_2 Height_i + \varepsilon_i$ (i=1,...,n), considerando-se $\varepsilon_i \sim N(0, \sigma^2)$ e $Cov(\varepsilon_i, \varepsilon_{i'}) = 0$, $i \neq i'$.

   (b) Construir e interpretar os **gráficos das variáveis adicionadas** e **gráficos de resíduos mais componente** para as variáveis circunferência e altura. Há algum problema?

   (c) Verificar se há a necessidade de transformação da variável resposta por meio do teste da razão de verossimilhança para λ da família Box-Cox de transformações.

   (d) Verificar se há a necessidade de transformação da variável resposta por meio do gráfico da variável adicionada para a variável construída para a transformação de Box-Cox. Este gráfico também é chamado **gráfico da variável construída** para a transformação de Box-Cox.

   (e) Verificar se há a necessidade de transformação das variáveis preditoras. (Transformação de Box-Tidwell)

Capítulo 5

Correlações lineares simples e parciais

5.1 Correlação linear simples

5.1.1 Introdução

Até esse ponto o interesse estava em se estudar por meio da relação linear, qual a influência de uma variável fixa X, ou um conjunto de variáveis fixas $X_1, X_2, \ldots, X_k$, sobre uma variável aleatória Y. Assim, enquanto que na análise de regressão é indispensável identificar a variável dependente, nos problemas de correlação isto não se faz necessário.

Aqui, o interesse está em se estudar o grau de relacionamento entre as variáveis X e Y, isto é, uma *medida* de covariabilidade entre elas. Assim, análise de correlação difere da análise de regressão em dois pontos básicos:

i) Em primeiro lugar não existe a idéia de que uma das variáveis é dependente de uma outra ou de um conjunto de outras variáveis. A correlação é considerada como uma medida de influência mútua ou conjunta entre variáveis, ou seja, não se está preocupado em verificar quem influencia ou quem é influenciado. A análise de regressão, como vem sendo tratada até aqui, tem por objetivo encontrar equações que indiquem o valor médio de Y para valores fixados de X.

ii) Na análise de correlação todas as variáveis são, geralmente, aleatórias e a amostra é considerada proveniente de uma distribuição conjunta dessas variáveis. Evidentemente, a distribuição normal bidimensional é, geralmente, suposta para a variável bidimensional (X, Y).

A princípio poderia ser tomada a covariância entre X e Y como uma medida para essa relação. No entanto, pela própria definição

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)]$$

nota-se que a covariância pode assumir qualquer valor real. Desse modo, torna-se difícil interpretar seus valores: o que é uma relação fraca ou forte? Uma medida alternativa para isso é o coeficiente de correlação de Pearson.

5.1.2 Distribuição normal bidimensional

É interessante, inicialmente, estudar algumas das propriedades da distribuição normal bidimensional. Seja $\boldsymbol{X} = (X_1, X_2)^T$ um vetor aleatório com distribuição normal bivariada com vetor de médias $\boldsymbol{\mu}$ e matriz de variâncias e covariâncias $\boldsymbol{\Sigma}$ (simétrica, positiva definida), isto é,

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_{X_1} \\ \mu_{X_2} \end{bmatrix} \quad \text{e} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2_{X_1} & \sigma_{X_1X_2} \\ \sigma_{X_1X_2} & \sigma^2_{X_2} \end{bmatrix}$$

Função de densidade conjunta - A função de densidade conjunta da distribuição normal bidimensional é dada por

$$\begin{aligned} f_{X_1,X_2}(x_1,x_2) &= \frac{1}{2\pi\sigma_{X_1}\sigma_{X_2}(1-\rho^2)^{1/2}} \exp\left[-\frac{1}{2(1-\rho^2)} \left(\frac{X_1-\mu_{X_1}}{\sigma_{X_1}} \right)^2 \right. \\ &- \left. 2\rho \left(\frac{X_1-\mu_{X_1}}{\sigma_{X_1}} \right) \left(\frac{X_2-\mu_{X_2}}{\sigma_{X_2}} \right) + \left(\frac{X_2-\mu_{X_2}}{\sigma_{X_2}} \right)^2 \right] \end{aligned}$$

Fazendo-se

$$Z_{X_1} = \frac{X_1-\mu_{X_1}}{\sigma_{X_1}} \quad \text{e} \quad Z_{X_2} = \frac{X_2-\mu_{X_2}}{\sigma_{X_2}}$$

tem-se

$$f_{Z_{X_1},Z_{X_2}}(z_{x_1}, z_{x_2}) = \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp\left[-\frac{1}{2(1-\rho^2)} (z^2_{X_1} - 2\rho z_{X_1} z_{X_2} + z^2_{X_2}) \right]$$

variáveis centradas padronizadas e ρ o coeficiente de correlação entre X_1 e X_2.

Geometricamente, essa função é uma superfície em forma de sino cujo formato depende das variâncias e do coeficiente de correlação.

Funções de densidades marginais - A função de densidade marginal de X_j, $j = 1, 2$, pode ser obtida através de:

$$f_{Z_{X_1}}(z_{x_1}) = \frac{1}{2\pi(1-\rho^2)^{1/2}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2(1-\rho^2)} (z^2_{X_1} - 2\rho z_{X_1} z_{X_2} + z^2_{X_2}) \right] dz_{X_2}.$$

Mas

$$z^2_{X_1} - 2\rho z_{X_1} z_{X_2} + z^2_{X_2} = z^2_{X_1} - 2\rho z_{X_1} z_{X_2} + z^2_{X_2} + \rho^2 z^2_{X_1} - \rho^2 z^2_{X_1} = (z_{X_2} - \rho z_{X_1})^2 + z^2_{X_1}(1-\rho^2)$$

Portanto,

$$\begin{aligned} f_{Z_{X_1}}(z_{x_1}) &= \frac{1}{2\pi(1-\rho^2)^{1/2}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2} z_{X_1}^2\right] \exp\left[-\frac{1}{2(1-\rho^2)}(z_{X_2} - \rho z_{X_1})^2\right] dz_{X_2} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} z_{X_1}^2\right] \frac{1}{\sqrt{2\pi}(1-\rho^2)^{1/2}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2(1-\rho^2)}(z_{X_2} - \rho z_{X_1})^2\right] dz_{X_2} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} z_{X_1}^2\right] \end{aligned} \tag{5.1}$$

pois, para $u = \frac{1}{\sqrt{(1-\rho^2)}}(z_{X_2} - \rho z_{X_1})$, tem-se $du = \frac{1}{\sqrt{(1-\rho^2)}} dz_{X_2}$, $z_{X_2} \to -\infty \Rightarrow u \to -\infty$, $z_{X_2} \to \infty \Rightarrow u \to \infty$ e, então,

$$\frac{1}{\sqrt{2\pi}\sigma_{X_2}(1-\rho^2)^{1/2}} \int_{-\infty}^{\infty} exp\left[-\frac{1}{2(1-\rho^2)}(z_{X_2} - \rho z_{X_1})^2\right] dz_{X_2} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2} u^2\right] du = 1.$$

Logo,

$$f_{X_1}(x_1) = \frac{1}{\sqrt{2\pi}\sigma_{X_1}} \exp\left[-\frac{1}{2}\left(\frac{X_1 - \mu_{X_1}}{\sigma_{X_1}}\right)^2\right].$$

Tem-se, portanto, que $X_j \sim \mathrm{N}(\mu_{X_j}, \sigma_{X_j}^2)$.

Funções de densidades condicionais - A função de densidade condicional de X_j dado $X_{j'} = x'_j$, $j \neq j' = 1, 2$, pode ser obtida através de:

$$\begin{aligned} f_{X_1|X_2=x_2}(x_1|x_2) &= \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)} \\ &= \frac{1}{\sqrt{2\pi}\sigma_{X_1}(1-\rho^2)^{1/2}} \exp\left\{-\frac{1}{2}\left[\frac{1}{(1-\rho^2)}(z_{X_1}^2 - 2\rho z_{X_1} z_{X_2} + z_{X_2}^2) - z_{X_2}^2\right]\right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma_{X_1}(1-\rho^2)^{1/2}} \exp\left[-\frac{1}{2(1-\rho^2)}(z_{X_1} - \rho z_{X_2})^2\right] \\ &= \frac{1}{\sqrt{2\pi}\sigma_{X_1}(1-\rho^2)^{1/2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{X_1 - \mu_{X_1}}{\sigma_{X_1}} - \rho\frac{X_2 - \mu_{X_2}}{\sigma_{X_2}}\right)^2\right] \\ &= \frac{1}{\sqrt{2\pi}\sigma_{X_1}(1-\rho^2)^{1/2}} \exp\left\{-\frac{1}{2(1-\rho^2)\sigma_{X_1}^2}\left[X_1 - \left(\mu_{X_1} + \rho\frac{\sigma_{X_1}}{\sigma_{X_2}}(X_2 - \mu_{X_2})\right)\right]^2\right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma_{X_1|X_2}} \exp\left\{-\frac{1}{2\sigma_{X_1|X_2}^2}(X_1 - \mu_{X_1|X_2})^2\right\} \end{aligned}$$

em que $\mu_{X_1|X_2} = \mu_{X_1} + \rho\frac{\sigma_{X_1}}{\sigma_{X_2}}(X_2 - \mu_{X_2})$ e $\sigma^2_{X_1|X_2} = \sigma_{X_1}(1-\rho^2)$. Tem-se, portanto, que $X_j|X_{j'} \sim \mathrm{N}(\mu_{X_j|X_{j'}}, \sigma^2_{X_j|X_{j'}})$.

Definição: Dadas duas variáveis aleatórias X_1 e X_2, chama-se de **regressão** de X_1 em relação a X_2, à média da distribuição condicional de X_1 dado $X_2 = x_2$, isto é, a

$$\mu_{X_1|X_2} = \mathrm{E}(X_1|X_2) = \mu_{X_1} + \rho\frac{\sigma_{X_1}}{\sigma_{X_2}}(X_2 - \mu_{X_2}).$$

5.1.3 Momentos da distribuição normal bivariada

Define-se momento de ordem r em relação a X_1 e de ordem s em relação a X_2 como:

$$\mu_{r,s} = \mathrm{E}(X_1^r X_2^s) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1^r x_2^s f_{X_1,X_2}(x_1,x_2)dx_1dx_2,$$

sendo r e s inteiros e não negativos. Para a obtenção desses momentos pode-se usar a função geradora de momentos dada por:

$$m_{X_1,X_2}(t_1,t_2) = \mathrm{E}(e^{t_1X_1}e^{t_2X_2}) = \mathrm{E}(e^{t_1X_1+t_2X_2}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t_1x_1+t_2x_2} f_{X_1,X_2}(x_1,x_2)dx_1dx_2,$$

que no caso da distribuição normal bivariada resulta em:

$$m_{X_1,X_2}(t_1,t_2) = e^{t_1\mu_{X_1}+t_2\mu_{X_2}+\frac{1}{2}(t_1^2\sigma^2_{X_1}+t_2^2\sigma^2_{X_2}+2\rho t_1t_2\sigma_{X_1}\sigma_{X_2})}, \quad -h < t_1, t_2 < h, \quad h > 0.$$

Os momentos de ordem r em relação a X_1 e de ordem s em relação a X_2 são obtidos por:

$$\mu_{r,s} = \mathrm{E}(X_1^r X_2^s) = \frac{\partial^{r+s} m(t_1,t_2)}{\partial^r t_1 \partial^s t_2}|_{t_1=0,t_2=0}$$

É fácil verificar que:

$$\mu_{X_1} = \mathrm{E}(X_1) = \mu_{(1,0)} = \frac{\partial m(t_1,t_2)}{\partial t_1}|_{t_1=0,t_2=0}$$

$$\mathrm{E}(X_1^2) = \mu_{(2,0)} = \frac{\partial^2 m(t_1,t_2)}{\partial^2 t_1}|_{t_1=0,t_2=0} = \sigma^2_{X_1} + \mu^2_{X_1}$$

e, portanto, $\mathrm{Var}(X_1) = \mathrm{E}(X_1^2) - [\mathrm{E}(X_1)]^2 = \sigma^2_{X_1}$. Tem-se, ainda, que:

$$\mathrm{E}(X_1X_2) = \mu_{(1,1)} = \frac{\partial^2 m(t_1,t_2)}{\partial t_1 \partial t_2}|_{t_1=0,t_2=0} = \rho\sigma_{X_1}\sigma_{X_2} + \mu_{X_1}\mu_{X_2}$$

e, portanto, $\mathrm{Cov}(X_1,X_2) = \mathrm{E}(X_1X_2) - E(X_1)\mathrm{E}(X_2) = \rho\sigma_{X_1}\sigma_{X_2}$ de onde se tem o coeficiente de correlação linear entre X_1 e X_2, isto é,

$$\rho = \frac{\mathrm{Cov}(X_1,X_2)}{\sigma_{X_1}\sigma_{X_2}}.$$

5.1.4 Correlação linear simples na população

Define-se o coeficiente de correlação linear ρ entre as variáveis X_1 e X_2 como

$$\rho_{X_1,X_2} = \frac{\text{Cov}(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}} = \frac{\text{E}[(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})]}{\sigma_{X_1}\sigma_{X_2}}.$$

A grande vantagem do uso do coeficiente de correlação em lugar da covariância resume-se no fato de que $-1 \leq \rho \leq 1$, simplificando sobremaneira sua interpretação.

De fato,

$$\rho_{X_1,X_2} = \frac{\text{Cov}(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}} = \text{E}\left[\frac{(X_1 - \mu_{X_1})}{\sigma_{X_1}}\frac{(X_2 - \mu_{X_2})}{\sigma_{X_2}}\right],$$

que é a esperança do produto de duas variáveis centradas padronizadas, isto é, de

$$Z_{X_1} = \frac{X_1 - \mu_{X_1}}{\sigma_{X_1}} \quad \text{e} \quad Z_{X_2} = \frac{X_2 - \mu_{X_2}}{\sigma_{X_2}}.$$

sendo $\text{E}(Z_{X_1}) = \text{E}(Z_{X_2}) = 0$, $\text{Var}(Z_{X_1}) = \text{Var}(Z_{X_2}) = 1$. Logo

$$\text{Var}(Z_{X_1} + Z_{X_2}) = \text{Var}(Z_{X_1}) + \text{Var}(Z_{X_2}) + 2\text{Cov}(Z_{X_1}, Z_{X_2}) = 2 + 2\rho_{Z_{X_1}Z_{X_2}} \geq 0 \Rightarrow \rho_{Z_{X_1}Z_{X_2}} \geq -1$$

e

$$\text{Var}(Z_{X_1} - Z_{X_2}) = \text{Var}(Z_{X_1}) + \text{Var}(Z_{X_2}) - 2\text{Cov}(Z_{X_1}, Z_{X_2}) = 2 - 2\rho_{Z_{X_1}Z_{X_2}} \geq 0 \Rightarrow \rho_{Z_{X_1}Z_{X_2}} \leq 1$$

e, portanto,

$$-1 \leq \rho_{Z_{X_1}Z_{X_2}} \leq 1.$$

Mas $\rho_{Z_{X_1}Z_{X_2}} = \rho_{X_1X_2}$, pois

$$\begin{aligned}\rho_{X_1X_2} &= \text{E}\left[\frac{(X_1 - \mu_{X_1})}{\sigma_{X_1}}\frac{(X_2 - \mu_{X_2})}{\sigma_{X_2}}\right] = \text{E}(Z_{X_1}Z_{X_2}) = \text{E}\left[\frac{(Z_{X_1} - 0)}{1}\frac{(Z_{X_2} - 0)}{1}\right] \\ &= \text{E}\left[\frac{(Z_{X_1} - \mu_{Z_{X_1}})}{\sigma_{Z_{X_1}}}\frac{(Z_{X_2} - \mu_{Z_{X_2}})}{\sigma_{Z_{X_2}}}\right] = \rho_{Z_{X_1}Z_{X_2}},\end{aligned}$$

e, portanto,

$$-1 \leq \rho_{X_1X_2} \leq 1.$$

5.1.5 Estimação dos parâmetros da distribuição normal bivariada

Dada uma amostra aleatória, $(X_{11}, X_{12}), (X_{21}, X_{22}), \ldots, (X_{n1}, X_{n2})$, a estimação dos parâmetros μ_{X_1}, μ_{X_2}, σ_{X_1}, σ_{X_2} e ρ pode ser feita utilizando-se o método da máxima verossimilhança. O logaritmo da função de verossilhança, considerando a distribuição normal bivariada fica:

$$\begin{aligned} l &= -nlog(2\pi) - \frac{n}{2}[log(\sigma^2_{X_1}) + log(\sigma^2_{X_2}) + log(1-\rho^2)] \\ &- \frac{1}{2(1-\rho^2)} \sum_{i=1}^{n} \left[\left(\frac{X_{i1} - \mu_{X_1}}{\sigma_{X_1}} \right)^2 - 2\rho \frac{X_{i1} - \mu_{X_1}}{\sigma_{X_1}} \frac{X_{i2} - \mu_{X_2}}{\sigma_{X_2}} + \left(\frac{X_{i2} - \mu_{X_2}}{\sigma_{X_2}} \right)^2 \right] \end{aligned}$$

Derivando-se l em relação a cada um dos parâmetros, igualando-se a zero e resolvendo-se o sistema de equações, obtêm-se:

$$\hat{\mu}_{X_j} = \bar{X}_j = \frac{\sum_{i=1}^{n} X_{ij}}{n}, \quad \hat{\sigma}^2_j = \frac{1}{n} \left[\sum_{i=1}^{n} X^2_{ij} - \frac{\left(\sum_{i=1}^{n} X_{ij}\right)^2}{n} \right]$$

e

$$\hat{\rho} = \frac{\sum_{i=1}^{n} X_{i1}X_{i2} - \frac{\sum_{i=1}^{n} X_{i1} \sum_{i=1}^{n} X_{i2}}{n}}{\sqrt{\left[\sum_{i=1}^{n} X^2_{i1} - \frac{\left(\sum_{i=1}^{n} X_{i1}\right)^2}{n} \right] \left[\sum_{i=1}^{n} X^2_{i2} - \frac{\left(\sum_{i=1}^{n} X_{i2}\right)^2}{n} \right]}}.$$

5.1.6 Correlação linear simples na amostra

Um estimador r para ρ é dado por:

$$r = \frac{\widehat{Cov}(X_1, X_2)}{\hat{\sigma}_{X_1}\hat{\sigma}_{X_2}} = \frac{\sum_{i=1}^{n}(X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{\sqrt{\sum_{i=1}^{n}(X_{i1} - \bar{X}_1)^2 \sum_{i=1}^{n}(X_{i2} - \bar{X}_2)^2}} = \frac{\sum_{i=1}^{n} x_{i1}x_{i2}}{\sqrt{\sum_{i=1}^{n} x^2_{i1} \sum_{i=1}^{n} x^2_{i2}}}.$$

Naturalmente $-1 \leq r \leq 1$. Pode ser notado que

$$r = \frac{\sum_{i=1}^{n}(X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{\sum_{i=1}^{n}(X_{i1} - \bar{X})^2} \frac{\sqrt{\sum_{i=1}^{n}(X_{i1} - \bar{X}_1)^2}}{\sqrt{\sum_{i=1}^{n}(X_{i2} - \bar{X}_2)^2}} = \hat{\beta}_{X_2|X_1} \frac{\hat{\sigma}_{X_1}}{\hat{\sigma}_{X_2}}.$$

ou

$$r = \frac{\sum_{i=1}^{n}(X_{i1} - \bar{X})(X_{i2} - \bar{X}_2)}{\sum_{i=1}^{n}(X_{i2} - \bar{X}_2)^2} \frac{\sqrt{\sum_{i=1}^{n}(X_{i2} - \bar{X}_2)^2}}{\sqrt{\sum_{i=1}^{n}(X_{i1} - \bar{X}_1)^2}} = \hat{\beta}_{X_1|X_2} \frac{\hat{\sigma}_{X_2}}{\hat{\sigma}_{X_1}}.$$

e, portanto,

$$r^2 = \hat{\beta}_{X_2|X_1}\hat{\beta}_{X_1|X_2}$$

o que mostra a relação entre a correlação linear e os coeficientes de regressão. Vê-se, também, que o sinal de r é o mesmo do coeficiente de regressão $\hat{\beta}$. Nota-se, além disso, a importância das variações individuais de X_1 e de X_2.

5.1.7 Testes de hipóteses

Caso 1: Pode-se mostrar que a distribuição de r é normal apenas sob a suposição de que $\rho = 0$, e, portanto, o teste da hipótese:

$$H_0 : \rho = 0 \quad versus \quad \begin{cases} H_{a_1} : \rho < 0 \\ H_{a_2} : \rho > 0 \\ H_{a_3} : \rho \neq 0 \end{cases}$$

é obtido a partir de:

$$\frac{\hat{\rho} - \rho_0}{\sqrt{\hat{V}(\hat{\rho})}} \sim t_{n-2}.$$

Assim, obtém-se:

$$\boxed{t_{calc} = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}}$$

e, a um nível de $100\gamma\%$ de probabilidade, rejeita-se H_0, em favor de:

$H_{a_1} : \rho < 0$ se $t_{calc} < -t_{n-2;\gamma}$ ou se $P(t_{n-2} < t_{calc}) < \gamma$;

$H_{a_2} : \rho > 0$ se $t_{calc} > t_{n-2;\gamma}$ ou se $P(t_{n-2} > t_{calc}) < \gamma$ e

$H_{a_3} : \rho \neq 0$ se $|t_{calc}| > t_{n-2;\frac{\gamma}{2}}$ ou se $P(|t_{n-2}| > |t_{calc}|) < \gamma$,

isto é, as regiões de rejeição de H_0 são dadas pelos intervalos de t correspondentes às áreas hachuradas nas Figuras 2.6, 2.7 e 2.8, respectivamente.

Observação: No caso particular de $H_a : \rho \neq 0$, tem-se que o teste t é equivalente ao teste F, pois $t^2_{calc} = F_{calc}$, isto é,

$$F_{calc} = \frac{SQReg}{QMRes} = \frac{r^2(n-2)}{1-r^2}$$

pois,

$$r^2 = \frac{[\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})]^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \frac{1}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{SQRL_{Y|X}}{SQTotal_Y}$$

ou,

$$r^2 = \frac{[\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})]^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{SQRL_{X|Y}}{SQTotal_X}$$

Caso 2: Um segundo tipo de hipóteses de interesse é dado por:

$$H_0 : \rho = \rho_0 \quad versus \quad \begin{cases} H_{a_1} : \rho < \rho_0 \\ H_{a_2} : \rho > \rho_0 \\ H_{a_3} : \rho \neq \rho_0 \end{cases}$$

A distribuição de r é simétrica apenas sob a suposição de que $\rho = 0$, e portanto, a distribuição normal para r não é adequada para testar valores de $\rho \neq 0$. Fisher mostrou que, assintoticamente,

$$U = \frac{1}{2}\log\frac{1+r}{1-r} = arctanh(r) \sim N(\delta, \sigma_U^2)$$

sendo $\delta = \frac{1}{2}\log\frac{1+\rho}{1-\rho}$ e $\sigma_U^2 = \frac{1}{n-3}$. Logo,

$$Z = (U - \delta_0)\sqrt{(n-3)} \sim N(0,1)$$

e, portanto, para testar $H_0 : \rho = \rho_0$, calcule:

i) $U = \frac{1}{2}\log\frac{1+r}{1-r}$

ii) $\delta_0 = \frac{1}{2}\log\frac{1+\rho_0}{1-\rho_0}$

iii) $z_c = (U - \delta_0)\sqrt{(n-3)}$

e, a um nível de $100\gamma\%$ de probabilidade, rejeita-se H_0, em favor de:

$H_{a_1} : \rho < \rho_0$ se $z_c < -z_\gamma$ ou se $P(Z < z_c) < \gamma$;

$H_{a_2} : \rho > 0$ se $z_c > z_\gamma$ ou se $P(Z > t_c) < \gamma$ e

$H_{a_3} : \rho \neq 0$ se $|z_c| < z_{\frac{\gamma}{2}}$ ou se $P(|Z| > |z_c|) < \gamma$.

Caso 3: No caso de várias variáveis o interesse pode estar no teste de:

$$H_0 : \rho_1 = \rho_2 = \ldots = \rho_m = \rho_0 \quad versus \quad H_a : \text{não } H_0$$

Fisher mostrou que, assintoticamente,

$$U_i = \frac{1}{2}\log\frac{1+r_i}{1-r_i} \sim N(\delta_i, \sigma_{U_i}^2)$$

sendo $\delta_i = \frac{1}{2}\log\frac{1+\rho_i}{1-\rho_i}$ e $\sigma_{U_i}^2 = \frac{1}{n_i - 3}$. Sob H_0, $\delta_i = \delta_0 = \frac{1}{2}\log\frac{1+\rho_0}{1-\rho_0}$. Logo,

$$X_c^2 = \sum_{i=1}^{m} Z_i^2 = \sum_{i=1}^{m} (U_i - \delta_i)^2 (n_i - 3) \sim \chi_m^2$$

e, portanto, para testar $H_0 : \rho_1 = \rho_2 = \ldots = \rho_m = \rho_0$, calcule:

i) $U_i = \frac{1}{2} \log \frac{1 + r_i}{1 - r_i}$

ii) $\delta_0 = \frac{1}{2} \log \frac{1 + \rho_0}{1 - \rho_0}$

iii) $X_c^2 = \sum_{i=1}^{m} Z_i^2 = \sum_{i=1}^{m} (U_i - \delta_0)^2 (n_i - 3)$

Rejeita-se H_0 se $X_c^2 \geq \chi_{\gamma,m}^2$.

Caso 4: Ainda, no caso de várias variáveis o interesse pode estar no teste de:

$$H_0 : \rho_1 = \rho_2 = \ldots = \rho_m = \rho \ \ (\rho \ \text{ não especificado}) \ \ \textit{versus} \ \ H_a : \text{não } H_0$$

i) $U_i = \frac{1}{2} \log \frac{1 + r_i}{1 - r_i}$

ii) $\bar{U} = \frac{\sum_{i=1}^{m} (n_i - 3) U_i}{\sum_{i=1}^{m} (n_i - 3)}$

iii) $Z_i = (U_i - \bar{U}) \sqrt{n_i - 3}$

iv) $X_c^2 = \sum_{i=1}^{m} Z_i^2 = \sum_{i=1}^{m} (U_i - \bar{U})^2 (n_i - 3)$

Rejeita-se H_0 se $X_c^2 \geq \chi_{\gamma,m-1}^2$.

Nota: No caso de não rejeição da hipótese H_0, o estimador r do coeficiente de correlação comum será:

$$r = \frac{e^{\bar{U}} - e^{-\bar{U}}}{e^{\bar{U}} + e^{-\bar{U}}} = \frac{e^{2\bar{U}} - 1}{e^{2\bar{U}} + 1} = tanh(\bar{U}).$$

5.1.8 Intervalo de confiança para ρ

O método utilizado aqui para a construção de um intervalo de confiança será o método da quantidade pivotal. Se $Q = q(Y_1, Y_2, \ldots, Y_n; \theta)$, isto é, uma função da amostra aleatória $Y_1, Y_2, \ldots, Y_n$ e de θ, o parâmetro de interesse, e tem uma distribuição que independe de θ, então Q é uma quantidade pivotal. Logo, para qualquer γ fixo, tal que $0 < \gamma < 1$, existem q_1 e q_2, dependendo de γ, tais que

$$P[q_1 < Q < q_2] = 1 - \gamma.$$

Dado que

$$Z = (U - \delta)\sqrt{(n-3)} \sim N(0,1)$$

sendo $U = \frac{1}{2}\log\frac{1+r}{1-r}$ e $\delta = \frac{1}{2}\log\frac{1+\rho}{1-\rho}$, então, um intervalo de confiança para δ, com um coeficiente de confiança $1-\gamma$, é dado por,

$$IC[\rho]_{1-\gamma} : \frac{1}{2}\log\frac{1+r}{1-r} \pm z_{\frac{\gamma}{2}}\sqrt{\frac{1}{n-3}}.$$

Outra forma de se obterem intervalos de confiança para ρ é através dos ábacos de David, que podem ser encontrados no Apêndice A.11 de Steel & Torrie (1980), páginas 594 e 595.

5.2 Correlações parciais

5.2.1 Introdução

Coeficientes de correlação parcial e múltipla são estritamente aplicáveis somente quando o vetor $(X_{i1}, X_{i2}, \ldots, X_{ik})$ é aleatório. Entretanto, a despeito da falta de aleatorização do vetor é sempre útil calcular esses coeficientes.

A definição de correlação parcial entre duas variáveis segue o mesmo princípio do coeficiente de regressão parcial na regressão múltipla, ou seja, define-se a correlação parcial entre duas variáveis aleatórias como a correlação simples entre essas duas variáveis ajustadas para o efeito linear das demais variáveis. Assim, por exemplo, o símbolo $r_{12.34}$ é usado para indicar a correlação simples entre X_1 e X_2 ajustados para X_3 e X_4. Desde que o vetor é aleatório e o coeficiente é, principalmente, descritivo, não há necessidade de uma determinada variável ser dependente.

Usando a definição do coeficiente de correlação linear simples, tem-se, então, que

$$r_{12.34} = \frac{\hat{\sigma}_{12.34}}{\sqrt{\hat{\sigma}_{11.34}\hat{\sigma}_{22.34}}} = \frac{\widehat{\mathrm{Cov}}(X_1|X_3,X_4;X_2|X_3,X_4)}{\sqrt{\widehat{\mathrm{Var}}(X_1|X_3,X_4)\widehat{\mathrm{Var}}(X_2|X_3,X_4)}},$$

sendo que,

$$X_{1.34} = X_1 - \hat{X}_1, \quad \text{com} \quad \hat{X}_1 = \hat{\beta}_{01} + \hat{\beta}_{11}X_3 + \hat{\beta}_{21}X_4,$$

$$X_{2.34} = X_2 - \hat{X}_2, \quad \text{com} \quad \hat{X}_2 = \hat{\beta}_{02} + \hat{\beta}_{12}X_3 + \hat{\beta}_{22}X_4,$$

$$\hat{\sigma}_{11.34} = \frac{\sum(X_1 - \hat{X}_1)^2}{n-3}, \quad \hat{\sigma}_{22.34} = \frac{\sum(X_2 - \hat{X}_2)^2}{n-3} \quad \text{e} \quad \hat{\sigma}_{12.34} = \frac{\sum(X_1 - \hat{X}_1)(X_2 - \hat{X}_2)}{n-3}.$$

5.2.2 Definição

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)^T$ um vetor aleatório com distribuição normal k-variada com vetor de médias $\boldsymbol{\mu}$ e matriz de variâncias e covariâncias $\boldsymbol{\Sigma}$ (simétrica, positiva definida), isto é,

$$\boldsymbol{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \cdots \\ X_k \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdots \\ \mu_k \end{bmatrix} \quad \text{e} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_{kk} \end{bmatrix}$$

A função de densidade conjunta da distribuição normal k-variada é dada por:

$$f(X_1, X_2, \ldots, X_k) = \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})\right].$$

Considere as partições:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad \text{e} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix},$$

sendo

$$\boldsymbol{X}_1 = \begin{bmatrix} X_1 \\ X_2 \\ \cdots \\ X_m \end{bmatrix}, \quad \boldsymbol{X}_2 = \begin{bmatrix} X_{m+1} \\ X_{m+2} \\ \cdots \\ X_k \end{bmatrix}, \quad \boldsymbol{\mu}_1 = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdots \\ \mu_m \end{bmatrix}, \quad \boldsymbol{\mu}_2 = \begin{bmatrix} \mu_{m+1} \\ \mu_{m+2} \\ \cdots \\ \mu_k \end{bmatrix},$$

$$\boldsymbol{\Sigma}_{11} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_{mm} \end{bmatrix}, \quad \boldsymbol{\Sigma}_{22} = \begin{bmatrix} \sigma_{m+1,m+1} & \sigma_{m+1,m+2} & \cdots & \sigma_{m+1,k} \\ \sigma_{m+2,m+1} & \sigma_{m+2,m+2} & \cdots & \sigma_{m+2,k} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{k,m+1} & \sigma_{k,m+2} & \cdots & \sigma_{kk} \end{bmatrix}$$

e

$$\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^T \begin{bmatrix} \sigma_{1,m+1} & \sigma_{1,m+2} & \cdots & \sigma_{1k} \\ \sigma_{2,m+1} & \sigma_{2,m+2} & \cdots & \sigma_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{m,m+1} & \sigma_{m,m+2} & \cdots & \sigma_{mk} \end{bmatrix}.$$

Demonstra-se que

$$\boldsymbol{X}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \;\text{ e }\; \boldsymbol{X}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}) \;\text{ (distribuições marginais)}$$

e que

$$\boldsymbol{X}_1|\boldsymbol{X}_2 \sim N(\boldsymbol{\mu}_{1.2}, \boldsymbol{\Sigma}_{11.2}) \;\text{ e }\; \boldsymbol{X}_2|\boldsymbol{X}_1 \sim N(\boldsymbol{\mu}_{2.1}, \boldsymbol{\Sigma}_{22.1}) \;\text{ (distribuições condicionais),}$$

sendo

$$\boldsymbol{\mu}_{1.2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{X}_2 - \boldsymbol{\mu}_2), \quad \boldsymbol{\Sigma}_{11.2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

e

$$\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{X}_1 - \boldsymbol{\mu}_1) \;\text{ e }\; \boldsymbol{\Sigma}_{22.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}.$$

Exemplo 1: Normal bidimensional

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix},$$

$$\mu_{1.2} = \mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(X_2 - \mu_2), \quad \sigma_{11.2} = \sigma^2_{1.2} = \sigma_{11} - \frac{\sigma^2_{12}}{\sigma_{22}} = \sigma^2_1 - \frac{\sigma^2_{12}}{\sigma^2_2}$$

e

$$\mu_{2.1} = \mu_2 + \frac{\sigma_{21}}{\sigma_{11}}(X_1 - \mu_1), \quad \sigma_{22.1} = \sigma^2_{2.1} = \sigma_{22} - \frac{\sigma^2_{12}}{\sigma_{11}} = \sigma^2_2 - \frac{\sigma^2_{12}}{\sigma^2_1}.$$

Exemplo 2: Normal tridimensional, sendo fixada a variável X_3

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix},$$

$$\boldsymbol{\mu}_{1.2} = \begin{bmatrix} \mu_{1.3} \\ \mu_{2.3} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \sigma_{13} \\ \sigma_{23} \end{bmatrix} \frac{1}{\sigma^2_3} \begin{bmatrix} X_3 - \mu_3 \end{bmatrix} = \begin{bmatrix} \mu_1 + \dfrac{\sigma_{13}}{\sigma^2_3}(X_3 - \mu_3) \\ \mu_2 + \dfrac{\sigma_{23}}{\sigma^2_3}(X_3 - \mu_3) \end{bmatrix}$$

e

$$\boldsymbol{\Sigma}_{1.2} = \begin{bmatrix} \sigma_{11.3} & \sigma_{12.3} \\ \sigma_{12.3} & \sigma_{22.3} \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} - \begin{bmatrix} \sigma_{13} \\ \sigma_{23} \end{bmatrix} \frac{1}{\sigma^2_3} \begin{bmatrix} \sigma_{13} & \sigma_{23} \end{bmatrix} = \begin{bmatrix} \sigma^2_1 - \dfrac{\sigma^2_{13}}{\sigma^2_3} & \sigma_{12} - \dfrac{\sigma_{13}\sigma_{23}}{\sigma^2_3} \\ \sigma_{12} - \dfrac{\sigma_{13}\sigma_{23}}{\sigma^2_3} & \sigma^2_2 - \dfrac{\sigma^2_{23}}{\sigma^2_3} \end{bmatrix}.$$

Exemplo 3: Normal com 5 variáveis, sendo fixadas as variáveis X_3, X_4 e X_5. Portanto,

$$\boldsymbol{\Sigma}_{1.2} = \begin{bmatrix} \sigma_{11.345} & \sigma_{12.345} \\ \sigma_{12.345} & \sigma_{22.345} \end{bmatrix}.$$

Por definição, o coeficiente de correlação parcial entre as variáveis X_i e X_j, $1 \leq i \leq j \leq m$, mantidas constantes as variáveis $X_{m+1}, X_{m+2}, \ldots, X_k$ ($k = 5$, $m = 2$ e $k - m = 3$) é dado por:

$$\rho_{ij.m+1,m+2,\ldots,k} = \frac{\sigma_{ij.m+1,m+2,\ldots,k}}{\sqrt{\sigma_{ii.m+1,m+2,\ldots,k}\ \sigma_{jj.m+1,m+2,\ldots,k}}}.$$

O número de graus de liberdade associado a $\sigma_{ij.m+1,m+2,\ldots,k}$ é dado por $n - (k - m)$. No Exemplo 3, a correlação parcial entre X_1 e X_2, mantidas contantes as variáveis X_3, X_4 e X_5, é dada por:

$$\rho_{12.345} = \frac{\sigma_{12.345}}{\sqrt{\sigma_{11.345}\ \sigma_{22.345}}}.$$

No Exemplo 2, tem-se:

$$\rho_{12.3} = \frac{\sigma_{12.3}}{\sqrt{\sigma_{11.3}\ \sigma_{22.3}}} = \frac{\sigma_{12} - \frac{\sigma_{13}\sigma_{23}}{\sigma_{33}}}{\sqrt{(\sigma_{11} - \frac{\sigma_{13}^2}{\sigma_{33}})(\sigma_{22} - \frac{\sigma_{23}^2}{\sigma_{33}})}} = \frac{\frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}} - \frac{\sigma_{13}}{\sqrt{\sigma_{11}\sigma_{33}}}\frac{\sigma_{23}}{\sqrt{\sigma_{22}\sigma_{33}}}}{\sqrt{(1 - \frac{\sigma_{13}^2}{\sigma_{11}\sigma_{33}})(1 - \frac{\sigma_{23}^2}{\sigma_{22}\sigma_{33}})}} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{(1-\rho_{13}^2)(1-\rho_{23}^2)}},$$

o que mostra a relação entre o coeficiente de correlação parcial e os coeficientes de correlação simples.

De uma forma genérica, tem-se que dada a matriz de correlações simples e sua inversa, isto é,

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \dots & \rho_{1k} \\ \rho_{21} & 1 & \rho_{23} & \dots & \rho_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ \rho_{k1} & \rho_{k2} & \rho_{k3} & \dots & 1 \end{bmatrix} \quad \text{e} \quad \boldsymbol{R}^{-1} = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1k} \\ c_{21} & c_{22} & \dots & c_{2k} \\ \dots & \dots & \dots & \dots \\ c_{k1} & c_{k2} & \dots & c_{kk} \end{bmatrix}$$

define-se

$$\rho_{ij.1,\dots,i-1,i+1,\dots,j-1,j+1,\dots,k} = \frac{-c_{ij}}{\sqrt{c_{ii}c_{jj}}}.$$

5.2.3 Estimativa do coeficiente de correlação parcial

Considerando-se k variáveis X_1, X_2, ..., X_k com n observações, tem-se que a estimativa de $\boldsymbol{\Sigma}$ pelo método da máxima verossimilhança é dada por:

$$\hat{\Sigma} = \begin{bmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} & \dots & \hat{\sigma}_{1k} \\ \hat{\sigma}_{21} & \hat{\sigma}_{22} & \dots & \hat{\sigma}_{2k} \\ \dots & \dots & \dots & \dots \\ \hat{\sigma}_{k1} & \hat{\sigma}_{k2} & \dots & \hat{\sigma}_{kk} \end{bmatrix} = \frac{1}{n}\begin{bmatrix} \sum x_1^2 & \sum x_1x_2 & \dots & \sum x_1x_k \\ \sum x_1x_2 & \sum x_2^2 & \dots & \sum x_2x_k \\ \dots & \dots & \dots & \dots \\ \sum x_1x_k & \sum x_2x_k & \dots & \sum x_k^2 \end{bmatrix}$$

que é uma estimativa viesada. Isso, porém, não importa na determinação do coeficiente de correlação, pois os denominadores se cancelam.

No Exemplo 2, em que

$$\rho_{12.3} = \frac{\sigma_{12.3}}{\sqrt{\sigma_{11.3}\sigma_{22.3}}} = \frac{\sigma_{12} - \frac{\sigma_{13}\sigma_{23}}{\sigma_{33}}}{\sqrt{(\sigma_{11} - \frac{\sigma_{13}^2}{\sigma_{33}})(\sigma_{22} - \frac{\sigma_{23}^2}{\sigma_{33}})}} = \frac{\sigma_{12}\sigma_{33} - \sigma_{13}\sigma_{23}}{\sqrt{(\sigma_{11}\sigma_{33} - \sigma_{13}^2)(\sigma_{22}\sigma_{33} - \sigma_{23}^2)}}$$

sua estimativa é dada por

$$r_{12.3} = \frac{\hat{\sigma}_{12}\hat{\sigma}_{33} - \hat{\sigma}_{13}\hat{\sigma}_{23}}{\sqrt{(\hat{\sigma}_{11}\hat{\sigma}_{33} - \hat{\sigma}_{13}^2)(\hat{\sigma}_{22}\hat{\sigma}_{33} - \hat{\sigma}_{23}^2)}} = \frac{\sum x_1x_2 \sum x_3^2 - \sum x_1x_3 \sum x_2x_3}{\sqrt{\left[\sum x_1^2 \sum x_3^2 - \left(\sum x_1x_3\right)^2\right]\left[\sum x_2^2 \sum x_3^2 - \left(\sum x_2x_3\right)^2\right]}}.$$

5.2.4 Testes de hipóteses

Em função do que já foi visto, tem-se que a estatística para o teste da hipótese:

$$H_0 : \rho_{ij.m+1,m+2,\ldots,k} = 0 \quad versus \quad \begin{cases} H_{a_1} : \rho_{ij.m+1,m+2,\ldots,k} < 0 \\ H_{a_2} : \rho_{ij.m+1,m+2,\ldots,k} > 0 \\ H_{a_3} : \rho_{ij.m+1,m+2,\ldots,k} \neq 0 \end{cases}$$

é dada por:

$$t_{calc} = \frac{\hat{\rho}_{ij.m+1,m+2,\ldots,k}}{\sqrt{1 - \hat{\rho}^2_{ij.m+1,m+2,\ldots,k}}} \sqrt{n - (k-m) - 2}$$

e, a um nível de $100\gamma\%$ de probabilidade, rejeita-se H_0, em favor de:

$H_{a_1} : \rho < 0$ se $t_{calc} < -t_{n-(k-m)-2;\gamma}$ ou se $P(t_{n-(k-m)-2} < t_{calc}) < \gamma$;

$H_{a_2} : \rho > 0$ se $t_{calc} > t_{n-(k-m)-2;\gamma}$ ou se $P(t_{n-(k-m)-2} > t_{calc}) < \gamma$ e

$H_{a_3} : \rho \neq 0$ se $|t_{calc}| > t_{n-(k-m)-2;\frac{\gamma}{2}}$ ou se $P(|t_{n-(k-m)-2}| > |t_{calc}|) < \gamma$.

5.3 Exemplo

Os dados da Tabela 5.1 referem-se a um estudo sobre a resposta da cultura do milho como função da quantidade de fosfato, porcentagem de saturação de bases (X_2) e sílica (X_3) em solos ácidos. A resposta (Y), em porcentagem, foi medida como a diferença entre as produções (em lb/acre) nas parcelas recebendo fosfato e aquelas não recebendo fosfato (X_1), dividida pelas produções das parcelas recebendo fosfato, e multiplicadas por 100. Considerando-se esses dados, foi obtida a variável produtividade Y_1 das parcelas recebendo fosfato, dada por $Y_1 = X_1(1 + \frac{Y}{100})$.

Nesse caso, as variáveis X_1, X_2 e X_3 são aleatórias, e o interesse do pesquisador está, principalmente no estudo de correlações entre as variáveis. Na Figura 5.1 podem ser vistos os gráficos de dispersão para as variáveis duas a duas. Observa-se que existe uma correlação linear grande e positiva entre as variáveis X_1 e X_2.

Os gráficos da diagonal principal da Figura 5.2 mostram as densidades não paramétricas que revelam distribuições simétricas não muito diferentes da distribuição normal. Os gráficos fora da diagonal, por sua vez, apresentam elipses de contorno de distribuições normais bivariadas com estimativas dos parâmetros dadas pela matriz de variâncias e covariâncias dos dados. Fornecem, desta forma, uma visualização gráfica da correlação linear entre as variáveis observadas. As variáveis Y_1 e X_1, por exemplo, mostram-se positiva e altamente correlacionadas. Para se obterem as matrizes de

Tabela 5.1: Dados de resposta da cultura do milho (Y) ao fosfato, em porcentagem, produtividade na testemunha (X_1), em lb/acre, porcentagem de saturação de bases (X_2) e sílica (pH do solo, X_3)

Y	X_1	X_2	X_3	Y	X_1	X_2	X_3
88	844	67	5,75	18	1262	74	6,10
80	1678	57	6,05	18	4624	69	6,05
42	1573	39	5,45	4	5249	76	6,15
37	3025	54	5,70	2	4258	80	5,55
37	653	46	5,55	2	2943	79	6,40
20	1991	62	5,00	-2	5092	82	6,55
20	2187	69	6,40	-7	4496	85	6,50

Fonte: Steel & Torrie (1980).

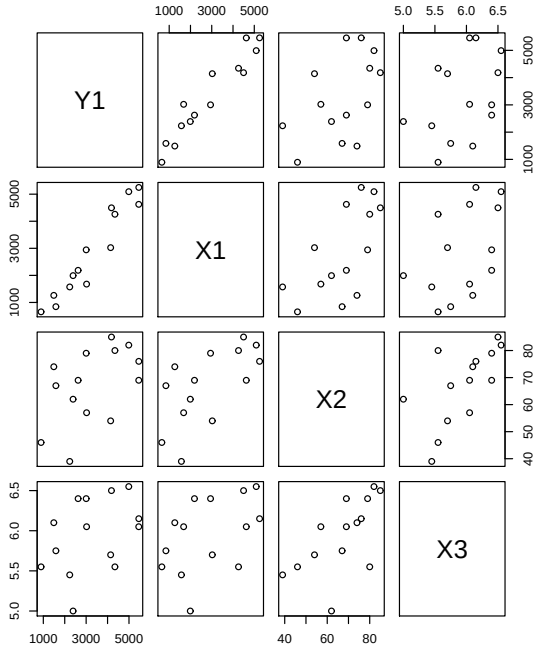


Figura 5.1: Gráficos de dispersão para as variáveis duas a duas.

variâncias e covariâncias e de correlações de Pearson entre as variáveis Y_1, X_1, X_2 e X_3, utilizam-se os comandos `cov` e `cor`, do seguinte modo:

```
> cov(Dados6a)   # Matriz de variâncias e covariâncias #
             Y1           X1           X2          X3
Y1 2255764.0780 2334575.5465 10591.389341 267.7395659
X1 2334575.5465 2631861.7198 14693.060440 340.2478022
```

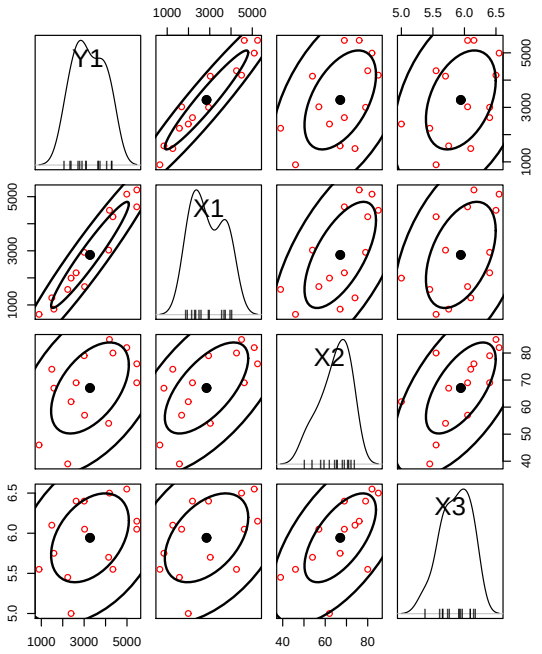


Figura 5.2: Gráficos de dispersão das variáveis Y_1, X_1, X_2 e X_3, duas a duas, com respectivas elipses de contorno de distribuições normais bivariadas com parâmetros estimados a partir dos dados e densidades não paramétricas.

```
X2   10591.3893   14693.0604   193.763736   4.0351648
X3     267.7396     340.2478     4.035165   0.2095604
> cor(Dados6a)   # Matriz de correlações              #
          Y1        X1        X2        X3
Y1 1.0000000 0.9581413 0.5066054 0.3894136
X1 0.9581413 1.0000000 0.6506456 0.4581515
X2 0.5066054 0.6506456 1.0000000 0.6332430
X3 0.3894136 0.4581515 0.6332430 1.0000000
```

que confirmam a correlação linear grande entre as variáveis Y_1 e X_1 ($r_{X_1,Y_1} = 0,9581413$). Para se testar a hipótese $H_0 : \rho_{X_1,Y_1} = 0$ contra a hipótese alternativa $H_0 : \rho_{X_1,Y_1} \neq 0$, utiliza-se o comando `cor.test`, da seguinte forma:

```
> cor.test(X1,Y1)

        Pearson's product-moment correlation

data:  X1 and Y1
t = 11.5933, df = 12, p-value = 7.095e-08
```

Tabela 5.2: Coeficientes de correlação de Pearson entre as variáveis Y_1, X_1, X_2 e X_3, e respectivos valores-p dos testes para a presença de correlação entre parênteses.

	Y_1	X_1	X_2	X_3
Y_1	1,0000000	0,9581413 (7,095e-08)	0,5066054 (0,0645)	0,3894136 (0,1687)
X_1	0,9581413 (7,095e-08)	1,0000000	0,6506456 (0,01174)	0,4581515 (0,09946)
X_2	0,5066054 (0,0645)	0,6506456 (0,01174)	1,0000000	0,6332430 (0,01506)
X_3	0,3894136 (0,1687)	0,4581515 (0,09946)	0,6332430 (0,01506)	1,0000000

```
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8696823 0.9869731
sample estimates:
      cor
0.9581413
```

e dessa forma, como o valor-p (7,095e-08) é menor do que 0,05 rejeita-se a hípótese de nulidade H_0 a um nível de significância 5% de probabilidade, ou seja, há evidências para dizer que a correlação linear entre as variáveis X_1 e Y_1 difere de zero. Pode-se dizer, portanto, que as variáveis produtividade sem fosfato e com fosfato estão positivamente correlacionadas e, ainda, que o intervalo de 95% de confiança para ρ_{X_1,Y_1} é $[0,8696823; 0,9869731]$. Na tabela 5.3 apresentam-se todos os coeficientes de correlação envolvidos e respectivos valores-p dos testes para correlação.

Pode-se, por outro lado, estar interessado em obter a correlação entre as variáveis produção com fosfato (Y_1) e porcentagem de saturação de bases (X_2), considerando-se fixa a produtividade na testemunha (X_1). Em outras palavras, tem-se interesse em obter a correlação parcial entre Y_1 e X_2 ajustada para X_1, denotada por $r_{Y_1,X_2.X_1}$. Embora não haja uma função que calcule diretamente essa correlação, pode-se, facilmente, obtê-la calculando-se a correlação entre os resíduos da regressão linear entre Y_1 e X_1 e os resíduos da regressão entre X_2 e X_1, o que pode ser feito, no R, da seguinte forma:

```
> cor(residuals(lm(Y1~X1)),residuals(lm(X2~X1))) # r Y1,X2|X1
[1] -0.5372623
```

Embora o sinal dessa correlação tenha sido oposto ao da correlação entre Y_1 e X_2 (0,5066054), um estudo mais detalhado mostraria que essas correlações não diferem estatisticamente de 0. Analogamente, podemos estar interessados na correlação linear entre X_2 e X_3, ajustada para as demais variáveis, ou seja, no coeficiente de correlação parcial $r_{X_2,X_3.Y_1,X_1}$, obtido da seguinte forma:

```
> cor(residuals(lm(X2~Y1+X1)),residuals(lm(X3~Y1+X1))) # r X2,X3|Y1,X1
[1] 0.4737149
```

cujo gráfico de dispersão entre as variáveis X_2 e X_3 ajustadas para Y_1 e X_1, é obtido por meio de:

```
> plot(residuals(lm(X2~Y1+X1)),residuals(lm(X3~Y1+X1)))
```

Esse gráfico, apresentado na Figura 5.3, também é chamado de gráfico de regressão parcial ("partial-regression plot") de X_3 em função de X_2, ajustadas para as demais variáveis, Y_1 e X_1, ou gráfico da variável adicionada ("added-variable plot"), que pode ser obtido, alternativamente, por meio de:

```
> av.plot(lm(X3~X2+Y1+X1),X2)
```

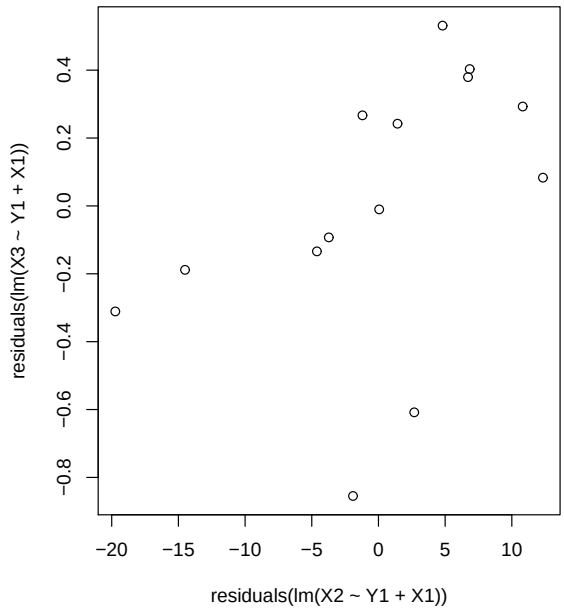


Figura 5.3: Gráfico de dispersão entre as variáveis porcentagem de saturação de bases (X_2) e pH do solo (X_3), ajustadas para as variáveis Y_1 e X_1.

5.4 Exercícios

1. Dada a matriz de correlações simples

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix}$$

   obtenha $\boldsymbol{R}^{-1}$ e $\rho_{12.3}$, $\rho_{13.2}$ e $\rho_{23.1}$.

2. Usando os dados do exercício 7 da 1ª aula prática, pede-se:

(a) Obter os coeficientes de correlação simples r_{XY}, r_{ZY} e r_{ZX}, onde $X_i = X_{1i}$ e $Z_i = X_{2i}$.

(b) Testar a hipótese $H_0 : \rho_{XZ} = 0$ *vs* $H_a : \rho_{XZ} \neq 0$, com um nível de significância 5%.

3. Dado a tabela 10.1, de coeficientes de correlação estimados, verificar as hipóteses:

(a) $H_0 : \rho_1 = 0$ *vs* $H_a : \rho_1 \neq 0$

(b) $H_0 : \rho_4 = 0,3$ *vs* $H_a : \rho_4 \neq 0,3$

(c) $H_0 : \rho_1 = \rho_2 = 0$ *vs* H_a : não H_0

(d) $H_0 : \rho_1 = \rho_2 = \rho_3 = \rho_4 = \rho$ *vs* H_a : não H_0. Estimar ρ

Tabela 10.1. Coeficientes de correlação estimados e respectivos tamanhos de amostra.

Amostra (i)	n_i	r_i	Parâmetro
1	24	0,028	ρ_1
2	28	0,054	ρ_2
3	24	0,407	ρ_3
4	20	0,381	ρ_4

4. Baseando-se nos dados apresentados na tabela 10.2, pede-se:

(a) Calcular os coeficientes de correlação simples entre T e L, entre T e N e entre L e N.

(b) Calcular os intervalos de 95% de confiança para ρ_{TL}, ρ_{TN} e ρ_{LN}.

(c) Testar, com um nível de significância 5%, as seguintes hipóteses:

i. $H_0 : \rho_{TL} = 0,5$ *vs* $H_a : \rho_{TL} \neq 0,5$.

ii. $H_0 : \rho_{TL} = \rho_{TN} = \rho_{LN} = 0,7$ *vs* H_a : não H_0.

iii. $H_0 : \rho_{TL} = \rho_{TN} = \rho_{LN} = \rho$ *vs* H_a : não H_0. Estimar ρ

Tabela 10.2. Comprimento do caule (T), do ramo (L) e do caule basal (N) de *Nicotiana*.

i	T	L	N	i	T	L	N
1	49	27	19	10	45	21	21
2	44	24	16	11	41	22	14
3	32	12	12	12	48	25	22
4	42	22	17	13	45	23	22
5	32	13	10	14	39	18	15
6	53	29	19	15	40	20	14
7	36	14	15	16	34	15	15
8	39	20	14	17	37	20	15
9	37	16	15	18	35	13	16

5. Seja (X_1, X_2, X_3, X_4) uma variável aleatória de dimensão 4 com distribuição normal multivariada. De uma amostra de tamanho $n = 20$, calculou-se a estimativa $\hat{\Sigma}$ da matriz de variâncias e covariâncias Σ, obtendo-se:

$$\hat{\Sigma} = \begin{bmatrix} 10 & 9 & -1 & -16 \\ 9 & 20 & -3 & -16 \\ -1 & -3 & 5 & 3 \\ -16 & -16 & 3 & 27 \end{bmatrix}.$$

Pede-se estimar $\rho_{12.4}$, $\rho_{13.4}$, $\rho_{34.12}$ e $\rho_{23.14}$

6. Mostre que o valor da estatística F para o teste da hipótese $H_0 : \beta_1 = \beta_2 = 0$, na análise de variância, pode ser reduzido a
$$F = \frac{R^2_{Y.12}}{1 - R^2_{Y.12}} \frac{n-k-1}{k}.$$

7. Os dados que se seguem referem-se a um estudo da resposta da cultura do milho (Y) ao fosfato, porcentagem de saturação de bases e sílica em solos ácidos. A resposta, em porcentagem, foi medida como a diferença entre as produções nas parcelas recebendo fosfato e aquelas não recebendo fosfato (X_1), dividida pelas produções das parcelas recebendo fosfato, e multiplicadas por 100. Portanto, uma correlação entre Y e X_1 foi introduzida nos cálculos.

Y	X_1	X_2	X_3	Y	X_1	X_2	X_3
88	844	67	5,75	18	1262	74	6,10
80	1678	57	6,05	18	4624	69	6,05
42	1573	39	5,45	4	5249	76	6,15
37	3025	54	5,70	2	4258	80	5,55
37	653	46	5,55	2	2943	79	6,40
20	1991	62	5,00	-2	5092	82	6,55
20	2187	69	6,40	-7	4496	85	6,50

Y=resposta ao fosfato, em porcentagem.
X_1=produtividade na testemunha, em lb/acre.
X_2=porcentagem de saturação de bases.
X_3=pH do solo.

Fonte: Steel, R.G.D & Torrie, J.H (1980). *Principles and Procedures os Statistics. A Biometrical Approach.* 2ª ed. Ed. McGraw-Hill, p.324.

Pede-se:

(a) Ajustar, a esses dados, o modelo $E(Y) = \beta_0 + \beta_1 X_1 + \beta_3 X_3$

(b) Testar, com um nível de significância 5%, as seguintes hipóteses:

i. $H_0 : \beta_1 = \beta_3 = 0$ *vs* H_a : não H_0

ii. $H_0 : \beta_1 = 0$ *vs* $H_a : \beta_1 \neq 0$

iii. $H_0 : \beta_3 = 0$ *vs* $H_a : \beta_3 \neq 0$

(c) Obter $r_{Y1.3}$, $r_{Y3.1}$, $r_{13.Y}$ e $R^2_{Y.13}$.

(d) Contruir intervalos de 95% de confiança para $\rho_{Y1.3}$ e $\rho_{Y3.1}$. Esses intervalos incluem o zero? Qual a conclusão?

(e) Testar, com um nível de significância 5%, a hipótese $H_0 : \rho_{Y1.3} = 0$ *vs* $H_a : \rho_{Y1.3} \neq 0$.

Um programa no SAS pode ser:

```
options nodate ps=25;
data aula9;
input X1 X2 X3 Y1;
cards;
```

```
-2    2   -2    8.5
-1   -1    0    1
-1    0    0    4
-1    0    0    4
-1    1    0    5
 1   -1    0    3
 1    0    0    6
 1    0    0    6
 1    1    0    7
 0    0   -1    5
 0    0    0    5
 0    0    0    5
 0    0    1    3
 2   -2    2    0.5
;
```

1. Seja (X, Y) uma variável aleatória bidimensional com distribuição apresentada na tabela 1, com parâmetros p_1 e p_2. Pede-se:

   (a) Calcular $E(X)$, $E(Y)$, $E(Y|X = x)$, $V(X)$, $V(Y)$, $Cov(X, Y)$ e ρ_{XY}.

   (b) Construir um gráfico que apresente o espaço paramétrico.

   (c) Obter os estimadores de máxima verossimilhança de p_1 e p_2, para uma amostra aleatória simples de tamanho n, extraída de uma população com essa distribuição.

   (d) Obter um estimador não viesado para ρ_{XY} para uma amostra aleatória simples de tamanho n, extraída de uma população com essa distribuição.

   (e) Como você testaria a hipótese $H_0 : \rho_{XY} = 0$ contra $H_a : \rho_{XY} > 0$, com um nível de significância $\gamma = 5\%$? Essa hipótese é equivalente à hipótese $H_0 : p_1 = \frac{1}{9}$ contra $H_a : p_1 \neq \frac{1}{9}$?

   (f) Exemplificar os procedimentos desenvolvidos nos itens anteriores utilizando, para isso, a seguinte amostra $\{(-1; -1), (0; -1), (0; 0), (0; 0), (0; 1), (1; 1)\}$.

Tabela 1. Distribuição de (X, Y)

	X		
Y	-1	0	1
-1	p_1	p_2	p_2
0	p_2	p_1	p_2
1	p_2	p_2	p_1

2. O arquivo *trees*, disponível no pacote R, contém os dados de 31 cerejeiras (*Black cherry*) da Floresta Nacional de Allegheny, relativos a três variáveis: volume de madeira útil ($Volume$),

em pés cúbicos; altura (*Height*), em pés, e circunferência (*Girth*) a $4,5$ pés ($1,37$ metros) de altura. Baseando-se nesse conjunto de dados, pede-se, usando o pacote R:

(a) Estimar o coeficiente de correlação linear ρ_{AC} entre as variáveis altura e circunferência.

(b) Testar a hipótese $H_0 : \rho_{AC} = 0$ contra $H_a : \rho_{AC} \neq 0$, considerando o nível de significância 5%.

(c) Testar a hipótese $H_0 : \rho_{AC} = 0,9$ contra $H_a : \rho_{AC} \neq 0,9$, considerando o nível de significância 5%.

(d) Construir o intervalo de 95% de confiança para ρ_{AC}.

3. Dada a tabela 2, de coeficientes de correlação estimados, testar as seguintes hipóteses:

(a) $H_0 : \rho_1 = 0$ contra $H_\text{a} : \rho_1 \neq 0$

(b) $H_0 : \rho_4 = 0,3$ contra $H_\text{a} : \rho_4 > 0,3$

(c) $H_0 : \rho_1 = \rho_2 = 0$ contra H_a : não H_0

(d) $H_0 : \rho_1 = \rho_2 = \rho_3 = \rho_4 = \rho$ contra H_a : não H_0. Estimar ρ

Tabela 2. Coeficientes de correlação estimados e respectivos tamanhos de amostra.

Amostra (i)	n_i	r_i	Parâmetro
1	24	0,028	ρ_1
2	28	0,054	ρ_2
3	24	0,407	ρ_3
4	20	0,381	ρ_4

4. Baseando-se nos dados apresentados na tabela 3, pede-se:

(a) Calcular os coeficientes de correlação simples entre T e L, entre T e N e entre L e N.

(b) Calcular os intervalos de 95% de confiança para ρ_{TL}, ρ_{TN} e ρ_{LN}.

(c) Testar, com um nível de significância 5%, as seguintes hipóteses:

i. $H_0 : \rho_{TL} = 0,5$ contra $H_\text{a} : \rho_{TL} \neq 0,5$.

ii. $H_0 : \rho_{TL} = \rho_{TN} = \rho_{LN} = 0,7$ contra H_a : não H_0.

iii. $H_0 : \rho_{TL} = \rho_{TN} = \rho_{LN} = \rho$ contra H_a : não H_0. Estimar ρ

Tabela 3. Comprimento do caule (T), do ramo (L) e do caule basal (N) de *Nicotiana*.

i	T	L	N	i	T	L	N
1	49	27	19	10	45	21	21
2	44	24	16	11	41	22	14
3	32	12	12	12	48	25	22
4	42	22	17	13	45	23	22
5	32	13	10	14	39	18	15
6	53	29	19	15	40	20	14
7	36	14	15	16	34	15	15
8	39	20	14	17	37	20	15
9	37	16	15	18	35	13	16

5. Em experimentos na área de Fitopatologia, freqüentemente estamos interessados em avaliar visualmente o grau de infestação de plantas por doenças. Essa avaliação, no entanto, exige um treinamento específico para a cultura e doença em questão. De modo a facilitar esse treinamento, foram desenvolvidos programas computacionais que geram imagens de folhas com diferentes porcentagens de infestação para o pesquisador estimar visualmente e apresentam, em seguida, a porcentagem de infestação real. Com base nos resultados, o pesquisador pode verificar se está bem treinado ou não. Considerando-se os resultados de um desses testes, apresentado na tabela 1:

   (a) Ajuste o modelo de regressão linear simples e faça a análise de resíduos. Há algum problema com relação às pressuposições necessárias para realizar testes de hipóteses?

   (b) Calcule o coeficiente de correlação linear de Pearson entre as porcentagens reais e estimadas visualmente. Você acha que esse coeficiente de correlação quantifica o quanto o pesquisador está bem treinado? Discuta.

   (c) Faça o teste das hipóteses $H_0 : \rho = 0$ contra $H_a : \rho \neq 0$ e $H_0 : \rho = 0,9$ contra $H_a : \rho \neq 0,9$ considerando o nível de significância 5%. Estes testes avaliam se o pesquisador está bem treinado? Discuta.

   (d) Faça o teste da hipótese $H_0 : \beta_0 = 0,\ \beta_1 = 1$ contra H_a : Não H_0, considerando o nível de significância 5%. Você acha que este teste serve para avaliar se o pesquisador está bem treinado?

Tabela 1. Porcentagens de área foliar com ferrugem reais e estimadas visualmente por um pesquisador, para 10 folhas de amendoim.

	Porcentagem de área foliar com ferrugem	
Nº da folha	Estimada visualmente	Real
1	10	19
2	10	17
3	8	18
4	40	34
5	40	41
6	15	17
7	30	26
8	80	51
9	60	49
10	60	50

Fonte: Dados obtidos através de simulação a partir do programa Disease.Pro Version 5.1

6. Seja (X, Y) uma variável aleatória com distribuição normal bivariada. Se a intenção de se coletarem pares de observações de (X, Y) é a de quantificar a concordância entre eles, Lin (1989) sugere utilizar a medida chamada *correlação de concordância* ρ_C, dada por:

$$\rho_C = \frac{2Cov(X,Y)}{\sigma_X^2 + \sigma_Y^2 + (\mu_X - \mu_Y)^2}$$

e estimada por

$$r_C = \frac{2\hat{Cov}(X,Y)}{s_X^2 + s_Y^2 + (\bar{X} - \bar{Y})^2} = \frac{2\sum xy}{\sum x^2 + \sum y^2 + (n-1)(\bar{X} - \bar{Y})^2}.$$

Este coeficiente pode variar de -1 a 1, e seu valor absoluto não pode ser maior do que o coeficiente de correlação de Pearson, r, ou seja, $-1 \leq -|r| \leq r_C \leq |r| \leq 1$. Considerando-se os dados da tabela 1, pede-se:

(a) Calcular o coeficiente de correlação de concordância de Lin.

(b) Obter o intervalo de 95% de confiança para ρ_C usando o procedimento apresentado por Lin (1989) ou Zar (1999, seção 19.13).

Referências:

Lin, L. I-K. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45: 255-268.

Zar, J. H. 1999. *Biostatistical Analysis.* 4ªed. Editora Prentice-Hall, New Jersey.

7. Seja $\boldsymbol{X} \sim N_k(\boldsymbol{\mu}; \Sigma)$ e $M = [\boldsymbol{X}_1 \boldsymbol{X}_2 \cdots \boldsymbol{X}_n]^T$, uma amostra de tamanho n extraída dessa população. Verifique que a estimativa de Σ pode ser obtida por meio de

$$\hat{\Sigma} = \frac{1}{n-1}\left(M^T M - \frac{1}{n} M^T \mathbf{1} M\right) = \frac{1}{n-1} M^T \left(I - \frac{1}{n}\mathbf{1}\right) M,$$

sendo $\mathbf{1}$ a matriz $n \times n$ de uns e I, a matriz identidade, de tamanho n. Apresente um exemplo.

1. Seja (X_1, X_2, X_3, X_4) uma variável aleatória de dimensão 4 com distribuição normal multivariada. De uma amostra de tamanho $n = 20$, foi obtida a estimativa $\hat{\Sigma}$ da matriz de variâncias e covariâncias Σ, dada por:

$$\hat{\Sigma} = \begin{bmatrix} 10 & 9 & -1 & -16 \\ 9 & 20 & -3 & -16 \\ -1 & -3 & 5 & 3 \\ -16 & -16 & 3 & 27 \end{bmatrix}.$$

Pede-se:

(a) Estimar $\rho_{12.4}$, $\rho_{13.4}$, $\rho_{34.12}$ e $\rho_{23.14}$

(b) Testar, considerando-se o nível de significância $\gamma = 5\%$, as hipóteses:

i. $H_0 : \rho_{12.4} = 0$ contra $H_0 : \rho_{12.4} \neq 0$

ii. $H_0 : \rho_{13.4} = 0$ contra $H_0 : \rho_{13.4} \neq 0$

iii. $H_0 : \rho_{34.12} = 0$ contra $H_0 : \rho_{34.12} \neq 0$

iv. $H_0 : \rho_{23.14} = 0$ contra $H_0 : \rho_{23.14} \neq 0$

2. Considerando-se os dados apresentados na Tabela 1,

(a) Construa os gráficos de dispersão relacionando todos os pares de variáveis observadas.

(b) Obtenha a matriz de variâncias e covariâncias e a matriz de correlações entre as variáveis.

(c) Considere os resultados dos itens anteriores. Há algum problema para aparente para a realização de testes de hipóteses sobre esses coeficientes?

(d) Refaça os dois primeiros itens transformando-se todas variáveis por meio da tranformação logarítmica ($\log_e$)

(e) Construa o intervalo de 95% de confiança para o coeficiente de correlação entre o logaritmo da condutividade elétrica e o logaritmo da proporção de cinzas.

(f) Obtenha 5000 amostras (com reposição) de tamanho 96 dos pares de observações ($\log_e Condut.$; $\log_e Cinzas$) e calcule, para cada uma delas, o coeficiente de correlação linear r de Pearson. Com base nos 5000 valores obtidos:

i. Construa um histograma (distribuição empírica *bootstrap* de r).

ii. Calcule a média (estimativa *bootstrap* de ρ).

iii. Construa o intervalo formado pelos quantis de ordem 2,5 e 97,5 (intervalo *bootstrap* de 95% de confiança para ρ, baseado nos percentis). Obs.: existem metodologias *bootstrap* mais adequadas para a construção de intervalos de confiança.

iv. Compare os resultados obtidos por meio da metodologia *bootstrap*, com os do item 2e.

(g) Seja $X_1 = log_e Condut$, $X_3 = log_e N$ e $X_2 = log_e Cinzas$. Pede-se:

i. Estimar o coeficiente de correlação linear entre X_1 e X_2, ajustado para X_3 ($\rho_{12.3}$), estimar $\rho_{13.2}$ e $\rho_{23.1}$.

ii. Testar, considerando-se o nível de significância $\gamma = 5\%$, $H_0 : \rho_{12.3} = 0$ contra $H_0 : \rho_{12.3} \neq 0$

iii. Calcular o coeficiente de correlação linear r^o entre os resíduos dos modelos de regressão $X_1 = \beta_0 + \beta_1 X_3 + \varepsilon$ e $X_2 = \beta_0 + \beta_1 X_3 + \varepsilon$ e verificar que $r_{12.3} = r^o$.

3. Os dados que se seguem referem-se a um estudo da resposta da cultura do milho (Y) ao fosfato, porcentagem de saturação de bases e sílica em solos ácidos. A resposta, em porcentagem, foi medida como a diferença entre as produções nas parcelas recebendo fosfato e aquelas não recebendo fosfato (X_1), dividida pelas produções das parcelas recebendo fosfato, e multiplicadas por 100. Portanto, uma correlação entre Y e X_1 foi introduzida nos cálculos.

Y	X_1	X_2	X_3	Y	X_1	X_2	X_3
88	844	67	5,75	18	1262	74	6,10
80	1678	57	6,05	18	4624	69	6,05
42	1573	39	5,45	4	5249	76	6,15
37	3025	54	5,70	2	4258	80	5,55
37	653	46	5,55	2	2943	79	6,40
20	1991	62	5,00	-2	5092	82	6,55
20	2187	69	6,40	-7	4496	85	6,50

Y=resposta ao fosfato, em porcentagem.
X_1=produtividade na testemunha, em lb/acre.
X_2=porcentagem de saturação de bases.
X_3=pH do solo.

Fonte: Steel & Torrie (1980), p.324.

Pede-se:

(a) Ajustar, a esses dados, o modelo $E(Y) = \beta_0 + \beta_1 X_1 + \beta_3 X_3$

(b) Testar, com um nível de significância 5%, as seguintes hipóteses:

i. $H_0 : \beta_1 = \beta_3 = 0$ *vs* H_a : não H_0

ii. $H_0 : \beta_1 = 0$ *vs* $H_a : \beta_1 \neq 0$

iii. $H_0 : \beta_3 = 0$ *vs* $H_a : \beta_3 \neq 0$

(c) Obter $r_{Y1.3}$, $r_{Y3.1}$, $r_{13.Y}$ e $R^2_{Y.13}$.

(d) Testar, com um nível de significância 5%, a hipótese $H_0 : \rho_{Y1.3} = 0$ *vs* $H_a : \rho_{Y1.3} \neq 0$.

4. Mostre que o valor da estatística F para o teste da hipótese $H_0 : \beta_1 = \beta_2 = 0$, na análise de variância, pode ser reduzido a

$$F = \frac{R^2_{Y.12}}{1 - R^2_{Y.12}} \frac{n-k-1}{k}.$$

Tabela 1. Análise físico química de méis silvestres, produzidos por *Apis mellifera* L. 1758 (Hymenoptera: Apidae) em 1999, provenientes de 94 localidades do estado de São Paulo.

Local	*Condut.*	*N*	*Cinzas*	*Local*	*Condut.*	*N*	*Cinzas*
1	341,00	0,1735	0,1329	48	508,67	0,1671	0,3135
2	1177,67	0,2291	0,6721	49	951,00	0,1013	0,3086
3	614,33	0,2335	0,2266	50	902,33	0,1443	0,3786
4	751,67	0,1227	0,4187	51	1204,67	0,2298	0,6252
5	507,67	0,3238	0,1621	52	222,00	0,0543	0,0636
6	1110,67	0,4834	0,6120	53	544,00	0,1438	0,2920
7	290,00	0,2168	0,1368	54	349,67	0,1295	0,2012
8	329,00	0,2081	0,1174	55	340,33	0,0953	0,2085
9	367,67	0,2236	0,1382	56	405,00	0,0785	0,1454
10	846,67	0,3159	0,9188	57	803,00	0,1499	0,2869
11	626,00	0,2707	0,3721	58	235,67	0,0815	0,0325
12	592,33	0,2981	0,3518	59	191,00	0,0536	0,0578
13	216,67	0,2187	0,0999	60	525,67	0,1614	0,1503
14	743,67	0,4118	0,4004	61	649,00	0,0912	0,1326
15	626,00	0,4061	0,2136	62	769,33	0,2032	0,2792
16	371,67	0,2080	0,0916	63	801,00	0,1343	0,2085
17	387,00	0,1967	0,0807	64	561,33	0,1213	0,2829
18	391,67	0,2808	0,0838	65	389,00	0,0647	0,1296
19	595,67	0,2869	0,2144	66	977,67	0,1716	0,4197
20	947,33	0,3343	0,3816	67	660,33	0,2007	0,1967
21	329,67	0,1975	0,1264	68	649,00	0,0685	0,2610
22	230,67	0,1659	0,0897	69	465,33	0,0756	0,2024
23	160,67	0,1323	0,1296	70	736,67	0,1018	0,3426
24	382,67	0,1611	0,1578	71	1097,33	0,1568	0,3861
25	241,00	0,1757	0,1338	72	853,67	0,1114	0,3751
26	281,33	0,1220	0,1577	73	452,33	0,1154	0,2476
27	405,00	0,2301	0,1672	74	542,67	0,3075	0,2870
28	500,33	0,2345	0,2392	75	328,00	0,1263	0,1251
29	469,00	0,5833	0,1872	76	814,00	0,2042	0,3784
30	273,33	0,2264	0,0668	77	240,00	0,1203	0,1038
31	786,00	0,0986	0,2570	78	807,67	0,2499	0,4004
32	1175,33	0,1379	0,3985	79	639,00	0,3813	0,2109
33	1257,67	0,1219	0,5937	80	927,33	0,3303	0,3802
34	280,67	0,0778	0,0514	81	887,00	0,2491	0,2545
35	214,67	0,0848	0,0972	82	206,00	0,0485	0,0896
36	191,33	0,0507	0,1179	83	590,67	0,3019	0,2629
37	874,00	0,1457	0,5010	84	541,67	0,2739	0,2270
38	891,67	0,1220	0,3947	85	406,67	0,1877	0,3260
39	355,33	0,0658	0,1647	86	1064,33	0,2247	0,2771
40	287,00	0,1150	0,1404	87	844,67	0,1857	0,2766
41	198,33	0,0736	0,0930	88	869,00	0,3262	0,2289
42	647,67	0,1446	0,2002	89	738,00	0,2852	0,2795
43	599,00	0,1627	0,2336	90	263,67	0,1295	0,0893
44	508,67	0,1999	0,2307	91	400,33	0,0844	0,1913
45	391,00	0,1033	0,1025	92	677,00	0,1360	0,3007
46	471,00	0,1104	0,1879	93	580,33	0,1600	0,3218
47	261,00	0,1930	0,1822	94	397,67	0,2786	0,0956

Condut. = condutividade elétrica, em μS.
N = proporção de nitrogênio proteico.
Cinzas = proporção de cinzas.

Fonte: Assessoria estatística, prestada por S.S. Zocchi (LCE/ESALQ/USP) a Gleuber M. Teixeira , em 25/02/1999.

Capítulo 6

Métodos de Seleção de Variáveis

6.1 Introdução

O ajuste de um modelo aos dados pode ser encarado como uma maneira de substituir um conjunto de dados observados Y por um conjunto de valores estimados $\hat{\mu}$ a partir de um modelo com um número, relativamente pequeno, de parâmetros. Logicamente, os $\hat{\mu}$'s não serão exatamente iguais aos Y's, e a questão, então que aparece é em quanto eles diferem. Isso porque, uma discrepância pequena pode ser tolerável enquanto que uma discrepância grande, não. Assim, o objetivo é determinar quantos termos são necessários na estrutura linear para uma descrição razoável dos dados.

Um número grande de variáveis explanatórias (ou covariáveis) pode levar a um modelo que explique bem os dados mas com um aumento de complexidade na interpretação. Por outro lado, um número pequeno de variáveis explanatórias (ou covariáveis) pode levar a um modelo de interpretação fácil, porém, que se ajuste pobremente aos dados. Em geral, deseja-se um modelo intermediário que explique bem os dados e que seja parcimonioso, isto é, com o menor número possível de parâmetros.

Dadas n observações, a elas podem ser ajustados modelos contendo até n parâmetros. Seja $M_1,\ M_2,\ \ldots, M_r$ uma sequência de modelos encaixados com $p_1 < p_2 < \ldots < p_r$ parâmetros, matrizes de modelos $\boldsymbol{X}_1,\ \boldsymbol{X}_2\ ,\ \ldots,\ \boldsymbol{X}_r$. Tem-se que

$$SQRes_1 \geq SQRes_2 \geq \ldots \geq SQRes_r \geq 0.$$

O modelo mais simples é o **modelo nulo** que tem um único parâmetro, representado por um valor comum a todas as observações. A matriz do modelo, então, reduz-se a um vetor coluna, formado de 1's. Esse modelo atribui toda a variação entre os Y's ao componente aleatório. No outro extremo está o **modelo saturado** ou completo que tem n parâmetros, um para cada observação. Ele atribui toda a variação ao componente sistemático e, portanto, ajusta-se perfeitamente, reproduzindo os próprios dados.

Na prática, o modelo nulo é simples demais e o modelo saturado não é informativo pois não resume os dados, mas simplesmente os repete. Existem, contudo, dois outros modelos limitantes, porém, menos extremos. Certos parâmetros têm que estar no modelo como é o caso, por exemplo, de efeitos de blocos. O modelo contendo apenas esses parâmetros é chamado **modelo minimal** pois é aquele que contém os termos obrigatórios. Por outro lado, o modelo que contém o maior número

de termos que podem ser considerados é chamado de **modelo maximal** e, em geral, fornecerá a estimativa de σ^2. Os termos desses modelos extremos são, geralmente, obtidos por interpretações *a priori*, da estrutura dos dados. Como exemplo, pode-se usar um experimento em blocos casualizados com os tratamentos sendo doses de um nutriente. Tem-se, então:

Modelo nulo: $\mathrm{E}(Y) = \beta_0$
Modelo minimal: $\mathrm{E}(Y) = \beta_0 + \beta_j$
Modelo maximal: $\mathrm{E}(Y) = \beta_0 + \beta_j + \alpha_i$
Modelo saturado: $\mathrm{E}(Y) = \beta_0 + \beta_j + \alpha_i + \beta\alpha_{ij}$
Modelo sob pesquisa: $\mathrm{E}(Y) = \beta_0 + \beta_j + \gamma X_i$

sendo β_0 uma constante, β_j o efeito de blocos, α_i o efeito de doses de adubos consideradas como efeito qualitativo (fator) e γ o coeficiente de regressão linear, considerando as doses X_i do nutriente como efeito quantitativo.

No caso de modelos de regressão linear múltipla, deseja-se modelar uma variável resposta Y em função das variáveis explicativas $X_1, X_2, \ldots, X_k$, podendo envolver funções dessas variáveis como, por exemplo, quadrados e produtos delas. A seleção de um modelo pode, então, estar baseada em duas direções:

(i) um modelo com o maior número possível de variáveis, de tal forma que a predição seja a melhor possível e

(ii) um modelo com o menor número possível de variáveis necessárias para um bom ajuste, pois a obtenção de informações a respeito de um número grande de variáveis aumenta o custo.

Entretanto, selecionar um modelo para um conjunto de variáveis não é uma tarefa simples e um problema adicional que pode aparecer é a existência de colinearidade entre as variáveis. Algumas vezes, a natureza dos dados mostra o que está ocorrendo, mas, em geral, há necessidade da ajuda de testes e gráficos para escolher modelos que representem bem os dados. Em geral, escolhem-se os modelos com menor $SQRes$, ou de forma equivalente maior R^2. Dois tipos de procedimentos podem ser podem ser adotados.

(i) métodos de seleção de subconjuntos das variáveis e

(ii) métodos de seleção das variáveis passo a passo - incluem ou excluem variáveis, sequencialmente.

6.2 Critérios usados na seleção de variáveis

A seleção de modelos pode estar baseada em diferentes critérios, como, por exemplo, na escolha de subconjuntos de tamanho pré-determinado, ou, então, na comparação de estatísticas com valores tabelados de referência. Dentre as estatísticas, destacam-se:

(i) Teste F - A comparação entre dois modelos M_p e M_q com $p < q$ parâmetros pode ser feita usando-se:

$$F = \frac{SQRes_p - SQRes_q}{(q-p)QMRes}$$

sendo $QMRes$ uma estimativa de σ^2, em geral, obtida a partir do modelo maximal (ou, então, alguma estimativa do *erro puro*). Geralmente, usado nos métodos de regressão passo a passo.

(ii) Coeficiente de determinação e coeficiente de determinação ajustado

Dados, respectivamente, por:

$$R^2 = 1 - \frac{SQRes}{SQTotal}$$

e

$$\bar{R}^2 = R^2 - \frac{1}{n-p}(1-R^2),$$

geralmente, usados no método de todas as regressões possíveis.

(iii) Estatística de Mallows - Mallows (1973) propôs para a seleção de modelos uma estatística baseada na razão entre a SQRes do modelo sob estudo, com $k = p-1$ variáveis explanatórias, e uma estimativa para σ^2, isto é,

$$C_p = \frac{SQRes_p}{\hat{\sigma}^2} - (n-2p).$$

Mostra-se que C_p é um estimador de

$$\Gamma_p = \mathrm{E}\left(\frac{SQRes_p}{\sigma^2}\right) - (n-2p) \approx (n-p) - (n-2p) = p.$$

Para um dado número $k = p-1$ de variáveis selecionadas, valores grandes de C_p indicam modelos com $QMRes$ grandes. Qualquer modelo com $C_p > p$ é uma indicação de que existe um viés devido a um modelo mal especificado (número insuficiente de termos). Se, por outro lado, $C_p < p$, tem-se que o modelo maximal está superparametrizado, isto é, contém muitas variáveis.

Mallows recomenda que se faça o gráfico de C_p versus p e se escolha como o melhor modelo aquele em que o mínimo de C_p aproxima-se de p (sugere-se fazer a reta que passa pelos pontos (p,p). As magnitudes das diferenças da estatística C_p entre o ótimo e a vizinhança do ótimo para cada submodelo, são, também, de interesse.

(iv) Critério de informação de Akaike (AIC)

$$AIC = -2\log L + 2(\text{número de parâmetros ajustados})$$

sendo L a função de verossimilhança.

(v) Critério de informação de Bayes (BIC)

$$BIC = -2\log L + \log[n(\text{número de parâmetros ajustados})]$$

sendo L a função de verossimilhança. O critério BIC penaliza mais fortemente modelos com um maior número de parâmetros do que o AIC tendendo, dessa forma, a selecionar modelos com um menor número de parâmetros.

6.3 Métodos de seleção de variáveis

Todas as regressões possíveis e melhor subconjunto

Esse método é bastante trabalhoso, pois envolve o ajuste de 2^k modelos. Os modelos, são separados em grupos de modelos com r variáveis, $r = 1, 2, \ldots, k$, sendo cada grupo ordenado de acordo com algum critério, por exemplo, R^2, escolhendo-se os modelos com maior R^2 de cada grupo ou, então, menor $QMRes$. Se dois modelos apresentam valores próximos de R^2 e de $QMRes$, escolhe-se aquele com menor número de parâmetros. É usada, também, a estatística C_p de Mallows.

Como o número de regressões possíveis cresce com o aumento do número de parâmetros, foram propostos os métodos de regressão passo a passo, que embora não sejam *ótimos*, requerem um tempo computacional bem menor do que o de todas as regressões possíveis.

Método do passo atrás (*backward*)

Esse método consiste em ajustar, inicialmente, o modelo completo e, a seguir, eliminar variáveis, uma a uma, com

(i) menor correlação parcial com a resposta Y;

(ii) menor diminuição no R^2 ou

(iii) menor diminuição significativa no teste F parcial ou no teste t parcial,

de acordo com algum critério de parada.

Os passos, baseados no teste F parcial, a serem seguidos, são:

Passo 1

Ajustar o modelo completo com l variáveis e obter $SQRes_c$ com $n - l$ graus de liberdade.

Passo 2

Para cada uma das l variáveis do modelo completo do **Passo 1**, considerar o modelo reduzido, com a retirada de uma variável e calcular $SQRes_r$ com $n - l + 1$ graus de liberdade e

$$F = \frac{SQRes_r - SQRes_c}{QMRes_c}$$

Passo 3

Obter $F_{\text{mín}}$

Passo 4

Comparar $F_{\text{mín}}$ com F_{sai} (percentil da tabela de F com 1 e $n - l$ graus de liberdade a um nível de significância α; usa-se, em geral, $\alpha = 0,10$)

(i) se $F_{\text{mín}} > F_{\text{sai}}$, não eliminar nenhuma variável e parar o processo, ficando o modelo completo com l variáveis;

(i) se $F_{\text{mín}} < F_{\text{sai}}$, eliminar a variável com $F_{\text{mín}}$ e voltar ao **Passo 1** com novo modelo completo com $l = l - 1$ variáveis.

Método do passo a frente (*forward*)

Esse método consiste em incluir, inicialmente, no modelo a variável com maior coeficiente de correlação simples com a variável resposta e, a seguir, variáveis, uma a uma, com

(i) maior correlação parcial com a resposta Y;

(ii) maior aumento no R^2 ou

(iii) maior aumento significativo no teste F parcial ou no teste t parcial,

de acordo com algum critério de parada.

Os passos, baseados no teste F parcial, a serem seguidos são, então,

Passo 1

Ajustar o modelo reduzido com m variáveis e obter $SQRes_r$ com $n-m$ graus de liberdade.

Passo 2

Para cada uma das variáveis não pertencentes ao modelo do **Passo 1**, considerar o modelo completo, com a adição de uma variável extra e calcular $SQRes_c$ com $n-m-1$ graus de liberdade e

$$F = \frac{SQRes_r - SQRes_c}{QMRes_c}$$

Passo 3

Obter $F_{\text{máx}}$

Passo 4

Comparar $F_{\text{máx}}$ com F_{en} (percentil da tabela de F com 1 e $n-m-1$ graus de liberdade a um nível de significância α; usa-se, em geral, $\alpha = 0,10$)

(i) se $F_{\text{máx}} > F_{\text{en}}$, incluir a variável com $F_{\text{máx}}$ e voltar ao **Passo 1** com novo modelo reduzido com $m = m + 1$ variáveis;

(i) se $F_{\text{máx}} < F_{\text{en}}$, não incluir a variável com o $F_{\text{máx}}$ e parar o processo, ficando o modelo completo com m variáveis.

Método do passo a frente passo atrás (*stepwise*)

Esse método consiste na mistura dos dois anteriores. Os passos, baseados no teste F parcial, a serem seguidos são, então,

Passo 1

Ajustar o modelo reduzido com m variáveis e obter $SQRes_r$ com $n-m$ graus de liberdade.

Passo 2

Para cada uma das variáveis não pertencentes ao modelo do **Passo 1**, considerar o modelo completo, com a adição de uma variável extra e calcular $SQRes_c$ com $n-m-1$ graus de liberdade e

$$F = \frac{SQRes_r - SQRes_c}{QMRes_c}$$

Passo 3

Obter $F_{\text{máx}}$

Passo 4

Comparar $F_{\text{máx}}$ com F_{en} (percentil da tabela de F com 1 e $n-m-1$ graus de liberdade a um nível de significância α; usa-se, em geral, $\alpha = 0,10$)

(i) se $F_{\text{máx}} > F_{\text{en}}$, incluir a variável com $F_{\text{máx}}$ e passar para o **Passo 5** com modelo completo com $l = m + 1$ variáveis;

(i) se $F_{\text{máx}} < F_{\text{en}}$, não incluir a variável com $F_{\text{máx}}$ e passar para o **Passo 5** com modelo completo com $l = m$ variáveis.

Passo 5

Ajustar o modelo completo com l variáveis e obter $SQRes_c$ com $n - l$ graus de liberdade.

Passo 6

Para cada uma das l variáveis do modelo completo do **Passo 4**, considerar o modelo reduzido, com a retirada de uma variável e calcular $SQRes_r$ com $n - l + 1$ graus de liberdade e

$$F = \frac{SQRes_r - SQRes_c}{QMRes_c}$$

Passo 7

Obter $F_{\text{mín}}$

Passo 8

Comparar $F_{\text{mín}}$ com F_{sai} (percentil da tabela de F com 1 e $n - l$ graus de liberdade a um nível de significância α; usa-se, em geral, $\alpha = 0,10$)

(i) se $F_{\text{mín}} > F_{\text{sai}}$, não eliminar nenhuma variável e voltar ao **Passo 1** com novo modelo reduzido com $m = l$ variáveis e parar o processo se no **Passo 4** nenhuma variável for incluída;

(i) se $F_{\text{mín}} < F_{\text{sai}}$, eliminar a variável com $F_{\text{mín}}$ e voltar ao **Passo 1** com novo modelo reduzido com $m = l - 1$ variáveis.

6.4 Exemplo

Considere os dados da Tabela 5.1, referentes a um estudo sobre a resposta da cultura do milho como função da quantidade de fosfato, porcentagem de saturação de bases (X_2) e sílica (X_3) em solos ácidos. Suponha que a variável resposta de interesse seja Y_1 e as demais, X_1, X_2 e X_3, sejam variáveis explicativas. Suponha, ainda, que se deseja selecionar o "melhor" subgrupo de variáveis explicativas. Uma alternativa seria ajustar todos os possíveis modelos (no caso $2^3 = 8$ modelos) e escolher o melhor segundo algum critério. No caso de o número de possíveis modelos ser muito elevado, podem-se utilizar algoritmos como o método "branch-and-bound" que só faz o ajuste dos modelos mais prováveis de serem os melhores. Entre os critérios mais utilizados, estão os seguintes:

1. Critério de informação de Akaike (AIC)
2. Critério de informação de Bayes (BIC)
3. R^2 ajustado
4. Estatística C_p de Mallow

No R, a seleção pode ser feita usando-se a função `step` da seguinte maneira:

```
> ML1<-lm(Y1~X1+X2+X3)
> step(ML1)
Start:  AIC= 171.84
 Y1 ~ X1 + X2 + X3

       Df Sum of Sq      RSS      AIC
- X3    1     16542  1709817      170
<none>               1693275      172
- X2    1    619186  2312461      174
- X1    1  19874994 21568268      205

Step:  AIC= 169.98
 Y1 ~ X1 + X2

       Df Sum of Sq      RSS      AIC
<none>               1709817      170
+ X3    1     16542  1693275      172
- X2    1    693808  2403625      173
- X1    1  20088899 21798716      204

Call:
lm(formula = Y1 ~ X1 + X2)

Coefficients:
(Intercept)           X1           X2
   1864.283        1.009      -21.855
```

O algoritmo de escolha padrão dessa função é semelhante ao "stepwise" considerando, no entanto, o valor de AIC para a escolha do modelo. Nesse método, inicialmente parte-se do modelo completo, isto é, contendo todas as 3 variáveis explicativas (X_1, X_2 e X_3), cujo valor de AIC é 171,84. Em seguida, são ajustados os modelos com uma variável a menos (2 variáveis) e é escolhido o com menor AIC, se for menor ou igual ao com todas as 3 variáveis. No caso, vê-se que a retirada da variável X_3 do modelo reduz o AIC a 169,98. Partindo desse último modelo (com X_1 e X_2), vê-se, então, se

há a necessidade de incluir X_3, excluir X_2 ou excluir X_1. No caso, nenhuma dessas alternativas reduz o AIC, levando à escolha do modelo final, com X_1 e X_2.

Uma alternativa bastante utilizada é ajustar todos os possíveis modelos e escolher o que possui o menor número de parâmetros e valor da estatística C_p de Mallow menor ou igual ao do número de parâmetros do modelo. No R, pode-se fazê-lo por meio de:

```
> selecao<-leaps(x=cbind(X1,X2,X3), y=Y1, method=c("Cp"))
> npar<-selecao$size
> Cp<-selecao$Cp
> cbind(npar,Cp,selecao$which)
  npar         Cp 1 2 3
1    2   4.195131 1 0 0
1    2 118.737040 0 1 0
1    2 136.922502 0 0 1
2    3   2.097695 1 1 0 <== modelo indicado
2    3   5.656740 1 0 1
2    3 119.376079 0 1 1
3    4   4.000000 1 1 1
```

No caso, vê-se que o modelo mais adequado é o modelo (1 1 0), isto é, o que inclui as variáveis explicativas X_1 e X_2. É bastante comum, também, para facilitar a escolha do modelo, construir o gráfico de C_p em função do número p de parâmetros do modelo, o que pode ser feito do seguinte modo:

```
> plot(npar,Cp,xlab="Nº de parâmetros (p)", ylab="Cp",ylim=c(0,15))
> abline(0,1)
> identify(npar,Cp)
[1] 4
> selecao$which[4,]
    1     2     3
 TRUE  TRUE FALSE
```

gerando o gráfico da esquerda da figura 6.1.

Uma outra alternativa seria escolher o modelo com o maior coeficiente de determinação ajustado, o que pode ser feito de forma análoga ao caso anterior, do seguinte modo:

```
> selecao<-leaps(x=cbind(X1,X2,X3), y=Y1, method=c("adjr2"))
> npar<-selecao$size
> r2.aj<-selecao$adjr2
> cbind(npar,R2.aj,selecao$which)
  npar     r2.aj 1 2 3
1    2 0.9112043 1 0 0
```

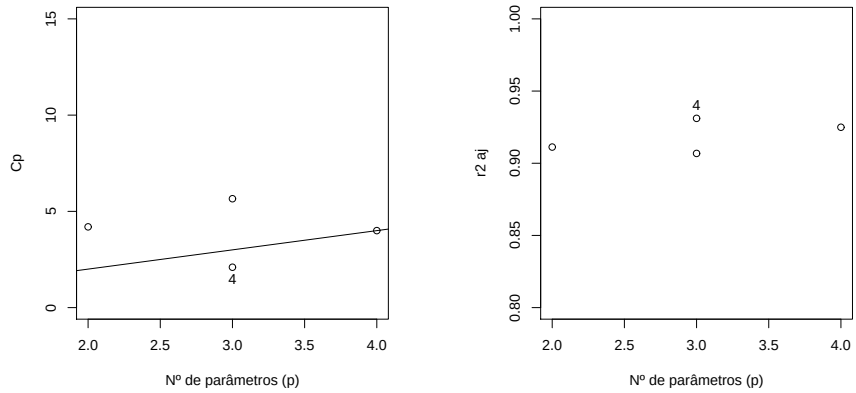


Figura 6.1: Gráficos da estatística C_p de Mallow em função do número p de parâmetros do modelo, com respectiva reta de referência $C_p = p$, e do valor do coeficiente de determinação ajustado em função de p, respectivamente.

```
1    2 0.1947032 0 1 0
1    2 0.0809465 0 0 1
2    3 0.9310930 1 1 0  <== Modelo com maior R2 ajustado
2    3 0.9068060 1 0 1
2    3 0.1307816 0 1 1
3    4 0.9249357 1 1 1
> plot(npar,r2.aj,xlab="Nº de parâmetros (p)", ylab="r2 aj",ylim=c(0.8,1))
> abline(0,1)
> identify(npar,r2.aj)
[1] 4
> selecao$which[4,]
    1     2     3
 TRUE  TRUE FALSE
```

gerando o gráfico da direita da figura 6.1.

6.5 Exercícios

1. Considerando-se os dados apresentados na Tabela 8.1,

   (a) Selecione as variáveis regressoras utilizando-se os métodos:

      i. Passo atrás (*backward*).

ii. Passo a frente (*forward*).

iii. Passo a passo (*stepwise*).

(b) Calcule a estatística C_p de Mallows para todos os possíveis modelos.

(c) Construa o gráfico de C_p *vs* p e interprete-o.

(d) Compare os resultados.

2. Repita o exercício anterior, considerando-se os dados apresentados na Tabela 9.1. e dizer qual é o "modelo correto".

Tabela 9.1. Valores de X_{1i}, X_{2i}, X_{3i} e Y_i $(i = 1, \ldots, 5)$.

i	X_{1i}	X_{2i}	X_{3i}	Y_i
1	1	1004	6,0	5
2	200	806	7,3	6
3	-50	1058	11,0	8
4	909	100	13,0	9
5	506	505	13,1	11

Fonte: WEISBERG, S. (1985). *Applied Linear Regression* 2ª ed. Ed. Wiley, p.221.

```
options nodate ps=25;
data aula9;
input X1 X2 X3 Y1;
cards;
-2    2   -2    8.5
-1   -1    0    1
-1    0    0    4
-1    0    0    4
-1    1    0    5
 1   -1    0    3
 1    0    0    6
 1    0    0    6
 1    1    0    7
 0    0   -1    5
 0    0    0    5
 0    0    0    5
 0    0    1    3
 2   -2    2    0.5
;

proc print data=aula9;
run;

proc glm data=aula9;
     model Y1=X1 X2 X3/ss1 ss2;
run;

proc reg data=aula9;
     model Y1=X1 X2 X3;
     ex9_1i:  test X1=0, X2=0;
     ex9_1ii: test X1=X2;
     ex9_1iii:test X1=X3;
     ex9_1iv: test X1=1,X1+X2+X3=1;
     ex9_1v:  test X2=-2*X3, 3*X1=2*X2;
run;


proc reg data=aula9;
     model Y1=X1 X2 X3/selection=backward;
run;

proc reg data=aula9;
     model Y1=X1 X2 X3/selection=forward;
run;

proc reg data=aula9;
     model Y1=X1 X2 X3/selection=stepwise;
run;

proc reg data=aula9;
     model Y1=X1 X2 X3/selection=rsquare;
run;

proc reg data=aula9 graphics;
     model Y1=X1 X2 X3/selection=rsquare noprint;
     plot cp.*np./chocking=red cmallows=blue vaxis=0 to 20 by 5;
run;
```

1. Os dados que se seguem referem-se a um estudo da resposta da cultura do milho (Y) ao fosfato, porcentagem de saturação de bases e sílica em solos ácidos. A resposta, em porcentagem, foi medida como a diferença entre as produções nas parcelas recebendo fosfato e aquelas não recebendo fosfato (X_1), dividida pelas produções das parcelas recebendo fosfato, e multiplicadas por 100. Considerando-se esses dados,

   (a) Selecione as variáveis regressoras utilizando-se os métodos:

      i. Passo atrás ("Backward").

      ii. Passo a frente ("Forward").

      iii. Passo a passo ("Stepwise").

   (b) Calcule a estatística C_p de Mallows para todos os possíveis modelos.

   (c) Construa o gráfico de C_p *vs* p e interprete-o.

   (d) Compare os resultados.

Y	X_1	X_2	X_3	Y	X_1	X_2	X_3
88	844	67	5,75	18	1262	74	6,10
80	1678	57	6,05	18	4624	69	6,05
42	1573	39	5,45	4	5249	76	6,15
37	3025	54	5,70	2	4258	80	5,55
37	653	46	5,55	2	2943	79	6,40
20	1991	62	5,00	-2	5092	82	6,55
20	2187	69	6,40	-7	4496	85	6,50

Fonte: Steel, R.G.D & Torrie, J.H (1980). *Principles and Procedures os Statistics. A Biometrical Approach.* 2ª ed. Ed. McGraw-Hill, p.324.

em que Y=resposta ao fosfato, em porcentagem, X_1=produtividade na testemunha, em lb/acre, X_2=porcentagem de saturação de bases e X_3=pH do solo.

2. Repita o exercício anterior, considerando-se os dados apresentados na tabela 11.2. e dizer qual é o "modelo correto".

Tabela 11.2. Valores de X_{1i}, X_{2i}, X_{3i} e Y_i $(i = 1, \ldots, 5)$.

i	X_{1i}	X_{2i}	X_{3i}	Y_i
1	1	1004	6,0	5
2	200	806	7,3	6
3	-50	1058	11,0	8
4	909	100	13,0	9
5	506	505	13,1	11

Fonte: WEISBERG, S. (1985). *Applied Linear Regression* 2ª ed. Ed. Wiley, p.221.

Programas no R

```
rm(list=ls(all=TRUE))

# Exemplo usando os dados de Hoffman #
#####################################
Hoffman<-read.table("c:/Hoffman1.txt",sep="\t",header=TRUE) #
Entrada dos dados attach(Hoffman) pairs(Hoffman) # Gráficos de
dispersão

library(leaps) # Biblioteca de funções para a seleção de variáveis

Sintaxe da função leaps:
      leaps(x=, y=, wt=rep(1, NROW(x)), int=TRUE, method=c("Cp", "adjr2", "r2"),
      nbest=10, names=NULL, df=NROW(x), strictly.compatible=T)

seleção<-leaps(x=cbind(X1,X2,X3),y=Y1,method=c("Cp"))
npar<-seleção$size Cp<-seleção$Cp cbind(npar,Cp)
plot(npar,Cp,xlab="N° de parâmetros (p)", ylab="Cp",ylim=c(0,30))
abline(0,1)


Sintaxe da função regsubsets:

      regsubsets(x=, y=, weights=rep(1, length(y)), nbest=1, nvmax=8, force.in=NULL,
      force.out=NULL, intercept=TRUE, method=c("exhaustive", "backward", "forward", "seqrep"),
      really.big=FALSE,...)

seleção<-regsubsets(x=cbind(X1,X2,X3),y=Y1,method=c("exhaustive"))
seleção

     Subset selection object
     3 Variables  (and intercept)
        Forced in Forced out
     X1     FALSE      FALSE
     X2     FALSE      FALSE
     X3     FALSE      FALSE
     1 subsets of each size up to 3
     Selection Algorithm: exhaustive

par(mfrow=c(1,2)) plot(seleção,scale="r2")
plot(seleção,scale="Cp")


# Exemplo usando os dados Swiss #
################################## ?swiss # Fornece informações
sobre os dados data(swiss) attach(swiss) pairs(swiss, panel =
panel.smooth, main = "swiss data",
      col = 3 + (swiss$Catholic > 50))
lm1<-lm(Fertility ~ . , data = swiss) summary(lm1)
seleção<-regsubsets(x=swiss[,2:6],y=swiss[,1],method=c("exhaustive"))
summary(seleção) par(mfrow=c(1,2)) plot(seleção,scale="r2")
plot(seleção,scale="Cp")

seleção<-leaps(x=swiss[,2:6],y=swiss[,1],method=c("Cp")) seleção
seleção$which[26,] npar<-seleção$size Cp<-seleção$Cp
cbind(npar,Cp) plot(npar,Cp,xlab="N° de parâmetros (p)",
```

```
ylab="Cp",ylim=c(0,30)) abline(0,1)


Função step - Select a formula-based model by AIC.
      step(object, scope, scale = 0,
      direction = c("both", "backward", "forward"),
      trace = 1, keep = NULL, steps = 1000, k = 2, ...)

slm1 <- step(lm1) summary(slm1) slm1$anova

# Exemplo usando os dados "fitness" do SAS #
##########################################
fitness<-read.table("c:/Fitness.txt",sep="\t",header=TRUE) #
Entrada dos dados attach(fitness) pairs(fitness,panel =
panel.smooth) # Gráficos de dispersão lm1<-lm(oxy ~ . , data =
fitness) # Análise de regressão parcial summary(lm1)

      Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
      (Intercept) 102.93448   12.40326   8.299 1.64e-08 ***
      age          -0.22697    0.09984  -2.273  0.03224 *
      weight       -0.07418    0.05459  -1.359  0.18687
      runtime      -2.62865    0.38456  -6.835 4.54e-07 ***
      rstpulse     -0.02153    0.06605  -0.326  0.74725
      runpulse     -0.36963    0.11985  -3.084  0.00508 **
      maxpulse      0.30322    0.13650   2.221  0.03601 *
      ---
      Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

      Residual standard error: 2.317 on 24 degrees of freedom
      Multiple R-Squared: 0.8487,     Adjusted R-squared: 0.8108
      F-statistic: 22.43 on 6 and 24 DF,  p-value: 9.715e-09

library(car) av.plots(lm1,ask=FALSE) # Gráficos das variáveis
adicionadas

# Seleção de variáveis # library(leaps) # Biblioteca de funções
para a seleção de variáveis
seleção<-leaps(x=fitness[,c(1,2,4,5,6,7)],y=fitness[,3],
         method=c("Cp"),nbest=10,names=names(fitness[c(1,2,4,5,6,7)]))
  # nbest=10 => apresenta somente os 10 melhores grupos
  #             de k variáveis, para cada k

seleção n.par<-seleção$size Cp<-seleção$Cp
cbind(n.par,Cp,seleção$which) seleção$which[27,]

# Gráfico de Cp vs p # plot(n.par,Cp,xlab="No. de parâmetros (p)",
ylab="Cp",ylim=c(0,20)) abline(0,1,col="blue") identify(n.par,Cp)


Função step - Select a formula-based model by AIC.
      step(object, scope, scale = 0,
      direction = c("both", "backward", "forward"),
      trace = 1, keep = NULL, steps = 1000, k = 2, ...)

      k: the multiple of the number of degrees of freedom used for the
         penalty. Only ‘k = 2’ gives the genuine AIC: ‘k = log(n)’ is
```

```
         sometimes referred to as BIC or SBC.

?step slm1 <- step(lm1) summary(slm1) slm1$anova

library(MASS) ?stepAIC slm1 <- stepAIC(lm1) summary(slm1)
slm1$anova

Função regsubsets - Generic function for regression subset
selection with methods for
                    formula and matrix arguments.
      regsubsets(x=, y=, weights=rep(1, length(y)), nbest=1, nvmax=8, force.in=NULL,
      force.out=NULL, intercept=TRUE, method=c("exhaustive", "backward", "forward", "seqrep"),
      really.big=FALSE,...)

seleção<-regsubsets(x=fitness[,c(1,2,4,5,6,7)],y=fitness[,3],nbest=1,method=c("exhaustive"))
seleção summary(seleção) par(mfrow=c(1,2)) plot(seleção)
plot(seleção,scale="Cp")
```

Programas no SAS

```
/*******************************************/
/* Exemplo do SAS: Aerobic Fitness Prediction*/
/*******************************************/
  data fitness;
     input age weight oxy runtime rstpulse runpulse maxpulse;
     cards;
  44 89.47  44.609 11.37 62 178 182
  40 75.07  45.313 10.07 62 185 185
  44 85.84  54.297  8.65 45 156 168
  42 68.15  59.571  8.17 40 166 172
  38 89.02  49.874  9.22 55 178 180
  47 77.45  44.811 11.63 58 176 176
  40 75.98  45.681 11.95 70 176 180
  43 81.19  49.091 10.85 64 162 170
  44 81.42  39.442 13.08 63 174 176
  38 81.87  60.055  8.63 48 170 186
  44 73.03  50.541 10.13 45 168 168
  45 87.66  37.388 14.03 56 186 192
  45 66.45  44.754 11.12 51 176 176
  47 79.15  47.273 10.60 47 162 164
  54 83.12  51.855 10.33 50 166 170
  49 81.42  49.156  8.95 44 180 185
  51 69.63  40.836 10.95 57 168 172
  51 77.91  46.672 10.00 48 162 168
  48 91.63  46.774 10.25 48 162 164
  49 73.37  50.388 10.08 67 168 168
  57 73.37  39.407 12.63 58 174 176
  54 79.38  46.080 11.17 62 156 165
  52 76.32  45.441  9.63 48 164 166
  50 70.87  54.625  8.92 48 146 155
  51 67.25  45.118 11.08 48 172 172
  54 91.63  39.203 12.88 44 168 172
  51 73.71  45.790 10.47 59 186 188
```

```
   57 59.08  50.545  9.93 49 148 155
   49 76.32  48.673  9.40 56 186 188
   48 61.24  47.920 11.50 52 170 176
   52 82.78  47.467 10.50 53 170 172
   ;
   proc reg data=fitness graphics;
      model oxy=age weight runtime runpulse rstpulse maxpulse
            / selection=forward;
      model oxy=age weight runtime runpulse rstpulse maxpulse
            / selection=backward;
      model oxy=age weight runtime runpulse rstpulse maxpulse
            / selection=maxr;
      model oxy=age weight runtime runpulse rstpulse maxpulse
            / selection=rsquare cp;
      plot cp.*np./chocking=red cmallows=blue vaxis=0 to 20 by 5;
   run;
```

==========

Capítulo 7

Polinômios Ortogonais

7.1 Introdução

No ajuste de um modelo de regressão polinomial de grau k do tipo

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_k X_i^k + \varepsilon_i$$

a n pares de observações, (X_i, Y_i), $i = 1, 2, \ldots, n$, a matriz $\boldsymbol{X'X}$ é frequentemente mal condicionada, isto é, quase singular, com problemas na estimação dos parâmetros e da matriz de variâncias e covariâncias. Para se evitar isso, é possível a construção de outros polinômios que sejam ortogonais entre si de tal forma que, no estudo da regressão, os coeficientes dos polinômios ortogonais sejam calculados independentemente. Decorre disso que cada coeficiente pode ser estimado isoladamente e consequentemente, testado isoladamente, sendo o nível conjunto de significância dado por $\alpha' = 1 - (1 - \alpha)^k$.

Seja, então, o modelo

$$Y_i = \alpha_0 + \alpha_1 P_{1i} + \alpha_2 P_{2i} + \cdots + \alpha_k P_{ki} + \varepsilon_i \tag{7.1}$$

com os polinômios

$$\begin{aligned}
P_{1i} &= \gamma_{10} + \gamma_{11} x_i \\
P_{2i} &= \gamma_{20} + \gamma_{21} x_i + \gamma_{22} x_i^2 \\
P_{3i} &= \gamma_{30} + \gamma_{31} x_i + \gamma_{32} x_i^2 + \gamma_{33} x_i^3 \qquad (7.2) \\
&\cdots \\
P_{ki} &= \gamma_{k0} + \gamma_{k1} x_i + \gamma_{k2} x_i^2 + \cdots + \gamma_{kk} x_i^k
\end{aligned}$$

sendo, em geral, $x_i = X_i - \bar{X}$ ou então no caso de níveis equidistantes, $x_i = \dfrac{X_i - \bar{X}}{q}$, para $\bar{X} = \dfrac{\sum_i X_i}{n}$

e q a distância entre os níveis. Na forma matricial, $\boldsymbol{Y} = \boldsymbol{X\theta} + \boldsymbol{\varepsilon}$, tem-se

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & P_{11} & P_{21} & \cdots & P_{k1} \\ 1 & P_{12} & P_{22} & \cdots & P_{k2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & P_{1n} & P_{2n} & \cdots & P_{kn} \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \cdots \\ \alpha_k \end{bmatrix} \quad \text{e} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{bmatrix}.$$

Do capítulo 3, tem-se que o sistema de equações normais é dado por:

$$\boldsymbol{X'X}\hat{\boldsymbol{\theta}} = \boldsymbol{X'Y}$$

sendo $\boldsymbol{\theta}$ estimado por:

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$$

com matriz de variâncias e covariâncias dada por

$$V(\hat{\boldsymbol{\theta}}) = V[(\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}] = (\boldsymbol{X'X})^{-1}\boldsymbol{X'}V(\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X'X})^{-1} = (\boldsymbol{X'X})^{-1}\sigma^2,$$

sendo que

$$\boldsymbol{X'X} = \begin{bmatrix} n & \sum_i P_{1i} & \sum_i P_{2i} & \cdots & \sum_i P_{ki} \\ \sum_i P_{1i} & \sum_i P_{1i}^2 & \sum_i P_{1i}P_{2i} & \cdots & \sum_i P_{1i}P_{ki} \\ \sum_i P_{2i} & \sum_i P_{2i}P_{1i} & \sum_i P_{2i}^2 & \cdots & \sum_i P_{2i}P_{ki} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_i P_{ki} & \sum_i P_{ki}P_{1i} & \sum_i P_{ki}P_{2i} & \cdots & \sum_i P_{ki}^2 \end{bmatrix} = \begin{bmatrix} n & 0 & 0 & \cdots & 0 \\ 0 & \sum_i P_{1i}^2 & 0 & \cdots & 0 \\ 0 & 0 & \sum_i P_{2i}^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \sum_i P_{ki}^2 \end{bmatrix}$$

e

$$\boldsymbol{X'Y} = \begin{bmatrix} \sum_i Y_i \\ \sum_i P_{1i}Y_i \\ \sum_i P_{2i}Y_i \\ \cdots \\ \sum_i P_{ki}Y_i \end{bmatrix},$$

pois, como os P_{ij} são construídos para serem ortogonais tem-se que a covariância entre eles deve ser igual a 0 e, portanto, $\sum_i P_{ji}P_{j'i} = 0$ para $j \neq j'$ e $\sum_i P_{ji} = 0$. Logo

$$(\boldsymbol{X'X})^{-1} = \begin{bmatrix} \frac{1}{n} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\sum_i P_{1i}^2} & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{\sum_i P_{2i}^2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \frac{1}{\sum_i P_{ki}^2} \end{bmatrix} \quad \text{e} \quad \hat{\boldsymbol{\theta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y} = \begin{bmatrix} \frac{1}{n}\sum_i Y_i \\ \frac{\sum_i P_{1i}Y_i}{\sum_i P_{1i}^2} \\ \frac{\sum_i P_{2i}Y_i}{\sum_i P_{2i}^2} \\ \cdots \\ \frac{\sum_i P_{ki}Y_i}{\sum_i P_{ki}^2} \end{bmatrix}.$$

Tem-se, portanto, que

$$\hat{\alpha_0} = \bar{Y} \qquad \text{e} \qquad \hat{\alpha_j} = \frac{\sum_i P_{ji}Y_i}{\sum_i P_{ji}^2}, \qquad j = 1, \ldots, k,$$

obtendo-se, então,

$$\hat{Y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 P_{1i} + \hat{\alpha}_2 P_{2i} + \cdots + \hat{\alpha}_k P_{ki}$$

que facilmente pode ser transformada em função da variável original X_i. Além disso,

$$V(\hat{\alpha_0}) = \frac{\sigma^2}{n}, \qquad V(\hat{\alpha_j}) = \frac{\sigma^2}{\sum_i P_{ji}^2}, \quad j = 1, \ldots, k \qquad \text{e} \qquad Cov(\hat{\alpha_j}, \hat{\alpha_{j'}}) = 0, \quad j \neq j'.$$

Conforme já foi visto, uma estimativa para σ^2 é dada por $s^2 = QMRes$ e intervalos de confiança e de previsão podem ser obtidos. Resta, entretanto, estimar os polinômios.

7.2 Construção dos polinômios

Ao se estabelecerem os k polinômios em 7.2 ficam definidos $\frac{k(k+3)}{2}$ parâmetros e a condição de ortogonalidade impõe $\frac{k(k+1)}{2}$ restrições, isto é, define um sistema de $\frac{k(k+1)}{2}$ equações. Logo, esse sistema é inconsistente e há necessidade de outras k restrições. O que normalmente se faz, por facilidade de cálculo, é supor que $\gamma_{jj} = 1$ e, portanto, os polinômios usados em (7.2) ficam:

$$\begin{array}{ll} 1^{\text{o}}\text{grau}: & P_{1i} = \gamma_{10} + x_i \\ 2^{\text{o}}\text{grau}: & P_{2i} = \gamma_{20} + \gamma_{21}x_i + x_i^2 \\ 3^{\text{o}}\text{grau}: & P_{3i} = \gamma_{30} + \gamma_{31}x_i + \gamma_{32}x_i^2 + x_i^3 \\ & \cdots \\ k^{\text{o}}\text{grau}: & P_{ki} = \gamma_{k0} + \gamma_{k1}x_i + \gamma_{k2}x_i^2 + \cdots + x_i^k \end{array}$$

Será desenvolvida, a seguir, a teoria para o cálculo de polinômios ortogonais levando em consideração que os níveis de X são equidistantes, teoria esta que pode ser generalizada para níveis não equidistantes. Considerando-se os polinômios dados em (7.2), e lembrando que com variáveis centradas e os níveis de X equidistantes, tem-se que $\sum_i x_i^{2j-1} = 0$ $(j = 1, \ldots, k)$, ou seja, $\sum_i x_i = \sum_i x_i^3 = \sum_i x_i^5 = \cdots = 0$. Além disso, o fato de se ter $\sum_i P_{ji} = 0$ levará sempre à obtenção de contrastes.

Determinação de P_{1i}

Usando $\sum_i P_{1i} = 0$, tem-se

$$\sum_i P_{1i} = \sum_i (\gamma_{10} + x_i) = n\gamma_{10} + \sum_i x_i = n\gamma_{10} = 0 \quad \Rightarrow \quad \gamma_{10} = 0.$$

Portanto,

$$P_{1i} = x_i.$$

Determinação de P_{2i}

$$\begin{aligned}\sum_i P_{2i} &= \sum_i(\gamma_{20}+\gamma_{21}x_i+x_i^2)\\ &= n\gamma_{20}+\gamma_{21}\sum_i x_i+\sum_i x_i^2\\ &= n\gamma_{20}+\sum_i x_i^2=0 \quad\Rightarrow\quad \gamma_{20}=-\frac{\sum_i x_i^2}{n}.\end{aligned}$$

e

$$\begin{aligned}\sum_i P_{1i}P_{2i} &= \sum_i x_i(\gamma_{20}+\gamma_{21}x_i+x_i^2)\\ &= \gamma_{20}\sum_i x_i+\gamma_{21}\sum_i x_i^2+\sum_i x_i^3\\ &= \gamma_{21}\sum_i x_i^2=0 \quad\Rightarrow\quad \gamma_{21}=0.\end{aligned}$$

Portanto,

$$P_{2i}=x_i^2-\frac{\sum_i x_i^2}{n}.$$

Determinação de P_{3i}

$$\sum_i P_{3i} = \sum_i(\gamma_{30}+\gamma_{31}x_i+\gamma_{32}x_i^2+x_i^3) \tag{7.3}$$

$$= n\gamma_{30}+\gamma_{31}\sum_i x_i+\gamma_{32}\sum_i x_i^2+\sum_i x_i^3=0$$

$$\Rightarrow\quad n\gamma_{30}+\gamma_{32}\sum_i x_i^2=0, \tag{7.4}$$

$$\begin{aligned}\sum_i P_{1i}P_{3i} &= \sum_i x_i(\gamma_{30}+\gamma_{31}x_i+\gamma_{32}x_i^2+x_i^3)\\ &= \gamma_{30}\sum_i x_i+\gamma_{31}\sum_i x_i^2+\gamma_{32}\sum_i x_i^3+\sum_i x_i^4\\ &= \gamma_{31}\sum_i x_i^2+\sum_i x_i^4=0 \quad\Rightarrow\quad \gamma_{31}=-\frac{\sum_i x_i^4}{\sum_i x_i^2}.\end{aligned}$$

e

$$\begin{aligned}\sum_i P_{2i}P_{3i} &= \sum_i\left(x_i^2-\frac{\sum_i x_i^2}{n}\right)(\gamma_{30}+\gamma_{31}x_i+\gamma_{32}x_i^2+x_i^3)\\ &= \gamma_{30}\sum_i x_i^2+\gamma_{31}\sum_i x_i^3+\gamma_{32}\sum_i x_i^4+\sum_i x_i^5\\ &\quad -\gamma_{30}\sum_i x_i^2-\frac{\gamma_{31}}{n}\sum_i x_i^2\sum_i x_i-\frac{\gamma_{32}}{n}(\sum_i x_i^2)^2-\frac{1}{n}\sum_i x_i^3\sum_i x_i^2\\ &= \gamma_{32}\sum_i x_i^4-\frac{\gamma_{32}}{n}(\sum_i x_i^2)^2=\gamma_{32}\left[\sum_i x_i^4-\frac{(\sum_i x_i^2)^2}{n}\right]=0 \quad\Rightarrow\quad \gamma_{32}=0. \qquad (7.5)\end{aligned}$$

Substituindo (7.5) em (7.4), tem-se $\gamma_{30}=0$ e, portanto,

$$P_{3i}=x_i^3-\frac{\sum_i x_i^4}{\sum_i x_i^2}x_i.$$

A determinação de P_{ji} para $j=4,\ldots,k$ pode ser obtida de modo análogo aos demais. Pode-se adotar, alternativamente, o processo apresentado por Nogueira (1978), que serve tanto para níveis equidistantes quanto para não equidistantes.

Considerando-se $x_i=\dfrac{X_i-\bar{X}}{q}$, verifica-se que x_i são os termos de uma progressão aritmética e que se $\dfrac{X_i}{q}$ tem correspondência biunívoca com os números naturais, então, pode-se escrever

$$x_i=\frac{X_i-\bar{X}}{q}=i-\frac{n+1}{2}, \quad i=1,\ 2,\ \ldots,\ n$$

e, então,

$$\sum_{i=1}^n x_i^2=\sum_{i=1}^n(i-\frac{n+1}{2})^2=\sum_{i=1}^n i^2-(n-1)\sum_{i=1}^n i+\frac{n(n+1)^2}{4}\sum_{i=1}^n x_i^2=\frac{n(n^2-1)}{12}$$

pois, $\sum_{i=1}^n i=\dfrac{n(n+1)}{2}$ e $\sum_{i=1}^n i^2=\dfrac{n(n+1)(2n+1)}{6}$. De forma semelhante, outras somas podem ser obtidas e, assim, encontrar as expressões para os polinômios ortogonais, com níveis equidistantes,

$x_i = \dfrac{X_i - \bar{X}}{q}$, para $\bar{X} = \dfrac{\sum_i X_i}{n}$ e q a distância entre os níveis, como mostradas em Pimentel Gomes (2000):

$$\begin{aligned}
P_{1i} &= x_i \\
P_{2i} &= x_i^2 - \frac{n^2-1}{12} \\
P_{3i} &= x_i^3 - \frac{3n^2-7}{20}x_i \\
P_{4i} &= x_i^4 - \frac{3n^2-13}{14}x_i^2 + \frac{3(n^2-1)(n^2-9)}{560} \\
P_{5i} &= x_i^5 - \frac{5(n^2-7)}{18}x_i^3 + \frac{15n^4-230n^2+407)}{1008}x_i
\end{aligned} \tag{7.6}$$

7.3 Análise de Variância

Usando-se os resultados do capítulo 3, tem-se que

$$\boldsymbol{Y'Y} = \hat{\boldsymbol{\theta}}'\boldsymbol{X'Y} + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}},$$

ou seja,

$$\text{SQTotal} = \text{SQParâmetros} + \text{SQResíduo} = \text{SQP} + \text{SQRes}.$$

No caso em questão,

$$\begin{aligned}
\hat{\boldsymbol{\theta}}'\boldsymbol{X'Y} &= \begin{bmatrix} \frac{1}{n}\sum_i Y_i & \frac{\sum_i P_{1i}Y_i}{\sum_i P_{1i}^2} & \frac{\sum_i P_{2i}Y_i}{\sum_i P_{2i}^2} & \cdots & \frac{\sum_i P_{ki}Y_i}{\sum_i P_{ki}^2} \end{bmatrix} \begin{bmatrix} \sum_i Y_i \\ \sum_i P_{1i}Y_i \\ \sum_i P_{2i}Y_i \\ \cdots \\ \sum_i P_{ki}Y_i \end{bmatrix} \\
&= \frac{1}{n}(\sum_i Y_i)^2 + \frac{(\sum_i P_{1i}Y_i)^2}{\sum_i P_{1i}^2} + \frac{(\sum_i P_{2i}Y_i)^2}{\sum_i P_{2i}^2} + \cdots + \frac{(\sum_i P_{ki}Y_i)^2}{\sum_i P_{ki}^2},
\end{aligned}$$

ou seja,

$$\text{SQParâmetros} = \text{Correção} + \text{SQ Regressão}.$$

Assim, a soma de quadrados da regressão de grau j é dada por

$$\text{SQReg.(grau } j) = \frac{(\sum_i P_{ji}Y_i)^2}{\sum_i P_{ji}^2}$$

e o quadro da análise de variância fica

Causas de variação	GL	SQ
Regressão linear	1	$(\sum_i P_{1i}Y_i)^2/\sum_i P_{1i}^2$
Regressão quadrática	1	$(\sum_i P_{2i}Y_i)^2/\sum_i P_{2i}^2$
Regressão cúbica	1	$(\sum_i P_{3i}Y_i)^2/\sum_i P_{3i}^2$
$\cdots$	$\cdots$	$\cdots$
Regressão de grau k	1	$(\sum_i P_{ki}Y_i)^2/\sum_i P_{ki}^2$
Regressão	k	$\hat{\boldsymbol{\theta}}'\boldsymbol{X}'\boldsymbol{Y}-$Correção
Resíduo	$n-k-1$	$\boldsymbol{Y}'\boldsymbol{Y}-\hat{\boldsymbol{\theta}}'\boldsymbol{X}'\boldsymbol{Y}$
Total corrigido	$n-1$	$\boldsymbol{Y}'\boldsymbol{Y}-$Correção

sendo Correção$=\dfrac{(\sum_i Y_i)^2}{n}$.

7.4 Dados com repetição

Sejam n pares de observações, (X_i, Y_{iu}), $i = 1, 2, \cdots, t$ e $u = 1, 2, \cdots, r_i$, isto é, cada nível de X_i é repetido r_i vezes, aos quais se ajusta um modelo de regressão polinomial de grau k do tipo

$$Y_{iu} = \alpha_0 + \alpha_1 P_{1i} + \alpha_2 P_{2i} + \cdots + \alpha_k P_{ki} + \varepsilon_{iu}.$$

Do capítulo 3 tem-se que o sistema de equações normais é dado por:

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{X}'\boldsymbol{Y}$$

sendo $\boldsymbol{\theta}$ estimado por:

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

com matriz de variâncias e covariâncias dada por

$$V(\hat{\boldsymbol{\theta}}) = V[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'V(\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2,$$

sendo

$$\boldsymbol{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ \cdots \\ Y_{1r_1} \\ \cdots \\ Y_{t1} \\ Y_{t2} \\ \cdots \\ Y_{tr_t} \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & P_{11} & P_{21} & \cdots & P_{k1} \\ 1 & P_{11} & P_{21} & \cdots & P_{k1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & P_{11} & P_{21} & \cdots & P_{k1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & P_{1t} & P_{2t} & \cdots & P_{kt} \\ 1 & P_{1t} & P_{2t} & \cdots & P_{kt} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & P_{1t} & P_{2t} & \cdots & P_{kt} \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \cdots \\ \alpha_k \end{bmatrix} \quad \text{e} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \cdots \\ \varepsilon_{1r_1} \\ \cdots \\ \varepsilon_{t1} \\ \varepsilon_{t2} \\ \cdots \\ \varepsilon_{tr_t} \end{bmatrix}$$

Nesse caso, o procedimento de obtenção dos polinômios é bastante semelhante. Dado que $\sum_i r_i P_{ji} = 0$ e $\sum_i r_i P_{ji} P_{j'i} = 0$, $j \neq j'$, então,

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} n = \sum_i r_i & 0 & 0 & \cdots & 0 \\ 0 & \sum_i r_i P_{1i}^2 & 0 & \cdots & 0 \\ 0 & 0 & \sum_i r_i P_{2i}^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \sum_i r_i P_{ki}^2 \end{bmatrix}, \boldsymbol{X}'\boldsymbol{Y} = \begin{bmatrix} \sum_i \sum_u Y_{iu} \\ \sum_i P_{1i} Y_{i\cdot} \\ \sum_i P_{2i} Y_{i\cdot} \\ \cdots \\ \sum_i P_{ki} Y_{i\cdot} \end{bmatrix},$$

e, portanto, tem-se que

$$\hat{\alpha_0} = \bar{Y} \qquad \text{e} \quad \hat{\alpha_j} = \frac{\sum_i P_{ji} Y_{i.}}{\sum_i r_i P_{ji}^2}, \quad j = 1, \ldots, k,$$

sendo $Y_{i.} = \sum_u Y_{iu}$. Logo,

$$\text{SQReg.(grau } j) = \frac{(\sum_i P_{ji} Y_{i.})^2}{\sum_i r_i P_{ji}^2}$$

Quadro Comparativo

	Dados pareados	Repetições iguais	Repetições diferentes
P_{1i}	x_i	x_i	x_i
P_{2i}	$x_i^2 - \frac{\sum_i x_i^2}{n}$	$x_i^2 - r\frac{\sum_i x_i^2}{n}$	$x_i^2 - \frac{\sum_i r_i x_i^2}{n}$
P_{3i}	$x_i^3 - \frac{\sum_i x_i^4}{\sum_i x_i^2} x_i$	$x_i^3 - \frac{\sum_i x_i^4}{\sum_i x_i^2} x_i$	$x_i^3 - \frac{\sum_i r_i x_i^4}{\sum_i r_i x_i^2} x_i$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\hat{\alpha_0}$	$\bar{Y}$	$\bar{Y}$	$\bar{Y}$
$\hat{\alpha_j}$	$\frac{\sum_i P_{ji} Y_i}{\sum_i P_{ji}^2}$	$\frac{\sum_i P_{ji} Y_{i.}}{r \sum_i P_{ji}^2}$	$\frac{\sum_i P_{ji} Y_{i.}}{\sum_i r_i P_{ji}^2}$

7.5 Dados não equidistantes

Procedimento análogo ao anterior, exceto pelo fato de que agora $\sum_i x_i^{2j-1} \neq 0$, resultando em cálculos mais complicados. Alternativa mais simples pode ser obtida usando Nogueira (1978).

7.6 Equivalência das fórmulas obtidas e as usadas por Pimentel Gomes (2000)

Foi visto que:

$$\hat{\alpha_0} = \bar{Y} \qquad \text{e} \quad \hat{\alpha_j} = \frac{\sum_i P_{ji} Y_{i.}}{r \sum_i P_{ji}^2} = \frac{\sum_i P_{ji} T_i}{r \sum_i P_{ji}^2}, \quad j = 1, \ldots, k.$$

em que $T_i = Y_{i.}$ é o total do tratamento i. Considerando M_j o valor que torna os polinômios inteiros, tem-se que:

$$\hat{\alpha_j} = \frac{\sum_i P_{ji} T_i}{r \sum_i P_{ji}^2} = \frac{M_j^2 \sum_i P_{ji} T_i}{M_j^2 r \sum_i P_{ji}^2} = M_j \frac{\sum_i M_j P_{ji} T_i}{r \sum_i (M_j P_{ji})^2} = M_j \frac{\sum_i C_{ji} T_i}{r \sum_i C_{ji}^2} = M_j \frac{\sum_i C_{ji} T_i}{r K_j} = M_j B_j,$$

sendo $C_{ji} = M_j P_{ji}$, $K_j = \sum_i C_{ji}^2$ e $B_j = \frac{\sum_i C_{ji} T_i}{r K_j}$, valores tabelados em Pimentel Gomes (2000) para diferentes números de níveis.

As somas de quadrados ficam:

$$\text{SQReg.(grau } j) = \frac{(\sum_i P_{ji}T_i)^2}{r\sum_i P_{ji}^2} = \frac{M_j^2(\sum_i P_{ji}T_i)^2}{M_j^2 r\sum_i P_{ji}^2} = \frac{(\sum_i C_{ji}T_i)^2}{rK_j}$$

e o modelo de regressão estimado

$$\hat{Y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 P_{1i} + \hat{\alpha}_2 P_{2i} + \cdots + \hat{\alpha}_k P_{ki}$$

transforma-se em:

$$\hat{Y}_i = \bar{Y} + B_1 M_1 P_{1i} + B_2 M_2 P_{2i} + \cdots + B_k M_k P_{ki}.$$

7.7 Exemplo

Os dados da Tabela 7.1 referem-se a produções de milho (Y), em kg/parcela, de um experimento casualizado em blocos de adubação de milho com diferentes doses (X) de P_2O_5.

Tabela 7.1: Produção de milho em kg/parcela, de um experimento de adubação de milho

Blocos	0	25	50	75	100	Totais
I	3,38	7,15	10,07	9,55	9,14	39,29
II	5,77	9,78	9,73	8,95	10,17	44,40
III	4,90	9,99	7,92	10,24	9,75	42,80
IV	4,54	10,10	9,48	8,66	9,50	42,28
Totais	18,59	37,02	37,20	37,40	38,56	168,77

Fonte: Pimentel Gomes (2000).

Considerando-se Doses de P_2O_5 como variável qualitativa têm-se os resultados apresentados na parte de baixo da Tabela 7.2, enquanto que na parte de cima é feito o desdobramento do número de graus de liberdade, usando-se Doses como variável quantitativa e polinômios ortogonais. Vê-se que um polinômio de grau 3 explica o comportamento das médias de tratamentos. A equação estimada é dada por:

$$\hat{Y}_i = 4,712 + 0,276X - 0,00483X^2 + 0,0000256X^3.$$

A Figura 7.1 apresenta os valores observados e a curva ajustada. Um modelo não linear talvez desse uma melhor explicação do ponto de vista prático.

O programa em R usado para os cálculos foi:

```
milho<-read.table("a:milho.dat", header=TRUE)
attach(milho)

dose<-rep(seq(0,100,25),4)
```

Tabela 7.2: Esquema de análise de variância

Causas de variação	G.L.	S.Q.	Q.M.	F
Reg. Linear	1	$40,64$	$40,64$	$44,66**$
Reg. Quadrática	1	$21,28$	$21,28$	$23,38**$
Reg. de grau 3	1	$9,23$	$9,23$	$10,14**$
Reg. de grau 4	1	$1,07$	$1,07$	$1,18$ ns
(Doses)	(4)	$(72,22)$	$18,06$	$19,84**$
Blocos	3	$2,73$		
Resíduo	12	$10,92$	$0,91$	
Total	19	$85,87$		

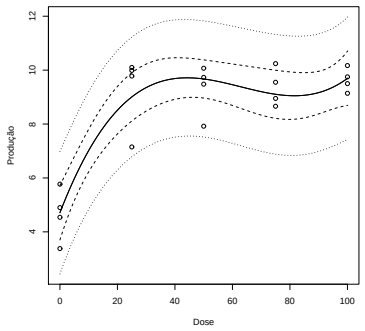


Figura 7.1: Valores observados de produção de milho e curva ajustada.

```
bloco<-factor(gl(4,5))

## Modelos ajustados ##
mod1<-lm(Y~bloco+as.factor(dose))
anova(mod1)
mod2<-lm(Y~bloco+poly(dose,4))
anova(mod2)
mod2<-lm(Y~bloco-1+poly(dose,4))
summary(mod2)
mod3<-lm(Y~poly(dose,3))
dose2<-dose*dose
dose3<-dose2*dose
mod4<-lm(Y~dose+dose2+dose3)
summary(mod4)

#Gráfico com curva ajustada e valores observados
plot(c(0,100), c(0,12), type="n", xlab="dose", ylab="produção")
```

```
points(X,Y)
d<-seq(0,100,1)
lp<-predict(mod3,data.frame(X=d))
lines(d,lp,lty=1)
title(sub="Figura 1. Curva ajustada e valores observados")

#Gráfico com intervalos
par(mfrow=c(1,1)) # um unico grafico
novoX<-data.frame(X=seq(min(X), max(X), by=((max(X)-min(X))/100)))
pred.w.clim<-predict(mod3, novoX, level=0.95, interval=c("confidence"))
pred.w.plim<-predict(mod3, novoX, level=0.95, interval=c("prediction"))
matplot(novoX, cbind(pred.w.clim, pred.w.plim[,-1]), lty=c(1,2,2,3,3),
col=c(1,1,1,1,1), type="l", xlab="Dose", ylab="Produção")
points(X,Y)
```

Os dados que se seguem referem-se aos totais de produtividade de cana-de-açúcar, em ton/ha, obtidos de um experimento casualizado em blocos com 6 repetições e 5 níveis de P_2O_5, em kg/ha.

Níveis de P_2O_5	Totais de produtividade
0	445,80
50	482,70
100	508,32
150	489,78
200	463,50

Fonte: Campos, H. (1984). pág. 251.

Quadro da análise de variância.

Causas de variação	GL	SQ	QM	F
Blocos	5	735,6497		
Níveis de P	4	387,8748	96,9687	10,92(**)
Resíduo	20	177,6581		
Total corrigido	29			

7.8 Exercícios

1. Considere o modelo de regressão

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$$

e que os valores de Y são observados para $X = 1, 2, 3, 4, 5$. Pede-se:

   (a) Obter a matriz do delineamento.

   (b) Obter os polinômios ortogonais até 3º grau por meio da ortogonalização de Gram Schmidt das colunas da matriz do delineamento.

(c) Verificar que os polinômios ortogonais fornecidos por meio da função poly() do R podem ser obtidos por meio da normalização das colunas (exceto a primeira) da matriz obtida no item anterior.

2. Repita o exercício anterior, considerando-se que há

   (a) Duas repetições para cada nível de X

   (b) Duas repetições para $X = 1$ e 2, e uma para $X = 3, 4, 5$.

   Os polinômios ortogonais são os mesmos?

3. Em um experimento de adubação em eucalipto (*Eucalyptus grandis*) conduzido em casa de vegetação, foram usadas 4 doses de K (0, 30, 60 e 90 ppm), obtendo-se as alturas, em cm, apresentadas na tabela 12.1. Considerando-se que o experimento foi conduzido segundo o delineamento inteiramente ao acaso com 3 repetições, pede-se:

   (a) Obter os polinômios ortogonais até 3º grau.

   (b) Verificar o ajuste de um polinômio de 3º grau a esses dados.

   (c) Obter a equação de regressão polinomial mais adequada.

Tabela 12.1. Dose de K, em ppm, e altura, em cm, de *Eucalyptus grandis* envasados.

Dose de K	Altura		
0	80	86	71
30	144	151	97
60	151	127	117
90	70	85	92

Fonte: Cicolim, R.A. e J.M. Gonçalves (1992).

4. Repetir o exercício anterior considerando-se que a terceira observação relativa à dose 30 ppm tenha sido perdida.

Programas

```
############### # Exercício 1 # ###############

# No R # x<-c(1,2,3,4,5) X<-cbind(1,x,x^2,x^3) cbind(1,poly(x,3))

# No Maple V # with(linalg): u1:=vector([1,1,1,1,1]);
u2:=vector([1,2,3,4,5]); u3:=vector([1,4,9,16,25]);
u4:=vector([1,8,27,64,125]); GS:=GramSchmidt([u1,u2,u3,u4]);
normalize(GS[2]);evalf(%);
normalize(GS[3]);evalf(%);
normalize(GS[4]);evalf(%);


############### # Exercício 2 # ###############
x1<-c(1,1,2,2,3,3,4,4,5,5) X1<-cbind(1,x1,x1^2,x1^3)
cbind(1,poly(x1,3))

x1<-c(1,1,2,2,3,4,5) X1<-cbind(1,x1,x1^2,x1^3) cbind(1,poly(x1,3))

############### # Exercício 3 # ###############
K<-c(0,0,0,30,30,30,60,60,60,90,90,90) Altura<-c(80,86,71,
144,151,97, 151,127,117, 70,85,92) M1<-lm(Altura~K+I(K^2)+I(K^3))

anova(M1) # Análise seqüencial

      Analysis of Variance Table

      Response: Altura
                Df    Sum Sq   Mean Sq   F value     Pr(>F)
      K          1      18.2      18.2    0.0537   0.822502
      I(K^2)     1    7650.8    7650.8   22.6521   0.001428 **
      I(K^3)     1 1.667e-02 1.667e-02 4.935e-05   0.994567
      Residuals  8    2702.0     337.8
      ---
      Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

M2<-lm(Altura~poly(K,3)) # Análise utilizando polinômios
ortogonais summary(M2) # Corresponde à análise seqüencial

      Call:
      lm(formula = Altura ~ poly(K, 3))

      Coefficients:
                  Estimate Std. Error t value   Pr(>|t|)
      (Intercept) 105.9167     5.3053  19.964   4.13e-08 ***
      poly(K, 3)1   4.2603    18.3780   0.232   0.82250
      poly(K, 3)2 -87.4686    18.3780  -4.759   0.00143 ** ==> Modelo quadrático
      poly(K, 3)3   0.1291    18.3780   0.007   0.99457    ==> No caso, corresponde ao
      ---                                                      teste da falta de ajuste
      Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

      Residual standard error: 18.38 on 8 degrees of freedom
      Multiple R-Squared: 0.7395,     Adjusted R-squared: 0.6418
      F-statistic: 7.569 on 3 and 8 DF,  p-value: 0.01008

lm(Altura~K+I(K^2)) # Ajuste do modelo quadrático
```

```
      Coefficients:
      (Intercept)            K       I(K^2)
         79.01667      2.56167     -0.02806

plot(K,Altura)
curve(79.01667+2.56167*x-0.02806*x^2,x=c(0,90),add=TRUE)

################ #  Exercício 4 # #################
K<-c(0,0,0,30,30,60,60,60,90,90,90) #  Dados sem a terceira
observação
                                    #  relativa à dose X=30
Altura<-c(80,86,71, 144,151, 151,127,117, 70,85,92)
M1<-lm(Altura~K+I(K^2)+I(K^3)) anova(M1) # Análise seqüencial

      Analysis of Variance Table

      Response: Altura
                Df Sum Sq Mean Sq F value   Pr(>F)
      K          1    9.2     9.2  0.0642   0.8072
      I(K^2)     1 8956.7  8956.7 62.5825 9.79e-05  ***
      I(K^3)     1  316.4   316.4  2.2108   0.1806
      Residuals  7 1001.8   143.1
      ---
      Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

M2<-lm(Altura~poly(K,3)) # Análise utilizando polinômios
ortogonais summary(M2) # Corresponde à análise seqüencial

      Call:
      lm(formula = Altura ~ poly(K, 3))

      Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
      (Intercept)  106.727      3.607  29.589 1.30e-08 ***
      poly(K, 3)1    3.032     11.963   0.253    0.807
      poly(K, 3)2  -94.640     11.963  -7.911 9.79e-05 *** ==> Modelo quadrático
      poly(K, 3)3   17.788     11.963   1.487    0.181
      ---
      Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

      Residual standard error: 11.96 on 7 degrees of freedom
      Multiple R-Squared: 0.9026,     Adjusted R-squared: 0.8608
      F-statistic: 21.62 on 3 and 7 DF,  p-value: 0.0006458

lm(Altura~K+I(K^2)) # Ajuste do modelo quadrático

      Call:
      lm(formula = Altura ~ K + I(K^2))

      Coefficients:
      (Intercept)            K       I(K^2)
         81.07483      2.85896     -0.03187

plot(K,Altura)
curve(81.07483+2.85896*x-0.03187*x^2,x=c(0,90),add=TRUE)
```