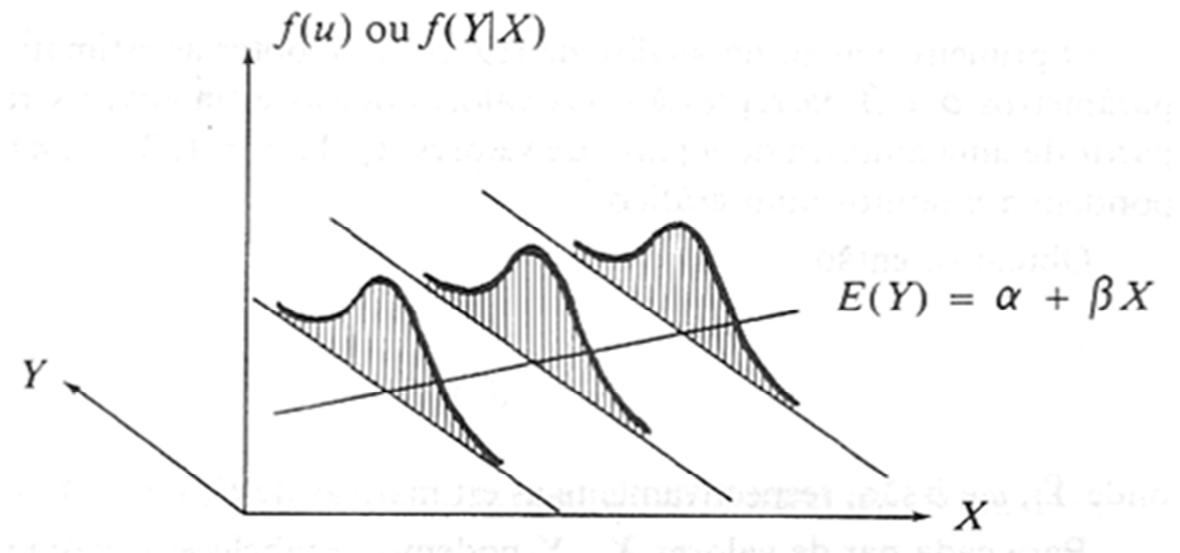


# **RODOLFO HOFFMANN**



# **ANÁLISE DE REGRESSÃO**

## **Uma Introdução à Econometria**

**Portal de Livros Abertos  
da USP  
2016**



**RODOLFO HOFFMANN**

**ANÁLISE DE REGRESSÃO**  
**Uma Introdução à**  
**Econometria**

Esta é uma nova versão do livro de mesmo título (quarta edição) publicado pela Editora HUCITEC em 2006, com edição esgotada em 2014.

**Piracicaba**  
**Edição do Autor**  
**2016**

DOI: 10.11606/9788592105709

**Dados Internacionais de Catalogação na Publicação**  
**DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP**

Hoffmann, Rodolfo  
Análise de regressão: uma introdução à econometria [recurso eletrônico] / Rodolfo Hoffmann. - - 5. ed. Piracicaba: O Autor, 2016.  
393 p. : il.

ISBN: 978-85-921057-0-9

1. Análise de regressão 2. Econometria I.Título

CDD 330.18  
H711a

DOI: 10.11606/9788592105709

Autorizo a reprodução parcial ou total desta obra, para fins acadêmicos,  
desde que citada a fonte

## SUMÁRIO

1. INTRODUÇÃO E CONCEITOS ESTATÍSTICOS BÁSICOS .....	1
1.1. Econometria e análise de regressão .....	1
1.2. Modelo matemático e modelo estatístico .....	1
1.3. Variável aleatória .....	4
1.4. Esperança matemática .....	5
1.5. Variância e covariância .....	5
1.6. Estimador não-tendencioso .....	10
1.7. Estimador de variância mínima .....	15
1.8. Estimadores de mínimos quadrados .....	19
1.9. Estimadores de máxima verossimilhança .....	21
1.10. Propriedades assintóticas dos estimadores .....	24
1.11. O limite inferior de Cramér-Rao e as propriedades assintóticas dos estimadores de máxima verossimilhança .....	32
1.12. Teste de hipóteses .....	34
Exercícios .....	40
 2. REGRESSÃO LINEAR SIMPLES .....	44
2.1. modelo estatístico de uma regressão linear simples .....	44
2.2. Estimativa dos parâmetros.....	47
2.3. O modelo simplificado e um exemplo numérico .....	50
2.4. Demonstração de que os estimadores de mínimos quadrados são estimadores lineares não-tendenciosos .....	53
2.5. Variâncias e covariâncias das estimativas dos parâmetros .....	55
2.6. Demonstração de que $b$ é um estimador linear não-tendencioso de variância mínima .....	58
2.7. Decomposição da soma de quadrados total .....	61
2.8. Esperanças das somas de quadrados .....	63
2.9. Análise de variância da regressão .....	65
2.10. O coeficiente de determinação corrigido para graus de liberdade e o coeficiente de variação .....	68

2.11. Estimativas das variâncias das estimativas dos parâmetros, teste de hipóteses a respeito dos parâmetros e respectivos intervalos de confiança .....	69
2.12. Variância de $\hat{Y}_i$ e intervalo de previsão .....	72
2.13. O problema da especificação e as funções que se tornam lineares por anamorfose .....	77
2.14. Estimativa de máxima verossimilhança .....	80
2.15. Análise de regressão quando $X$ é uma variável aleatória .....	81
Exercícios .....	82
 3. CORRELAÇÃO .....	103
3.1. O coeficiente de correlação simples para uma amostra .....	103
3.2. Aplicação da análise de regressão a uma população com distribuição normal bidimensional .....	110
Exercícios .....	112
 4. REGRESSÃO LINEAR MÚLTIPLA.....	120
4.1. O modelo estatístico de uma regressão linear múltipla .....	120
4.2. Estimativas dos parâmetros de acordo com o método dos mínimos quadrados .....	121
4.3. Variâncias e covariâncias das estimativas dos parâmetros.....	124
4.4. Variância de uma combinação linear das estimativas dos parâmetros .....	125
4.5. Análise de variância da regressão linear múltipla .....	126
4.6. Demonstração de que $b$ é um estimador linear não-tendencioso de variância mínima .....	130
4.7. O uso das variáveis centradas .....	132
4.8. Exemplo de uma regressão linear múltipla com duas variáveis explanatórias .....	135
4.9. Previsão e teste de hipóteses a respeito do valor de combinações lineares dos parâmetros .....	139
4.10. Interpretação dos coeficientes de regressão de uma regressão linear múltipla com duas variáveis explanatórias .....	143
4.11. Os coeficientes de correlação parcial .....	146
4.12. Intervalos de confiança e regiões de confiança para os parâmetros.....	154
4.13. Exemplo de regressão linear múltipla com três variáveis explanatórias .....	162

4.14. Problemas de especificação.....	168
4.15. Transformação das variáveis para obter a matriz de correlações simples .....	171
4.16. Regressões que se tornam lineares por anamorfose .....	173
4.17. Ortogonalidade e multicolinearidade na matriz <b>X</b> .....	173
4.18. Teste de hipóteses no modelo linear .....	178
4.19. Interpretação geométrica da análise de regressão linear de acordo com o método de mínimos quadrados .....	181
Exercícios .....	194
<b>5. USO DE VARIÁVEIS BINÁRIAS .....</b>	<b>219</b>
5.1. Níveis de medida .....	219
5.2. Uso de variáveis binárias para distinguir as categorias de uma variável nominal.....	220
5.3. Uso de variáveis binárias para ajustar poligonais .....	226
5.4. Mudança estrutural .....	230
5.5. Análise de variância de dados com vários tratamentos e o teste para "falta de ajustamento" .....	236
Exercícios .....	240
<b>6. HETEROCEDASTICIA .....</b>	<b>254</b>
6.1. O caso de uma regressão linear simples em que o desvio padrão do erro é proporcional a <i>X</i> .....	254
6.2. O método dos mínimos quadrados ponderados .....	255
6.3. Conseqüências do uso de estimadores de mínimos quadrados ordinários quando existe heterocedasticidade .....	257
6.4. Testes para a homocedasticidade e obtenção de estimativas dos parâmetros quando a matriz <b>V</b> é desconhecida .....	261
6.5. O estimador de White para variância quando há heterocedasticidade .....	267
Exercícios .....	268
<b>7. MÍNIMOS QUADRADOS GENERALIZADOS E AUTOCORRELAÇÃO NOS RESÍDUOS .....</b>	<b>275</b>
7.1. Mínimos quadrados generalizados .....	275
7.2. Autocorrelação nos resíduos .....	278
7.3. O teste de Durbin-Watson .....	283
Exercícios .....	285

8. VARIÁVEIS INSTRUMENTAIS E ERROS NAS VARIÁVEIS EXPLANATÓRIAS .....	291
8.1. Introdução .....	291
8.2. A consistência dos estimadores de mínimos quadrados ordinários .....	291
8.3. A inconsistência dos estimadores de mínimos quadrados quando os erros estão assintoticamente correlacionados com uma ou mais das variáveis explanatórias .....	294
8.4. O uso de variáveis instrumentais para obter estimativas consistentes .....	295
8.5. Regressão linear simples com as duas variáveis sujeitas a erros de medida .....	298
8.6. O método da variável instrumental .....	301
8.7. Outro método .....	303
Exercícios .....	305
9. EQUAÇÕES SIMULTÂNEAS .....	308
9.1. Introdução .....	308
9.2. Um exemplo numérico .....	311
9.3. O estimador de variável instrumental .....	312
9.4. Mínimos quadrados indiretos .....	312
9.5. Mínimos quadrados em dois estágios .....	315
9.6. Variáveis conjuntamente determinadas e variáveis predeterminadas .....	317
9.7. Notação geral .....	318
9.8. Variáveis instrumentais .....	319
9.9. Identificação .....	321
9.10. Estimação dos parâmetros em caso de superidentificação .....	327
9.11. Outras maneiras de obter o estimador de mínimos quadrados em dois estágios .....	328
9.12. Um exemplo numérico .....	329
9.13. Um segundo exemplo numérico .....	333
9.14. Terceiro exemplo .....	334
9.15. Uma visão global .....	340
Exercícios .....	342
10. SÉRIES TEMPORAIS .....	352
10.1. Processos estocásticos .....	352
10.2. Ruído branco .....	354
10.3. Modelos de regressão .....	355

10.4. Modelos de decomposição .....	355
10.5. Modelos ARMA .....	355
10.6. Análise do AR(1) .....	357
10.7. O passeio aleatório com deslocamento .....	358
10.8. Transformando modelos AR em modelos MA e vice-versa .....	362
10.9. Raiz unitária e modelos ARIMA .....	364
10.10. Função de autocorrelação .....	365
10.11. Os testes de Dickey-Fuller .....	367
10.12. Modelo de correção de erro e co-integração .....	368
Exercícios .....	373
APÊNDICE .....	376
BIBLIOGRAFIA .....	383
ÍNDICE ANALÍTICO.....	387

## PREFÁCIO

Este livro reflete o esforço do autor em preparar material didático para disciplinas de econometria e análise de regressão ministradas na ESALQ-USP e, a partir de 1997, no Instituto de Economia da UNICAMP.

O interesse na aprendizagem desses métodos estatísticos se deve, em grande parte, ao uso que deles se faz em pesquisas econômicas. Mas a análise de regressão também é largamente aplicada em outras áreas, como biologia, física ou engenharia. Não é exagero afirmar que muitas vezes a condução e a avaliação de uma pesquisa dependem do conhecimento do pesquisador sobre econometria e análise de regressão, inclusive no que tange a suas potencialidades e a suas limitações.

Um aspecto didaticamente importante, neste livro, é a apresentação de exercícios numéricos que não exigem, para serem resolvidos, nem mesmo uma máquina de calcular. Dessa maneira o aluno pode, sem dispensar muito tempo em cálculo, testar sua aprendizagem e usar os conhecimentos recém-adquiridos. Aliás, a idéia de minimizar cálculos não é nova. Basta lembrarmos de que, quando aprendemos a resolver equações do 2º grau, trabalhamos com exercícios do tipo

$$2x^2 - 7x + 3 = 0$$

e não do tipo

$$-0,072150x^2 + 1,481099x - 470,1902 = 0$$

Não há dúvida, entretanto, que técnicas mais avançadas e recentes exigem o uso do computador. O próprio desenvolvimento dos métodos estatísticos nas últimas décadas está muito associado ao uso do computador como poderoso instrumento de fazer cálculos.

Nesta quarta edição foi acrescentado um capítulo sobre séries temporais. Também foram incorporados novos exercícios e novas seções em capítulos anteriores, sempre procurando melhorar a apresentação dos temas, deixando para um outro volume a análise de regressão não-linear e modelos de lógite e próbita.

Seria difícil listar todos os colegas e alunos que, com suas críticas e sugestões muito contribuíram para que versões anteriores deste livro fossem sucessivamente melhoradas. A Profa. Sonia Vieira foi co-autora das edições anteriores. A Profa. Angela A. Kageyama fez cuidadosa revisão da 1ª edição. A Profa. Rosângela Ballini fez várias

sugestões e correções nesta 4<sup>a</sup> edição. E a tarefa de digitar todo o texto novamente foi realizada com muita competência e cuidado por Joselene Rodrigues da Silva.

Cabe, finalmente, registrar as boas condições de trabalho fornecidas pelas instituições onde trabalhei e trabalho, a ESALQ-USP e o IE-UNICAMP, e agradecer o apoio recebido da FAPESP e do CNPq.

Para esta nova edição em meio digital, de 2016, contei com a indispensável colaboração de Helena Aparecida Cardoso.

Sugestões, correções ou dúvidas podem ser enviadas para o e-mail do autor:  
[hoffmannr@usp.br](mailto:hoffmannr@usp.br).

## **1. INTRODUÇÃO E CONCEITOS ESTATÍSTICOS BÁSICOS**

### **1.1. Econometria e análise de regressão**

A econometria consiste na aplicação de métodos matemáticos e estatísticos a problemas de economia. O econometrista combina conhecimentos de três ramos científicos: Economia, Matemática e Estatística.

A análise de regressão é o método mais importante da econometria.

Sempre é interessante conhecer os efeitos que algumas variáveis exercem, ou que parecem exercer, sobre outras. Mesmo que não exista relação causal entre as variáveis podemos relacioná-las por meio de uma expressão matemática, que pode ser útil para se estimar o valor de uma das variáveis quando conhecemos os valores das outras (estas de mais fácil obtenção ou antecessoras da primeira no tempo), sob determinadas condições.

Genericamente, tais relações funcionais podem ser representadas por

$$Y = f(X_1, X_2, \dots, X_k)$$

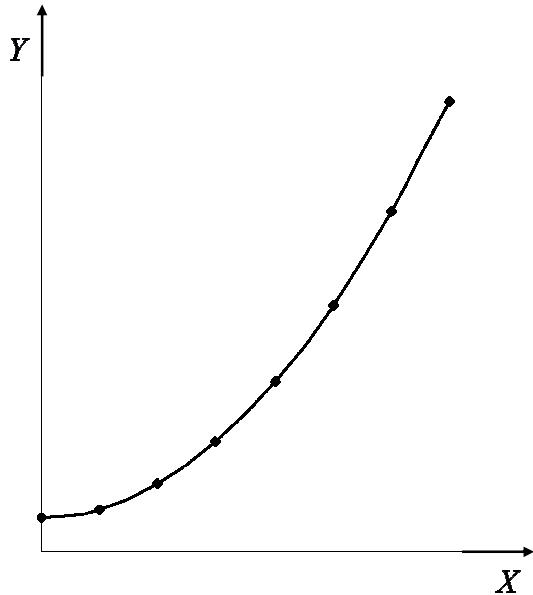
onde  $Y$  representa a variável dependente e os  $X_h$  ( $h = 1, 2, \dots, k$ ) representam as variáveis explanatórias.

São exemplos de relações funcionais entre variáveis:

- a) crescimento da população ou do PNB de um país ( $Y$ ) em função dos anos ( $X$ );
- b) variação da produção ( $Y$ ) obtida numa cultura conforme a quantidade de nitrogênio ( $X_1$ ), fósforo ( $X_2$ ) e potássio ( $X_3$ ) utilizada na adubação;
- c) variação do preço ( $Y$ ) de um produto no mercado em função da quantidade oferecida ( $X$ ).

### **1.2. Modelo matemático e modelo estatístico**

Consideremos duas variáveis,  $X$  e  $Y$ , relacionadas por uma função matemática  $Y = f(X)$ . Dado um conjunto de valores  $X_i$  ( $i = 1, 2, \dots, n$ ) e os correspondentes valores de  $Y_i = f(X_i)$ , se colocarmos os pontos  $(X_i, Y_i)$  em um gráfico verificaremos que eles pertencem à curva que representa o modelo matemático que relaciona as duas variáveis, como mostra a figura 1.1.



*Figura 1.1.* Modelo matemático:  $Y_i = f(X_i)$

É comum, entretanto, que a variável dependente seja afetada por outros fatores, além dos considerados no modelo adotado. Admitamos que a variável dependente sofra a influência de  $k + m$  variáveis, isto é,

$$Y = f(X_1, X_2, \dots, X_k, X_{k+1}, \dots, X_{k+m})$$

e que por vários motivos (não disponibilidade dos valores, impossibilidade de mensuração, para simplificar a análise etc.) não consideramos a influência das variáveis  $X_{k+1}, \dots, X_{k+m}$ . Ao analisarmos  $Y$  como função das  $k$  primeiras variáveis permanece, então, um resíduo ou erro.

Admitindo que esse erro seja aditivo, o modelo estatístico fica

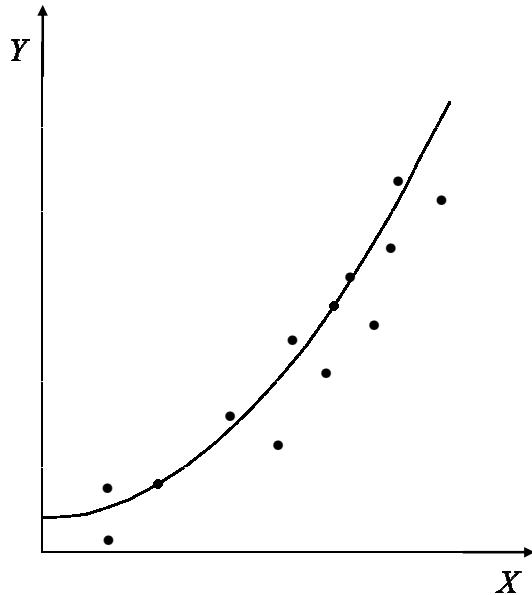
$$Y_i = f(X_{1i}, X_{2i}, \dots, X_{ki}) + u_i \quad (i = 1, \dots, n)$$

Se apenas uma das variáveis independentes é considerada, temos

$$Y_i = f(X_i) + u_i$$

Neste caso, o conjunto de pares de valores  $(X_i, Y_i)$  corresponde a um conjunto de pontos, dispersos em torno da curva representativa da função, como mostra a figura

1.2. Dizemos que as duas variáveis estão relacionadas de acordo com um modelo estatístico.



*Figura 1.2. Modelo estatístico:  $Y_i = f(X_i) + u_i$*

Outra justificativa para a existência do erro ( $u_i$ ) em um modelo estatístico é dada pelos erros de mensuração da variável dependente. Se os verdadeiros valores ( $V_i$ ) da variável dependente são uma função matemática das variáveis explanatórias, isto é,

$$V_i = f(X_{1i}, X_{2i}, \dots, X_{ki})$$

e se os valores observados ( $Y_i$ ) da variável dependente apresentam erros de mensuração ( $u_i$ ), isto é,

$$Y_i = V_i + u_i,$$

a relação entre  $Y_i$  e os  $X_{hi}$  ( $h = 1, 2, \dots, k$ ) fica

$$Y_i = f(X_{1i}, X_{2i}, \dots, X_{ki}) + u_i$$

Em casos reais geralmente existem tanto erros de mensuração como efeitos de outras variáveis. Nestes casos, o erro residual do modelo será a soma desses dois tipos de erro.

Desde que existam erros de mensuração, é lógico admitir que os valores das variáveis explanatórias também são afetados; os problemas que isso acarreta serão discutidos mais adiante; numa primeira etapa admitiremos apenas um erro residual devido à existência de fatores não incluídos no modelo e/ou erros de mensuração apenas na variável dependente.

Nas próximas seções deste capítulo faremos uma revisão de alguns conceitos básicos de estatística.<sup>1</sup>

### 1.3. Variável aleatória

Dizemos que uma variável discreta  $X$  é aleatória, se a cada um de seus valores se associa uma probabilidade  $P(X)$ . O conjunto dos valores da variável e das respectivas probabilidades é a distribuição de  $X$ .

Vejamos um exemplo. Se uma moeda é lançada 5 vezes, o número de vezes que se obtém “cara” é uma variável aleatória discreta, que pode assumir valores inteiros de 0 a 5, inclusive. Essa variável tem distribuição binomial. Demonstra-se que, se  $p$  é a probabilidade de obter “cara” em um único lançamento da moeda, a probabilidade de ocorrerem  $X = k$  caras, em 5 lançamentos da moeda, é

$$P(X = k) = \binom{5}{k} p^k (1-p)^{5-k}$$

Esta é a função de probabilidade da distribuição binomial para  $n = 5$ , onde  $n$  é o número de ensaios.

Se a variável aleatória é contínua, a probabilidade de obtermos exatamente um determinado valor  $k$  é zero, isto é:

$$P(X = k) = 0$$

---

<sup>1</sup> Um desenvolvimento mais detalhado da maioria dos temas abordados nesta revisão pode ser encontrado em HOFFMANN (1980).

Entretanto, desde que seja definida a função de densidade  $f(X)$ , podemos obter a probabilidade de a variável aleatória assumir valores no intervalo  $(a, b)$ , isto é,

$$P(a < X < b) = \int_a^b f(X) dX$$

O valor de  $f(X)$  também é denominado densidade de probabilidade.

Se a variável contínua tem distribuição normal com média  $\mu$  e variância  $\sigma^2$ , a função de densidade é

$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(X-\mu)^2}{2\sigma^2}\right\}$$

#### 1.4. Esperança matemática

Por definição, se a variável aleatória é discreta, a esperança de  $X$  é

$$\mu = E(X) = \sum X_i P(X_i)$$

e, se a variável aleatória é contínua, a esperança de  $X$  é

$$\mu = E(X) = \int_{-\infty}^{+\infty} xf(X) dX$$

Pode-se demonstrar, dadas as variáveis aleatórias  $X$  e  $Y$  e a constante  $K$ , que a esperança apresenta as seguintes propriedades:

- a)  $E(K) = K$
- b)  $E(X + K) = E(X) + K$
- c)  $E(KX) = KE(X)$
- d)  $E(X + Y) = E(X) + E(Y)$

e, se  $X$  e  $Y$  são independentes,

$$e) E(XY) = E(X) \cdot E(Y)$$

#### 1.5. Variância e covariância

Por definição, a variância de uma variável aleatória  $X$ , de população infinita, é

$$\sigma^2 = V(X) = E[X - E(X)]^2 = E(X - \mu)^2$$

A variância é uma medida de dispersão da distribuição.

Demonstremos, a seguir, que, se  $K$  é uma constante,  $V(KX) = K^2V(X)$ .

Temos

$$\begin{aligned} V(KX) &= E[KX - E(KX)]^2 = \\ &= E[KX - KE(X)]^2 = \\ &= E\{K^2[X - E(X)]^2\} = \\ &= K^2E[X - E(X)]^2 = \\ &= K^2V(X), \quad \text{c.q.d.} \end{aligned}$$

Dadas duas variáveis aleatórias,  $X$  e  $Y$ , a covariância entre  $X$  e  $Y$  é, por definição:

$$\begin{aligned} \text{cov}(X, Y) &= E[X - E(X)][Y - E(Y)] = \\ &= E(X - \mu_X)(Y - \mu_Y) \end{aligned}$$

Demonstremos, a seguir, que

$$V(X + Y) = V(X) + V(Y) + 2 \text{cov}(X, Y)$$

Temos

$$V(X + Y) = E[(X + Y) - E(X + Y)]^2$$

Então

$$\begin{aligned} V(X + Y) &= E\{(X - E(X)) + (Y - E(Y))\}^2 = \\ &= E[(X - \mu_X)^2 + (Y - \mu_Y)^2 + 2(X - \mu_X)(Y - \mu_Y)] = \\ &= V(X) + V(Y) + 2 \text{cov}(X, Y) \end{aligned}$$

É fácil verificar que

$$V(X - Y) = V(X) + V(Y) - 2 \text{cov}(X, Y)$$

Se  $X$  e  $Y$  são duas variáveis aleatórias independentes temos

$$\begin{aligned} \text{cov}(X, Y) &= E(X - \mu_X)(Y - \mu_Y) = \\ &= E(X - \mu_X) \cdot E(Y - \mu_Y) = 0 \end{aligned}$$

Segue-se que, no caso de variáveis independentes,

$$V(X \pm Y) = V(X) + V(Y)$$

Para exemplificar, consideremos que um tetraedro regular, feito de material homogêneo, em cujas faces estão marcados os números 0, 2, 4 e 6, é lançado. Seja  $X$  a variável aleatória que representa o valor marcado na face que ficar em contato com a mesa. Os sucessivos lançamentos desse tetraedro geram uma população infinita, em que a cada um dos 4 diferentes valores está associada a probabilidade 1/4.

Então

$$\mu_X = E(X) = \sum_{i=1}^4 X_i P(X_i) = 0 \cdot \frac{1}{4} + 2 \cdot \frac{1}{4} + 4 \cdot \frac{1}{4} + 6 \cdot \frac{1}{4} = 3$$

e

$$\begin{aligned} \sigma_X^2 &= V(X) = E(X - 3)^2 = \\ &= (-3)^2 \cdot \frac{1}{4} + (-1)^2 \cdot \frac{1}{4} + 1^2 \cdot \frac{1}{4} + 3^2 \cdot \frac{1}{4} = 5 \end{aligned}$$

Consideremos, agora, que temos dois tetraedros, um azul e outro branco. Sejam  $X$  e  $Y$  as variáveis aleatórias que representam os valores obtidos nos tetraedros azul e branco, respectivamente.

Temos

$$\begin{aligned} \mu_X &= \mu_Y = 3 \\ \sigma_X^2 &= \sigma_Y^2 = 5 \end{aligned}$$

Uma vez que  $X$  e  $Y$  são, obviamente, variáveis independentes, devemos verificar que  $\text{cov}(X, Y) = 0$ .

Na tabela 1.1 são dados os valores do produto  $(X - \mu_X)(Y - \mu_Y)$  a serem utilizados no cálculo da  $\text{cov}(X, Y)$ .

TABELA 1.1. Valores de  $(X - \mu_X)(Y - \mu_Y) = (X - 3)(Y - 3)$

$Y$	$X$			
	0	2	4	6
0	9	3	-3	-9
2	3	1	-1	-3
4	-3	-1	1	3
6	-9	-3	3	9

Verificamos então que

$$\begin{aligned}\text{cov}(X, Y) &= E(X - \mu_X)(Y - \mu_Y) = \\ &= 9 \cdot \frac{1}{16} + 3 \cdot \frac{1}{16} + \dots + 9 \cdot \frac{1}{16} = 0\end{aligned}$$

Seja  $Z = X + Y$

$$\text{Então } V(Z) = V(X) + V(Y) + 2 \text{ cov}(X, Y) = 5 + 5 = 10$$

Verifiquemos este resultado calculando  $V(Z)$  diretamente da definição. Na tabela 1.2 são apresentados os valores de  $Z = X + Y$ .

TABELA 1.2. Soma dos valores obtidos lançando dois tetraedros

Y	X				
	0	2	4	6	
0	0	2	4	6	
2	2	4	6	8	
4	4	6	8	10	
6	6	8	10	12	

Temos que

$$E(Z) = E(X + Y) = E(X) + E(Y) = 3 + 3 = 6$$

Esse valor também pode ser obtido calculando a média dos valores obtidos na tabela 1.2, como segue:

$$E(Z) = 0 \cdot \frac{1}{16} + 2 \cdot \frac{1}{16} + 4 \cdot \frac{1}{16} + \dots + 12 \cdot \frac{1}{16} = 6$$

Finalmente, obtemos

$$\begin{aligned}V(Z) &= E[Z - E(Z)]^2 = \\ &= (0 - 6)^2 \cdot \frac{1}{16} + (2 - 6)^2 \cdot \frac{1}{16} + \dots + (12 - 6)^2 \cdot \frac{1}{16} = 10,\end{aligned}$$

confirmando o resultado obtido anteriormente.

Devemos ressaltar que, embora  $\text{cov}(X, Y) = 0$  sempre que  $X$  e  $Y$  são variáveis aleatórias independentes, o inverso não é verdadeiro, isto é, se  $\text{cov}(X, Y) = 0$ , não podemos concluir que  $X$  e  $Y$  são independentes. Na tabela 1.3 apresentamos uma distribuição conjunta em que  $\text{cov}(X, Y) = 0$  e as variáveis não são independentes, pois

$$P(X_i, Y_j) \neq P(X_i) \cdot P(Y_j)$$

TABELA 1.3. Valores de  $P(X_i, Y_j)$  para a distribuição conjunta de duas variáveis dependentes com  $\text{cov}(X, Y) = 0$

Y	X			$P(Y)$
	-1	0	1	
-1	0,10	0,30	0,10	0,50
1	0,25	0	0,25	0,50
$P(X)$	0,35	0,30	0,35	1,00

Entretanto, é possível demonstrar que, se as variáveis têm distribuição normal, o fato de a covariância ser igual a zero é condição suficiente para podermos afirmar que são variáveis independentes.

Vejamos, a seguir, um exemplo de duas variáveis com covariância não nula. No lançamento do tetraedro descrito anteriormente, seja  $X$  o valor marcado na face que fica em contato com a mesa e seja  $W$  a soma dos valores marcados nas outras 3 faces. A tabela 1.4 mostra os valores de  $X$  e de  $W$ , bem como do produto  $[X - E(X)][W - E(W)]$

TABELA 1.4. Valores necessários para o cálculo da  $\text{cov}(X, W)$

X	W	$[X - E(X)][W - E(W)]$
0	12	-9
2	10	-1
4	8	-1
6	6	-9

Temos que

$$E(X) = 3,$$

$$E(W) = 9$$
 e

$$\begin{aligned}\text{cov}(X, W) &= [X - E(X)][W - E(W)] = \\ &= (-9)\frac{1}{4} + (-1)\frac{1}{4} + (-1)\frac{1}{4} + (-9)\frac{1}{4} = -5\end{aligned}$$

Como exercício, o leitor pode verificar que  $V(W - X) = 20$ .

Pode-se demonstrar que, se  $K$  é uma constante e se  $X$ ,  $Y$  e  $Z$  são variáveis aleatórias, a covariância apresenta as seguintes propriedades:

- a)  $\text{cov}(X + Y, Z) = \text{cov}(X, Z) + \text{cov}(Y, Z)$
- b)  $\text{cov}(KX, Y) = \text{cov}(X, KY) = K \text{cov}(X, Y)$
- c)  $\text{cov}(K, X) = \text{cov}(X, K) = 0$

Segue-se que, se  $\alpha_1$ ,  $\beta_1$ ,  $\gamma_1$ ,  $\alpha_2$ ,  $\beta_2$  e  $\gamma_2$  são constantes,

$$\begin{aligned}\text{cov}(\alpha_1 + \beta_1 X + \gamma_1 Y, \alpha_2 + \beta_2 X + \gamma_2 Y) &= \\ &= \beta_1 \beta_2 V(X) + (\gamma_1 \beta_2 + \beta_1 \gamma_2) \text{cov}(X, Y) + \gamma_1 \gamma_2 V(Y)\end{aligned}$$

Como caso particular temos:

$$\text{cov}(X, \alpha + \beta X) = \beta V(X)$$

Este último resultado pode ser utilizado para obter a covariância entre as variáveis  $X$  e  $W$  da tabela 1.4. Como a soma de todos os valores marcados no tetraedro é sempre igual a 12, temos que  $W = 12 - X$ . Então

$$\text{cov}(X, W) = \text{cov}(X, 12 - X) = -V(X) = -5,$$

confirmando o resultado obtido anteriormente.

## 1.6. Estimador não tendencioso

Por definição,  $a$  é um estimador não-tendencioso (não-viesado ou imparcial) do parâmetro  $\alpha$  da população se

$$E(a) = \alpha$$

É importante lembrar que o estimador  $a$  é uma variável, isto é, ele representa uma dada fórmula de cálculo que fornecerá valores que serão diferentes, conforme a amostra selecionada.

Para exemplificar, consideremos, novamente, a população infinita gerada pelo lançamento do tetraedro regular em cujas faces estão marcados os valores 0, 2, 4 e 6.

Já vimos que  $\mu = E(X) = 3$  e  $\sigma^2 = V(X) = 5$

Lançando o tetraedro duas vezes, podemos obter amostras com  $n = 2$  elementos dessa população. Na tabela 1.5 apresentamos as dezesseis amostras de tamanho  $n = 2$ , que podem ser obtidas, e as respectivas estimativas dos parâmetros  $\mu$  e  $\sigma^2$ . Os estimadores são

$$\bar{X} = \frac{\sum X_i}{n} = \frac{X_1 + X_2}{2}$$

e

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1} = (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2$$

Calculamos, também, as estimativas da variância da média da amostra. Esta variância é definida por

$$\sigma_{\bar{X}}^2 = V(\bar{X}) = E[\bar{X} - E(\bar{X})]^2$$

Temos

$$V(\bar{X}) = V\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{1}{n^2} V(X_1 + X_2 + \dots + X_n)$$

Uma vez que as observações de uma amostra aleatória de uma população infinita são independentes, segue-se que

$$V(\bar{X}) = \frac{1}{n^2} n \sigma^2 = \frac{\sigma^2}{n}$$

O estimador da variância média é  $s_{\bar{X}}^2 = \frac{s^2}{n}$

Obviamente, cada uma das dezesseis amostras tem probabilidade 1/16 de ser selecionada.

TABELA 1.5. Valores de  $\bar{X}$ ,  $s^2$ ,  $s_{\bar{X}}^2$  e  $(\bar{X} - \mu)^2$  para as 16 amostras que podem ser obtidas lançando duas vezes o tetraedro.

Amostra	$\bar{X}$	$s^2$	$s_{\bar{X}}^2$	$(\bar{X} - \mu)^2$
0 e 0	0	0	0	9
0 e 2	1	2	1	4
0 e 4	2	8	4	1
0 e 6	3	18	9	0
2 e 0	1	2	1	4
2 e 2	2	0	0	1
2 e 4	3	2	1	0
2 e 6	4	8	4	1
4 e 0	2	8	4	1
4 e 2	3	2	1	0
4 e 4	4	0	0	1
4 e 6	5	2	1	4
6 e 0	3	18	9	0
6 e 2	4	8	4	1
6 e 4	5	2	1	4
6 e 6	6	0	0	9

Verificamos que

$$E(\bar{X}) = 0 \cdot \frac{1}{16} + 1 \cdot \frac{1}{16} + 2 \cdot \frac{1}{16} + \dots + 5 \cdot \frac{1}{16} + 6 \cdot \frac{1}{16} = \frac{48}{16} = \mu,$$

Ou seja,  $\bar{X}$  é um estimador não-tendencioso (não viesado, não-viciado ou imparcial) de  $\mu$ . Isto pode ser facilmente demonstrado:

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \\ &= \frac{1}{n}[E(X_1) + E(X_2) + \dots + E(X_n)] = \frac{n\mu}{n} = \mu \end{aligned}$$

Verificamos, também, que

$$E(s^2) = 0 \cdot \frac{1}{16} + 2 \cdot \frac{1}{16} 8 \cdot \frac{1}{16} + \dots + 2 \cdot \frac{1}{16} 0 \cdot \frac{1}{16} = \frac{80}{16} = 5 = \sigma^2,$$

ou seja,  $s^2$  é um estimador não-tendencioso de  $\sigma^2$ .

A variância da média da amostra pode ser obtida através da expressão

$$V(\bar{X}) = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} = \frac{5}{2}$$

ou diretamente, a partir da definição, utilizando os valores da última coluna da tabela 1.5, como segue:

$$\begin{aligned} V(\bar{X}) &= E[\bar{X} - E(\bar{X})]^2 = E(\bar{X} - \mu)^2 = 9 \cdot \frac{1}{16} + \\ &+ 4 \cdot \frac{1}{16} + \dots + 9 \cdot \frac{1}{16} = \frac{40}{16} = \frac{5}{2} \end{aligned}$$

Considerando os valores de  $s_{\bar{X}}^2$  apresentados na tabela 1.5, verificamos que

$$E(s_{\bar{X}}^2) = 0 \cdot \frac{1}{16} + 1 \cdot \frac{1}{16} + 4 \cdot \frac{1}{16} + \dots + 1 \cdot \frac{1}{16} + 0 \cdot \frac{1}{16} = \frac{40}{16} = \frac{5}{2},$$

ou seja,  $s_{\bar{X}}^2$  é um estimador não-tendencioso de  $\sigma_{\bar{X}}^2$ .

Devemos ressaltar que o exemplo apresentado refere-se a uma *população infinita*. As mesmas fórmulas serão válidas se, de uma *população finita*, tirarmos amostras *com reposição* dos elementos.

Consideremos, agora, o caso de uma *população finita* (com  $m$  elementos) da qual se tiram amostras (de  $n$  elementos) *sem reposição*.

A média da população é

$$\mu = E(X) = \frac{1}{m} \sum_{i=1}^m X_i$$

A variância de  $X$  é definida por (ver Cochran, 1965, p. 42)

$$V(X) = S^2 = \frac{1}{m-1} \sum_{i=1}^m (X_i - \mu)^2$$

Demonstra-se que (ver Cochran, 1965, p. 44)

$$V(\bar{X}) = \sigma_{\bar{X}}^2 = \frac{S^2}{n} \left(1 - \frac{n}{m}\right)$$

Dada uma amostra (sem reposição) de  $n$  elementos, uma estimativa não-tendenciosa de  $\mu$  é dada por

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

As estimativas não-tendenciosas de  $S^2$  e  $\sigma_{\bar{X}}^2$  são dadas, respectivamente, por

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \quad \text{e} \quad s_{\bar{X}}^2 = \frac{s^2}{n} \left(1 - \frac{n}{m}\right)$$

Vejamos um exemplo numérico simples, embora artificial. Seja uma população de apenas 4 elementos ( $m = 4$ ), onde  $X_i$  assume os valores 0, 2, 4 e 6. Temos que

$$\mu = \frac{0+2+4+6}{4} = 3$$

e

$$S^2 = \frac{\sum(X_i - \mu)^2}{m-1} = \frac{(0-3)^2 + (1-3)^2 + (4-3)^2 + (6-3)^2}{3} = \frac{20}{3}$$

Consideremos as  $\binom{4}{2} = 6$  diferentes amostras de 2 elementos ( $n = 2$ ) que podemos tirar dessa população. Essas amostras estão discriminadas na tabela 1.6, com os correspondentes valores de  $\bar{X}$ ,  $s^2$ ,  $s_{\bar{X}}^2$  e  $(\bar{X} - \mu)^2$ .

TABELA 1.6. Valores de  $X_i$ ,  $\bar{X}$ ,  $s^2$ ,  $s_{\bar{X}}^2$  e  $(\bar{X} - \mu)^2$  para as 6 possíveis amostras de 2 elementos (sem reposição).

Valores de $X_i$	$\bar{X}$	$s^2$	$s_{\bar{X}}^2$	$(\bar{X} - \mu)^2$
0 e 2	1	2	1/2	4
0 e 4	2	8	2	1
0 e 6	3	18	9/2	0
2 e 4	3	2	1/2	0
2 e 6	4	8	2	1
4 e 6	5	2	1/2	4

Para amostras com  $n = 2$  elementos, temos

$$V(\bar{X}) = \sigma_{\bar{X}}^2 = \frac{S^2}{n} \left(1 - \frac{n}{m}\right) = \frac{20}{6} \left(1 - \frac{2}{4}\right) = \frac{5}{3}$$

O mesmo resultado pode ser obtido a partir da definição de variância, utilizando os valores da última coluna da tabela 1.6. Como as 6 diferentes amostras são igualmente prováveis, temos

$$\sigma_{\bar{X}}^2 = E(\bar{X} - \mu)^2 = \frac{4+1+0+0+1+4}{6} = \frac{10}{6} = \frac{5}{3}$$

Verificamos que:

$$E(\bar{X}) = \frac{1}{6}(1+2+\dots+5) = \frac{18}{6} = 3,$$

$$\text{ou seja, } E(\bar{X}) = \mu$$

$$E(s^2) = \frac{1}{6}(2+8+\dots+2) = \frac{40}{6} = \frac{20}{3},$$

$$\text{ou seja, } E(s^2) = S^2$$

$$E(s_{\bar{X}}^2) = \frac{1}{6} \left( \frac{1}{2} + 2 + \dots + \frac{1}{2} \right) = \frac{1}{6} \cdot \frac{20}{2} = \frac{5}{3}$$

$$\text{ou seja, } E(s_{\bar{X}}^2) = \sigma_{\bar{X}}^2$$

## 1.7. Estimador de variância mínima

A não-tendenciosidade ou ausência de viés é uma qualidade desejável para os estimadores. Entretanto, essa qualidade é insuficiente como critério para selecionar um estimador. Assim, por exemplo, no caso da média de uma população, podemos verificar que qualquer média ponderada dos valores de uma amostra é um estimador não tendencioso de  $\mu$ .

Consideremos a média ponderada

$$m = \sum_{i=1}^n \pi_i X_i, \text{ com } \sum \pi_i = 1$$

Temos que

$$E(m) = \sum \pi_i E(X_i) = \mu \sum \pi_i = \mu$$

Isso mostra que qualquer média ponderada dos valores observados em uma amostra aleatória é um estimador não-tendencioso de  $\mu$ . Portanto, existem infinitos estimadores não-tendenciosos de  $\mu$ .

Dados dois estimadores não-tendenciosos de  $\alpha$ ,  $a_1$  e  $a_2$ , por definição a eficiência relativa de  $a_2$ , em comparação com  $a_1$ , é igual a

$$\frac{V(a_1)}{V(a_2)}$$

Assim, por exemplo, dada uma amostra aleatória com 2 elementos,  $X_1$  e  $X_2$ , de uma população infinita, consideremos 2 estimadores não-tendenciosos da média da população:

a) a média aritmética  $\bar{X} = \frac{X_1 + X_2}{2} = \frac{1}{2}X_1 + \frac{1}{2}X_2$  e

b) a média ponderada  $m = \frac{1}{4}X_1 + \frac{3}{4}X_2$

Temos

$$V(\bar{X}) = \frac{\sigma^2}{2}$$

e

$$V(m) = \frac{1}{16}\sigma^2 + \frac{9}{16}\sigma^2 = \frac{5}{8}\sigma^2$$

A eficiência de  $m$  em relação a  $\bar{X}$  é

$$\frac{\frac{1}{2}\sigma^2}{\frac{5}{8}\sigma^2} = \frac{4}{5} = 0,8 \text{ ou } 80\%$$

É fácil provar que, dada uma amostra com 2 observações ( $X_1$  e  $X_2$ ), dentre os estimadores da classe

$$m = \theta X_1 + (1 - \theta) X_2,$$

o mais eficiente é a média aritmética, ou seja, o caso em que  $\theta = \frac{1}{2}$ .

Temos

$$V(m) = \theta^2 \sigma^2 + (1 - \theta)^2 \sigma^2 = (1 - 2\theta + 2\theta^2) \sigma^2$$

Igualando a zero a derivada em relação a  $\theta$  e simplificando, obtemos

$$-2 + 4\theta = 0$$

Donde

$$\theta = \frac{1}{2}$$

A derivada segunda é positiva, confirmando que a variância é mínima quando

$$\theta = \frac{1}{2}.$$

Generalizando esse resultado, demonstraremos que, dada uma variável aleatória  $X$  de população infinita com média  $\mu$  e variância  $\sigma^2$ , a média aritmética de uma amostra aleatória de  $n$  observações é, dentre os estimadores lineares não-tendenciosos, o estimador de variância mínima.

Dizemos que um estimador é linear quando ele é uma combinação linear dos valores da amostra. Como exemplo, consideremos o seguinte estimador linear de  $\mu$ :

$$m = \sum_{i=1}^n \pi_i X_i$$

Temos que

$$E(m) = \mu \sum \pi_i$$

Para que  $m$  seja estimador não-tendencioso de  $\mu$ , devemos ter

$$\sum \pi_i = 1$$

Temos, também, que

$$V(m) = \sigma^2 \sum \pi_i^2$$

Para minimizar  $V(m)$  devemos minimizar  $\sum \pi_i^2$ , considerando a restrição  $\sum \pi_i = 1$ . Utilizando o método do multiplicador de Lagrange, definimos a função

$$\phi = \sum \pi_i^2 - \lambda(\sum \pi_i - 1)$$

Igualando a zero as derivadas parciais em relação a  $\pi_i$  e  $\lambda$ , obtemos o sistema de equações

$$2\pi_i - \lambda = 0, \quad i = 1, 2, \dots, n \quad (1.1)$$

$$\sum \pi_i = 1 \quad (1.2)$$

De (1.1), obtemos

$$\pi_i = \frac{\lambda}{2} \quad (1.3)$$

Substituindo (1.3) em (1.2), obtemos

$$\frac{n\lambda}{2} = 1$$

Donde

$$\frac{\lambda}{2} = \frac{1}{n}$$

Comparando esse resultado com (1.3) concluímos que

$$\pi_i = \frac{1}{n}, \text{ c.q.d.}$$

Não há necessidade de verificar a condição de 2<sup>a</sup> ordem para mínimo por se tratar de uma soma de quadrados.

## 1.8. Estimadores de mínimos quadrados

Pode parecer óbvio que o estimador da média de uma variável seja a média dos valores observados em uma amostra. Mas em situações um pouco mais complicadas será necessário recorrer a um método geral de determinação de estimadores, como o método dos mínimos quadrados ou o método da máxima verossimilhança (que será descrito na próxima seção).

O método dos mínimos quadrados consiste em adotar os estimadores que minimizam a soma dos quadrados dos desvios entre valores estimados e valores observados na amostra.

Mostraremos que a média aritmética dos valores da amostra é um estimador de mínimos quadrados. Para tanto, determinemos o valor de  $a$  que minimiza  $\sum_{i=1}^n (X_i - a)^2$ .

Derivando em relação a  $a$  e igualando a zero, obtemos:

$$2 \sum (X_i - a)(-1) = 0$$

$$\sum X_i - na = 0$$

Donde

$$a = \frac{\sum X_i}{n} = \bar{X}, \text{ c.q.d.}$$

É interessante notar que o método de mínimos quadrados conduz à média aritmética, mas que existem outros critérios associados às demais medidas de tendência central. Assim, para minimizar o valor absoluto do maior desvio, devemos adotar o ponto central entre os extremos (o ponto médio entre o menor e o maior valor); para maximizar o número de desvios iguais a zero devemos adotar a moda da amostra; e para minimizar a soma dos valores absolutos dos desvios devemos adotar a mediana. Para verificar essa última afirmativa, consideremos a distribuição de freqüências apresentada na tabela 1.7.

TABELA 1.7. Distribuição de freqüências com 13 distribuições

$X$	:	0	1	2	3	4	5	6	7	8
Freqüência:		1	5	1	1	1	1	2	0	1

É fácil verificar que a moda é 1, a mediana é 2, a média aritmética é 3 e o ponto central entre os extremos é 4.

A soma dos valores absolutos dos desvios em relação à mediana é 27 (7 para os valores abaixo da mediana e 20 para os valores acima da mediana). Para mostrar que a mediana é o ponto que minimiza a soma dos valores absolutos dos desvios, consideremos um ponto abaixo da mediana diferindo desta de menos de 1 unidade, isto é, o ponto de abcissa  $2 - \Delta$ , com  $0 < \Delta < 1$ . Para os 6 pontos abaixo da mediana, os desvios ficam aumentados de  $\Delta$ ; para os 6 pontos localizados acima da mediana, os desvios ficam aumentados de  $\Delta$  e para o ponto cuja abcissa é igual à mediana surge um desvio igual a  $\Delta$  em valor absoluto. A soma dos valores absolutos dos desvios em relação ao ponto de abcissa  $2 - \Delta$  é, portanto,

$$7 - 6\Delta + 20 + 6\Delta + \Delta = 27 + \Delta > 27$$

Raciocínio semelhante mostra que a soma dos valores absolutos dos desvios em relação a um ponto acima da mediana também é maior do que 27. Concluímos, então, que essa soma é mínima quando referida à mediana.

Vejamos um exemplo onde o uso da média aritmética, como medida de tendência central, parece ser mais razoável do que o uso da mediana, o que implica em afirmar que o critério de mínimos quadrados parece ser mais razoável do que a minimização da soma dos desvios absolutos. Consideremos uma amostra com 3 observações, onde  $X_1 = X_2 = 0$  e  $X_3 \neq 0$ . A mediana é igual a zero, qualquer que seja o valor de  $X_3$ , isto é, o valor da mediana independe de  $X_3$ . Entretanto, a média aritmética é igual a  $\frac{1}{3}X_3$ .

Para uma outra ilustração da aplicação do método de mínimos quadrados, consideremos a determinação do estimador do parâmetro  $p$  de uma distribuição binomial, sabendo que numa amostra de  $n$  observações foram constatados  $X$  casos favoráveis e  $n - X$  casos contrários. Como os valores esperados são de  $np$  casos

favoráveis e  $n(1 - p)$  casos contrários, queremos, de acordo com o método de mínimos quadrados, o valor de  $p$  que minimize

$$(X - np)^2 + [(n - X) - n(1 - p)]^2$$

Deixamos para o leitor verificar que a solução é

$$\hat{p} = \frac{X}{n}$$

### 1.9. Estimadores de máxima verossimilhança

De acordo com o método da máxima verossimilhança adotamos, como estimativas dos parâmetros, os valores que maximizam a probabilidade (no caso da variável aleatória ser discreta) ou a densidade de probabilidade (no caso de variável contínua) de ser obtida a amostra observada. Para obter estimadores de máxima verossimilhança é necessário conhecer ou pressupor qual é a distribuição da variável em estudo.

Para exemplificar, consideremos que cada uma das faces de um tetraedro regular são pintadas de branco ou de azul, e que, ao lançar o tetraedro, o resultado é considerado sucesso se a face que ficar em contato com a mesa for azul. Vamos supor que o tetraedro foi lançado 4 vezes, sem que soubéssemos se o número de faces azuis do tetraedro era 0, 1, 2, 3 ou 4. Somos então informados de que, nas 4 tentativas, foi obtido sucesso apenas uma vez. Qual é a estimativa de máxima verossimilhança para o número de faces azuis no tetraedro utilizado?

Na tabela 1.8 apresentamos a probabilidade de obter apenas um sucesso em 4 tentativas, para cada um dos casos possíveis.

TABELA 1.8. A função de verossimilhança.

Número de faces azuis	Probabilidade ( $p$ ) de obter sucesso em uma tentativa	Probabilidade de obter apenas um sucesso em 4 tentativas = $4p(1-p)^3$
0	0	0
1	1/4	27/64
2	1/2	1/4 = 16/64
3	3/4	3/64
4	1	0

A simples observação da tabela 1.8 mostra que o valor de  $p$  que maximiza a probabilidade de obter um sucesso em 4 tentativas é  $p = 1/4$ . Então, essa é a estimativa de máxima verossimilhança para a probabilidade de obter sucesso em um lançamento, ou seja, o tetraedro utilizado deve ter apenas uma face azul.

Se  $p$  varia continuamente, a estimativa de máxima verossimilhança pode ser obtida através das condições necessárias e suficientes do cálculo diferencial. Desejamos o valor de  $p$  que maximize

$$P(X) = \binom{n}{X} p^X (1-p)^{n-X},$$

onde  $X$  é o número de sucessos obtidos em  $n$  tentativas.

Como o logaritmo é uma função monotônica crescente, o valor de  $p$  que maximiza  $P(X)$  também maximiza

$$Z = \ln P(X) = \ln \binom{n}{X} + X \ln p + (n - X) \ln (1-p)$$

Igualando a zero a derivada em relação a  $p$ , obtemos

$$\frac{X}{\hat{p}} - \frac{n-X}{1-\hat{p}} = 0$$

cuja solução é  $\hat{p} = \frac{X}{n}$ , que é o estimador já obtido na seção anterior pelo método de mínimos quadrados.

Como

$$\frac{d^2Z}{dp^2} = -\frac{X}{p^2} - \frac{n-X}{(1-p)^2} < 0,$$

a condição de segunda ordem para máximo é satisfeita.

Como mais um exemplo, consideremos a determinação dos estimadores de máxima verossimilhança da média ( $\mu$ ) e da variância ( $\sigma^2$ ) de uma variável aleatória ( $X$ ), com distribuição normal, com base em uma amostra aleatória de  $n$  elementos.

Neste caso, a densidade de probabilidade de obter um valor  $X_i$  na amostra é

$$f(X_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(X_i - \mu)^2}{2\sigma^2}\right\}$$

Como as observações são independentes, a densidade de probabilidade de obter os valores  $X_1, X_2, \dots, X_n$  da amostra é

$$L(X_1, X_2, \dots, X_n; \mu, \sigma^2) = f(X_1) \cdot f(X_2) \cdot \dots \cdot f(X_n) =$$

$$= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(X_i - \mu)^2}{2\sigma^2}\right\} =$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{\sum(X_i - \mu)^2}{2\sigma^2}\right\}$$

Essa é a função de verossimilhança da amostra. É usual representá-la por  $L$  porque a palavra inglesa para verossimilhança é *likelihood*.

Os estimadores de máxima verossimilhança de  $\mu$  e  $\sigma^2$  são os valores que maximizam o valor de  $L(\mu, \sigma^2 | X_1, X_2, \dots, X_n)$ . Como o logaritmo é uma função monotônica crescente, os valores de  $\mu$  e  $\sigma^2$  que maximizam  $L$  também maximizam

$$\ln L = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{\sum(X_i - \mu)^2}{2\sigma^2}$$

Igualando a zero as derivadas parciais em relação a  $\mu$  e  $\sigma^2$  obtemos o sistema de equações

$$\begin{cases} \frac{2\sum(X_i - \hat{\mu})}{2\hat{\sigma}^2} = 0 \\ -\frac{n}{2\hat{\sigma}^2} + \frac{\sum(X_i - \hat{\mu})^2}{2\hat{\sigma}^4} = 0 \end{cases} \quad (1.4)$$

$$\begin{cases} \frac{2\sum(X_i - \hat{\mu})}{2\hat{\sigma}^2} = 0 \\ -\frac{n}{2\hat{\sigma}^2} + \frac{\sum(X_i - \hat{\mu})^2}{2\hat{\sigma}^4} = 0 \end{cases} \quad (1.5)$$

De (1.4) obtemos

$$\hat{\mu} = \frac{\sum X_i}{n} = \bar{X} \quad (1.6)$$

Já vimos que  $\bar{X}$  é um estimador de mínimos quadrados, não-tendencioso e de variância mínima. Sabemos agora que, se  $X$  tem distribuição normal,  $\bar{X}$  é, também, um estimador de máxima verossimilhança.

De (1.5) e (1.6) obtemos

$$\hat{\sigma}^2 = \frac{\sum(X_i - \bar{X})^2}{n}$$

É interessante notar que o estimador de máxima verossimilhança da variância é tendencioso, uma vez que o estimador não-tendencioso é

$$s^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

## 1.10. Propriedades assintóticas dos estimadores

Seja  $a_n$  o estimador de um parâmetro  $\alpha$ , obtido com base em uma amostra com  $n$  observações. Em geral  $a_n$  é uma variável aleatória cuja distribuição é caracterizada pela função de densidade  $f(a_n)$ , com média  $E(a_n)$  e variância  $V(a_n) = E[a_n - E(a_n)]^2$ . Variando o tamanho da amostra, temos várias seqüências:

a) a seqüência dos estimadores:

$$\{a_n\} = a_1, a_2, \dots, a_n, \dots \quad (1.7)$$

b) a seqüência das médias:

$$\{E(a_n)\} = E(a_1), E(a_2), \dots, E(a_n), \dots \quad (1.8)$$

c) a seqüência das variâncias:

$$\{V(a_n)\} = V(a_1), V(a_2), \dots, V(a_n), \dots \quad (1.9)$$

d) a seqüência das funções de densidade:

$$\{f(a_n)\} = f(a_1), f(a_2), \dots, f(a_n), \dots \quad (1.10)$$

A teoria assintótica dos estimadores se destina a estabelecer o comportamento dessas seqüências quando  $n$  tende para infinito.

Denominamos *esperança assintótica* de  $a_n$  ao valor do  $\lim_{n \rightarrow \infty} E(a_n)$ . Se

$\lim_{n \rightarrow \infty} E(a_n) = \alpha$ , dizemos que  $a_n$  é um estimador *assintoticamente não-tendencioso*.

Poderíamos pensar em definir a variância assintótica de  $a_n$  como  $\lim_{n \rightarrow \infty} V(a_n)$ .

Entretanto, esse limite é freqüentemente igual a zero, porque a distribuição de  $a_n$  se concentra em um único ponto. Para exemplificar, consideremos a média ( $\bar{X}$ ) de uma amostra aleatória com  $n$  observações da variável  $X$ , de média  $\mu$  e variância  $\sigma^2$ . De  $V(\bar{X}) = \sigma^2 / n$  segue-se que

$$\lim_{n \rightarrow \infty} V(\bar{X}) = 0$$

Pode-se demonstrar que, quando  $n$  cresce, a distribuição da mediana ( $m$ ) da amostra se concentra em torno de  $\mu$  e o limite de sua variância também é zero, isto é,

$$\lim_{n \rightarrow \infty} V(m) = 0$$

Para verificar qual de dois estimadores é assintoticamente mais eficiente, poderíamos pensar em comparar os limites das variâncias desses estimadores, quando  $n$  tende para infinito. Entretanto, se esses limites são iguais a zero a eficiência relativa não é definida.

O problema é resolvido definindo *variância assintótica* como

$$n^{-1} \lim_{n \rightarrow \infty} E\left\{\sqrt{n}[a_n - E(a_n)]\right\}^2 \quad (1.11)$$

Para o estimador  $\bar{X}$  temos

$$V(\bar{X}) = E(\bar{X} - \mu)^2 = \frac{\sigma^2}{n}$$

Então

$$E[\sqrt{n}(X - \mu)]^2 = \sigma^2$$

e a variância assintótica de  $\bar{X}$  é

$$n^{-1} \lim_{n \rightarrow \infty} E[\sqrt{n}(X - \mu)]^2 = \frac{\sigma^2}{n}$$

Pode-se demonstrar que, se  $X$  tem distribuição normal, a variância assintótica da mediana ( $m$ ) da amostra é

$$n^{-1} \lim_{n \rightarrow \infty} E[\sqrt{n}(m - \mu)]^2 = \frac{\pi\sigma^2}{2n}$$

Como  $(\pi/2) > 1$ , concluímos que a média ( $\bar{X}$ ) é um estimador de  $\mu$  assintoticamente mais eficiente do que a mediana ( $m$ ).

Ao analisar a seqüência (1.7) é importante ter em mente que, fixado o valor de  $n$ ,  $a_n$  é uma variável aleatória. Por isso não tem sentido falar no limite de  $a_n$  quando  $n$  tende a infinito. É necessário, então, introduzir o conceito de *convergência em probabilidade*.

Dizemos que uma seqüência de variáveis aleatórias  $\{a_n\} = a_1, a_2, \dots, a_n, \dots$  converge em probabilidade para uma constante  $\alpha$  se, para qualquer  $\varepsilon > 0$ , arbitrariamente pequeno,

$$\lim_{n \rightarrow \infty} P(|a_n - \alpha| > \varepsilon) = 0, \quad (1.12)$$

indicando-se

$$a_n \xrightarrow{P} \alpha$$

ou

$$\text{plim } a_n = \alpha,$$

que se lê: “o limite em probabilidade de  $a_n$  é igual a  $\alpha$ ”.

Dada uma amostra de  $n$  observações,  $a_n$  é um estimador *consistente* do parâmetro  $\alpha$  da população se  $\text{plim } a_n = \alpha$ .

Antes de prosseguir vamos analisar melhor esse conceito. A expressão (1.12) pode ser escrita

$$\lim_{n \rightarrow \infty} P(\alpha - \varepsilon < a_n < \alpha + \varepsilon) = 1 \quad (1.13)$$

Na figura 1.3 representamos a distribuição de  $a_n$  para  $n = 10$  e  $n = 100$  e assinalamos, por meio de traços verticais, os limites  $\alpha - \varepsilon$  e  $\alpha + \varepsilon$ . De acordo com (1.13), para que  $a_n$  seja um estimador consistente de  $\alpha$ , a probabilidade de termos  $\alpha - \varepsilon < a_n < \alpha + \varepsilon$  deve tender para um quando  $n$  tende para infinito. Em outras palavras, dados  $\varepsilon$  e  $\omega$ , positivos e arbitrariamente pequenos, deve existir  $n_o$  tal que para todo  $n > n_o$  temos

$$P(\alpha - \varepsilon < a_n < \alpha + \varepsilon) > 1 - \omega$$

Em termos da figura 1.3, à medida que  $n$  cresce, a distribuição de  $a_n$  deve se concentrar em torno de  $\alpha$ , de maneira que quase toda a distribuição fique compreendida entre os limites  $\alpha - \varepsilon$  e  $\alpha + \varepsilon$ .

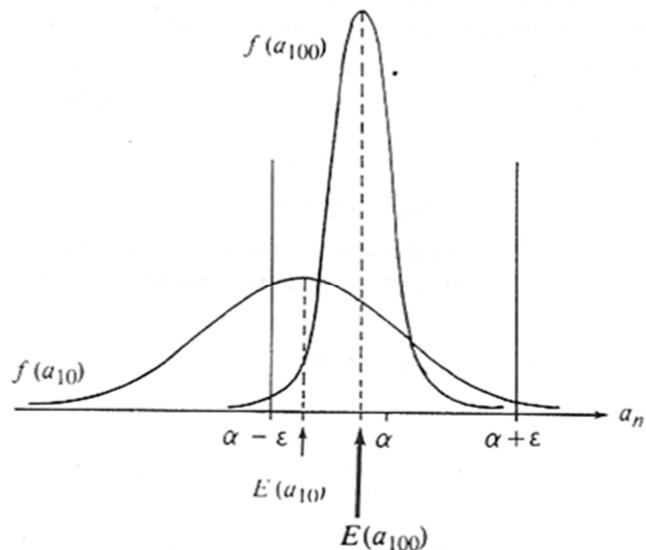


Figura 1.3. O conceito de estimador consistente

Prosseguindo no estudo das propriedades assintóticas dos estimadores, vejamos o conceito de convergência em média quadrática. Dizemos que uma série de variáveis aleatórias  $\{a_n\} = a_1, a_2, \dots, a_n, \dots$  converge em média quadrática para uma constante  $\alpha$  se

$$\lim_{n \rightarrow \infty} E(a_n - \alpha)^2 = 0 \quad (1.14)$$

Demonstraremos adiante que a convergência em média quadrática é condição suficiente para que tenhamos convergência em probabilidade. Para isso vamos deduzir, preliminarmente, a desigualdade de Chebyshev.

Consideremos uma variável aleatória  $Z \geq 0$ , com média finita, e um número real  $\theta > 0$ . Definimos a variável aleatória  $Y$  da seguinte maneira:

$$Y = 0, \text{ se } Z < \theta$$

e

$$Y = \theta, \text{ se } Z \geq \theta$$

Então,

$$P(Y = 0) = P(Z < \theta)$$

e

$$P(Y = \theta) = P(Z \geq \theta)$$

Segue-se que

$$E(Y) = 0 \cdot P(Y = 0) + \theta \cdot P(Y = \theta) = \theta \cdot P(Z \geq \theta) \quad (1.15)$$

Da definição de  $Y$ , segue-se que

$$Y \leq Z$$

Então,

$$E(Y) \leq E(Z)$$

Considerando (1.15) temos:

$$\theta \cdot P(Z \geq \theta) \leq E(Z)$$

ou

$$P(Z \geq \theta) \leq \frac{E(Z)}{\theta} \quad (1.16)$$

Consideremos agora a variável aleatória  $X$ , com média  $\mu$  e variância  $\sigma^2$ . Aplicando a relação (1.16) à variável aleatória  $(X - \mu)^2 \geq 0$  e ao número  $k^2$ , obtemos

$$P[(X - \mu)^2 \geq k^2] \leq \frac{E(X - \mu)^2}{k^2} = \frac{\sigma^2}{k^2} \quad (1.17)$$

Donde, com  $k > 0$ ,

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2},$$

que é a desigualdade de Chebyshev.

Demonstremos agora que a convergência em média quadrática é condição suficiente para que tenhamos convergência em probabilidade. Aplicando a relação (1.16) à variável  $(a_n - \alpha)^2$  e ao número  $\varepsilon^2$ , obtemos

$$P[(a_n - \alpha)^2 \geq \varepsilon^2] \leq \frac{E(a_n - \alpha)^2}{\varepsilon^2}$$

Então

$$\lim_{n \rightarrow \infty} P[(a_n - \alpha)^2 \geq \varepsilon^2] \leq \lim_{n \rightarrow \infty} \frac{E(a_n - \alpha)^2}{\varepsilon^2}$$

Se  $a_n$  converge em média quadrática para  $\alpha$ , temos

$$\lim_{n \rightarrow \infty} E(a_n - \alpha)^2 = 0$$

Segue-se que

$$\lim_{n \rightarrow \infty} P[(a_n - \alpha)^2 \geq \varepsilon^2] = 0$$

Lembrando que para uma variável aleatória contínua a probabilidade de se observar um determinado valor é nula, podemos escrever

$$\lim_{n \rightarrow \infty} P[(a_n - \alpha)^2 > \varepsilon^2] = 0$$

ou

$$\lim_{n \rightarrow \infty} P(|a_n - \alpha| > \varepsilon) = 0$$

isto é,

$$\text{plim } a_n = \alpha$$

Demonstremos, também, que

$$E(a_n - \alpha)^2 = V(a_n) + [E(a_n) - \alpha]^2 \quad (1.18)$$

Temos

$$\begin{aligned} E(a_n - \alpha)^2 &= E\{[a_n - E(a_n)] + [E(a_n) - \alpha]\}^2 = \\ &= E\{[a_n - E(a_n)]^2 + [E(a_n) - \alpha]^2 + 2[a_n - E(a_n)][E(a_n) - \alpha]\} = \\ &= V(a_n) + [E(a_n) - \alpha]^2, \quad \text{c.q.d.} \end{aligned}$$

Vamos resumir as definições e resultados obtidos até esse ponto.

Para que o estimador  $a_n$ , baseado numa amostra de  $n$  observações, seja um estimador consistente de  $\alpha$ , isto é, para que

$$\text{plim } a_n = \alpha,$$

é suficiente que

$$\lim_{n \rightarrow \infty} E(a_n - \alpha)^2 = 0$$

Para que isso aconteça, por sua vez, é suficiente, de acordo com (1.18), que

$$\lim_{n \rightarrow \infty} V(a_n) = 0$$

e

$$E(a_n) = \alpha$$

ou

$$\lim_{n \rightarrow \infty} [E(a_n)] = \alpha$$

Concluímos então que um estimador não-tendencioso ou assintoticamente não-tendencioso é consistente se o limite da sua variância, quando o tamanho da amostra tende para infinito, é igual a zero.

Vejamos um exemplo. Sabemos que  $\bar{X}$  é um estimador não-tendencioso de  $\mu$  e que  $V(\bar{X}) = \frac{\sigma^2}{n}$ .

Como

$$\lim_{n \rightarrow \infty} V(\bar{X}) = 0,$$

concluímos que  $\text{plim } \bar{X} = \mu$ , isto é,  $\bar{X}$  é um estimador consistente de  $\mu$ .

Vimos que os estimadores devem ser não-tendenciosos e eficientes. É desejável, também, que sejam consistentes e assintoticamente eficientes, isto é, que apresentem variância assintótica mínima. A não-tendenciosidade e a eficiência são denominadas propriedades de amostra pequena, porque sua validade não depende do tamanho da amostra, isto é, quando um estimador apresenta tais propriedades, elas são igualmente válidas para amostras grandes e para amostras pequenas. Por outro lado, as propriedades definidas em termos de limites, quando o tamanho ( $n$ ) da amostra tende para infinito, são denominadas propriedades de amostra grande ou propriedades assintóticas.

A seguir são apresentadas, sem demonstração, algumas propriedades da convergência em probabilidade.

Se  $\text{plim } a = \alpha$  e  $F(a)$  é uma função contínua de  $a$ , então  $\text{plim } F(a) = F(\alpha)$ . Em particular, temos  $\text{plim } (a^2) = (\text{plim } a)^2$  e  $\text{plim } (a^{-1}) = (\text{plim } a)^{-1}$ . O teorema se estende ao caso de uma função contínua de duas ou mais variáveis, isto é, se  $\text{plim } a = \alpha$ ,  $\text{plim } b = \beta$  e  $F(a, b)$  é uma função contínua, temos

$\text{plim } F(a, b) = F(\alpha, \beta)$ . Temos, por exemplo,  $\text{plim } (a + b) = \text{plim } a + \text{plim } b$ ,  $\text{plim } (ab) = (\text{plim } a)(\text{plim } b)$  e, se  $\text{plim } b \neq 0$ ,  $\text{plim } (a/b) = (\text{plim } a)/(\text{plim } b)$ .

Essas propriedades facilitam a determinação do valor para o qual converge em probabilidade uma função de estimadores. Note que, conhecida a esperança matemática de várias variáveis, não é geralmente tão imediata a determinação da esperança matemática de expressões envolvendo tais variáveis. Dado que  $E(a) = \alpha$  e  $E(b) = \beta$ , sabemos que  $E(a + b) = \alpha + \beta$ , mas nada podemos dizer, de imediato, sobre o valor de  $E(a^2)$ ,  $E(ab)$  ou  $E(a/b)$ .

Para introduzir a idéia de convergência em distribuição, vamos considerar, novamente, a distribuição da média ( $\bar{X}$ ) de uma amostra aleatória com  $n$  observações, com  $E(X) = \mu$  e  $V(X) = \sigma^2$ , mas sem que se conheça a forma da distribuição de  $X$ . Já vimos que  $V(\bar{X})$  tende a zero quando  $n$  cresce. Dizemos que, no limite, a distribuição de  $\bar{X}$  degenera, concentrando-se em um ponto. Então é conveniente analisar o que ocorre com a distribuição de  $\sqrt{n}\bar{X}$ . O *teorema do limite central* estabelece que, em condições bastante gerais, no limite, quando  $n$  tende a infinito, a distribuição de  $\sqrt{n}\bar{X}$  é uma distribuição normal com média  $\sqrt{n}\mu$  e variância  $\sigma^2$ . Esse é um exemplo de convergência em distribuição, indicando-se

$$\sqrt{n}\bar{X} \xrightarrow{d} N(\sqrt{n}\mu, \sigma^2)$$

Dizemos, então, que a distribuição assintótica de  $\bar{X}$  é uma distribuição normal com média  $\mu$  e variância  $\sigma^2/n$ .

### 1.11. O limite inferior de Cramér-Rao e as propriedades assintóticas dos estimadores de máxima verossimilhança

Consideremos uma amostra aleatória de  $n$  observações  $(X_1, X_2, \dots, X_n)$  de uma variável cuja distribuição é caracterizada por um parâmetro  $\alpha$  cujo valor é desconhecido. Se  $f(X)$  é uma função de densidade de  $\Pi$ , a função de verossimilhança dessa amostra é

$$L(X_1, X_2, \dots, X_n; \alpha) = \prod_{i=1}^n f(X_i)$$

Seja  $a$  um estimador não-tendencioso de  $\alpha$ . Se a função de densidade  $f(X)$  obedecer a certas condições de regularidade relativas à integração e diferenciação e se existe a variância de  $a$ , então pode-se demonstrar que<sup>2</sup>

$$V(a) \geq \frac{1}{-E\left(\frac{d^2 \ln L}{d\alpha^2}\right)} = \frac{1}{E\left(\frac{d \ln L}{d\alpha}\right)^2} \quad (1.19)$$

O valor do 2º membro dessa desigualdade é denominado limite inferior de Cramér-Rao. A desigualdade (1.19) estabelece que não existe estimador não-tendencioso cuja variância seja menor do que o limite inferior de Cramér-Rao.

Para exemplificar, consideremos uma variável  $X$  com distribuição normal de média  $\mu$ , desconhecida, e variância igual a um. Dada uma amostra aleatória com  $n$  observações  $(X_1, X_2, \dots, X_n)$ , a função de verossimilhança é

$$\begin{aligned} L(X_1, X_2, \dots, X_n; \mu) &= \prod_{i=1}^n (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(X_i - \mu)^2\right\} = \\ &= (2\pi)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum(X_i - \mu)^2\right\} \end{aligned}$$

Então

$$\ln L = -\frac{n}{2} \ln 2\pi - \frac{1}{2} \sum(X_i - \mu)^2$$

Segue-se que

$$\frac{d \ln L}{d\mu} = \sum(X_i - \mu)$$

e

---

<sup>2</sup> A demonstração pode ser encontrada em Theil (1971), p. 384-387.

$$\frac{d^2 \ln L}{d\mu^2} = -n$$

De acordo com (1.19), obtemos

$$V(m) \geq \frac{1}{n}$$

onde  $m$  é qualquer estimador não-tendencioso de  $\mu$ . Sabemos que, com  $\sigma^2 = 1$ , a variância de  $\bar{X}$  é igual a  $1/n$ , isto é, a média aritmética dos valores da amostra é um estimador com variância igual ao limite inferior de Cramér-Rao.

Convém ressaltar que há casos nos quais o limite inferior de Cramér-Rao não é atingido, isto é, há casos onde não existe estimador não-tendencioso com variância igual ao limite inferior de Cramér-Rao.

Entretanto, existe um teorema que afirma, em condições bastante gerais, que, se  $\hat{\alpha}$  é o estimador de máxima verossimilhança de  $\alpha$  então  $\hat{\alpha}$  apresenta distribuição assintoticamente normal com média  $\alpha$  e variância igual ao limite inferior de Cramér-Rao, isto é, os estimadores de máxima verossimilhança são consistentes e assintoticamente eficientes.<sup>3</sup>

## 1.12. Teste de hipóteses

Dada uma hipótese de nulidade ( $H_0$ ), define-se como erro tipo I o erro que consiste em rejeitar  $H_0$ , dado que  $H_0$  é verdadeira. Define-se como erro tipo II o erro que consiste em não rejeitar  $H_0$ , dado que  $H_0$  é falsa.

A hipótese da nulidade, quando dada em termos quantitativos, é, necessariamente, uma igualdade.

Usa-se a letra grega  $\alpha$  para indicar a probabilidade de cometer erro tipo I, que é o nível de significância do teste, e a letra grega  $\beta$  para indicar a probabilidade de cometer erro tipo II.

Podemos definir ainda o poder do teste, que é a probabilidade de rejeitar  $H_0$ , dado que  $H_0$  é falsa.

---

<sup>3</sup> A demonstração deste teorema pode ser encontrada em Theil (1971), p. 392-395.

Evidentemente, o poder do teste é igual a  $1 - \beta$ .

Para exemplificar, consideremos 2 tetraedros regulares, feitos de material homogêneo, sendo que um deles tem uma face azul e 3 brancas e o outro tem 2 faces azuis e 2 brancas. Quando esses tetraedros são lançados, o resultado é considerado sucesso se a face em contato com a mesa for azul. Então, a probabilidade de obter sucesso em um lançamento é, para o primeiro tetraedro,  $p = 1/4$  e, para o segundo tetraedro,  $p = 1/2$ .

O número ( $X$ ) de sucessos, obtidos em  $n$  lançamentos de um desses tetraedros é uma variável aleatória discreta com distribuição binomial. A tabela 1.9 apresenta a distribuição de  $X$  para cada um dos dois tetraedros, no caso de  $n = 2$  lançamentos.

TABELA 1.9. Distribuição do número de sucessos obtidos em dois lançamentos, para cada um dos dois tetraedros

$X$	$P(X)$	
	para $p = 1/4$	para $p = 1/2$
0	9/16	1/4
1	6/16	2/4
2	1/16	1/4

Consideremos a seguinte situação: suponhamos que um dos tetraedros (não sabemos qual) foi lançado duas vezes e que fomos informados sobre o número ( $X$ ) de sucessos ( $X$  pode assumir os valores 0, 1 ou 2); com base nessa informação, devemos decidir qual dos dois tetraedros foi utilizado, ou seja, devemos decidir entre

$$H_o : p = 1/4 \quad \text{e} \quad H_A : p = 1/2$$

Para a solução deste problema, devemos proceder a um teste de hipóteses. Então, antes de conhecer o valor assumido por  $X$ , devemos estabelecer a regra de decisão a ser adotada, isto é, devemos estabelecer para que valores de  $X$  devemos rejeitar  $H_o$ . Para este problema podemos estabelecer qualquer uma das quatro regras de decisão que constam na tabela 1.10. Nesta tabela também são dados os valores de  $\alpha$  e  $\beta$ , relativos a cada regra de decisão, e a relação  $\Delta\beta/\Delta\alpha$ , isto é, a razão entre o incremento em  $\beta$  e o incremento em  $\alpha$ , quando se passa de uma regra de decisão para a seguinte.

TABELA 1.10. Valores de  $\alpha$  e  $\beta$  relativos às possíveis regras de decisão e relação

Regra de decisão	$\alpha$	$\beta$	$\Delta\beta/\Delta\alpha$
Nunca rejeitar $H_o$	0	1	-4
Rejeitar $H_o$ se $X = 2$	$1/16 = 0,0625$	$3/4 = 0,75$	$-4/3$
Rejeitar $H_o$ se $X \geq 1$	$7/16 = 0,4375$	$1/4 = 0,25$	$-4/9$
Sempre rejeitar $H_o$	1	0	

Indiquemos por  $\beta = \phi(\alpha)$  a relação funcional decrescente que existe entre  $\alpha$  e  $\beta$ .

A figura 1.4 mostra essa relação para o problema descrito. Neste exemplo, a função  $\beta = \phi(\alpha)$  é descontínua porque o teste de hipótese é baseado em uma variável aleatória discreta. Se o teste de hipótese for baseado em uma variável aleatória contínua, a função  $\beta = \phi(\alpha)$  também será contínua.

Como escolher a regra de decisão, ou seja, como escolher o nível de significância do teste? Isso implica escolher o “ponto ótimo” sobre a função  $\beta = \phi(\alpha)$ .

Admitamos que a probabilidade *a priori* de  $H_o$  ser verdadeira seja  $\theta$  (Essa probabilidade deve ser determinada com base em outras informações que não as que estão sendo utilizadas para fazer o teste). Então, podemos obter, como constam na tabela 1.11, os valores da receita líquida  $U$  (num contexto mais geral, os valores  $U$  seriam os níveis de utilidade) associados a cada uma das 4 situações possíveis (quando a hipótese alternativa é simples), e as respectivas probabilidades.

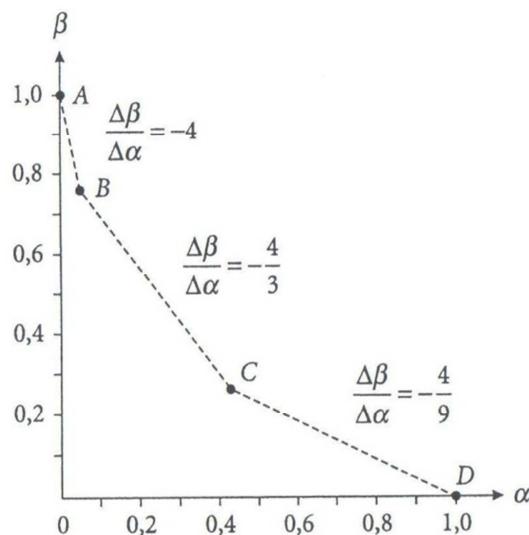


Figura 1.4. Relação entre  $\alpha$  e  $\beta$

TABELA 1.11. A tabela de resultados

Situação real	Decisão tomada	
	não rejeitar $H_o$	rejeitar $H_o$
$H_o$ é verdadeira (probab. = $\theta$ )	$U_{11}$ $p_{11} = \theta(1 - \alpha)$	$U_{12}$ $p_{12} = \theta\alpha$
$H_o$ é falsa (probab. = $1 - \theta$ )	$U_{21}$ $p_{21} = (1 - \theta)\beta$	$U_{22}$ $p_{22} = (1 - \theta)(1 - \beta)$

Se todas essas informações estivessem disponíveis, poderíamos escolher o nível de significância que maximiza a receita líquida esperada, dada por

$$L = E(U) = \theta(1 - \alpha)U_{11} + \theta\alpha U_{12} + (1 - \theta)\beta U_{21} + (1 - \theta)(1 - \beta)U_{22} \quad (1.20)$$

Essa relação pode ser escrita

$$\beta = \frac{\theta U_{11} + (1 - \theta)U_{22} - L}{(1 - \theta)(U_{22} - U_{21})} - \frac{\theta(U_{11} - U_{12})}{(1 - \theta)(U_{22} - U_{21})}\alpha \quad (1.21)$$

A diferença  $U_{11} - U_{12} = C_I > 0$  representa o custo de cometer erro tipo I e a diferença  $U_{22} - U_{21} = C_{II} > 0$  representa o custo de cometer erro tipo II.

Dados os valores de  $\theta$ ,  $U_{11}$ ,  $U_{12}$ ,  $U_{21}$ ,  $U_{22}$ , a relação (1.21) corresponde a um feixe de retas paralelas num sistema de eixos cartesianos com coordenadas  $\alpha$  e  $\beta$ . O coeficiente angular é sempre igual a

$$-\frac{\theta C_I}{(1 - \theta)C_{II}} \quad (1.22)$$

e o coeficiente linear é tanto menor quanto maior for o valor de  $L = E(U)$ . Para maximizar  $L = E(U)$  devemos determinar o ponto de  $\beta = \phi(\alpha)$  que pertença a uma reta com declividade dada por (1.22) e coeficiente linear mínimo.

Para exemplificar, consideremos a relação  $\beta = \phi(\alpha)$  representada na figura 1.4 e admitamos que  $\theta = 0,5$ . Neste caso, temos:

- a) se  $4 < \frac{C_I}{C_{II}} < \infty$ , o ponto ótimo é  $A$ , isto é, nunca devemos rejeitar  $H_o$ .
- b) se  $\frac{C_I}{C_{II}} = 4$ , é indiferente utilizar a regra de decisão correspondente ao ponto  $A$  ou ao ponto  $B$ .
- c) se  $\frac{4}{3} < \frac{C_I}{C_{II}} < 4$ , o ponto ótimo é  $B$ , isto é, devemos rejeitar  $H_o$  se  $X = 2$ , fazendo um teste com nível de significância  $\alpha = 0,0625$ .
- d) se  $\frac{C_I}{C_{II}} = \frac{4}{3}$ , é indiferente utilizar a regra de decisão correspondente ao ponto  $B$  ou ao ponto  $C$ .
- e) se  $\frac{4}{9} < \frac{C_I}{C_{II}} < \frac{4}{3}$ , o ponto ótimo é  $C$ , isto é, devemos rejeitar  $H_o$  se  $X \geq 1$ , fazendo um teste com nível de significância  $\alpha = 0,4375$ .
- f) se  $\frac{C_I}{C_{II}} = \frac{4}{9}$ , é indiferente utilizar a regra de decisão correspondente ao ponto  $C$  ou ao ponto  $D$ , e
- g) se  $0 < \frac{C_I}{C_{II}} < \frac{4}{9}$ , o ponto ótimo é  $D$ , isto é, devemos rejeitar  $H_o$  sempre, qualquer que seja o valor observado de  $X$ .

Se a função  $\beta = \phi(\alpha)$  for contínua, o ponto que maximiza a receita líquida pode ser determinado igualando a zero a derivada de  $L$  em relação a  $\alpha$ .

De (1.20) obtemos

$$\begin{aligned} \frac{dL}{d\alpha} &= -\theta(U_{11} - U_{12}) - (1-\theta)(U_{22} - U_{21}) \frac{d\beta}{d\alpha} = \\ &= -\theta C_I - (1-\theta)C_{II} \frac{d\beta}{d\alpha} \end{aligned} \quad (1.23)$$

Segue-se que  $\frac{dL}{d\alpha} = 0$  implica

$$\frac{d\beta}{d\alpha} = -\frac{\theta C_I}{(1-\theta)C_{II}} \quad (1.24)$$

O ponto de  $\beta = \phi(\alpha)$  que satisfaz essa condição corresponde a um máximo de  $L = E(U)$  se

$$\frac{d^2 L}{d\alpha^2} < 0$$

De (1.23) obtemos

$$\frac{d^2 L}{d\alpha^2} = -(1 - \theta)C_{II} \frac{d^2 \beta}{d\alpha^2},$$

mostrando que a condição de segunda ordem para máximo é satisfeita se  $\frac{d^2 \beta}{d\alpha^2} > 0$ , isto é, se a função  $\beta = \phi(\alpha)$  for convexa em relação à origem.

Sendo  $\beta = \phi(\alpha)$  uma função decrescente e convexa em relação à origem, o nível de significância ótimo estabelecido através de (1.24) será tanto menor quanto maior for  $\theta$  (a probabilidade *a priori* de  $H_0$  ser verdadeira) e quanto maior for a relação  $\frac{C_I}{C_{II}}$  (o custo de cometer erro tipo I em comparação com o custo de cometer erro tipo II).

Em problemas práticos é geralmente impossível determinar o nível de significância ótimo da maneira indicada, porque não se tem nem a probabilidade ( $\theta$ ) de  $H_0$  ser verdadeira *a priori*, nem o valor exato da relação  $\frac{C_I}{C_{II}}$ . Além disso, a hipótese alternativa é, geralmente, composta; a determinação rigorosa de um nível de significância ótimo exigiria, neste caso, o conhecimento da distribuição *a priori* dos valores possíveis para a hipótese alternativa, com os respectivos valores do custo de cometer erro tipo II.

Por isso, a escolha do nível de significância tem muito de arbitrário.

A finalidade da discussão feita é deixar claro o sentido em que deve ser ajustado o nível de significância conforme mudem a probabilidade *a priori* de  $H_0$  ser verdadeira e a relação entre os custos de cometer erro tipo I e erro tipo II.

É usual que a hipótese alternativa não se refira a um valor específico. É comum, por exemplo, testar se um parâmetro  $\gamma$  é igual a zero ( $H_0 : \gamma = 0$ ) contra a hipótese alternativa de que é diferente de zero ( $H_A : \gamma \neq 0$ ). Neste caso pode-se fixar o nível de

significância do teste ( $\alpha$ ), mas o poder do teste ( $1 - \beta$ ) não é um valor único. Pode-se construir a curva de poder do teste, que mostra como esse varia em função de valores alternativos do parâmetro. É claro que o poder do teste se aproxima do nível de significância quando o valor alternativo do parâmetro se aproxima do valor estabelecido pela hipótese da nulidade, fazendo com que, fixado um baixo nível de significância, o poder do teste seja baixo para tais valores alternativos do parâmetro. Note-se como, nestas condições, não há simetria entre as decisões de “rejeitar” e “aceitar” a hipótese da nulidade. Ao rejeitar a hipótese da nulidade estaremos tomando uma decisão de maneira que a probabilidade de estar cometendo erro (tipo I) é conhecida e pequena. Mas se o resultado do teste é não-significativo e “aceitamos” a hipótese da nulidade, a probabilidade de cometer erro tipo II é desconhecida e tende a ser elevada para valores do parâmetro próximos ao estabelecido pela hipótese da nulidade. A linguagem usada na interpretação do resultado de um teste de hipóteses deve refletir essa assimetria. Se, ao testar  $(H_0 : \gamma = 0)$  contra  $(H_A : \gamma \neq 0)$ , o resultado do teste é significativo, *rejeitamos* a hipótese da nulidade. Se o resultado for não-significativo, a conclusão é que os dados da amostra utilizada não permitem rejeitar a hipótese da nulidade. Note-se a natureza “provisória” da conclusão. A afirmativa de que “aceita-se  $H_0$ ” não reflete adequadamente a indeterminação da probabilidade de cometer erro tipo II quando a hipótese alternativa é composta (não estabelece um único valor alternativo para o parâmetro).

## Exercícios

- 1.1. Seja  $X$  o resultado obtido no lançamento de um dado (hexaedro regular) não-chumbado. Seja  $Y$  a soma dos resultados obtidos em 100 lançamentos desse dado. Determine  $E(X)$ ,  $V(X)$ ,  $E(Y)$  e  $V(Y)$ .

- 1.2. Com base na distribuição conjunta de  $X$  e  $Y$ , apresentada na tabela ao lado, determine a  $E(X)$ , a  $E(Y)$ , a  $V(X)$ , a  $V(Y)$  e a  $\text{cov}(X, Y)$ . As variáveis  $X$  e  $Y$  são independentes?

Valores de  $P(X_i, Y_j)$  para a distribuição conjunta das variáveis  $X_i$  e  $Y_j$ .

$Y_j$	$X_i$		
	1	2	3
4	0,3	0	0,3
8	0	0,4	0

- 1.3. A tabela ao lado mostra a distribuição conjunta de  $X$  e  $Y$ .

- a) Essas variáveis são independentes? (Justifique sua resposta).
- b) Determine  $E(X)$  e  $E(Y)$ .
- c) Determine  $V(X)$  e  $V(Y)$ .
- d) Determine  $\text{cov}(X, Y)$  e a correlação ( $\rho$ ) entre as duas variáveis.

Valores de  $P(X_i, Y_j)$  para a distribuição conjunta das variáveis  $X_i$  e  $Y_j$ .

$Y_j$	$X_i$		
	2	4	6
4	0,2	0,1	0
5	0,1	0,2	0,1
6	0	0,1	0,2

- 1.4. Temos duas urnas, aparentemente idênticas, com 63 bolas no interior de cada uma.

Essas bolas são marcadas com números ( $X$ ) de zero a 5. Na urna  $A$  há  $2^x$  bolas com o número  $X$ , isto é, há uma bola com o nº 0, duas bolas com o nº 1, 4 bolas com o nº 2, e assim por diante, até 32 bolas com o nº 5. Na urna  $B$  há  $2^{5-x}$  bolas com o número  $X$ , isto é, há 32 bolas com o nº 0, 16 bolas com o nº 1, 8 bolas com o nº 2, e assim por diante, até uma bola com o nº 5. Uma dessas urnas, escolhida ao acaso, é entregue a um estatístico, que deve decidir se é a urna  $A$  ou se é a urna  $B$ , retirando, ao acaso, uma única bola da urna. Ele especifica a hipótese da nulidade como

$$H_0 : \text{trata-se da urna } A$$

e a hipótese alternativa como

$$H_A : \text{trata-se da urna } B$$

O estatístico decide, também; que a regra de decisão será rejeitar  $H_0$  (em favor de  $H_A$ ) se a bola retirada da urna apresentar número menor do que 3. Determine: (a) o nível de significância do teste; (b) a probabilidade ( $\beta$ ) de cometer erro tipo II; (c) o poder do teste.

Refaça o problema considerando, agora, que a regra de decisão é rejeitar  $H_0$  se o número ( $X$ ) marcado na bola retirada for menor ou igual a 1.

- 1.5. Temos duas urnas, aparentemente idênticas, com 55 bolas no interior de cada uma. Na urna  $A$  há uma bola com o nº 0, duas bolas com o nº 1, 3 bolas com o nº 2, e assim por diante, até 10 bolas com o nº 9. Na urna  $B$  há 1 bola com o nº 9, 2 bolas com o nº 8, 3 bolas com o nº 7, e assim por diante, até 10 bolas com o nº 0. Uma dessas urnas, escolhida ao acaso, é entregue a um estatístico, que deve decidir se é a urna  $A$  ou se é a urna  $B$  examinando uma única bola retirada da urna, ao acaso. Ele especifica a hipótese da nulidade como

$$H_0 : \text{trata-se da urna } A$$

e a hipótese alternativa como

$$H_A : \text{trata-se da urna } B$$

O estatístico adota a seguinte regra de decisão: rejeitar  $H_0$  (em favor de  $H_A$ ) se a bola retirada da urna apresentar número menor do que 5. Determine:

- a) o nível de significância do teste
- b) a probabilidade ( $\beta$ ) de cometer erro tipo II
- c) o poder do teste.

Refaça o problema considerando, agora, que a regra de decisão é rejeitar  $H_0$  se o número marcado na bola retirada for menor ou igual a 3.

- 1.6. Temos dois tetraedros regulares de material homogêneo. Um deles tem uma face azul e três faces brancas. O outro tem três faces azuis e uma branca. Uma pessoa pega, ao acaso, um desses tetraedros e o lança  $n$  vezes. Seja  $X$  o número de vezes em que o resultado foi “face azul”. Com base no valor de  $X$  devemos testar a hipótese

$$H_0 : \text{“foi utilizado o tetraedro com uma face azul”}$$

contra a hipótese alternativa

$$H_A : \text{“foi utilizado o tetraedro com três faces azuis”}$$

Seja  $\alpha$  o nível de significância do teste e seja  $\beta$  a probabilidade de cometer erro tipo II.

- a) Considerando as diferentes regras de decisão, faça uma tabela e um gráfico mostrando como  $\beta$  varia em função de  $\alpha$  para  $n = 3$ .
- b) Qual é o nível de significância para um teste com  $n = 5$ , mantendo  $\beta = \alpha$ ?

### Respostas

1.1.  $E(X) = 3,5$ ,  $V(X) = \frac{17,5}{6} = 2,9167$ ,  $E(Y) = 350$  e  $V(Y) = \frac{1750}{6} = 291,67$

1.2.  $E(X) = 2$ ,  $E(Y) = 5,6$ ,  $V(X) = 0,6$ ,  $V(Y) = 3,84$ ,  $\text{cov}(X, Y) = 0$ . As variáveis  $X$  e  $Y$  não são independentes.

- 1.3. a) Não      b)  $E(X) = 4$  e  $E(Y) = 5$       b)  $V(X) = 2,4$  e  $V(Y) = 0,6$   
c)  $\text{cov}(X, Y) = 0,8$  e  $\rho = 0,667$ .

1.4. Para a regra de decisão “Rejeitar  $H_0$  se  $X < 3$ ” obtemos  $\alpha = 7/63 = 1/9 = 0,111$ ,  
 $\beta = 1/9 = 0,111$  e  $1 - \beta = 8/9 = 0,889$

Para a regra de decisão “Rejeitar  $H_0$  se  $X \leq 1$ ” obtemos  $\alpha = 3/63 = 1/21 = 0,0476$ ,  
 $\beta = 15/63 = 5/21 = 0,238$  e  $1 - \beta = 16/21 = 0,762$ .

1.5. Rejeitar  $H_0$  se número < 5:  $\alpha = 3/11$ ,  $\beta = 3/11$  e  $1 - \beta = 8/11$ .

Rejeitar  $H_0$  se número  $\leq 3$ :  $\alpha = 2/11$ ,  $\beta = 21/55$  e  $1 - \beta = 34/55$ .

1.6.

---

a) Regra de decisão:  
rejeitar  $H_0$  :  $p = 1/4$  se

	$\alpha$	$\beta$
$X \geq 0$ (sempre)	1	0
$X \geq 1$	$37/64$	$1/64$
$X \geq 2$	$10/64$	$10/64$
$X \geq 3$	$1/64$	$37/64$
$X > 3$ (nunca)	0	1

---

b)  $\alpha = \beta = \frac{53}{512} = 0,1035$

## 2. REGRESSÃO LINEAR SIMPLES

### 2.1. O modelo estatístico de uma regressão linear simples

Dados  $n$  pares de valores de duas variáveis,  $X_i$ ,  $Y_i$  (com  $i = 1, 2, \dots, n$ ), se admitirmos que  $Y$  é função linear de  $X$ , podemos estabelecer uma regressão linear simples, cujo modelo estatístico é

$$Y_i = \alpha + \beta X_i + u_i ,$$

onde  $\alpha$  e  $\beta$  são parâmetros,  $X$  é a variável explanatória e  $Y$  é a variável dependente.

O coeficiente angular da reta ( $\beta$ ) é também denominado coeficiente de regressão e o coeficiente linear da reta ( $\alpha$ ) é também conhecido como termo constante da equação de regressão.

A análise de regressão também pode ser aplicada às relações não-lineares. Inicialmente, estudaremos apenas o caso da reta. Veremos adiante o caso das relações não-lineares.

Ao estabelecer o modelo de regressão linear simples, pressupomos que:

- I) A relação entre  $X$  e  $Y$  é linear.
- II) Os valores de  $X$  são fixos, isto é,  $X$  não é uma variável aleatória.
- III) A média do erro é nula, isto é,  $E(u_i) = 0$ .
- IV) Para um dado valor de  $X$ , a variância do erro  $u$  é sempre  $\sigma^2$ , denominada variância residual, isto é,

$$E(u_i^2) = \sigma^2$$

ou

$$E[Y_i - E(Y_i | X_i)]^2 = \sigma^2$$

Dizemos, então, que o erro é homocedástico ou que temos homocedasticia (do erro ou da variável dependente).

- V) O erro de uma observação é não-correlacionado com o erro em outra observação, isto é,  $E(u_i u_j) = 0$  para  $i \neq j$ .

VI) Os erros têm distribuição normal.

Combinando as pressuposições III, IV e VI, temos que

$$u_i \sim N(0, \sigma^2)$$

Devemos, ainda, verificar se o número de observações disponíveis é maior do que o número de parâmetros da equação de regressão. Para ajustar uma regressão linear simples precisamos ter, no mínimo, 3 observações. Se só dispomos de 2 observações (2 pontos), a determinação da reta é um problema de geometria analítica; não é possível, neste caso, fazer nenhuma análise estatística.

Veremos adiante que as pressuposições I, II e III são necessárias para que se possa demonstrar que as estimativas dos parâmetros obtidas pelo método dos mínimos quadrados são não-tendenciosas ou imparciais, isto é, que

$$E(a) = \alpha$$

e

$$E(b) = \beta$$

onde  $a$  e  $b$  são as estimativas de mínimos quadrados de  $\alpha$  e  $\beta$ , respectivamente.

Veremos também que, com base nas cinco primeiras pressuposições, é possível demonstrar que as estimativas dos parâmetros obtidas pelo método dos mínimos quadrados são estimativas lineares não-tendenciosas de variância mínima.

É interessante assinalar que a pressuposição II não é, na verdade, essencial. Veremos, no fim deste capítulo, que em certas condições, se  $X$  for uma variável aleatória, os resultados obtidos pressupondo que os valores de  $X$  são fixos continuam válidos. Entretanto, tendo em vista a simplicidade das demonstrações de vários teoremas, essa pressuposição será adotada durante a maioria das seções do capítulo. Devemos observar que, se os pares de valores  $X_i, Y_i$  (com  $i = 1, 2, \dots, n$ ) foram obtidos experimentalmente e  $X$  é uma variável controlada (fixada) pelo pesquisador, a pressuposição II é válida.

A pressuposição III exclui, por exemplo, a existência de erros sistemáticos de medida da variável  $Y$ .

Quando não é razoável supor que os erros são homocedásticos (pressuposição IV), isto é, quando existe heterocedasticidade, devemos utilizar o método dos mínimos quadrados ponderados, que será examinado no capítulo 6.

Na figura 2.1 está representado o modelo estatístico de uma regressão linear simples, considerando as pressuposições de I a IV. As pressuposições I, II e III permitem escrever

$$E(Y_i) = \alpha + \beta X_i,$$

ou seja, as médias das distribuições de  $Y | X$  estão sobre a reta  $\alpha + \beta X$ . À pressuposição IV corresponde, na figura 2.1, o fato de as distribuições de  $Y$  para diferentes valores de  $X$  apresentarem todas a mesma dispersão.

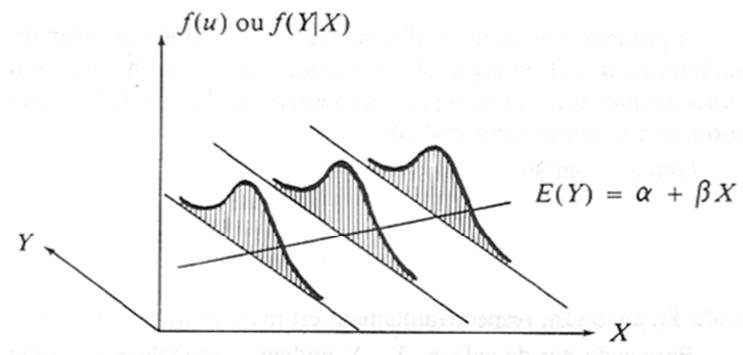


Figura 2.1. Representação do modelo estatístico de uma regressão linear simples.

Se os pares de valores  $X_i, Y_i$  foram obtidos através de amostragem aleatória de uma população infinita, fica garantida a independência entre observações. Se, além disso, a esperança do erro é igual a zero, temos, com  $i \neq j$ ,

$$E(u_i u_j) = E(u_i) \cdot E(u_j) = 0$$

Entretanto, a pressuposição V geralmente não é obedecida quando trabalhamos com séries cronológicas de dados. Dizemos, então, que há autocorrelação nos resíduos. Temos autocorrelação positiva se  $E(u_i u_{i+1}) > 0$  e autocorrelação negativa se  $E(u_i u_{i+1}) < 0$ . No capítulo 7 veremos os métodos que podem ser utilizados quando há autocorrelação nos resíduos.

A pressuposição VI é necessária para que possamos utilizar as distribuições de  $t$  e de  $F$  para testar hipóteses a respeito dos valores dos parâmetros ou construir intervalos de confiança. Em alguns casos, é possível justificar essa pressuposição com base no teorema do limite central. Esse teorema, na sua versão mais geral, estabelece que a soma de um grande número de variáveis aleatórias independentes tem distribuição aproximadamente normal, desde que nenhuma delas seja dominante. Vimos que o erro ( $u_i$ ) do modelo estatístico de uma regressão linear pode ser devido à influência de todas as variáveis que afetam a variável dependente e que não foram incluídas no modelo. Uma vez que as variáveis que não foram consideradas devem ser as menos importantes, seus efeitos devem ser todos relativamente pequenos. Considerando que o número de fatores que podem afetar certa variável dependente é bastante grande, e desde que seus efeitos sejam aditivos e independentes, podemos concluir, com base no teorema do limite central, que o erro residual tem distribuição aproximadamente normal.

## 2.2. Estimativa dos parâmetros

O primeiro passo, na análise de regressão, é obter as estimativas  $a$  e  $b$  dos parâmetros  $\alpha$  e  $\beta$  da regressão. Os valores dessas estimativas serão obtidos a partir de uma amostra de  $n$  pares de valores  $X_i$ ,  $Y_i$  (com  $i = 1, 2, \dots, n$ ), que correspondem a  $n$  pontos num gráfico.

Obtemos, então

$$\hat{Y}_i = a + bX_i$$

onde  $\hat{Y}_i$ ,  $a$  e  $b$  são, respectivamente estimativas de  $E(Y_i) = \alpha + \beta X_i$ ,  $\alpha$  e  $\beta$ .

Para cada par de valores  $X_i$ ,  $Y_i$  podemos estabelecer o desvio

$$e_i = Y_i - \hat{Y}_i = Y_i - (a + bX_i)$$

O método dos mínimos quadrados consiste em adotar como estimativas dos parâmetros os valores que minimizam a soma dos quadrados dos desvios

$$Z = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n [Y_i - (a + bX_i)]^2$$

A função  $Z$  terá mínimo quando suas derivadas parciais em relação a  $a$  e  $b$  forem nulas:

$$\frac{\partial Z}{\partial a} = -2 \sum [Y_i - (a + bX_i)] = 0 \quad (2.1)$$

$$\frac{\partial Z}{\partial b} = 2 \sum [Y_i - (a + bX_i)] (-X_i) = 0 \quad (2.2)$$

Por se tratar de uma soma de quadrados de desvios, o ponto extremo será necessariamente um ponto de mínimo da função. Formalmente, pode-se verificar que as condições de segunda ordem para mínimo são satisfeitas.

Simplificando as equações (2.1) e (2.2), chegamos ao sistema de *equações normais*

$$\begin{cases} na + b \sum X_i = \sum Y_i \\ a \sum X_i + b \sum X_i^2 = \sum X_i Y_i \end{cases} \quad (2.3)$$

$$a \sum X_i + b \sum X_i^2 = \sum X_i Y_i \quad (2.4)$$

Resolvendo o sistema, obtemos:

$$a = \frac{(\sum X^2)(\sum Y) - (\sum X)(\sum XY)}{n(\sum X^2) - (\sum X)^2}$$

$$b = \frac{n \sum XY - (\sum X)(\sum Y)}{n \sum X^2 - (\sum X)^2}$$

Note que, por simplicidade, omitimos o índice. Para sermos explícitos deveríamos escrever, por exemplo,  $\sum_{i=1}^n X_i$ , onde escrevemos, simplesmente,  $\sum X$ .

Na prática determinamos  $b$  em primeiro lugar e da equação (2.3) obtemos

$$a = \frac{\sum Y}{n} - b \frac{\sum X}{n}$$

ou

$$a = \bar{Y} - b \bar{X}$$

É fácil verificar que a fórmula para o cálculo de  $b$  pode ser escrita de diversos modos, quais sejam:

$$\begin{aligned}
 b &= \frac{n \sum XY - (\sum X)(\sum Y)}{n \sum X^2 - (\sum X)^2} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} = \\
 &= \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2} = \frac{\sum(X - \bar{X})Y}{\sum(X - \bar{X})^2} = \\
 &= \frac{\sum X(Y - \bar{Y})}{\sum(X - \bar{X})^2} = \frac{\sum xy}{\sum x^2} = \frac{\sum xY}{\sum x^2} = \frac{\sum Xy}{\sum x^2}
 \end{aligned}$$

onde

$$\bar{X} = \frac{\sum X}{n}, \quad \bar{Y} = \frac{\sum Y}{n}, \quad x = X - \bar{X} \quad \text{e} \quad y = Y - \bar{Y}$$

Assinalemos duas relações bastante úteis que podem ser obtidas a partir das equações (2.1) e (2.2). Lembrando que

$Y_i - (a + bX_i) = Y_i - \hat{Y}_i = e_i$ , tais equações ficam:

$$\sum e_i = 0 \tag{2.5}$$

e

$$\sum X_i e_i = 0 \tag{2.6}$$

Temos, também, que

$$\sum \hat{Y}_i e_i = \sum(a + bX_i)e_i = a \sum e_i + b \sum X_i e_i$$

De acordo com (2.5) e (2.6), concluímos que

$$\sum \hat{Y}_i e_i = 0 \tag{2.7}$$

As relações (2.5), (2.6) e (2.7) mostram, respectivamente, que

- a) a soma dos desvios é igual a zero,
- b) a soma dos produtos dos desvios pelos correspondentes valores da variável independente é igual a zero, e
- c) a soma dos produtos dos desvios pelos respectivos valores estimados da variável dependente é igual a zero.

Estas relações podem ser utilizadas para verificar se as estimativas dos parâmetros foram corretamente calculadas e para verificar o efeito dos erros de arredondamento.

Como  $Y_i = \hat{Y}_i + e_i$ , de (2.5) concluímos que

$$\frac{\sum Y_i}{n} = \frac{\sum \hat{Y}_i}{n} = \bar{Y}, \quad (2.8)$$

isto é, a média dos valores observados de  $Y$  é igual à média dos valores estimados de  $Y$ .

### 2.3. O modelo simplificado e um exemplo numérico

Uma simplificação conveniente dos cálculos é obtida quando usamos a variável centrada  $x_i = X_i - \bar{X}$ . Na representação gráfica, isso corresponde a tomar a média da variável  $X_i$  como origem do eixo das abscissas. Nesse caso, o modelo estatístico fica

$$Y_i = A + \beta x_i + u_i$$

Representando por  $\hat{A}$  a estimativa de mínimos quadrados do parâmetro  $A$ , temos

$$Y_i = \hat{A} + b x_i + e_i$$

Como  $\sum x_i = 0$ , as equações normais ficam

$$\begin{cases} n\hat{A} &= \sum Y_i \\ b \sum x_i^2 &= \sum x_i Y_i \end{cases}$$

Donde

$$\hat{A} = \frac{\sum Y}{n} = \bar{Y}$$

e

$$b = \frac{\sum xy}{\sum x^2}$$

Então a reta de regressão estimada é

$$\hat{Y}_i = \bar{Y} + bx_i$$

ou

$$\hat{y}_i = bx_i \quad (2.9)$$

onde

$$\hat{y}_i = \hat{Y}_i - \bar{Y}$$

Temos

$$\sum \hat{y}_i e_i = \sum (\hat{Y}_i - \bar{Y}) e_i = \sum \hat{Y}_i e_i - \bar{Y} \sum e_i$$

Lembrando (2.5) e (2.7) concluímos que

$$\sum \hat{y}_i e_i = 0 \quad (2.10)$$

Para exemplificar, consideremos a amostra de 10 pares de valores  $X_i$ ,  $Y_i$  da tabela 2.1, representados graficamente na figura 2.2.

TABELA 2.1. Valores de  $X_i$  e  $Y_i$  ( $i = 1, \dots, 10$ )

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

Os números apresentados são artificiais e foram escolhidos de maneira a simplificar os cálculos. O estudante de economia pode imaginar, por exemplo, que  $Y$  é o custo total de produção correspondente à quantidade produzida  $X$ , para empresas de certa indústria; esse estudante poderá verificar que, neste caso,  $\alpha$  representa o custo fixo (custo existente mesmo quando  $X = 0$ ) e  $\beta$  representa o custo marginal. Também é possível imaginar que  $Y$  é a quantidade de algum produto oferecida em certo mercado, e

$X$  é o respectivo preço, ou ainda, que  $Y$  é o logaritmo do consumo semanal de carne de uma família e  $X$  é o logaritmo da renda mensal dessa família.

São dados, a seguir, os resultados de alguns cálculos intermediários para a obtenção das estimativas  $a$  e  $b$ .

$$\sum X = 30, \quad \bar{X} = 3$$

$$\sum X^2 = 126$$

$$\sum x^2 = \sum X^2 - \frac{(\sum X)^2}{n} = 126 - 90 = 36$$

$$\sum Y = 50, \quad \bar{Y} = 5$$

$$\sum XY = 186$$

$$\sum xy = \sum XY - \frac{(\sum X)(\sum Y)}{n} = 186 - 150 = 36$$

Destes resultados obtemos

$$b = \frac{\sum xy}{\sum x^2} = \frac{36}{36} = 1$$

$$a = \bar{Y} - b\bar{X} = 5 - 3 = 2$$

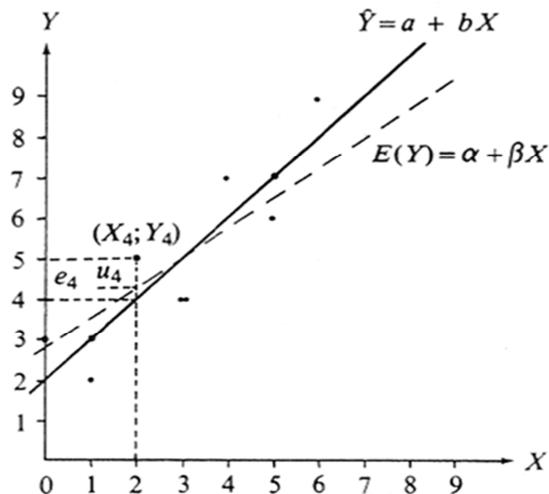


Figura 2.2. Representação gráfica dos pares de valores da tabela 2.1, a reta ajustada ( $\hat{Y} = a + bX$ ) e a reta verdadeira [ $E(Y) = \alpha + \beta X$ ].

A reta de regressão estimada é

$$\hat{Y} = 5 + x$$

ou

$$\hat{Y} = 2 + X$$

## 2.4. Demonstração de que os estimadores de mínimos quadrados são estimadores lineares não-tendenciosos

Demonstraremos, inicialmente, que  $b = \frac{\sum x_i Y}{\sum x_i^2}$  é um estimador linear não-tendencioso de  $\beta$ .

Temos que

$$b = \frac{\sum x_i Y}{\sum x_i^2} = \frac{x_1}{\sum x_i^2} Y_1 + \frac{x_2}{\sum x_i^2} Y_2 + \dots + \frac{x_n}{\sum x_i^2} Y_n$$

Uma vez que os valores de  $X_i$  são fixos, de acordo com a pressuposição II,  $\frac{x_i}{\sum x_i^2}$  são, também, valores fixos. Então,  $b$  é uma combinação linear dos valores de  $Y_i$ .

Como

$$Y_i = \alpha + \beta X_i + u_i,$$

obtemos

$$\begin{aligned} b &= \frac{1}{\sum x_i^2} \sum x_i (\alpha + \beta X_i + u_i) = \\ &= \frac{1}{\sum x_i^2} (\alpha \sum x_i + \beta \sum x_i X_i + \sum x_i u_i) \end{aligned}$$

Como  $\sum x_i = 0$  e  $\sum x_i^2 = \sum x_i X_i$ ,

$$b = \beta + \frac{\sum x_i u_i}{\sum x_i^2} \tag{2.11}$$

Lembrando que, de acordo com a pressuposição III,  $E(u_i) = 0$ , obtemos

$$E(b) = \beta$$

Isto é,  $b = \frac{\sum x_i Y}{\sum x_i^2}$  é um estimador não-tendenciosos ou imparcial de  $\beta$ .

Demonstraremos, agora, que

$$a = \bar{Y} - b\bar{X}$$

é um estimador linear não-tendenciosos de  $\alpha$ .

Temos que

$$a = \bar{Y} - b\bar{X} = \frac{\sum Y_i}{n} - \bar{X} \frac{\sum x_i Y}{\sum x_i^2} = \sum \left( \frac{1}{n} - \frac{\bar{X}x_i}{\sum x_i^2} \right) Y_i$$

Essa última expressão mostra que  $a$  é uma função linear dos valores de  $Y_i$ .

Como

$$Y_i = \alpha + \beta X_i + u_i,$$

obtemos

$$\begin{aligned} a &= \sum \left( \frac{1}{n} - \frac{\bar{X}x_i}{\sum x_i^2} \right) (\alpha + \beta X_i + u_i) = \\ &= \alpha - \frac{\alpha \bar{X} \sum x_i}{\sum x_i^2} + \beta \frac{\sum X_i}{n} - \beta \frac{\bar{X} \sum x_i X_i}{\sum x_i^2} + \sum \left( \frac{1}{n} - \frac{\bar{X}x_i}{\sum x_i^2} \right) u_i \end{aligned}$$

e, como  $\sum x_i = 0$  e  $\sum x_i^2 = \sum x_i X_i$ ,

$$a = \alpha + \sum \left( \frac{1}{n} - \frac{\bar{X}x_i}{\sum x_i^2} \right) u_i \quad (2.12)$$

Lembrando que  $E(u_i) = 0$ , obtemos

$$E(a) = \alpha$$

Isto é,  $a = \bar{Y} - b\bar{X}$  é um estimador não-tendencioso de  $\alpha$ .

## 2.5. Variâncias e covariâncias das estimativas dos parâmetros

Determinaremos, inicialmente, a expressão da variância de  $b$ .

Como  $E(b) = \beta$ , por definição temos

$$V(b) = E(b - \beta)^2 \quad (2.13)$$

De (2.11), obtemos

$$b - \beta = \frac{\sum x_i u_i}{\sum x_i^2}$$

Substituindo esse resultado em (2.13), obtemos

$$V(b) = \frac{E(\sum x_i u_i)^2}{(\sum x_i^2)^2}$$

Mas

$$\begin{aligned} E(\sum x_i u_i)^2 &= E(x_1 u_1 + x_2 u_2 + \dots + x_n u_n)^2 = \\ &= E(x_1^2 u_1^2 + x_2^2 u_2^2 + \dots + x_n^2 u_n^2 + 2x_1 x_2 u_1 u_2 + \\ &\quad + \dots + 2x_1 x_n u_1 u_n + \dots) = \\ &= x_1^2 \sigma^2 + x_2^2 \sigma^2 + \dots + x_n^2 \sigma^2 = \\ &= \sigma^2 \sum x_i^2 \end{aligned} \quad (2.14)$$

uma vez que, de acordo com as pressuposições IV e V,  $E(u_i^2) = \sigma^2$  e  $E(u_i u_j) = 0$ , para  $i \neq j$ .

Então

$$V(b) = \frac{\sigma^2 \sum x_i^2}{(\sum x_i^2)^2} = \frac{\sigma^2}{\sum x_i^2} \quad (2.15)$$

Determinemos, a seguir, a variância de  $a$

De (2.12) segue-se que

$$a - \alpha = \sum \left( \frac{1}{n} - \frac{\bar{X} x_i}{\sum x_i^2} \right) u_i$$

Então

$$V(a) = E(a - \alpha)^2 = E\left[\sum\left(\frac{1}{n} - \frac{\bar{X}x_i}{\sum x_i^2}\right)u_i\right]^2$$

Lembrando que  $E(u_i^2) = \sigma^2$  e  $E(u_i u_j) = 0$ , para  $i \neq j$ , podemos obter

$$\begin{aligned} V(a) &= \sum\left(\frac{1}{n} - \frac{\bar{X}x_i}{\sum x_i^2}\right)^2 \sigma^2 = \\ &= \sum\left(\frac{1}{n^2} + \frac{\bar{X}^2 x_i^2}{(\sum x_i^2)^2} - \frac{2\bar{X}x_i}{n \sum x_i^2}\right) \sigma^2 = \\ &= \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2}\right) \sigma^2 \end{aligned} \tag{2.16}$$

Notando que

$$\sum x_i^2 = \sum X_i^2 - \frac{(\sum X_i)^2}{n} = \sum X_i^2 - n\bar{X}^2$$

também podemos obter

$$V(a) = \frac{\sum X_i^2}{n \sum x_i^2} \sigma^2 \tag{2.17}$$

Antes de deduzir a expressão da covariância entre  $a$  e  $b$ , determinemos as variâncias e covariâncias das estimativas dos parâmetros do modelo simplificado analisado na seção 2.3. Vimos ali que essas estimativas são

$$\hat{A} = \bar{Y} \quad \text{e} \quad b = \frac{\sum xY}{\sum x^2}$$

Temos que

$$\bar{Y} = \frac{\sum Y_i}{n} = \frac{\sum(\alpha + \beta X_i + u_i)}{n} = \alpha + \beta \bar{X} + \bar{u}$$

onde  $\bar{u} = (\sum u_i)/n$

Como  $E(\bar{u}) = 0$ , podemos escrever

$$E(\bar{Y}) = \alpha + \beta \bar{X}$$

Então, é fácil verificar que

$$\bar{Y} - E(\bar{Y}) = \bar{u}, \quad (2.18)$$

onde obtemos

$$V(\bar{Y}) = E[\bar{Y} - E(\bar{Y})]^2 = E(\bar{u}^2) = E\left(\frac{\sum u_i}{n}\right)^2$$

e, finalmente,

$$V(\bar{Y}) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \quad (2.19)$$

Determinemos agora a covariância entre  $\bar{Y}$  e  $b$

$$\text{cov}(\bar{Y}, b) = E[\bar{Y} - E(\bar{Y})](b - \beta)$$

Considerando (2.11) e (2.18), obtemos

$$\begin{aligned} \text{cov}(\bar{Y}, b) &= E\left(\bar{u} \frac{\sum x_i u_i}{\sum x_i^2}\right) = \frac{E(\sum u_i \sum x_i u_i)}{n \sum x_i^2} = \\ &= \frac{E(u_1 + u_2 + \dots + u_n)(x_1 u_1 + x_2 u_2 + \dots + x_n u_n)}{n \sum x_i^2} = \\ &= \frac{E[x_1(u_1^2 + u_1 u_2 + \dots + u_1 u_n) + x_2(u_2 u_1 + u_2^2 + \dots + u_2 u_n) + \dots]}{n \sum x_i^2} \end{aligned}$$

Lembrando que  $E(u_i^2) = \sigma^2$  e  $E(u_i u_j) = 0$ , para  $i \neq j$ , segue-se que

$$\text{cov}(\bar{Y}, b) = \frac{\sigma^2 \sum x_i}{n \sum x_i^2} = 0 \quad (2.20)$$

Determinemos, finalmente, a covariância entre  $a$  e  $b$ .

Como  $a = \bar{Y} - b\bar{X}$ , temos, de acordo com as propriedades da covariância dadas na seção 1.5, que

$$\text{cov}(a, b) = \text{cov}(\bar{Y}, b) - \bar{X}V(b)$$

Considerando (2.15) e (2.20), obtemos

$$\text{cov}(a, b) = -\frac{\bar{X}\sigma^2}{\sum x_i^2} \quad (2.21)$$

## 2.6. Demonstração de que $b$ é um estimador linear não-tendencioso de variância mínima

Consideremos um estimador linear qualquer de  $\beta$ ,

$$B = \sum c_i Y_i$$

Para que esse estimador seja não-tendencioso, isto é, para que tenhamos  $E(B) = \beta$ , as constantes  $c_i$  devem ter certas propriedades, que serão deduzidas a seguir.

Temos

$$\begin{aligned} B &= \sum c_i Y_i = \\ &= \sum c_i (\alpha + \beta X_i + u_i) = \\ &= \alpha \sum c_i + \beta \sum c_i X_i + \sum c_i u_i \end{aligned} \quad (2.22)$$

Lembrando que  $E(u_i) = 0$ , podemos escrever

$$E(B) = \alpha \sum c_i + \beta \sum c_i X_i$$

Para termos  $E(B) = \beta$ , necessariamente

$$\sum c_i = 0 \quad (2.23)$$

e

$$\sum c_i X_i = 1 \quad (2.24)$$

Nestas condições e de acordo com (2.22), temos que

$$B = \beta + \sum c_i u_i$$

ou

$$B - \beta = \sum c_i u_i$$

Então,

$$V(B) = E(B - \beta)^2 = E(\sum c_i u_i)^2$$

Lembrando que  $E(u_i^2) = \sigma^2$  e  $E(u_i u_j) = 0$  para  $i \neq j$ , obtemos

$$V(B) = \sigma^2 \sum c_i^2$$

Nosso problema é determinar os valores de  $c_i$  que minimizem  $V(B)$ , ou seja, que minimizem  $\sum c_i^2$ , sujeitos às condições (2.23) e (2.24). Aplicando o método do multiplicador de Lagrange, definimos a função

$$F = \sum c_i^2 - \lambda \sum c_i - \theta (\sum c_i X_i - 1)$$

No ponto de mínimo condicionado devemos ter

$$\frac{\partial F}{\partial c_i} = 2c_i - \lambda - \theta X_i = 0 \quad (2.25)$$

para  $i = 1, 2, \dots, n$

Somando essas  $n$  igualdades obtemos

$$2 \sum c_i - n\lambda - \theta \sum X_i = 0$$

Pela condição (2.23) segue-se que

$$\lambda + \theta \bar{X} = 0$$

Adicionando essa igualdade a cada uma das relações em (2.25) obtemos

$$2c_i - \theta(X_i - \bar{X}) = 0$$

$$2c_i = \theta x_i$$

$$c_i = \frac{\theta}{2} x_i, \text{ para } i = 1, 2, \dots, n \quad (2.26)$$

Multiplicando cada uma dessas relações pelo respectivo valor de  $X_i$  e somando, obtemos

$$\sum c_i X_i = \frac{\theta}{2} \sum x_i X_i = \frac{\theta}{2} \sum x_i^2$$

Lembrando a condição (2.24), segue-se que

$$1 = \frac{\theta}{2} \sum x_i^2$$

ou

$$\frac{\theta}{2} = \frac{1}{\sum x_i^2}$$

Substituindo esse resultado em (2.26), obtemos

$$c_i = \frac{x_i}{\sum x_i^2}, \quad i = 1, 2, \dots, n$$

Concluímos então que o estimador linear não-tendencioso de variância mínima que procuramos é

$$B = \sum c_i Y_i = \frac{\sum x_i Y_i}{\sum x_i^2},$$

que é o estimador de mínimos quadrados.

Demonstração análoga pode ser feita em relação à estimativa do parâmetro  $\alpha$ . Concluímos, então, que os estimadores dos parâmetros de uma regressão linear simples, obtidos pelo método dos mínimos quadrados, são estimadores lineares não-tendenciosos de variância mínima. Esse é um caso particular do teorema de Gauss-Markov. Note que esse resultado depende de certas pressuposições a respeito do erro do modelo (pressuposições III a V).

## 2.7. Decomposição da soma de quadrados total

Demonstraremos que

$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2$$

ou

$$\sum y_i^2 = \sum e_i^2 + \sum \hat{y}_i^2,$$

isto é, que a soma de quadrados total (S.Q.Total) é igual à soma de quadrados residual (S.Q.Res.), também chamada soma de quadrados dos desvios, mais a soma de quadrados da regressão (S.Q.Regr.).

Partimos da identidade

$$Y_i - \bar{Y} = Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y}$$

ou

$$y_i = e_i + \hat{y}_i$$

Elevando ao quadrado e somando, obtemos

$$\sum y_i^2 = \sum e_i^2 + \sum \hat{y}_i^2 + 2 \sum \hat{y}_i e_i$$

Lembrando (2.10), concluímos que

$$\sum y_i^2 = \sum e_i^2 + \sum \hat{y}_i^2 \quad (2.27)$$

Essa relação mostra que a variação dos valores de  $Y$  em torno da sua média ( $\sum y_i^2$ ) pode ser dividida em duas partes: uma ( $\sum \hat{y}_i^2$ ) que é “explicada” pela regressão e outra ( $\sum e_i^2$ ) devida ao fato de que nem todos os pontos estão sobre a reta de regressão, que é a parte não “explicada” pela regressão. O *coeficiente de determinação*, definido por

$$r^2 = \frac{\text{S.Q.Regr.}}{\text{S.Q.Total}} = \frac{\sum \hat{y}_i^2}{\sum y_i^2},$$

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

Se estamos interessados em estimar valores de  $Y$  a partir de valores de  $X$ , a regressão será tanto mais útil quanto mais próximo de um estiver o valor de  $r^2$ .

Verificamos, facilmente, que

$$r^2 = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = b^2 \frac{\sum x^2}{\sum y^2} = \frac{(\sum xy)^2}{(\sum x^2)(\sum y^2)} \quad (2.28)$$

e que

$$\text{S.Q.Regr.} = \sum \hat{y}^2 = b^2 \sum x^2 = b \sum xy = \frac{(\sum xy)^2}{\sum x^2}$$

Vamos, agora, verificar a decomposição da soma de quadrados total e calcular o valor do coeficiente de determinação para o exemplo apresentado anteriormente.

Da tabela 2.1 obtemos

$$\text{S.Q.Total} = \sum y^2 = \sum Y^2 - \frac{(\sum Y)^2}{n} = 294 - 250 = 44$$

$$\text{S.Q.Regr.} = b \sum xy = 36$$

$$\text{S.Q.Res.} = \text{S.Q.Total} - \text{S.Q.Regr.} = 44 - 36 = 8$$

Esta é a maneira usual de obter os valores das várias somas de quadrados. Para que o aluno compreenda melhor o que está sendo feito, vamos calcular a S.Q.Regr. e a S.Q.Res. diretamente da sua definição; para isso precisamos obter, inicialmente, os valores de  $\hat{Y}_i$  e  $e_i = Y_i - \hat{Y}_i$ , apresentados na tabela 2.2.

As relações (2.5), (2.6), (2.7) e (2.10), deduzidas anteriormente, que são  $\sum e_i = 0$ ,  $\sum X_i e_i = 0$ ,  $\sum \hat{Y}_i e_i = 0$  e  $\sum \hat{y}_i e_i = 0$ , são facilmente verificadas neste exemplo.

Como a soma dos desvios é nula, a média aritmética de  $\hat{Y}_i$  é  $\bar{Y} = 5$ . Obtidos os valores  $\hat{y}_i = \hat{Y}_i - \bar{Y}$ , calculamos

$$\text{S.Q.Regr.} = \sum \hat{y}_i^2 = 36,$$

que é o mesmo valor obtido anteriormente, pela expressão

$$\text{S.Q.Regr.} = b \sum xy$$

TABELA 2.2. Valores de  $X_i, Y_i, \hat{Y}_i, \hat{y}_i$  e  $e_i$

$X_i$	$Y_i$	$\hat{Y}_i = 2 + X$	$\hat{y}_i$	$e_i = Y_i - \hat{Y}_i$
0	3	2	-3	+1
1	2	3	-2	-1
1	3	3	-2	0
2	5	4	-1	+1
3	4	5	0	-1
3	4	5	0	-1
4	7	6	+1	+1
5	6	7	+2	-1
5	7	7	+2	0
6	9	8	+3	+1

O valor da soma de quadrados residual, obtido anteriormente por diferença, pode agora ser obtido diretamente:

$$\text{S.Q.Res.} = \sum e_i^2 = 8$$

O leitor pode verificar que aplicando qualquer uma das fórmulas de (2.28) o valor do coeficiente de determinação obtido é

$$r^2 = \frac{9}{11} = 0,818 \text{ ou } 81,8\%$$

## 2.8. Esperanças das somas de quadrados

Vejamos, inicialmente, a esperança da soma de quadrados de regressão.

Temos

$$Y_i = \alpha + \beta X_i + u_i$$

e

$$\bar{Y} = \alpha + \beta \bar{X} + \bar{u}$$

Subtraindo esta equação da anterior, obtemos

$$y_i = \beta x_i + u_i - \bar{u} \quad (2.29)$$

Sabemos que

$$S.Q.Regr. = b \sum x_i y_i = \frac{(\sum x_i y_i)^2}{\sum x_i^2}$$

Aplicando esperança temos

$$E(S.Q.Regr.) = \frac{E(\sum x_i y_i)^2}{\sum x_i^2} \quad (2.30)$$

De (2.29) podemos obter

$$\sum x_i y_i = \beta \sum x_i^2 + \sum x_i u_i - \bar{u} \sum x_i = \beta \sum x_i^2 + \sum x_i u_i$$

Então

$$(\sum x_i y_i)^2 = \beta^2 (\sum x_i^2)^2 + (\sum x_i u_i)^2 + 2\beta \sum x_i^2 \sum x_i u_i$$

Aplicando esperança, vem

$$E(\sum x_i y_i)^2 = \beta^2 (\sum x_i^2)^2 + \sigma^2 \sum x_i^2 \quad (2.31)$$

Substituindo esse resultado em (2.30), obtemos

$$E(S.Q.Regr.) = \beta^2 \sum x_i^2 + \sigma^2 \quad (2.32)$$

Determinemos, a seguir, a esperança da soma de quadrados total.

Já vimos que

$$S.Q.Total = \sum y_i^2$$

Considerando (2.29), segue-se que

$$\begin{aligned} S.Q.Total &= \sum (\beta x_i + u_i - \bar{u})^2 = \\ &= \sum [\beta^2 x_i^2 + (u_i - \bar{u})^2 + 2\beta x_i (u_i - \bar{u})] = \\ &= \beta^2 \sum x_i^2 + \sum (u_i - \bar{u})^2 + 2\beta \sum x_i u_i - 2\beta \bar{u} \sum x_i = \\ &= \beta^2 \sum x_i^2 + \sum (u_i - \bar{u})^2 + 2\beta \sum x_i u_i \end{aligned}$$

Donde

$$E(S.Q.Total) = \beta^2 \sum x_i^2 + E[\sum (u_i - \bar{u})^2] \quad (2.33)$$

Mas

$$\sum (u_i - \bar{u})^2 = \sum (u_i^2 + \bar{u}^2 - 2u_i \bar{u}) =$$

$$\begin{aligned}
&= \sum u_i^2 + n \left( \frac{\sum u_i}{n} \right)^2 - 2 \frac{(\sum u_i)^2}{n} = \\
&= \sum u_i^2 - \frac{(\sum u_i)^2}{n}
\end{aligned}$$

Então

$$E[\sum(u_i - \bar{u})^2] = n\sigma^2 - \frac{n\sigma^2}{n} = (n-1)\sigma^2$$

Substituindo esse resultado em (2.33) obtemos

$$E(S.Q.Total) = \beta^2 \sum x_i^2 + (n-1)\sigma^2 \quad (2.34)$$

Resta determinar a esperança da soma de quadrados residual.

Como

$$S.Q.Res. = S.Q.Total - S.Q.Regr.,$$

temos

$$E(S.Q.Res.) = E(S.Q.Total) - E(S.Q.Regr.)$$

De acordo com (2.32) e (2.34) segue-se que

$$E(S.Q.Res.) = (n-2)\sigma^2 \quad (2.35)$$

## 2.9. Análise de variância da regressão

Os valores das esperanças das somas de quadrados, deduzidas no item anterior, justificam que se associe às somas de quadrados total, de regressão e residual  $n-1$ , 1 e  $n-2$  graus de liberdade, respectivamente.

Por definição, os quadrados médios são obtidos dividindo as somas de quadrados pelos respectivos graus de liberdade.

Então, para o caso de uma regressão linear simples, temos

$$Q.M.Regr. = S.Q.Regr.$$

e

$$Q.M.Res. = \frac{S.Q.Res.}{n-2}$$

Lembrando (2.32) e (2.35), obtemos

$$E(\text{Q.M.Regr.}) = \beta^2 \sum x_i^2 + \sigma^2$$

e

$$E(\text{Q.M.Res.}) = \sigma^2$$

De posse destes resultados, podemos conduzir a análise de variância da regressão linear simples, conforme o esquema seguinte:

Análise da Variância			
Causas de Variação	Graus de Liberdade	Somas de Quadrados	Quadrados Médios
Regressão	1	$b \sum x_i y_i$	$b \sum x_i y_i$
Resíduo	$n - 2$	$\sum y_i^2 - b \sum x_i y_i$	$(\sum y_i^2 - b \sum x_i y_i) / (n - 2)$
Total	$n - 1$	$\sum y_i^2$	

Considerando as diferentes amostras aleatórias de tamanho  $n$  que poderiam ser obtidas a partir da população de pares de valores ( $X, Y$ ), e sendo verdadeiras as 6 pressuposições dada na seção 2.1, concluímos que:

- a) o Q.M.Res. é uma estimativa não-tendenciosa da variância do erro ( $\sigma^2$ )
- b) o Q.M.Regr. é, em média, igual a essa mesma variância residual ( $\sigma^2$ ) somada ao produto de  $\sum x_i^2$  pelo quadrado do parâmetro  $\beta$ . É claro que se  $\beta = 0$ , o Q.M.Regr. é, em média, igual a  $\sigma^2$

Não faremos aqui, mas pode-se demonstrar que, se os erros têm distribuição normal e se  $\beta = 0$ , o cociente

$$F = \frac{\text{Q.M.Regr.}}{\text{Q.M.Res.}}$$

tem distribuição de  $F$  com 1 e  $n - 2$  graus de liberdade.

Então, para testar a hipótese

$$H_0 : \beta = 0,$$

ao nível de significância adotado, podemos utilizar a estatística  $F$ . Nesse caso, o procedimento consiste em rejeitar  $H_0$  para todo  $F$  maior ou igual ao  $F$  crítico, com  $1$  e  $n - 2$  graus de liberdade, relativo ao nível de significância adotado.

Note que, se essa hipótese é verdadeira, tanto o Q.M.Regr. como o Q.M.Res. são, em média, iguais a  $\sigma^2$  e o valor de  $F$  tende a  $1$ . Para  $\beta \neq 0$  teremos  $E(\text{Q.M.Regr.}) > E(\text{Q.M.Res.})$ , e o valor de  $F$  tende a ser superior a  $1$ .

Para ilustrar a aplicação desses conceitos, voltemos a considerar o exemplo numérico da tabela 2.1. Para este exemplo, obtemos a seguinte tabela de análise da variância:

Análise da Variância				
C.V.	G.L.	S.Q.	Q.M.	F
Regressão	1	36	36	36
Resíduo	8	8	1	
Total	9	44		

Ao nível de significância de  $5\%$  e para  $1$  e  $8$  graus de liberdade, o valor crítico de  $F$  é  $5,32$  (ver tabela de valores críticos de  $F$ ). O valor de  $F$  calculado, sendo superior ao valor crítico, é significativo ao nível de  $5\%$ . Conseqüentemente, rejeitamos a hipótese  $H_0 : \beta = 0$  em favor da hipótese alternativa  $H_A : \beta \neq 0$ , a esse nível de significância.

Um bom programa de análise de regressão para computador, além de calcular o valor de  $F$ , apresenta, ao lado, a probabilidade de, na distribuição teórica, ocorrer um valor de  $F$  maior do que o calculado. Trata-se da probabilidade associada à cauda da distribuição, acima do valor calculado. Parece apropriado denominá-la probabilidade caudal do teste.<sup>4</sup> Conhecendo essa probabilidade caudal, para saber se o resultado do teste é ou não significativo, basta compará-lo com o nível de significância adotado. O resultado é significativo sempre que a probabilidade caudal for igual ou menor do que o nível de significância. Torna-se desnecessário, então, obter o valor crítico da tabela apropriada. Para o exemplo numérico apresentado, o computador informa que a probabilidade caudal associada ao valor  $36$ , na distribuição de  $F$  com  $1$  e  $8$  graus de

---

<sup>4</sup> Nos textos em inglês essa probabilidade é denominada “p-value”, o que tem sido traduzido por “valor-p”.

liberdade, é 0,0003. O valor calculado é, portanto, significativo ao nível de 5% (é significativo mesmo que tivesse sido adotado um nível de significância de 0,1%).

## 2.10. O coeficiente de determinação corrigido para graus de liberdade e o coeficiente de variação

Na seção 2.7 vimos que o coeficiente de determinação é uma medida descritiva da qualidade do ajustamento obtido. Entretanto, o valor do coeficiente de determinação depende do número de observações da amostra, tendendo a crescer quando  $n$  diminui; no limite, para  $n = 2$ , teríamos sempre  $r^2 = 1$ , pois dois pontos determinam uma reta e os desvios são, portanto, nulos. Veremos a seguir como, numa tentativa de superar esse inconveniente, é definido o coeficiente de determinação corrigido para graus de liberdade, indicado por  $\bar{r}^2$ . Sabemos que

$$1 - r^2 = 1 - \frac{\text{S.Q.Regr.}}{\text{S.Q.Total}} = \frac{\text{S.Q.Res.}}{\text{S.Q.Total}}$$

Considerando o número de graus de liberdade associado à S.Q.Res. e à S.Q.Total, o coeficiente de determinação corrigido é definido por

$$1 - \bar{r}^2 = \frac{\frac{1}{n-2}(\text{S.Q.Res.})}{\frac{1}{n-1}(\text{S.Q.Total})} = \frac{n-1}{n-2}(1 - r^2) \quad (2.36)$$

ou

$$\bar{r}^2 = r^2 - \frac{1}{n-2}(1 - r^2) \quad (2.37)$$

Excluindo o caso em que  $r^2 = 1$ , temos  $\bar{r}^2 < r^2$ . Note que  $\bar{r}^2$  pode ser negativo.

Um outro indicador da qualidade do ajustamento obtido é o coeficiente de variação, definido por

$$CV = \frac{s}{\bar{Y}}, \quad (2.38)$$

onde  $s = \sqrt{Q.M.Res.}$ . O coeficiente de variação mede a dispersão relativa das observações, porque, por definição, é o cociente entre a medida da dispersão dos pontos

em torno da reta ( $s$ ) e o valor médio da variável dependente ( $\bar{Y}$ ). O resultado é tanto melhor quanto menor for o coeficiente de variação.

Veja o exercício 2.23 para uma análise comparativa dos valores do coeficiente de determinação e do coeficiente de variação, em vários casos.

## 2.11. Estimativas das variâncias das estimativas dos parâmetros, teste de hipóteses a respeito dos parâmetros e respectivos intervalos de confiança

Na seção 2.5 deduzimos que

$$V(b) = \frac{\sigma^2}{\sum x_i^2}$$

e que

$$V(a) = \left( \frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2} \right) \sigma^2$$

As respectivas estimativas são obtidas substituindo  $\sigma^2$  por  $s^2 = \text{Q.M.Res.}$ , ou seja,

$$\hat{V}(b) = s^2(b) = \frac{s^2}{\sum x_i^2} \quad (2.39)$$

e

$$\hat{V}(a) = s^2(a) = \left( \frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2} \right) s^2 \quad (2.40)$$

As estimativas dos desvios padrões  $s(a)$  e  $s(b)$  são obtidas extraindo a raiz quadrada das respectivas estimativas de variância.

Demonstra-se que, sendo válidas as seis pressuposições apresentadas na seção 2.1, inclusive a que estabelece a normalidade da distribuição dos erros, então os cocientes

$$t(b) = \frac{b - \beta}{s(b)} \quad \text{e} \quad t(a) = \frac{a - \alpha}{s(a)}$$

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

Vamos indicar algumas das etapas da demonstração no caso de  $t(b)$ . De (2.11), obtemos

$$b - \beta = \sum \left( \frac{x_i}{\sum x_i^2} \right) u_i,$$

mostrando que  $b - \beta$  é uma combinação linear dos  $u_i$ . Se os erros têm distribuição normal com média zero, segue-se que  $b - \beta$  também tem distribuição normal com média zero. Indicando o desvio padrão de  $b$  por  $\sigma_b = \sqrt{V(b)}$ , conclui-se que

$$Z = \frac{b - \beta}{\sigma_b},$$

tem distribuição normal reduzida.

O número de graus de liberdade associados a  $t(b)$  deve ser relacionado com o fato de  $s(b)$  ser obtido a partir do quadrado médio residual, que, conforme demonstramos, tem  $n - 2$  graus de liberdade.

Os valores  $t(b)$  e  $t(a)$  podem ser utilizados para testar hipóteses sobre os valores dos parâmetros, como ilustraremos a seguir com base no exemplo numérico que estamos desenvolvendo.

Calculemos, inicialmente, as estimativas das variâncias de  $b$  e de  $a$ .

$$\begin{aligned}\hat{V}(b) &= \frac{s^2}{\sum x_i^2} = \frac{1}{36} \\ \hat{V}(a) &= \left( \frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2} \right) s^2 = \frac{1}{10} + \frac{9}{36} = 0,35\end{aligned}$$

As estimativas dos desvios padrões são

$$s(b) = \frac{1}{6}$$

e

$$s(a) = \sqrt{0,35}$$

Para testar a hipótese  $H_0 : \beta = 0$ , contra a hipótese alternativa  $H_A : \beta \neq 0$ , ao nível de significância de 5%, calculamos

$$t(b) = \frac{1-0}{1/6} = 6$$

Para um teste bilateral, o valor crítico de  $t$  com 8 graus de liberdade, ao nível de significância de 5%, é 2,306 (ver tabela de valores críticos de  $t$ ). Portanto, o valor calculado  $t(b)$  é significativo ao nível de 5%, ou seja, rejeitamos  $H_0$  em favor de  $H_A$ , a esse nível de significância.

Note que esse teste é perfeitamente equivalente ao teste  $F$  feito na análise de variância, uma vez que o valor de  $F$  calculado é igual ao quadrado do valor de  $t$  calculado e que o valor crítico de  $F$  é igual ao quadrado do valor crítico de  $t$ .

No caso dos testes  $t$ , os programas de computador usualmente fornecem a probabilidade de o valor teórico de  $t$  ser, *em valor absoluto*, maior do que o valor calculado. Trata-se, portanto, de uma probabilidade caudal apropriada para testes bilaterais. Se o teste de hipóteses for unilateral, é necessário dividir por 2 a probabilidade caudal fornecida pelo computador.

Consideremos, agora, que se deseja testar a hipótese  $H_0 : \alpha = 3$  contra a hipótese alternativa  $H_A : \alpha < 3$ , ao nível de significância de 5%. Para isso calculamos

$$t(a) = \frac{2-3}{\sqrt{0,35}} = -1,690$$

A região de rejeição para este teste é  $t < -1,860$ . Como o valor calculado não pertence a esse intervalo, ele não é significativo, ou seja, não rejeitamos, ao nível de significância de 5%, a hipótese  $H_0 : \alpha = 3$ .

Também podem ser obtidos intervalos de confiança para os parâmetros. Sendo  $t_0$  o valor crítico de  $t$  com  $n - 2$  graus de liberdade e ao nível de confiança estabelecido, os intervalos de confiança para  $\beta$  e para  $\alpha$  são, respectivamente,

$$b - t_0 s(b) < \beta < b + t_0 s(b)$$

e

$$a - t_0 s(a) < \alpha < a + t_0 s(a)$$

Vamos determinar, no exemplo numérico que estamos desenvolvendo, o intervalo de confiança para  $\beta$  ao nível de confiança de 90%. O valor crítico de  $t$  para 8 graus de liberdade é 1,860. Então o intervalo de 90% de confiança é

$$1 - 1,860 \frac{1}{6} < \beta < 1 + 1,860 \frac{1}{6}$$

$$0,69 < \beta < 1,31$$

## 2.12. Variância de $\hat{Y}_i$ e intervalo de previsão

Vimos na seção 2.3 que

$$\hat{Y}_i = \bar{Y} + bx_i$$

Então

$$V(\hat{Y}_i) = V(\bar{Y}) + x_i^2 V(b) + 2x_i \text{cov}(\bar{Y}, b)$$

Considerando (2.15), (2.19) e (2.20), segue-se que

$$V(\hat{Y}_i) = \frac{\sigma^2}{n} + \frac{x_i^2 \sigma^2}{\sum x_i^2} = \left( \frac{1}{n} + \frac{x_i^2}{\sum x_i^2} \right) \sigma^2 \quad (2.41)$$

Donde

$$\hat{V}(\hat{Y}_i) = s^2(\hat{Y}_i) = \left( \frac{1}{n} + \frac{x_i^2}{\sum x_i^2} \right) s^2$$

Podemos estimar o valor de  $Y$  correspondente a um valor de  $X$  que não existe na amostra. Se reservarmos o índice  $i$  para indicar os elementos pertencentes à amostra, devemos introduzir aqui um outro índice ( $h$ ) para indicar outros valores de  $X$ . O novo valor,  $X_h$ , pode coincidir ou não com um dos valores ( $X_i$ ) da amostra. Temos

$$\hat{Y}_h = a + bX_h = \bar{Y} + bx_h$$

e

$$\hat{V}(\hat{Y}_h) = s^2(\hat{Y}_h) = \left( \frac{1}{n} + \frac{x_h^2}{\sum x_i^2} \right) s^2$$

Extraindo a raiz quadrada desse valor obtemos a estimativa dos desvio padrão de  $\hat{Y}_h$ . Sendo  $t_0$  o valor crítico de  $t$  com  $n - 2$  graus de liberdade e ao nível de confiança estabelecido, o intervalo de confiança para  $E(Y_h) = \alpha + \beta X_h$  é

$$\hat{Y}_h - t_0 s(\hat{Y}_h) < E(Y_h) < \hat{Y}_h + t_0 s(\hat{Y}_h)$$

Freqüentemente, temos interesse em estimar o valor de uma nova observação ( $Y_h$ ), relativa ao valor de  $X_h$  da variável independente, isto é, queremos prever o valor da variável dependente em uma nova observação com  $X = X_h$ . O estimador de  $Y_h = \alpha + \beta X_h + u_h$  é  $\hat{Y}_h = a + bX_h$ . O erro de previsão é

$$\hat{Y}_h - Y_h = (a - \alpha) + (b - \beta)X_h - u_h$$

Dizemos que  $\hat{Y}_h$  é uma previsão não-tendenciosa do valor de  $Y_h$  porque a esperança do erro de previsão é igual a zero. Verifica-se, também, que  $E(\hat{Y}_h) = E(Y_h)$ . Note, entretanto, que  $E(\hat{Y}_h) = \alpha + \beta X_h \neq Y_h$ .

Para avaliar a precisão  $\hat{Y}_h$  como previsão do valor da nova observação, determinamos o intervalo de previsão, como mostraremos a seguir. Note, inicialmente, que tanto  $\hat{Y}_h$  como  $Y_h$  são variáveis aleatórias e que, de acordo com a pressuposição V, suas distribuições são independentes. Uma vez que, para determinado valor ( $X_h$ ) da variável independente, os valores de  $Y$  variam em torno de sua verdadeira média, isto é, em torno de  $E(Y_h)$ , com variância  $\sigma^2$ , a variância que nos interessa é

$$\sigma^2 + V(\hat{Y}_h) = \left(1 + \frac{1}{n} + \frac{x_h^2}{\sum x_i^2}\right)\sigma^2 \quad (2.42)$$

O intervalo de previsão para a nova observação ( $Y_h$ ) é

$$\hat{Y}_h - t_0 \left[ \left(1 + \frac{1}{n} + \frac{x_h^2}{\sum x_i^2}\right)s^2\right]^{1/2} < Y_h < \hat{Y}_h + t_0 \left[ \left(1 + \frac{1}{n} + \frac{x_h^2}{\sum x_i^2}\right)s^2\right]^{1/2}$$

O conceito de intervalo de previsão é análogo ao de intervalo de confiança, com a diferença de que, enquanto o intervalo de confiança se refere a uma constante (o

parâmetro  $\beta$ , por exemplo), o intervalo de previsão se refere a uma variável aleatória ( $Y_h$ , no caso).

Consideremos, para exemplificar a aplicação dessas fórmulas, os pares de valores da tabela 2.1. Já vimos que para esses dados  $\bar{Y} = 5$  e  $b = 1$ .

Então

$$\hat{Y}_h = 5 + x_h \quad (2.43)$$

O valor crítico de  $t$ , com 8 graus de liberdade, ao nível de confiança de 95%, é 2,306.

Lembrando que, no exemplo numérico que estamos desenvolvendo,  $n = 10$ ,  $\sum x_i^2 = 36$  e  $s^2 = 1$ , verificamos que os limites do intervalo de confiança para  $E(Y_h)$ , ao nível de confiança de 95%, são dados por

$$\hat{Y}_h \pm 2,306 \sqrt{\frac{1}{10} + \frac{x_h^2}{36}} \quad (2.44)$$

Os limites do intervalo de previsão para uma nova observação ( $Y_h$ ) são dados por

$$\hat{Y}_h \pm 2,306 \sqrt{1 + \frac{1}{10} + \frac{x_h^2}{36}} \quad (2.45)$$

Utilizando as expressões (2.43), (2.44) e (2.45) obtivemos os valores de  $\hat{Y}_h$  e os limites dos intervalos de confiança e de previsão que estão na tabela 2.4 e na figura 2.3. Note que consideramos alguns valores de  $X$  fora do intervalo para o qual dispúnhamos de dados, ou seja, estamos fazendo extrapolação.

TABELA 2.4 Valores de  $\hat{Y}_h$ , limites do intervalo de confiança para  $E(Y_h)$ , ao nível de confiança de 95%.

$X_h$	$\hat{Y}_h$	Intervalo de confiança para $E(Y_h)$	Intervalo de previsão para $Y_h$
0	2	0,64	3,36
1	3	1,94	4,06
2	4	3,18	4,82
3	5	4,27	5,73
4	6	5,18	6,82
5	7	5,94	8,06
6	8	6,64	9,36
7	9	7,30	10,70
8	10	7,94	12,06
9	11	8,58	13,42

A análise da expressão

$$\hat{Y}_h \pm t_0 \left( \frac{1}{n} + \frac{x_h^2}{\sum x_i^2} \right)^{1/2} s,$$

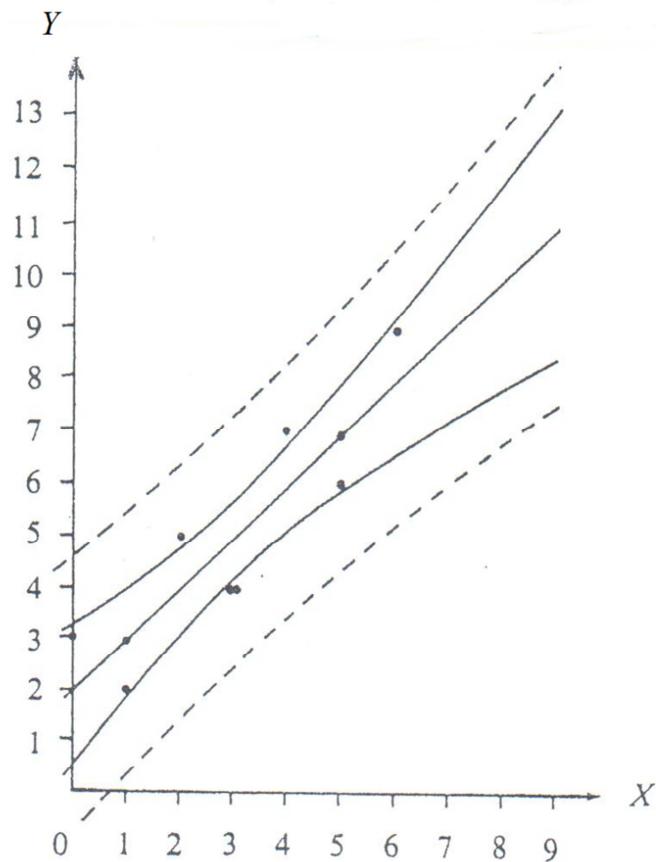
que nos dá os limites do intervalo de confiança para  $E(Y_h)$ , permite afirmar que a precisão da estimativa de  $Y$  é tanto maior quanto:

- a) menor for  $s$ , isto é, quanto menor for a dispersão dos valores observados de  $Y$  em torno da reta de regressão.
- b) maior for  $n$
- c) maior for  $\sum x_i^2$ , isto é, quanto maior for a dispersão dos valores de  $X$  em torno da respectiva média.

Podemos então concluir que:

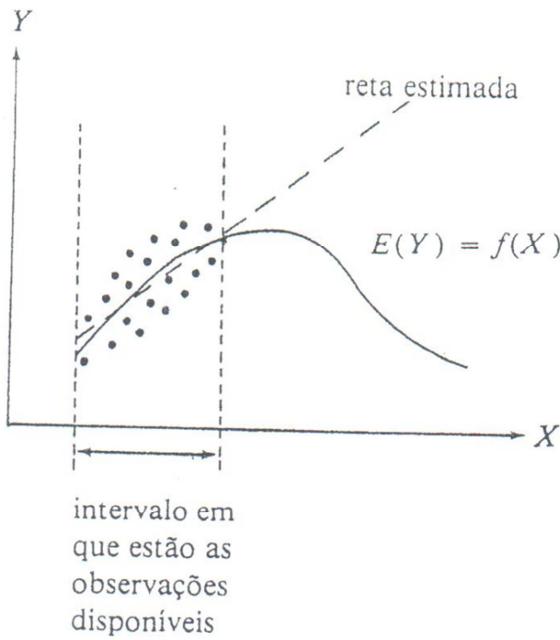
- a) O número de observações ( $n$ ) deve ser o maior possível.
- b) Se possível, devemos escolher valores de  $X$  que conduzem a um elevado valor para  $\sum x_i^2$ .

Devemos notar, ainda, que o intervalo de confiança aumenta à medida que  $X_h$  se afasta de  $\bar{X}$ .



*Figura 2.3.* A reta de regressão estimada, o intervalo de confiança para  $E(Y_h)$  e o intervalo de previsão para  $Y_h$ .

Ao fazer uma extrapolação é necessário considerar, ainda, um outro problema, provavelmente mais sério que o crescimento da amplitude do intervalo de confiança (e também do intervalo de previsão) à medida que  $X_h$  se afasta de  $\bar{X}$ . Freqüentemente o modelo (linear) ajustado é razoável para o intervalo coberto pela amostra, mas é absolutamente inapropriado para uma extrapolação. A figura 2.4 ilustra a questão. As observações da amostra pertencem ao intervalo em que a relação entre  $E(Y)$  e  $X$  é aproximadamente linear. Entretanto, se utilizarmos a reta estimada para prever valores à direita desse intervalo, os resultados estarão totalmente “fora do alvo”.



*Figura 2.4. O perigo da extrapolação*

### 2.13. O problema da especificação e as funções que se tornam lineares por anamorfose

Quando aplicamos análise de regressão ao estudo da relação funcional entre duas variáveis, o problema da especificação consiste em determinar a forma matemática da função que será ajustada. Podemos escolher, por exemplo:

$$\text{I)} \quad Y_i = \alpha + \beta X_i$$

$$\text{II)} \quad Y_i = \alpha \beta^{X_i}$$

$$\text{III)} \quad Y_i = \alpha X_i^\beta$$

$$\text{IV)} \quad Y_i = \alpha + \frac{\beta}{X_i}$$

$$\text{V)} \quad Y_i = \alpha + \beta X_i + \gamma X_i^2$$

$$\text{VI)} \quad Y_i = \alpha + \beta \rho^{X_i}$$

Onde  $\alpha$ ,  $\beta$ ,  $\gamma$  e  $\rho$  são parâmetros a serem estimados.

A determinação da forma matemática da função pode ser feita de duas formas diferentes, muitas vezes complementares:

- a) utilizando o conhecimento que temos *a priori* sobre o fenômeno.

- b) Empregando o conhecimento adquirido pela inspeção dos dados numéricos disponíveis. É muito útil fazer um gráfico com os pontos  $(X_i, Y_i)$  e, eventualmente, gráficos com os pontos  $(\ln X_i, Y_i)$ ,  $(X_i, \ln Y_i)$  ou  $(\ln X_i, \ln Y_i)$ .

Freqüentemente, ajustamos mais de um modelo e escolhemos, com base nos resultados estatísticos obtidos (coeficientes de determinação, quadrados médios residuais etc.), o modelo que melhor se ajusta aos dados.

Admitindo um erro aditivo o modelo I fica

$$Y_i = \alpha + \beta X_i + u_i,$$

que é o modelo estatístico estudado até aqui.

Mostraremos agora que os modelos II, III e IV são exemplos de modelos não-lineares que se transformam em funções lineares por anamorfose, isto é, por substituição dos valores de uma ou mais variáveis por funções destas variáveis.

No caso do modelo II (função exponencial), admitindo um erro multiplicativo  $\varepsilon_i$ , obtemos o modelo estatístico

$$Y_i = \alpha \beta^{X_i} \varepsilon_i$$

Aplicando logaritmos, obtemos

$$\log Y_i = \log \alpha + X_i \log \beta + \log \varepsilon_i$$

Fazendo

$$\log Y_i = Z_i,$$

$$\log \alpha = A,$$

$$\log \beta = B$$

e

$$\log \varepsilon_i = u_i$$

temos

$$Z_i = A + BX_i + u_i$$

que é o modelo estatístico de uma regressão linear simples de  $Z_i = \log Y_i$  em relação a  $X_i$ . Caso o erro  $u_i = \log \varepsilon_i$  obedeça às pressuposições dadas na seção 2.1, podemos aplicar à amostra de pares de valores  $X_i$  e  $Z_i$  os métodos de análise de regressão já estudados. Obtidas as estimativas dos parâmetros  $A$  e  $B$  é fácil determinar as correspondentes estimativas de  $\alpha$  e de  $\beta$ .

No caso do modelo III (função potência), conhecido entre economistas como função de Cobb-Douglas, o correspondente modelo estatístico é

$$Y_i = \alpha X_i^\beta \varepsilon_i$$

Aplicando logaritmos, obtemos

$$\log Y_i = \log \alpha + \beta \log X_i + \log \varepsilon_i$$

ou

$$Z_i = A + \beta V_i + u_i$$

onde

$$Z_i = \log Y_i,$$

$$A = \log \alpha,$$

$$V_i = \log X_i$$

e

$$u_i = \log \varepsilon_i$$

Verificamos que a função potência corresponde a um modelo de regressão linear simples nos logaritmos das duas variáveis.

Admitindo um erro aditivo o modelo IV (hipérbole) dá origem ao seguinte modelo estatístico

$$Y_i = \alpha + \frac{\beta}{X_i} + u_i$$

Basta fazer a anamorfose

$$V_i = \frac{1}{X_i}$$

para obter o modelo de uma regressão linear simples,

$$Y_i = \alpha + \beta V_i + u_i$$

Os métodos de ajustamento dos modelos V e VI serão visto adiante. O modelo V é ajustado como uma regressão múltipla com duas variáveis explanatórias e o modelo VI será estudado no capítulo sobre regressão não-linear.

## 2.14. Estimativa de máxima verossimilhança

Para determinar as estimativas de máxima verossimilhança dos parâmetros  $\alpha$  e  $\beta$  da regressão

$$Y_i = \alpha + \beta X_i + u_i \quad (2.46)$$

admitiremos que os erros ( $u_i$ ) são variáveis aleatórias independentes com média zero, variância  $\sigma^2$  e distribuição normal, ou seja,

$$u_i \sim N(0, \sigma^2)$$

Então, o valor da densidade de probabilidade para certo valor  $Y_1$  é

$$f(Y_1) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}[Y_1 - (\alpha + \beta X_1)]^2\right\}$$

Consideremos uma amostra de  $n$  pares de valores  $X_i, Y_i$ . Se os valores de  $X$  são fixos e as observações são independentes, a densidade de probabilidade do modelo (2.46) ter gerado tais valores de  $Y_i$  é dada por

$$\begin{aligned} L(X_1, \dots, X_n; \alpha, \beta, \sigma^2) &= \\ &= \prod_{i=1}^n (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}[Y_i - (\alpha + \beta X_i)]^2\right\} = \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n [Y_i - (\alpha + \beta X_i)]^2\right\} \end{aligned} \quad (2.47)$$

Os estimadores de máxima verossimilhança de  $\alpha$ ,  $\beta$  e  $\sigma^2$  são os valores que maximizam  $L(\alpha, \beta, \sigma^2 | X_1, \dots, X_n)$ , que é a função de verossimilhança. Uma vez que  $\alpha$  e  $\beta$  só aparecem no expoente negativo de (2.47), concluímos que o máximo dessa função corresponde ao mínimo de

$$\sum [Y_i - (\alpha + \beta X_i)]^2$$

Portanto, as estimativas de máxima verossimilhança dos parâmetros  $\alpha$  e  $\beta$  coincidem com as estimativas de mínimos quadrados, desde que a distribuição dos erros seja normal.

O leitor pode verificar que a estimativa de máxima verossimilhança de  $\sigma^2$  é

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n}$$

## 2.15. Análise de regressão quando $X$ é uma variável aleatória

Consideremos o modelo

$$Y_i = \alpha + \beta X_i + u_i$$

onde  $X$  é uma variável aleatória cuja distribuição não depende de  $\alpha$ ,  $\beta$  ou  $\sigma^2$ .

Vamos manter as pressuposições I, III, IV, V e VI, dadas na seção 2.1, e pressupor ainda que os erros  $u_i$  são independentes dos valores de  $X$ . Então  $X_i$  e  $u_i$  são não-correlacionados, isto é,

$$E[(X_i - \mu_X)u_i] = 0.$$

Considerando as distribuições condicionais dos  $Y_i$  (ou dos  $u_i$ ), dados  $X_i$ , pode-se verificar que todos os resultados obtidos são válidos, desde que interpretados como condicionados aos valores de  $X$  observados.

Entretanto, tais resultados condicionais podem ser insatisfatórios para o problema em estudo. É o caso, por exemplo, de um intervalo de confiança para  $\beta$ , válido apenas para o conjunto dos valores de  $X$  observados.

Devemos salientar, no entanto, que mesmo sendo  $X$  uma variável aleatória, se forem verdadeiras as pressuposições enunciadas (e é crucial que as distribuições de  $X_i$  e  $u_i$  sejam independentes), é possível demonstrar que os estimadores de mínimos quadrados são não-tendenciosos e coincidem com os estimadores de máxima verossimilhança. Também se demonstra que os procedimentos para a realização de testes de hipóteses e para a determinação de intervalos de confiança, apresentados na seção 2.11, continuam válidos. É interessante notar que a interpretação do intervalo de 95% de confiança para  $\beta$ , por exemplo, passa a ser a seguinte: para um grande número de amostras, onde tanto os valores de  $Y_i$  como os valores de  $X_i$  variam de uma amostra para outra, em aproximadamente 95% dos casos os intervalos de confiança, obtidos da maneira indicada na seção 2.11, conteriam o valor verdadeiro ( $\beta$ ).

## Exercícios

- 2.1. É dada uma amostra de 10 pares de valores

$X$	$Y$
−2	0
−2	0
−1	2
−1	3
0	4
0	4
1	5
1	6
2	8
2	8

Admite-se que as variáveis  $X$  e  $Y$  estão relacionadas de acordo com o modelo  $Y_i = \alpha + \beta X_i + u_i$ , onde os  $u_i$  são variáveis aleatórias independentes com distribuição normal de média zero e variância  $\sigma^2$ .

- a) Determine as estimativas dos parâmetros da regressão linear.
  - b) Teste  $H_0 : \beta = 0$  ao nível de significância de 5%.
  - c) Calcule o coeficiente de determinação.
  - d) Determine a estimativa de  $Y$  para  $X = 3$  e o respectivo intervalo de confiança ao nível de confiança de 95%.
- 2.2. Seja  $b$  o coeficiente de regressão de  $Y$  em relação a  $X$ . Seja  $d$  o coeficiente de regressão de  $Z$  em relação a  $X$ . Sabendo que  $Z = m + nY$ , determine a relação existente entre as estimativas de mínimos quadrados de  $b$  e  $d$ .
- 2.3. Demonstre que numa regressão linear simples o valor de  $F$  da análise de variância da regressão é igual ao quadrado do valor de  $t(b)$ , relativo à hipótese da nulidade  $\beta = 0$  (onde  $\beta$  é o coeficiente de regressão).
- 2.4. Admite-se que as variáveis  $X$  e  $Y$  estão relacionadas de acordo com o modelo  $Y_i = \alpha + \beta X_i + u_i$ , onde os  $u_i$  são variáveis aleatórias independentes com distribuição normal de média zero e variância  $\sigma^2$ .

São dados os seguintes valores, obtidos de uma amostra aleatória de 14 observações:

$$\begin{aligned}\sum X_i &= 140 & \sum Y_i &= 728 \\ \sum X_i^2 &= 1456 & \sum Y_i^2 &= 39424 \\ \sum X_i Y_i &= 7504\end{aligned}$$

- a) Determine as estimativas dos parâmetros da regressão de  $Y$  em relação a  $X$  e os respectivos desvios padrões.
- b) Calcule o coeficiente de determinação da regressão.
- c) Teste a hipótese  $H_0 : \beta = 0$  contra a hipótese alternativa  $H_A : \beta > 0$ , ao nível de significância de 0,5%.

2.5. É dada uma amostra de 5 pares de valores:

$X$	$Y$
1	3,0
2	7,5
3	7,0
4	11,5
5	11,0

Admite-se que as variáveis  $X$  e  $Y$  estão relacionadas de acordo com o modelo  $Y_i = \alpha + \beta X_i + u_i$ , onde os  $u_i$  são variáveis aleatórias independentes com distribuição normal de média zero e variância  $\sigma^2$ .

- a) Determine as estimativas dos parâmetros da regressão linear.
- b) Calcule o coeficiente de determinação e faça a análise de variância da regressão, considerando o nível de significância de 5%.
- c) Teste, ao nível de significância de 0,5%, a hipótese  $H_0 : \beta = -2$  contra a hipótese alternativa  $H_A : \beta \neq -2$ .
- d) Teste, ao nível de significância de 0,5%, a hipótese  $H_0 : \alpha = 13$  contra a hipótese alternativa  $H_A : \alpha < 13$ .
- e) Determine a estimativa de  $Y$  para  $X = 5$  e o respectivo intervalo de confiança ao nível de confiança de 95%.

2.6. São dados os 5 pares de valores:

X	Y
0	2
2	3
4	14
6	15
8	26

Obtemos  $\sum X = 20$ ,  $\sum Y = 60$ ,  $\sum X^2 = 120$ ,  $\sum Y^2 = 1110$  e  $\sum XY = 360$

Admite-se que as variáveis  $X$  e  $Y$  estão relacionadas de acordo com o modelo  $Y_i = \alpha + \beta X_i + u_i$  e que são válidas as pressuposições usuais a respeito dos erros  $u_i$ .

- a) Determine as estimativas dos parâmetros da regressão linear.
- b) Calcule o coeficiente de determinação.
- c) Verifique se o coeficiente de determinação é estatisticamente diferente de zero ao nível de significância de 1%.
- d) Teste a hipótese  $H_0 : \beta = 6$  contra a hipótese alternativa  $H_A : \beta > 6$ , ao nível de significância de 1%.
- e) Teste a hipótese  $H_0 : \alpha = 6$  contra a hipótese alternativa  $H_A : \alpha \neq 6$ , ao nível de significância de 10%.
- f) Determine o intervalo de previsão para uma nova observação de  $Y$  com  $X = 10$ , ao nível de confiança de 95%.

- 2.7. Com base em 52 pares de valores das variáveis  $X$  e  $Y$  foi obtida a equação de regressão

$$\hat{Y}_i = -0,4 + X_i$$

- A estimativa do desvio padrão da estimativa do coeficiente de regressão é 0,1. Calcule o coeficiente de determinação e teste a hipótese de que o coeficiente angular da equação é igual a zero, ao nível de significância de 1%.

- 2.8. Mostre que  $b$  é um estimador consistente de  $\beta$ .

- 2.9. A partir de  $n$  pares de valores  $X_i, Y_i$  obtemos, pelo método de mínimos quadrados, a equação de regressão  $\hat{Y}_i = a + bX_i$ . Sendo  $Z_i = Y_i + X_i$ , temos  $n$  pares de valores  $Z_i, X_i$ , a partir dos quais obtemos a equação de regressão:

$$\hat{Z}_i = c + dX_i$$

Que relação existe entre  $b$  e  $d$ ? E entre  $a$  e  $c$ ? Sendo  $b$  e  $d$  as estimativas dos parâmetros  $\beta$  e  $\delta$ , respectivamente, demonstre que o valor de  $t$  relativo à hipótese da nulidade  $H_0 : \beta = 0$  é igual ao valor de  $t$  relativo à hipótese da nulidade  $H_0 : \delta = 1$ , ou seja, que

$$\frac{b}{s(b)} = \frac{d - 1}{s(d)}$$

- 2.10. Dados os pares de valores  $X$  e  $Y$  abaixo, qual é o modelo que você usaria e como faria para obter uma relação que lhe permitisse estimar  $Y$  a partir de valores de  $X$ ?

$X$	$Y$
10	2,0
12	8,2
14	31,0
16	130,0
18	510,0

- 2.11. A partir de uma amostra de 27 pares de valores foi obtida a equação de regressão de  $Y$  em relação a  $X$

$$\hat{Y} = 25,0 + 2,00X$$

Sabendo que  $s = 1,50$  ( $s^2 = \text{Q.M.Res.}$ ), que a estimativa do desvio padrão de  $X$  é  $s(X) = 3,00$  e que  $\bar{X} = 7,50$ ,

- a) determine o intervalo de confiança do coeficiente de regressão ao nível de confiança de 95%.
- b) Teste, ao nível de significância de 1%, a hipótese de que o coeficiente de regressão da população é 1,70.

2.12. Para aumentar a precisão da estimativa do coeficiente de regressão, que devemos fazer com relação à escolha dos valores  $X$  que serão utilizados na análise de regressão?

2.13. Discuta, rapidamente, os problemas relacionados com a extrapolação, especialmente no campo socioeconômico.

2.14. Suponha que uma fábrica dispõe dos seguintes dados:

Quantidade Produzida	105	130	141	159	160	172
Custo (R\$)	2042	2301	2421	2518	2606	2718

Por meio de uma análise de regressão, estabeleça:

- O valor mais provável dos custos fixos e o respectivo intervalo de confiança, ao nível de confiança de 95%.
- A quantidade para a qual o lucro é nulo, admitindo um preço de venda de R\$ 18,00 por unidade.

2.15. Com base em 11 pares de valores das variáveis  $X$  e  $Y$  foi obtida a equação de regressão

$$\hat{Y} = 20 - X ,$$

com  $r^2 = 0,64$ . Sabe-se que a estimativa não-tendenciosa da variância de  $X$  é 64. Teste, ao nível de significância de 0,5%, a hipótese  $H_0 : \beta = 0$ , contra a hipótese alternativa  $H_A : \beta > 0$ .

2.16. Numa análise de regressão ( $Y_i = \alpha + \beta X_i + u_i$ ) foram obtidos, a partir de uma amostra de 6 pares de valores  $X$  e  $Y$ , os seguintes resultados:

$$r^2 = \frac{16}{25}; \quad s(X) = 3; \quad s(Y) = 5; \quad \bar{X} = 3 \quad \text{e} \quad \bar{Y} = 10$$

- Determine o intervalo de 95% de confiança para  $\beta$ , sabendo que  $Y$  é uma função crescente de  $X$ .
- Teste, ao nível de significância de 5%, a hipótese  $H_0 : \alpha = 0$  contra a hipótese alternativa  $H_A : \alpha > 0$ .

2.17. Admitindo que as variáveis  $X$  e  $Y$  estão relacionadas conforme o modelo

$$Y_i = \alpha + \frac{\beta}{X_i} + u_i,$$

onde  $u_i$  representa erros aleatórios independentes com média zero e variância constante, determine as estimativas dos parâmetros  $\alpha$  e  $\beta$ , com base nos seguintes dados:

$X$	$Y$
12	9,0
15	8,5
20	8,5
30	6,5
60	5,0

2.18. Sejam as variáveis  $X$  e  $Y$ , relacionadas de acordo com o modelo

$$Y_i = \alpha + \beta X_i + u_i$$

onde  $u_i$  são erros aleatórios. Dados os resultados abaixo, obtidos de uma amostra aleatória com 50 observações,

$$\sum Y_i = 100 \quad \sum Y_i^2 = 212 \quad \sum X_i Y_i = 21$$

$$\sum X_i = 10 \quad \sum X_i^2 = 2,1$$

- a) Calcule as estimativas dos parâmetros  $\alpha$  e  $\beta$  e as estimativas das respectivas variâncias. Que pressuposições devem ser feitas para que estas estimativas sejam imparciais e de variância mínima?
- b) Teste as seguintes hipóteses ao nível de significância de 5%:
  - I)  $H_0 : \alpha = 2$  contra a hipótese alternativa  $H_A : \alpha \neq 2$
  - II)  $H_0 : \beta = 12$  contra a hipótese alternativa  $H_A : \beta < 12$
 Que pressuposição adicional deve ser feita para testar essas hipóteses?
- c) Calcule o valor do coeficiente de determinação da regressão e interprete o resultado.

- 2.19. São dados os seguintes valores, obtidos de uma amostra aleatória com 10 observações:

$X$	$Y$
0	2,5
1	1
2	2
3	0
4	0,5
	3,5

Admite-se que as variáveis  $X$  e  $Y$  estão relacionadas de acordo com o modelo  $Y_i = \alpha + \beta X_i + u_i$ , onde  $u_i$  são variáveis aleatórias homocedásticas, normalmente distribuídas e com média zero. Pode-se verificar que  $\sum x^2 = 20$ ,  $\sum Y^2 = 55$ ,  $\sum y^2 = 15$ ,  $\sum xY = -10$  e  $\bar{Y} = 2$ .

- a) Determine a reta de regressão de  $Y$  em relação a  $X$ , de acordo com o método dos mínimos quadrados.
- b) Calcule o coeficiente de determinação e verifique se é estatisticamente diferente de zero, através do teste  $F$ , considerando um nível de significância de 5%.
- c) Teste a hipótese  $H_0 : \beta = 0$  contra a hipótese alternativa  $H_A : \beta > 0$ , ao nível de significância de 5%.
- d) Teste a hipótese  $H_0 : \alpha = 1$  contra a hipótese alternativa  $H_A : \alpha \neq 1$ , ao nível de significância de 1%.
- e) Determine o intervalo de previsão para uma nova observação de  $Y$  com  $X = 2 + \sqrt{10}$ , ao nível de confiança de 95%.
- f) Considere que a equação de regressão obtida é a demanda de um produto em certo mercado, sendo  $X$  o preço e  $Y$  a quantidade procurada. Se o produto em questão é vendido por um monopolista e seu custo médio de produção é constante e igual a Cr\$ 2,00, que preço o monopolista deve estabelecer para maximizar sua renda líquida? Estime a quantidade que será vendida a esse preço e determine o intervalo de confiança correspondente, ao nível de confiança de 95%.

- 2.20. Seja  $X$  a quantidade de certo produto, em milhares de unidades, e  $Y$  o respectivo custo total de produção em milhares de cruzeiros. Admite-se que o custo marginal seja constante. É dada a seguinte amostra de 10 pares de valores

(extraídos de H.W. GUTHRIE. *Statistical Methods in Economics*. Richard D. Irwin, 1966, p. 108-109):

$X$	$Y$
(1 000 unidades)	(Cr\$ 1 000,00)
1	7
2	11
3	15
4	14
5	18
6	21
7	23
8	30
9	32
10	34

Pode-se verificar que  $\sum X = 55$ ,  $\sum Y = 205$ ,  $\sum X^2 = 385$ ,  $\sum Y^2 = 4965$  e  $\sum XY = 1375$ .

- a) Estime a função de custo total.
- b) Teste, ao nível de significância de 1%, a hipótese de que o custo marginal é nulo.
- c) Determine o intervalo de 95% de confiança para o valor dos custos fixos.
- d) Calcule o coeficiente de determinação da regressão.
- e) Se o produtor vende em regime de competição perfeita ao preço de Cr\$ 3,50 por unidade, quantas unidades deve produzir para que sua renda líquida seja de Cr\$ 2.000,00?
- f) Determine a estimativa de  $Y$  para  $X = 10$  e o respectivo intervalo de confiança, ao nível de confiança de 95%.

2.21. É dada uma amostra de 4 pares de valores:

$X$	$Y$
2	6
1	8
1	9
4	13

Admite-se que as variáveis  $X$  e  $Y$  estão relacionadas de acordo com o modelo  $Y_i = \alpha + \beta X_i + u_i$ , onde os  $u_i$  são erros independentes, de média zero, variância constante e distribuição normal.

- a) Determine as estimativas dos parâmetros da regressão linear.
  - b) Calcule o coeficiente de determinação da regressão.
  - c) Teste, ao nível de significância de 5%, a hipótese  $H_0 : \beta = 5$  contra a hipótese alternativa  $H_A : \beta \neq 5$ .
  - d) Determine a estimativa de  $Y$  para  $X = 2$  e o intervalo de confiança para  $E(Y | X = 2)$ , ao nível de confiança de 95%.
- 2.22. A tabela ao lado mostra os valores de  $X$  e  $Y$  em uma amostra com 8 observações. Admite-se que essas variáveis estão relacionadas de acordo com o modelo usual de regressão linear simples. Pode-se verificar que  $\sum X = 32$ ,  $\sum Y = 96$ ,  $\sum X^2 = 168$ ,  $\sum Y^2 = 1340$  e  $\sum XY = 304$ .
- | $X$ | $Y$ |
|-----|-----|
| 1   | 15  |
| 1   | 19  |
| 3   | 16  |
| 3   | 14  |
| 5   | 9   |
| 5   | 13  |
| 7   | 6   |
| 7   | 4   |

- a) Determine a equação de regressão de  $Y$  contra  $X$  de acordo com o método de mínimos quadrados.
- b) Obtenha uma estimativa não-tendenciosa da variância do erro do modelo.
- c) Verifique se a influência de  $X$  sobre  $Y$  é estatisticamente significativa ao nível de 1%.
- d) Determine o intervalo de 90% de confiança para a variação esperada de  $Y$  quando  $X$  diminui 3 unidades ( $\Delta X = -3$ ).

- 2.23. São dados 3 conjuntos de 6 pares de valores  $(X_i, Y_i, i = 1, \dots, 6)$

Conjunto A		Conjunto B		Conjunto C	
X	Y	X	Y	X	Y
1	5,5	1	10,5	1	0,5
2	6,5	2	9,5	2	1,5
3	9,0	3	10,0	3	4,0
4	10,5	4	9,5	4	5,5
5	13,5	5	10,5	5	8,5
6	15,0	6	10,0	6	10,0

Para cada um desses conjuntos, obtenha as estimativas de mínimos quadrados dos parâmetros da regressão linear de  $Y$  contra  $X$ . Calcule os valores do coeficiente de determinação e do coeficiente de variação e analise comparativamente os resultados. Para melhor visualização, faça, para cada conjunto, um gráfico mostrando os pontos observados e a reta de regressão ajustada.

- 2.24. Dada uma amostra de  $n$  pares de valores  $(X_i, Y_i, i = 1, \dots, n)$ , mostre que a estimativa dos coeficiente angular da reta, obtida através do método dos mínimos quadrados ( $b$ ), é uma média ponderada das declividades das retas que passam pelos pontos  $(X_i, Y_i)$  e pelo ponto central da amostra  $(\bar{X}, \bar{Y})$ .

- 2.25. A partir de uma amostra de 7 pares de valores, foi obtida a equação de regressão

$$\hat{Y} = 30 + 5X ,$$

com um coeficiente de determinação  $r^2 = \frac{2}{3}$

A estimativa do desvio padrão de  $X$  é  $s(X) = 2$ .

- a) Determine o intervalo de confiança do coeficiente de regressão, ao nível de confiança de 95%.
- b) Teste, ao nível de significância de 5%, a hipótese de que o coeficiente de regressão da população é 8,5, considerando a hipótese alternativa de que o coeficiente de regressão da população é menor do que 8,5.

- 2.26. Seja  $Y$  o custo de carregamento mecânico por tonelada de cana-de-açúcar. Seja  $X$  o número de toneladas carregadas, por carregadeira e por ano. Suponha que um pesquisador levantou os custos de carregamento mecânico da cana-de-açúcar em diversas propriedades, obtendo uma amostra de pares de valores  $(X_i, Y_i)$ .

Admitindo que os custos totais de carregamento mecânico por ano sejam constituídos por uma parte fixa (que não varia com  $X$ ) e por uma parte variável (de tal maneira que o custo variável por tonelada seja constante), que modelo matemático deve ser usado para estudar, por meio da análise de regressão, a variação do custo de carregamento por tonelada de cana-de-açúcar em função do número de toneladas carregadas? Que anamorfose deverá ser feita?

- 2.27. Analisando a série de valores do Produto Nacional Bruto (PNB) de determinado país, durante um período de 10 anos, verificou-se que são aproximadamente constantes os incrementos anuais relativos do PNB. Qual é a equação (modelo matemático) que deve ser usada na análise de regressão desses dados? Que transformação de variáveis (anamorfose) deve ser feita para determinar as estimativas dos parâmetros através do método dos mínimos quadrados? Sabendo que, utilizando logaritmos decimais, a estimativa do coeficiente de regressão é 0,0193 e a estimativa do respectivo desvio padrão é 0,0010, teste, ao nível de significância de 5%, a hipótese de que a taxa de crescimento é 4% ao ano (sabe-se que  $\log 104 = 2,0170$ ).
- 2.28. Admitindo que as variáveis  $X$  e  $Y$  estão relacionadas conforme o modelo  $Y_i = \alpha X_i^\beta \varepsilon_i$ , onde  $\varepsilon_i$  são erros multiplicativos, determine as estimativas dos parâmetros  $\alpha$  e  $\beta$  com base nos seguintes dados:

$X$	$Y$
1	1
1	10
100	1.000
100	1.000
10.000	100
10.000	1.000

- 2.29. Suponha que um pesquisador está determinando a função de demanda do produto  $A$  em determinado mercado, com base em uma série de 8 pares de valores  $X_i$ ,  $Y_i$ , onde  $Y_i$  é o preço pelo qual foi vendida a quantidade  $X_i$  do produto em determinado intervalo de tempo. Admitindo que a elasticidade-preço da demanda do produto é constante, qual é a equação (modelo matemático) que o pesquisador deve usar? Que transformação de variável (anamorfose) deverá ser feita para determinar as estimativas dos parâmetros através do método dos

mínimos quadrados? Sabendo que a estimativa do coeficiente de regressão obtida é  $-1,24$ , com um desvio padrão estimado em  $0,10$ , teste, ao nível de significância de  $5\%$ , a hipótese de que a elasticidade-preço é igual a  $-1$ .

- 2.30. Seja  $Y$  uma grandeza econômica qualquer e seja  $X$  o tempo, em anos. Se admitirmos que a taxa geométrica de crescimento de  $Y$  é constante, que modelo de regressão deve ser adotado? Sabendo que, em cinco anos consecutivos,  $Y$  assumiu os valores  $4, 4, 32, 64$  e  $32$ , qual é a estimativa da taxa de crescimento de acordo com o método dos mínimos quadrados? Faça a análise de variância da regressão.
- 2.31. Em estudos da variação do consumo de certos produtos em função da renda da família tem sido usada a função  $Y = \exp\left(\alpha - \frac{\beta}{X}\right)$ , onde  $Y$  é o dispêndio com o produto considerado e  $X$  é a renda da família.  
Mostre as anamorfoses que devem ser feitas para que as fórmulas de regressão linear simples sejam usadas para ajustar essa função, utilizando dados obtidos de uma amostra aleatória.
- 2.32. a) Deduza, de acordo com o método dos mínimos quadrados, a fórmula para estimar o parâmetro do modelo

$$Y_i = \beta X_i + u_i \quad \text{com } i = 1, \dots, n,$$

onde

$$E(u_i) = 0,$$

$$E(u_i^2) = \sigma^2$$

e

$$E(u_i u_j) = 0 \quad \text{para } i \neq j$$

- b) Prove que o estimador obtido (b) é não-tendencioso
- c) Prove que a variância da estimativa obtida é  $V(b) = \frac{\sigma^2}{\sum X_i^2}$
- d) Prove que o estimador obtido é um estimador linear não-tendencioso de variância mínima
- e) Mostre que a soma de quadrados residual é dada por

$$\sum Y_i^2 - \frac{(\sum X_i Y_i)^2}{\sum X_i^2} = \sum Y_i^2 - b \sum X_i Y_i$$

- f) Demonstre que a esperança da soma de quadrados residual é igual a  $(n-1)\sigma^2$
- g) Admitindo que  $Y$  é a receita de uma empresa comercial em certo intervalo de tempo e que  $X$  é a quantidade vendida (em unidades físicas), ajuste aos pares de valores, dados a seguir, uma reta que passe pela origem dos eixos. Teste, ao nível de significância de 5%, a hipótese  $H_0 : \beta = 0$ .

$X$	$Y$
2	5
3	7
4	11
4	5
5	9

2.33. Dados um conjunto de pares de valores  $X_{ij}$ ,  $Y_{ij}$  ( $i = 1, \dots, m; j = 1, \dots, n$ ), ajustase um conjunto de  $m$  retas paralelas

$$\hat{Y}_{ij} = a_i + bX_{ij}$$

Mostre que as estimativas dos parâmetros, de acordo com o método dos mínimos quadrados, são dadas por

$$b = \frac{\sum_i \sum_j (X_{ij} - \bar{X}_i)(Y_{ij} - \bar{Y}_i)}{\sum_i \sum_j (X_{ij} - \bar{X}_i)^2}$$

e

$$a_i = \bar{Y}_i - b\bar{X}_i$$

$$\text{onde } \bar{X}_i = \frac{1}{n} \sum_{j=1}^n X_{ij} \text{ e } \bar{Y}_i = \frac{1}{n} \sum_{j=1}^n Y_{ij}$$

(Extraído de DRAPER e SMITH, 1996, p. 38).

2.34. É dada uma amostra de 12 pares de valores

$X_i$	$Y_i$	$X_i$	$Y_i$
1	2	4	9
1	4	4	13
1	3	5	11
1	5	5	10
2	8	5	16
2	6	5	9

Obtemos  $\sum Y_i = 96$ ,  $\sum Y_i^2 = 962$ ,  $\sum y_i^2 = 194$

Admite-se que as variáveis  $X$  e  $Y$  estão relacionadas de acordo com o modelo  $Y_i = \alpha + \beta X_i + u_i$ , onde os  $u_i$  são variáveis aleatórias independentes com distribuição normal de média zero e variância  $\sigma^2$ .

- a) Determine as estimativas dos parâmetros da regressão linear.
- b) Calcule o coeficiente de determinação da regressão e faça a análise de variância, interpretando o teste  $F$  realizado. Considere um nível de significância de 1%.
- c) Teste, ao nível de significância de 1%, a hipótese de que  $\beta = 0$  contra a hipótese de que  $\beta > 0$ .
- d) Determine a estimativa de  $Y$  para  $X = 6$  e o intervalo de confiança para  $E(Y | X = 6)$ , ao nível de significância de 99%.
- e) Determine o valor da estimativa da variação em  $E(Y)$ , isto é, estime  $\theta = E(\Delta Y)$ , quando o valor de  $X$  aumenta de 2 unidades ( $\Delta X = 2$ ). Qual é a variância de  $\hat{\theta}$ ? Teste, ao nível de significância de 5%, a hipótese de que  $\theta = 2,5$  contra a hipótese alternativa de que  $\theta > 2,5$ .

2.35. Mostre que a covariância entre duas estimativas de  $Y$  ( $\hat{Y}_1$  e  $\hat{Y}_2$ , para  $X = X_1$  e  $X = X_2$ , respectivamente) é

$$\text{cov}(\hat{Y}_1, \hat{Y}_2) = \left( \frac{1}{n} + \frac{x_1 x_2}{\sum x_i^2} \right) \sigma^2$$

2.36. Seja  $X$  a quantidade de adubo colocada no solo, em doses por hectare, e seja  $Y$  a produtividade obtida, em toneladas por hectare. Admite-se que essas variáveis estão relacionadas de acordo com a função

$$Y_i = \alpha + \beta \sqrt{X_i} + u_i,$$

onde  $\alpha$  e  $\beta$  são parâmetros e os  $u_i$  são variáveis aleatórias independentes com distribuição normal de média zero e variância  $\sigma^2$ .

Um experimento com 5 parcelas forneceu os seguintes resultados:

$X_i$	$Y_i$
0	2,7
1	4,4
4	5,3
9	5,4
16	7,2

- a) Determine as estimativas de mínimos quadrados ( $a$  e  $b$ ) de  $\alpha$  e  $\beta$ .
  - b) Calcule o coeficiente de determinação da regressão.
  - c) Determine o intervalo de 95% de confiança para  $\alpha$ .
  - d) Determine o intervalo de 95% de confiança para  $\beta$ .
  - e) Determine a dose econômica de adubo admitindo que o preço da tonelada do produto seja igual ao dobro do preço da dose de adubo (já considerados os custos de colocação do adubo, juros e/ou subsídios, etc.).
  - f) Determine o intervalo de 95% de confiança para o verdadeiro valor ( $\chi$ ) da dose econômica, para as condições do item anterior (Sugestão: determine inicialmente, os limites do intervalo de confiança para  $\sqrt{\chi}$ , e depois eleve ao quadrado).
  - g) Determine o intervalo de previsão para a produção de uma nova parcela com  $X = 1$ , considerando um nível de confiança de 95%.
- 2.37. Considere o modelo  $Y_i = \beta X_i + u_i$  com  $X_i$  fixos,  $E(u_i) = 0$ ,  $E(u_i^2) = \sigma^2$  e  $E(u_i u_j) = 0$  para  $i \neq j$ .

Sabe-se que o estimador de mínimos quadrados para  $\beta$  é  $b = \frac{\sum X_i Y_i}{\sum X_i^2}$ , não-

tendencioso, com  $V(b) = \frac{\sigma^2}{\sum X_i^2}$  (ver exercício 2.32).

Um estimador alternativo para  $\beta$  é  $\hat{\beta} = \bar{Y} / \bar{X}$ , que é a inclinação da reta unindo a origem do sistema de eixos ao ponto  $\bar{X}, \bar{Y}$ .

- a) Prove que  $\hat{\beta}$  é um estimador linear não-tendencioso.
- b) Deduza a expressão que dá  $V(\hat{\beta})$  em função de  $\sigma^2$  e dos valores de  $X$ .
- c) Prove (sem utilizar o teorema de Gaus-Markov) que  $V(\hat{\beta}) \geq V(b)$ . Em que condições tem-se  $V(\hat{\beta}) = V(b)$ ?
- 2.38. Considere o modelo de regressão linear simples  $Y_i = \alpha + \beta X_i + u_i$ , onde  $u_i$  são erros aleatórios independentes com média zero e variância  $\sigma^2$ , e os  $X_i$  são fixos, com  $X_i = i$  para  $i = 1, 2, \dots, 9$ .
- Sejam  $\bar{X}_1$  e  $\bar{X}_2$  as médias de  $X$  para as  $h$  primeiras e as  $h$  últimas observações, isto é,

$$\bar{X}_1 = \frac{1}{h} \sum_{i=1}^h X_i \quad \text{e} \quad \bar{X}_2 = \frac{1}{h} \sum_{i=10-h}^9 X_i$$

Verifique os valores de  $\bar{X}_1$  e  $\bar{X}_2$  apresentados na tabela abaixo

$h$	$\bar{X}_1$	$\bar{X}_2$
1	1	9
2	1,5	8,5
3	2	8
4	2,5	7,5

Sejam  $\bar{Y}_1$  e  $\bar{Y}_2$  as médias dos correspondentes valores de  $Y$ , isto é,

$$\bar{Y}_1 = \frac{1}{h} \sum_{i=1}^h Y_i \quad \text{e} \quad \bar{Y}_2 = \frac{1}{h} \sum_{i=10-h}^9 Y_i$$

Define-se o seguinte estimador de  $\beta$ :

$$b_{*h} = \frac{\bar{Y}_2 - \bar{Y}_1}{\bar{X}_2 - \bar{X}_1}$$

a) Prove que  $b_{*h} = \beta + \frac{\bar{u}_2 - \bar{u}_1}{\bar{X}_2 - \bar{X}_1}$ ,

onde  $\bar{u}_1 = \frac{1}{h} \sum_{i=1}^h u_i$  e  $\bar{u}_2 = \frac{1}{h} \sum_{i=10-h}^9 u_i$

b) Mostre que  $b_{*h}$  é um estimador não-tendencioso de  $\beta$ .

c) Demonstre que  $V(b_{*h}) = \frac{2\sigma^2}{h(\bar{X}_2 - \bar{X}_1)^2}$

d) Faça uma tabela mostrando os valores de  $V(b_{*h})$  para  $h = 1, 2, 3, 4$ .

Apresente, na mesma tabela, para cada valor de  $h$ , a eficiência relativa de  $b_{*h}$  em comparação com o estimador de mínimos quadrados.

## Respostas

2.1. a)  $\hat{Y} = 4 + 1,9X$

b)  $F = 320,89$ , rejeita-se  $H_0 : \beta = 0$

c)  $r^2 = 0,976$

d)  $\hat{Y} = 9,7$ . Os limites do intervalo de confiança são 8,89 e 10,51.

2.4. a)  $\hat{Y} = 12 + 4X ; s(b) = 1; s(a) = 10,2$

b)  $r^2 = \frac{4}{7} = 0,571$

c)  $t = 4$ , significativo

2.5. a)  $\hat{Y} = 2 + 2X$

b)  $r^2 = 0,842 ; F = 16$ , significativo

c)  $t = 8$ , significativo

d)  $t = -6,63$ , significativo

e)  $\hat{Y} = 12; 8,10$  a  $15,90$

2.6. a)  $\hat{Y} = 3X$

b)  $r^2 = \frac{12}{13} = 0,923$

c)  $F = 36$ , significativo

d)  $t = -6$ , não-significativo

e)  $t = -2,45$ , significativo

f) 15,42 a 44,58

2.7.  $r^2 = \frac{2}{3}; F = 100$ , significativo

2.8. Uma vez que  $E(b) = \beta$ , basta mostrar que  $\lim_{n \rightarrow \infty} V(b) = 0$ , o que acontece desde que

$\sum x^2$  cresce indefinidamente quando  $n$  cresce.

2.9.  $d = b + 1$  e  $c = a$

2.10. Notando que os acréscimos relativos ( $\Delta Y / Y$ ) de  $Y$  são aproximadamente constantes, conclui-se que o modelo matemático apropriado é  $Y = \alpha\beta^X$ . A mesma conclusão é obtida notando que os pontos ( $X, Y$ ) estão aproximadamente alinhados, quando marcados em um gráfico com o eixo das ordenadas em escala logarítmica, isto é, notando que os pontos de coordenadas  $\log Y$  e  $X$  estão aproximadamente alinhados.

2.11. a)  $2 \pm 0,202$

b)  $t = 3,059$ , significativo

2.14. a)  $1\,025 \pm 250$

b) 124,3

2.15.  $t = -4$ , não-significativo

2.16. a)  $-0,06$  a  $2,72$

b)  $t = 2,954$ , significativo

2.17.  $\hat{Y} = 4,5 + \frac{60}{X}$

2.18. a)  $a = 0, b = 10$ ,  $\hat{V}(a) = 0,0175$ ,  $\hat{V}(b) = 0,417$

b) I)  $t = -15,12$ , significativo

II)  $t = -3,10$ , significativo

c)  $r^2 = 0,833$  ou 83,3%

2.19. a)  $\hat{Y} = 3 - 0,5X$

b)  $r^2 = \frac{1}{3}; F = 4$ , não-significativo

c)  $t = -2$ , não-significativo

d)  $t = 3,27$ , não-significativo

e)  $-2,84$  a  $3,68$

f)  $X = 4; \hat{Y} = 1; -0,41$  a  $2,41$

2.20. a)  $\hat{Y} = 4 + 3X$

b)  $t = 17,23$ , significativo; rejeita-se  $H_0 : \beta = 0$

c)  $1,51 < \alpha < 6,49$

d)  $r^2 = 0,974$

e)  $X = 12$ , isto é, 12 000 unidades

f)  $34 \pm 2,14$

2.21. a)  $\hat{Y} = 6 + 1,5X$

b)  $r^2 = \frac{27}{52} = 0,519$

c)  $t = -3,429$ , não-significativo ( $t_0 = 4,303$ )

d)  $\hat{Y} = 9$

$$3,62 < E(Y | X = 2) < 14,38$$

2.22. a)  $\hat{Y} = 20 - 2X$

b)  $s^2 = \frac{14}{3} = 4,667$

c)  $F = 34,29$ , significativo ( $F_0 = 13,7$ )

d)  $4,009 < E(\Delta Y) < 7,991$

2.23.

Estatística	Conjunto		
	A	B	C
a	3	10	-2
b	2	0	2
$\bar{Y}$	10	10	5
S.Q.Res.	1	1	1
S.Q.Regr.	70	0	70
$r^2$	98,6%	0	98,6%
CV	5%	5%	10%

2.25. a)  $0,94 < \beta < 9,06$

b)  $t = -2,21$ , significativo

2.26. Anamorfose:  $\frac{1}{X} = V$

2.27. Anamorfose:  $Y = \log(PNB)$

$t = 2,30$ , não-significativo

2.28.  $\hat{Y} = 10X^{0,5}$

2.29.  $Y = AX^B$

Anamorfoses:  $Z = \log Y$  e  $V = \log X$

$t = -2,40$ , não-significativo

2.30.  $Y_i = \alpha\beta^{X_i}\varepsilon_i$

Adotando como origem do tempo ( $X = 0$ ) o ano em que foi efetuada a terceira das observações consideradas, obtemos

$$\hat{Y} = 16 \cdot 2^X$$

A taxa de crescimento é 100% ao ano

$$F = 7,5$$

2.31. Anamorfoses:  $Z = \ln Y$  e  $V = \frac{1}{X}$

2.32. g)  $\hat{Y} = 2X$ ;  $t = 7,303$ , significativo.

2.34. a)  $\hat{Y} = 2 + 2X$

b)  $r^2 = 0,742$ ;  $F = 28,8$ , significativo ( $F_0 = 10,04$ )

c)  $t = 5,37$ , significativo ( $t_0 = 2,76$ )

d)  $\hat{Y} = 14$ ;  $9,91 < E(Y|X = 6) < 18,09$

e)  $\hat{\theta} = 4$ ;  $V(\hat{\theta}) = (\Delta X)^2 V(b) = \frac{\sigma^2}{9}$ ;

$$\hat{V}(\hat{\theta}) = \frac{5}{9} \text{ e } t = 2,01, \text{ significativo } (t_0 = 1,81)$$

2.36. a)  $a = 3$ ,  $b = 1$

b)  $r^2 = 0,931$

c)  $1,78 < \alpha < 4,22$

d)  $0,50 < \beta < 1,50$

e) Uma dose por hectare

f)  $0,50 < \sqrt{\chi} < 1,50$

ou  $0,25 < \chi < 2,25$

g)  $2,20 < Y_h < 5,80$

2.37. b)  $V(\hat{\beta}) = \frac{n\sigma^2}{(\sum X)^2}$

c) As variâncias são iguais apenas quando todos os valores de  $X$  forem iguais.

$$2.38. V(b) = \frac{\sigma^2}{60}$$

$h$	$\bar{X}_2 - \bar{X}_1$	$V(b_{*_h})$	Eficiência relativa
1	8	$\sigma^2 / 32$	0,533
2	7	$\sigma^2 / 49$	0,817
3	6	$\sigma^2 / 54$	0,900
4	5	$\sigma^2 / 50$	0,833

### 3. CORRELAÇÃO

Vimos que numa análise de regressão linear simples, se determina, através de estimativas dos parâmetros, como uma variável  $X$  exerce, ou parece exercer, efeito sobre uma outra variável  $Y$ .

Na análise de correlação, que veremos aqui, se procura determinar o grau de relacionamento entre duas variáveis, ou seja, se procura medir a covariabilidade entre elas.

Na análise de regressão é necessário distinguir a variável dependente e a variável explanatória; na análise de correlação, tal distinção não é necessária.

#### 3.1. O coeficiente de correlação simples para uma amostra

Inicialmente, desenvolveremos o conceito do coeficiente de correlação ( $r$ ) para uma amostra de  $n$  pares de valores  $X_i, Y_i$  ( $i = 1, 2, \dots, n$ ).

Para obter uma medida de correlação sem a influência da média (tendência central) e da variância (dispersão), vamos utilizar variáveis reduzidas, definidas por

$$v_i = \frac{X_i - \bar{X}}{s(X)} = \frac{x_i}{\sqrt{\frac{\sum x_i^2}{n-1}}} \quad (3.1)$$

e

$$z_i = \frac{Y_i - \bar{Y}}{s(Y)} = \frac{y_i}{\sqrt{\frac{\sum y_i^2}{n-1}}} \quad (3.2)$$

Como as variáveis reduzidas não têm dimensão, esta transformação também elimina qualquer influência da unidade de medida.

As figuras 3.1, 3.2 e 3.3 apresentam três diferentes resultados que poderiam ser obtidos quando colocamos os pontos  $(v_i, z_i)$  em um gráfico.

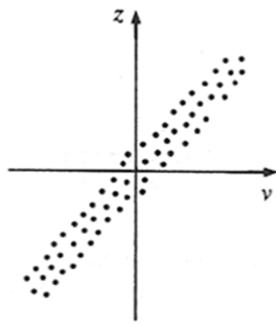


Figura 3.1. Correlação Positiva

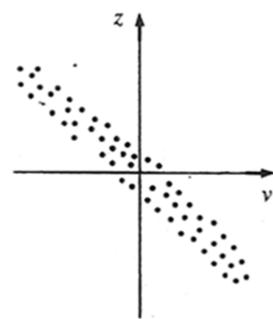


Figura 3.2. Correlação negativa

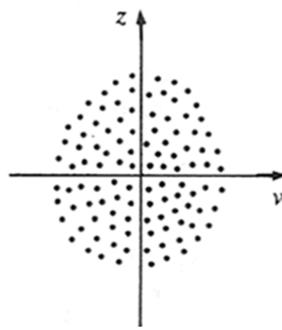


Figura 3.3. Correlação aproximadamente igual a zero

Se  $X$  e  $Y$  estão positivamente correlacionados, isto é, se  $X$  e  $Y$  tendem a variar no mesmo sentido, então a maioria dos pontos  $(v_i, z_i)$  estará no 1º e no 3º quadrantes, como ocorre na figura 3.1. Uma vez que, para pontos localizados nesses quadrantes, o produto  $v_i z_i$  é positivo, o valor de  $\sum v_i z_i$  será, neste caso, positivo e relativamente alto.

Se  $X$  e  $Y$  estão negativamente correlacionados, isto é, se  $X$  e  $Y$  tendem a variar em sentidos opostos, então a maioria dos pontos  $(v_i, z_i)$  estará no 2º e no 4º quadrantes, como ocorre na figura 3.2. Uma vez que, para pontos localizados nesses quadrantes, o produto  $v_i z_i$  é negativo, o valor de  $\sum v_i z_i$  será, neste caso, negativo e de valor absoluto relativamente alto.

Se não existe correlação, os pontos  $(v_i, z_i)$  estarão distribuídos pelos quatro quadrantes, como ocorre na figura 3.3. Então  $\sum v_i z_i$  será igual a zero ou terá valor absoluto pequeno, pois as parcelas positivas (correspondendo a pontos no 1º e 3º quadrantes) são anuladas pelas parcelas negativas (correspondendo a pontos no 2º e 4º quadrantes).

Portanto, o valor de  $\sum v_i z_i$  pode ser utilizado como medida de correlação. Entretanto, em termos absolutos, esse valor tende a crescer com o número de observações. Então, o coeficiente de correlação simples é definido por

$$r = \frac{\sum v_i z_i}{n-1}$$

Considerando (3.1) e (3.2), obtemos

$$r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} \quad (3.3)$$

Comparando (3.3) com (2.28), verificamos que o quadrado do coeficiente de correlação é igual ao coeficiente de determinação da regressão linear simples.

Já vimos que

$$0 \leq r^2 \leq 1$$

Então,

$$-1 \leq r \leq 1$$

É importante assinalar que um coeficiente de correlação igual a zero não implica em ausência de relação entre as duas variáveis. Isso é mostrado na figura 3.4, onde, apesar de o coeficiente de correlação ser nulo, é evidente que existe uma relação parabólica entre  $X$  e  $Y$ . Portanto, um coeficiente de correlação nulo somente implica ausência de relação *linear* entre as duas variáveis.

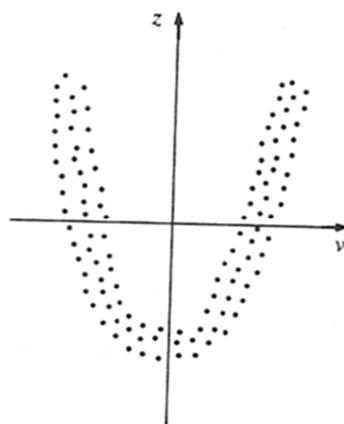


Figura 3.4. Relação parabólica entre  $X$  e  $Y$ , onde  $r=0$

Para exemplificar, consideremos os 6 pares de valores dados na tabela 3.1 e representados na figura 3.5. Pode-se imaginar que cada par de valores são as notas tiradas por um aluno em duas disciplinas.

Tabela 3.1 Amostra de 6 pares de valores  $X_i$ ,  $Y_i$

$X_i$	$Y_i$	$X_i$	$Y_i$
4	6	6	8
4	7	8	7
6	6	8	8

Obtemos

$$\bar{X} = \frac{36}{6} = 6; \quad \bar{Y} = \frac{42}{6} = 7$$

$$\sum x^2 = \sum X_i^2 - \frac{(\sum X_i)^2}{n} = 232 - \frac{36^2}{6} = 16$$

$$\sum y^2 = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n} = 298 - \frac{42^2}{6} = 4$$

$$\sum x_i y_i = \sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n} = 256 - \frac{36 \cdot 42}{6} = 4$$

$$r = \frac{4}{\sqrt{16 \cdot 4}} = 0,5$$

Vejamos a relação que existe entre o coeficiente de correlação e o coeficiente de regressão.

Como

$$b = \frac{\sum xy}{\sum x^2} = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} \sqrt{\frac{\sum y^2}{\sum x^2}},$$

verificamos, considerando (3.3), que

$$b = r \sqrt{\frac{\sum y^2}{\sum x^2}} = r \frac{s(Y)}{s(X)} \quad (3.4)$$

onde

$$s(X) = \sqrt{\frac{\sum x_i^2}{n-1}} \quad \text{e} \quad s(Y) = \sqrt{\frac{\sum y_i^2}{n-1}}$$

Mostraremos agora que o quadrado do coeficiente de correlação é igual ao produto das estimativas dos coeficientes de regressão de  $Y$  em relação a  $X$  e de  $X$  em

relação a  $Y$ . Representando essas estimativas por  $b_{Y \cdot X}$  e  $b_{X \cdot Y}$  respectivamente, podemos escrever

$$b_{Y \cdot X} = \frac{\sum xy}{\sum x^2} \quad \text{e} \quad b_{X \cdot Y} = \frac{\sum xy}{\sum y^2}$$

Segue-se, imediatamente, que

$$r^2 = b_{Y \cdot X} b_{X \cdot Y} \quad (3.5)$$

Para a amostra apresentada na tabela 3.1, temos:

$$b_{Y \cdot X} = \frac{\sum xy}{\sum x^2} = \frac{4}{16} = 0,25, \quad b_{X \cdot Y} = \frac{\sum xy}{\sum y^2} = \frac{4}{4} = 1$$

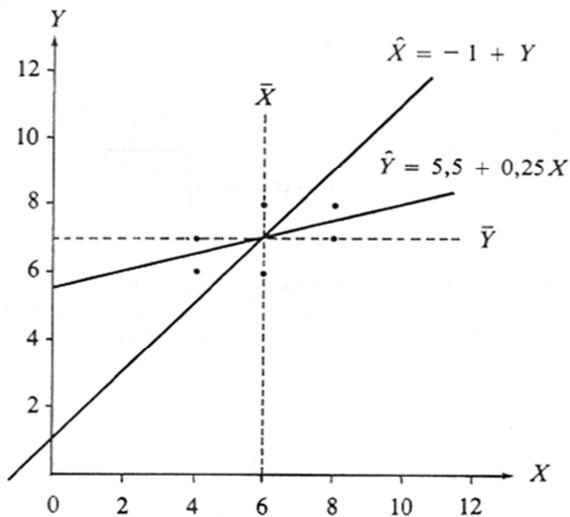
$$\text{e} \quad b_{Y \cdot X} b_{X \cdot Y} = 0,25 = r^2$$

Também podemos obter as retas de regressão de  $Y$  em relação a  $X$  e de  $X$  em relação a  $Y$ , que são, respectivamente,

$$\hat{Y} = 5,5 + 0,25X$$

e

$$\hat{X} = -1 + Y$$



*Figura 3.5.* Retas de regressão de  $Y$  em relação a  $X$  e de  $X$  em relação a  $Y$ , para os dados da tabela 3.1.

Para ilustrar melhor o conceito de correlação, consideremos um outro exemplo. A tabela 3.2, transcrita de Yule e Kendall (1940), apresenta as freqüências (em centenas) de casamentos na Inglaterra e na Irlanda, em 1933, conforme as idades do marido ( $X$ ) e da mulher ( $Y$ ).

TABELA 3.2. Número de casamentos em função da idade do marido e da mulher, na Inglaterra e na Irlanda, em 1933.

Idade da mulher	Idade do marido em anos (limite inferior do intervalo)													Total
	15	20	25	30	35	40	45	50	55	60	65	70	75	
15	33	189	56	8	2	—	—	—	—	—	—	—	—	288
20	18	682	585	106	19	5	2	1	—	—	—	—	—	1418
25	1	140	511	179	40	14	6	3	1	1	—	—	—	896
30	—	11	75	101	42	20	10	5	2	1	1	—	—	268
35	—	2	10	24	28	19	13	8	5	2	1	—	—	112
40	—	—	1	5	9	14	12	10	6	4	2	1	—	64
45	—	—	—	1	3	5	9	9	7	4	3	1	—	42
50	—	—	—	—	—	1	3	7	6	5	3	1	—	26
55	—	—	—	—	—	—	1	3	5	4	3	1	—	17
60	—	—	—	—	—	—	—	1	1	4	3	2	—	11
65	—	—	—	—	—	—	—	—	1	1	3	2	1	8
70	—	—	—	—	—	—	—	—	—	—	1	1	1	3
Total	52	1024	1238	424	143	78	56	47	34	26	20	9	2	3153

Fonte: Yule e Kendall (1940), p. 198.

Podemos, para facilitar os cálculos, utilizar as seguintes variáveis auxiliares:

$$V_i = \frac{X_i - 27,5}{5}, \quad i = 1, 2, \dots, 13 \quad (3.6)$$

e

$$Z_j = \frac{Y_j - 27,5}{5}, \quad j = 1, 2, \dots, 12 \quad (3.7)$$

Devemos ressaltar que o coeficiente de correlação entre  $V$  e  $Z$  é igual ao coeficiente de correlação entre  $X$  e  $Y$  (ver exercício 3.9).

Note que essas variáveis auxiliares assumem valor zero no ponto médio da classe de 25 a 30 anos e são medidas em unidades de 5 anos.

Representando por  $f_{ij}$  as freqüências em cada cela, por  $F_i$  as freqüências totais para cada classe de idade do marido e por  $G_i$  as freqüências totais para cada classe de idade da mulher, temos:

$$\bar{V} = \frac{\sum V_i F_i}{n} = \frac{922}{3153} = 0,2924$$

$$\bar{Z} = \frac{\sum Z_j G_j}{n} = \frac{-742}{3153} = -0,2353$$

$$\sum v_i^2 F_i = \sum V_i^2 F_i - \frac{(\sum V_i F_i)^2}{n} = 9708 - \frac{922^2}{3153} = 9438,39$$

$$\sum z_j^2 G_j = \sum Z_j^2 G_j - \frac{(\sum Z_j G_j)^2}{n} = 7090 - \frac{742^2}{3153} = 6915,38$$

$$\sum \sum v_i z_j f_{ij} = \sum \sum V_i Z_j f_{ij} - \frac{(\sum V_i F_i)(\sum Z_j G_j)}{n} = 6256 - \frac{922(-742)}{3153} = 6472,98$$

$$r = \frac{\sum \sum v_i z_j f_{ij}}{\sqrt{(\sum v_i^2 F_i)(\sum z_j^2 G_j)}} = 0,8012$$

As retas de regressão de  $Z$  em relação a  $V$  e de  $V$  em relação a  $Z$  são, respectivamente,

$$\hat{Z} = -0,2353 + 0,6858(V - 0,2924)$$

e

$$\hat{V} = 0,2924 + 0,9360(Z + 0,2353)$$

Considerando (3.6) e (3.7) obtemos, após simplificações, as equações de regressão de  $Y$  em relação a  $X$  e de  $X$  em relação a  $Y$ :

$$\hat{Y} = 6,5 + 0,686X$$

e

$$\hat{X} = 4,3 + 0,936Y$$

É interessante assinalar, na tabela 3.2, as celas modais das distribuições condicionais de  $Y | X$ ; elas mostram, grosseiramente, a posição da reta de regressão de  $Y$  em relação a  $X$ . Da mesma maneira, as celas modais das distribuições de  $X | Y$  mostram, aproximadamente, a posição da reta de regressão de  $X$  em relação a  $Y$ .

### 3.2. Aplicação da análise de regressão a uma população com distribuição normal bidimensional

O coeficiente de correlação de uma população é definido por

$$\rho = E\left(\frac{X - \mu_X}{\sigma_X} \cdot \frac{Y - \mu_Y}{\sigma_Y}\right) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\text{cov}(X, Y)}{\sqrt{V(X)V(Y)}}$$

Devemos lembrar, aqui, que:

- Se  $X$  e  $Y$  são independentes, temos  $\text{cov}(X, Y) = 0$  e, portanto,  $\rho = 0$ .
- Dados  $\text{cov}(X, Y) = 0$  e  $\rho = 0$ , não é possível concluir, em geral, que as variáveis são independentes. Isto é mostrado no exemplo apresentado na tabela 1.3 (seção 1.5) e no caso ilustrado na figura 3.4.
- Se as variáveis têm distribuição normal, demonstra-se que  $\text{cov}(X, Y) = \rho = 0$  é condição suficiente para que as variáveis sejam independentes.

Para fazer inferência estatística a respeito de  $\rho$ , partindo do coeficiente de correlação ( $r$ ) da amostra, pressupomos que as variáveis  $X$  e  $Y$  apresentam uma distribuição normal bidimensional. Então estamos excluindo casos como o representado na figura 3.4.

Para testar a hipótese da nulidade  $H_0 : \rho = 0$  contra a hipótese alternativa  $H_A : \rho \neq 0$ , utilizamos o teste

$$F = \frac{r^2(n-2)}{1-r^2}, \text{ com } 1 \text{ e } n-2 \text{ graus de liberdade.}$$

Pode-se verificar que o valor de  $F$ , obtido por essa fórmula, é igual ao valor de  $F$  da análise de variância da regressão, obtido dividindo o quadrado médio de regressão pelo quadrado médio residual. Portanto, testar a hipótese  $H_0 : \rho = 0$  equivale a testar a hipótese  $H_0 : \beta = 0$ .

A função de densidade de uma distribuição normal bidimensional corresponde a uma superfície cujas seções, tanto na direção do eixo dos  $X$  como na direção do eixo dos  $Y$ , são curvas normais. As seções horizontais dessa superfície são elipses de isoprobabilidade, duas das quais estão traçadas na figura 3.6.

Vamos mostrar agora que os pontos  $C$ ,  $E$ ,  $F$  e  $G$  da figura 3.6, onde retas paralelas ao eixo dos  $Y$  tangenciam elipses de isoprobabilidade, são os pontos médios das distribuições condicionais de  $Y$ . Consideremos, particularmente, o plano perpendicular ao eixo dos  $X$  passando por  $A$ ; esse plano secciona infinitas elipses de isoprobabilidade, mas todas elas de nível inferior ao da elipse de isoprobabilidade que tangencia o plano em  $C$ , que é, portanto, a moda da curva normal definida pela intersecção do plano em questão com a superfície de densidade da população bidimensional. Como a moda de uma distribuição normal coincide com a média, concluímos que o ponto  $C$  corresponde à média da distribuição de  $Y$ , dado  $X = \overline{OA}$ . Para os pontos  $E$ ,  $F$  e  $G$  vale, evidentemente, o mesmo raciocínio. Considerando, ainda, que, se pode demonstrar que as distribuições condicionais de  $Y$  têm variância constante, concluímos que a reta  $GC$  é a verdadeira reta de regressão de  $Y$  em relação a  $X$ , ou seja, é a reta  $E(Y | X) = \alpha + \beta X$ .

Pode-se mostrar, analogamente, que a reta  $PL$  é a verdadeira reta de regressão de  $X$  em relação a  $Y$ .

A reta  $E(Y | X) = \alpha + \beta X$  (ou uma estimativa obtida de uma amostra) poderia ser usada para, dado um valor de  $X$ , prever o correspondente valor de  $Y$ .

Consideremos, por exemplo, que  $X$  e  $Y$  são, respectivamente, as notas de Matemática e Estatística obtidas por alunos dessas duas disciplinas. Se um aluno obteve nota  $\overline{OA}$  em Matemática, prevê-se que obtenha nota  $\overline{AC}$  em Estatística. É interessante notar que  $\overline{AC}$  é uma média ponderada de  $\overline{AD}$  e  $\overline{AB} = \mu_Y$ , com os pesos dependendo do valor de  $\rho$ . Para entender isso, consideremos, inicialmente, os casos extremos de  $\rho = 1$  e  $\rho = 0$ . À medida que  $\rho$  aumenta, as elipses de isoprobabilidade se alongam na direção do seu eixo principal e o ângulo entre as retas de regressão  $GC$  e  $PL$  diminui, de maneira que no limite, quando  $\rho = 1$ , as retas de regressão  $GC$  e  $PL$  coincidem com o eixo principal e a melhor estimativa de  $Y$  para  $X = \overline{OA}$  seria  $\overline{AD}$ . Por outro lado, quando  $\rho = 0$ , isto é, quando não existe correlação, a melhor estimativa de  $Y$  será  $\mu_Y$ , qualquer que seja o valor de  $X$ . Num caso intermediário em que  $0 < \rho < 1$ , a melhor estimativa de  $Y$  para  $X = \overline{OA}$  estará entre os valores  $\overline{AB} = \mu_Y$  e  $\overline{AD}$ , sendo próxima de  $\overline{AB}$  quando  $\rho$  for pequeno e se aproximando de  $\overline{AD}$  à medida que  $\rho$  se aproxima de 1.

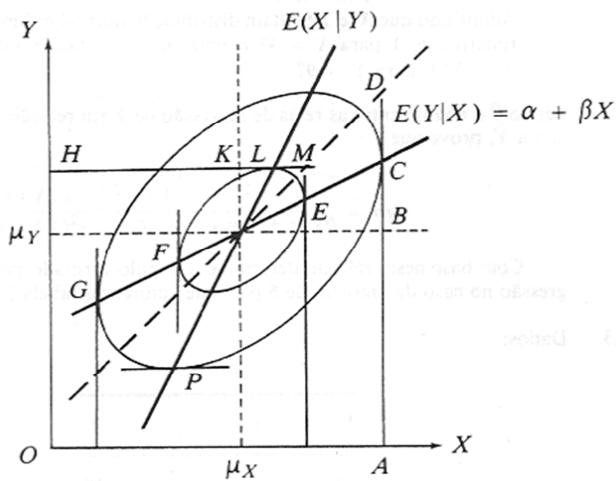


Figura 3.6. As elipses de isoprobabilidade de uma distribuição normal bidimensional e as retas de regressão de  $Y$  em relação a  $X$  e de  $X$  em relação a  $Y$ .

### EXERCÍCIOS

- 3.1. Calcule o coeficiente de correlação para a seguinte amostra de 10 pares de valores  $X_i, Y_i$ .

$X_i$	$Y_i$	$X_i$	$Y_i$
5	1	7	7
5	3	8	5
6	1	8	9
6	5	9	7
7	3	9	9

- a) Teste, ao nível de significância de 1%, a hipótese de nulidade  $H_0 : \rho = 0$  contra a hipótese alternativa  $\rho \neq 0$ .
- b) Determine as retas de regressão de  $Y$  em relação a  $X$  e de  $X$  em relação a  $Y$ . Verifique que  $b_{Y|X} b_{X|Y} = r^2$ .
- c) Admitindo que  $X$  e  $Y$  tenham distribuição normal bidimensional, qual é a estimativa de  $Y$  para  $X = 4$ ? E para  $X = 9$ ? Qual é a estimativa de  $X$  para  $Y = 3$ ? E para  $Y = 9$ ?

3.2. Sendo  $\theta$  o ângulo entre as retas de regressão de  $Y$  em relação a  $X$  e de  $X$  em relação a  $Y$ , prove que

$$\operatorname{tg} \theta = \frac{1 - r^2}{b_{Y.X} + b_{X.Y}} = \frac{1 - r^2}{r} \cdot \frac{s(X)s(Y)}{s^2(X) + s^2(Y)}$$

Com base nesta relação determine o ângulo formado pelas duas retas de regressão no caso da amostra de 6 pares de valores da tabela 3.1.

3.3. Dados:

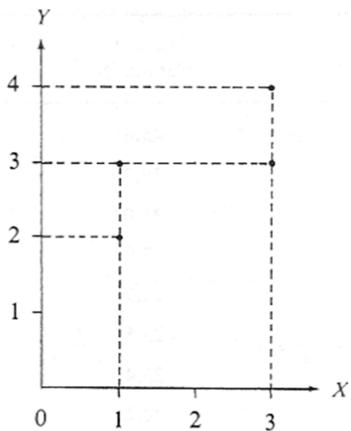
$X$	$Y$
2	18
4	12
5	10
6	8
8	7
11	5

$$\sum X = 36; \sum X^2 = 266; \sum Y = 60; \sum Y^2 = 706; \sum XY = 293$$

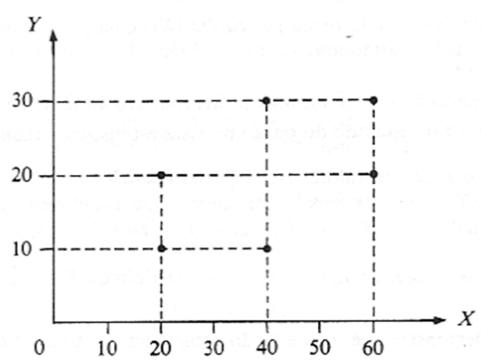
- a) Determine o coeficiente de correlação entre  $X$  e  $Y$ .
- b) Determine as estimativas dos parâmetros da equação de regressão linear de  $Y$  em relação a  $X$ .
- c) Admitindo que as variáveis  $X$  e  $Y$  estão relacionadas de acordo com o modelo  $Y_i = \alpha + \beta X_i + u_i$ , onde  $u_i$  são erros com média zero, variância constante e distribuição normal, teste a hipótese da nulidade  $H_0 : \beta = 1$ , contra a hipótese alternativa  $H_A : \beta > 1$ , considerando um nível de significância de 5%.

3.4. Com base no gráfico dado a seguir, determine geometricamente (sem usar as fórmulas comuns de análise de regressão):

- a) A reta de regressão de  $Y$  em relação a  $X$ .
- b) A reta de regressão de  $X$  em relação a  $Y$ .
- c) O coeficiente de correlação ( $r$ ).
- d) O valor estimado de  $Y$  para  $X = 1$ .
- e) O valor estimado de  $X$  para  $Y = 1$ .



- 3.5. Com base no gráfico dado a seguir, determine geometricamente (sem usar as fórmulas comuns de análise de regressão):
- A reta de regressão de  $Y$  em relação a  $X$ .
  - A reta de regressão de  $X$  em relação a  $Y$ .
  - O coeficiente de correlação ( $r$ ).
  - O valor estimado de  $Y$  para  $X = 60$ .
  - O valor estimado de  $X$  para  $Y = 30$ .



- 3.6. Os dados a seguir foram apresentados em defesa da tese de que dietas com alto teor de proteína reduzem a fertilidade. *a)* Estabeleça, sem calcular, o sinal do coeficiente de correlação entre as duas variáveis. *b)* Discuta se dados desse tipo são apropriados para estabelecer relações de “causa-e-efeito” entre essas variáveis.

Pais	Taxa de Natalidade	Teor de proteína na dieta
Formosa	45,6	4,7
Malaia	39,7	7,5
Índia	33,0	8,7
Japão	27,0	9,7
Iugoslávia	25,9	11,2
Grécia	23,5	15,2
Itália	23,4	15,2
Bulgária	22,2	16,8
Alemanha	20,0	37,3
Irlanda	19,1	46,7
Dinamarca	18,3	56,1
Austrália	18,0	59,9
EUA	17,9	61,4
Suécia	15,0	62,6

- 3.7. Com base em uma amostra de 200 pares de valores para as variáveis  $X$  e  $Y$  obtivemos um coeficiente de correlação igual a 0,02. Podemos afirmar que não existe relação entre essas variáveis? Explique.
- 3.8. Com base nos valores de renda *per capita* ( $X_1$ ) e da porcentagem de analfabetos ( $X_2$ ) para 20 países latino-americanos em 1950, obtivemos o coeficiente de correlação  $r = -0,6$ .
- Esse resultado é estatisticamente significativo ao nível de 1%?
  - Interprete o resultado do ponto de vista estatístico e econômico-social.
- 3.9. São dados  $n$  pares de valores  $X_i$ ,  $Y_i$  cujo coeficiente de correlação é  $r$ . Sendo  $Z_i = a + bY_i$  e  $V_i = k + hX_i$ , demonstre que o coeficiente de correlação entre  $V_i$  e  $Z_i$  é igual a  $r$ , se  $bh > 0$ , e é igual a  $-r$ , se  $bh < 0$  ( $a$ ,  $b$ ,  $k$  e  $h$  são constantes).
- 3.10. São dados os valores de  $Z_i$  ( $i = 1, \dots, n$ ). Definimos  $X_i = \frac{a}{Z_i}$  e  $Y_i = \frac{b}{Z_i}$ . Demonstre que, se  $a$  e  $b$  são constantes positivas, o coeficiente de correlação entre  $X_i$  e  $Y_i$  é igual a 1.

3.11. O coeficiente de correlação entre as variáveis  $X$  e  $Y$  é  $r = 0,60$ . Sabendo que  $s(X) = 1,50$ ,  $s(Y) = 2,00$ ,  $\bar{X} = 10$  e  $\bar{Y} = 20$ , determine a equação de regressão de  $Y$  em relação a  $X$ .

3.12. Com base em uma amostra de 27 pares de valores foi obtido o coeficiente de correlação  $r = 0,40$ . Teste, ao nível de significância de 5%, a hipótese de que o coeficiente de correlação das variáveis é zero.

3.13. O coeficiente de correlação obtido de uma amostra de  $n$  pares de valores  $X_i$ ,  $Y_i$  é  $r = 4/5$ . Sabendo que  $s(X) = 3$  e  $s(Y) = 5$ , determine o coeficiente de regressão de  $Y$  em relação a  $X$ .

3.14. A partir de uma amostra aleatória com  $n$  observações, foi obtida a equação de regressão

$$\hat{Y} = 10 - 0,28X$$

Determine o coeficiente de correlação entre  $X$  e  $Y$  sabendo que

$$s^2(X) = \frac{\sum x^2}{n-1} = 25 \quad \text{e} \quad s^2(Y) = \frac{\sum y^2}{n-1} = 4$$

3.15. Para duas variáveis negativamente correlacionadas, foram obtidos:  $\bar{X} = 0$ ,  $\bar{Y} = 12$ ,  $s(X) = 8$ ,  $s(Y) = 10$  e  $r^2 = 0,64$ . Determine a equação de regressão de  $Y$  em relação a  $X$ .

3.16. Dados

$X$	$Y$
0	2
2	2
4	4
6	8

- a) Determine as estimativas dos parâmetros do modelo  $Y_i = \alpha + \beta X_i + u_i$ .
- b) Calcule o coeficiente de determinação da regressão ( $r_{yx}^2$ ).
- c) Calcule os 4 valores de  $\hat{Y}_i$  e determine o valor do quadrado do coeficiente de correlação ( $r_{yy}^2$ ) entre  $Y_i$  e  $\hat{Y}_i$ . Verifique que esse valor é igual ao valor do coeficiente de determinação, calculado no item (b).

- d) Demonstre que o coeficiente de determinação de uma regressão ( $r_{yx}^2$ ) é sempre igual ao quadrado do coeficiente de correlação entre os valores observados e os valores estimados da variável dependente ( $r_{y\hat{y}}^2$ ).

3.17. Seja  $r$  o coeficiente de correlação entre as variáveis  $X_i$  e  $Y_i$  em uma amostra com  $n$  observações. Definimos as variáveis reduzidas

$$w_i = \frac{x_i}{s(X)}, \text{ com } x_i = X_i - \bar{X} \text{ e } s(X) = \sqrt{\frac{\sum x_i^2}{n-1}}, \text{ e}$$

$$z_i = \frac{y_i}{s(Y)}, \text{ com } y_i = Y_i - \bar{Y} \text{ e } s(Y) = \sqrt{\frac{\sum y_i^2}{n-1}}$$

Seja  $c$  a estimativa do coeficiente de regressão de  $z_i$  contra  $w_i$ , de acordo com o método de mínimos quadrados.

- a) Deduza a relação entre  $r$  e  $c$ .
- b) Deduza expressões que dêem a S.Q.Total e a S.Q.Regressão da regressão de  $z_i$  contra  $w_i$  em função de  $r$  e  $n$ .

## RESPOSTAS

3.1. a)  $r = 0,8$ ;  $F = 14,22$ , significativo

b)  $\hat{Y} = -6,2 + 1,6X$  e  $\bar{X} = 5 + 0,4Y$

c) Para  $X = 4$ , temos  $\hat{Y} = 0,2$  e para  $X = 9$ , temos  $\hat{Y} = 8,2$

Para  $Y = 3$ , temos  $\hat{X} = 6,2$  e para  $Y = 9$ , temos  $\hat{X} = 8,6$

3.2.  $\theta = 30^\circ 58'$

3.3. a)  $r = -0,92$

b)  $\hat{Y} = 18,04 - 1,34X$

c) não se rejeita  $H_0 : \beta = 1$

3.4. a)  $\hat{Y} = 2 + 0,5X$

b)  $\hat{X} = -1 + Y$

c)  $r = 0,7071$

d)  $\hat{Y} = 2,5$

e)  $\hat{X} = 0$

3.5. a)  $\hat{Y} = 10 + 0,25X$

b)  $\hat{X} = 20 + Y$

c)  $r = 0,5$

d)  $\hat{Y} = 25$

e)  $\hat{X} = 50$

3.6. a) A correlação é negativa.

b) Tais dados não permitem estabelecer relação de “causa-e-efeito”. Outras variáveis, como renda *per capita*, deveriam ser consideradas na análise. Os dados podem ser úteis para sugerir pesquisas médico-biológicas sobre possíveis relações causais entre consumo de proteína e fertilidade.

3.7. Correlação linear igual a zero não implica ausência de relação entre as variáveis.

3.8. a)  $F = 10,12$  ou  $t = -3,18$ , significativos.

Renda *per capita* e proporção de analfabetos se mostram negativamente correlacionados. Esse resultado estatístico não prova a existência de uma relação de causa-e-efeito. No caso, sabemos que existe causação nos dois sentidos. Analfabetismo implica baixo nível tecnológico, baixa produtividade e baixa renda *per capita*. Pobreza, por outro lado, significa falta de recursos, dificultando a analfabetização.

3.11.  $\hat{Y} = 12 + 0,8X$

3.12.  $F = 4,76$ , significativo

3.13.  $b = 4/3$

3.14.  $r = -0,7$

3.15.  $\hat{Y} = 12 - X$

3.16. a)  $\hat{Y} = 1 + X$

b)  $r^2 = 5/6$

3.17. a)  $r = c$

$$b) \text{ S.Q.Total} = \sum z_i^2 = n - 1$$

$$\text{S.Q.Regr.} = r^2(n - 1)$$

## 4. REGRESSÃO LINEAR MÚLTIPLA

### 4.1. O modelo estatístico de uma regressão linear múltipla

Temos uma regressão linear múltipla quando admitimos que o valor da variável dependente é função linear de duas ou mais variáveis explanatórias. O modelo estatístico de uma regressão linear múltipla com  $k$  variáveis explanatórias é:

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \dots + \beta_k X_{kj} + u_j, \quad j = 1, \dots, n$$

ou

$$Y_j = \alpha + \sum_{i=1}^k \beta_i X_{ij} + u_j \quad (4.1)$$

Utilizando notação matricial o modelo fica

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \quad (4.2)$$

onde

$$\begin{aligned} \mathbf{y} &= \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} & \mathbf{X} &= \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{k1} \\ 1 & X_{12} & X_{22} & \dots & X_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{1n} & X_{2n} & \dots & X_{kn} \end{bmatrix} \\ \boldsymbol{\beta} &= \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} & \mathbf{u} &= \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \end{aligned}$$

Mantemos, com algumas modificações, as pressuposições apresentadas na seção 2.1:

- I) a variável dependente ( $Y_j$ ) é função linear das variáveis explanatórias ( $X_{ij}$ ,  $i = 1, \dots, k$ );
- II) os valores das variáveis explanatórias são fixos;
- III)  $E(u_j) = 0$ , ou seja,  $E(\mathbf{u}) = \mathbf{0}$ , onde  $\mathbf{0}$  representa um vetor de zeros;
- IV) os erros são homocedásticos, isto é,  $E(u_j^2) = \sigma^2$ ;

- V) os erros são não-correlacionados entre si, isto é,  $E(u_j u_h) = 0$  para  $j \neq h$ ;  
VI) os erros têm distribuição normal.

Combinando as pressuposições IV e V temos

$$E(\mathbf{u}\mathbf{u}') = \mathbf{I}\sigma^2 \quad (4.3)$$

Seja  $p = k + 1$  o número de parâmetros a serem estimados ( $\alpha, \beta_1, \dots, \beta_k$ ). Se dispomos de apenas  $p$  observações, a determinação dos parâmetros se reduz a um problema matemático de resolução de um sistema de  $p$  equações com  $p$  incógnitas, não sendo possível fazer qualquer análise estatística. Devemos, portanto, ter  $n > p$ . Além disso, veremos que para obter as estimativas de mínimos quadrados dos parâmetros a matriz  $\mathbf{X}'\mathbf{X}$  deve ser não-singular, isto é, sua característica deve ser igual a  $p$ . Isso significa que a característica de  $\mathbf{X}$  deve ser igual a  $p$ . Nas deduções que se seguem admitiremos que essas condições são observadas, isto é, admitiremos que  $\mathbf{X}$  tem característica  $p = k + 1 < n$ .

Da mesma maneira que na regressão linear simples, as pressuposições I, II e III são necessárias para demonstrar que os estimadores de mínimos quadrados são não-tendenciosos e as cinco primeiras pressuposições permitem demonstrar que tais estimadores são estimadores lineares não-tendenciosos de variância mínima (teorema de Gauss-Markov). A pressuposição VI é necessária para realizar testes de hipótese e para construir intervalos de confiança para os parâmetros.

## 4.2. Estimativas dos parâmetros de acordo com o método dos mínimos quadrados

Sejam  $\mathbf{b}$  e  $\mathbf{e}$  os vetores das estimativas dos parâmetros e dos desvios, respectivamente, isto é,

$$\mathbf{b} = \begin{bmatrix} a \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} \quad \text{e} \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Temos

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} = \hat{\mathbf{y}} + \mathbf{e} \quad (4.4)$$

e

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b} = \mathbf{y} - \hat{\mathbf{y}}$$

onde

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix}$$

A soma dos quadrados dos desvios é dada por

$$Z = \mathbf{e}'\mathbf{e} = (\mathbf{y}' - \mathbf{b}'\mathbf{X}')(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b} - \mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}$$

As matrizes  $\mathbf{y}'\mathbf{X}\mathbf{b}$  e  $\mathbf{b}'\mathbf{X}'\mathbf{y}$  são iguais, pois uma é a transposta da outra e cada uma tem apenas um elemento. Então

$$Z = \mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} \quad (4.5)$$

A função  $Z$  apresenta ponto de mínimo para os valores de  $\mathbf{b}$  que tornem sua diferencial identicamente nula, isto é:

$$dZ = -2(d\mathbf{b}')\mathbf{X}'\mathbf{y} + (d\mathbf{b}')\mathbf{X}'\mathbf{X}\mathbf{b} + \mathbf{b}'\mathbf{X}'\mathbf{X}(d\mathbf{b}) \equiv 0$$

Como  $(d\mathbf{b}')\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{b}'\mathbf{X}'\mathbf{X}(d\mathbf{b})$ , por serem matrizes com apenas um elemento e uma ser a transposta da outra, segue-se que

$$-2(d\mathbf{b}')\mathbf{X}'\mathbf{y} + 2(d\mathbf{b}')\mathbf{X}'\mathbf{X}\mathbf{b} \equiv 0$$

ou

$$(d\mathbf{b}')(\mathbf{X}'\mathbf{X}\mathbf{b} - \mathbf{X}'\mathbf{y}) \equiv 0$$

Portanto, a diferencial de  $Z$  será identicamente nula para

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y} \quad (4.6)$$

que é o sistema de equações normais.

Se  $\mathbf{X}'\mathbf{X}$  é não singular, existe a matriz inversa  $(\mathbf{X}'\mathbf{X})^{-1}$ . Pré-multiplicando os dois membros de (4.6) por  $(\mathbf{X}'\mathbf{X})^{-1}$ , obtemos

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (4.7)$$

A primeira etapa dos cálculos para obtenção das estimativas dos parâmetros é a construção das matrizes

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum X_{1j} & \sum X_{2j} & \dots & \sum X_{kj} \\ \sum X_{1j} & \sum X_{1j}^2 & \sum X_{1j}X_{2j} & \dots & \sum X_{1j}X_{kj} \\ \sum X_{2j} & \sum X_{1j}X_{2j} & \sum X_{2j}^2 & \dots & \sum X_{2j}X_{kj} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum X_{kj} & \sum X_{1j}X_{kj} & \sum X_{2j}X_{kj} & & \sum X_{kj}^2 \end{bmatrix}$$

e

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} \sum Y_j \\ \sum X_{1j}Y_j \\ \sum X_{2j}Y_j \\ \vdots \\ \sum X_{kj}Y_j \end{bmatrix}$$

Veremos, adiante, que essas matrizes são necessárias em várias outras fases da análise de regressão linear múltipla.

Do sistema de equações normais podemos obter outros resultados de interesse.

De (4.6) segue-se que

$$\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{0}$$

onde  $\mathbf{0}$  representa um vetor cujos elementos são todos iguais a zero.

Então

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{0}$$

ou

$$\mathbf{X}'\mathbf{e} = \mathbf{0} \quad (4.8)$$

Essa relação matricial significa que

$$\sum e_j = 0$$

e

$$\sum X_{ij}e_j = 0 \text{ para } i = 1, \dots, k$$

Note-se que a nulidade da soma dos desvios ( $\sum e_j = 0$ ) decorre do fato de o modelo ter um termo constante ( $\alpha$ ), fazendo com que a primeira coluna de  $\mathbf{X}$  seja um vetor com todos os elementos iguais a 1.

Sendo nula a soma dos desvios, concluímos que

$$\sum Y_j = \sum \hat{Y}_j \quad (4.9)$$

Mostraremos, a seguir, que  $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$  é um estimador não-tendencioso de  $\boldsymbol{\beta}$ . Substituindo (4.2) em (4.7) obtemos

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u})$$

ou

$$\mathbf{b} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \quad (4.10)$$

Lembrando as pressuposições II e III, verificamos que

$$E(\mathbf{b}) = \boldsymbol{\beta}, \text{ c.q.d.}$$

### 4.3. Variâncias e covariâncias das estimativas dos parâmetros

A matriz

$$E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})']$$

é por definição, a matriz das variâncias e covariâncias das estimativas dos parâmetros, pois

$$E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})'] =$$

$$= \begin{bmatrix} E(a - \alpha)^2 & E(a - \alpha)(b_1 - \beta_1) & \dots & E(a - \alpha)(b_k - \beta_k) \\ E(a - \alpha)(b_1 - \beta_1) & E(b_1 - \beta_1)^2 & \dots & E(b_1 - \beta_1)(b_k - \beta_k) \\ \vdots & \vdots & & \vdots \\ E(a - \alpha)(b_k - \beta_k) & E(b_1 - \beta_1)(b_k - \beta_k) & \dots & E(b_k - \beta_k)^2 \end{bmatrix}$$

Considerando (4.10) e notando que a matriz  $\mathbf{X}'\mathbf{X}$  é simétrica e, portanto, igual à sua transposta, obtemos

$$E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})'] = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]$$

De acordo com (4.3) e com a pressuposição II, segue-se que

$$E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})'] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{I}\sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

ou

$$E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})'] = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2 \quad (4.11)$$

#### 4.4. Variância de uma combinação linear das estimativas dos parâmetros

Seja  $\mathbf{c}'$  um vetor-linha com  $p = k + 1$  constantes:

$$\mathbf{c}' = [c_0 \quad c_1 \quad c_2 \quad \dots \quad c_k]$$

Determinemos a variância de  $\mathbf{c}'\mathbf{b}$ .

Sabemos que

$$E(\mathbf{b}) = \boldsymbol{\beta}$$

Então

$$E(\mathbf{c}'\mathbf{b}) = \mathbf{c}'\boldsymbol{\beta}$$

Desde que  $\mathbf{c}'\mathbf{b}$  é uma matriz com um único elemento, temos

$$\begin{aligned} V(\mathbf{c}'\mathbf{b}) &= E(\mathbf{c}'\mathbf{b} - \mathbf{c}'\boldsymbol{\beta})^2 = \\ &= E[\mathbf{c}'(\mathbf{b} - \boldsymbol{\beta})]^2 = \\ &= E[\mathbf{c}'(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})'\mathbf{c}] \end{aligned}$$

Considerando (4.11) obtemos

$$V(\mathbf{c}'\mathbf{b}) = \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}\sigma^2 \quad (4.12)$$

Uma aplicação importante desse resultado é a determinação da variância da estimativa ( $\hat{Y}_h$ ) de um valor da variável dependente.

Considerando o modelo de regressão linear múltipla

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \dots + \beta_k X_{kj} + u_j, \quad j = 1, \dots, n$$

ou

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

a estimativa do valor de  $Y$ , dados os valores  $X_{1h}, X_{2h}, \dots, X_{kh}$  das variáveis explanatórias, é

$$\hat{Y}_h = a + b_1 X_{1h} + b_2 X_{2h} + \dots + b_k X_{kh}$$

ou

$$\hat{Y}_h = \mathbf{x}'_h \mathbf{b}, \quad (4.13)$$

onde

$$\mathbf{x}'_h = [1 \quad X_{1h} \quad X_{2h} \quad \dots \quad X_{kh}]$$

O vetor  $\mathbf{x}'_h$  pode ou não ser uma das linhas da matriz  $\mathbf{X}$ .

Em (4.13) notamos que  $\hat{Y}_h$  é uma combinação linear das estimativas dos parâmetros. Então, de acordo com (4.12), obtemos

$$V(\hat{Y}_h) = \mathbf{x}'_h (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_h \sigma^2 \quad (4.14)$$

#### 4.5. Análise de variância da regressão linear múltipla

De (4.5) e (4.6) segue-se que a soma de quadrados dos desvios, ou soma de quadrados residual, é dada por

$$\mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{y}$$

ou

$$\text{S.Q.Res.} = \mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y} \quad (4.15)$$

Sabemos que a soma de quadrados total é dada por

$$\text{S.Q.Total} = \sum y_j^2 = \sum Y_j^2 - \frac{(\sum Y_j)^2}{n} = \mathbf{y}'\mathbf{y} - \frac{(\sum Y_j)^2}{n} \quad (4.16)$$

A soma de quadrados de regressão é dada por

$$\begin{aligned} \text{S.Q.Regr.} &= \sum(\hat{Y}_j - \bar{Y})^2 = \sum \hat{y}_j^2 = \\ &= \sum \hat{Y}_j^2 - \frac{(\sum \hat{Y}_j)^2}{n} = \\ &= \hat{\mathbf{y}}'\hat{\mathbf{y}} - \frac{(\sum \hat{Y}_j)^2}{n} = \\ &= (\mathbf{X}\mathbf{b})'\mathbf{X}\mathbf{b} - \frac{(\sum \hat{Y}_j)^2}{n} = \\ &= \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} - \frac{(\sum \hat{Y}_j)^2}{n} \end{aligned}$$

Considerando (4.6) e (4.9) segue-se que

$$S.Q.Regr. = \mathbf{b}'\mathbf{X}'\mathbf{y} - \frac{(\sum Y_j)^2}{n} \quad (4.17)$$

De (4.15), (4.16) e (4.17), concluímos que

$$S.Q.Res. = (S.Q.Total) - (S.Q.Regr.)$$

Sendo  $p = k + 1$  o número de parâmetros da regressão, podemos demonstrar que  $E(S.Q.Res.) = (n - p)\sigma^2$ .

Para isso definimos, inicialmente, as matrizes

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

e

$$\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{I} - \mathbf{H}$$

As matrizes  $\mathbf{H}$  e  $\mathbf{M}$  são simétricas e idempotentes, isto é,

$$\mathbf{H}' = \mathbf{H},$$

$$\mathbf{M}' = \mathbf{M},$$

$$\mathbf{HH} = \mathbf{H}$$

$$e \qquad \qquad \qquad \mathbf{MM} = \mathbf{M} \quad (4.18)$$

Verifica-se, também que

$$\mathbf{HX} = \mathbf{X} \quad \text{ou} \quad \mathbf{X}'\mathbf{H} = \mathbf{X}'$$

e

$$\mathbf{MX} = \mathbf{0} \quad \text{ou} \quad \mathbf{X}'\mathbf{M} = \mathbf{0}, \quad (4.19)$$

onde  $\mathbf{0}$  é uma matriz de zeros.

Temos que

$$\begin{aligned} \mathbf{e} &= \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\mathbf{b} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \\ &= [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y} = \\ &= \mathbf{My} = \mathbf{M}(\mathbf{X}\beta + \mathbf{u}) \end{aligned}$$

Considerando (4.19), segue-se que

$$\mathbf{e} = \mathbf{Mu} \quad (4.20)$$

De (4.18) e (4.20) segue-se que

$$S.Q.Res. = \mathbf{e}'\mathbf{e} = \mathbf{u}'\mathbf{Mu} \quad (4.21)$$

Como  $\mathbf{e}'\mathbf{e}$  é uma matriz com apenas um elemento temos que

$$\mathbf{e}'\mathbf{e} = \text{tr}(\mathbf{u}'\mathbf{M}\mathbf{u})$$

Lembrando o teorema de álgebra de matrizes que estabelece que o traço de um produto de matrizes não é afetado por uma mudança na ordem dos fatores, desde que o novo produto também seja definido, obtemos

$$\mathbf{e}'\mathbf{e} = \text{tr}(\mathbf{u}\mathbf{u}'\mathbf{M})$$

Considerando (4.3) e a pressuposição II, segue-se que

$$E(\text{S.Q.Res.}) = E(\mathbf{e}'\mathbf{e}) = \sigma^2 \text{tr}(\mathbf{M})$$

Mas

$$\text{tr}(\mathbf{M}) = \text{tr}[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = n - (k + 1) = n - p$$

Então

$$E(\text{S.Q.Res.}) = (n - p)\sigma^2, \quad (4.22)$$

c.q.d.

Nos exercícios 4.25 e 4.26 é indicada a maneira como podemos obter as expressões para  $E(\text{S.Q.Total.})$  e  $E(\text{S.Q.Regr.})$ .

Esses resultados nos levam à construção do seguinte esquema de análise de variância:

#### Análise de Variância

C.V.	G.L.	S.Q.
Régressão	$k = p - 1$	$\mathbf{b}'\mathbf{X}'\mathbf{y} - \frac{(\sum Y_j)^2}{n}$
Resíduo	$n - p$	$\mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y}$
Total	$n - 1$	$\mathbf{y}'\mathbf{y} - \frac{(\sum Y_j)^2}{n}$

O quadrado médio residual, dado pelo cociente  $\text{S.Q.Res.}/(n - p)$ , é, portanto, uma estimativa não-tendenciosa da variância do erro ( $\sigma^2$ ). Substituindo  $\sigma^2$  por

$s^2 = \text{Q.M.Res.}$  na expressão (4.11) obtemos a matriz das estimativas das variâncias e covariâncias das estimativas dos parâmetros:

$$\hat{V}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}s^2 \quad (4.23)$$

É possível demonstrar que, se os erros  $u_j$  têm distribuição normal e se  $\beta_1 = \beta_2 = \dots = \beta_k = 0$ , o cociente

$$F = \frac{\text{Q.M.Regr.}}{\text{Q.M.Res.}}$$

tem distribuição de  $F$  com  $k$  e  $n - p$  graus de liberdade. Então, o valor  $F$  assim obtido é utilizado para testar a hipótese

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

Obtidas as estimativas dos desvios padrões das estimativas dos parâmetros, dadas pelas raízes quadradas dos elementos da diagonal principal da matriz  $(\mathbf{X}'\mathbf{X})^{-1}s^2$ , podemos utilizar o valor

$$t = \frac{b_i - \beta_i}{s(b_i)}, \quad (4.24)$$

associado a  $n - p$  graus de liberdade, para testar hipóteses a respeito dos valores dos parâmetros.

Podemos, ainda, construir intervalos de confiança para os parâmetros. Escolhido o nível de confiança, e sendo  $t_0$  o correspondente valor crítico de  $t$ , o intervalo de confiança para  $\beta_i$  é

$$b_i - t_0 s(b_i) < \beta_i < b_i + t_0 s(b_i) \quad (4.25)$$

Devemos ressaltar que tanto o teste  $t$  como o intervalo de confiança só são válidos se os erros  $u_j$  tiverem distribuição normal.

O coeficiente de determinação múltipla é definido por

$$R^2 = \frac{\text{S.Q.Regr.}}{\text{S.Q.Total}}$$

e mostra a proporção da soma de quadrados total que é “explicada” pela regressão múltipla.

Temos que

$$1 - R^2 = \frac{\text{S.Q.Res.}}{\text{S.Q.Total}}$$

O coeficiente de determinação corrigido para graus de liberdade é definido por

$$1 - \bar{R}^2 = \frac{\frac{1}{n-p}(\text{S.Q.Res.})}{\frac{1}{n-1}(\text{S.Q.Total})} = \frac{n-1}{n-p}(1 - R^2)$$

ou

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

#### 4.6. Demonstração de que $\mathbf{b}$ é um estimador linear não-tendencioso de variância mínima

Para demonstrar que os estimadores de mínimos quadrados são estimadores lineares não-tendenciosos de variância mínima, vamos considerar, inicialmente, a combinação linear  $\mathbf{c}'\boldsymbol{\beta}$  dos parâmetros. Um estimador de  $\theta = \mathbf{c}'\boldsymbol{\beta}$  é  $\mathbf{c}'\mathbf{b} = \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ . Na seção 4.4 vimos que  $\mathbf{c}'\mathbf{b}$  é um estimador não-tendencioso de  $\mathbf{c}'\boldsymbol{\beta}$  e que, de acordo com (4.12), sua variância é

$$V(\mathbf{c}'\mathbf{b}) = \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}\sigma^2$$

Consideremos um estimador linear qualquer  $\hat{\theta} = \mathbf{g}'\mathbf{y}$  de  $\theta = \mathbf{c}'\boldsymbol{\beta}$ . Note que  $\mathbf{c}'\mathbf{b}$  é um caso particular de  $\hat{\theta}$ , com  $\mathbf{g}' = \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

Temos

$$\hat{\theta} = \mathbf{g}'\mathbf{y} = \mathbf{g}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = \mathbf{g}'\mathbf{X}\boldsymbol{\beta} + \mathbf{g}'\mathbf{u} \quad (4.26)$$

Então, para que  $\hat{\theta}$  seja um estimador não-tendencioso de  $\mathbf{c}'\boldsymbol{\beta}$ , isto é, para que tenhamos  $E(\hat{\theta}) = \mathbf{c}'\boldsymbol{\beta}$ , devemos ter

$$\mathbf{g}'\mathbf{X} = \mathbf{c}' \quad (4.27)$$

De acordo com (4.12) e (4.27), obtemos

$$V(\mathbf{c}'\mathbf{b}) = \mathbf{g}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{g}\sigma^2 \quad (4.28)$$

De (4.26), obtemos

$$\hat{\theta} - E(\hat{\theta}) = \mathbf{g}'\mathbf{u}$$

Donde

$$\begin{aligned} V(\hat{\theta}) &= E[\hat{\theta} - E(\hat{\theta})]^2 = \\ &= E(\mathbf{g}'\mathbf{u}\mathbf{u}'\mathbf{g}) \end{aligned}$$

Como  $E(\mathbf{u}\mathbf{u}') = \mathbf{I}\sigma^2$ , podemos escrever

$$V(\hat{\theta}) = \mathbf{g}'\mathbf{g}\sigma^2 \quad (4.29)$$

De (4.28) e (4.29) segue-se que

$$\begin{aligned} V(\hat{\theta}) - V(\mathbf{c}'\mathbf{b}) &= [\mathbf{g}'\mathbf{g} - \mathbf{g}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{g}]\sigma^2 = \\ &= \mathbf{g}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{g}\sigma^2 = \\ &= \mathbf{g}'\mathbf{M}\mathbf{g}\sigma^2 \end{aligned}$$

Vimos, em (4.21), que  $\mathbf{e}'\mathbf{e} = \mathbf{u}'\mathbf{M}\mathbf{u}$ . Ora,  $\mathbf{e}'\mathbf{e} \geq 0$  porque é uma soma de quadrados. Portanto  $\mathbf{M}$  é uma matriz semidefinida positiva e  $\mathbf{g}'\mathbf{M}\mathbf{g} \geq 0$ . Concluímos então que

$$V(\hat{\theta}) \geq V(\mathbf{c}'\mathbf{b}), \quad (4.30)$$

onde  $\hat{\theta} = \mathbf{g}'\mathbf{y}$  é qualquer estimador linear não-tendencioso de  $\mathbf{c}'\beta$ .

Consideremos o caso particular em que

$$\mathbf{c}' = [0 \ \dots \ 0 \ 1 \ 0 \ \dots \ 0],$$

isto é, o  $i$ -ésimo elemento do vetor  $\mathbf{c}'$  é igual a um, e todos os outros são iguais a zero. Então a desigualdade (4.30) fica

$$V(\hat{\theta}) \geq V(b_i)$$

onde  $\hat{\theta} = \mathbf{g}'\mathbf{y}$  é qualquer estimador linear não-tendencioso de  $\beta_i$ .

Esse resultado mostra que, dentre os estimadores lineares não-tendenciosos,  $b_i$  é o que tem menor variância, isto é, os estimadores de mínimos quadrados são estimadores lineares não-tendenciosos de variância mínima.

#### 4.7. O uso das variáveis centradas

Para simplificar os cálculos, muitas vezes trabalhamos com as variáveis centradas

$$x_{ij} = X_{ij} - \bar{X}_i, \quad i = 1, 2, \dots, k$$

onde

$$\bar{X}_i = \frac{1}{n} \sum_{j=1}^n X_{ij}$$

Neste caso o modelo estatístico fica

$$Y_j = \beta_0 + \beta_1 x_{1j} + \dots + \beta_k x_{kj} + u_j, \quad j = 1, 2, \dots, n$$

ou em notação matricial,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

com as devidas mudanças nas definições das matrizes  $\mathbf{X}$  e  $\boldsymbol{\beta}$ .

As matrizes  $\mathbf{X}'\mathbf{X}$  e  $\mathbf{X}'\mathbf{y}$  ficam

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & 0 & 0 & \dots & 0 \\ 0 & \sum x_{1j}^2 & \sum x_{1j}x_{2j} & \dots & \sum x_{1j}x_{kj} \\ 0 & \sum x_{1j}x_{2j} & \sum x_{2j}^2 & \dots & \sum x_{2j}x_{kj} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & \sum x_{1j}x_{kj} & \sum x_{2j}x_{kj} & \dots & \sum x_{kj}^2 \end{bmatrix}$$

e

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} \sum Y_j \\ \sum x_{1j}Y_j \\ \sum x_{2j}Y_j \\ \vdots \\ \sum x_{kj}Y_j \end{bmatrix}$$

Decompondo a matriz  $\mathbf{X}'\mathbf{X}$  apropriadamente, o elemento igual a  $n$  pode ser invertido separadamente. Então a estimativa de  $\beta_0$  é

$$b_0 = \frac{\sum Y_j}{n} = \bar{Y}$$

Verifica-se que a expressão (4.15) pode ser escrita como segue:

$$\text{S.Q.Res.} = \sum_{j=1}^n Y_j^2 - \bar{Y} \sum_{j=1}^n Y_j - \sum_{i=1}^k b_i \sum_{j=1}^n x_{ij} Y_j$$

Como

$$\sum Y_j^2 - \bar{Y} \sum Y_j = \sum Y_j^2 - \frac{(\sum Y_j)^2}{n} = \text{S.Q.Total},$$

concluímos que

$$\text{S.Q.Regr.} = \sum_{i=1}^k b_i \sum_{j=1}^n x_{ij} Y_j \quad (4.31)$$

Determinadas as estimativas dos parâmetros do modelo simplificado, se quisermos escrever a equação estimada com as variáveis na forma original, basta calcular a estimativa de  $\alpha$ , dada por

$$\alpha = \bar{Y} - \sum b_i \bar{X}_i \quad (4.32)$$

Às vezes, os cálculos são feitos com todas as variáveis centradas, inclusive a variável dependente, ou seja, utilizamos

$$y_j = Y_j - \bar{Y}$$

Se somarmos, membro a membro, as relações (4.1), para  $j = 1, 2, \dots, n$ , e dividirmos por  $n$ , obtemos

$$\bar{Y} = \alpha + \sum_{i=1}^k \beta_i \bar{X}_i + \bar{u} \quad (4.33)$$

onde

$$\bar{u} = \frac{1}{n} \sum u_j$$

Subtraindo (4.33) das relações (4.1) obtemos

$$y_j = \sum_{i=1}^k \beta_i x_{ij} + u_j - \bar{u}$$

ou

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} - \bar{\mathbf{u}} \quad (4.34)$$

onde

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} x_{11} & x_{21} & \dots & x_{k1} \\ x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & & \vdots \\ x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

e  $\bar{\mathbf{u}}$  é um vetor-coluna com  $n$  elementos iguais a

$$\bar{u} = \frac{1}{n} \sum u_j$$

É fácil verificar que, neste caso, a matriz  $\mathbf{X}'\mathbf{X}$  é igual à do modelo onde apenas as variáveis independentes são centradas excluindo a primeira linha e a primeira coluna, e a matriz  $\mathbf{X}'\mathbf{y}$  é igual à do mesmo modelo, excluindo apenas o primeiro elemento. Os termos  $\mathbf{y}'\mathbf{y}$  e  $\mathbf{b}'\mathbf{X}'\mathbf{y}$  de (4.15) correspondem, respectivamente, à soma de quadrados total e à soma de quadrados de regressão, de maneira que o coeficiente de determinação é

$$R^2 = \frac{\mathbf{b}'\mathbf{X}'\mathbf{y}}{\mathbf{y}'\mathbf{y}}$$

As propriedades dos estimadores não são afetadas pelo uso de variáveis centradas. Assim, substituindo (4.34) na expressão

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

obtemos

$$\begin{aligned} \mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u} - \bar{\mathbf{u}}) = \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\bar{\mathbf{u}} \end{aligned} \quad (4.35)$$

À primeira vista, o resultado obtido em (4.35) é diferente da expressão (4.10), obtida quando as variáveis não são centradas. Entretanto, os elementos do vetor  $\mathbf{X}'\bar{\mathbf{u}}$ , com variáveis centradas, são

$$\sum_j x_{ij}\bar{u} = \bar{u} \sum_j x_{ij} = 0$$

Então, a expressão (4.35) fica

$$\mathbf{b} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u},$$

que é a relação (4.10).

Assim, da mesma maneira que no modelo sem centrar as variáveis, temos:

$$E(\mathbf{b}) = \boldsymbol{\beta}$$

e

$$V(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1} \sigma^2$$

#### 4.8. Exemplo de uma regressão linear múltipla com duas variáveis explanatórias

Na tabela 4.1 apresentamos os valores de uma amostra de 5 observações das variáveis  $Y_j$ ,  $X_{1j}$  e  $X_{2j}$ .

TABELA 4.1. Valores de três variáveis em uma amostra de 5 observações.

$Y_j$	$X_{1j}$	$X_{2j}$
16,5	1,0	2
14,0	3,5	3
6,0	4,0	4
10,0	7,5	5
3,5	9,0	6

Obtemos:

$$\begin{array}{lll} \sum Y_j = 50 & \sum X_{1j} = 25 & \sum X_{2j} = 20 \\ \bar{Y} = 10 & \bar{X}_1 = 5 & \bar{X}_2 = 4 \\ \sum y_j^2 = 116,5 & \sum x_{1j}^2 = 41,5 & \sum x_{2j}^2 = 10 \\ \sum x_{1j}Y_j = -54 & \sum x_{2j}Y_j = -30 & \sum x_{1j}x_{2j} = 20 \end{array}$$

Tendo em vista o modelo

$$Y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + u_j,$$

construímos as matrizes

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & 0 & 0 \\ 0 & \sum x_{1j}^2 & \sum x_{1j}x_{2j} \\ 0 & \sum x_{1j}x_{2j} & \sum x_{2j}^2 \end{bmatrix} = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 41,5 & 20 \\ 0 & 20 & 10 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} \sum Y_j \\ \sum x_{1j}Y_j \\ \sum x_{2j}Y_j \end{bmatrix} = \begin{bmatrix} 50 \\ -54 \\ -30 \end{bmatrix}$$

A seguir, determinamos as estimativas dos parâmetros

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} \frac{1}{5} & 0 & 0 \\ 0 & \frac{2}{3} & -\frac{4}{3} \\ 0 & -\frac{4}{3} & \frac{83}{30} \end{bmatrix} \begin{bmatrix} 50 \\ -54 \\ -30 \end{bmatrix} = \begin{bmatrix} 10 \\ 4 \\ -11 \end{bmatrix}$$

A equação estimada é, então,

$$\hat{Y}_j = 10 + 4x_{1j} - 11x_{2j}$$

Como

$$x_{1j} = X_{1j} - 5$$

e

$$x_{2j} = X_{2j} - 4$$

obtemos

$$\hat{Y}_j = 34 + 4X_{1j} - 11X_{2j}$$

De acordo com (4.31), temos

$$\begin{aligned} \text{S.Q.Regr.} &= b_1 \sum x_{1j}Y_j + b_2 \sum x_{2j}Y_j = \\ &= 4 \cdot (-54) - 11 \cdot (-30) = 114 \end{aligned}$$

Então

$$\text{S.Q.Res.} = \sum y_j^2 - 114 = 116,5 - 114 = 2,5$$

Poderíamos, também, ter obtido o valor da soma de quadrados residual de (4.15):

$$\begin{aligned} \text{S.Q.Res.} &= \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y} = \\ &= 616,5 - 614 = 2,5 \end{aligned}$$

Com esses resultados podemos construir a tabela de análise da variância.

TABELA 4.2. Análise da Variância

C.V.	G.L.	S.Q.	Q.M.	F
Regressão	2	114,0	57	45,6
Resíduo	2	2,5	1,25	
Total	4	116,5		

Para 2 e 2 graus de liberdade e ao nível de significância de 5%, o valor crítico de  $F$  é 19,00. Portanto, o resultado é significativo, isto é, rejeita-se, a esse nível de significância, a hipótese  $H_0 : \beta_1 = \beta_2 = 0$ . Um bom programa para computador informa que a probabilidade caudal associada a  $F = 45,6$ , com 2 e 2 graus de liberdade, é 0,0215, permitindo concluir que o resultado é significativo ao nível de 5%, sem necessidade de obter o valor crítico de  $F$ .

O coeficiente de determinação múltipla é

$$R^2 = \frac{114}{116,5} = 0,9785$$

isto é, 97,85% da soma de quadrados total é “explicada” pela regressão linear ajustada.

Conforme a definição apresentada no final da seção 4.5, podemos verificar que o coeficiente de determinação corrigido para graus de liberdade é  $\bar{R}^2 = 0,9571$ .

Como  $s^2 = 1,25$  temos, de acordo com (4.23), as seguintes estimativas das variâncias e covariâncias das estimativas dos parâmetros:

$$\hat{V}(b_0) = s^2(b_0) = s^2(\bar{Y}) = \frac{1,25}{5} = 0,25$$

$$\hat{V}(b_1) = s^2(b_1) = \frac{2}{3} \cdot 1,25 = \frac{5}{6} = 0,8333$$

$$\hat{V}(b_2) = s^2(b_2) = \frac{83}{30} \cdot 1,25 = \frac{83}{24} = 3,4583$$

$$\text{cov}(\bar{Y}, b_1) = \text{cov}(\bar{Y}, b_2) = 0$$

$$\text{cov}(b_1, b_2) = -\frac{4}{3} \cdot 1,25 = -\frac{5}{3}$$

Temos que

$$a = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2$$

Então

$$V(a) = V(\bar{Y}) + \bar{X}_1^2 V(b_1) + \bar{X}_2^2 V(b_2) - 2\bar{X}_1 \text{cov}(\bar{Y}, b_1) - \\ - 2\bar{X}_2 \text{cov}(\bar{Y}, b_2) + 2\bar{X}_1 \bar{X}_2 \text{cov}(b_1, b_2)$$

e

$$\hat{V}(a) = 0,25 + 5^2 \cdot \frac{5}{6} + 4^2 \cdot \frac{83}{24} + 2 \cdot 5 \cdot 4 \cdot \left( -\frac{5}{3} \right) = 9,75$$

Se tivéssemos utilizado o modelo com as variáveis não centradas, a estimativa da variância de  $a$  seria dada pelo primeiro elemento da diagonal principal de  $(\mathbf{X}'\mathbf{X})^{-1} s^2$ .

Podemos, agora, testar hipóteses a respeito dos valores dos parâmetros. Adotando o nível de significância de 5%, consideremos as seguintes hipóteses:

a)  $H_0 : \alpha = 50$  contra  $H_A : \alpha < 50$

Temos

$$t = \frac{34 - 50}{\sqrt{9,75}} = -5,124$$

O resultado é significativo, pois a região de rejeição para esse teste unilateral é  $t \leq -2,920$ . Portanto, ao nível de significância de 5%, rejeitamos a hipótese  $H_0 : \alpha = 50$ , em favor da hipótese  $H_A : \alpha < 50$ .

b)  $H_0 : \beta_1 = 0$  contra  $H_A : \beta_1 \neq 0$

Calculamos

$$t = \frac{4 - 0}{\sqrt{0,8333}} = 4,382$$

Como o valor crítico de  $t$  para 2 graus de liberdade e ao nível de significância de 5% é 4,303, o resultado obtido é significativo, isto é, rejeitamos, a esse nível, a hipótese de que  $\beta_1 = 0$ . Um bom programa de computador fornece a probabilidade caudal associada ao  $t$  calculado ( $t = 4,382$ ), isto é, a probabilidade de, na distribuição de  $t$  com 2 graus de liberdade, essa variável assumir valor absoluto maior do que 4,382. Essa probabilidade é 0,0483, permitindo concluir que o resultado é significativo ao nível de 5%, sem necessidade de obter o valor crítico de  $t$ .

c)  $H_0 : \beta_2 = 0$  contra  $H_A : \beta_2 \neq 0$

Obtemos

$$t = \frac{-11 - 0}{\sqrt{3,4583}} = -5,915, \text{ significativo}$$

Neste exemplo rejeitamos, ao nível de significância de 5%, a hipótese  $H_0 : \beta_1 = \beta_2 = 0$  e também rejeitamos, ao mesmo nível de significância, tanto a hipótese  $H_0 : \beta_1 = 0$  como a hipótese  $H_0 : \beta_2 = 0$ . Quando o teste  $F$  da análise de variância de uma regressão linear múltipla é significativo (rejeitando-se a hipótese de que  $\beta_1 = \beta_2 = \dots = \beta_k = 0$ ), é comum que pelo menos um dos valores de

$$t = \frac{b_i}{s(b_i)} , \quad i = 1, 2, \dots, k$$

seja significativo, considerando-se um teste bilateral com o mesmo nível de significância. Mas nem sempre isso acontece, podendo ocorrer que, apesar de o teste  $F$  da análise de variância da regressão ser significativo, nenhum dos testes  $t$  para as hipóteses  $H_0 : \beta_i = 0$ , (para  $i = 1, 2, \dots, k$ ) seja significativo, como mostra o exemplo apresentado na seção 4.12, na qual esse assunto será melhor analisado.

#### 4.9. Previsão e teste de hipóteses a respeito do valor de combinações lineares dos parâmetros

Consideremos o modelo de regressão linear múltipla

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \dots + \beta_k X_{kj} + u_j, \quad j = 1, \dots, n$$

ou

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

Na seção 4.4 vimos que, dados os valores  $X_{1h}, X_{2h}, \dots, X_{kh}$  das variáveis explanatórias, a estimativa de

$$E(Y_h) = \alpha + \beta_1 X_{1h} + \beta_2 X_{2h} + \dots + \beta_k X_{kh} = \mathbf{x}'_h \boldsymbol{\beta}$$

é

$$\hat{Y}_h = a + b_1 X_{1h} + b_2 X_{2h} + \dots + b_k X_{kh} = \mathbf{x}'_h \mathbf{b}$$

onde

$$\mathbf{x}'_h = [1 \quad X_{1h} \quad X_{2h} \quad \dots \quad X_{kh}]$$

Devemos ressaltar que o vetor  $\mathbf{x}'_h$  pode ou não ser uma das linhas da matriz  $\mathbf{X}$ .

De acordo com (4.14), temos

$$\hat{V}(\hat{Y}_h) = \mathbf{x}'_h (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_h s^2 \quad (4.36)$$

Obtida a estimativa da variância de  $\hat{Y}_h$ , dada por (4.36), podemos construir o intervalo de confiança para  $E(Y_h) = \mathbf{x}'_h \mathbf{\beta}$ . Sendo  $t_0$  o valor crítico de  $t$  com  $n - p$  graus de liberdade e ao nível de confiança adotado, o intervalo de confiança é

$$\mathbf{x}'_h \mathbf{b} - t_0 \sqrt{\mathbf{x}'_h (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_h s^2} < E(Y_h) < \mathbf{x}'_h \mathbf{b} + t_0 \sqrt{\mathbf{x}'_h (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_h s^2} \quad (4.37)$$

Consideremos, mais uma vez, o exemplo numérico da tabela 4.1. Tendo em vista o modelo<sup>5</sup>

$$Y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + u_j$$

obtivemos, anteriormente, as matrizes

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \frac{1}{5} & 0 & 0 \\ 0 & \frac{2}{3} & -\frac{4}{3} \\ 0 & -\frac{4}{3} & \frac{83}{30} \end{bmatrix}$$

e

$$\mathbf{b} = \begin{bmatrix} 10 \\ 4 \\ -11 \end{bmatrix}$$

Consideremos  $X_{1h} = 7$  e  $X_{2h} = 1$ . Uma vez que estamos fazendo os cálculos tendo em vista o modelo com as variáveis explanatórias centradas, e lembrando que  $\bar{X}_1 = 5$  e  $\bar{X}_2 = 4$ , obtemos  $x_{1h} = 2$  e  $x_{2h} = -3$  e fazemos

$$\mathbf{x}'_h = [1 \quad 2 \quad -3]$$

Então,

---

<sup>5</sup> Se considerarmos o modelo em que todas as variáveis, incluindo a dependente, são centradas, obteremos, através de (4.14), a variância  $\hat{y}_h = \hat{Y}_h - \bar{Y}$ . Como as covariâncias entre  $\bar{Y}$  e  $b_i$  ( $i = 1, 2, \dots, k$ ) são nulas, a variância de  $\hat{Y}_h$  é dada por

$$V(\hat{Y}_h) = V(\hat{y}_h) + \frac{\sigma^2}{n}$$

$$\hat{Y}_h = \mathbf{x}'_h \mathbf{b} = 51$$

Lembrando que  $s^2 = \text{Q.M.Res.} = 1,25$ , obtemos, de acordo com (4.36),

$$\hat{V}(\hat{Y}_h) = \mathbf{x}'_h (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_h s^2 =$$

$$= [1 \ 2 \ -3] \begin{bmatrix} \frac{1}{5} & 0 & 0 \\ 0 & \frac{2}{3} & -\frac{4}{3} \\ 0 & -\frac{4}{3} & \frac{83}{30} \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ -3 \end{bmatrix} 1,25 = 54,708$$

Para um nível de confiança de 95%, o valor crítico de  $t$  com 2 graus de liberdade é 4,303. Então, o intervalo de confiança para  $E(Y_h) = \alpha + 7\beta_1 + \beta_2$  é

$$51 - 4,303\sqrt{54,708} < E(Y_h) < 51 + 4,303\sqrt{54,708}$$

ou

$$19,17 < E(Y_h) < 82,83$$

Consideremos, agora, que desejamos prever o valor da variável dependente ( $Y_h$ ) para uma nova observação e que as variáveis independentes assumem os valores  $X_{1h}$ ,  $X_{2h}$ , ...,  $X_{kh}$ .

O estimador de  $Y_h = \mathbf{x}'_h \boldsymbol{\beta} + u_h$  é  $\hat{Y}_h = \mathbf{x}'_h \mathbf{b}$ . O erro de previsão é

$$\hat{Y}_h - Y_h = \mathbf{x}'_h (\mathbf{b} - \boldsymbol{\beta}) - u_h \quad (4.38)$$

Dizemos que  $\hat{Y}_h$  é uma previsão não-tendenciosa do valor de  $Y_h$  porque a esperança do erro de previsão é igual a zero.

Para avaliar a precisão de  $\hat{Y}_h$  como previsão do valor da nova observação, determinamos o intervalo de previsão, como mostraremos a seguir. Para isso devemos considerar a variância do erro de previsão, dado por (4.38). Uma vez que, de acordo com a pressuposição V, o erro ( $u_h$ ) da nova observação é independente dos erros ( $u_j$ ,  $j = 1, \dots, n$ ) das observações da amostra utilizada para obter a estimativa ( $\mathbf{b}$ ) de  $\boldsymbol{\beta}$ , de (4.38) obtemos

$$V(\hat{Y}_h - Y_h) = V[\mathbf{x}'_h(\mathbf{b} - \boldsymbol{\beta})] + \sigma^2$$

De acordo com (4.14), segue-se que

$$\begin{aligned} V(\hat{Y}_h - Y_h) &= \sigma^2 + \mathbf{x}'_h(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_h\sigma^2 = \\ &= [1 + \mathbf{x}'_h(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_h]\sigma^2 \end{aligned}$$

Sendo  $t_0$  o valor crítico de  $t$  com  $n - p$  graus de liberdade e ao nível de confiança adotado, o intervalo de previsão para a nova observação é

$$\mathbf{x}'_h\mathbf{b} - t_0\sqrt{[1 + \mathbf{x}'_h(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_h]s^2} < Y_h < \mathbf{x}'_h\mathbf{b} + t_0\sqrt{[1 + \mathbf{x}'_h(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_h]s^2}$$

No exemplo numérico que estamos analisando, a estimativa da variância do erro de previsão para  $X_{1h} = 7$  e  $X_{2h} = 1$  é

$$1,25 + 54,708 = 55,958$$

e o intervalo de previsão, ao nível de confiança de 95%, para uma nova observação com esses valores de  $X_{1h}$  e  $X_{2h}$  é

$$51 - 4,303\sqrt{55,958} < Y_h < 51 + 4,303\sqrt{55,958}$$

ou

$$18,81 < Y_h < 83,19$$

Note a grande amplitude do intervalo de previsão, apesar do elevado coeficiente de determinação ( $R^2 = 0,9785$ ) da equação ajustada.

A previsão do valor da variável dependente para uma nova observação pode ser feita para valores de  $\mathbf{x}_h$  fora da região onde estão os valores das variáveis explanatórias da amostra, isto é, pode ser feita uma extrapolação. Da mesma maneira que no caso da regressão linear simples (ver seção 2.12), a validade da equação estimada, fora do intervalo das observações, deve ser cuidadosamente examinada.

A expressão (4.14), que dá a variância de  $\hat{Y}_h$ , é um caso particular de (4.12). Uma outra aplicação de (4.12) é o teste de hipóteses a respeito de combinações lineares dos parâmetros. Admitamos que se queira testar, no exemplo que estamos desenvolvendo, a hipótese  $H_0 : 2\beta_1 + \beta_2 = 0$  contra a hipótese  $H_A : 2\beta_1 + \beta_2 < 0$ , considerando um nível de significância de 5%. A hipótese da nulidade pode ser escrita

$$H_0 : \mathbf{c}'\boldsymbol{\beta} = 0$$

onde

$$\mathbf{c}' = [0 \ 2 \ 1]$$

Para testar essa hipótese calculamos, de acordo com (4.12),

$$\hat{V}(\mathbf{c}'\mathbf{b}) = \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}s^2 = 0,125$$

A seguir, obtemos

$$t = \frac{\mathbf{c}'\mathbf{b} - 0}{\sqrt{\hat{V}(\mathbf{c}'\mathbf{b})}} = \frac{-3}{\sqrt{0,125}} = -8,485$$

O resultado é significativo, isto é, rejeitamos  $H_0: \beta_2 = -2\beta_1$  ao nível de significância de 5%, pois a região de rejeição para esse teste unilateral é  $t \leq -2,920$ .

#### **4.10. Interpretação dos coeficientes de regressão de uma regressão linear múltipla com duas variáveis explanatórias**

Consideremos o modelo de uma regressão linear com duas variáveis explanatórias, com todas as variáveis centradas:

$$y_j = \beta_1 x_{1j} + \beta_2 x_{2j} + u_j - \bar{u} \quad (4.39)$$

Neste caso, temos

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \sum x_{1j}^2 & \sum x_{1j}x_{2j} \\ \sum x_{1j}x_{2j} & \sum x_{2j}^2 \end{bmatrix}, \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} \sum x_{1j}y_j \\ \sum x_{2j}y_j \end{bmatrix}$$

e

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{\sum x_{1j}^2 \sum x_{2j}^2 - (\sum x_{1j}x_{2j})^2} \begin{bmatrix} \sum x_{2j}^2 & -\sum x_{1j}x_{2j} \\ -\sum x_{1j}x_{2j} & \sum x_{1j}^2 \end{bmatrix}$$

De  $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ , obtemos

$$b_1 = \frac{\sum x_{2j}^2 \sum x_{1j}y_j - \sum x_{1j}x_{2j} \sum x_{2j}y_j}{\sum x_{1j}^2 \sum x_{2j}^2 - (\sum x_{1j}x_{2j})^2} \quad (4.40)$$

e

$$b_2 = \frac{\sum x_{1j}^2 \sum x_{2j}y_j - \sum x_{1j}x_{2j} \sum x_{1j}y_j}{\sum x_{1j}^2 \sum x_{2j}^2 - (\sum x_{1j}x_{2j})^2} \quad (4.41)$$

Vamos indicar os desvios da regressão de  $x_{1j}$  em relação a  $x_{2j}$  por  $v_j$  e os desvios da regressão de  $y_j$  em relação a  $x_{2j}$  por  $z_j$ . Seja  $\hat{\theta}$  a estimativa do coeficiente de regressão de  $z_j$  em relação a  $v_j$ . Demonstraremos que  $b_1 = \hat{\theta}$ , isto é, que a estimativa do coeficiente de regressão de  $x_{1j}$  numa regressão linear com duas variáveis explanatórias mede como  $y_j$  varia em função de  $x_{1j}$ , após eliminar dessas variáveis a influência linear de  $x_{2j}$ . Analogamente,  $b_2$  é uma estimativa de como  $y_j$  varia em função de  $x_{2j}$ , descontando-se, previamente, as variações de  $y_j$  e  $x_{2j}$  que possam ser devidas à influência linear de  $x_{1j}$ .

Em outras palavras,  $b_1$  estima o efeito linear de  $X_1$ , sobre  $Y$  depois que essas variáveis são depuradas da influência linear de  $X_2$ . Analogamente,  $b_2$  estima o efeito linear de  $X_2$  sobre  $Y$  depois que essas variáveis são depuradas da influência linear de  $X_1$ .

Sabemos que para uma regressão linear simples de  $y_j$  em relação a  $x_j$  temos

$$\text{desvio} = y_j - \hat{y}_j = y_j - bx_j = y_j - \left( \frac{\sum x_j y_j}{\sum x_j^2} \right) x_j$$

Segue-se que

$$v_j = x_{1j} - \left( \frac{\sum x_{2j} x_{1j}}{\sum x_{2j}^2} \right) x_{2j} \quad (4.42)$$

e

$$z_j = y_j - \left( \frac{\sum x_{2j} y_j}{\sum x_{2j}^2} \right) x_{2j} \quad (4.43)$$

Como  $x_{1j}$ ,  $x_{2j}$  e  $y_j$  têm média igual a zero,  $v_j$  e  $z_j$  também têm média igual a zero. Então, a estimativa do coeficiente de regressão de  $z_j$  em relação a  $v_j$  é

$$\hat{\theta} = \frac{\sum v_j z_j}{\sum v_j^2} \quad (4.44)$$

Mas

$$\sum v_j z_j = \sum \left[ x_{1j} - \left( \frac{\sum x_{2j} x_{1j}}{\sum x_{2j}^2} \right) x_{2j} \right] \left[ y_j - \left( \frac{\sum x_{2j} y_j}{\sum x_{2j}^2} \right) x_{2j} \right]$$

Desenvolvendo e simplificando, obtemos

$$\sum v_j z_j = \sum x_{1j} y_j - \frac{\sum x_{1j} x_{2j} \sum x_{2j} y_j}{\sum x_{2j}^2} \quad (4.45)$$

Analogamente, obtemos

$$\sum v_j^2 = \sum x_{1j}^2 - \frac{(\sum x_{1j} x_{2j})^2}{\sum x_{2j}^2} \quad (4.46)$$

Substituindo (4.45) e (4.46) em (4.44) e multiplicando numerador e denominador por  $\sum x_{2j}^2$ , obtemos

$$\hat{\theta} = \frac{\sum x_{2j}^2 \sum x_{1j} y_j - \sum x_{1j} x_{2j} \sum x_{2j} y_j}{\sum x_{1j}^2 \sum x_{2j}^2 - (\sum x_{1j} x_{2j})^2} \quad (4.47)$$

Comparando (4.47) com (4.40), verificamos que

$$b_1 = \hat{\theta}, \text{ c.q.d.}$$

Para melhor esclarecer o assunto vamos desenvolver essas etapas no exemplo numérico da tabela 4.1.

Vamos calcular, inicialmente, os desvios ( $v_j$ ) da regressão de  $x_{1j}$  em relação a  $x_{2j}$ , dados por

$$v_j = x_{1j} - 2x_{2j}$$

e os desvios ( $z_j$ ) da regressão de  $y_j$  em relação a  $x_{2j}$ , dados por

$$z_j = y_j + 3x_{2j}$$

Tais valores constam na tabela 4.3.

TABELA 4.3. Valores de  $y_j$ ,  $x_{1j}$ ,  $x_{2j}$ ,  $v_j$  (desvios da regressão de  $x_{1j}$  contra  $x_{2j}$ ) e  $z_j$  (desvios da regressão de  $y_j$  contra  $x_{2j}$ ) obtidos com base nos dados da tabela 4.1.

$y_j$	$x_{1j}$	$x_{2j}$	$v_j$	$z_j$
6,5	-4	-2	0	0,5
4	-1,5	-1	0,5	1
-4	-1	0	-1	-4
0	2,5	1	0,5	3
-6,5	4	2	0	-0,5

A estimativa do coeficiente de regressão de  $z_j$  em relação a  $v_j$  é

$$\hat{\theta} = \frac{\sum v_j z_j}{\sum v_j^2} = \frac{6}{1,5} = 4,$$

que é o valor que já havíamos obtido para  $b_1$ .

A análise que fizemos para uma regressão com duas variáveis explanatórias pode ser generalizada para o caso de uma regressão linear múltipla com  $k$  variáveis explanatórias. Pode-se demonstrar que a estimativa ( $b_1$ ) do coeficiente de uma variável  $X_{ij}$  de uma regressão linear múltipla, normalmente obtida através de (4.7), poderia ser obtida percorrendo-se as seguintes etapas:

- a) cálculo dos resíduos ( $v_j$ ) da regressão de  $X_{ij}$  contra todas as outras variáveis explanatórias;
- b) cálculo dos resíduos ( $z_j$ ) da regressão de  $Y_j$  em relação a essas mesmas variáveis, isto é, as variáveis explanatórias exclusive  $X_{ij}$ ;
- c) determinação da estimativa do coeficiente de regressão de  $z_j$  em relação a  $v_j$ , que é igual a  $b_i$ .

Resumindo, podemos afirmar que, ao ajustarmos uma regressão linear múltipla através do método dos mínimos quadrados, a estimativa do coeficiente de uma variável  $X_i$  mede o efeito linear de  $X_i$  sobre  $Y$  depois de terem sido “descontadas” de ambas essas variáveis as “influências” lineares de todas as outras variáveis explanatórias consideradas no modelo.

#### 4.11. Os coeficientes de correlação parcial

Consideremos o caso de uma regressão linear múltipla com duas variáveis explanatórias, cujo modelo estatístico, utilizando todas as variáveis centradas, é (4.39).

Lembrando que indicamos por  $v_j$  os desvios da regressão linear de  $x_{1j}$  contra  $x_{2j}$  e por  $z_j$  os desvios da regressão de  $y_j$  contra  $x_{2j}$ , o coeficiente de correlação parcial entre  $y_j$  e  $x_{1j}$  ( $r_{Y12}$ ) é, por definição, o coeficiente de correlação entre  $z_j$  e  $v_j$ , isto é,

$$r_{Y1:2} = \frac{\sum v_j z_j}{\sqrt{\sum v_j^2 \sum z_j^2}} \quad (4.48)$$

Deduziremos agora a expressão que relaciona o coeficiente de correlação parcial  $r_{Y1:2}$  com os coeficientes de correlação simples entre  $y_j$  e  $x_{1j}$ , entre  $y_j$  e  $x_{2j}$  e entre  $x_{1j}$  e  $x_{2j}$ , indicados por  $r_{Y1}$ ,  $r_{Y2}$  e  $r_{12}$ , respectivamente.

Considerando (4.42) e (4.43) temos, por analogia com (4.46), que

$$\sum z_j^2 = \sum y_j^2 - \frac{(\sum x_{2j} y_j)^2}{\sum x_{2j}^2} \quad (4.49)$$

Substituindo (4.45), (4.46) e (4.49) em (4.48), obtemos

$$r_{Y1:2} = \frac{\sum x_{1j} x_{2j} \sum x_{2j} y_j}{\sqrt{\left[ \sum x_{1j}^2 - \frac{(\sum x_{1j} x_{2j})^2}{\sum x_{2j}^2} \right] \left[ \sum y_j^2 - \frac{(\sum x_{2j} y_j)^2}{\sum x_{2j}^2} \right]}} \quad (4.50)$$

Dividindo o numerador e o denominador por  $\sqrt{\sum x_{1j}^2 \sum y_j^2}$ , verifica-se que

$$r_{Y1:2} = \frac{r_{Y1} - r_{12} r_{Y2}}{\sqrt{(1 - r_{12}^2)(1 - r_{Y2}^2)}} \quad (4.51)$$

Analogamente, temos

$$r_{Y2:1} = \frac{r_{Y2} - r_{12} r_{Y1}}{\sqrt{(1 - r_{12}^2)(1 - r_{Y1}^2)}} \quad (4.52)$$

Consideremos o exemplo numérico da tabela 4.1. Vamos calcular, inicialmente, os valores dos coeficientes de correlação simples:

$$r_{Y1} = \frac{\sum x_{1j} y_j}{\sqrt{\sum x_{1j}^2 \sum y_j^2}} = \frac{-54}{\sqrt{41,5 \cdot 116,5}} = -0,776617$$

$$r_{Y2} = \frac{\sum x_{2j} y_j}{\sqrt{\sum x_{2j}^2 \sum y_j^2}} = \frac{-30}{\sqrt{10 \cdot 116,5}} = -0,878938$$

$$r_{12} = \frac{\sum x_{1j} x_{2j}}{\sqrt{\sum x_{1j}^2 \sum x_{2j}^2}} = \frac{20}{\sqrt{41,5 \cdot 10}} = 0,981761$$

Essas correlações simples são apresentadas com grande número de decimais para evitar erros de arredondamento nos próximos cálculos. Substituindo esses valores em (4.51), obtemos

$$r_{Y1:2} = 0,952$$

Esse coeficiente de correlação parcial também pode ser obtido calculando-se o coeficiente de correlação simples entre os valores de  $v_j$  e  $z_j$ , da tabela 4.3.

$$r_{Y1:2} = \frac{\sum v_j z_j}{\sqrt{\sum v_j^2 \sum z_j^2}} = \frac{6}{\sqrt{1,5 \cdot 26,5}} = 0,952$$

É importante verificar a relação existente entre um coeficiente de correlação parcial e o correspondente coeficiente de regressão. De (4.50), obtemos

$$r_{Y1:2} = \frac{\left( \sum x_{1j} y_j - \frac{\sum x_{1j} x_{2j} \sum x_{2j} y_j}{\sum x_{2j}^2} \right)}{\left[ \sum x_{1j}^2 - \frac{(\sum x_{1j} x_{2j})^2}{\sum x_{2j}^2} \right]} \cdot \sqrt{\frac{\sum x_{1j}^2 - \frac{(\sum x_{1j} x_{2j})^2}{\sum x_{2j}^2}}{\sum y_j^2 - \frac{(\sum x_{2j} y_j)^2}{\sum x_{2j}^2}}}$$

Multiplicando o numerador e denominador da primeira fração por  $\sum x_{2j}^2$  e lembrando (4.40), segue-se que

$$r_{Y1:2} = b_1 \cdot \frac{\sqrt{\sum x_{1j}^2 - \frac{(\sum x_{1j} x_{2j})^2}{\sum x_{2j}^2}}}{\sqrt{\sum y_j^2 - \frac{(\sum x_{2j} y_j)^2}{\sum x_{2j}^2}}} = b_1 \frac{\sqrt{(\text{S.Q.Res. } x_{1j} | x_{2j})}}{\sqrt{(\text{S.Q.Res. } y_j | x_{2j})}}$$

ou

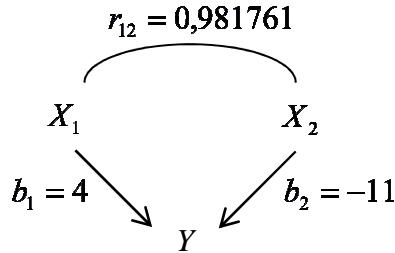
$$b_1 = r_{Y1:2} \frac{\sqrt{(\text{S.Q.Res. } y_j | x_{2j})}}{\sqrt{(\text{S.Q.Res. } x_{1j} | x_{2j})}} \quad (4.53)$$

Analogamente,

$$b_2 = r_{Y2:1} \frac{\sqrt{(\text{S.Q.Res. } y_j | x_{1j})}}{\sqrt{(\text{S.Q.Res. } x_{2j} | x_{1j})}} \quad (4.54)$$

As relações (4.53) e (4.54) devem ser comparadas com a relação (3.4).

Essas relações mostram que um coeficiente de correlação parcial sempre tem sinal igual ao do respectivo coeficiente de regressão na regressão múltipla. Mas o correspondente coeficiente de correlação simples pode ter sinal oposto, como ocorre com  $r_{Y_1}$  e  $r_{Y_{1,2}}$  no exemplo analisado. Vimos que  $r_{Y_1} = -0,776617$  e  $r_{Y_{1,2}} = 0,952$ . O esquema a seguir procura mostrar como isso é possível



O efeito direto de  $X_1$  sobre  $Y$  é positivo, como mostra o valor de  $b_1$  ou o valor de  $r_{Y_{1,2}}$ . Mas  $X_1$  tem forte correlação positiva com  $X_2$ , que, por sua vez, tem forte efeito negativo sobre  $Y$ . A correlação simples entre  $X_1$  e  $Y$  “mistura” os efeitos direto e indireto (via  $X_2$ ) de  $X_1$  sobre  $Y$ . Neste exemplo o efeito indireto é negativo e supera o efeito direto positivo, fazendo com que a correlação simples ( $r_{Y_1}$ ) seja negativa.

Veremos, a seguir, uma outra maneira de interpretar os coeficientes de correlação parcial.

Vamos considerar ainda o caso de uma regressão múltipla com duas variáveis explanatórias, ou seja, a regressão de  $y_j$  em relação a  $x_{1j}$  e  $x_{2j}$ . De acordo com (4.31), a soma de quadrados de regressão é dada por

$$(\text{S.Q.Regr.de } y_j | x_{1j} \text{ e } x_{2j}) = b_1 \sum x_{1j} y_j + b_2 \sum x_{2j} y_j$$

ou, lembrando a definição do coeficiente de determinação múltipla ( $R^2$ ),

$$(\text{S.Q.Regr. de } y_j | x_{1j} \text{ e } x_{2j}) = R^2 \sum y_j^2$$

A “contribuição de  $x_{1j}$ ” para essa soma de quadrados é a diferença entre esse valor e a soma de quadrados de regressão da regressão linear simples de  $y_j$  contra  $x_{2j}$ , isto é,

$$\begin{aligned} (\text{Contribuição de } x_{1j}) &= (\text{S.Q.Regr. de } y_j | x_{1j} \text{ e } x_{2j}) - (\text{S.Q.Regr. de } y_j | x_{2j}) = \\ &= R^2 \sum y_j^2 - r_{Y2}^2 \sum y_j^2 \end{aligned} \tag{4.55}$$

Para medir a importância da “contribuição de  $x_{1j}$ ”, comparamos seu valor com a soma de quadrados residual da regressão linear simples de  $y_j$  contra  $x_{2j}$ , por meio do cociente

$$\phi = \frac{(\text{Contribuição de } x_{1j})}{(\text{S.Q.Res. de } y_j | x_{2j})} = \frac{R^2 \sum y_j^2 - r_{Y2}^2 \sum y_j^2}{\sum y_j^2 - r_{Y2}^2 \sum y_j^2} = \frac{R^2 - r_{Y2}^2}{1 - r_{Y2}^2} \quad (4.56)$$

Esse cociente, chamado coeficiente de determinação parcial entre  $y_j$  e  $x_{1j}$ , é, no máximo, igual a um. Isso ocorre quando a introdução da variável  $x_{1j}$  “explicar” (em termos de soma de quadrados) tudo o que  $x_{2j}$  deixou de explicar. Está claro que, neste caso, o coeficiente de determinação múltipla também será igual a um.

Para obter a expressão do cociente  $\phi$  em função apenas dos coeficientes de correlação simples, utilizaremos a relação

$$R^2 = \frac{r_{Y1}^2 - 2r_{12}r_{Y1}r_{Y2} + r_{Y2}^2}{1 - r_{12}^2} \quad (4.57)$$

que pode ser obtida (após várias passagens algébricas que podem ser desenvolvidas, como exercício) substituindo (4.40) e (4.41) em

$$R^2 = \frac{b_1 \sum x_1 y_1 + b_2 \sum x_{2j} y_j}{\sum y_j^2}$$

Substituindo (4.57) em (4.56), simplificando e fatorando, obtemos

$$\phi = \frac{(r_{Y1} - r_{12}r_{Y2})^2}{(1 - r_{12}^2)(1 - r_{Y2}^2)}$$

Comparando esse resultado com (4.51), concluímos que o coeficiente de determinação parcial entre  $y_j$  e  $x_{1j}$  é igual ao quadrado do coeficiente de correlação parcial entre as mesmas variáveis. Então (4.56) fica

$$r_{Y1-2}^2 = \frac{(\text{Contribuição de } x_{1j})}{(\text{S.Q.Res. de } y_j | x_{2j})} = \frac{R^2 - r_{Y2}^2}{1 - r_{Y2}^2} \quad (4.58)$$

Analogamente,

$$r_{Y_{2:1}}^2 = \frac{(\text{Contribuição de } x_{2j})}{(\text{S.Q.Res. de } y_j | x_{1j})} = \frac{R^2 - r_{Y_1}^2}{1 - r_{Y_1}^2} \quad (4.59)$$

Para o exemplo numérico da tabela 4.1, temos

$$(\text{S.Q.Regr.de } y_j | x_{2j}) = \frac{(\sum x_{2j} y_j)^2}{\sum x_{2j}^2} = \frac{(-30)^2}{10} = 90$$

e

$$(\text{S.Q.Res.de } y_j | x_{2j}) = 116,5 - 90 = 26,5$$

Anteriormente já havíamos obtido (ver tabela 4.2)

$$(\text{S.Q.Regr.de } y_j | x_{1j} \text{ e } x_{2j}) = 114$$

Então

$$(\text{Contribuição de } x_{1j}) = 114 - 90 = 24$$

De acordo com (4.58) temos

$$r_{Y_{1:2}}^2 = \frac{24}{26,5} = 0,905660$$

Extraindo a raiz quadrada e adotando, de acordo com (4.53), o sinal de  $b_1$ , temos

$$r_{Y_{1:2}} = 0,952$$

que já obtivemos anteriormente de (4.51).

Podemos verificar a significância estatística da “contribuição de  $x_{1j}$ ” por meio de um teste  $F$ , como é mostrado na tabela 4.4.

TABELA 4.4. Análise de Variância

C.V.	G.L.	S.Q.	Q.M.	$F$
Regr. de $y_j   x_{2j}$	1	90		
Contribuição de $x_{1j}$	1	24	24	19,2
Regr. de $y_j   x_{1j} \text{ e } x_{2j}$	2	114	57	45,6
Resíduo	2	2,5	1,25	
Total	4	116,5		

Dividindo o quadrado médio referente à “contribuição de  $x_{1j}$ ” (que é igual à respectiva soma de quadrados, pois esta tem 1 grau de liberdade) pelo quadrado médio residual da regressão múltipla obtivemos  $F = 19,2$ .

Ao nível de significância de 5%, o valor crítico de  $F$ , para 1 e 2 graus de liberdade, é 18,51. Portanto, o resultado obtido é significativo.

É importante observar que esse teste  $F$  é equivalente ao teste  $t$ , feito anteriormente na seção 4.8, para testar a hipótese  $H_0 : \beta_1 = 0$  contra  $H_A : \beta_1 \neq 0$ . Note que o valor de  $F$  obtido (19,2) é igual ao quadrado do valor de  $t$  calculado para testar essa hipótese (4,382).

Até aqui analisamos o conceito de correlação parcial para o caso de uma regressão linear com duas variáveis explanatórias. O conceito pode, entretanto, ser generalizado para o caso de uma regressão linear múltipla com  $k$  variáveis explanatórias. Apenas para facilidade de notação, consideremos o coeficiente de correlação parcial entre  $y_j$  e  $x_{1j}$  ( $r_{Y1:23...k}$ ). Sendo  $v_j$  os desvios da regressão múltipla de  $x_{1j}$  contra  $x_{2j}, x_{3j}, \dots, x_{kj}$  e  $z_i$  os desvios da regressão múltipla de  $y_i$  contra  $x_{2j}, x_{3j}, \dots, x_{kj}$ , o coeficiente de correlação parcial entre  $y_j$  e  $x_{1j}$  é, por definição, o coeficiente de correlação simples entre  $v_j$  e  $z_j$ . Pode-se demonstrar que o mesmo resultado é obtido de<sup>6</sup>

$$r_{Y1:23...k}^2 = \frac{(\text{Contribuição de } x_{1j})}{(\text{S.Q.Res. de } y_j | x_{2j}, x_{3j}, \dots, x_{kj})}$$

onde

$$\begin{aligned} & (\text{Contribuição de } x_{1j}) = \\ & = (\text{S.Q.Regr. de } y_j | x_{1j}, x_{2j}, x_{3j}, \dots, x_{kj}) - (\text{S.Q.Regr. de } y_j | x_{2j}, x_{3j}, \dots, x_{kj}) \end{aligned}$$

A significância estatística da “contribuição de  $x_{1j}$ ” pode ser testada por meio de uma decomposição da soma de quadrados de regressão, como está indicado no esquema a seguir.

---

<sup>6</sup> Ver a seção 5.3 (p. 132-135) de Johnston (1972), para uma outra maneira de obter os coeficientes de correlação parcial.

### Esquema da Análise de Variância

C.V.	G.L.
Regr. de $y_j   x_{2j}, x_{3j}, \dots, x_{kj}$	$k - 1$
“Contribuição de $x_{1j}$ ”	1
Regr. de $y_j   x_{1j}, x_{2j}, x_{3j}, \dots, x_{kj}$	$k = p - 1$
Resíduo	$n - p$
Total	$n - 1$

Pode-se demonstrar que o teste  $F$  para a “contribuição de  $x_{hj}$ ” é sempre equivalente ao teste  $t$  da hipótese  $H_0 : \beta_h = 0$  contra  $H_A : \beta_h \neq 0$ , isto é, se  $t_h = b_h / s(b_h)$ , temos

$$t_h^2 = \frac{(\text{Contribuição de } x_{hj})}{s^2}$$

e  $(\text{Contribuição de } x_{hj}) = t_h^2 s^2$

A soma de quadrados residual da regressão completa é  $(n - p)s^2$ . Então a soma de quadrados residual da regressão sem  $x_{hj}$  é  $(n - p)s^2 + (\text{Contribuição de } x_{hj})$  e o coeficiente de determinação parcial entre  $y_j$  e  $x_{hj}$  é

$$r_{Yh:i \neq h}^2 = \frac{(\text{Contribuição de } x_{hj})}{(n - p)s^2 + (\text{Contribuição de } x_{hj})}$$

Segue-se que

$$r_{Yh:i \neq h}^2 = \frac{t_h^2 s^2}{(n - p)s^2 + t_h^2 s^2}$$

ou

$$r_{Yh:i \neq h}^2 = \frac{t_h^2}{t_h^2 + n - p}$$

Essa expressão permite obter com facilidade um coeficiente de determinação parcial a partir do valor de  $t$  referente à hipótese de nulidade do coeficiente correspondente na regressão múltipla.

#### 4.12. Intervalos de confiança e regiões de confiança para os parâmetros

Consideremos, novamente, o modelo de regressão linear com duas variáveis explanatórias:

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + u_j$$

Nesta seção vamos utilizar um novo exemplo numérico, baseado na amostra de 6 observações apresentada na tabela 4.5. Pode-se verificar que  $\bar{Y} = 9,5$ ,  $\bar{X}_1 = 1,5$  e  $\bar{X}_2 = 3$ . A mesma tabela mostra os valores das variáveis centradas  $y_j$ ,  $x_{1j}$  e  $x_{2j}$ .

TABELA 4.5. Valores de três variáveis em uma amostra com 6 observações.

$Y_j$	$X_{1j}$	$X_{2j}$	$y_j$	$x_{1j}$	$x_{2j}$
1,5	0	0	-8,0	-1,5	-3
6,5	1	2	-3,0	-0,5	-1
10,0	1	4	0,5	-0,5	1
11,0	2	2	1,5	0,5	-1
11,5	2	4	2,0	0,5	1
16,5	3	6	7,0	1,5	3

Tendo em vista o modelo com todas as variáveis centradas, obtemos

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \sum x_{1j}^2 & \sum x_{1j}x_{2j} \\ \sum x_{1j}x_{2j} & \sum x_{2j}^2 \end{bmatrix} = \begin{bmatrix} 5,5 & 9 \\ 9 & 22 \end{bmatrix},$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} \sum x_{1j}y_j \\ \sum x_{2j}y_j \end{bmatrix} = \begin{bmatrix} 25,5 \\ 49 \end{bmatrix},$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \frac{11}{20} & -\frac{9}{40} \\ -\frac{9}{40} & \frac{11}{80} \end{bmatrix} = \begin{bmatrix} 0,55 & -0,225 \\ -0,225 & 0,1375 \end{bmatrix}$$

e

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

A equação estimada é

$$\hat{y}_j = 3x_{1j} + x_{2j}$$

ou

$$\hat{Y}_j = 2 + 3X_{1j} + X_{2j}$$

A tabela 4.6 mostra a análise de variância da regressão, verificando-se que  $R^2 = 0,977$ .

TABELA 4.6. Análise da Variância

C.V.	G.L.	S.Q.	Q.M.	F
Regressão	2	125,5	62,75	62,75
Resíduo	3	3	1	
Total	5	128,5		

Como  $s^2 = 1$ , temos

$$\hat{V}(b_1) = 0,55 \quad \text{e} \quad \hat{V}(b_2) = 0,1375$$

Seguindo o procedimento apresentado na seção 4.8, verifica-se que

$$\hat{V}(a) = \frac{37}{60}$$

Adotando um nível de significância de 1%, o valor crítico de  $F$ , com 2 e 3 graus de liberdade, é 30,82. Portanto, o resultado é significativo, isto é, rejeita-se, a esse nível de significância, a hipótese  $H_0 : \beta_1 = \beta_2 = 0$ .

Para testar a hipótese  $H_0 : \beta_1 = 0$ , contra  $H_A : \beta_1 \neq 0$ , calculamos

$$t = \frac{3}{\sqrt{0,55}} = 4,045$$

Como o valor crítico de  $t$  para 3 graus de liberdade, ao nível de significância de 1%, é 5,841, o resultado obtido não é significativo, isto é, não rejeitamos a hipótese  $H_0 : \beta_1 = 0$ .

Para testar a hipótese  $H_0 : \beta_2 = 0$  contra a  $H_A : \beta_2 \neq 0$ , calculamos

$$t = \frac{1}{\sqrt{0,1375}} = 2,697, \text{ não significativo}$$

É interessante notar que neste exemplo, embora se rejeite, ao nível de significância de 1%, a hipótese  $H_0 : \beta_1 = \beta_2 = 0$ , não se rejeita, ao mesmo nível de significância, nem a hipótese  $H_0 : \beta_1 = 0$ , nem a hipótese  $H_0 : \beta_2 = 0$ . Veremos, adiante, porque isso pode ocorrer.

Na seção 4.5 vimos que o intervalo de confiança para o parâmetro  $\beta_i$  de uma regressão linear múltipla é

$$b_i - t_0 s(b_i) < \beta_i < b_i + t_0 s(b_i)$$

Vamos determinar os intervalos de confiança, ao nível de confiança de 99%, para os parâmetros  $\alpha$ ,  $\beta_1$  e  $\beta_2$ , com base nos dados da tabela 4.5. O valor crítico de  $t$  é, neste caso, 5,841.

Temos

$$a = 2, b_1 = 3 \text{ e } b_2 = 1.$$

O intervalo de confiança para  $\alpha$  é

$$2 - 5,841\sqrt{\frac{37}{60}} < \alpha < 2 + 5,841\sqrt{\frac{37}{60}}$$

ou

$$-2,59 < \alpha < 6,59$$

O intervalo de confiança para  $\beta_1$  é

$$3 - 5,841\sqrt{0,55} < \beta_1 < 3 + 5,841\sqrt{0,55}$$

ou

$$-1,33 < \beta_1 < 7,33$$

O intervalo de confiança para  $\beta_2$  é

$$1 - 5,841\sqrt{0,1375} < \beta_2 < 1 + 5,841\sqrt{0,1375}$$

ou

$$-1,17 < \beta_2 < 3,17$$

Esses intervalos de confiança devem ser interpretados com cuidado. Consideremos, por simplicidade, apenas os parâmetros  $\beta_1$  e  $\beta_2$ . Na figura 4.1 assinalamos os intervalos de confiança e traçamos a elipse que delimita a região de confiança, ao nível de confiança de 99%, para esses parâmetros. Note que, apesar de tanto o intervalo de confiança para  $\beta_1$  como o intervalo de confiança para  $\beta_2$  incluírem o valor zero, o ponto  $(\beta_1 = 0, \beta_2 = 0)$  não pertence à região de confiança para  $\beta_1$  e  $\beta_2$ . É por isso que, embora os valores de  $t$  obtidos não nos levem a rejeitar, ao nível de significância de 1%, as hipóteses  $H_0 : \beta_1 = 0$  ou  $H_0 : \beta_2 = 0$ , o teste  $F$  permite rejeitar, a esse mesmo nível de significância, a hipótese  $H_0 : \beta_1 = \beta_2 = 0$ .

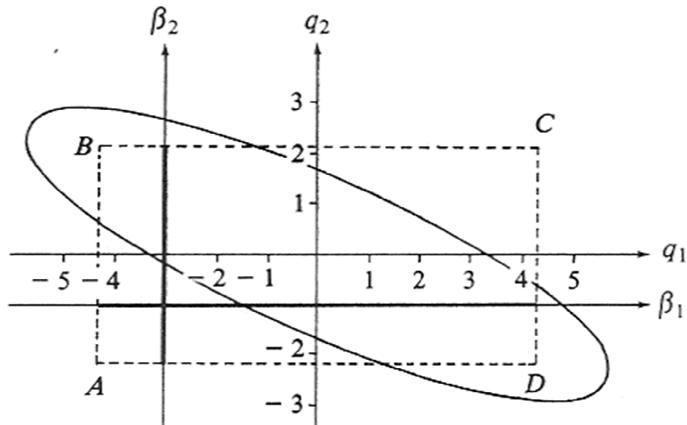


Figura 4.1. A região de confiança para os parâmetros  $\beta_1$  e  $\beta_2$ .

O conjunto de pontos do retângulo  $ABCD$  corresponde aos valores de  $\beta_1$  e  $\beta_2$  que pertencem aos respectivos intervalos de confiança. Poder-se-ia pensar que esse retângulo seria a região de confiança para  $\beta_1$  e  $\beta_2$ . Entretanto, embora esse retângulo e a região de confiança correta (elíptica) tenham uma área em comum, é fácil verificar que existem pontos do retângulo que não pertencem à região de confiança (como é o caso de  $\beta_1 = \beta_2 = 0$ ) e pontos da região de confiança que não pertencem ao retângulo.

Vejamos, agora, como foi obtida a elipse da figura 4.1.

Para o modelo de regressão linear múltipla, com todas as pressuposições apresentadas na seção 4.1, pode-se demonstrar que

$$(\mathbf{C}\mathbf{b} - \mathbf{C}\boldsymbol{\beta})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}(\mathbf{C}\mathbf{b} - \mathbf{C}\boldsymbol{\beta}) \frac{1}{ms^2} = F \quad (4.60)$$

onde  $m$  é a característica da matriz de constantes  $\mathbf{C}$  e  $F$  está associado a  $m$  e  $n - p$  graus de liberdade.

A relação (4.60) é muito geral. Mostraremos, inicialmente, que o teste  $F$  relativo à hipótese  $H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$  é um de seus casos particulares.

Consideremos o modelo de regressão linear múltipla com todas as variáveis centradas:

$$y_j = \sum_{i=1}^k \beta_i x_{ij} + u_j - \bar{u} \quad (4.61)$$

Fazendo  $\mathbf{C} = \mathbf{I}_k$ , onde  $\mathbf{I}_k$  é uma matriz identidade com característica  $m = k$ , a hipótese da nulidade fica

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$$

Em (4.60), fazendo  $\mathbf{C} = \mathbf{I}_k$  e  $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ , obtemos

$$F = \frac{\mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}}{ks^2}$$

Lembrando que  $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ , obtemos

$$F = \frac{\mathbf{b}'\mathbf{X}'\mathbf{y}}{ks^2} = \frac{\mathbf{b}'\mathbf{X}'\mathbf{y}}{k} = \frac{s^2}{s^2} = \frac{\text{Q.M.Regr.}}{\text{Q.M.Res.}},$$

que é, como sabemos, o valor de  $F$  usualmente calculado para verificar a significância estatística da regressão ajustada.

Um outro caso particular de (4.60) é o teste  $F$  para a “contribuição” de uma variável que, como vimos na seção 4.11, equivale ao teste  $t$  relativo à hipótese  $H_0 : \beta_i = 0$ . Consideremos o modelo (4.61) e admitamos, para facilitar a indicação, que se deseja verificar se a contribuição da variável  $x_{1j}$  é significativa, ou seja, vamos testar a hipótese  $H_0 : \beta_1 = 0$ .

Fazemos

$$\mathbf{C} = [1 \quad 0 \quad 0 \quad \dots \quad 0], \text{ cuja característica é } m = 1.$$

Então, a hipótese da nulidade pode ser escrita como segue:

$$H_0 : \mathbf{C}\boldsymbol{\beta} = 0$$

Temos, também, que

$$\mathbf{C}\mathbf{b} = b_1$$

e

$$\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}' = w_{11},$$

onde  $w_{11}$  é o primeiro elemento da diagonal principal de  $(\mathbf{X}'\mathbf{X})^{-1}$ .

Substituindo esses resultados em (4.60) e considerando que  $\mathbf{C}\boldsymbol{\beta} = 0$ , obtemos

$$F = \frac{b_1(w_{11})^{-1}b_1}{s^2} = \frac{b_1^2}{w_{11}s^2}$$

Como  $w_{11}s^2 = s^2(b_1)$  segue-se que

$$|t| = \sqrt{F} = \frac{|b_1|}{s(b_1)}$$

Uma vez que o quadrado de um teste  $t$  é sempre igual a um teste  $F$  com numerador associado a um grau de liberdade, pode-se verificar que a relação (4.60) engloba, como caso particular, qualquer  $t$  relativo a uma hipótese sobre o valor de um parâmetro ou sobre o valor de uma combinação linear de parâmetros, incluindo-se, neste último caso, um teste a respeito de  $E(Y_h)$ .

Se, escolhido um nível de confiança, substituirmos  $F$ , em (4.60), pelo seu valor crítico  $F_0$ , essa relação nos fornecerá os limites de um intervalo ou de uma região de confiança (dependendo de como é definida a matriz  $\mathbf{C}$ ). Os pontos pertencentes ao intervalo ou região de confiança obedecem à desigualdade

$$(\mathbf{Cb} - \mathbf{C}\beta)'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}(\mathbf{Cb} - \mathbf{C}\beta) < F_0 ms^2 \quad (4.62)$$

Consideremos, por exemplo, que se deseja obter o intervalo de confiança para  $\beta_1$ . Tendo em vista o modelo (4.61), fazemos

$$\mathbf{C} = [1 \quad 0 \quad 0 \quad \dots \quad 0], \quad \text{cuja característica é } m = 1.$$

Então,

$$\mathbf{Cb} = b_1, \quad \mathbf{C}\beta = \beta_1 \quad \text{e} \quad \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}' = w_{11}$$

Substituindo esses resultados em (4.62), obtemos

$$\begin{aligned} (b_1 - \beta_1)(w_{11})^{-1}(b_1 - \beta_1) &< F_0 s^2 \\ (\beta_1 - b_1)^2 &< F_0 w_{11} s^2 \\ -\sqrt{F_0 w_{11} s^2} &< \beta_1 - b_1 < \sqrt{F_0 w_{11} s^2} \\ b_1 - \sqrt{F_0 w_{11} s^2} &< \beta_1 < b_1 + \sqrt{F_0 w_{11} s^2} \end{aligned}$$

Finalmente, como  $\sqrt{F_0} = t_0$  e  $\sqrt{w_{11} s^2} = s(b_1)$ , temos

$$b_1 - t_0 s(b_1) < \beta_1 < b_1 + t_0 s(b_1)$$

Verificamos, assim, que o intervalo de confiança para  $\beta_1$  pode ser considerado como um caso particular de (4.62).

Determinemos, agora, a região de confiança para os parâmetros  $\beta_1$  e  $\beta_2$  de uma regressão linear múltipla com 2 variáveis explanatórias. Tendo em vista o modelo

$$Y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + u_j$$

fazemos

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

cuja característica é  $m = 2$ . Então

$$\mathbf{C}\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \quad \text{e} \quad \mathbf{C}\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

Fazendo

$$\beta_1 - b_1 = q_1 \quad \text{e} \quad \beta_2 - b_2 = q_2 \quad (4.63)$$

segue-se que

$$\mathbf{C}\mathbf{b} - \mathbf{C}\boldsymbol{\beta} = -\begin{bmatrix} q_1 \\ q_2 \end{bmatrix} \quad (4.64)$$

Temos, também, que

$$\begin{aligned} \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}' &= \\ &= \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{n} & 0 & 0 \\ 0 & w_{11} & w_{12} \\ 0 & w_{12} & w_{22} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} \\ w_{12} & w_{22} \end{bmatrix} \end{aligned}$$

Donde

$$[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1} = \begin{bmatrix} \sum x_{1j}^2 & \sum x_{1j}x_{2j} \\ \sum x_{1j}x_{2j} & \sum x_{2j}^2 \end{bmatrix} \quad (4.65)$$

Substituindo (4.64) e (4.65) em (4.62) e lembrando que  $m = 2$ , obtemos

$$\begin{bmatrix} q_1 & q_2 \end{bmatrix} \begin{bmatrix} \sum x_{1j}^2 & \sum x_{1j}x_{2j} \\ \sum x_{1j}x_{2j} & \sum x_{2j}^2 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} < 2F_0 s^2 \quad (4.66)$$

No caso do exemplo numérico apresentado no início desta seção, temos

$$\sum x_{1j}^2 = 5,5 \quad \sum x_{2j}^2 = 22$$

$$\sum x_{1j}x_{2j} = 9 \quad s^2 = 1$$

Para o nível de confiança de 99%, o valor crítico de  $F$  com 2 e 3 graus de liberdade é 30,82.

Substituindo esses resultados em (4.66), obtemos

$$[q_1 \quad q_2] \begin{bmatrix} 5,5 & 9 \\ 9 & 22 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} < 2 \cdot 30,82$$

ou

$$5,5q_1^2 + 18q_1q_2 + 22q_2^2 - 61,64 < 0$$

Essa desigualdade é satisfeita pelos pontos delimitados pela elipse

$$5,5q_1^2 + 18q_1q_2 + 22q_2^2 - 61,64 = 0,$$

que é a elipse traçada na figura 4.1.

Para mais uma aplicação da relação (4.60), consideremos que, no exemplo numérico que estamos desenvolvendo, desejamos testar, ao nível de significância de 5%, a hipótese

$$H_0 : \beta_1 = 4 \text{ e } \beta_2 = 2$$

Ressaltemos que esta é uma única hipótese envolvendo, concomitantemente, os valores de dois parâmetros e que o teste dessa hipótese não equivale a fazer dois testes consecutivos, um para a hipótese  $H_0 : \beta_1 = 4$  e outro para a hipótese  $H_0 : \beta_2 = 2$ .

Fazendo

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

a hipótese  $H_0 : \beta_1 = 4 \text{ e } \beta_2 = 2$  pode ser indicada como segue:

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$$

Como

$$\mathbf{Cb} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix},$$

temos que

$$\mathbf{Cb} - \mathbf{C}\boldsymbol{\beta} = \begin{bmatrix} 3-4 \\ 1-2 \end{bmatrix} = -\begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{4.67}$$

Substituindo (4.65) e (4.67) em (4.60) e lembrando que  $m = 2$  e  $s^2 = 1$ , obtemos

$$F = \frac{1}{2} [1 \quad 1] \begin{bmatrix} 5,5 & 9 \\ 9 & 22 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 22,75$$

Como o valor crítico de  $F$  com 2 e 3 graus de liberdade e ao nível de significância de 5% é 9,55, o resultado é significativo, isto é, rejeita-se a hipótese  $H_0 : \beta_1 = 4$  e  $\beta_2 = 2$ .

Se tivéssemos adotado o nível de significância de 1%, não rejeitaríamos  $H_0 : \beta_1 = 4$  e  $\beta_2 = 2$ , pois, neste caso, o valor crítico de  $F$  é 30,82. Isso pode ser verificado na figura 4.1, notando que o ponto  $(\beta_1 = 4, \beta_2 = 2)$  pertence à região de 99% de confiança para  $\beta_1$  e  $\beta_2$ .

#### **4.13. Exemplo de uma regressão linear múltipla com três variáveis explanatórias**

Nesta seção, desenvolveremos um exemplo de regressão linear múltipla com três variáveis explanatórias, como ilustração do que já foi vista neste capítulo.

Na tabela 4.7 são apresentados 14 valores das variáveis  $Y_j$ ,  $X_{1j}$ ,  $X_{2j}$  e  $X_{3j}$ . Note que as três variáveis explanatórias já são centradas.

TABELA 4.7. Amostra de 14 observações para 4 variáveis.

$Y_j$	$X_{1j} = x_{1j}$	$X_{2j} = x_{2j}$	$X_{3j} = x_{3j}$
8,5	-2	2	-2
1,0	-1	-1	0
4,0	-1	0	0
4,0	-1	0	0
5,0	-1	1	0
3,0	1	-1	0
6,0	1	0	0
6,0	1	0	0
7,0	1	1	0
5,0	0	0	-1
5,0	0	0	0
5,0	0	0	0
3,0	0	0	1
0,5	2	-2	2

Os valores básicos a serem calculados são:

$$\begin{array}{lll} \sum Y_j = 63 & \sum x_{1j}^2 = 16 & \bar{Y} = 4,5 \\ \sum x_{2j}^2 = 12 & \sum y_j^2 = 61 & \sum x_{3j}^2 = 10 \\ \sum x_{1j}Y_j = -8 & \sum x_{1j}x_{2j} = -8 & \sum x_{2j}Y_j = 24 \\ \sum x_{1j}x_{3j} = 8 & \sum x_{3j}Y_j = -18 & \sum x_{2j}x_{3j} = -8 \end{array}$$

Tendo em vista o modelo

$$Y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + u_j, \quad j = 1, 2, \dots, 14$$

construímos as matrizes

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 14 & 0 & 0 & 0 \\ 0 & 16 & -8 & 8 \\ 0 & -8 & 12 & -8 \\ 0 & 8 & -8 & 10 \end{bmatrix} \text{ e } \mathbf{X}'\mathbf{y} = \begin{bmatrix} 63 \\ -8 \\ 24 \\ -18 \end{bmatrix}$$

A seguir, obtemos

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \frac{1}{14} & 0 & 0 & 0 \\ 0 & \frac{7}{64} & \frac{1}{32} & -\frac{1}{16} \\ 0 & \frac{1}{32} & \frac{3}{16} & \frac{1}{8} \\ 0 & -\frac{1}{16} & \frac{1}{8} & \frac{1}{4} \end{bmatrix}$$

e

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \begin{bmatrix} 4,5 \\ 1 \\ 2 \\ -1 \end{bmatrix}$$

A equação estimada é

$$\hat{Y}_j = 4,5 + x_{1j} + 2x_{2j} - x_{3j}$$

ou, uma vez neste exemplo  $x_{ij} = X_{ij}$  ( $i = 1, 2, 3$ ),

$$\hat{Y}_j = 4,5 + X_{1j} + 2X_{2j} - X_{3j}$$

Temos

$$\begin{aligned} \text{S.Q.Regr.} &= \sum_i b_i \sum_j x_{ij} Y_j = \\ &= 1 \cdot (-8) + 2 \cdot 24 + (-1) \cdot (-18) = 58 \end{aligned}$$

A análise da variância é dada na tabela 4.8.

TABELA 4.8. Análise de Variância

C.V.	G.L.	S.Q.	Q.M.	F
Regressão	3	58	19,33	64,44
Resíduo	10	3	0,30	
Total	13	61		

Ao nível de significância de 1% e com 3 e 10 graus de liberdade, o valor crítico de  $F$  é 6,55. O resultado obtido é, portanto, significativo, isto é, rejeita-se, a esse nível de significância, a hipótese  $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ .

O coeficiente de determinação múltipla é

$$R^2 = \frac{58}{61} = 0,951$$

Sabemos que as estimativas das variâncias e covariâncias das estimativas dos parâmetros são dadas por  $(\mathbf{X}'\mathbf{X})^{-1} s^2$ .

Assim, por exemplo, a estimativa da variância de  $b_2$  é

$$\hat{V}(b_2) = \frac{9}{160}$$

Então

$$s(b_2) = \sqrt{\frac{9}{160}} = 0,237$$

O intervalo de confiança para  $\beta_2$ , ao nível de confiança de 95%, é

$$2 - 2,228 \cdot 0,237 < \beta_2 < 2 + 2,228 \cdot 0,237$$

ou

$$1,47 < \beta_2 < 2,53$$

Para testar, ao nível de significância de 1%, a hipótese  $H_0 : \beta_2 = 0$ , contra a hipótese alternativa  $H_A : \beta_2 \neq 0$ , calculamos

$$t = \frac{b_2 - 0}{s(b_2)} = \frac{2}{\sqrt{\frac{9}{160}}} = 8,433$$

Como o valor crítico de  $t$  para 10 graus de liberdade é 3,169, o resultado obtido é significativo, isto é, rejeita-se, ao nível de significância de 1%, a hipótese de que  $\beta_2 = 0$ , em favor da hipótese de que  $\beta_2 \neq 0$ .

Passemos, agora, à determinação do coeficiente de correlação parcial entre  $Y_j$  e  $X_{3j}$ , dados  $X_{1j}$  e  $X_{2j}$ . Para isso obteremos, inicialmente, as estimativas dos coeficientes da regressão de  $Y_j$  em relação a  $x_{1j}$  e  $x_{2j}$ . Sendo  $c_0$ ,  $c_1$  e  $c_2$  as estimativas dos coeficientes dessa regressão e  $z_j$  os desvios, temos

$$Y_j = c_0 + c_1 x_{1j} + c_2 x_{2j} + z_j$$

De acordo com (4.7), obtemos

$$\begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 14 & 0 & 0 \\ 0 & 16 & -8 \\ 0 & -8 & 12 \end{bmatrix}^{-1} \begin{bmatrix} 63 \\ -8 \\ 24 \end{bmatrix} =$$

$$= \begin{bmatrix} \frac{1}{14} & 0 & 0 \\ 0 & \frac{3}{32} & \frac{1}{16} \\ 0 & \frac{1}{16} & \frac{1}{8} \end{bmatrix} \begin{bmatrix} 63 \\ -8 \\ 24 \end{bmatrix} = \begin{bmatrix} 4,5 \\ 0,75 \\ 2,5 \end{bmatrix}$$

A soma de quadrados de regressão dessa regressão é

$$(S.Q.Regr. de Y_j | x_{1j}, x_{2j}) = 0,75 \cdot (-8) + 2,5 \cdot 24 = 54$$

Então

$$(S.Q.Res. de Y_j | x_{1j}, x_{2j}) = 61 - 54 = 7$$

Como

$$(S.Q.Regr. de Y_j | x_{1j}, x_{2j}, x_{3j}) = 58,$$

segue-se que

$$(\text{Contribuição de } x_{3j}) = 58 - 54 = 4$$

Então, o coeficiente de determinação parcial entre  $Y_j$  e  $X_{3j}$  é

$$r_{Y312}^2 = \frac{(\text{contribuição de } x_{3j})}{(\text{S.Q.Res. de } Y_j | x_{1j}, x_{2j})} = \frac{4}{7}$$

Lembrando que  $b_3$  é negativo, concluímos que o coeficiente de correlação parcial entre  $Y_j$  e  $X_{3j}$  é

$$r_{Y312} = -\sqrt{\frac{4}{7}} = -0,756$$

Os demais coeficientes de correlação parcial podem ser obtidos de maneira análoga.

Façamos, agora, o teste da hipótese

$$H_0 : \beta_1 + \beta_2 + \beta_3 = 1$$

contra a hipótese alternativa

$$H_A : \beta_1 + \beta_2 + \beta_3 \neq 1,$$

ao nível de significância de 5%.

Fazendo

$$\mathbf{c}' = [0 \ 1 \ 1 \ 1],$$

a hipótese da nulidade fica

$$H_0 : \mathbf{c}' \boldsymbol{\beta} = 1$$

De acordo com (4.12) temos

$$\hat{V}(\mathbf{c}' \mathbf{b}) = \mathbf{c}' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{c} s^2 =$$

$$= [0 \ 1 \ 1 \ 1] \begin{bmatrix} \frac{1}{14} & 0 & 0 & 0 \\ 0 & \frac{7}{64} & \frac{1}{32} & -\frac{1}{16} \\ 0 & \frac{1}{32} & \frac{3}{16} & \frac{1}{8} \\ 0 & -\frac{1}{16} & \frac{1}{8} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} 0,30 = \frac{141}{640}$$

A seguir, obtemos

$$t = \frac{\mathbf{c}'\mathbf{b} - \mathbf{c}'\boldsymbol{\beta}}{\sqrt{\hat{V}(\mathbf{c}'\mathbf{b})}} = \frac{2-1}{\sqrt{\frac{141}{640}}} = 2,130$$

Como, ao nível de significância de 5% e para 10 graus de liberdade, o valor crítico de  $t$  é 2,228, o resultado obtido não é significativo.

Determinemos, agora, a região de 95% de confiança para os parâmetros  $\beta_2$  e  $\beta_3$

Fazemos

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

cuja característica é  $m = 2$ .

Temos que

$$\mathbf{C}\boldsymbol{\beta} = \begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix}, \quad \mathbf{C}\mathbf{b} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

e

$$\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}' = \begin{bmatrix} \frac{3}{16} & \frac{1}{18} \\ \frac{1}{8} & \frac{1}{4} \end{bmatrix}$$

Substituindo esses resultados em (4.62) obtemos

$$\begin{bmatrix} 2-\beta_2 \\ -1-\beta_3 \end{bmatrix}' \begin{bmatrix} \frac{3}{16} & \frac{1}{8} \\ \frac{1}{8} & \frac{1}{4} \end{bmatrix}^{-1} \begin{bmatrix} 2-\beta_2 \\ -1-\beta_3 \end{bmatrix} < F_0 \cdot 2 \cdot 0,30$$

Como o valor crítico de  $F$  com 2 e 10 graus de liberdade, para um nível de confiança de 95%, é 4,10, segue-se que os pontos pertencentes à região de confiança obedecem à desigualdade

$$\begin{bmatrix} \beta_2 - 2 \\ \beta_3 + 1 \end{bmatrix}' \begin{bmatrix} 8 & -4 \\ -4 & 6 \end{bmatrix} \begin{bmatrix} \beta_2 - 2 \\ \beta_3 + 1 \end{bmatrix} < 4,10 \cdot 2 \cdot 0,30$$

Sabemos que essa região de confiança é delimitada por uma elipse com centro no ponto  $\beta_2 = 2$  e  $\beta_3 = -1$ .

A região de confiança para  $\beta_1$ ,  $\beta_2$  e  $\beta_3$  é um elipsóide num espaço com 3 dimensões. A visualização desse elipsóide exigiria a determinação e o traçado de elipses que resultam da intersecção da superfície do elipsóide com planos perpendulares a um dos três eixos coordenados, para vários valores do parâmetro correspondente. Portanto, é fácil ver que a quantidade de cálculos exigida para a determinação das regiões de confiança dos parâmetros de uma regressão cresce muito rapidamente com o número de parâmetros envolvidos e que a visualização dessas regiões de confiança se torna difícil para mais de dois parâmetros.

#### 4.14. Problemas de especificação

Vimos, no Capítulo 2, que no caso de uma regressão linear simples, o problema de especificação da relação entre as duas variáveis consiste em escolher o tipo de função, isto é, o modelo matemático.

Surge outro problema de especificação quando mais de uma variável explanatória pode estar afetando a variável dependente. Então, além de escolher o tipo de função, é necessário determinar quais as variáveis explanatórias que devem ser consideradas no modelo.

Vamos analisar o que ocorre com as estimativas dos coeficientes quando se cometem erros de especificação da matriz  $\mathbf{X}$ .

Admitamos que a relação verdadeira seja

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \quad (4.68)$$

e que o pesquisador, erroneamente, utilize, em lugar de  $\mathbf{X}$ , uma matriz  $\mathbf{V}$ . É óbvio que, geralmente, as matrizes  $\mathbf{X}$  e  $\mathbf{V}$  têm algumas colunas em comum.

De acordo com o método de mínimos quadrados, o pesquisador em questão obterá

$$\mathbf{g} = (\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'\mathbf{y} \quad (4.69)$$

enquanto que as estimativas corretas seriam

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Substituindo (4.68) em (4.69) obtemos

$$\mathbf{g} = (\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'\mathbf{u}$$

e

$$E(\mathbf{g}) = \mathbf{P}\boldsymbol{\beta}, \quad (4.70)$$

onde

$$\mathbf{P} = (\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'\mathbf{X} \quad (4.71)$$

Para mais facilmente explicar o problema, consideremos um caso em que o modelo correto é

$$y_j = \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + u_j - \bar{u}$$

e o pesquisador, erroneamente, obtém a equação estimada

$$\hat{y}_j = g_1 x_{1j} + g_2 x_{2j}$$

Neste caso

$$\mathbf{V}'\mathbf{V} = \begin{bmatrix} \sum x_{1j}^2 & \sum x_{1j}x_{2j} \\ \sum x_{1j}x_{2j} & \sum x_{2j}^2 \end{bmatrix}$$

e

$$\mathbf{V}'\mathbf{X} = \begin{bmatrix} \sum x_{1j}^2 & \sum x_{1j}x_{2j} & \sum x_{1j}x_{3j} \\ \sum x_{1j}x_{2j} & \sum x_{2j}^2 & \sum x_{2j}x_{3j} \end{bmatrix}$$

Substituindo esses resultados em (4.17), obtemos

$$\mathbf{P} = (\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'\mathbf{X} = \begin{bmatrix} 1 & 0 & \hat{\theta}_1 \\ 0 & 1 & \hat{\theta}_2 \end{bmatrix} \quad (4.72)$$

onde  $\hat{\theta}_1$  e  $\hat{\theta}_2$  são estimativas dos coeficientes de regressão de  $x_{3j}$  contra  $x_{1j}$  e  $x_{2j}$ .

De (4.70) e (4.72), segue-se que

$$E(\mathbf{g}) = \begin{bmatrix} 1 & 0 & \hat{\theta}_1 \\ 0 & 1 & \hat{\theta}_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} \beta_1 + \hat{\theta}_1\beta_3 \\ \beta_2 + \hat{\theta}_2\beta_3 \end{bmatrix}$$

ou seja,

$$E(g_1) = \beta_1 + \hat{\theta}_1\beta_3$$

e

$$E(g_2) = \beta_2 + \hat{\theta}_2 \beta_3$$

Verificamos que as estimativas dos coeficientes obtidos ( $g_1$  e  $g_2$ ) com o modelo erroneamente especificado são tendenciosas. O viés de  $g_i$  como estimativa de  $\beta_i$  depende do valor do parâmetro da variável excluída ( $\beta_3$ , no caso acima) e do valor da estimativa do coeficiente relativo a  $x_{ij}$  na regressão da variável excluída contra as variáveis incluídas.

Consideremos, agora, o caso em que o modelo correto seria

$$y_j = \beta_1 x_{1j} + \beta_2 x_{2j} + u_j - \bar{u}$$

e o pesquisador, erroneamente, ajusta a regressão

$$\hat{y}_j = g_1 x_{1j} + g_2 x_{2j} + g_3 x_{3j}$$

Neste caso

$$\mathbf{V}'\mathbf{V} = \begin{bmatrix} \sum x_{1j}^2 & \sum x_{1j}x_{2j} & \sum x_{1j}x_{3j} \\ \sum x_{1j}x_{2j} & \sum x_{2j}^2 & \sum x_{2j}x_{3j} \\ \sum x_{1j}x_{3j} & \sum x_{2j}x_{3j} & \sum x_{3j}^2 \end{bmatrix}$$

e

$$\mathbf{V}'\mathbf{X} = \begin{bmatrix} \sum x_{1j}^2 & \sum x_{1j}x_{2j} \\ \sum x_{1j}x_{2j} & \sum x_{2j}^2 \\ \sum x_{1j}x_{3j} & \sum x_{2j}x_{3j} \end{bmatrix}$$

Então, de acordo com (4.71),

$$\mathbf{P} = (\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

Substituindo esse resultado em (4.70), obtemos

$$E(\mathbf{g}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ 0 \end{bmatrix},$$

ou seja,

$$E(g_1) = \beta_1,$$

$$E(g_2) = \beta_2$$

e

$$E(g_3) = 0$$

É interessante notar que quando incluímos uma variável desnecessária, as estimativas dos coeficientes permanecem não-tendenciosas, diferentemente do que ocorre quando deixamos de incluir uma das variáveis explanatórias importantes. Isso mostra que é preferível incluir uma variável desnecessária que não incluir uma variável relevante. Entretanto, a inclusão de variáveis desnecessárias também é prejudicial, pois em geral faz com que aumente a variância dos estimadores.

Há, também, o perigo de um controle inapropriado mascarar o efeito que se deseja captar. Considere um pesquisador que deseja avaliar o efeito das transferências de renda do Programa Bolsa Família sobre a pobreza, utilizando dados por Unidade de Federação. A variável dependente é a redução da pobreza e a variável explanatória fundamental é o montante de transferências per capita em certo período. Devem ser controladas características específicas de cada Unidade da Federação que condicionam o efeito das transferências sobre a pobreza, mas é um absurdo incluir, nesses controles, mudanças na renda média e no índice de Gini da distribuição da renda em cada Unidade da Federação. Aumentando a renda dos pobres, as transferências contribuem para reduzir a desigualdade e aumentar um pouco a renda média da população. Usando uma medida de desigualdade e a renda média como controles o pesquisador torna praticamente impossível captar o efeito das transferências sobre a pobreza<sup>7</sup>. O exercício 4.39 apresenta dados numéricos artificiais que ilustram a questão.

O problema dos “maus controles” é discutido em Angrist e Pischke (2009, p. 64-68). Eles assinalam que nem sempre mais controle é melhor, que variáveis medidas antes que a variável explanatória de interesse tenha sido determinada são geralmente bons controles e que é necessário verificar se alguma variável de controle é, ela própria, determinada pela variável explanatória de interesse.

#### **4.15. Transformação das variáveis para obter a matriz de correlações simples**

Consideremos o modelo

$$y_j = \sum_i \beta_i x_{ij} + u_j - \bar{u}$$

---

<sup>7</sup> Como ocorre frequentemente, isso pode parecer óbvio depois de assinalado. Mas em dois artigos publicados na Revista Brasileira de Economia há erro de especificação semelhante, incluindo o índice de Gini e o PIB per capita de cada Unidade da Federação em modelos destinados a captar o efeito de transferências do governo federal sobre a pobreza: Marinho e Araujo (2010) e Marinho et al. (2011).

ou

$$z_j = \sum_i \gamma_i v_{ij} + \varepsilon_j + \omega \quad (4.73)$$

onde

$$z_j = \frac{y_j}{\sqrt{\sum y_j^2}}, \quad v_{ij} = \frac{x_{ij}}{\sqrt{\sum x_{ij}^2}}, \quad i = 1, 2, \dots, k \quad (4.74)$$

$$\gamma_i = \beta_i \frac{\sqrt{\sum x_{ij}^2}}{\sqrt{\sum y_j^2}}, \quad (4.75)$$

$$\varepsilon_j = \frac{u_j}{\sqrt{\sum y_j^2}} \quad \text{e} \quad \omega = \frac{\bar{u}}{\sqrt{\sum y_j^2}}$$

Indicando por  $\mathbf{V}$  a matriz das variáveis explanatórias e por  $\mathbf{z}$  o vetor dos valores da variável dependentes, temos

$$\mathbf{V}'\mathbf{V} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1k} \\ r_{12} & 1 & \dots & r_{2k} \\ \vdots & \vdots & & \vdots \\ r_{1k} & r_{2k} & \dots & 1 \end{bmatrix} \quad \text{e} \quad \mathbf{V}'\mathbf{z} = \begin{bmatrix} r_{Y1} \\ r_{Y2} \\ \vdots \\ r_{Yk} \end{bmatrix}$$

pois  $\sum_i v_{ij}^2 = 1$ ,  $\sum_i v_{ij} v_{hj} = r_{ih}$  com  $i \neq h$ , e  $\sum_i v_{ij} y_j = r_{Yi}$

As estimativas de mínimos quadrados dos  $\gamma_i$  são dadas por

$$\mathbf{c} = (\mathbf{V}'\mathbf{V})^{-1} \mathbf{V}'\mathbf{z}$$

Geralmente, os programas de computador para ajuste de regressões múltiplas fazem, no início, as transformações (4.74). Note que os elementos das matrizes  $\mathbf{V}'\mathbf{V}$  e  $\mathbf{V}'\mathbf{z}$ , que passam a ser utilizadas em lugar de  $\mathbf{X}'\mathbf{X}$  e  $\mathbf{X}'\mathbf{y}$ , variam apenas de -1 a +1. Isso contribui para diminuir os efeitos dos erros de arredondamento. Obtidas as estimativas de  $\gamma_i$ , as estimativas dos parâmetros  $\beta_i$  são, de acordo com (4.75), dadas por

$$b_i = c_i \frac{\sqrt{\sum y_j^2}}{\sqrt{\sum x_{ij}^2}}$$

#### 4.16. Regressões que se tornam lineares por anamorfose

Vários modelos estatísticos podem ser facilmente transformados em modelos de regressão linear múltipla.

Assim, a regressão quadrática

$$Y_j = \alpha + \beta_1 X_j + \beta_2 X_j^2 + u_j$$

pode ser encarada como uma regressão linear múltipla com duas variáveis explanatórias, fazendo  $X_j = X_{1j}$  e  $X_j^2 = X_{2j}$ .

De maneira análoga, qualquer regressão polinomial pode ser ajustada como uma regressão linear múltipla.

Em pesquisas econômicas, é freqüentemente utilizado o modelo

$$Y_j = \alpha X_{1j}^{\beta_1} X_{1j}^{\beta_2} \dots X_{kj}^{\beta_k} \varepsilon_j$$

Aplicando logaritmos, obtemos

$$\log Y_j = \log \alpha + \sum_i \beta_i \log X_{ij} + \log \varepsilon_j,$$

que é um modelo de regressão linear múltipla nos logaritmos das variáveis. Neste caso, desde que  $u_j = \log \varepsilon_j$  obedeça às pressuposições vistas na seção 4.1, as estimativas de mínimos quadrados têm as propriedades estatísticas desejáveis.

#### 4.17. Ortogonalidade e multicolinearidade na matriz $\mathbf{X}$

Vejamos o que ocorre quando todas as colunas da matriz  $\mathbf{X}$  são ortogonais entre si, isto é,

$$\sum_i x_{ij} x_{hj} = 0 \text{ para } i \neq h$$

A matriz  $\mathbf{X}'\mathbf{X}$  é, então, uma matriz diagonal.

Tendo em vista o modelo

$$y_j = \sum_i \beta_i x_{ij} + u_j + \bar{u}$$

temos, neste caso, que

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \frac{1}{\sum x_{1j}^2} & 0 & 0 \\ 0 & \frac{1}{\sum x_{2j}^2} & 0 \\ 0 & 0 & \frac{1}{\sum x_{kj}^2} \end{bmatrix}$$

e

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} \frac{\sum x_{1j}y_j}{\sum x_{1j}^2} \\ \frac{\sum x_{2j}y_j}{\sum x_{2j}^2} \\ \frac{\sum x_{kj}y_j}{\sum x_{kj}^2} \end{bmatrix}$$

Portanto, as estimativas dos parâmetros da regressão múltipla coincidem com as estimativas dos coeficientes das regressões lineares simples de  $Y_j$  contra cada uma das variáveis explanatórias e a soma de quadrados de regressão, dada por

$$\text{S.Q.Regr.} = \sum_i b_i \sum_j x_{ij} y_j,$$

é, neste caso, igual à soma das somas de quadrados de regressão das regressões lineares simples de  $Y_j$  contra cada uma das variáveis explanatórias. O coeficiente de determinação múltipla é, portanto, igual à soma dos coeficientes de determinação das regressões lineares simples mencionadas.

Como vimos, a ortogonalidade entre as colunas da matriz  $\mathbf{X}$  facilita bastante a análise.

Vejamos, agora, o que ocorre quando há multicolinearidade perfeita, isto é, quando existem, na matriz  $\mathbf{X}$ , colunas linearmente dependentes.

Consideremos, inicialmente, o caso de uma regressão linear múltipla com duas variáveis independentes perfeitamente correlacionadas entre si

$$Y_j = \beta_1 x_{1j} + \beta_2 x_{2j} + u_j - \bar{u}$$

com

$$r_{12}^2 = \frac{(\sum x_{ij}x_{2j})^2}{\sum x_{1j}^2 \sum x_{2j}^2} = 1$$

Neste caso, temos

$$|\mathbf{X}'\mathbf{X}| = \sum x_{1j}^2 \sum x_{2j}^2 - (\sum x_{1j}x_{2j})^2 = 0,$$

isto é, o determinante da matriz  $\mathbf{X}'\mathbf{X}$  é igual a zero. Não é possível, então, inverter a matriz  $\mathbf{X}'\mathbf{X}$  e, consequentemente, é impossível obter as estimativas de  $\beta_1$  e  $\beta_2$ .

Geometricamente, o que ocorre é que os pontos  $(x_{1j}, x_{2j}, y_j)$  estão todos sobre um plano  $\psi$ , perpendicular ao plano definido pelos eixos de  $x_{1j}$  e de  $x_{2j}$ . O método dos mínimos quadrados permite determinar apenas uma reta no plano  $\psi$ ; qualquer que seja o plano que contenha essa reta, a soma dos quadrados dos desvios assume o mesmo valor; portanto, existe indeterminação.

É importante compreender que no caso de uma regressão com mais de duas variáveis explanatórias pode existir multicolinearidade perfeita, mesmo que nenhum dos coeficientes de determinação simples seja igual a um (ver exercício 4.12).

Freqüentemente, a matriz  $\mathbf{X}$  apresenta multicolinearidade elevada, embora não perfeita. As principais consequências desse fato são as seguintes:

- 1) As variâncias e covariâncias das estimativas dos parâmetros serão muito elevadas, isto é, as estimativas obtidas podem ter erros muito grandes e esses erros podem estar altamente correlacionados entre si. A baixa precisão das estimativas torna difícil, ou até mesmo impossível, distinguir as influências das diversas variáveis explanatórias;
- 2) Um pesquisador pode ser levado a eliminar variáveis da análise porque os coeficientes não se mostraram estatisticamente diferentes de zero; essas variáveis podem, na realidade, ser importantes e a amostra disponível é que não permite detectar sua influência;
- 3) As estimativas dos coeficientes variam muito de amostra para amostra. A adição de algumas observações à amostra pode alterar muito o valor da estimativa obtida.

Para mostrar como a multicolinearidade afeta a precisão das estimativas consideremos, novamente, uma regressão múltipla com apenas duas variáveis explanatórias:

$$y_j = \sum_i \beta_i x_{ij} + u_j + \bar{u}$$

Fazendo

$$z_j = \frac{y_j}{\sqrt{\sum y_j^2}}, \quad v_{1j} = \frac{x_{1j}}{\sqrt{\sum x_{1j}^2}}, \quad v_{2j} = \frac{x_{2j}}{\sqrt{\sum x_{2j}^2}},$$

$$\gamma_1 = \beta_1 \frac{\sqrt{\sum x_{1j}^2}}{\sqrt{\sum y_j^2}}, \quad \gamma_2 = \beta_2 \frac{\sqrt{\sum x_{2j}^2}}{\sqrt{\sum y_j^2}},$$

$$\varepsilon_j = \frac{u_j}{\sqrt{\sum y_j^2}} \quad \text{e} \quad \omega = \frac{\bar{u}}{\sqrt{\sum y_j^2}},$$

obtemos

$$z_j = \gamma_1 v_{1j} + \gamma_2 v_{2j} + \varepsilon_j + \omega \quad (4.76)$$

A matriz de variâncias e covariâncias das estimativas ( $c_1$  e  $c_2$ ) dos parâmetros de (4.76) é

$$\begin{bmatrix} \sum v_{1j}^2 & \sum v_{1j} v_{2j} \\ \sum v_{1j} v_{2j} & \sum v_{2j}^2 \end{bmatrix}^{-1} \sigma^2,$$

ou, de acordo com o que foi visto na seção 4.15,

$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}^{-1} \sigma^2 = \begin{bmatrix} \frac{1}{1-r_{12}^2} & \frac{-r_{12}}{1-r_{12}^2} \\ \frac{-r_{12}}{1-r_{12}^2} & \frac{1}{1-r_{12}^2} \end{bmatrix} \sigma^2,$$

Concluímos que

$$V(c_1) = V(c_2) = \frac{\sigma^2}{1-r_{12}^2} \quad (4.77)$$

e

$$\text{cov}(c_1, c_2) = \frac{-r_{12}\sigma^2}{1-r_{12}^2} \quad (4.78)$$

As expressões (4.77) e (4.78) mostram que as variâncias e o valor absoluto da covariância das estimativas dos parâmetros crescem rapidamente quando  $r_{12}^2$  se aproxima de um, isto é, quando aumenta o grau de multicolinearidade.

Se  $r_{12}$  for positivo, verifica-se, pela expressão (4.78), que a covariância das estimativas dos parâmetros é negativa.

É interessante examinar o valor do erro de estimativa para uma determinada amostra. De acordo com (4.10) temos que

$$\mathbf{c} - \boldsymbol{\gamma} = \begin{bmatrix} c_1 - \gamma_1 \\ c_2 - \gamma_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{1-r_{12}^2} & \frac{-r_{12}}{1-r_{12}^2} \\ \frac{-r_{12}}{1-r_{12}^2} & \frac{1}{1-r_{12}^2} \end{bmatrix} \begin{bmatrix} \sum v_{1j} \boldsymbol{\varepsilon}_j \\ \sum v_{2j} \boldsymbol{\varepsilon}_j \end{bmatrix} \quad (4.79)$$

Indicando por  $d_j$  os desvios da regressão linear simples de  $v_{2j}$  contra  $v_{1j}$  temos

$$v_{2j} = r_{12} v_{1j} + d_j, \quad j = 1, \dots, n \quad (4.80)$$

De acordo com (2.6),

$$\sum v_{1j} d_j = 0$$

Multiplicando cada uma das igualdades em (4.80) por  $\boldsymbol{\varepsilon}_j$  e somando, obtemos

$$\sum v_{2j} \boldsymbol{\varepsilon}_j = r_{12} \sum v_{1j} \boldsymbol{\varepsilon}_j + \sum d_j \boldsymbol{\varepsilon}_j \quad (4.81)$$

De (4.79) e (4.81) obtemos

$$c_1 - \gamma_1 = \frac{\sum v_{1j} \boldsymbol{\varepsilon}_j - r_{12}^2 \sum v_{1j} \boldsymbol{\varepsilon}_j - r_{12} \sum d_j \boldsymbol{\varepsilon}_j}{1-r_{12}^2} = \sum v_{1j} \boldsymbol{\varepsilon}_j - \frac{r_{12} \sum d_j \boldsymbol{\varepsilon}_j}{1-r_{12}^2} \quad (4.82)$$

e

$$c_2 - \gamma_2 = \frac{-r_{12} \sum v_{1j} \boldsymbol{\varepsilon}_j + r_{12} \sum v_{1j} \boldsymbol{\varepsilon}_j + \sum d_j \boldsymbol{\varepsilon}_j}{1-r_{12}^2} = \frac{\sum d_j \boldsymbol{\varepsilon}_j}{1-r_{12}^2} \quad (4.83)$$

essas expressões mostram que, se  $r_{12}$  é positivo e se aproxima de um, os erros de estimação dos parâmetros são grandes e de sinais opostos; se  $c_1$  superestima  $\gamma_1$ , então  $c_2$  subestima  $\gamma_2$ , e vice-versa.

Pode-se demonstrar que não há razão para que a multicolinearidade afete seriamente a estimativa da variância residual.<sup>8</sup> Portanto, as estimativas das variâncias das estimativas dos parâmetros são indicadores adequados da existência de multicolinearidade. O efeito de uma variável explanatória pode ser suficientemente forte, de tal maneira que o respectivo coeficiente se mostre estatisticamente diferente de zero apesar da multicolinearidade; uma multicolinearidade bastante alta impedirá, entretanto, que se detecte a influência de variáveis importantes.

#### **4.18. Teste de hipóteses no modelo linear**

Nesta seção vamos apresentar uma maneira geral de encarar os testes de hipóteses no modelo de regressão linear.

Uma hipótese sobre os parâmetros de um modelo de regressão corresponde, sempre, a uma *restrição*. O modelo que incorpora a restrição, denominado modelo restrito, será menos flexível do que o modelo original.

Dada uma amostra de dados, seja  $S_U$  a soma de quadrados residual obtida ajustando o modelo original, irrestrito (o índice do símbolo se deve à palavra inglesa para irrestrito: *unrestricted*), e seja  $S_R$  a soma de quadrados residual obtida ajustando o modelo restrito. Como o modelo restrito é menos flexível, a respectiva soma de quadrados residual tende a ser maior, isto é,

$$S_R \geq S_U \quad (4.84)$$

Se  $S_R$  for igual ou pouco maior do que  $S_U$ , indicando que o modelo restrito se ajusta aos dados quase tão bem como o modelo irrestrito, não há razão para rejeitar a hipótese. Por outro lado, se  $S_R$  for substancialmente maior que  $S_U$ , indicando que o ajustamento do modelo restrito (que incorpora a hipótese) é claramente pior, deveremos rejeitar a hipótese formulada, pois ela entra em choque com os dados.

---

<sup>8</sup> Ver Johnston (1972), p. 163.

Para saber se a diferença  $S_R - S_U$  é ou não “substancial”, o respectivo quadrado médio é comparado com o quadrado médio residual do modelo irrestrito, calculando-se

$$F = \frac{\frac{S_R - S_U}{g_R - g_U}}{\frac{S_U}{g_U}}, \quad (4.85)$$

onde  $g_U$  e  $g_R$  são os graus de liberdade associados a  $S_U$  e  $S_R$ , respectivamente. É claro que a explicação apresentada para obter a expressão (4.85) é informal, procurando apenas mostrar a lógica mais geral da sua fundamentação. Formalmente, é necessário mostrar que  $S_U / \sigma^2$  tem distribuição de Qui-quadrado com  $g_U$  graus de liberdade, e que, se a hipótese da nulidade for verdadeira,  $(S_R - S_U) / \sigma^2$  tem distribuição de Qui-quadrado com  $g_R - g_U$  graus de liberdade, independente da distribuição de  $S_U / \sigma^2$ . Pode-se, então, concluir que a variável obtida em (4.85) tem distribuição de  $F$  com  $g_R - g_U$  e  $g_U$  graus de liberdade.

Note-se que o denominador de (4.85) é o quadrado médio do resíduo do modelo original (irrestrito), que tem sido indicado por  $s^2$ . Então a expressão fica

$$F = \frac{S_R - S_U}{(g_R - g_U)s^2} \quad (4.86)$$

Se uma hipótese sobre os parâmetros de uma regressão linear é formulada como  $H_0 : \mathbf{C}\beta = \boldsymbol{\theta}$ , onde  $\mathbf{C}$  e  $\boldsymbol{\theta}$  são matrizes com elementos numéricos dados, a expressão (4.85) é equivalente à expressão (4.60), podendo-se deduzir (4.60) de (4.85).

Vejamos alguns exemplos de aplicação de (4.85). O procedimento sempre envolve três etapas: (a) obter a soma de quadrados residual do modelo irrestrito ( $S_U$ ) e os respectivos graus de liberdade ( $g_U$ ); (b) construir o modelo restrito e obter a correspondente soma de quadrados residual ( $S_R$ ) e seus graus de liberdade ( $g_R$ ); (c) aplicar (4.85) e interpretar o resultado.

Consideremos os dados da tabela 4.1 e o modelo

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + u_j$$

Vamos admitir que queremos testar a hipótese  $H_0 : \beta_1 = 0$ . Então o modelo restrito é

$$Y_j = \alpha + \beta_2 X_j + u_j$$

Na seção 4.8 já obtivemos a soma de quadrados residual para o modelo original (irrestrito), que é

$$S_U = 2,5, \text{ com } g_U = 2 \quad (4.87)$$

Na seção 4.11 (tabela 4.4) vimos que a soma de quadrados de regressão de uma regressão de  $Y$  contra  $X_2$  é igual a 90, podendo-se verificar, então, que a soma de quadrados residual do modelo restrito é

$$S_R = 26,5, \text{ com } g_R = 3 \quad (4.88)$$

Substituindo os resultados (4.87) e (4.88) em (4.85), obtemos

$$F = \frac{\frac{26,5 - 2,5}{3 - 2}}{\frac{2,5}{2}} = \frac{24}{1,25} = 19,2$$

Como não podia deixar de ser, esse é o valor de  $F$  para “contribuição de  $X_1$ ” na tabela 4.4, que é o quadrado do valor de  $t$  referente à hipótese  $H_0 : \beta_1 = 0$  obtido na seção 4.8 ( $t = 4,382$ ).

Como segundo exemplo, vamos usar (4.85) para testar a hipótese

$$H_0 : \beta_1 = 4 \text{ e } \beta_2 = 2 \quad (4.89)$$

com base nos dados da tabela 4.5 (seção 4.12). O modelo irrestrito é

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + u_j \quad (4.90)$$

Na seção 4.12 vimos que a soma de quadrados residual para esse modelo é

$$S_U = 3, \text{ com } g_U = 3 \quad (4.91)$$

O modelo restrito fica

$$Y_j = \alpha + 4X_{1j} + 2X_{2j} + u_j$$

ou

$$Y_j - 4X_{1j} - 2X_{2j} = \alpha + u_j \quad (4.92)$$

Note-se que o modelo sempre deve ser escrito de maneira que no segundo membro fiquem apenas o erro e os termos com parâmetros a serem estimados.

Fazendo

$$W_j = Y_j - 4X_{1j} - 2X_{2j}, \quad (4.93)$$

o modelo (4.92) fica

$$W_j = \alpha + u_j \quad (4.94)$$

Usando (4.93), podemos calcular os 6 valores de  $W_j$  para a amostra apresentada na tabela 4.5. A estimativa de  $\alpha$  no modelo (4.94) é, simplesmente  $\bar{W}$ , que é igual a -2,5. Os desvios dessa “regressão” são  $w_j = W_j - \bar{W}$ , pois como não há coeficientes de regressão, a soma de quadrados residual é a própria soma de quadrados total ( $\sum w_j^2 = 48,5$ , com  $6 - 1 = 5$  graus de liberdade). Portanto, para o modelo restrito (4.92) temos

$$S_R = 48,5, \text{ com } g_R = 5 \quad (4.95)$$

Substituindo os resultados (4.89) e (4.95) na expressão (4.85), obtemos

$$F = \frac{\frac{48,5 - 3}{3}}{\frac{5 - 3}{3}} = \frac{45,5}{2} = 22,75 \quad (4.96)$$

Adotando um nível de significância de 5% , o valor crítico de  $F$  para 2 e 3 graus de liberdade é 9,55. O resultado é significativo, levando a rejeitar a hipótese (4.89). Note-se que, como não podia deixar de ser, o resultado obtido em (4.96) é idêntico ao obtido no final da seção 4.12, usando a expressão (4.60).

#### **4.19. Interpretação geométrica da análise de regressão linear de acordo com o método de mínimos quadrados**

Do estudo da álgebra vetorial sabemos que o vetor

$$\mathbf{x} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} 11 \\ 2 \\ 10 \end{bmatrix} \quad (4.97)$$

pode ser representado, graficamente, por uma seta que vai da origem do sistema de eixos ao ponto (11, 2, 10), num espaço tridimensional, como mostra a figura 4.2.

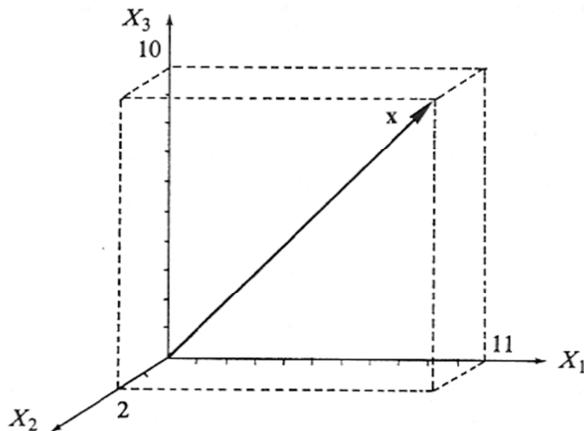


Figura 4.2. Representação gráfica de um vetor no espaço tridimensional.

Genericamente, dizemos que

$$\mathbf{x} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \quad \text{e} \quad \mathbf{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

são vetores no espaço  $n$ -dimensional.

O produto escalar ou produto interno de dois vetores de mesma dimensão,  $\mathbf{x}$  e  $\mathbf{y}$ , que indicamos  $\mathbf{x} \cdot \mathbf{y}$ , é, por definição, igual a  $\sum X_i Y_i$ . Temos

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{y} \cdot \mathbf{x} = \mathbf{x}'\mathbf{y} = \mathbf{y}'\mathbf{x} = \sum_{i=1}^n X_i Y_i$$

Verifica-se facilmente que o produto escalar de vetores é distributivo em relação à soma, isto é, se  $\mathbf{x}$ ,  $\mathbf{y}$  e  $\mathbf{v}$  são três vetores de mesma dimensão,

$$\mathbf{x} \cdot (\mathbf{y} + \mathbf{v}) = \mathbf{x} \cdot \mathbf{y} + \mathbf{x} \cdot \mathbf{v}$$

O módulo ou comprimento do vetor  $\mathbf{x}$  é dado por

$$|\mathbf{x}| = \sqrt{\mathbf{x} \cdot \mathbf{x}}$$

Para vetores com duas ou três dimensões pode-se verificar, através do teorema de Pitágoras, que esta definição de comprimento de um vetor corresponde ao comprimento da seta que o representa. O comprimento do vetor dado em (4.97) é

$$|\mathbf{x}| = \sqrt{11^2 + 2^2 + 10^2} = 15$$

Por definição, dois vetores são ortogonais se seu produto escalar é igual a zero, isto é,  $\mathbf{x}$  é ortogonal a  $\mathbf{y}$  se, e somente se,  $\mathbf{x} \cdot \mathbf{y} = 0$ . Assim, por exemplo, os vetores

$$\mathbf{x} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \quad \text{e} \quad \mathbf{y} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

são ortogonais entre si, pois  $\mathbf{x} \cdot \mathbf{y} = 0$ . Esses vetores estão representados na figura 4.3, que mostra que as setas representativas de dois vetores ortogonais são perpendiculares entre si.

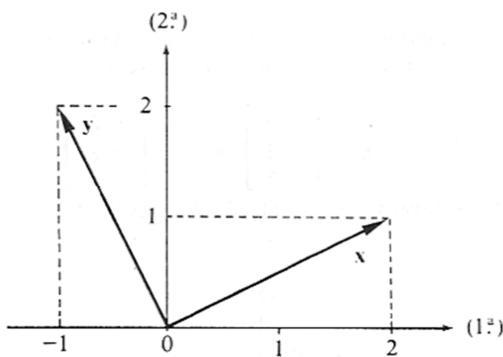


Figura 4.3. Dois vetores ortogonais

Seja  $\lambda$  um escalar. Então, se  $\mathbf{x}$  é um vetor com elementos  $X_1, X_2, \dots, X_n$ , os elementos do vetor  $\lambda\mathbf{x}$  são  $\lambda X_1, \lambda X_2, \dots, \lambda X_n$ . Os vetores  $\mathbf{x}$  e  $\lambda\mathbf{x}$  têm a mesma direção, isto é, são vetores colineares (as setas estão sobre a mesma reta-suporte). Se  $\lambda > 0$ , a orientação ou sentido dos vetores  $\mathbf{x}$  e  $\lambda\mathbf{x}$  é o mesmo e, se  $\lambda < 0$ , esses vetores têm sentidos opostos.

Temos

$$|\lambda\mathbf{x}| = \sqrt{(\lambda\mathbf{x}) \cdot (\lambda\mathbf{x})} = \sqrt{\lambda^2 \mathbf{x} \cdot \mathbf{x}} = |\lambda| |\mathbf{x}|,$$

isto é, o comprimento de  $\lambda\mathbf{x}$  é igual ao comprimento de  $\mathbf{x}$  multiplicado pelo valor absoluto de  $\lambda$ .

Para exemplificar, lembremos o vetor  $\mathbf{x}$  dado em (4.97). Pode-se verificar que os vetores  $-2\mathbf{x}$  e  $2\mathbf{x}$  são colineares com  $\mathbf{x}$  e têm comprimento igual a 30, isto é, o dobro do comprimento de  $\mathbf{x}$ .

Vamos supor agora que, com base em uma amostra com apenas 3 observações, foram obtidos os pares de valores  $(X_i, Y_i)$  apresentados na tabela 4.9. As médias das duas variáveis nessa amostra são iguais a zero, isto é,  $x_i = X_i$  e  $y_i = Y_i$  (com  $i = 1, 2, 3$ ).

TABELA 4.9. Valores de  $X_i$  e  $Y_i$  para uma amostra com 3 observações.

$X_i = x_i$	$Y_i = y_i$
0	-2
-2	-2
2	4

Seja  $\mathbf{x}$  o vetor cujos elementos são os valores da variável centrada  $x_i$  e seja  $\mathbf{y}$  o vetor cujos elementos são os valores da variável centrada  $y_i$ , isto é,

$$\mathbf{x} = \begin{bmatrix} 0 \\ -2 \\ 2 \end{bmatrix} \quad \text{e} \quad \mathbf{y} = \begin{bmatrix} -2 \\ -2 \\ 4 \end{bmatrix} \quad (4.98)$$

A figura 4.4 mostra a representação desses vetores no espaço tridimensional.

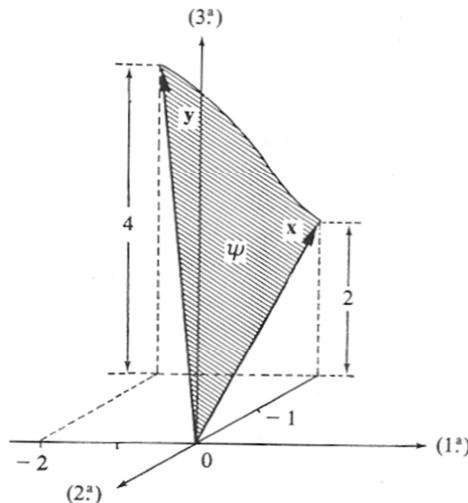


Figura 4.4. O plano ( $\psi$ ) dos vetores  $\mathbf{x}$  e  $\mathbf{y}$ .

É importante notar a diferença entre esse tipo de representação gráfica e aquele das figuras 2.2 e 3.1. Nas figuras 2.2 e 3.1 cada eixo corresponde a uma variável, ao passo que na figura 4.4 cada eixo corresponde a uma observação. Lá utilizamos o espaço das variáveis e aqui estamos considerando o espaço das observações.

Temos  $\mathbf{x} \cdot \mathbf{x} = 8$  e  $\mathbf{y} \cdot \mathbf{y} = 24$ . Portanto o comprimento do vetor  $\mathbf{x}$  é  $|\mathbf{x}| = \sqrt{8} = 2\sqrt{2}$  e o comprimento do vetor  $\mathbf{y}$  é  $|\mathbf{y}| = \sqrt{24} = 2\sqrt{6}$ .

Na figura 4.4, seja  $\psi$  o plano dos vetores  $\mathbf{x}$  e  $\mathbf{y}$ . O plano  $\psi$  é um subespaço bidimensional do espaço tridimensional. Qualquer que seja o número de observações da amostra, isto é, qualquer que seja a dimensão ( $n$ ) de  $\mathbf{x}$  e  $\mathbf{y}$ , desde que esses vetores não

sejam colineares, eles definem um plano (um subespaço bidimensional) no espaço  $n$ -dimensional. Na figura 4.5 os vetores  $\mathbf{x}$  e  $\mathbf{y}$  estão representados nesse subespaço bidimensional.

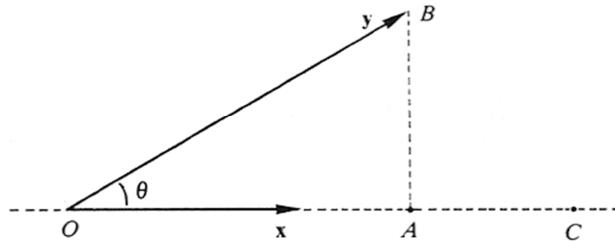


Figura 4.5. A projeção vertical de  $\mathbf{y}$  sobre a reta-suporte de  $\mathbf{x}$  e o ângulo entre  $\mathbf{x}$  e  $\mathbf{y}$ .

Na figura 4.5 o ângulo entre os vetores  $\mathbf{x}$  e  $\mathbf{y}$  foi denominado  $\theta$ . Foi obtida a projeção vertical  $\overrightarrow{OA} = \mathbf{y}^*$  do vetor  $\mathbf{y}$  sobre a reta-suporte do vetor  $\mathbf{x}$ . Como  $\mathbf{y}^*$  e  $\mathbf{x}$  são colineares, existe um escalar  $\lambda$  tal que

$$\overrightarrow{OA} = \mathbf{y}^* = \lambda \mathbf{x} \quad (4.99)$$

Temos

$$\mathbf{y}^* + \overrightarrow{AB} = \mathbf{y}$$

ou

$$\overrightarrow{AB} = \mathbf{y} - \mathbf{y}^*$$

Uma vez que  $\mathbf{x}$  e  $\overrightarrow{AB} = \mathbf{y} - \mathbf{y}^*$  são vetores ortogonais entre si, temos

$$\mathbf{x} \cdot (\mathbf{y} - \mathbf{y}^*) = 0$$

Então

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x} \cdot \mathbf{y}^* \quad (4.100)$$

Substituindo (4.99) em (4.100), obtemos

$$\mathbf{x} \cdot \mathbf{y} = \lambda(\mathbf{x} \cdot \mathbf{x})$$

ou

$$\lambda = \frac{\mathbf{x} \cdot \mathbf{y}}{\mathbf{x} \cdot \mathbf{x}} \quad (4.101)$$

Na figura 4.5, como  $OAB$  é um triângulo retângulo com hipotenusa igual a  $|\mathbf{y}|$ , temos

$$\cos \theta = \frac{|\mathbf{y}^*|}{|\mathbf{y}|} = \frac{\sqrt{\mathbf{y}^* \cdot \mathbf{y}^*}}{\sqrt{\mathbf{y} \cdot \mathbf{y}}}$$

Lembrando (4.99), segue-se que

$$\cos \theta = \frac{\lambda \sqrt{\mathbf{x} \cdot \mathbf{x}}}{\sqrt{\mathbf{y} \cdot \mathbf{y}}} \quad (4.102)$$

Substituindo (4.101) em (4.102) obtemos

$$\cos \theta = \frac{\mathbf{x} \cdot \mathbf{y}}{\sqrt{(\mathbf{x} \cdot \mathbf{x})(\mathbf{y} \cdot \mathbf{y})}} \quad (4.103)$$

Se os elementos de  $\mathbf{x}$  e de  $\mathbf{y}$  são, respectivamente  $x_i$  e  $y_i$  (com  $i = 1, 2, \dots, n$ ), isto é, são os valores das variáveis centradas obtidos de uma amostra com  $n$  observações, verifica-se que

$$\cos \theta = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} = r, \quad (4.104)$$

onde  $r$  é o coeficiente de correlação entre  $X$  e  $Y$  na amostra.

Para a amostra dada na tabela 4.9, que corresponde aos vetores  $\mathbf{x}$  e  $\mathbf{y}$  definidos em (4.98) e representados nas figuras 4.4 e 4.5, temos

$$\cos \theta = r = \frac{12}{\sqrt{8 \cdot 24}} = \frac{\sqrt{3}}{2}$$

Donde

$$\theta = 30^\circ$$

Vamos agora considerar a análise de regressão de  $Y$  contra  $X$  de acordo com o modelo

$$Y_i = \beta X_i + u_i$$

ou

$$\mathbf{y} = \beta \mathbf{x} + \mathbf{u},$$

onde

$$\mathbf{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \text{e} \quad \mathbf{x} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

Para dar um exemplo numérico, vamos considerar os valores de  $X_i$  e  $Y_i$  dados na tabela 4.9. Então  $\mathbf{x}$  e  $\mathbf{y}$  são os vetores tridimensionais definidos em (4.98) e representados nas figuras 4.4 e 4.5. Devemos ressaltar que o raciocínio apresentado a seguir não depende da dimensão dos vetores  $\mathbf{x}$  e  $\mathbf{y}$ , pois estaremos considerando apenas o plano (subespaço bidimensional) definido por esses vetores (admitindo que  $\mathbf{x}$  e  $\mathbf{y}$  não sejam colineares).

Se  $b$  é a estimativa de  $\beta$ , o vetor dos desvios é

$$\mathbf{e} = \mathbf{y} - b\mathbf{x}$$

De acordo com o método de mínimos quadrados, devemos determinar o valor de  $b$  que minimiza a soma dos quadrados dos desvios, dada por  $\mathbf{e}'\mathbf{e}$  ou  $\mathbf{e} \cdot \mathbf{e}$ . Mas  $\mathbf{e} \cdot \mathbf{e} =$ , também, o quadrado do comprimento do vetor  $\mathbf{e} = \mathbf{y} - b\mathbf{x}$ . Devemos, portanto, determinar o valor de  $b$  que minimiza o comprimento do vetor  $\mathbf{e} = \mathbf{y} - b\mathbf{x}$ .

Uma vez que  $b$  é um escalar,  $b\mathbf{x}$  é um vetor colinear com  $\mathbf{x}$ , como os vetores  $\overrightarrow{OA}$  e  $\overrightarrow{OC}$  na figura 4.5. Se fizermos  $b\mathbf{x} = \overrightarrow{OA}$ , teremos  $\mathbf{e} = \overrightarrow{AB}$  e, se fizermos  $b\mathbf{x} = \overrightarrow{OC}$ , teremos  $\mathbf{e} = \overrightarrow{CB}$ . Por outro lado, sabemos que a menor distância de um ponto a uma reta é dada pela perpendicular baixada do ponto sobre a reta. Uma vez que, na figura 4.5,  $\overrightarrow{OA} = \lambda\mathbf{x}$  é, por construção, a projeção vertical de  $\mathbf{y}$  sobre a reta-suporte do vetor  $\mathbf{x}$ , concluímos que, para minimizar o comprimento de  $\mathbf{e} = \mathbf{y} - b\mathbf{x}$ , devemos fazer  $b\mathbf{x} = \overrightarrow{OA} = \lambda\mathbf{x}$ . Portanto, o estimador de mínimos quadrados de  $\beta$  é

$$b = \lambda$$

Lembrando (4.101), obtemos

$$b = \frac{\mathbf{x} \cdot \mathbf{y}}{\mathbf{x} \cdot \mathbf{x}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$$

Esta relação é um caso particular de (4.7).

Para os valores de  $X_i$  e  $Y_i$  da tabela 4.9, obtemos  $\mathbf{x} \cdot \mathbf{x} = 8$ ,  $\mathbf{x} \cdot \mathbf{y} = 12$  e  $b = 12/8 = 1,5$ .

Como  $\overrightarrow{OA} = b\mathbf{x}$  é o vetor com os valores estimados de  $Y_i$ , passaremos a indicá-lo por  $\hat{\mathbf{y}}$ , de acordo com a notação utilizada em seções anteriores.

O vetor dos desvios da regressão é

$$\overrightarrow{AB} = \mathbf{e} = \mathbf{y} - \overrightarrow{OA} = \mathbf{y} - \hat{\mathbf{y}}$$

No triângulo retângulo  $OAB$  da figura 4.5, o teorema de Pitágoras estabelece que

$$\overline{OB}^2 = \overline{OA}^2 + \overline{AB}^2$$

ou

$$\mathbf{y} \cdot \mathbf{y} = \hat{\mathbf{y}} \cdot \hat{\mathbf{y}} + \mathbf{e} \cdot \mathbf{e}$$

Também podemos escrever

$$\mathbf{y}'\mathbf{y} = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \mathbf{e}'\mathbf{e}$$

ou

$$S.Q.\text{total} = (S.Q.\text{Regr.}) + (S.Q.\text{Res.}),$$

que é uma relação já demonstrada anteriormente, mas de outra maneira.

Uma vez que  $\mathbf{x}$  e  $\mathbf{e}$  são vetores ortogonais entre si, temos

$$\mathbf{x} \cdot \mathbf{e} = \mathbf{x}'\mathbf{e} = 0$$

Esta relação é um caso particular de (4.8).

É fácil verificar que se  $\mathbf{x}$  e  $\mathbf{y}$  fossem colineares, teríamos  $\theta = 0$ ,  $r = \cos \theta = 1$ ,  $\mathbf{y} = \hat{\mathbf{y}} = b\mathbf{x}$ ,  $\mathbf{e} = \mathbf{0}$  e  $S.Q.\text{Res.} = \mathbf{e}'\mathbf{e} = 0$ .

Vamos considerar, a seguir, o modelo de regressão linear múltipla

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} \quad (4.105)$$

onde

$$\begin{aligned} \mathbf{y} &= \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, & \mathbf{X} &= \begin{bmatrix} X_{11} & X_{21} \\ X_{12} & X_{22} \\ \vdots & \vdots \\ X_{1n} & X_{2n} \end{bmatrix}, \\ \boldsymbol{\beta} &= \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \quad \text{e} \quad \mathbf{u} &= \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \end{aligned}$$

Se indicarmos por  $\mathbf{x}_1$  e  $\mathbf{x}_2$  os vetores constituídos pela primeira e pela segunda coluna da matriz  $\mathbf{X}$ , respectivamente, isto é, se fizermos

$$\mathbf{x}_1 = \begin{bmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1n} \end{bmatrix} \quad \text{e} \quad \mathbf{x}_2 = \begin{bmatrix} X_{21} \\ X_{22} \\ \vdots \\ X_{2n} \end{bmatrix},$$

temos

$$\mathbf{y} = \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \mathbf{u} \quad (4.106)$$

Vamos admitir que os vetores  $\mathbf{y}$ ,  $\mathbf{x}_1$  e  $\mathbf{x}_2$  têm dimensão igual ou superior a 3, não são colineares nem estão todos em um mesmo plano, isto é, vamos admitir que  $\mathbf{y}$ ,  $\mathbf{x}_1$  e  $\mathbf{x}_2$  são linearmente independentes.<sup>9</sup> Observada esta condição, qualquer que seja a dimensão ( $n \geq 3$ ) dos vetores  $\mathbf{x}_1$ ,  $\mathbf{x}_2$  e  $\mathbf{y}$ , tais vetores geram um espaço tridimensional. A figura 4.6 mostra os 3 vetores nesse espaço. Seja  $\psi$  o plano (subespaço bidimensional) gerado por  $\mathbf{x}_1$  e  $\mathbf{x}_2$ .

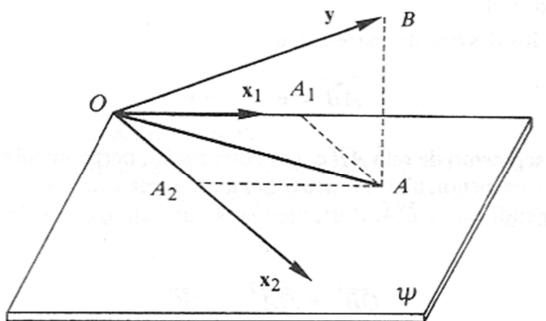


Figura 4.6. Os vetores  $\mathbf{y}$ ,  $\mathbf{x}_1$  e  $\mathbf{x}_2$ .

De acordo com o método de mínimos quadrados, devemos determinar  $\hat{\mathbf{y}} = b_1\mathbf{x}_1 + b_2\mathbf{x}_2$  de maneira a minimizar  $\mathbf{e} \cdot \mathbf{e} = (\mathbf{y} - \hat{\mathbf{y}}) \cdot (\mathbf{y} - \hat{\mathbf{y}})$ , que é o quadrado do comprimento do vetor  $\mathbf{y} - \hat{\mathbf{y}}$ . Devemos, portanto, determinar  $\hat{\mathbf{y}} = b_1\mathbf{x}_1 + b_2\mathbf{x}_2$  de maneira a minimizar o comprimento do vetor  $\mathbf{y} - \hat{\mathbf{y}}$ . Necessariamente  $\hat{\mathbf{y}} = b_1\mathbf{x}_1 + b_2\mathbf{x}_2$  é um vetor no plano  $\psi$ , porque é uma combinação linear dos vetores  $\mathbf{x}_1$  e  $\mathbf{x}_2$ .

O ponto do plano  $\psi$  que está mais próximo do ponto  $B$  (a extremidade do vetor  $\mathbf{y}$ ) é a projeção vertical de  $B$  sobre  $\psi$ , isto é, o ponto  $A$  na figura 3.6. Portanto  $\hat{\mathbf{y}} = \overline{OA}$  é a combinação linear ( $\hat{\mathbf{y}} = b_1\mathbf{x}_1 + b_2\mathbf{x}_2$ ) de  $\mathbf{x}_1$  e  $\mathbf{x}_2$  que minimiza o comprimento do vetor  $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ .

Obtido o vetor  $\hat{\mathbf{y}} = \overline{OA}$ , podemos determinar os vetores  $b_1\mathbf{x}_1$  e  $b_2\mathbf{x}_2$ . Para isso, devemos traçar por  $A$  retas paralelas a  $\mathbf{x}_1$  e  $\mathbf{x}_2$ , construindo o paralelogramo  $OA_1AA_2$ . Temos

---

<sup>9</sup> Dado um conjunto de vetores, dizemos que eles são linearmente independentes se nenhum deles pode ser expresso como uma combinação linear dos demais. Dados dois vetores linearmente independentes (não colineares),  $\mathbf{x}_1$  e  $\mathbf{x}_2$ , uma combinação linear  $\lambda_1\mathbf{x}_1 + \lambda_2\mathbf{x}_2$  é sempre um vetor no plano definido por  $\mathbf{x}_1$  e  $\mathbf{x}_2$  e, reciprocamente, todo vetor (ponto) neste plano é uma combinação linear de  $\mathbf{x}_1$  e  $\mathbf{x}_2$ ; dizemos, então, que os vetores  $\mathbf{x}_1$  e  $\mathbf{x}_2$  geram o plano (subespaço bidimensional). Analogamente, 3 vetores linearmente independentes geram um subespaço tridimensional.

$$\hat{\mathbf{y}} = \overrightarrow{OA} = \overrightarrow{OA_1} + \overrightarrow{OA_2},$$

$$\overrightarrow{OA_1} = b_1 \mathbf{x}_1$$

e

$$\overrightarrow{OA_2} = b_2 \mathbf{x}_2$$

O valor absoluto de  $b_1$  é o cociente da divisão do comprimento de  $\overrightarrow{OA_1}$  pelo comprimento de  $\mathbf{x}_1$  e o valor absoluto de  $b_2$  é o cociente da divisão do comprimento de  $\overrightarrow{OA_2}$  pelo comprimento de  $\mathbf{x}_2$ . No caso da figura 4.6 verifica-se que  $b_1 > 1$  e  $0 < b_2 < 1$ .

O vetor dos desvios da regressão é

$$\overrightarrow{AB} = \mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$$

Como o segmento de reta  $\overrightarrow{AB}$  é, por construção, perpendicular ao plano  $\psi$ , tal segmento é perpendicular ou ortogonal a toda reta desse plano. Em particular,  $\overrightarrow{AB}$  é perpendicular a  $\overrightarrow{OA}$ . Portanto  $OAB$  é um triângulo retângulo e, consequentemente

$$\overrightarrow{OB}^2 = \overrightarrow{OA}^2 + \overrightarrow{AB}^2$$

ou

$$\mathbf{y} \cdot \mathbf{y} = \hat{\mathbf{y}} \cdot \hat{\mathbf{y}} + \mathbf{e} \cdot \mathbf{e}$$

ou, ainda,

$$S.Q.Total = (S.Q.Regr.) + (S.Q.Res.)$$

Como  $\mathbf{e} = \overrightarrow{AB}$  é ortogonal a  $\mathbf{x}_1$  e  $\mathbf{x}_2$ , temos que

$$\mathbf{x}_1 \cdot \mathbf{e} = \mathbf{x}'_1 \mathbf{e} = 0$$

e

$$\mathbf{x}_2 \cdot \mathbf{e} = \mathbf{x}'_2 \mathbf{e} = 0$$

De acordo com essas relações, podemos escrever

$$\mathbf{X}' \mathbf{e} = \mathbf{0}, \quad (4.107)$$

que é a relação (4.8).

Temos

$$\mathbf{e} = \mathbf{y} - \mathbf{Xb}, \quad (4.108)$$

onde

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

Substituindo (4.108) em (4.107), obtemos

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{0}$$

ou

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y},$$

que é o sistema de equações normais.

É interessante examinar o que ocorre quando os vetores  $\mathbf{x}_1$  e  $\mathbf{x}_2$  são colineares. Nesse caso esses vetores geram apenas um subespaço unidimensional, que é a reta que os contém, e não um plano. Seja  $\hat{\mathbf{y}} = \overrightarrow{OA}$  a projeção vertical de  $\mathbf{y}$  sobre essa reta, como mostra a figura 4.7.

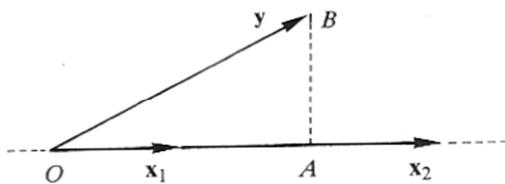


Figura 4.7. Vectors  $\mathbf{x}_1$  and  $\mathbf{x}_2$  collinear.

Existem infinitas combinações lineares de  $\mathbf{x}_1$  e  $\mathbf{x}_2$  que produzem o vetor  $\hat{\mathbf{y}} = \overrightarrow{OA}$ . No caso da figura 4.7, admitindo que  $|\mathbf{x}_1| = 1$ ,  $|\mathbf{x}_2| = 3$  e  $|\overrightarrow{OA}| = 2$ , temos, por exemplo,

$$\overrightarrow{OA} = 2\mathbf{x}_1 + 0\mathbf{x}_2,$$

$$\overrightarrow{OA} = 0\mathbf{x}_1 + \frac{2}{3}\mathbf{x}_2$$

ou

$$\overrightarrow{OA} = -\mathbf{x}_1 + \mathbf{x}_2$$

Isso mostra que os valores de  $b_1$  e  $b_2$  em  $\hat{\mathbf{y}} = b_1\mathbf{x}_1 + b_2\mathbf{x}_2$  são indeterminados. Esse é o problema da multicolinearidade perfeita, já examinado, sob outro enfoque, na seção 4.17.

Vários outros problemas de análise de regressão podem ser examinados com o auxílio da interpretação geométrica apresentada nesta seção.<sup>10</sup> Como ilustração final, vamos considerar o exercício 2.9, onde se pede para comparar a análise de regressão

---

<sup>10</sup> Uma exposição didática do assunto pode ser encontrada em Wonnacott e Wonnacott (1976), parte II.

linear simples de  $Y_i$  contra  $X_i$  com a análise de regressão linear simples de  $Z_i$  contra  $X_i$ , com  $Z_i = Y_i + X_i$ . As equações estimadas de acordo com o método de mínimos quadrados são indicadas respectivamente por

$$\hat{Y}_i = a + bX_i$$

e

$$\hat{Z}_i = c + dX_i$$

Vamos demonstrar que  $c = a$  e  $d = b + 1$ .

Sejam  $\mathbf{y}$  e  $\hat{\mathbf{y}}$  os vetores com os valores de  $Y_i$  e  $\hat{Y}_i$ , respectivamente; seja  $\mathbf{x}_1$  o vetor com os valores de  $X_i$ ; sejam  $\mathbf{z}$  e  $\hat{\mathbf{z}}$  os vetores com os valores de  $Z_i$  e  $\hat{Z}_i$ , respectivamente, e seja  $\mathbf{x}_0$  um vetor cujos elementos são todos iguais a 1. Temos

$$\mathbf{z} = \mathbf{y} + \mathbf{x}_1$$

$$\hat{\mathbf{y}} = a\mathbf{x}_0 + b\mathbf{x}_1$$

e

$$\hat{\mathbf{z}} = c\mathbf{x}_0 + d\mathbf{x}_1$$

Na figura 4.8 estão representados os vetores  $\mathbf{x}_0, \mathbf{x}_1$  e  $\mathbf{y}$ . O plano gerado por  $\mathbf{x}_0$  e  $\mathbf{x}_1$  é denominado plano  $\psi$ . Indicando a extremidade de  $\mathbf{y}$  por  $B$  e a projeção vertical de  $B$  sobre  $\psi$  por  $A$ , temos que

$$\overrightarrow{OA} = \hat{\mathbf{y}}$$

e

$$\overrightarrow{AB}^2 = (\text{S.Q.Res. } \mathbf{y} \mid \mathbf{x}_0 \text{ e } \mathbf{x}_1) \quad (4.109)$$

Traçando pelo ponto  $A$  retas paralelas a  $\mathbf{x}_0$  e a  $\mathbf{x}_1$ , obtemos

$$\overrightarrow{OA_0} = a\mathbf{x}_0 \quad \text{e} \quad \overrightarrow{OA_1} = b\mathbf{x}_1,$$

de tal maneira que

$$\overrightarrow{OA} = \hat{\mathbf{y}} = a\mathbf{x}_0 + b\mathbf{x}_1$$

Uma vez que neste caso  $a$  e  $b$  são positivos, segue-se que

$$a = \frac{\overrightarrow{OA_0}}{|\mathbf{x}_0|} \quad (4.110)$$

e

$$b = \frac{\overline{OA_1}}{|\mathbf{x}_1|} \quad (4.111)$$

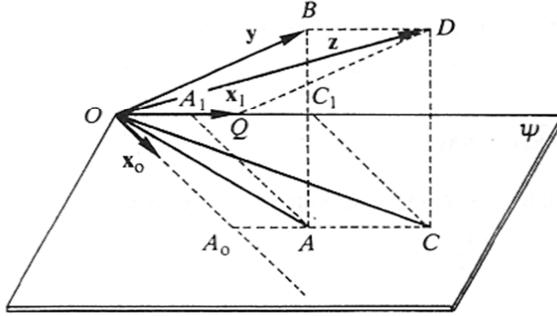


Figura 4.8. Regressão linear simples de  $Y$  contra  $X$  e de  $Z = Y + X$  contra  $X$ .

Para obter o vetor  $\mathbf{z} = \mathbf{x}_1 + \mathbf{y}$  construímos o paralelogramo  $OB\bar{D}Q$ . Então  $\overline{BD}$  é um segmento de reta paralelo ao plano  $\psi$  com comprimento igual a  $|\mathbf{x}_1|$  e  $\overrightarrow{OD} = \mathbf{z}$ .

Sendo  $C$  a projeção vertical de  $D$  sobre  $\psi$ , temos

$$\overrightarrow{OC} = \hat{\mathbf{z}}$$

e

$$\overline{CD}^2 = (\text{S.Q.Res. } \mathbf{z} | \mathbf{x}_0 \text{ e } \mathbf{x}_1) \quad (4.112)$$

Como  $\overline{BD}$  é paralelo a  $\mathbf{x}_1$  e ao plano  $\psi$  e  $\overline{AC}$  é a projeção de  $\overline{BD}$  sobre  $\psi$ , temos que  $\overline{AC} = \overline{BD} = |\mathbf{x}_1|$ , com  $\overline{AC}$  paralelo a  $\mathbf{x}_1$  e, portanto, colinear com o segmento de reta  $\overline{A_0A}$ .

Além disso, temos  $\overline{CD} = \overline{AB}$ . Lembrando (4.109) e (4.112), concluímos que

$$(\text{S.Q.Res. } \mathbf{z} | \mathbf{x}_0 \text{ e } \mathbf{x}_1) = (\text{S.Q.Res. } \mathbf{y} | \mathbf{x}_0 \text{ e } \mathbf{x}_1).$$

Traçamos  $\overline{C_1C}$  paralelo a  $\mathbf{x}_0$ . Então

$$\overline{A_1C_1} = \overline{AC} = |\mathbf{x}_1| \quad (4.113)$$

e

$$\overrightarrow{OC} = \hat{\mathbf{z}} = \overrightarrow{OA_0} + \overrightarrow{OC_1},$$

com

$$\overrightarrow{OA_0} = c\mathbf{x}_0$$

e

$$\overrightarrow{OC_1} = d\mathbf{x}_1$$

Uma vez que neste caso  $c$  e  $d$  são positivos, segue-se que

$$c = \frac{\overline{OA_0}}{|\mathbf{x}_0|} \quad (4.114)$$

e

$$d = \frac{\overline{OC_1}}{|\mathbf{x}_1|} \quad (4.115)$$

Comparando (4.114) com (4.110) concluímos que

$$c = a$$

Examinando a figura 4.8 e lembrando (4.113), temos que

$$\overline{OC_1} = \overline{OA_1} + \overline{A_1C_1} = \overline{OA_1} + |\mathbf{x}_1|$$

Substituindo esse resultado em (4.115) obtemos

$$d = \frac{\overline{OA_1}}{|\mathbf{x}_1|} + 1$$

Lembrando (4.111) concluímos que

$$d = b + 1, \text{ c.q.d.}$$

## Exercícios

4.1. Mostre que as fórmulas para regressão linear simples, deduzidas no Capítulo 2, são casos particulares das expressões gerais:

- a)  $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
- b)  $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$ , que é a matriz de variâncias e covariâncias das estimativas dos parâmetros.
- c) S.Q.Res. =  $\mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y}$
- d)  $V(\hat{Y}_h) = \mathbf{x}_h(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_h\sigma^2$

4.2. São dados os valores de  $X_1$ ,  $X_2$  e  $Y$  da tabela a seguir:

$X_1$	$X_2$	$Y$
0	0	-1
0	2	3
0	4	5
0	6	5
2	0	4
2	2	10
2	4	12
2	6	10

Admite-se que as variáveis estão relacionadas de acordo com o modelo  $Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + u_j$ , onde os  $u_j$  são variáveis aleatórias independentes, homocedásticas, com média zero e distribuição normal.

- Determine as estimativas dos parâmetros da regressão linear múltipla de  $Y$  em relação a  $X_1$  e  $X_2$ .
- Faça a análise de variância da regressão.
- Determine a contribuição de cada variável para a soma de quadrados de regressão. Verifique que o respectivo teste  $F$  é igual ao quadrado do teste  $t$  correspondente à hipótese de que seja nulo o valor do coeficiente de regressão da variável em questão.

#### 4.3. Idem, para

$X_1$	$X_2$	$X_3$	$Y$
-1	-1	0	5
-1	0	0	7
-1	1	0	3
1	-1	0	7
1	0	0	9
1	1	0	5
0	0	-1	3
0	0	0	8
0	0	1	7

#### 4.4. Idem para

$X_1$	$X_2$	$Y$
0	1	1
1	0	1
1	1	4
1	2	5
1	3	4

Além disso, determine:

d) o valor dos coeficientes de correlação parcial  $r_{Y1.2}$  e  $r_{Y2.1}$ .

e) o intervalo de 95% de confiança para cada um dos três parâmetros do modelo linear

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + u_j,$$

admitindo que os  $u_j$  são variáveis independentes com distribuição normal de média zero e variância  $\sigma^2$ .

f) a estimativa de  $Y$  para  $X_1 = 1$  e  $X_2 = 2$ , e o intervalo de 95% de confiança para  $E(Y | X_1 = 1 \text{ e } X_2 = 2)$ .

g) Idem, para  $X_1 = 2$  e  $X_2 = 4$  (uma extração).

#### 4.5. São dados os seguintes valores, obtidos de uma amostra aleatória com 10 observações:

$X_1$	$X_2$	$X_3$	$Y$
-1	-1	0	5
-1	0	0	7
-1	1	0	3
1	-1	0	7
1	0	0	9
1	1	0	5
0	0	-1	3
0	0	0	8
0	0	1	7
1	1	0	6

Admite-se que as variáveis estão relacionadas de acordo com o modelo

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + u_j,$$

onde os  $u_j$  são erros independentes com distribuição normal de média zero e variância  $\sigma^2$ .

- a) Determine as estimativas dos parâmetros.
- b) Faça a análise de variância da regressão.
- c) Teste a hipótese  $H_0 : \alpha = 0$  contra  $H_A : \alpha \neq 0$ , ao nível de significância de 1%.
- d) Teste a hipótese  $H_0 : \beta_3 = 0$  contra  $H_A : \beta_3 > 0$ , ao nível de significância de 5%.
- e) Determine os valores dos coeficientes de determinação parcial  $r_{Y1.23}^2$ ,  $r_{Y2.13}^2$  e  $r_{Y3.12}^2$ .
- f) Determine a estimativa de  $Y_j$  para  $X_1 = 0$ ,  $X_2 = 1$  e  $X_3 = 1$ , e o respectivo intervalo de 95% de confiança.

- 4.6. No caso do exemplo apresentado na seção 4.13, teste, através do valor de  $t$ , a hipótese  $H_0 : \beta_3 = 0$  contra a hipótese  $H_A : \beta_3 \neq 0$ , considerando um nível de significância de 5%. Obtenha, também, o valor do teste  $F$  para a “contribuição de  $X_3$ ”, verificando que este valor é igual ao quadrado do valor de  $t$  obtido anteriormente. Determine, para o mesmo exemplo, a estimativa de  $Y_j$  para  $X_1 = X_2 = X_3 = 1$  e o respectivo intervalo de 95% de confiança.

- 4.7. Considerando o modelo de regressão múltipla

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + u_j - \bar{u},$$

com todas as variáveis centradas, demonstre que

$$R^2 = \frac{r_{Y1}^2 - 2r_{12}r_{Y1}r_{Y2} + r_{Y2}^2}{1 - r_{12}^2},$$

onde  $R^2$  é o coeficiente de determinação múltipla.

- 4.8. Considerando o resultado do problema anterior, mostre que, se  $r_{Y1} = r_{Y2} = r_{12} = r \neq 1$ , obtemos

$$R^2 = \frac{2r^2}{1+r}$$

Discuta o caso em que  $r_{Y1} = r_{Y2} = r_{12} = r = 1$ .

- 4.9. Considerando o modelo do problema 4.7 mostre que, se  $r_{Y1,2} = 1$ , temos  $r_{Y2,1}^2 = 1$  e  $R^2 = 1$ .

- 4.10. São dados os valores de  $Y_j, X_{1j}$  e  $X_{2j}$  da tabela a seguir:

$Y_j$	$X_{1j}$	$X_{2j}$
3	0	0
0	0	3
6	1	1
9	3	0

Considerando o modelo  $Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + u_j$ , onde os  $u_j$  são erros independentes com distribuição normal de média zero e variância  $\sigma^2$ ,

- a) Determine a equação de regressão de  $Y$  em relação a  $X_1$  e  $X_2$ .
- b) Calcule o valor do teste  $F$  para a regressão (para testar a hipótese  $H_0 : \beta_1 = \beta_2 = 0$ ).
- c) Calcule os valores de  $R^2$  e  $r_{Y1,2}^2$ .
- d) Teste a hipótese  $H_0 : \beta_1 = \beta_2$  contra a hipótese  $H_A : \beta_1 > \beta_2$ , considerando o nível de significância de 5%.
- e) Determine a estimativa de  $Y$  para  $X_1 = X_2 = 2$  e o respectivo intervalo de 90% de confiança.
- f) Calcule os valores de  $\hat{Y}$  para as observações da amostra e, a seguir, o quadrado do coeficiente de correlação entre  $Y$  e  $\hat{Y}$  (note que esse valor é igual a  $R^2$ ).

- 4.11. São dados os valores de  $X_1$ ,  $X_2$  e  $Y$  da tabela a seguir:

$Y$	$X_1$	$X_2$
-4	0	3
5	1	1
4	2	2
11	3	0

Admite-se que as variáveis estão relacionadas de acordo com o modelo  $Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + u_j$  ( $j = 1, \dots, 4$ ), onde os  $u_j$  são erros independentes, homocedásticas, com média zero e distribuição normal.

- Determine a equação de regressão de  $Y$  em relação a  $X_1$  e  $X_2$ .
- Teste, ao nível de significância de 5%, a hipótese  $H_0 : \beta_1 = \beta_2 = 0$ .
- Calcule o valor do coeficiente de determinação múltipla.
- Calcule o valor de  $r_{Y1.2}^2$ .
- Teste, ao nível de significância de 5%, a hipótese  $H_0 : \beta_2 = 4$ .
- Teste, ao nível de significância de 5%, a hipótese  $H_0 : \beta_1 = -2$  e  $\beta_2 = 4$ .
- Delimite a região de 95% de confiança para  $\beta_1$  e  $\beta_2$ .
- Teste, ao nível de significância de 5%, a hipótese  $H_0 : \beta_1 = \beta_2$ , contra a hipótese  $H_A : \beta_1 > \beta_2$ .
- Calcule a estimativa de  $Y$  para  $X_1 = X_2 = 2,5$  e determine o respectivo intervalo de 90% de confiança.
- Calcule os valores de  $\hat{Y}$  para as observações da amostra e, a seguir, o quadrado do coeficiente de correlação entre  $Y$  e  $\hat{Y}$  (note que esse valor é igual a  $R^2$ ).

4.12. Com a finalidade de ajustar o modelo  $Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \beta_4 X_{4j} + u_j$  foi obtida uma amostra de 8 observações. Os valores das variáveis explanatórias constam da tabela a seguir:

$X_1$	$X_2$	$X_3$	$X_4$
-1	-1	-1	-3
-1	-1	1	-1
-1	1	-1	-1
-1	1	1	1
1	-1	-1	-1
1	-1	1	1
1	1	-1	1
1	1	1	3

Verifique que, embora o valor dos coeficientes de correlação entre pares de variáveis independentes seja sempre inferior a 0,58, existe multicolinearidade perfeita.

4.13. As variáveis explanatórias  $X_1$  e  $X_2$  assumem os valores  $-3, -1, +1$  e  $+3$ . Temos 16 observações de  $Y = f(X_1, X_2)$  correspondendo a todas as combinações possíveis para  $X_1$  e  $X_2$ . Decidiu-se ajustar, a esses dados, uma regressão múltipla com um termo constante e todos os possíveis termos do 1º, 2º, 3º e 4º graus em  $X_1$  e  $X_2$ . Não foi possível obter as estimativas dos parâmetros  $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ . Por quê? Decidiu-se, então, ignorar a variável  $X_2$  e se tornou a ajustar, aos mesmos dados, um polinômio do quarto grau em  $X_1$ . Novamente não foi possível obter as estimativas dos parâmetros. Por quê? (Extraído de DRAPER e SMITH, 1966, p. 160).

4.14. Sejam 3 regressões onde o número de observações e os valores das variáveis independentes são os mesmos; numa das regressões a variável dependente é  $Y_{1j}$ , na outra é  $Y_{2j}$ , e na terceira é  $Y_{3j} = Y_{1j} + Y_{2j}$ . Sendo  $\mathbf{b}_1$ ,  $\mathbf{b}_2$  e  $\mathbf{b}_3$  os vetores das estimativas dos parâmetros da primeira, da segunda e da terceira regressão, respectivamente, prove que  $\mathbf{b}_3 = \mathbf{b}_1 + \mathbf{b}_2$ .

4.15. São dados os pares de valores  $X, Y$  da tabela a seguir:

$X$	$Y$
-2	0,9
-1	6,4
0	8,4
1	10,4
2	8,9

Admite-se que as variáveis estão relacionadas de acordo com o modelo  $Y_i = \alpha + \beta X_i + \gamma X_i^2 + u_i$ , onde os  $u_i$  são os erros independentes com distribuição normal de média zero e variância  $\sigma^2$ .

- Determine as estimativas dos parâmetros.
- Faça a análise de variância da regressão, testando, ao nível de significância de 5%, a hipótese  $H_0 : \beta = \gamma = 0$ .
- Calcule o valor do coeficiente de determinação da regressão ajustada.

- d) Teste, ao nível de significância de 5%, a hipótese  $H_0 : \gamma = 0$  contra  $H_A : \gamma < 0$ .
- e) Determine o valor da contribuição do termo quadrático para a soma de quadrados de regressão. Verifique se o respectivo teste  $F$  é significativo ao nível de 5% (note que o valor de  $F$  obtido é igual ao quadrado do valor de  $t$  calculado no item anterior).
- 4.16. Com base em uma amostra com 34 observações foi estimada a equação de regressão de  $Y$  contra  $X_1$ ,  $X_2$  e  $X_3$ , considerando o modelo
- $$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + u_j$$
- O coeficiente de determinação parcial de  $Y$  e  $X_1$ , dados  $X_2$  e  $X_3$ , é igual a 0,25. Teste, ao nível de significância de 1%, a hipótese de que  $\beta_1 = 0$ .
- 4.17. Demonstre que o coeficiente de determinação múltipla de uma regressão linear múltipla qualquer (definido como o cociente da divisão da S.Q.Regr. pela S.Q.Total) é igual ao quadrado do coeficiente de correlação entre  $Y_j$  e  $\hat{Y}_j$  (Para facilitar a demonstração considere que a regressão tenha sido ajustada com todas as variáveis centradas).
- 4.18. Admite-se que as variáveis  $X_1$ ,  $X_2$  e  $Y$  estão relacionadas conforme o modelo

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + u_j$$

onde  $u_j$  são erros independentes com  $E(u_j) = 0$  e  $E(u_j^2) = \sigma^2$

Para uma amostra de 9 observações, obtivemos:

$$\begin{array}{llll} \bar{X}_1 = \bar{X}_2 = 0 & \sum x_1 y = -8 & \sum x_1^2 = 8 & \sum x_2 y = -8 \\ \sum x_2^2 = 4 & \sum y^2 = 42 & \sum x_1 x_2 = 0 & \bar{Y} = 6 \end{array}$$

- a) Calcule as estimativas de mínimos quadrados para  $\alpha$ ,  $\beta_1$  e  $\beta_2$  ( $a$ ,  $b_1$  e  $b_2$ , respectivamente).
- b) Ache a estimativa da matriz de variâncias e covariâncias de  $a$ ,  $b_1$  e  $b_2$ .
- c) Determine o intervalo de previsão, ao nível de confiança de 95%, para uma nova observação de  $Y$  com  $X_1 = X_2 = 1$ .
- d) Teste a hipótese  $H_0 : \beta_2 = \beta_1$  contra a hipótese alternativa  $H_A : \beta_2 < \beta_1$ .

- e) Teste a hipótese  $H_0 : \alpha = 10$  e  $\beta_1 - 2\beta_2 = 0$ .

Adote, nos testes, o nível de significância de 5%.

4.19. Um ensaio de adubação forneceu os seguintes resultados:

$X = \text{dose de adubo por hectare}$	$Y = \text{produção por hectare}$
0	6; 8
1	16; 18
2	18; 20
3	12; 14

Pode-se verificar que  $\sum Y_i = 112$ ,  $\bar{Y} = 14$  e  $\sum(Y_i - \bar{Y})^2 = 176$

- Admitindo que a função de produção seja uma parábola do 2º grau, determine as estimativas dos parâmetros dessa função de acordo com o método de mínimos quadrados.
- Faça a análise de variância da regressão e calcule o coeficiente de determinação.
- Sabendo que a relação entre o preço da dose de adubo e o preço do produto é igual a 2, determine a quantidade economicamente ótima de adubo a ser aplicada.
- Verifique se o coeficiente do termo quadrático é estatisticamente diferente de zero. Pressupõe-se que a lei dos rendimentos marginais decrescentes seja válida.
- Teste a hipótese de que a produção máxima é obtida aplicando-se 2 doses de adubo por hectare.

Considere, nos testes estatísticos, um nível de significância de 1%.

4.20. Foi estabelecido o seguinte modelo de função de produção:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \quad (i = 1, \dots, n)$$

onde o índice  $i$  indica a empresa agropecuária,  $Y_i$  é o logaritmo do valor da produção,  $X_{1i}$  é o logaritmo da mão-de-obra utilizada e  $X_{2i}$  é o logaritmo do capital utilizado.

A amostra tem 23 observações e são conhecidas as seguintes matrizes (relativas ao modelo com todas as variáveis centradas, inclusive a dependente)

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 12 & 8 \\ 8 & 12 \end{bmatrix} \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} 10 \\ 8 \end{bmatrix} \quad \mathbf{y}'\mathbf{y} = 10$$

- a) Determine as estimativas dos coeficientes de regressão e dos respectivos desvios padrões
- b) Calcule o valor de  $R^2$ .
- c) Teste se os rendimentos à escala são constantes, isto é, teste a hipótese  $H_0 : \beta_1 + \beta_2 = 1$ , considerando um nível de significância de 5%.

4.21. São dados os valores das variáveis  $X_{1j}$ ,  $X_{2j}$  e  $Y_j$  para uma amostra com 6 observações:

$X_{1j}$	$X_{2j}$	$Y_j$
0	0	-0,5
1	1	3,5
1	2	7,0
2	1	7,0
2	2	7,5
3	3	11,5

- a) Determine a equação de regressão linear de  $Y$  em relação a  $X_1$  e  $X_2$ .
- b) Calcule o valor do coeficiente de determinação da regressão.
- c) Teste, ao nível de significância de 5%, a hipótese  $H_0 : \beta_1 = \beta_2 + 4$  contra a hipótese alternativa  $H_A : \beta_1 < \beta_2 + 4$ .
- d) Teste, ao nível de significância de 1%, a hipótese  $H_0 : \beta_1 = 3$  e  $\beta_2 = 5$ .
- e) Calcule o coeficiente de determinação parcial entre  $Y_j$  e  $X_{2j}(r_{Y2,1}^2)$  e verifique se é estatisticamente diferente de zero, considerando um nível de significância de 1%.
- f) Calcule a estimativa de  $Y$  para  $X_1 = 0,5$  e  $X_2 = 2,5$  e determine o respectivo intervalo de 90% de confiança.
- g) O problema da multicolinearidade é mais sério no exercício anterior ou neste? Justifique.

4.22. Numa análise da demanda de certo produto, baseada em dados anuais para um período de 17 anos, foram obtidos os seguintes valores, referentes às variáveis  $Y$  (logaritmo da quantidade consumida *per capita*),  $X_1$  (logaritmo da renda *per capita*) e  $X_2$  (logaritmo do preço do produto).

Médias	Estimativas dos desvios padrões	Coeficiente de correlação
$\bar{Y} = 4,5$	$s_y = 3\sqrt{5}$	$r_{Y1} = \frac{5}{\sqrt{30}}$
$\bar{X}_1 = 1$	$s_1 = \sqrt{6}$	$r_{Y2} = -\frac{4}{\sqrt{30}}$
$\bar{X}_2 = 1$	$s_2 = \sqrt{6}$	$r_{12} = -0,5$

- a) Calcule as estimativas dos coeficientes de regressão relativos a  $X_1$  e a  $X_2$  na regressão linear de  $Y$  em relação a  $X_1$  e  $X_2$ .
- b) Teste, ao nível de significância de 1%, a hipótese  $H_0 : \beta_2 = 0$  contra a hipótese alternativa  $H_0 : \beta_2 < 0$ .

4.23. Admite-se que as variáveis  $Y$ ,  $X$  e  $T$  estão relacionadas de acordo com o modelo

$$Y_j = \alpha + \beta X_j + \gamma T_j + u_j,$$

onde  $T$  é o tempo, medido em anos, e  $u_j$  são erros aleatórios com média zero e variância constante. Sejam  $a$ ,  $b$  e  $c$  as estimativas de mínimos quadrados de  $\alpha$ ,  $\beta$  e  $\gamma$ , obtidas a partir de uma amostra com  $n$  observações anuais de  $X$  e  $Y$ . Mostre que o valor de  $b$ , obtido através das fórmulas usuais de regressão múltipla, é igual à estimativa do coeficiente de regressão de  $Y$  em relação a  $v$ , sendo  $v$  os desvios da regressão linear simples de  $X$  em relação a  $T$ .

4.24. Numa análise da demanda de certo produto, baseada numa série temporal de dados, foram estimados os parâmetros da regressão de  $Y$  (logaritmo da quantidade consumida *per capita*) em relação a  $X$  (logaritmo do preço do produto) e  $T$  (tempo, em anos).

Explique o significado econômico da inclusão do tempo como variável explanatória.

4.25. Considere a matriz

$$\mathbf{A} = \mathbf{I} - \frac{1}{n} \mathbf{u}' ,$$

onde  $\mathbf{I}$  é uma matriz unitária de ordem  $n$  e  $\mathbf{t}$  é um vetor-coluna com  $n$  elementos, todos iguais a 1.

- a) Mostre que  $\mathbf{A}$  é uma matriz quadrada, simétrica e idempotente.
- b) Demonstre que  $\text{tr}(\mathbf{A}) = n - 1$ .

4.26. Considere a matriz  $\mathbf{A}$ , definida no exercício anterior, e o modelo de regressão linear múltipla apresentado no início do capítulo 4, isto é,

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

onde  $\mathbf{y}$  é um vetor-coluna com as  $n$  observações da variável dependente,  $\mathbf{X}$  é uma matriz  $n \times p$  de valores fixos,  $\boldsymbol{\beta}$  é um vetor-coluna com  $p$  parâmetros (incluindo  $\alpha$ , o termo constante da equação de regressão) e  $\mathbf{u}$  é o vetor-coluna dos erros, com  $E(\mathbf{u}) = \mathbf{0}$  e  $E(\mathbf{u}\mathbf{u}') = \mathbf{I}\sigma^2$ .

Demonstre que:

- a) S.Q.Total =  $\mathbf{y}'\mathbf{A}\mathbf{y}$
- b) S.Q.Total =  $\mathbf{u}'\mathbf{A}\mathbf{u} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\mathbf{X}\boldsymbol{\beta} + 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\mathbf{u}$
- c) S.Q.Total =  $\text{tr}(\mathbf{A}\mathbf{u}\mathbf{u}') + \boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\mathbf{X}\boldsymbol{\beta} + 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\mathbf{u}$
- d)  $E(\text{S.Q.Total}) = (n - 1)\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\mathbf{X}\boldsymbol{\beta}$

Finalmente, considerando (4.22) e lembrando que

$$\text{S.Q.Regr.} = (\text{S.Q.Total}) - (\text{S.Q.Res.}),$$

Deduza que

$$E(\text{S.Q.Regr.}) = \boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\mathbf{X}\boldsymbol{\beta} + (p - 1)\sigma^2$$

Verifique, ainda, que (2.32) é um caso particular desse resultado e que o valor de  $\boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\mathbf{X}\boldsymbol{\beta}$  não depende de  $\alpha$ .

4.27. Considere, novamente, o modelo de regressão linear múltipla do exercício anterior e lembre as propriedades das matrizes  $\mathbf{H}$  e  $\mathbf{M}$  definidas na seção 4.5.

- a) Se  $\mathbf{x}_j$  é a  $j$ -ésima coluna de  $\mathbf{X}'$ , verifique que o  $j$ -ésimo elemento da diagonal de  $\mathbf{H}$  é

$$h_j = \mathbf{x}'_j (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_j$$

- b) Utilizando (4.20), deduza que  $E(\mathbf{e}) = \mathbf{0}$  e que a matriz de variâncias e covariâncias do vetor de desvios é

$$V(\mathbf{e}) = E(\mathbf{e}\mathbf{e}') = \mathbf{M}\sigma^2$$

- c) Demonstre que a variância do desvio da  $j$ -ésima observação é

$$V(e_j) = (1 - h_j)\sigma^2$$

e que a respectiva estimativa é

$$\hat{V}(e_j) = (1 - h_j)s^2$$

- d) Para o exemplo numérico apresentado na seção 4.8, verifique que  $h_1 = 0,6$  e  $\hat{V}(e_1) = 0,5$ . Muitas vezes se considera que o quadrado médio do resíduo ( $s^2$ ), que é a estimativa não-tendenciosa da variância do erro, é, também, a estimativa da variância dos desvios. Note como neste caso (com  $n$  pequeno) a estimativa correta é muito diferente.

#### 4.28. Considere o modelo

$$Y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t \quad (t = 1, \dots, n)$$

com

$$E(u_t) = 0 \text{ para todo } t$$

$$E(u_t^2) = \sigma^2 \text{ para todo } t$$

$$E(u_t u_s) = 0 \text{ se } t \neq s$$

$$\text{e } \sum_t x_{2t} = \sum_t x_{3t} = 0$$

Seja  $b_3$  uma estimativa não-tendenciosa de  $\beta_3$ , obtida de dados independentes. Sabendo que a variância dessa estimativa é  $V(b_3) = v^2$ , demonstre que a variância da estimativa do coeficiente de regressão de  $(Y_t - b_3 x_{3t})$  contra  $x_{2t}$ , de acordo com o método de mínimos quadrados, é

$$\frac{\sigma^2}{\sum x_{2t}^2} + v^2 b_{32}^2,$$

onde  $b_{32}$  é a estimativa do coeficiente angular da regressão linear simples de  $x_{3t}$  em relação a  $x_{2t}$ .

- 4.29. Foi proposto o seguinte modelo para analisar o crescimento de uma espécie vegetal (baseado em AIGNER, 1971, p. 107-108):

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + u_j,$$

onde  $Y$  é a altura da planta, em cm,  $X_1$  é o tempo decorrido, em semanas,  $X_2 = X_1^2$  e os  $u_j$  são erros aleatórios independentes com distribuição normal de média zero e variância  $\sigma^2$ .

A partir de uma amostra com  $n = 13$  observações semanais, com a variável  $X_1$  assumindo os valores  $-6, -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5$  e  $6$ , foram obtidos os seguintes resultados:

$$\bar{Y} = 32 \quad \sum y_j^2 = 3640 \quad r_{Y1} = 0,90$$

$$\bar{X}_1 = 0 \quad \sum x_{1j}^2 = 182 \quad r_{Y2} = -0,20$$

$$\bar{X}_2 = 14 \quad \sum x_{2j}^2 = 2002$$

- a) Calcule as estimativas de  $\alpha$ ,  $\beta_1$  e  $\beta_2$  de acordo com o método de mínimos quadrados.
- b) Calcule  $R^2$ .
- c) Teste, ao nível de significância de 5%, a hipótese  $H_0 : \beta_1 = \beta_2 = 0$ .
- d) Teste, ao nível de significância de 5%, a hipótese  $H_0 : \beta_2 = 0$ .

- 4.30. Considerando o exemplo numérico do capítulo 2, teste, ao nível de significância de 5%, a hipótese de que  $\alpha = 5$  e  $\beta = 0$ .

- 4.31. Considere o modelo  $Y_j = \beta_1 X_{1j} + \beta_2 X_{2j} + u_j$ , onde os  $u_j$  são erros aleatórios independentes com distribuição normal de média zero e variância  $\sigma^2$ .

- a) Utilizando os dados da tabela a seguir, teste, ao nível de significância de 5%, a hipótese  $H_0 : \beta_1 = \beta_2$ .

$X_1$	$X_2$	$Y$
0	-1	0
0	-1	0
0	1	1
0	1	3
1	0	5
1	0	3

- b) Verifique que sob  $H_0$ , isto é, com  $\beta_1 = \beta_2 = \beta$ , o modelo fica  $Y_j = \beta(X_{1j} + X_{2j}) + u_j$ . Seja  $S_1$ , com  $n - 2$  graus de liberdade, a soma de quadrados residual para o modelo  $Y_j = \beta_1 X_{1j} + \beta_2 X_{2j} + u_j$  e seja  $S_2$ , com  $n - 1$  graus de liberdade, a soma de quadrados residual para o modelo  $Y_j = \beta(X_{1j} + X_{2j}) + u_j$ . Calcule o valor de

$$\phi = \frac{S_2 - S_1}{S_1 / (n - 2)}$$

e compare com o valor de  $F$  relativo ao teste de hipótese da parte (a).

- 4.32. Com base em uma amostra aleatória com  $n$  observações foi estimada, pelo método de mínimos quadrados, a equação de regressão linear múltipla de  $Y$  contra  $X_1$  e  $X_2, \dots, X_k$  e foram obtidas, através da equação ajustada, as estimativas de  $Y$  para as  $n$  observações da amostra, isto é, foram calculados os valores de  $\hat{Y}_j$  para  $j = 1, 2, \dots, n$ .

A seguir o método de mínimos quadrados é novamente utilizado para estimar os parâmetros da equação de regressão linear simples de  $Y_j$  contra  $\hat{Y}_j$ , isto é, para obter as estimativas ( $c$  e  $d$ ) dos parâmetros  $\gamma$  e  $\delta$  do modelo

$$Y_j = \gamma + \delta \hat{Y}_j + \varepsilon_j$$

Deduza quais são os valores de  $c$  e  $d$ . Faça um gráfico ilustrando suas conclusões.

- 4.33. São dados os valores de  $Y$  (renda),  $X_1$  (escolaridade) e  $X_2$  (idade) observados em uma amostra aleatória de 5 indivíduos:

$Y_j$	$X_{1j}$	$X_{2j}$
10	6	28
20	12	40
17	10	32
12	8	36
11	9	34

Fonte: S.R. SEARLE. *Linear Models*. Wiley, 1971.

Verifica-se que

$$\sum Y_j = 70, \quad \bar{Y} = 14, \quad \sum X_{1j} = 45, \quad \bar{X}_1 = 9,$$

$$\sum X_{2j} = 170, \quad \bar{X}_2 = 34, \quad \sum x_{1j}^2 = 20,$$

$$\sum x_{2j}^2 = 80 \quad \text{e} \quad \sum x_{1j}x_{2j} = 32$$

- a) Qual é a estimativa de mínimos quadrados do coeficiente angular da regressão linear simples de  $Y$  contra  $X_1$ ?
- b) Qual é a estimativa de mínimos quadrados do coeficiente angular da regressão linear simples de  $Y$  contra  $X_2$ ?
- c) Determine as estimativas de mínimos quadrados ( $\alpha$ ,  $\beta_1$  e  $\beta_2$ ) dos parâmetros da equação de regressão linear múltipla de  $Y$  em relação a  $X_1$  e  $X_2$ .
- d) Considerando o modelo  $Y_j = \gamma + \delta X_{1j} + u_j$ , onde os  $u_j$  são erros independentes entre si com distribuição normal de média zero, variância  $\sigma^2$  e independente de  $X_{1j}$ , teste a hipótese  $H_0 : \delta = 0$  contra a hipótese  $H_A : \delta > 0$ , adotando um nível de significância de 5%.
- e) Considerando o modelo  $Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + u_j$ , onde os  $u_j$  são erros independentes entre si com distribuição normal de média zero, variância  $\sigma^2$  e independente de  $X_{1j}$ , teste a hipótese  $H_0 : \beta_1 = 0$  contra a hipótese  $H_A : \beta_1 > 0$ , adotando um nível de significância de 5%.
- f) Tanto no item (d) como no item (e) o teste se refere à influência de  $X_1$  sobre  $Y$ . Os resultados dos testes são diferentes? Por quê?
- g) É possível, com base na amostra dada, fazer o teste para “falta de ajustamento” relativa ao modelo do item (d)? Explique.

- h) É possível, com base na amostra dada, estimar os parâmetros do modelo a seguir? Justifique a resposta.

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{1j}^2 + \beta_4 X_{2j}^2 + u_j$$

4.34. Um pesquisador admite que a variável  $Y$  é linearmente dependente das variáveis  $X_1$  e  $X_2$ . Obtém, então, uma amostra aleatória com  $n$  observações de  $Y$ ,  $X_1$  e  $X_2$ . Os coeficientes de correlação simples são  $r_{Y1} = 0$ ,  $r_{Y2} = 0,8$  e  $r_{12} = 0,6$ . O pesquisador, que está procurando “explicar” as variações de  $Y$ , observando que o coeficiente de correlação simples entre  $Y$  e  $X_1$  é igual a zero ( $r_{Y1} = 0$ ), propõe que  $X_1$  deixe de ser considerada como variável explanatória. É verdade que, como sugere esta proposta, o coeficiente de determinação da regressão linear múltipla de  $Y$  em função de  $X_1$  e  $X_2$  é igual ao coeficiente de determinação da regressão linear de  $Y$  contra apenas  $X_2$ ? Calcule esses coeficientes de determinação.

4.35. Seja  $Q$  a quantidade demandada de certo produto, em determinado mercado, por unidade de tempo, seja  $P$  o respectivo preço e seja  $W$  a renda *per capita* dos consumidores. Admita que essas variáveis estão relacionadas de acordo com o modelo

$$Q_j = \gamma P_j^{\beta_1} W_j^{\beta_2} \varepsilon_j$$

e que  $u_j = \log \varepsilon_j$  são erros aleatórios independentes com média zero e variância constante. Dispomos das seguintes observações:

$P_j$	$W_j$	$Q_j$
1	1	10
1	10	100
10	1	10
10	10	100
10	100	100
100	100	100
100	10	10
100	1000	100
1000	1000	10
1000	100	10

- a) Que anamorfoses devem ser feitas para obtermos um modelo de regressão linear múltipla?

- b) Obtenha as estimativas de  $\gamma$ ,  $\beta_1$  e  $\beta_2$  (utilize logaritmos decimais).
- c) Faça a análise de variância da regressão e calcule o valor do coeficiente de determinação múltipla. Teste, ao nível de significância de 5%, a hipótese  $H_0 : \beta_1 = \beta_2 = 0$ .
- d) Teste, ao nível de significância de 5%, a hipótese de que o coeficiente de elasticidade-preço da demanda do produto é igual a -1, contra a hipótese alternativa de que o coeficiente de elasticidade-preço da demanda é, em valor absoluto, menor do que 1.
- e) Teste, ao nível de significância de 5%, a hipótese de que o coeficiente de elasticidade-renda da demanda do produto é igual a 0,85, contra a hipótese alternativa de que esse coeficiente é diferente de 0,85.
- f) Calcule a estimativa da variação percentual na quantidade, demandada quando o preço aumenta de 1% e a renda *per capita* cresce 3%. Obtenha, também, a estimativa do desvio padrão dessa estimativa.

Sugestão: (1º) Mostre que as estimativas de  $\beta_1$  e  $\beta_2$  não são alteradas se mudarmos a base dos logaritmos utilizados; portanto, essas estimativas seriam as mesmas se fossem utilizados logaritmos neperianos; (2º) Lembre que, com logaritmos neperianos temos  $\frac{\Delta Q}{Q} = \Delta(\log Q)$ , desde que as variações consideradas sejam pequenas.

4.36. Considere o modelo estatístico de uma função de produção tipo Cobb-Douglas

$$Z_j = \theta W_{1j}^{\beta_1} W_{2j}^{\beta_2} \varepsilon_j \quad (1)$$

onde  $Z$  é a produção de determinada cultura,  $W_1$  é a área cultivada, em hectares, e  $W_2$  representa o montante de despesas com mão-de-obra e demais insumos (sementes, fertilizantes, defensivos, etc.). Admita que se dispõe dos valores de  $Z$ ,  $W_1$  e  $W_2$  para uma amostra de  $n$  empresas agrícolas.

De (1), aplicando logaritmos e fazendo  $\log Z_j = Y_j$ ,  $\log W_{1j} = X_{1j}$ ,  $\log W_{2j} = X_{2j}$ ,  $\log \theta = \alpha$  e  $\log \varepsilon_j = u_j$ , obtemos o modelo de regressão linear múltipla

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + u_j \quad (1')$$

Sejam  $a$ ,  $b_1$  e  $b_2$  as estimativas de mínimos quadrados de  $\alpha$ ,  $\beta_1$  e  $\beta_2$ , respectivamente.

Vamos admitir, agora, que o pesquisador prefere analisar como a produtividade da cultura varia em função do montante de despesas com mão-de-obra e insumos por hectare, isto é, como  $Z/W_1$  varia em função de  $W_2/W_1$ . A área cultivada ( $W_1$ ) é mantida como variável independente no modelo para verificar se a escala de produção afeta a produtividade.

Então o modelo fica

$$\frac{Z}{W_{1j}} = \phi W_{1j}^{\delta_1} \left( \frac{W_{2j}}{W_{1j}} \right)^{\delta_2} \varepsilon_j \quad (2)$$

O modelo (2) pode ser obtido de (1) dividindo os dois membros da equação por  $W_{1j}$ . Verifica-se que  $\phi = \theta$ ,  $\delta_1 = \beta_1 + \beta_2 - 1$  e  $\delta_2 = \beta_2$ .

De (2), aplicando logaritmos e fazendo  $\log \phi = \gamma$ , obtemos o modelo de regressão linear múltipla

$$(Y_j - X_{1j}) = \gamma + \delta_1 X_{1j} + \delta_2 (X_{2j} - X_{1j}) + u_j \quad (2')$$

Sejam  $c$ ,  $d_1$  e  $d_2$  as estimativas de mínimos quadrados de  $\gamma$ ,  $\delta_1$  e  $\delta_2$ , respectivamente

- a) Demonstre que  $d_2 = b_2$ ,  $d_1 = b_1 + b_2 - 1$  e  $c = a$ .
- b) Prove que o quadrado médio do resíduo relativo à regressão (1') é igual ao quadrado médio do resíduo relativo à regressão (2').
- c) No modelo (1'), para testar a hipótese de que a função de produção é linearmente homogênea (rendimentos constantes à escala), ou seja, para testar  $H_0 : \beta_1 + \beta_2 = 1$ , devemos calcular

$$t_1 = \frac{b_1 + b_2 - 1}{\sqrt{\hat{V}(b_1 + b_2)}}$$

No modelo (2'), para testar  $H_0 : \delta_1 = 0$  devemos calcular

$$t_2 = \frac{d_1}{\sqrt{\hat{V}(d_1)}}$$

Mostre que essas hipóteses são equivalentes e demonstre que  $t_1 = t_2$ .

4.37. Admite-se que as variáveis  $X_1$ ,  $X_2$  e  $Y$  estão relacionadas de acordo com o modelo

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + u_j,$$

onde os  $u_j$  são erros aleatórios, com as pressuposições usuais. É dada a seguinte amostra:

$X_1$	$X_2$	$Y$
6	24	2
10	16	24
13	10	50
15	6	64

É possível estimar os parâmetros do modelo com base nesta amostra? O número de observações é insuficiente? Explique.

4.38. Admite-se que  $Y$  é uma função linear de  $X_1$ ,  $X_2$  e  $X_3$ , de acordo com o modelo

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + u_j,$$

em que o erro  $u_j$  tem as propriedades usuais. A partir de uma amostra com 14 observações, foram obtidas as seguintes matrizes, considerando todas as variáveis na sua forma original:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 14 & 56 & 84 & 0 \\ 56 & 248 & 368 & 0 \\ 84 & 368 & 600 & 0 \\ 0 & 0 & 0 & 14 \end{bmatrix}, \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} 168 \\ 656 \\ 1040 \\ 28 \end{bmatrix} \quad \text{e} \quad \mathbf{y}'\mathbf{y} = 2196$$

Com todas as variáveis centradas, as matrizes são

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 24 & 32 & 0 \\ 32 & 96 & 0 \\ 0 & 0 & 14 \end{bmatrix}, \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} -16 \\ 32 \\ 28 \end{bmatrix} \quad \text{e} \quad \mathbf{y}'\mathbf{y} = 180$$

- a) Qual o valor da média de  $X_2$  na amostra?
- b) Determine as estimativas de mínimos quadrados de  $\beta_1$ ,  $\beta_2$  e  $\beta_3$ .
- c) Teste, ao nível de significância de 1%, a hipótese  $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ .

- d) Teste, ao nível de significância de 1%, a hipótese  $H_0 : \beta_3 = 0$  .  
 e) Teste, ao nível de significância de 5%, a hipótese  $H_0 : \beta_1 = \beta_2 = 0$  .

4.39. A Tabela a seguir mostra os valores de  $X_1$ ,  $X_2$ ,  $X_3$  e  $Y$  em uma amostra com 8 observações:

$X_1$	$X_2$	$X_3$	$Y$
13	32	5	116
3	4	45	8
8	18	23	63
13	34	5	118
8	18	27	61
3	2	45	6
8	18	23	69
8	18	27	55

Fazendo uma regressão múltipla de  $Y$  contra  $X_1$ ,  $X_2$  e  $X_3$ , verifica-se que os respectivos coeficientes de regressão são 0, 1 e -2. Usando  $X_2$  e  $X_3$  como controles, não se constata *nenhum* efeito de  $X_1$  sobre  $Y$ .

Verifique, também, que a regressão simples de  $Y$  contra  $X_1$  fornece um coeficiente de regressão igual a 11 e que o teste  $t$  da respectiva hipótese de nulidade é igual a 26,42, fortemente significativo.

Esses dados foram construídos com  $X_2$  e  $X_3$  como funções de  $X_1$  (com erros independentes) e  $Y$  como função de  $X_2$  e  $X_3$ . Não há efeito direto de  $X_1$  sobre  $Y$ , mas há efeitos importantes associados a  $X_2$  e  $X_3$ .

### Respostas

- 4.2. a)  $\hat{Y} = 3X_1 + X_2$   
 b)  $F = 14$   
 c) (Contribuição de  $X_1$ ) = 72;  $F = 18$   
 (Contribuição de  $X_2$ ) = 40;  $F = 10$

- 4.3. a)  $\hat{Y} = 6 + X_1 - X_2 + 2X_3$   
 b)  $F = 5/3$   
 c) (Contribuição de  $X_1$ ) = 6;  $F = 5/3$   
 (Contribuição de  $X_2$ ) = 4;  $F = 10/9$   
 (Contribuição de  $X_3$ ) = 8;  $F = 20/9$

- 4.4. a)  $\hat{Y} = 2X_1 + X_2$   
 b)  $F = 2,5$   
 c) (Contribuição de  $X_1$ ) = 40/13;  $F = 20/13$   
 (Contribuição de  $X_2$ ) = 5;  $F = 2,5$   
 d)  $r_{Y1.2} = \sqrt{\frac{10}{23}} = 0,659$  ;  $r_{Y2.1} = \frac{\sqrt{5}}{3} = 0,745$   
 e)  $-6,67 < \alpha < 6,67$  ;  $-4,94 < \beta_1 < 8,94$  ;  $-1,72 < \beta_2 < 3,72$   
 f)  $4 \pm 3,33$   
 g)  $8 \pm 10,18$

- 4.5. a)  $Y = 6 + X_1 - X_2 + 2X_3$ ,  $s^2 = 3$   
 b)  $F = 2$   
 c)  $t = 10,792$ , significativo ( $t_0 = 3,707$ )  
 d)  $t = 1,633$ , não-significativo ( $t_0 = 1,943$ )  
 e)  $r_{Y1.23}^2 = \frac{55}{202}$ ;  $r_{Y2.13}^2 = \frac{55}{262}$  e  $r_{Y3.12}^2 = \frac{4}{13}$   
 f)  $7 \pm 3,73$

- 4.6.  $t = -3,65$ , significativo  
 $6,5 \pm 1,095$

- 4.10. a)  $\hat{Y} = 3,5 + 2X_1 - X_2$   
 b)  $F = 7$   
 c)  $R^2 = \frac{14}{15} = 0,9333$  ;  $r_{Y1.2}^2 = \frac{6}{7} = 0,8571$   
 d)  $t = 3,67$ , não-significativo  
 e)  $5,5 \pm 10,47$

4.11. a)  $\hat{Y} = 5,5 + 2X_1 - 3X_2$

b)  $F = 56,5$ , não-significativo ( $F_0 = 200$ )

c)  $R^2 = \frac{113}{114} = 0,9912$

d)  $r_{Y1.2}^2 = \frac{36}{41} = 0,8780$

e)  $t = -9,39$ , não-significativo

f)  $F = 274,50$ , significativo

g) A região de confiança é delimitada pela elipse  $5q_1^2 - 8q_1q_2 + 5q_2^2 = 400$

onde  $q_1 = \beta_1 - 2$  e  $q_2 = \beta_2 + 3$

h)  $t = 10,607$ , significativo

i)  $3 \pm 9,47$  ou  $-6,47 < E(Y | X_1 = X_2 = 2,5) < 12,47$

4.15. a)  $\hat{Y} = 9 + 2X - X^2$ ,  $s^2 = 0,35$

b)  $F = 77,14$ , significativo

c)  $R^2 = 0,9872$

d)  $t = -6,325$ , significativo

e) (Contribuição de  $X^2$ ) = 14

$F = 40$ , significativo

4.16.  $F = 10$ , significativo ( $F_0 = 7,56$ )

4.18. a)  $a = 6$ ,  $b_1 = -1$  e  $b_2 = -2$

b)  $(\mathbf{X}'\mathbf{X})^{-1}s^2 = \begin{bmatrix} 1/3 & 0 & 0 \\ 0 & 3/8 & 0 \\ 0 & 0 & 3/4 \end{bmatrix}$

c)  $-2,17 < Y_h < 8,17$

d)  $t = 0,943$ , não-significativo (A região de rejeição é  $t \geq 1,943$ )

e)  $F = 25,33$ , significativo ( $F_0 = 5,14$ )

4.19. a)  $\hat{Y} = 7 + 14X - 4X^2$

b)  $F = 52,5$ ;  $R^2 = 0,9545$

c) 1,5 doses

d)  $t = -8,944$ , significativo (A região de rejeição é  $t \leq -3,365$ )

e)  $F = 11,11$ , não-significativo ( $F_0 = 16,26$ )

4.20. a)  $b_1 = 0,7$ ,  $b_2 = 0,2$ ,  $s(b_1) = s(b_2) = 0,1025$

b)  $R^2 = 0,86$

c)  $t = -1,195$ , não-significativo ( $t_0 = 2,086$ )

4.21. a)  $\hat{Y} = 2X_1 + 2X_2$

b)  $R^2 = 0,964$ .

c)  $t = -2,828$ , significativo (A região de rejeição é  $t \leq -2,353$ )

d)  $F = 41$ , significativo ( $F_0 = 30,82$ )

e)  $r_{Y2:1}^2 = \frac{80}{113} = 0,708$ ;  $F = 7,27$ , não-significativo ( $F_0 = 34,12$ )

f)  $\hat{Y} = 6$

$$2,54 < E(Y | X_1 = 0,5; X_2 = 2,5) < 9,46$$

g) No exercício anterior temos  $r_{12} = 0,67$  e neste temos  $r_{12} = 0,82$ .

Portanto, a multicolinearidade é mais forte no exercício 4.21.

4.22. a)  $b_1 = 2$  e  $b_2 = -1$

b)  $s^2 = 24/7$ ;  $t = -4,583$ , significativo (a região de rejeição é  $t \leq -2,624$ ).

4.24. O tempo, em anos, é, neste caso, uma variável *proxy* (variável representativa) para várias variáveis sócio-econômicas, como a renda *per capita* e as mudanças nos hábitos de consumo (estas, por sua vez, associadas à crescente urbanização etc.).

4.29. a)  $b_1 = 4,025$ ,  $b_2 = -0,2697$  e  $a = 35,776$ .

b)  $R^2 = 0,85$ .

c)  $F = 28,33$ , significativo ( $F_0 = 4,10$ )

d)  $t = -1,633$ , não-significativo, ( $t_0 = 2,228$ )

4.30.  $F = 18$ , significativo ( $F_0 = 4,46$ )

4.31. a)  $b_1 = 4$  e  $b_2 = 1$ ;  $s^2 = 2$

$$F = 6, \text{ não-significativo } (F_0 = 7,71)$$

b)  $\phi = F = 6$ . Na realidade, o cálculo de  $\phi$ , como foi indicado neste item do exercício, é uma maneira de obter o valor de  $F$  para testar a hipótese  $H_0 : \beta_1 = \beta_2$ .

4.32.  $d = 1$  e  $c = 0$

4.33. a) 1,75

b) 0,625

c)  $a = \frac{7}{3}$ ,  $b_1 = \frac{25}{12}$  e  $b_2 = -\frac{5}{24}$

d)  $t = 3,796$ , significativo ( $t_0 = 2,353$ )

e)  $t = 2,331$ , não-significativo ( $t_0 = 2,920$ )

f) A inclusão de  $X_2$  afeta o valor da estimativa do coeficiente de regressão, o valor da estimativa ( $s^2$ ) da variância residual e o valor do coeficiente, obtido da matriz  $(\mathbf{X}'\mathbf{X})^{-1}$ , da variância da estimativa do parâmetro. Este último efeito, bastante importante neste exemplo, se deve à correlação entre  $X_1$  e  $X_2$ .

g) Não, porque não há repetições.

h) Sim. Entretanto, como temos apenas 5 observações e o modelo tem 5 parâmetros, não é possível fazer qualquer análise estatística (não há graus de liberdade para o resíduo).

4.34.  $R^2 = 1$  e  $r_{Y2}^2 = 0,64$ .

4.35. a)  $Y = \log Q$ ,  $X_1 = \log P$  e  $X_2 = \log W$ .

b)  $c = 10\sqrt{10}$ ,  $b_1 = -0,5$  e  $b_2 = 0,5$ .

c)  $R^2 = 0,6$ ;  $F = 5,25$ , significativo ( $F_0 = 4,74$ ).

d)  $t = 3$ , significativo ( $t_0 = 1,895$ ).

e)  $t = -2,10$ , não-significativo ( $t_0 = 2,365$ ).

f)  $\Delta\hat{Y} = 0,01$  ou um crescimento de 1% em  $Q$ .

$$s(\Delta\hat{Y}) = 0,00398 \text{ ou, aproximadamente, } 0,4\% \text{ de } Q.$$

4.37. Os parâmetros não podem ser estimados com base nessa amostra pois há multicolinearidade perfeita. Verifica-se que para as 4 observações  $2X_1 + X_2 = 36$ .

4.38. a)  $\bar{X}_2 = 6$                           b)  $\mathbf{b}' = [-2 \ 1 \ 2]$

c)  $F = \frac{40}{6} = 6,67$ , significativo ( $F_0 = 6,55$ )

d)  $t = 3,06$ , não-significativo ( $t_0 = 3,169$ )

e)  $F = \frac{32}{6} = 5,33$ , significativo ( $F_0 = 4,10$ )

## 5. USO DE VARIÁVEIS BINÁRIAS

Neste capítulo vamos examinar a utilização de variáveis binárias como variáveis explanatórias em análise de regressão. Uma variável binária (também denominada variável “dummy”) é aquela que só tem dois valores distintos, geralmente zero e um. Preliminarmente, vamos lembrar os diversos níveis de medida de uma variável.

### 5.1. Níveis de medida

Podemos distinguir os seguintes níveis de medida ou escalas:

- a) Escala nominal, quando temos apenas uma classificação em categorias. Exemplos: sexo ou religião das pessoas. Neste caso, se forem usados números para indicar as diferentes categorias, eles são apenas “etiquetas” ou “nomes”.
- b) Escala ordinal, quando é válida apenas a ordem dos números. Exemplo: uma escala de *status social*.
- c) Escala intervalar. Neste caso vale a ordem e também podemos comparar, numéricamente, intervalos (diferenças) entre valores. Mas a razão entre valores não tem sentido porque a origem é arbitrária. Exemplos: temperatura medida em graus centígrados ou graus Fahrenheit e ano (data). Tem sentido dizer que o período 1982-1990 (incluindo extremos) é três vezes mais longo do que o período 1979-1981, mas não tem sentido dizer que no ano de 2002 estávamos no “dobro” de 1001.
- d) Escala razão ou cardinal, quando são válidas todas as operações com os valores. Exemplos: comprimento, peso, idade, valor monetário.

Note-se que há um “enriquecimento” progressivo do significado dos números quando passamos de uma escala para a seguinte, na ordem em que foram apresentadas.

Pode-se argumentar que a escolaridade de uma pessoa é uma variável apenas ordinal. Mas é comum, em trabalhos econométricos, considerá-la como cardinal.

Note-se que o cálculo de uma simples média entre dois valores implica a comparação de dois intervalos e exige, portanto, que a medida tenha escala intervalar ou cardinal. Para uma escala nominal podemos determinar a moda (mas não a mediana ou

a média). Para uma escala ordinal podemos determinar tanto a moda como a mediana, mas não a média da variável.

Com uma variável binária ocorre algo interessante. Como ela tem apenas dois valores distintos, há um único intervalo e não podemos dizer que ela contrarie a condição para ser considerada intervalar. Isso permite que uma variável binária seja usada como variável explanatória em análise de regressão.

Cabe ressaltar que o modelo usual de regressão não permite que a variável dependente ( $Y$ ) seja binária. É óbvio que uma variável que inclui um erro com distribuição normal não pode ser binária. Há métodos especiais (como os modelos de lógite e próbite) para analisar variáveis dependentes binárias.

## 5.2. Uso de variáveis binárias para distinguir as categorias de uma variável nominal

Vamos admitir que uma variável nominal tenha  $k$  categorias. Podemos usar  $k - 1$  variáveis binárias para distinguir as  $k$  categorias.

A tabela 5.1 mostra uma maneira de distinguir as 5 regiões do Brasil usando 4 variáveis binárias, adotando o Nordeste como *base*.

TABELA 5.1. Uso de 4 variáveis binárias para distinguir 5 regiões.

Região	Variável binária			
	$Z_1$	$Z_2$	$Z_3$	$Z_4$
Nordeste	0	0	0	0
Norte	1	0	0	0
Sudeste	0	1	0	0
Sul	0	0	1	0
Centro-Oeste	0	0	0	1

Consideremos o modelo de regressão

$$Y_j = \alpha + \beta X_j + \gamma_1 Z_{1j} + \gamma_2 Z_{2j} + \gamma_3 Z_{3j} + \gamma_4 Z_{4j} + u_j \quad (5.1)$$

O parâmetro  $\gamma_3$ , por exemplo, é o valor esperado da mudança em  $Y_j$ , quando passamos do Nordeste para o Sul, para dado valor de  $X_j$ .

Para dado valor de  $X_j$ , a diferença entre o valor esperado de  $Y_j$  no Centro-Oeste e no Sudeste é  $\gamma_4 - \gamma_2$ .

Cabe assinalar que há várias outras maneiras corretas de distinguir as 5 regiões por meio de variáveis binárias. Uma alternativa, mantendo o Nordeste como *base*, é apresentada na tabela 5.2.

TABELA 5.2. Outra maneira de definir 4 variáveis binárias para distinguir 5 regiões.

Região	Variável binária			
	$Z_1$	$Z_2$	$Z_3$	$Z_4$
Nordeste	0	0	0	0
Norte	1	0	0	0
Sudeste	1	1	0	0
Sul	1	1	1	0
Centro-Oeste	1	1	1	1

Seja  $\delta_i$  o coeficiente de  $Z_{ij}$  ( $i = 1, 2, 3, 4$ ) no modelo com essa nova definição das variáveis. Neste caso o valor esperado da mudança em  $Y_j$  quando passamos do Nordeste para o Sul, dado  $X_j$ , é  $\delta_1 + \delta_2 + \delta_3$ . Pode-se provar que o uso de alternativas corretas na definição das variáveis binárias leva a resultados equivalentes. Assim, a estimativa de  $\delta_1 + \delta_2 + \delta_3$  será igual à estimativa de  $\gamma_3$  no modelo (5.1).

Se o modelo de regressão tiver um termo constante, não podemos utilizar  $k$  variáveis binárias para distinguir  $k$  categorias pois isso causaria um problema de multicolinearidade perfeita. Consideremos o esquema em que a variável  $Z_i$  (com  $i = 1, \dots, k$ ) é igual a 1 para a  $i$ -ésima categoria e é igual a zero para as demais categorias. Neste caso a primeira coluna da matriz  $\mathbf{X}$ , cujos elementos são todos iguais a 1, seria igual à soma das colunas referentes às  $k$  variáveis binárias e, consequentemente, a matriz  $\mathbf{X}'\mathbf{X}$  seria singular. Mas podemos usar uma binária para cada categoria se o termo constante for eliminado do modelo de regressão. Por exemplo, para captar as variações estacionais em uma série de preços mensais de um produto agrícola podemos utilizar o modelo

$$Y_t = \alpha + \sum_{i=2}^{12} \gamma_i Z_{it} + u_t \quad (5.2)$$

com  $Z_{it} = 1$  para o  $i$ -ésimo mês do ano e

$Z_{it} = 0$  para os demais meses,

ou o modelo

$$Y_t = \sum_{i=1}^{12} \gamma_i Z_{it} + u_t \quad (5.3)$$

Mas teríamos multicolinearidade perfeita se tentássemos usar o modelo

$$Y_t = \alpha + \sum_{i=1}^{12} \gamma_i Z_{it} + u_t \quad (5.4)$$

Vejamos um exemplo numérico de uso de uma variável binária para distinguir dois períodos utilizando a amostra de 8 observações das variáveis  $X_j$  e  $Y_j$  apresentada na tabela 5.3.

TABELA 5.3. Amostra de 8 pares de valores das variáveis  $X_j$  e  $Y_j$ .

Período	$X_j$	$Y_j$
I	1	8
	2	7
	3	7
	4	6
II	1	6
	2	5
	3	3
	4	2

Admitamos que as quatro primeiras observações se referem a um período com características distintas do período ao qual se referem as quatro últimas observações. Assim, por exemplo, se  $X_j$  é o preço de um produto e  $Y_j$  é a quantidade demandada, os dois períodos em questão poderiam ser inverno e verão, ou um período de guerra e um de paz, ou ainda, antes e depois de uma importante campanha publicitária.

Para avaliar as mudanças de um período para outro, consideramos uma variável binária ( $Z_j$ ) que assume o valor um ( $Z_j = 1$ ) quando a observação está no período I e o valor zero ( $Z_j = 0$ ) quando a observação está no período II.

O modelo estatístico é

$$Y_j = \alpha + \gamma Z_j + \beta X_j + u_j \quad (5.5)$$

Então, no período I a relação fica

$$Y_j = (\alpha + \gamma) + \beta X_j + u_j$$

e no período II a relação fica

$$Y_j = \alpha + \beta X_j + u_j$$

Portanto, estamos admitindo que o coeficiente de regressão ( $\beta$ ) é o mesmo nos dois períodos, só mudando o nível em que a reta se localiza. O modelo (5.5) corresponde, graficamente, a um par de retas paralelas.

A matriz  $\mathbf{X}$  relativa ao modelo (5.5) é

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 1 & 3 \\ 1 & 1 & 4 \\ 1 & 0 & 1 \\ 1 & 0 & 2 \\ 1 & 0 & 3 \\ 1 & 0 & 4 \end{bmatrix}$$

Obtemos

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 44 \\ 28 \\ 100 \end{bmatrix}, \quad \mathbf{X}'\mathbf{X} = \begin{bmatrix} 8 & 4 & 20 \\ 4 & 4 & 10 \\ 20 & 10 & 60 \end{bmatrix}$$

e

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \frac{7}{8} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{2} & 0 \\ -\frac{1}{4} & 0 & \frac{1}{10} \end{bmatrix}$$

Segue-se que o vetor das estimativas dos parâmetros é

$$\mathbf{b} = \begin{bmatrix} a \\ c \\ b \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} 6,5 \\ 3 \\ -1 \end{bmatrix}$$

A equação de regressão estimada é

$$\hat{Y}_j = 6,5 + 3Z_j - X_j$$

Na figura 5.1 está o par de retas paralelas correspondente a essa equação, juntamente com os pontos da amostra.

Para fazer a análise de variância da regressão, apresentada na tabela 5.4, calculamos

$$S.Q.Res. = \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y} = 272 - (6,5 \cdot 44 + 3 \cdot 28 - 1 \cdot 100) = 2$$

e

$$S.Q.Total = \mathbf{y}'\mathbf{y} - \frac{(\sum Y_j)^2}{n} = 272 - 242 = 30$$

TABELA 5.4. Análise de Variância

C.V.	G.L.	S.Q.	Q.M.	F
Regressão	2	28	14	35
Resíduo	5	2	0,4	
Total	7	30		

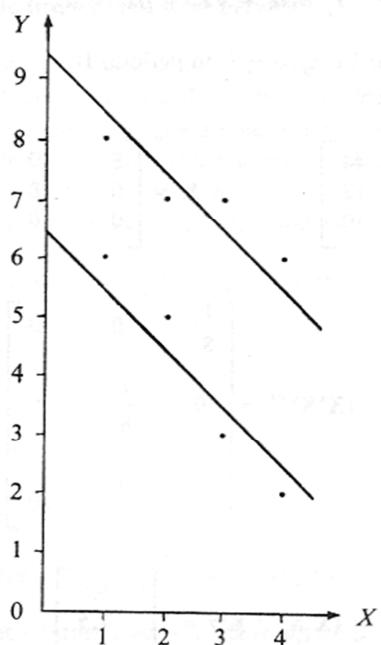


Figura 5.1. Retas paralelas ajustadas aos dados da tabela 5.3.

Vamos testar, ao nível de significância de 1%, a hipótese  $H_0 : \gamma = 0$  contra a hipótese alternativa  $H_A : \gamma > 0$ .

Temos

$$\hat{V}(c) = \frac{1}{2}0,4 = 0,2$$

e

$$t = \frac{c - \gamma}{s(c)} = \frac{3 - 0}{\sqrt{0,2}} = 6,708$$

A região de rejeição, para um teste unilateral com 5 graus de liberdade e ao nível de significância de 1%, é  $t \geq 3,365$ . O resultado obtido é, portanto, significativo, isto é, rejeita-se, ao nível de significância de 1%, a hipótese de que o termo constante da relação linear entre  $Y_j$  e  $X_j$  seja o mesmo nos dois períodos.

Os cálculos para ajustar um par de retas paralelas aos dados apresentados, com número igual de observações em cada período, ficam mais simples se considerarmos o modelo com a variável independente centrada:

$$Y_j = \alpha + \gamma Z_j + \beta X_j + u_j,$$

com  $Z_j = 1$  no período I e  $Z_j = -1$  no período II.

Neste caso, obtemos

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 44 \\ 12 \\ -10 \end{bmatrix}, \quad \mathbf{X}'\mathbf{X} = \begin{bmatrix} 8 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 10 \end{bmatrix},$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \frac{1}{8} & 0 & 0 \\ 0 & \frac{1}{8} & 0 \\ 0 & 0 & \frac{1}{10} \end{bmatrix}$$

e

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} 5,5 \\ 1,5 \\ -1 \end{bmatrix}$$

A equação estimada é

$$\hat{Y}_j = 5,5 + 1,5Z_j - x_j$$

ou

$$\hat{Y}_j = 8 + 1,5Z_j - X_j$$

No período I, com  $Z_j = 1$ , temos

$$\hat{Y}_j = 9,5 - X_j,$$

e no período II, com  $Z_j = -1$ , temos

$$\hat{Y}_j = 6,5 - X_j$$

O resultado, como era de se esperar, é o mesmo que obtivemos quando utilizamos o modelo (5.5).

### 5.3. Uso de variáveis binárias para ajustar poligonais

Neste caso as variáveis binárias são usadas para captar a mudança na inclinação entre segmentos consecutivos da poligonal.

O modelo geral para uma poligonal com  $k$  vértices ( $k+1$  segmentos) é

$$Y_j = \alpha + \beta X_j + \sum_{h=1}^k \gamma_h Z_{hj} (X_j - \theta_h) + u_j \quad (5.6)$$

onde  $\theta_h$  é a abcissa do  $h$ -ésimo vértice (que pressupomos conhecida) e  $Z_{hj}$  é uma variável binária tal que

$$Z_{hj} = 0 \text{ para } X_j \leq \theta_h$$

e

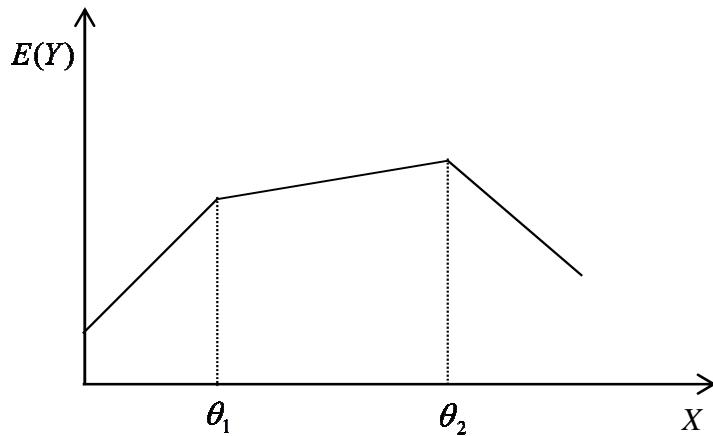
$$Z_{hj} = 1 \text{ para } X_j > \theta_h$$

Pode-se verificar que  $\gamma_h$  é a mudança na inclinação do  $h$ -ésimo segmento da poligonal, em relação à inclinação do segmento anterior.

Para uma poligonal com 3 segmentos o modelo fica

$$Y_j = \alpha + \beta X_j + \gamma_1 Z_{1j} (X_j - \theta_1) + \gamma_2 Z_{2j} (X_j - \theta_2) + u_j \quad (5.7)$$

A figura 5.2 mostra como poderia ser a forma da poligonal que mostra como  $E(Y_j)$  varia em função de  $X_j$  com  $\gamma_1 < 0$  e  $\gamma_2 < 0$ .



$$\begin{array}{lll} Z_1 = 0 & Z_1 = 1 & Z_1 = 1 \\ Z_2 = 0 & Z_2 = 0 & Z_2 = 1 \end{array}$$

Figura 5.2. Uma poligonal com 3 segmentos.

Para o 1º intervalo ( $X_j \leq \theta_1$ ) a reta é

$$E(Y_j) = \alpha + \beta X_j \quad (5.8)$$

No 2º intervalo ( $\theta_1 < X_j \leq \theta_2$ ) a reta é

$$E(Y_j) = \alpha - \gamma_1 \theta_1 + (\beta + \gamma_1) X_j \quad (5.9)$$

No 3º intervalo ( $X_j > \theta_2$ ) a reta é

$$E(Y_j) = \alpha - \gamma_1 \theta_1 - \gamma_2 \theta_2 + (\beta + \gamma_1 + \gamma_2) X_j \quad (5.10)$$

É interessante verificar que tanto (5.8) como (5.9) produzem a mesma ordenada para  $X_j = \theta_1$  (que é a ordenada do 1º vértice). Analogamente, (5.9) e (5.10) produzem a mesma ordenada para  $X_j = \theta_2$ .

Para obter a poligonal da figura acima devemos ter  $\alpha > 0$ ,  $\beta > 0$ ,  $\gamma_1 < 0$ ,  $\gamma_2 < 0$ ,  $\beta + \gamma_1 > 0$  e  $\beta + \gamma_1 + \gamma_2 < 0$ .

Vamos considerar um exemplo numérico para o qual o modelo tem apenas dois segmentos. Nesse caso o modelo fica

$$Y_j = \alpha + \beta X_j + \gamma Z_j (X_j - \theta) + u_j, \quad (5.11)$$

com  $Z_j = 0$  para  $X_j \leq \theta$  e  $Z_j = 1$  para  $X_j > \theta$ .

Consideremos, por simplicidade, que estamos analisando a tendência de uma variável ( $Y_j$ ) qualquer, ou seja, consideremos que a variável explanatória ( $X_j$ ) é o tempo, medido em anos, por exemplo.

Na tabela 5.5 são apresentados os valores de 6 observações consecutivas da variável  $Y_j$ . Vamos admitir que a inclinação da linha de tendência se modifique na 3<sup>a</sup> observação. Desejamos, portanto, ajustar uma poligonal com um vértice cuja abcissa é igual à abcissa da 3<sup>a</sup> observação.

Para facilitar os cálculos, vamos considerar que no instante correspondente à 3<sup>a</sup> observação temos  $X_j = 0$ . Dessa maneira temos  $\theta = 0$  e o modelo estatístico do problema em questão fica

$$Y_j = \alpha + \beta X_j + \gamma Z_j X_j + u_j \quad (5.12)$$

ou, fazendo  $W_j = Z_j X_j$ ,

$$Y_j = \alpha + \beta X_j + \gamma W_j + u_j$$

TABELA 5.5. Amostra de 6 observações consecutivas da variável  $Y_j$ .

Tempo ( $X_j$ )	$Z_j$	$Y_j$
-2	0	5,5
-1	0	5,0
0	0	1,0
1	1	3,5
2	1	4,5
3	1	4,5

Para as 3 primeiras observações, onde  $Z_j = 0$ , a relação é

$$E(Y_j) = \alpha + \beta X_j,$$

e para as 3 últimas observações, onde  $Z_j = 1$ , a equação da tendência passa a ser

$$E(Y_j) = \alpha + (\beta + \gamma) X_j$$

Confirma-se, portanto, que  $\gamma$  representa a mudança na tendência.

Para o exemplo apresentado, obtemos:

$$\mathbf{X} = \begin{bmatrix} 1 & -2 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 3 & 3 \end{bmatrix}, \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} 24 \\ 10 \\ 26 \end{bmatrix},$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 6 & 3 & 6 \\ 3 & 19 & 14 \\ 6 & 14 & 14 \end{bmatrix}, \quad (\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{114} \begin{bmatrix} 70 & 42 & -72 \\ 42 & 48 & -66 \\ -72 & -66 & 105 \end{bmatrix}$$

e

$$\mathbf{b} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \begin{bmatrix} 2 \\ -2 \\ 3 \end{bmatrix}$$

A equação ajustada é

$$\hat{Y}_j = 2 - 2X_j + 3Z_j X_j$$

A poligonal correspondente está traçada na figura 5.3. Note que a declividade no primeiro período é  $-2$  e no segundo período é  $(-2 + 3) = 1$ .

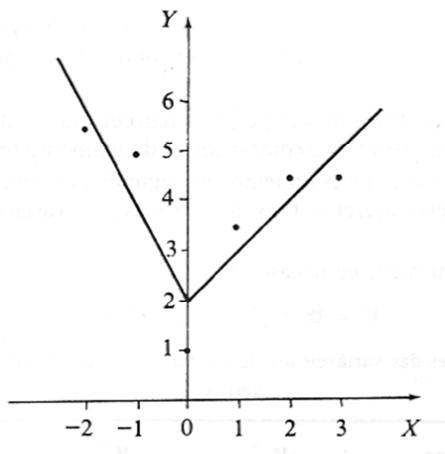


Figura 5.3. Poligonal ajustada aos dados da tabela 5.5.

A seguir calculamos

$$S.Q.Res. = \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y} = 109 - (2 \cdot 24 - 2 \cdot 10 + 3 \cdot 26) = 3$$

e fazemos a análise de variância da regressão, apresentada na tabela 5.6.

TABELA 5.6. Análise de Variância

C.V.	G.L.	S.Q.	Q.M.	F
Regressão	2	10	5	5
Resíduo	3	3	1	
Total	5	13		

Para verificar se a mudança de tendência após a terceira observação é estatisticamente significativa, fazemos o teste da hipótese  $H_0: \gamma = 0$ . Para isso calculamos

$$\hat{V}(c) = \frac{105}{114} = \frac{35}{38}$$

e

$$t = \frac{c - 0}{s(c)} = \frac{3}{\sqrt{\frac{35}{38}}} = 3,126$$

O valor crítico de  $t$  para um teste bilateral, ao nível de significância de 5% e com 3 graus de liberdade, é 3,182. O resultado obtido não é, portanto, significativo.

#### 5.4. Mudança estrutural

Vamos admitir que estamos analisando a relação entre duas variáveis ( $X_j$  e  $Y_j$ ) e temos duas situações (dois períodos, duas regiões ou duas categorias). Seja  $n_1$  o número de observações disponíveis para a situação I e seja  $n_2$  o número de observações disponíveis para a situação II. Podemos usar uma variável binária para distinguir as duas situações, fazendo  $Z_j = 0$  para as observações da situação I e  $Z_j = 1$  para as observações da situação II. Admitindo que tanto o “nível” como a inclinação da relação entre  $X_j$  e  $Y_j$  sejam diferentes nas duas situações, um modelo apropriado é

$$Y_j = \alpha + \beta X_j + \gamma Z_j + \delta Z_j X_j + u_j \quad (5.13)$$

A equação estimada com base nas  $n_1 + n_2$  observações é

$$\hat{Y} = a + bX + cZ + dZX \quad (5.14)$$

A respectiva soma de quadrados residual é

$$S.Q.Res. = S_U, \text{ com } n_1 + n_2 - 4 \text{ graus de liberdade}$$

Na situação I, com  $Z_j = 0$ , a relação entre  $X$  e  $Y$  é

$$E(Y_j) = \alpha + \beta X_j$$

Na situação II, com  $Z_j = 1$ , a relação fica

$$E(Y_j) = \alpha + \gamma + (\beta + \delta)X_j$$

Para verificar se existe “diferença estrutural” entre as duas situações testamos a hipótese

$$H_0 : \gamma = \delta = 0 \quad (5.15)$$

Seja  $F_E$  o valor de  $F$  calculado para testar essa hipótese.

O modelo *restrito* para a hipótese (5.15) é  $Y_j = \alpha + \beta X_j + u_j$ . Seja  $S_R$  a soma de quadrados residual obtida ajustando esse modelo restrito às  $n_1 + n_2$  observações. Então  $S_R$  está associado a  $n_1 + n_2 - 2$  graus de liberdade. De acordo com (4.85), sabemos que o valor de  $F_E$  pode ser obtido de

$$F_E = \frac{\frac{S_R - S_U}{2}}{\frac{S_U}{n_1 + n_2 - 4}} \quad (5.16)$$

Se admitimos que a relação entre  $X_j$  e  $Y_j$  é distinta nas duas situações analisadas, é lógico ajustar regressões separadamente para cada situação.

Para as  $n_1$  observações da situação I obtemos

$$\hat{Y} = a_1 + b_1 X \quad \text{e} \quad S.Q.Res. = S_1, \text{ com } n_1 - 2 \text{ graus de liberdade}$$

e para as  $n_2$  observações da situação II obtemos

$$\hat{Y} = a_2 + b_2 X \quad \text{e} \quad S.Q.Res. = S_2, \text{ com } n_2 - 2 \text{ graus de liberdade.}$$

Pode-se provar que  $a_1 = a$ ,  $b_1 = b$ ,  $a_2 = a + c$  e  $b_2 = b + d$ , isto é, que as duas retas estimadas separadamente são idênticas ao conjunto de duas retas estimado por meio do modelo (5.13). Consequentemente

$$S_1 + S_2 = S_U$$

Então o teste  $F$  para mudança estrutural pode ser obtido de

$$F_E = \frac{\frac{S_R - (S_1 + S_2)}{2}}{\frac{S_1 + S_2}{n_1 + n_2 - 4}} \quad (5.17)$$

Esse é o teste de Chow para mudança estrutural.

Genericamente, para um modelo com  $p$  parâmetros em cada uma das duas situações, temos

$$F_E = \frac{\frac{S_R - (S_1 + S_2)}{p}}{\frac{S_1 + S_2}{n_1 + n_2 - 2p}} \quad (5.18)$$

Para ilustrar o tema, consideremos um exercício apresentado em Draper e Smith (1966), no qual dispomos de 9 valores de uma variável  $Y_j$ , observados em 9 meses consecutivos. Vamos admitir que há uma tendência para as quatro primeiras observações, e que há outra tendência para as 5 últimas observações, com mudança tanto no termo constante como no coeficiente angular. Para captar essas mudanças definimos uma variável binária  $Z_j$  cujo valor é zero para as quatro primeiras observações e é 1 para as 5 últimas observações. Sendo  $X_j$  o número de ordem dos 9 meses, para simplificar um pouco as contas, vamos utilizar a variável centrada  $x_j = X_j - 5$ . O modelo fica

$$Y_j = \alpha + \beta x_j + \gamma Z_j + \delta Z_j x_j + u_j \quad (5.19)$$

A tabela 5.7 mostra os valores das variáveis que serão utilizadas para estimar a equação.

TABELA 5.7. Valores da variável  $Y_j$  e das variáveis explanatórias utilizados para ajustar um par de retas.

Tempo em meses ( $X_j$ )	$Y_j$	$x_j$	$Z_j$	$Z_j x_j$
1	1,0	-4	0	0
2	4,0	-3	0	0
3	6,0	-2	0	0
4	7,0	-1	0	0
5	9,5	0	1	0
6	11,0	1	1	1
7	11,5	2	1	2
8	13,0	3	1	3
9	13,5	4	1	4

Tendo em vista o modelo (5.19), obtemos

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 9 & 0 & 5 & 10 \\ 0 & 60 & 10 & 30 \\ 5 & 10 & 5 & 10 \\ 10 & 30 & 10 & 30 \end{bmatrix}, \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} 76,5 \\ 92 \\ 58,5 \\ 127 \end{bmatrix}$$

e

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 1,5 & 0,5 & -1,5 & -0,5 \\ 0,5 & 0,2 & -0,5 & -0,2 \\ -1,5 & -0,5 & 2,1 & 0,3 \\ -0,5 & -0,2 & 0,3 & 0,3 \end{bmatrix}$$

Então

$$\mathbf{b} = \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} 9,5 \\ 2 \\ 0,2 \\ -1 \end{bmatrix}$$

e

$$\hat{Y}_j = 9,5 + 2x_j + 0,2Z_j - Z_j x_j \quad (5.20)$$

No primeiro período, com  $Z_j = 0$ , a reta estimada é

$$\hat{Y}_j = 9,5 + 2x_j \quad \text{ou} \quad \hat{Y}_j = -0,5 + 2X_j$$

No segundo período, com  $Z_j = 1$ , a reta estimada é

$$\hat{Y}_j = 9,7 + x_j \quad \text{ou} \quad \hat{Y}_j = 4,7 + X_j$$

A figura 5.4 mostra os pontos da amostra e o par de retas ajustado.

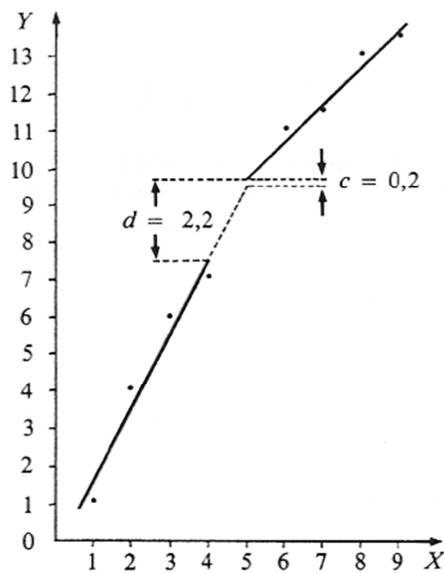


Figura 5.4. Retas ajustadas aos dados da tabela 5.7.

Para fazer a análise de variância da regressão, apresentada na tabela 5.8, devemos calcular

$$S.Q.Res. = \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y} = 796,75 - 795,45 = 1,3$$

TABELA 5.8. Análise de Variância

C.V.	G.L.	S.Q.	Q.M.	F
Regressão	3	145,2	48,4	186,15
Resíduo	5	1,3	0,26	
Total	8	146,5		

A tabela 5.9 apresenta, além das estimativas dos quatro parâmetros, as estimativas dos respectivos desvios padrões, o teste  $t$  e a correspondente probabilidade caudal (probabilidade associada a valores absolutos de  $t$  maiores do que o calculado).

TABELA 5.9. Estimativas dos parâmetros do modelo (5.19) e dos respectivos desvios padrões, o teste  $t$  e a correspondente probabilidade caudal.

Parâmetro	Estimativa	Desvio Padrão	Teste $t$	Probabilidade caudal
$\alpha$	9,5	0,6245	15,21	< 0,01%
$\beta$	2	0,2280	8,77	0,03%
$\gamma$	0,2	0,7389	0,27	79,75%
$\delta$	-1	0,2793	-3,58	1,59%

A estimativa de  $\gamma$  não é estatisticamente diferente de zero, mas, adotando um nível de significância de 5%, rejeitam-se as hipóteses de nulidade de  $\alpha$ ,  $\beta$  ou  $\delta$ .

É importante observar que o mesmo par de retas ajustado com base no modelo (5.19) é obtido fazendo-se duas regressões lineares simples.

Considerando os 4 primeiros pares de valores para as variáveis  $Y_j$  e  $X_j$  obtemos

$$\hat{Y}_j = -0,5 + 2X_j \quad (5.21)$$

com S.Q.Res. = 1.

Considerando os 5 últimos pares de valores para as variáveis  $Y_j$  e  $X_j$  obtemos

$$\hat{Y}_j = 4,7 + X_j \quad (5.22)$$

com S.Q.Res. = 0,3.

Note que a soma de quadrados de resíduos dessas duas regressões lineares simples é igual à soma de quadrados de resíduos da regressão múltipla ajustada anteriormente.

O ajustamento dessas duas regressões lineares simples exige menos cálculo do que o ajustamento de um modelo como (5.19). Entretanto, um modelo como (5.19) tem a vantagem de tornar relativamente mais fácil testar, posteriormente, hipótese envolvendo os valores dos parâmetros das duas retas.

Para fazer o teste de mudança estrutural da maneira indicada por Chow, além de obter as somas de quadrados de resíduos das equações (5.21) e (5.22), é necessário obter a soma de quadrados residual de uma regressão linear simples de  $Y_j$  contra  $X_j$  para as 9 observações. A equação estimada é

$$\hat{Y}_j = 0,8333 + 1,533X_j, \quad (5.23)$$

com S.Q.Res. =  $S_R = 5,4333$ , associada a 7 graus de liberdade. De acordo com (5.17), obtemos

$$F_E = \frac{\frac{5,4333 - (1 + 0,3)}{2}}{\frac{1 + 0,3}{5}} = \frac{2,0667}{0,26} = 7,95$$

Ao nível de significância de 5%, com 2 e 5 graus de liberdade, o valor crítico de  $F$  é 5,79. Portanto, rejeita-se a hipótese de que não houve mudança estrutural a partir do 5º ano, isto é, rejeita a hipótese  $H_0 : \gamma = \delta = 0$ . Cabe assinalar que o teste dessa hipótese também pode ser feito usando (4.60), obtendo-se exatamente o mesmo resultado.

Tendo concluído que há mudança estrutural, o ajustamento do modelo (5.19) permite que se especifique melhor a natureza da mudança. Nesse exemplo numérico, tendo em vista a tabela 5.9, verifica-se que a mudança ocorre, basicamente, no coeficiente angular da relação linear entre  $Y_j$  e  $X_j$ .

## 5.5. Análise de variância de dados com vários tratamentos e o teste para “falta de ajustamento”

Consideremos um total de  $n$  observações de uma variável, submetida a  $H$  diferentes tratamentos ( $h = 1, \dots, H$ ). Seja  $Y_{hi}$  o valor da  $i$ -ésima observação referente ao  $h$ -ésimo tratamento. A variável em questão pode ser o preço de um produto em diferentes regiões, a renda de indivíduos classificados conforme o nível de escolaridade, a produção de milho nas parcelas de um experimento de competição de variedades, etc.

Seja  $n_h$  o número de observações relativas ao  $h$ -ésimo tratamento, cujo total é

$$T_h = \sum_i Y_{hi} \quad (5.24)$$

O total geral é

$$G = \sum_h T_h = \sum_h \sum_i Y_{hi} \quad (5.25)$$

Para distinguir os  $H$  tratamentos vamos utilizar  $H$  variáveis binárias  $Z_h$  (com  $h = 1, \dots, H$ ), fazendo  $Z_h = 1$  para as observações do  $h$ -ésimo tratamento e  $Z_h = 0$  para as observações dos demais tratamentos. O modelo de regressão é

$$Y_{hi} = \gamma_1 Z_1 + \gamma_2 Z_2 + \dots + \gamma_H Z_h + u_{hi} \quad (5.26)$$

As linhas da matriz  $\mathbf{X}$  desse modelo têm todas um único elemento igual a 1, e os demais elementos iguais a zero. Assim, em todas as linhas temos

$$Z_1 + Z_2 + \dots + Z_H = 1 \quad (5.27)$$

Verifica-se que a matriz  $\mathbf{X}'\mathbf{X}$  é uma matriz diagonal com o número de observações de cada tratamento na diagonal e que o vetor-coluna  $\mathbf{X}'\mathbf{y}$  é formado pelos totais dos tratamentos. No caso de  $H = 3$  tratamentos, temos

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n_1 & 0 & 0 \\ 0 & n_2 & 0 \\ 0 & 0 & n_3 \end{bmatrix}, \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} T_1 \\ T_2 \\ T_3 \end{bmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \frac{1}{n_1} & 0 & 0 \\ 0 & \frac{1}{n_2} & 0 \\ 0 & 0 & \frac{1}{n_3} \end{bmatrix} \quad \text{e} \quad \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} \frac{T_1}{n_1} \\ \frac{T_2}{n_2} \\ \frac{T_3}{n_3} \end{bmatrix}$$

Segue-se que

$$\begin{aligned} \text{S.Q.Res.} &= \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y} = \sum_h \sum_i Y_{hi}^2 - \sum_h \frac{T_h^2}{n_h} = \\ &= \left( \sum_h \sum_i Y_{hi}^2 - \frac{G^2}{n} \right) - \left( \sum_h \frac{T_h^2}{n_h} - \frac{G^2}{n} \right) \end{aligned} \quad (5.28)$$

Na primeira expressão entre parênteses podemos reconhecer a soma de quadrados total:

$$\text{S.Q.Total} = \sum_h \sum_i \left( Y_{hi} - \frac{G}{n} \right)^2 = \sum_h \sum_i Y_{hi}^2 - \frac{G^2}{n} \quad (5.29)$$

com  $n - 1$  graus de liberdade.

A última expressão entre parênteses em (5.28) é a soma de quadrados de tratamentos:

$$\text{S.Q.Trat.} = \sum_h \frac{T_h^2}{n_h} - \frac{G^2}{n}, \quad \text{com } n - H \text{ graus de liberdade} \quad (5.30)$$

A soma de quadrados de tratamentos representa a parte da variação dos  $Y_{hi}$  devida às diferenças *entre* tratamentos, ou diferenças entre médias de  $Y_{hi}$  nos vários tratamentos ou categorias consideradas.

Note-se que, dados os valores de  $Y_{hi}$ , as expressões (5.28), (5.29) e (5.30) permitem calcular S.Q.Trat., S.Q.Total e S.Q.Res. sem que seja necessário estimar os parâmetros do modelo (5.26).

Vamos admitir, agora, que os diferentes tratamentos se distinguem apenas pelo valor de uma variável  $X$ , e seja  $X_h$  o valor correspondente ao  $h$ -ésimo tratamento. Pressupondo que o efeito de  $X_h$  sobre  $Y_{hi}$  seja linear, os parâmetros  $\gamma_h$  do modelo (5.26) podem ser substituídos por  $\alpha + \beta X_h$ . Com essa restrição, o modelo fica

$$Y_{hi} = (\alpha + \beta X_h)(Z_1 + Z_2 + \dots + Z_H) + u_{hi}$$

Lembrando (5.27), conclui-se que o modelo restrito é

$$Y_{hi} = \alpha + \beta X_h + u_{hi}, \quad (5.31)$$

que é o modelo de uma regressão linear simples de  $Y_{hi}$  contra  $X_h$ .

Em geral, se os tratamentos são caracterizados pelo valor de  $k$  variáveis explanatórias e pressupomos que seu efeito sobre  $Y$  é linear, obtermos um modelo de regressão linear múltipla com  $p = k + 1$  parâmetros. Se  $p < H$ , esse modelo será um modelo restrito em comparação com o modelo (5.26).

A soma de quadrados residual do modelo irrestrito é

$$S_U = \sum_h \sum_i Y_{hi}^2 - \sum_h \frac{T_h^2}{n_h}, \text{ com } n - H \text{ graus de liberdade} \quad (5.32)$$

e a soma de quadrados residual do modelo restrito é

$$S_R = (S.Q.Res. Y_{hi} | X_{1h}, \dots, X_{kh}), \text{ com } n - p \text{ graus de liberdade} \quad (5.33)$$

De acordo com (4.85), calculamos

$$F = \frac{\frac{S_R - S_U}{H - p}}{\frac{S_U}{n - H}} \quad (5.34)$$

Neste contexto, a diferença  $S_R - S_U$  é denominada soma de quadrados de “falta de ajustamento” e o correspondente teste  $F$  é denominado teste para “falta de ajustamento”, pois um valor elevado de  $S_R - S_U$  indica que o modelo onde se impõe a linearidade do efeito das variáveis explanatórias não se ajusta bem aos dados.

Só é razoável utilizar o modelo de regressão linear de  $Y_{hi}$  contra as  $k$  variáveis explanatórias se esse teste for não-significativo. Um valor de  $F$  significativo indica que devemos rejeitar a hipótese de linearidade do efeito das variáveis explanatórias sobre  $Y_{hi}$ . Neste caso deveremos nos limitar ao modelo (5.26) ou experimentar outras formas para a relação funcional entre  $Y_{hi}$  e as variáveis explanatórias.

Para exemplificar, consideremos os dados da tabela 5.8, com 8 observações e 3 tratamentos (3 valores distintos de uma única variável explanatória).

TABELA 5.8. Amostra de 8 valores de  $Y$ , com 3 valores distintos de  $X$ .

$X_h$	$n_h$	$Y_{hi}$	$T_h$
2	4	14, 11, 12 e 13	50
3	2	18 e 17	35
5	2	22 e 21	43

De acordo com (5.32), obtemos

$$S_U = 2168 - 2162 = 6 \quad , \text{ com } 5 \text{ graus de liberdade}$$

Ajustando a regressão linear simples de  $Y$  contra  $X$ , obtemos

$$\hat{Y} = 7 + 3X \tag{5.35}$$

e  $S_R = 12$ , com 6 graus de liberdade.

Substituindo esses valores em (5.34) verifica-se que o valor de  $F$  para “falta de ajustamento” é 5. Adotando um nível de significância de 5%, o valor crítico, para 1 e 5 graus de liberdade, é  $F_0 = 6,61$ . Então o resultado é não-significativo, não se rejeita a linearidade da relação entre  $X$  e  $Y$  e é válido usar a equação estimada (5.35).

Se fosse adotado um nível de significância de 10%, o valor crítico passaria a ser  $F_0 = 4,06$  e o valor obtido ( $F = 5$ ) seria significativo. Neste caso não seria razoável utilizar a equação (5.35) e deveríamos experimentar uma outra relação funcional entre  $Y$  e  $X$ . Uma alternativa seria considerar o modelo  $Y_j = \alpha + \beta_1 X_j + \beta_2 X_j^2 + u_j$ . Para esse

exemplo numérico não é possível fazer o teste de “falta de ajustamento” para esse novo modelo, pois o número de parâmetros ( $p = 3$ ) é igual ao número de tratamentos ( $H = 3$ ). A soma de quadrados residual da equação de segundo grau será, necessariamente, igual a  $S_U$ . Mas, se o número de tratamentos fosse maior, poderíamos fazer o teste de “falta de ajustamento” para a equação de regressão de segundo grau e verificar se esse novo modelo seria aceitável ou se seria necessário experimentar outras formas funcionais.

Cabe assinalar que o teste de “falta de ajustamento” só pode ser feito quando há, na amostra, mais de um valor de  $Y$  para determinadas combinações de valores das variáveis explanatórias (que definem os tratamentos), isto é, devemos ter  $n > H$ . Isso é comum em dados experimentais, mas não é comum em dados de amostras de levantamentos sócio-econômicos.

### Exercícios

5.1. Na análise da oferta de certo produto, admite-se que a função tem, conforme o período do ano, 2 posições distintas, mas com a mesma declividade. Foram observados os valores

Período	$X = \text{preço}$	$Y = \text{quantidade}$
I	1	2,0
	2	1,5
	3	2,5
II	1	3,0
	2	5,5
	3	6,5

Um pesquisador ajustou o modelo

$$Y_i = \alpha + \beta X_i + \gamma Z_i + u_i,$$

com  $Z$  tomando valor  $-1$  no período I e valor  $+1$  no período II.

- a) Estime os parâmetros.
- b) Qual é o deslocamento da oferta de um período para outro?
- c) Verifique se o deslocamento é estatisticamente significativo ao nível de significância de 1%.

5.2. Uma variável  $Y$  assume, em 5 anos consecutivos, os seguintes valores: 2,5; 3,0; 0; 3,0 e 2,5.

Ajuste a esses dados uma poligonal com vértice num ponto de abcissa igual à abcissa da 3<sup>a</sup> observação.

- a) Qual é a estimativa da declividade no 1º período (da 1<sup>a</sup> à 3<sup>a</sup> observação)?
  - b) Qual é a estimativa da declividade no 2º período (da 3<sup>a</sup> à 5<sup>a</sup> observação)?
  - c) Teste, ao nível de significância de 10%, a hipótese de que essas duas declividades são iguais.
- 5.3. Dois ensaios de adubação forneceram os seguintes resultados:

Ensaio	$X = \text{dose de nutriente}$	$Y = \text{produção}$
I	0	1,5
	1	6,0
	2	5,5
II	0	5,5
	1	7,0
	2	9,5

Os três primeiros pares de valores referem-se ao ensaio I e os três últimos pares ao ensaio II. Admite-se que a função de produção é uma parábola. Admite-se, também, que as funções de produção para os dois ensaios apresentam, para uma abcissa qualquer, a mesma declividade, embora a função correspondente ao ensaio II esteja em nível mais elevado.

- a) Adote um modelo e estime os seus parâmetros.
  - b) Qual é a estimativa da diferença de nível entre as duas funções?
  - c) Verifique se essa diferença é estatisticamente diferente de zero, ao nível de significância de 10%.
  - d) Teste a hipótese de que o coeficiente de regressão associado ao termo quadrático é igual a -3, considerando um nível de significância de 10%.
- 5.4. Dispomos de medidas da variável  $Y$  durante 5 anos consecutivos, com duas medidas (repetições) para cada ano.

Ano	Valores de $Y$
1	1 e 2
2	5 e 5
3	5 e 6
4	7 e 9
5	6 e 8

Admite-se que há uma tendência linear do 1º ao 4º ano e uma outra tendência linear do 4º para o 5º ano.

- a) Usando um modelo de regressão múltipla apropriado, ajuste aos dados uma linha poligonal com vértice no 4º ano. Qual é o significado das estimativas dos parâmetros obtidas?
  - b) Faça a análise de variância da regressão testando “falta de ajustamento”.
  - c) Teste a hipótese de que a declividade da linha no 2º período (4º ao 5º ano) é igual à declividade da linha no 1º período (1º ao 4º ano), considerando um nível de significância de 5%.
- 5.5. Consideremos duas amostras aleatórias, uma da variável  $Y_1$ , com  $n_1$  observações, e outra da variável  $Y_2$ , com  $n_2$  observações. Indiquemos por  $\mu_1$  e  $\mu_2$  as médias dessas variáveis, cujas estimativas são as médias das amostras  $\bar{Y}_1$  e  $\bar{Y}_2$ , respectivamente. Admitindo que as duas variáveis têm distribuições normais com variâncias iguais, a hipótese  $H_0 : \mu_1 = \mu_2$  pode ser testada, como sabemos, por meio do teste  $t$ ,

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)s^2}},$$

onde

$$s^2 = \frac{\sum_{i=1}^{n_1}(Y_{1i} - \bar{Y}_1)^2 + \sum_{i=1}^{n_2}(Y_{2i} - \bar{Y}_2)^2}{n_1 + n_2 - 2}$$

Demonstre que este teste é igual ao teste  $t$  relativo à hipótese  $H_0 : \beta = 0$ , sendo  $\beta$  o coeficiente de regressão do modelo

$$Y_{ki} = \alpha + \beta Z_{ki} + u_{ki},$$

onde:

- a)  $Z_{ki}$  é uma variável binária que assume valor  $-1$  para as observações de uma das amostras e valor  $+1$  no caso da outra amostra;
- b) o índice  $k = 1, 2$  indica que se trata de uma observação da variável  $Y_1$  ou da variável  $Y_2$ , e
- c) o índice  $i$  varia de 1 a  $n_1$  se  $k = 1$  e de 1 a  $n_2$  se  $k = 2$

- 5.6. Mostre que o teste  $t$  descrito no exercício anterior é, também, igual ao teste  $t$  relativo à hipótese  $H_0 : \beta = \gamma$ , sendo  $\beta$  e  $\gamma$  os coeficientes de regressão do modelo

$$Y_{ki} = \beta Z_{ki} + \gamma W_{ki} + u_{ki},$$

onde  $Z_{ki} = 1$  e  $V_{ki} = 0$  no caso das observações da amostra de  $Y_1$ , e  $Z_{ki} = 0$  e  $V_{ki} = 1$  quando se trata das observações da amostra de  $Y_2$ .

- 5.7. Para ajustar um par de retas a um conjunto de  $n_1 + n_2$  observações ( $n_1$  observações no grupo I e  $n_2$  observações no grupo II), podemos utilizar o modelo

$$Y_{ki} = \alpha_1 Z_{1k} + \alpha_2 Z_{2k} + \beta_1 Z_{1k} X_{ki} + \beta_2 Z_{2k} X_{ki} + u_{ki}, \\ k = 1, 2 \quad \text{e} \quad i = 1, 2, \dots, n_1 \text{ ou } n_2$$

com

$$Z_{11} = 1, \quad Z_{12} = 0, \quad Z_{21} = 0 \quad \text{e} \quad Z_{22} = 1$$

Demonstre que o valor de  $t$  para testar a hipótese  $H_0 : \beta_1 = \beta_2$  é

$$t = \frac{b_1 - b_2}{\sqrt{\left( \frac{1}{\sum x_{1i}^2} + \frac{1}{\sum x_{2i}^2} \right) \frac{Q_1 + Q_2}{n_1 + n_2 - 4}}}$$

onde, para  $k = 1, 2$ ,

$$b_k = \frac{\sum_i x_{ki} y_{ki}}{\sum_i x_{ki}^2},$$

$$x_{ki} = X_{ki} - \bar{X}_k,$$

$$y_{ki} = Y_{ki} - \bar{Y}_k$$

e  $Q_1$  e  $Q_2$  são as somas de quadrados de resíduo para as regressões lineares simples de  $Y_{ki}$  contra  $X_{ki}$ , para os grupos de observações I e II, respectivamente, isto é,

$$Q_k = \sum_i y_{ki}^2 - b_k \sum_i x_{ki} y_{ki}$$

- 5.8. Dados:

X	Valores de Y no		
	Tratamento 1	Tratamento 2	Tratamento 3
0	4	3	3
1	7	4	2
2	6	6	4
3	9	5	5
Totais	26	18	14

Admitimos que para cada tratamento existe uma relação linear entre  $X$  e  $Y$ , com o mesmo coeficiente angular, isto é, admitimos que a relação funcional entre  $Y$  e  $X$  pode ser representada por um feixe de 3 retas paralelas. Sejam  $\alpha_h$  ( $h = 1, 2, 3$ ) os coeficientes lineares das retas e seja  $\beta$  o coeficiente angular comum. Admitimos, também, que  $Y_i = E(Y_i) + u_i$ , com  $i = 1, 2, \dots, 12$ ,

onde  $u_i$  são variáveis aleatórias independentes, com média zero, variância  $\sigma^2$  e distribuição normal.

- Determine as estimativas de  $\alpha_h$  ( $h = 1, 2, 3$ ) e de  $\beta$  de acordo com o método de mínimos quadrados.
- Quais as propriedades dessas estimativas?
- Teste a hipótese  $H_0 : \beta = 0$ .
- Teste a hipótese  $H_0 : \alpha_1 = 0$  contra  $H_A : \alpha_1 > 0$ .
- Teste a hipótese  $H_0 : \alpha_1 = \frac{1}{2}(\alpha_2 + \alpha_3)$
- Teste a hipótese  $H_0 : \alpha_2 = \alpha_3$  e  $\alpha_1 = \frac{\alpha_2 + \alpha_3}{2} + 2$ .

Considere um nível de significância de 1%.

Sugestão: Adote, inicialmente, o modelo

$$Y_i = \sum_{h=1}^3 \alpha_h Z_{hi} + \beta X_i + u_i,$$

onde  $Z_{hi} = 1$  para toda observação do tratamento  $h$  e  $Z_{hi} = 0$  para as observações dos outros dois tratamentos (com  $h = 1, 2, 3$ ).

Dessa maneira  $\alpha_h$  é o intercepto da reta relativa ao  $h$ -ésimo tratamento.

A seguir mostre que o modelo inicialmente adotado é equivalente a

$$Y_i = \sum_{h=1}^3 \delta_h Z_{hi} + \beta x_i + u_i$$

onde  $\delta_h = \alpha_h + \beta \bar{X}$  e  $x_i = X_i - \bar{X}$

Os cálculos ficam bastante facilitados utilizando este último modelo, pois a correspondente matriz  $\mathbf{X}'\mathbf{X}$  será uma matriz diagonal.

5.9. Faça o teste para “falta de ajustamento” para a regressão linear simples do exercício 2.1.

5.10. Faça o teste para “falta de ajustamento” para a reta estimada no exercício 2.19.

5.11. É dada uma amostra com 12 pares de valores das variáveis  $X$  e  $Y$ :

$X$	$Y$	Temos
0	1; 1; 2; 2	$\sum Y = 36$ , $\sum X = 24$ , $\sum Y^2 = 126$
2	3; 4; 4; 5	
4	3; 3; 4; 4	$\sum x^2 = 32$ , $\sum y^2 = 18$ e $\sum xy = 16$

Admite-se que as variáveis estejam relacionadas de acordo com o modelo  $Y_j = \alpha + \beta X_j + u_j$ , onde os  $u_j$  são erros independentes, com  $E(u_j) = 0$ , variância constante e distribuição normal.

- Determine a reta de regressão de  $Y$  em relação a  $X$ , de acordo com o método dos mínimos quadrados.
- Calcule o coeficiente de determinação e faça a análise de variância da regressão, adotando um nível de significância de 1%.
- Verifique se há razões para rejeitar o modelo linear inicialmente proposto, considerando um nível de significância de 1%.

5.12. Verifique se há razões para rejeitar o modelo linear proposto no exercício 2.34.

5.13. Com base nos 8 pontos cujas coordenadas são dadas na tabela a seguir, ajuste um plano que passe pela origem dos eixos, considerando  $Y$  como variável dependente. Faça o teste para “falta de ajustamento”. Verifique se o coeficiente de  $X_2$  é estatisticamente diferente de zero, considerando um nível de significância de 5%

$X_1$	$X_2$	$Y$
1	1	3,5
1	1	4,5
1	2	5,5
1	2	4,5
2	1	4,5
2	1	3,5
2	2	5,0
2	2	6,0

5.14. A tabela ao lado mostra uma série de 9 valores quadrimestrais da variável  $Y$ . Admite-se que essa variável apresenta variações cíclicas estacionais. Verifica-se que  $\sum Y = 147$ ,  $\sum Y^2 = 2535$  e  $\sum y^2 = 134$ .

Ano	Quadrimestre	$Y$
1º	1º	15
	2º	22
	3º	15
2º	1º	11
	2º	18
	3º	16
3º	1º	10
	2º	20
	3º	20

- Estabeleça um modelo de regressão para captar as variações estacionais de  $Y$ , utilizando variáveis binárias. Construa a matriz  $\mathbf{X}$ .
- Estime os parâmetros do modelo.

- c) Teste, ao nível de significância de 5%, a hipótese de que não há variações estacionais [caso em que se tem  $E(Y) = \alpha$  ].
- d) Determine o intervalo de previsão para o valor de  $Y$  no 2º quadrimestre do 4º ano, ao nível de confiança de 95%.

5.15. É dada uma série de 9 valores anuais da variável  $Y$ .

Admite-se que  $Y$  varia linearmente em função do tempo (em anos), mas acredita-se que ocorreu uma mudança estrutural entre a 4ª e a 5ª observação, de maneira que haveria uma tendência linear durante os 4 primeiros anos da série e uma tendência linear distinta durante os 5 últimos anos.

Ano	$Y$
1º	39
2º	54
3º	63
4º	66
5º	96
6º	108
7º	111
8º	120
9º	135

Verifica-se que  $\sum Y = 792$ ,  $\sum Y^2 = 78588$  e  $\sum y^2 = 8892$ :

- a) Estime as taxas aritméticas de crescimento anual de  $Y$  nos dois períodos.
- b) Verifique se a mudança estrutural é estatisticamente significativa. Sugere-se fazer o teste com base nas regressões simples, como indicado por Chow.
- c) Há diferença estatisticamente significativa entre as taxas aritméticas de crescimento de  $Y$  nos dois períodos? Adote um nível de significância de 5% em todos os testes de hipóteses deste exercício.

5.16. Vamos admitir que temos os resultados de uma pesquisa de orçamentos familiares, sendo  $W$  a renda *per capita* e  $Q$  o consumo *per capita* de determinado alimento. Os respectivos logaritmos neperianos são

$$Y = \ln Q \quad \text{e} \quad X = \ln W$$

Admite-se que a elasticidade-renda do consumo é maior para os relativamente pobres do que para os relativamente ricos. Considera-se relativamente pobres as pessoas com  $X \leq 4$ . Para analisar como  $Y$  varia em função de  $X$  será adotado, então, um modelo que corresponde a uma poligonal com dois segmentos e vértice no ponto de abscissa ( $X$ ) igual a 4.

Dispomos de uma amostra com 6 pares de valores de  $X$  e  $Y$ :

- a) Estabeleça o modelo apropriado e estime seus parâmetros.
- b) Calcule o coeficiente de determinação da regressão.

$X$	$Y$
1	0,1
2	0,6
3	1,5
5	2,9
6	3,0
7	2,7

- c) Teste, ao nível de significância de 1%, a hipótese de que a elasticidade-renda do consumo desse alimento para os relativamente pobres é igual a zero, contra a hipótese alternativa de que essa elasticidade é positiva.
- d) Faça um teste bilateral, ao nível de significância de 1%, para a hipótese de que a elasticidade-renda para os relativamente ricos é igual a 1.

5.17. Temos uma amostra com 6 valores da variável econômica  $Y$  em duas regiões (3 observações em cada região), como mostra a tabela ao lado:

- a) Estabeleça um modelo de regressão com uma ou duas variáveis binárias para distinguir as duas regiões. Ressalte-se que o modelo vai captar apenas a diferença no valor de  $E(Y)$  nas duas regiões, incluindo um erro aleatório com as propriedades usuais.
- b) Com base na amostra, estime os dois parâmetros do modelo e mostre como essas estimativas estão associadas com a estimativa de  $E(Y)$  em cada região.
- c) Estime a variância do erro do modelo.
- d) Teste, ao nível de significância de 5%, a hipótese de que a  $E(Y)$  é a mesma nas duas regiões.

Região	$Y$
A	8
A	12
A	7
B	14
B	20
B	17

5.18. Dispomos dos 8 pares de valores das variáveis  $X_i$  e  $Y_i$  da tabela a seguir.

- a) Ajuste a equação de regressão linear simples de  $Y$  contra  $X$  e determine a respectiva soma de quadrados dos resíduos.
- b) Estabeleça um modelo cujas variáveis explanatórias são variáveis binárias que permitem distinguir os 4 diferentes valores de  $X$  observados. Estime os parâmetros do modelo, determine o valor da soma de quadrados dos resíduos e teste, ao nível de significância de 5%, a hipótese de que o valor esperado de  $Y$  é o mesmo para os 4 valores distintos de  $X$ .

$X_i$	$Y_i$
1	41
1	37
3	80
3	78
5	107
5	103
7	98
7	96

c) Considerando a regressão linear simples ajustada no item (a) como um modelo *restrito* em comparação com o modelo do item (b), faça um teste de “falta de ajustamento”, isto é, verifique, ao nível de significância de 1%, se deve ser rejeitada a hipótese de que o efeito de  $X$  sobre  $Y$  é linear.

d) Ajustando uma equação de segundo grau aos dados, foi obtida a equação

$$\hat{Y} = 7 + 34X - 3X^2,$$

com S.Q.Res. = 60. Faça um teste de “falta de ajustamento” para essa equação, isto é, verifique se podemos admitir que o efeito de  $X$  sobre  $Y$  obedece a uma equação de segundo grau, adotando um nível de significância de 5%.

e) Qual é a soma de quadrados dos desvios de uma equação de terceiro grau ajustada a esses dados? É possível fazer um teste de “falta de ajustamento” para a equação de terceiro grau? Justifique a resposta.

5.19. É dada uma série temporal de 8 valores

trimestrais da variável  $Y$ , cobrindo um período de dois anos. Admite-se que  $Y$  tenha variações cíclicas estacionais, além de erros  $u_j$  independentes com  $E(u_j) = 0$  e variância constante.

Ano	Trimestre	$Y$
1	1	29
1	2	18
1	3	13
1	4	22
2	1	27
2	2	22
2	3	11
2	4	18

- a) Estabeleça um modelo onde as variações estacionais de  $Y$  são captadas por meio de variáveis binárias e estime a equação.
- b) Teste, ao nível de significância de 5%, a hipótese de que o valor esperado de  $Y$  é o mesmo no terceiro e no quarto trimestres.
- c) Teste, ao nível de significância de 1%, a hipótese de que não há variações estacionais.

5.20. São dados os valores de  $X_2$  e  $Y$  para uma série de 13 anos, como mostra a tabela a seguir. Admitindo que haja uma “mudança estrutural” entre a 7<sup>a</sup> e a 8<sup>a</sup> observação (posição assinada na tabela pela linha tracejada), consideramos o modelo

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma Z + \delta_1 ZX_1 + \delta_2 ZX_2 + u$$

Ano ( $X_1$ )	( $X_2$ )	$Y$
1	12	70
2	4	50
3	8	54
4	8	66
5	8	62
6	4	66
7	12	94
<hr/>		
8	8	114
9	14	134
10	2	76
11	2	78
12	14	140
13	8	124

A equação estimada é

$$\hat{Y} = 26 + 4X_1 + 3X_2 + 24Z + 2ZX_1 + 2ZX_2,$$

com S.Q.Res. = 384,  $R^2 = 0,9670$  e

$$\bar{R}^2 = 0,9434.$$

Ajustando uma regressão múltipla de  $Y$  contra  $X_1$  e  $X_2$ , com as 13 observações, obtemos

$$\hat{Y} = 9,6923 + 6X_1 + 4,3846X_2,$$

com S.Q.Res. = 1069,54,  $R^2 = 0,9080$  e  $\bar{R}^2 = 0,8896$ .

- a) Qual é a equação estimada fazendo uma regressão de  $Y$  contra  $X_1$  e  $X_2$ , para as 7 primeiras observações?
- b) Qual é a equação estimada fazendo uma regressão de  $Y$  contra  $X_1$  e  $X_2$ , para as 6 últimas observações?
- c) Teste, ao nível de significância de 5% a hipótese de que ocorreu a suposta mudança estrutural (o que corresponde, no modelo inicial, a testar a hipótese de que  $\gamma = \delta_1 = \delta_2 = 0$ ).

- 5.21. A tabela a seguir mostra a escolaridade ( $X$ ) e o rendimento ( $Y$ ) de 4 pessoas ocupadas na agricultura e 4 pessoas ocupadas nos setores “urbanos” (indústria ou serviços). Define-se uma variável binária  $Z$  que é igual a zero para pessoas ocupadas na agricultura e é igual a 1 nos demais casos.

Z	X	Y
0	1	35
0	3	57
0	5	73
0	7	83
1	3	73
1	5	91
1	7	115
1	9	145

Para o modelo

$$Y_j = \alpha + \beta X_j + \gamma Z_j + \delta Z_j X_j + u_j$$

obteve-se

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 8 & 40 & 4 & 24 \\ 40 & 248 & 24 & 164 \\ 4 & 24 & 4 & 24 \\ 24 & 164 & 24 & 164 \end{bmatrix} \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} 672 \\ 3936 \\ 424 \\ 2784 \end{bmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 1,05 & -0,2 & -1,05 & 0,2 \\ -0,2 & 0,05 & 0,2 & -0,05 \\ -1,05 & 0,2 & 3,1 & -0,5 \\ 0,2 & -0,05 & -0,5 & 0,1 \end{bmatrix}$$

$$\hat{Y} = 30 + 8X + 4Z + 4ZX, \text{ S.Q.Res.} = 72 \quad \text{e} \quad s^2 = 18.$$

- a) Determine a equação de regressão de  $Y$  contra  $X$  e a respectiva S.Q.Res. para as 4 pessoas do setor agrícola.
- b) Idem, para as 4 pessoas do setor “urbano”.
- c) Teste  $H_0 : \gamma = 0$  ao nível de significância de 1%.
- d) Teste  $H_0 : \delta = 0$  ao nível de significância de 1%.
- e) Ao nível de significância de 1%, há diferença estrutural entre setor agrícola e setor “urbano” no que se refere à relação linear entre escolaridade e rendimento?
- f) Teste, ao nível de significância de 1%, a hipótese de que para o nível de escolaridade médio ( $X = 5$ ) não há diferença no rendimento esperado para pessoas ocupadas no setor agrícola e no setor “urbano”.

## Respostas

- 5.1. a)  $\hat{Y}_i = 1,5 + X + 1,5Z$   
 b) 3 unidades  
 c)  $t = 3,674$ , não-significativo ( $t_0 = 5,841$ )
- 5.2. a) -1 unidade por ano  
 b) +1 unidade por ano  
 c)  $t = 1,265$ , não-significativo ( $t_0 = 2,920$ )
- 5.3. a) Adotando o modelo  $Y_i = \alpha + \beta x_i^2 + \gamma x_i + \delta Z_i$ , com  $x_i = X_i - \bar{X}$ ,  $Z_1 = -1$  para o ensaio I e  $Z_1 = 1$  para o ensaio II, obtemos  

$$\hat{Y}_i = 6,5 - x_i^2 + 2x_i + 1,5Z_i$$
  
 ou  

$$\hat{Y}_i = 3,5 + 1,5Z_i + 4X_i - X_i^2$$
  
 b) 3 unidades  
 c)  $t = 3$ , significativo ( $t_0 = 2,920$ )  
 d)  $t = 1,886$ , não-significativo ( $t_0 = 2,920$ )
- 5.4. a) Sendo  $X$  o ano e  $Z$  uma variável binária que assume valor zero até o 4º ano e valor 1 no 5º ano, definimos  $V_1 = (1-Z)(X-4)$  e  $V_2 = Z(X-4)$ . Então  $V_1$  cresce de -3 para 0 nos primeiros anos e  $V_2$  cresce de 0 para 1 do 4º para o 5º ano. Obtemos  

$$\hat{Y} = 8 + 2V_1 - V_2$$
  
 As declividades no 1º e no 2º períodos são 2 e -1, respectivamente.  
 b) O valor de  $F$  para falta de ajustamento é igual a 1,5, não-significativo ao nível de 5%.  
 c) As declividades são estatisticamente diferentes ( $t = 2,510$ , significativo, pois  $t_0 = 2,365$ ).
- 5.8. a)  $b = 1$ ,  $a_1 = 5$ ,  $a_2 = 3$  e  $a_3 = 2$ .  
 b) São estimativas lineares não-tedenciosas de variância mínima, e consistentes.  
 São, também, estimativas de máxima verossimilhança.  
 c)  $t = 3,873$ , significativo ( $t_0 = 3,355$ )  
 d)  $t = 7,906$ , significativo ( $t_0 = 2,896$ )  
 e)  $t = 4,082$ , significativo ( $t_0 = 3,355$ )

f)  $F = 1,33$ , não-significativo ( $F_0 = 8,65$ )

5.9.  $F = 1,33$ , não-significativo ao nível de 10% ( $F_0 = 3,62$ )

5.10.  $F = 5/7$ , não-significativo.

5.11. a)  $\hat{Y} = 2 + 0,5X$

b)  $r^2 = 4/9 = 0,444$ ;  $F = 8$ , não-significativo ( $F_0 = 10,04$ )

c)  $F = 13,5$ , significativo ( $F_0 = 10,56$ ).

5.12.  $F = 0,55$ , não-significativo.

5.13.  $\hat{Y} = X_1 + 2X_2$

O valor de  $F$  para “falta de ajustamento” é 2,5, não-significativo ao nível de 10% ( $F_0 = 4,32$ ).

Para testar  $H_0: \beta_2 = 0$  obtemos  $t = 4,50$ , significativo ao nível de 5% ( $t_0 = 2,447$ ).

5.14. a)  $Y = \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + u$ , com  $Z_i = 1$  no  $i$ -ésimo quadrimestre e  $Z_i = 0$  nos demais quadrimestres.

b)  $\hat{Y} = 12Z_1 + 20Z_2 + 17Z_3$

c)  $F = 8,167$ , significativo ( $F_0 = 5,14$ )

d)  $13,08 < Y_{11} < 26,92$

5.15. a)  $b_1 = b_2 = 9$ .

b)  $F = 6,25$ , significativo ( $F_0 = 5,79$ )

c) As estimativas são iguais:  $t = 0$ , obviamente não-significativo.

5.16. a)  $Y = \alpha + \beta X + \gamma Z(X - 4) + u$ , com  $Z = 0$  para  $X \leq 4$  e  $Z = 1$  para  $X > 4$ .

$a = -1$ ,  $b = 0,9$  e  $c = -0,8$ .

b)  $R^2 = 0,9695$

c)  $t = 5,953$ , significativo ( $t_0 = 4,541$ )

d)  $t = -5,953$ , significativo ( $t_0 = 5,841$ ).

5.17. a)  $Y_i = \alpha + \beta Z_i + u_i$ , com  $Z_i = -1$  para a região A e  $Z_i = 1$  para a região B.

b)  $\hat{Y} = 13 + 4Z$

As estimativas de  $E(Y)$  nas regiões A e B são, respectivamente,  $13 - 4 = 9$  e  $13 + 4 = 17$ .

- c)  $s^2 = 8$ , com 4 graus de liberdade.
- d)  $t = 3,464$ , significativo ( $t_0 = 2,776$ ).

5.18. a)  $\hat{Y} = 40 + 10X$ , com S.Q.Res. = 1212

b)  $\hat{Y} = 39 + 40Z_2 + 66Z_3 + 58Z_4$

ou  $\hat{Y} = 39Z_1 + 79Z_2 + 105Z_3 + 97Z_4$ , com S.Q.Res. = 20

$F = 346,13$ , significativo ( $F_0 = 6,59$ )

c)  $F = \frac{596}{5} = 119,2$ , significativo ( $F_0 = 18,0$ )

d)  $F = \frac{40}{5} = 8$ , significativo ( $F_0 = 7,71$ )

e) S.Q.Res. = 20 (a mesma do item b). Não há grau de liberdade para o teste de “falta de ajustamento”.

5.19. a)  $Y = \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + u$ , com  $Z_i = 1$  para o  $i$ -ésimo trimestre e  $Z_i = 0$  para os demais trimestres ( $i = 1, 2, 3$  ou  $4$ ).

$\hat{Y} = 28Z_1 + 20Z_2 + 12Z_3 + 20Z_4$ .

b)  $t = 3,578$ , significativo ( $t_0 = 2,776$ ).

c)  $F = 17,07$ , significativo ( $F_0 = 16,69$ )

5.20. a)  $\hat{Y} = 26 + 4X_1 + 3X_2$ .

b)  $\hat{Y} = 50 + 2X_1 + 5X_2$

c)  $F = 4,166$ , não-significativo ( $F_0 = 4,35$ )

5.21. a)  $\hat{Y} = 30 + 8X$ , S.Q.Res. = 36

b)  $\hat{Y} = 34 + 12X$ , S.Q.Res. = 36

c)  $t = 0,535$ , não-significativo ( $t_0 = 4,604$ )

d)  $t = 2,981$ , não-significativo ( $t_0 = 4,604$ )

e)  $F = \frac{560}{18} = 31,11$ , significativo ( $F_0 = 18,0$ )

f)  $H_0 : \gamma + 5\delta = 0$ , com  $t = \frac{24}{\sqrt{10,8}} = 7,303$ , significativo ( $t_0 = 4,604$ )

## 6. HETEROCEDASTICIA

Veremos, neste capítulo, como obter as estimativas dos parâmetros de uma regressão linear quando a variância do erro não é constante, isto é, quando há heterocedasticia.

### 6.1. O caso de uma regressão linear simples em que o desvio padrão do erro é proporcional a $X$

Consideremos, inicialmente, o caso de uma regressão linear simples em que a variância do erro é proporcional ao valor de  $X^2$ . O modelo dessa regressão é

$$Y_j = \alpha + \beta X_j + u_j \quad (6.1)$$

com

$$E(u_j^2) = \sigma_j^2 = X_j^2 \sigma^2$$

Admitiremos que são válidas as demais pressuposições relativas ao modelo de regressão linear simples, vistas no capítulo 2.

O modelo (6.1) pode ser transformado em um modelo de regressão linear simples com homocedasticia. Para isso, basta dividir cada termo por  $X_j$ , obtendo

$$\frac{Y_j}{X_j} = \alpha \frac{1}{X_j} + \beta + \frac{u_j}{X_j}$$

ou

$$Z_j = \beta + \alpha V_j + \varepsilon_j, \quad (6.2)$$

onde

$$Z_j = \frac{Y_j}{X_j}, \quad V_j = \frac{1}{X_j} \quad \text{e} \quad \varepsilon_j = \frac{u_j}{X_j}$$

Convém ressaltar que

$$E(\varepsilon_j^2) = \frac{E(u_j^2)}{X_j^2} = \sigma^2,$$

ou seja, a variância do erro no modelo (6.2) é constante. O cálculo das estimativas dos parâmetros, a determinação de intervalos de confiança e os testes de hipóteses relativos ao

modelo (6.2) podem, portanto, ser feitos da maneira usual, utilizando as fórmulas de mínimos quadrados ordinários.

Os mesmos resultados podem ser obtidos através do raciocínio exposto a seguir. Sabemos que, no caso de um modelo homocedástico, as estimativas dos parâmetros são os valores que minimizam

$$\sum(Y_j - a - bX_j)^2$$

No caso de um modelo heterocedástico, com  $E(u_j^2) = \sigma_j^2$ , os quadrados dos desvios devem ser ponderados, sendo que o fator de ponderação deve ser inversamente proporcional à variância, isto é, devemos dar peso maior às observações de menor variância. As estimativas dos parâmetros são, então, os valores que minimizam

$$\sum \frac{1}{\sigma_j^2} (Y_j - a - bX_j)^2$$

No caso em que  $\sigma_j^2 = X_j^2 \sigma^2$  isso implica minimizar

$$\sum \frac{1}{X_j^2} (Y_j - a - bX_j)^2 = \sum \left( \frac{Y_j}{X_j} - b - a \frac{1}{X_j} \right)^2$$

Verificamos, portanto, que o resultado é o mesmo que o obtido aplicando o método dos mínimos quadrados ordinários (não ponderados) ao modelo (6.2).

## 6.2. O método dos mínimos quadrados ponderados

Consideremos o modelo

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \quad (6.3)$$

com  $E(\mathbf{u}) = \mathbf{0}$  e

$$E(\mathbf{u}\mathbf{u}') = \mathbf{V}\sigma^2 = \begin{bmatrix} v_1 & 0 & \dots & 0 \\ 0 & v_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & v_n \end{bmatrix} \sigma^2$$

Note que  $\mathbf{V}$  é uma matriz diagonal. Vamos admitir que sejam conhecidos os valores de  $v_j$ , que mostram como varia o valor da variância do erro. O fato de serem nulos os elementos fora da diagonal principal da matriz  $\mathbf{V}$  significa que é válida a pressuposição de

ausência de covariância entre os erros das várias observações, isto é, que  $E(u_j u_h) = 0$  para  $j \neq h$ .

Definimos a matriz diagonal

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}$$

onde

$$\lambda_j = \frac{1}{\sqrt{v_j}}, \quad j = 1, \dots, n$$

Dessa maneira, temos que

$$\Lambda \Lambda = \mathbf{V}^{-1} \quad (6.4)$$

e

$$\mathbf{V} = \Lambda^{-1} \Lambda^{-1} \quad (6.5)$$

Pré-multiplicando cada um dos termos de (6.3) por  $\Lambda$ , obtemos o modelo

$$\Lambda \mathbf{y} = \Lambda \mathbf{X} \boldsymbol{\beta} + \Lambda \mathbf{u} \quad (6.6)$$

No modelo (6.6) o vetor dos erros é  $\boldsymbol{\varepsilon} = \Lambda \mathbf{u}$  e uma vez que  $E(\mathbf{u}) = \mathbf{0}$ , temos  $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ .

Notando que  $\Lambda' = \Lambda$  e lembrando que  $E(\mathbf{u}\mathbf{u}') = \mathbf{V}\sigma^2$ , obtemos

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = E(\Lambda \mathbf{u} \mathbf{u}' \Lambda) = \Lambda \mathbf{V} \Lambda \sigma^2$$

De acordo com (6.5), segue-se que

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \Lambda \Lambda^{-1} \Lambda^{-1} \Lambda \sigma^2 = \mathbf{I} \sigma^2$$

ou seja, o modelo (6.6) é homocedástico. Podemos, então, aplicar a esse modelo as fórmulas de mínimos quadrados ordinários, deduzidas no capítulo 4. É óbvio que devemos tomar o cuidado de substituir, naquelas fórmulas, as matrizes  $\mathbf{y}$  e  $\mathbf{X}$  pelas matrizes  $\Lambda \mathbf{y}$  e  $\Lambda \mathbf{X}$ , respectivamente. Considerando (6.4) obtemos:

$$\mathbf{b} = (\mathbf{X}' \Lambda \Lambda \mathbf{X})^{-1} \mathbf{X}' \Lambda \Lambda \mathbf{y} = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1} \mathbf{y} \quad (6.7)$$

$$\text{S.Q.Res.} = \mathbf{y}' \Lambda \Lambda \mathbf{y} - \mathbf{b}' \mathbf{X}' \Lambda \Lambda \mathbf{y} = \mathbf{y}' \mathbf{V}^{-1} \mathbf{y} - \mathbf{b}' \mathbf{X}' \mathbf{V}^{-1} \mathbf{y} \quad (6.8)$$

$$E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})'] = (\mathbf{X}' \boldsymbol{\Lambda} \mathbf{X})^{-1} \sigma^2 = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \sigma^2 \quad (6.9)$$

Desde que  $E(\boldsymbol{\varepsilon}) = \mathbf{0}$  e  $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \mathbf{I}\sigma^2$ , isto é, desde que os erros  $\varepsilon_j$  são variáveis aleatórias independentes com média zero e variância constante, as estimativas dos parâmetros obtidas através de (6.7) são estimativas lineares não-tendenciosas de variância mínima, de acordo com o que foi visto no capítulo 4.

### 6.3. Consequências do uso de estimadores de mínimos quadrados ordinários quando existe heterocedasticidade

Vejamos, inicialmente, a perda de eficiência decorrente do uso das fórmulas de mínimos quadrados ordinários quando há heterocedasticidade.

Admitamos que o modelo correto seja (6.3) e que o pesquisador, erroneamente, admite que  $E(\mathbf{u}\mathbf{u}') = \mathbf{I}\sigma^2$ , calculando

$$\mathbf{b}_* = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y} \quad (6.10)$$

Substituindo (6.3) em (6.10), obtemos

$$\mathbf{b}_* = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' (\mathbf{X} \boldsymbol{\beta} + \mathbf{u})$$

ou

$$\mathbf{b}_* = \boldsymbol{\beta} + (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{u} \quad (6.11)$$

Aplicando esperança em (6.11) obtemos

$$E(\mathbf{b}_*) = \boldsymbol{\beta},$$

isto é, o estimador de mínimos quadrados ordinários é não-tendencioso. Entretanto,  $\mathbf{b}_*$  não é um estimador eficiente, já que o estimador de variância mínima é o dado por (6.7).

Determinemos a matriz de variâncias e covariâncias de  $\mathbf{b}_*$ . De (6.11) obtemos

$$\mathbf{b}_* - \boldsymbol{\beta} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{u}$$

Então

$$E[(\mathbf{b}_* - \boldsymbol{\beta})(\mathbf{b}_* - \boldsymbol{\beta})'] = E[(\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{u} \mathbf{u}' (\mathbf{X}' \mathbf{X})^{-1}]$$

Como  $E(\mathbf{u}\mathbf{u}') = \mathbf{V}\sigma^2$ , segue-se que

$$E[(\mathbf{b}_* - \boldsymbol{\beta})(\mathbf{b}_* - \boldsymbol{\beta})'] = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{V} \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \sigma^2 \quad (6.12)$$

Por simplicidade, consideremos o modelo

$$Y_j = \beta X_j + u_j, \quad j = 1, \dots, n \quad (6.13)$$

com

$$E(u_j) = 0, \quad E(u_j^2) = v_j \sigma^2 \text{ e } E(u_j u_h) = 0 \text{ para } h \neq j.$$

De acordo com (6.7), obtemos

$$b = \frac{\sum v_j^{-1} X_j Y_j}{\sum v_j^{-1} X_j^2} \quad (6.14)$$

O estimador de mínimos quadrados ordinários é

$$b_* = \frac{\sum X_j Y_j}{\sum X_j^2} \quad (6.15)$$

De acordo com (6.9) e (6.12) temos, respectivamente,

$$V(b) = \frac{\sigma^2}{\sum v_j^{-1} X_j^2} \quad (6.16)$$

e

$$V(b_*) = \frac{\sigma^2 \sum v_j X_j^2}{(\sum X_j^2)^2} \quad (6.17)$$

A eficiência relativa de  $b_*$ , em comparação com  $b$ , é

$$\phi = \frac{V(b)}{V(b_*)} = \frac{(\sum X_j^2)^2}{\sum v_j^{-1} X_j^2 \sum v_j X_j^2}$$

Para exemplificar, admitamos que a variância de  $Y$ , dado  $X$ , seja direta ou inversamente proporcional ao quadrado do valor de  $X$ , isto é,  $v_j = X_j^2$  ou  $v_j = X_j^{-2}$ . Em ambos os casos temos

$$\phi = \frac{(\sum X_j^2)^2}{n \sum X_j^4}$$

Se  $X_j = j - 1$ , com  $j = 1, \dots, 11$ , temos  $\phi = 0,532$ , isto é, a eficiência relativa do estimador de mínimos quadrados ordinários, em comparação com o estimador de mínimos quadrados ponderados, é de apenas 53,2%.

Outros exemplos, considerando o modelo  $Y_j = \alpha + \beta X_j + u_j$ , podem ser encontrados em Johnston (1971, p. 225-229).

É claro que o pesquisador que estivesse, inadvertidamente, utilizando as expressões de mínimos quadrados ordinários consideraria  $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$ , e não (6.12), como a matriz de variâncias e covariâncias de  $\mathbf{b}_*$ , e, para estimar  $\sigma^2$ , utilizaria

$$\begin{aligned} s_*^2 &= \frac{1}{n-p}(\mathbf{y}'\mathbf{y} - \mathbf{b}'_*\mathbf{X}'\mathbf{y}) = \\ &= \frac{1}{n-p}(\mathbf{y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}), \end{aligned} \quad (6.18)$$

em lugar do estimador correto que, de acordo com (6.8), é

$$s^2 = \frac{1}{n-p}(\mathbf{y}'\mathbf{V}^{-1}\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{y})$$

Lembrando (6.7), obtemos

$$s^2 = \frac{1}{n-p}\mathbf{y}'[\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}]\mathbf{y} \quad (6.19)$$

Substituindo (6.3) em (6.18) e (6.19), obtemos

$$s_*^2 = \frac{1}{n-p}\mathbf{u}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{u} \quad (6.20)$$

e

$$s^2 = \frac{1}{n-p}\mathbf{u}'[\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}]\mathbf{u} \quad (6.21)$$

De (6.20), notando que se trata de uma matriz com um único elemento, temos

$$\begin{aligned} s_*^2 &= \frac{1}{n-p} \text{tr}\{\mathbf{u}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{u}\} = \\ &= \frac{1}{n-p} \text{tr}\{\mathbf{u}\mathbf{u}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\} \end{aligned}$$

Como  $E(\mathbf{u}\mathbf{u}') = \mathbf{V}\sigma^2$  segue-se que

$$\begin{aligned} E(s_*^2) &= \frac{\sigma^2}{n-p} \text{tr}\{\mathbf{V}[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\} = \\ &= \frac{\sigma^2}{n-p} \{\text{tr}(\mathbf{V}) - \text{tr}[\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]\} \end{aligned} \quad (6.22)$$

Analogamente, de (6.21) obtemos

$$\begin{aligned}
E(s^2) &= \frac{\sigma^2}{n-p} \text{tr}\{\mathbf{V}[\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}]\} = \\
&= \frac{\sigma^2}{n-p} \{\text{tr}(\mathbf{I}_n) - \text{tr}[\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}]\} = \\
&= \frac{\sigma^2}{n-p} [\text{tr}(\mathbf{I}_n) - \text{tr}(\mathbf{I}_p)] = \sigma^2
\end{aligned} \tag{6.23}$$

Esse resultado já era esperado, pois (6.19) é o quadrado médio do resíduo relativo ao modelo (6.6), cujo vetor de erros ( $\boldsymbol{\varepsilon}$ ) é tal que  $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \mathbf{I}\sigma^2$ , que é, de acordo com o que vimos na seção 4.5, a condição necessária para demonstrar que o quadrado médio do resíduo é um estimador não-tendencioso da variância residual.

Por simplicidade, consideremos, novamente, o modelo (6.13). O pesquisador que, inadvertidamente, não considerasse a existência de heterocedasticidade, calcularia  $b_*$ , dado por (6.15), e, de acordo com (4.23), obteria

$$\hat{V}_*(b_*) = \frac{s_*^2}{\sum X_j^2} \tag{6.24}$$

onde, de acordo com (6.18),

$$s_*^2 = \frac{1}{n-1} \left[ \sum Y_j^2 - \frac{(\sum X_j Y_j)^2}{\sum X_j^2} \right]$$

Considerando (6.22), temos que

$$E[\hat{V}_*(b_*)] = \frac{\sigma^2}{(n-1)\sum X_j^2} \left( \sum v_j - \frac{\sum v_j X_j^2}{\sum X_j^2} \right) \tag{6.25}$$

De acordo com (6.17), o estimador correto da variância de  $b_*$  é

$$\hat{V}(b_*) = \frac{s^2 \sum v_j X_j^2}{(\sum X_j^2)^2}, \tag{6.26}$$

onde, de acordo com (6.19),

$$s^2 = \frac{1}{n-1} \left[ \sum v_j^{-1} Y_j^2 - \frac{(\sum v_j^{-1} X_j Y_j)^2}{\sum v_j^{-1} X_j^2} \right]$$

Considerando (6.23), temos que

$$E[\hat{V}(b_*)] = \frac{\sigma^2 \sum v_j X_j^2}{(\sum X_j^2)^2} \tag{6.27}$$

Comparando (6.27) com (6.17), verificamos que (6.26) é um estimador não-tendencioso da variância de  $b_*$ .

De (6.17) e (6.25) obtemos a tendenciosidade ou viés de (6.24) como estimador de  $V(b_*)$ , que é

$$\begin{aligned} E[\hat{V}_*(b_*)] - V(b_*) &= \\ &= \frac{\sigma^2}{\sum X_j^2} \left[ \frac{1}{n-1} \left( \sum v_j - \frac{\sum v_j X_j^2}{\sum X_j^2} \right) - \frac{\sum v_j X_j^2}{\sum X_j^2} \right] = \\ &= \frac{n\sigma^2}{(n-1)\sum X_j^2} \left( \frac{\sum v_j}{n} - \frac{\sum v_j X_j^2}{\sum X_j^2} \right) \end{aligned} \quad (6.28)$$

Nesta expressão temos, entre parênteses, a diferença entre a média aritmética e a média ponderada dos  $v_j$ , com  $X_j^2$  como fatores de ponderação. Se, por exemplo, os maiores valores dos  $v_j$  estiverem associados aos maiores valores absolutos dos  $X_j$ , a média ponderada é maior do que a média aritmética e, portanto,  $\hat{V}_*(b_*)$  é um estimador negativamente viesado de  $V(b_*)$ . É óbvio que, neste caso, não são válidos os intervalos de confiança ou testes de hipóteses feitos com base nos estimadores tendenciosos das variâncias obtidos de  $(\mathbf{X}'\mathbf{X})^{-1}s_*^2$ .

Entretanto, se não houver associação entre  $v_j$  e  $X_j^2$ , as duas médias dos  $v_j$  (sem ponderação e com ponderação por  $X_j^2$ ) em (6.28) tendem a ser iguais. Neste caso o pesquisador que usa as fórmulas de mínimos quadrados ordinários, embora não esteja usando o estimador mais eficiente, não será sistematicamente induzido a conclusões erradas ao efetuar teste de hipóteses.

#### **6.4. Testes para a homocedasticia e obtenção de estimativas dos parâmetros quando a matriz $\mathbf{V}$ é desconhecida**

Até aqui admitimos que a matriz  $\mathbf{V}$ , de  $E(\mathbf{u}\mathbf{u}') = \mathbf{V}\sigma^2$ , é conhecida.

Entretanto, em problemas práticos, freqüentemente desconhecemos se os erros são homocedásticos ou heterocedásticos. Vejamos então como podemos, com base nos dados da amostra, verificar se a variância dos erros é ou não homogênea.

Vamos admitir, inicialmente, que, na amostra disponível, temos repetições de conjuntos de valores das variáveis explanatórias, ou seja, dispomos de  $n_h > 1$  valores de  $Y$  para os valores  $X_{1h}, X_{2h}, \dots, X_{kh}$  das variáveis explanatórias. Se  $H$  é o número de diferentes conjuntos de valores de  $X_i$  existentes (ou número de vetores diferentes entre as linhas da matriz  $\mathbf{X}$ ), os  $n = \sum n_h$  valores da variável dependente podem ser indicados por  $Y_{hj}$  ( $j = 1, \dots, n_h ; h = 1, \dots, H$ ).

As estimativas das variâncias dentro de cada grupo são dadas por

$$s_h^2 = \frac{\sum_j (Y_{hj} - \bar{Y}_h)^2}{g_h}, \quad (6.29)$$

onde  $g_h = n_h - 1$  e  $\bar{Y}_h = \frac{1}{n_h} \sum Y_{hj}$

Sejam

$$U = (\sum g_h) \ln \frac{\sum g_h s_h^2}{\sum g_h} - \sum g_h \ln s_h^2 \quad (6.30)$$

e

$$G = 1 + \frac{1}{3(H-1)} \left( \sum \frac{1}{g_h} - \frac{1}{\sum g_h} \right) \quad (6.31)$$

Pode-se demonstrar que, se a variância de  $Y$  é homogênea, a variável  $U/G$  tem, aproximadamente, distribuição de qui-quadrado com  $H - 1$  graus de liberdade. Então, o valor de  $U/G$  pode ser utilizado para testar a hipótese  $H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_H^2$ , isto é, a hipótese de que a variância de  $Y$  é constante.<sup>11</sup>

Se o teste for não-significativo, é justificável pressupor que há homocedasticia. Neste caso, é razoável aplicar o método de mínimos quadrados ordinários. Ressaltemos, entretanto, que um resultado não-significativo não implica, necessariamente, que haja homocedasticia. O teste de hipótese pode apenas mostrar se é razoável manter ou não a pressuposição de homocedasticia, nunca provando sua veracidade. Em outras palavras, o fato de o teste para homocedasticia não ser significativo não tira o caráter de *pressuposição* da afirmação de que os erros são homocedásticos.

---

<sup>11</sup> Ver Hoel (1962, p. 225-227).

Se o teste resultar significativo, devemos usar o método de mínimos quadrados ponderados. Para isso, como a matriz  $\mathbf{V}$  é desconhecida, ela é substituída por uma matriz diagonal  $\hat{\mathbf{V}}$ , cujos elementos diferentes de zero são  $\hat{v}_{hj} = s_h^2$ , obtidos de (6.29). Quando utilizamos  $\hat{\mathbf{V}}$  em lugar de  $\mathbf{V}$ , temos  $\mathbf{b} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$ , que não é um estimador linear não-tendencioso de variância mínima. Pode-se demonstrar, entretanto, que, em certas condições,<sup>12</sup> esse é um estimador consistente e assintoticamente eficiente de  $\beta$ .

Consideremos, agora, o caso em que o número de valores de  $Y$  para um mesmo conjunto de valores de  $X_i$  é insuficiente para aplicar o procedimento anteriormente descrito. Então, para testar a homocedasticidade dos erros, pode-se usar o método proposto por Goldfeld e Quandt (1965), que descrevemos a seguir.

Dado o modelo  $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ , com  $E(\mathbf{u}) = \mathbf{0}$  e  $E(\mathbf{u}\mathbf{u}') = \mathbf{V}\sigma^2$ , onde  $\mathbf{V}$  é uma matriz diagonal cujos elementos diferentes de zero são  $v_j$  ( $j = 1, \dots, n$ ), a hipótese da nulidade é  $H_0 : v_j = \theta$ , onde  $\theta$  é uma constante, e a hipótese alternativa é que  $v_j$  é uma função monotonicamente crescente (ou decrescente) de  $X_{ij}$  (uma das variáveis independentes do modelo) ou de alguma outra variável cujos valores, para cada uma das observações da amostra, são conhecidos.

São as seguintes as etapas do teste:

- Ordenamos as observações de acordo com valores crescentes de  $X_{ij}$  ( $j = 1, \dots, n$ ).
- Eliminamos  $m$  observações centrais e ajustamos, pelo método de mínimos quadrados ordinários, uma equação de regressão para as primeiras  $(n - m)/2$  observações e uma outra equação de regressão para as últimas  $(n - m)/2$  observações. Simulações realizadas por Goldfeld e Quandt, considerando o caso em que  $v_j = X_{ij}^2$ , indicam que o poder do teste é maior quando  $m$  é igual a cerca de 1/4 de  $n$ .
- Sendo  $S_1$  e  $S_2$  as somas de quadrados de resíduos das regressões com os valores relativamente pequenos e relativamente grandes de  $X_{ij}$ , respectivamente, calculamos  $T_1 = S_2 / S_1$ , se a hipótese alternativa é que os  $v_j$  crescem com  $X_{ij}$  e

---

<sup>12</sup> Ver Theil (1971), p. 399.

calculamos  $T_2 = S_1 / S_2$  se a hipótese alternativa é que os  $v_j$  decrescem com  $X_{ij}$ . Se  $H_0$  é verdadeira,  $T_1$  (ou  $T_2$ ) tem distribuição de  $F(n-m-2p)/2$  e  $(n-m-2p)/2$  graus de liberdade, onde  $p$  é o número de parâmetros da equação de regressão. Se for verdadeira a hipótese alternativa de que o valor de  $v_j$  é uma função monotonicamente crescente (ou decrescente) de  $X_{ij}$ , o valor de  $T_1$  (ou  $T_2$ ) tende, obviamente, a ser elevado (maior do que 1).

O teste descrito pode ser utilizado para verificar se determinada hipótese a respeito da forma da heterocedasticidade é razoável.

Consideremos, por exemplo, que no modelo

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + u_j$$

desejamos verificar se é razoável pressupor que  $E(u_j^2) = X_{1j}^2 \sigma^2$ .

Dada essa pressuposição, o modelo

$$\frac{Y_j}{X_{1j}} = \alpha \frac{1}{X_{1j}} + \beta_1 + \beta_2 \frac{X_{2j}}{X_{1j}} + \frac{u_j}{X_{1j}}$$

é homocedástico, pois  $E(u_j/X_{1j})^2 = \sigma^2$ . Portanto, para verificar se a pressuposição de que  $E(u_j^2) = X_{1j}^2 \sigma^2$  é razoável, aplicamos o teste de Goldfeld e Quandt considerando a regressão de  $Y_j/X_{1j}$  contra  $1/X_{1j}$  e  $X_{2j}/X_{1j}$ .

Um outro teste para verificar a existência de heterocedasticidade foi proposto por Glejser (1969). Sejam  $e_j$  os desvios da equação de regressão estimada pelo método de mínimos quadrados ordinários. Admitindo que a variância do erro do modelo é uma função monotônica do valor da variável  $X_{ij}$ , ajustamos alguns modelos simples de regressão do módulo de  $e_j$  em função de  $X_{ij}$ . Glejser sugere que sejam considerados os modelos

$$|e_j| = \gamma_0 + \gamma_1 X_{ij}^\delta + \varepsilon_j,$$

onde  $\delta$  é igual a 1, -1, 1/2 ou -1/2.

Sejam  $c_0$  e  $c_1$  as estimativas de mínimos quadrados ordinários de  $\gamma_0$  e  $\gamma_1$ .

A seguir testamos as hipóteses  $H_0 : \gamma_0 = 0$  e  $H_0 : \gamma_1 = 0$ . Três possibilidades são consideradas pelo autor:

- a) Não se rejeita  $H_0 : \gamma_0 = 0$ ,  $c_0 > 0$  e rejeita-se  $H_0 : \gamma_1 = 0$ .

É razoável, então, pressupor que o desvio padrão de  $u_j$  é proporcional a  $X_{ij}^\delta$ , aplicando-se então o método de mínimos quadrados ponderados com  $v_j = X_{ij}^{2\delta}$ .

b) O valor de  $c_0 + c_1 X_{ij}^\delta$  é positivo no intervalo relevante de  $X_{ij}$  e rejeita-se  $H_0 : \gamma_0 = 0$  e  $H_0 : \gamma_1 = 0$ . Neste caso devemos aplicar o método de mínimos quadrados ponderados substituindo os  $v_j$  por  $\hat{v}_j = (c_0 + c_1 X_{ij}^\delta)^2$ .

c) Em qualquer outro caso, admite-se que os erros são homocedásticos e utiliza-se o método de mínimos quadrados ordinários.

Tanto o teste de Glejser como o teste de Goldfeld e Quandt exigem que a hipótese alternativa especifique que a variância do erro seja uma função monotônica de uma única variável (que pode ser ou não uma das variáveis explanatórias do modelo de regressão que desejamos estimar). O teste de Breusch-Pagan/Godfrey é mais geral, admitindo que a variância do erro ( $\sigma_j^2$ ) seja uma função de uma combinação linear de  $K$  variáveis  $Z_{1j}, Z_{2j}, \dots, Z_{Kj}$ , ou seja:

$$\sigma_j^2 = \psi \left( \delta_0 + \sum_{h=1}^K \delta_h Z_{hj} \right),$$

com  $\psi$  sendo uma função qualquer para a qual são definidas as duas primeiras derivadas. A hipótese de nulidade, que estabelece a homocedasticidade dos erros, corresponde a

$$H_0 : \delta_1 = \delta_2 = \dots = \delta_K = 0$$

O procedimento para efetuar esse teste pode ser dividido em 3 etapas.

a) Estimamos o modelo  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$  por mínimos quadrados ordinários e obtemos o vetor dos desvios  $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}$ , com elementos  $e_j (j = 1, \dots, n)$ . Determinamos a estimativa de máxima verossimilhança da variância do erro, dada por

$$\hat{\sigma}^2 = (\mathbf{e}'\mathbf{e})/n$$

e calculamos os valores de

$$g_j = \frac{e_j^2}{\hat{\sigma}^2} \text{ para } j = 1, \dots, n$$

b) Fazemos a regressão de  $g_j$  contra  $Z_{1j}, Z_{2j}, \dots, Z_{Kj}$ , incluindo um termo constante, e calculamos a respectiva soma de quadrados de regressão, que passamos a indicar por  $Q$ .

c) Se os erros  $u_j$  do modelo original tiverem distribuição normal e variância constante,  $Q/2$  tem, assintoticamente, distribuição de qui-quadrado com  $K$  graus de liberdade. A hipótese de nulidade (homocedasticia) será rejeitada se o valor de  $Q/2$  for igual ou maior do que o valor crítico de  $\chi^2_K$ , ao nível de significância escolhido.

Os testes de Goldfeld e Quandt, de Glejser e de Breusch-Pagan/Godfrey exigem que tenhamos alguma idéia sobre a natureza da heterocedasticia (sua associação com uma ou mais variáveis conhecidas). Isso não é necessário no teste de White (1980), cujo procedimento é descrito a seguir. Sejam  $e_j$  (com  $j = 1, \dots, n$ ) os desvios da regressão de  $Y_j$  contra  $X_{1j}, X_{2j}, \dots, X_{kj}$ , estimada pelo método de mínimos quadrados ordinários. Incluindo o termo constante, essa regressão inicial tem  $p = k + 1$  coeficientes. Calculamos os valores de  $e_j^2$  e fazemos uma regressão auxiliar dessa variável contra as mesmas variáveis explanatórias ( $X_{1j}, X_{2j}, \dots, X_{kj}$ ), seus quadrados ( $X_{1j}^2, X_{2j}^2, \dots, X_{kj}^2$ ) e todos os produtos de duas variáveis ( $X_{1j}X_{2j}, \dots, X_{1j}X_{kj}, \dots$ ). Pode-se verificar que em geral essa regressão auxiliar terá  $p(p+1)/2$  coeficientes. Mas o número de coeficientes poderá ser menor, pois é necessário eliminar possíveis repetições de variáveis. Se, por exemplo, uma das variáveis do modelo original for uma variável binária, seu quadrado não pode ser incluído na regressão auxiliar. Também será necessário eliminar duplicações se a regressão original já incluir o quadrado de uma variável (se, por exemplo,  $X_{2j} = X_{1j}^2$ ). Seja  $L$  o número de coeficientes remanescentes na regressão auxiliar, incluindo a constante, e seja  $R^2$  o seu coeficiente de determinação múltipla. Sob a hipótese de homocedasticia e pressupondo que o modelo original foi corretamente especificado,  $nR^2$  tem distribuição de qui-quadrado com  $L - 1$  graus de liberdade. Um valor de  $nR^2$  significativo (igual ou maior do que o valor crítico) indica a existência de heterocedasticia.

Um inconveniente do teste de White é que ele não é construtivo, no sentido de que não fornece indicação sobre a natureza da heterocedasticia, para possível aplicação do método de mínimos quadrados ponderados.

A generalidade do teste é uma vantagem, mas também pode fazer com que o teste de White seja pouco poderoso em comparação com um teste mais específico. Cabe

assinalar, ainda, que um resultado significativo pode ser consequência de um erro de especificação no modelo original.

Antes de encerrar esta seção, vejamos o procedimento a ser seguido se admitirmos que o desvio padrão do erro do modelo é proporcional a  $E(Y_j)$ , isto é,  $E(u_j^2) = [E(Y_j)]^2 \sigma^2$ . Uma vez que o valor de  $E(Y_j)$  é desconhecido, não é possível aplicar diretamente as fórmulas de mínimos quadrados ponderados. Então obtemos, inicialmente, os valores de  $Y$  estimados através do método de mínimos quadrados ordinários, que passamos a indicar por  $\hat{Y}_{*j}$  ( $j = 1, \dots, n$ ). Em seguida aplicamos as fórmulas de mínimos quadrados ponderados, substituindo  $v_j$  por  $\hat{v}_j = \hat{Y}_{*j}^2$ . Esse procedimento é sugerido por Theil (1971), que mostra que os estimadores obtidos são consistentes.

### 6.5. O estimador de White para variâncias quando há heterocedasticidade

Nesta sessão veremos como pode ser obtida uma estimativa consistente da matriz de variâncias e covariâncias de  $\mathbf{b}_*$  (o estimador de mínimos quadrados ordinários de  $\beta$ ) quando a matriz  $\mathbf{V}$  é desconhecida.

De acordo com (6.12), na presença de heterocedasticidade a matriz de variâncias e covariâncias das estimativas de mínimos quadrados ordinários dos parâmetros é

$$V(\mathbf{b}_*) = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \sigma^2 \quad (6.32)$$

ou

$$V(\mathbf{b}_*) = n(\mathbf{X}'\mathbf{X})^{-1} \mathbf{Q}(\mathbf{X}'\mathbf{X})^{-1}, \quad (6.33)$$

com

$$\mathbf{Q} = \frac{1}{n} \mathbf{X}'\mathbf{V}\mathbf{X}\sigma^2 \quad (6.34)$$

Como  $\mathbf{V}$  é uma matriz diagonal, se indicarmos as colunas de  $\mathbf{X}'$  por  $\mathbf{x}_j$  (e as linhas de  $\mathbf{X}$  por  $\mathbf{x}'_j$ ), temos

$$\mathbf{Q} = \frac{1}{n} \sum_{j=1}^n \mathbf{x}_j \mathbf{x}'_j v_j \sigma^2$$

ou

$$\mathbf{Q} = \frac{1}{n} \sum_{j=1}^n \mathbf{x}_j \mathbf{x}'_j \sigma_j^2 \quad (6.35)$$

$$\text{com } \sigma_j^2 = v_j \sigma^2 = E(u_j^2)$$

Sejam  $e_j$  os desvios da regressão de  $\mathbf{y}$  contra  $\mathbf{X}$ , estimada por mínimos quadrados ordinários. Se, em (6.35), substituirmos  $\sigma_j^2$  por  $e_j^2$ , obtemos

$$\mathbf{Q}_e = \frac{1}{n} \sum_{j=1}^n \mathbf{x}_j \mathbf{x}'_j e_j^2 \quad (6.36)$$

White (1980) demonstrou que  $\text{plim} \mathbf{Q}_e = \text{plim} \mathbf{Q}$ . Então, substituindo  $\mathbf{Q}$  por  $\mathbf{Q}_e$  em (6.33) obtemos o estimador consistente para heterocedasticidade (*heterocedasticity-consistent estimator*) ou estimador de White para  $V(\mathbf{b}_*)$ :

$$\hat{V}(\mathbf{b}_*) = (\mathbf{X}'\mathbf{X})^{-1} \sum_{j=1}^n \mathbf{x}_j \mathbf{x}'_j e_j^2 (\mathbf{X}'\mathbf{X})^{-1} \quad (6.37)$$

Cabe ressaltar que testes  $t$  ou  $F$  baseados nessas estimativas de variâncias são estritamente válidos apenas assintoticamente.

Conforme explica Greene (2000, p. 507), estudos baseados em simulação de dados mostram que o estimador de White tende a subestimar a variância correta, sendo aconselhável multiplicar o resultado de (6.37) por um fator  $n/(n-p)$  ou, tendo em vista que  $V(e_j) = (1-h_j)\sigma^2$  (ver exercício 4.27), substituir, na expressão (6.37),  $e_j^2$  por  $e_j^2/(1-h_j)$ , sendo  $h_j$  os elementos da diagonal principal de  $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ . Essas correções foram propostas por Davidson e Mackinnon (1993).

## Exercícios

6.1. Considere o modelo

$$Y_j = \alpha + \beta X_j + u_j,$$

$$\text{com } E(u_j) = 0, E(u_j^2) = X_j^2 \sigma^2 \text{ e } E(u_j u_h) = 0 \text{ para } h \neq j.$$

São dados os valores de  $X_j$  e  $Y_j$  observados em uma amostra aleatória com 5 observações:

$X_j$	$Y_j$
1	13
2	10
5	20
5	15
10	50

- a) Obtenha as estimativas lineares não-tendenciosas de variância mínima de  $\alpha$  e  $\beta$ .  
b) Teste, ao nível de significância de 10%, a hipótese  $H_0 : \beta = 1$  contra  $H_A : \beta > 1$ .

6.2. Estime os parâmetros do modelo

$$Y_j = \alpha + \beta X_j + u_j$$

com base nos seguintes dados

$X$	$Y$
-3	6,5
0	2,5
+3	0,5

Sabe-se que  $E(\mathbf{u}\mathbf{u}') = \begin{bmatrix} 0,5 & 0 & 0 \\ 0 & 0,25 & 0 \\ 0 & 0 & 0,5 \end{bmatrix} \sigma^2$ ,

onde  $\mathbf{u}$  representa o vetor dos erros.

Teste, ao nível de significância de 10%, a hipótese  $H_0 : \beta = 0$ .

6.3. Estime os parâmetros do modelo

$$Y_i = \alpha + \beta X_i + u_i$$

com base nos seguintes dados

$X$	$Y$
0	4
2	8
4	3
6	6
8	1

$$\text{Sabe-se que } E(\mathbf{u}\mathbf{u}') = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0,5 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0,5 \end{bmatrix} \sigma^2,$$

onde  $\mathbf{u}$  representa o vetor dos erros.

Teste, ao nível de significância de 10%, a hipótese  $H_0 : \beta = 0$ .

#### 6.4. Consideremos o modelo

$$Y_i = \beta X_i + u_i$$

com,  $E(u_i^2) = X_i \sigma^2$ ,  $E(u_i u_j) = 0$  para  $i \neq j$

Deduza, de acordo com o método de mínimos quadrados, a fórmula que dá a estimativa linear não-tendenciosa de variância mínima de  $\beta$ .

- 6.5. Considere a situação descrita no exercício 2.26, admitindo que a variância de  $Y | X$  é inversamente proporcional a  $X$  (Isso porque as empresas maiores, fazendo melhor controle dos custos, forneceram estimativas do custo médio mais precisas). Desenvolva, para esse caso, as fórmulas que dão as estimativas dos parâmetros da regressão de  $Y$  em relação a  $X$ .

- 6.6. É dada uma amostra com  $n$  observações das variáveis  $X_i$  e  $Y_i$ . Considere o seguinte modelo:

$$Y_i = \alpha + u_i \quad (i = 1, \dots, n)$$

onde  $u_i \sim N(0, \sigma^2 X_i)$ ,  $E(u_i u_j) = 0$  para  $i \neq j$  e os valores de  $X_i$  são fixos (não aleatórios). Determine a estimativa linear, não-tendenciosa e de variância mínima para  $\alpha$ , e sua variância.

- 6.7. Admite-se que as variáveis  $X$  e  $Y$  estão relacionadas de acordo com o modelo

$$Y_j = \beta X_j + u_j,$$

onde os  $u_j$  são erros aleatórios independentes, com média zero e variância  $W_j\sigma^2$ .

Note que há heterocedasticia e que o modelo não tem termo constante. É dada uma amostra de 3 valores de  $X_j$ ,  $W_j$  e  $Y_j$ :

$X_j$	$W_j$	$Y_j$
2	1	2
6	0,5	16
8	0,25	25

- a) Obtenha a estimativa linear não-tendenciosa de variância mínima para  $\beta$ .
- b) Teste, ao nível de significância de 1%, a hipótese de que  $\beta = 0$ .

#### 6.8. Considere o modelo

$$Y_i = \beta X_i + u_i \quad (i = 1, \dots, n)$$

onde os valores de  $X_i$  são fixados (não são aleatórios),  $E(u_i) = 0$ ,  $E(u_i u_j) = 0$  para  $i \neq j$  e, sendo  $\theta$  uma constante,  $E(u_i^2) = \theta [E(Y_i)]^2$ .

Obtenha o estimador de  $\beta$  de acordo com o método de mínimos quadrados ponderados para a seguinte amostra:

$X_i$	$Y_i$
1	2
2	6
3	6
4	12
5	10

#### 6.9. Temos uma amostra de 100 famílias de mesmo tamanho. Essas famílias foram classificadas em quatro estratos de renda familiar, como mostra a tabela a seguir:

Número de famílias ( $f_i$ ), renda familiar média ( $W_i$ ) e valor médio do logaritmo neperiano do consumo ( $\ln C_i$ ) de determinado produto em quatro estratos de renda familiar

Estrato	$f_i$	$W_i$	$\ln C_i$
1	40	1	0,9
2	30	2	2,3
3	20	5	2,6
4	10	10	2,3

Na última coluna dessa tabela está o valor do logaritmo neperiano do consumo de determinado produto, por família, para cada estrato.

Admite-se que o consumo desse produto varia com a renda familiar de acordo com o modelo

$$\ln C_i = \alpha + \frac{\beta}{W_i} + u_i,$$

onde  $u_i$  é um erro aleatório com valor esperado igual a zero, variância inversamente proporcional ao número de famílias do estrato e ausência de covariância entre erros de diferentes observações.

- a) Qual é a renda média das 100 famílias?
- b) Obtenha estimativas apropriadas de  $\alpha$  e  $\beta$ , levando em consideração o número de famílias de cada estrato.
- c) Qual é a estimativa da elasticidade-renda do consumo desse produto quando  $W = 2$ ? E quando  $W = 5$ ?
- d) Teste, ao nível de significância de 5%, a hipótese de que  $\beta$  é igual a zero.

6.10. Admite-se que  $Y$  é uma função de  $X$  com as seguintes características:

- a)  $E(Y) = 0$  quando  $X = 0$ ;
- b)  $E(Y)$  cresce linearmente com  $X$  para  $0 \leq X \leq 5$ ;
- c) Quando  $X = 5$  há uma redução da declividade, havendo um vértice na relação poligonal entre  $E(Y)$  e  $X$ ;
- d) Para  $5 < X < 10$  a relação entre  $E(Y)$  e  $X$  volta a ser linear, embora com declividade menor do que no primeiro intervalo;
- e) A variância do erro  $u = Y - E(Y)$  é  $\sigma^2$  quando  $0 \leq X \leq 5$  e é  $2\sigma^2$  quando  $5 < X < 10$ ; admite-se que os erros  $u_i$  têm média zero, não são correlacionados entre si e têm distribuição normal.

<i>X</i>	<i>Y</i>
3	10
4	10
6	19
7	17
8	17

É dada a seguinte amostra de valores de *X* e *Y*:

- a) Estabeleça um modelo de regressão linear para analisar como *Y* varia em função de *X* e determine as estimativas lineares não-tendenciosas de variância mínima dos seus dois parâmetros.
- b) Obtenha a estimativa de  $\sigma^2$ .
- c) Teste, ao nível de significância de 10%, a hipótese de que não há redução da declividade (teste unilateral).

### Respostas

6.1. a)  $a = 10$  e  $b = 2$

b)  $t = 0,777$ , não-significativo ( $t_0 = 1,638$ )

6.2.  $\hat{Y} = 3 - X$

$t = -4,243$ , não-significativo ( $t_0 = 6,314$ )

6.3.  $\hat{Y} = 6 - 0,5X$

$t = -1,245$ , não-significativo ( $t_0 = 2,353$ )

6.4.  $b = \frac{\sum Y_i}{\sum X_i} = \frac{\bar{Y}}{\bar{X}}$

6.5.  $a = \frac{\sum \frac{1}{X_i} \sum X_i Y_i - n \sum Y_i}{\sum X_i \sum \frac{1}{X_i} - n^2}$

$$b = \frac{\sum X_i \sum Y_i - n \sum X_i Y_i}{\sum X_i \sum \frac{1}{X_i} - n^2}$$

6.6.  $a = \frac{\sum \frac{Y_i}{X_i}}{\sum \frac{1}{X_i}}, \quad V(a) = \frac{\sigma^2}{\sum \frac{1}{X_i}}$

6.7. a)  $b = 3$

b)  $t = 14,61$ , significativo ( $t_0 = 9,925$ )

6.8.  $b = \frac{1}{n} \sum \frac{Y_i}{X_i} = 2,4$

6.9. a) 3

b)  $a = 3$  e  $b = -2$

c) 1 e 0,4

d)  $t = -4,209$ , não-significativo ( $t_0 = 4,303$ ).

6.10. a)  $E(Y) = \beta X + \gamma(X - 5)Z$ , com  $Z = 0$  para  $X \leq 5$  e  $Z = 1$  para  $X > 5$ . Estimativas dos parâmetros:  $b = 3$  e  $c = -2$

b)  $s^2 = 3,33$

c)  $t = -1,601$ , não-significativo (Região de rejeição:  $t \leq -1,638$ ).

## 7. MÍNIMOS QUADRADOS GENERALIZADOS E AUTOCORRELAÇÃO NOS RESÍDUOS

### 7.1. Mínimos quadrados generalizados

No capítulo anterior estudamos o procedimento a ser usado quando há heterocedasticidade, mas admitimos que as covariâncias entre erros de diferentes observações eram todas iguais a zero, fazendo com que a matriz de variâncias e covariâncias do vetor de erros  $\mathbf{u}$  fosse uma matriz diagonal. Neste capítulo vamos analisar o caso mais geral, em que se admite que haja covariâncias positivas ou negativas entre erros de diferentes observações.

Para isso, consideremos o modelo

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \quad (7.1)$$

com  $E(\mathbf{u}) = \mathbf{0}$  e  $E(\mathbf{u}\mathbf{u}') = \mathbf{V}\sigma^2$ ,

onde  $\mathbf{V}$  é uma matriz  $n \times n$ , simétrica e definida positiva. Então  $\mathbf{V}^{-1}$  também é simétrica e definida positiva e existe uma matriz  $\boldsymbol{\Lambda}$  tal que

$$\boldsymbol{\Lambda}'\boldsymbol{\Lambda} = \mathbf{V}^{-1} \quad (7.2)$$

Pré-multiplicando (7.1) por  $\boldsymbol{\Lambda}$ , obtemos

$$\boldsymbol{\Lambda}\mathbf{y} = \boldsymbol{\Lambda}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{u} \quad (7.3)$$

Para o modelo de regressão (7.3), de  $\boldsymbol{\Lambda}\mathbf{y}$  contra  $\boldsymbol{\Lambda}\mathbf{X}$ , com  $\boldsymbol{\varepsilon} = \boldsymbol{\Lambda}\mathbf{u}$ , temos

$$E(\boldsymbol{\varepsilon}) = \mathbf{0}$$

e

$$\begin{aligned} E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') &= E(\boldsymbol{\Lambda}\mathbf{u}\mathbf{u}'\boldsymbol{\Lambda}') = \\ &= \boldsymbol{\Lambda}\mathbf{V}\boldsymbol{\Lambda}'\sigma^2 = \\ &= \boldsymbol{\Lambda}(\boldsymbol{\Lambda}'\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'\sigma^2 = \\ &= \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1}(\boldsymbol{\Lambda}')^{-1}\boldsymbol{\Lambda}'\sigma^2 = \mathbf{I}\sigma^2 \end{aligned}$$

Podemos, portanto, aplicar ao modelo (7.3), de regressão de  $\boldsymbol{\Lambda}\mathbf{y}$  contra  $\boldsymbol{\Lambda}\mathbf{X}$ , as expressões conhecidas, obtendo

$$\mathbf{b} = (\mathbf{X}'\Lambda'\Lambda\mathbf{X})^{-1}\mathbf{X}'\Lambda'\Lambda\mathbf{y} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}, \quad (7.4)$$

que é o estimador linear não-tendencioso de variância mínima de  $\beta$ .

A matriz de variâncias e covariâncias de  $\mathbf{b}$  é

$$(\mathbf{X}'\Lambda'\Lambda\mathbf{X})^{-1}\sigma^2 = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\sigma^2 \quad (7.5)$$

A estimativa não-tendenciosa de  $\sigma^2$  é

$$s^2 = \frac{\mathbf{y}'\mathbf{V}^{-1}\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}}{n-p} \quad (7.6)$$

De maneira geral, as expressões relativas ao modelo básico de regressão múltipla são facilmente generalizadas, bastando substituir  $\mathbf{X}$  por  $\Lambda\mathbf{X}$  e  $\mathbf{y}$  por  $\Lambda\mathbf{y}$ , e lembrar que  $\Lambda'\Lambda = \mathbf{V}^{-1}$ .

Entretanto, no caso de os erros serem correlacionados, isto é, de  $\mathbf{V}$  ser não-diagonal, as expressões relativas à *predição* (estimação de novos valores) são afetadas de maneira mais complicada, já que é possível, então, que o erro da nova observação esteja correlacionado com os erros das observações da amostra.

Admitamos que seja calculado, incorretamente, o estimador de mínimos quadrados ordinários

$$\mathbf{b}_* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Considerando (7.1), obtemos

$$\begin{aligned} \mathbf{b}_* &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u}) = \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \end{aligned} \quad (7.7)$$

Então  $E(\mathbf{b}_*) = \beta$ , isto é, o estimador de mínimos quadrados ordinários é não-tendencioso.

Com  $\mathbf{V} \neq \mathbf{I}$ , entretanto,  $\mathbf{b}_* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$  é um estimador ineficiente, já que o estimador linear não-tendencioso de variância mínima é dado por (7.4).

De (7.7) obtemos

$$\mathbf{b}_* - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$$

Segue-se que

$$\begin{aligned} E[(\mathbf{b}_* - \boldsymbol{\beta})(\mathbf{b}_* - \boldsymbol{\beta})'] &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] = \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\sigma^2 \end{aligned} \quad (7.8)$$

Variâncias e covariâncias obtidas a partir da expressão  $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$  estariam, portanto, erradas se  $\mathbf{V} \neq \mathbf{I}$ .

Quem estivesse, erroneamente, aplicando mínimos quadrados ordinários utilizaria, como estimativa da variância residual, o valor dado por

$$s_*^2 = \frac{\mathbf{y}'\mathbf{y} - \mathbf{b}'_*\mathbf{X}'\mathbf{y}}{n-p} \quad (7.9)$$

Sabemos que  $\mathbf{y}'\mathbf{y} - \mathbf{b}'_*\mathbf{X}'\mathbf{y} = \mathbf{u}'\mathbf{M}\mathbf{u}$ , com  $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

Então

$$\begin{aligned} E(\mathbf{y}'\mathbf{y} - \mathbf{b}'_*\mathbf{X}'\mathbf{y}) &= E(\mathbf{u}'\mathbf{M}\mathbf{u}) = E[\text{tr}(\mathbf{u}'\mathbf{M}\mathbf{u})] = \\ &= E[\text{tr}(\mathbf{M}\mathbf{u}\mathbf{u}')] = \sigma^2 \text{tr}(\mathbf{M}\mathbf{V}) = \\ &= \sigma^2 \{\text{tr}(\mathbf{V}) - \text{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}]\} = \\ &= \sigma^2 \{\text{tr}(\mathbf{V}) - \text{tr}[\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]\} \end{aligned} \quad (7.10)$$

Uma vez que, com  $\mathbf{V} \neq \mathbf{I}$ , teremos, normalmente,  $\text{tr}(\mathbf{V}) - \text{tr}[\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \neq n-p$ , a expressão (7.9) dará uma estimativa tendenciosa da variância  $\sigma^2$ .

Note que o método de mínimos quadrados ponderados, estudado no capítulo anterior, é um caso particular de mínimos quadrados generalizados, em que a matriz  $\mathbf{V}$  é diagonal. Aliás, o modelo estudado no capítulo 4 também pode ser encarado como o caso particular em que  $\mathbf{V} = \mathbf{I}$ .

A matriz simétrica  $\mathbf{V}$  tem  $n(n+1)/2$  elementos possivelmente distintos que precisamos conhecer, para poder aplicar as fórmulas (7.4) e (7.6). Em aplicações práticas do método de mínimos quadrados generalizados, comumente há restrições que tornam viável a determinação da matriz  $\mathbf{V}$ . Na próxima seção vamos examinar uma aplicação de

mínimos quadrados generalizados na qual todos os elementos de  $\mathbf{V}$  são função de um único parâmetro.

## 7.2. Autocorrelação nos resíduos

Para ilustrar o problema da autocorrelação, consideremos o modelo

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} \quad \text{ou} \quad Y_t = \sum_{i=0}^k \beta_i X_{it} + u_t, \quad (t = 1, \dots, n) \quad (7.11)$$

com

$$u_t = \rho u_{t-1} + \varepsilon_t \quad (7.12)$$

Admitimos que  $-1 \leq \rho \leq 1$  e que  $\varepsilon_t$  é um ruído branco, isto é, uma série com as seguintes propriedades:

$$E(\varepsilon_t) = 0,$$

$$E(\varepsilon_t^2) = \sigma_\varepsilon^2,$$

$$E(\varepsilon_t \varepsilon_{t-h}) = 0 \quad \text{se } h \neq 0$$

Para que o modelo (7.11) tenha um termo constante devemos ter  $X_{0t} = 1$  para  $t = 1, \dots, n$ .

Aqui utilizamos a letra  $t$  para indicar o índice associado às diferentes observações porque o problema da autocorrelação dos resíduos surge, geralmente, quando estamos trabalhando com séries cronológicas de dados; então cada observação corresponde a um certo período de tempo (ano, mês ou semana, geralmente).

A relação (7.12) mostra que estamos admitindo que o erro da observação relativa a um período está correlacionado com o erro da observação anterior. Se  $\rho > 0$  dizemos que os erros estão positivamente autocorrelacionados e se  $\rho < 0$  dizemos que há autocorrelação negativa.

Se  $\rho = 0$  teremos, obviamente, o modelo de regressão linear múltipla estudado no capítulo 4, isto é, podemos aplicar mínimos quadrados ordinários.

Consideremos, inicialmente, o caso particular em que  $\rho = 1$ . Então

$$u_t = u_{t-1} + \varepsilon_t$$

ou

$$u_t - u_{t-1} = \varepsilon_t \quad (7.13)$$

De (7.11) podemos obter

$$Y_t - Y_{t-1} = \sum_{i=0}^k \beta_i (X_{it} - X_{i,t-1} + u_t - u_{t-1}) \quad (t = 2, \dots, n)$$

Considerando (7.13), segue-se que

$$Y_t - Y_{t-1} = \sum_{i=0}^k \beta_i (X_{it} - X_{i,t-1}) + \varepsilon_t \quad (t = 2, \dots, n) \quad (7.14)$$

Como os erros  $\varepsilon_t$  têm média zero, são não-correlacionados e homocedásticos, podemos aplicar, para o modelo (7.14), as fórmulas de mínimos quadrados ordinários. Note que no modelo (7.14) a variável dependente é  $\Delta Y_t = Y_t - Y_{t-1}$ , as variáveis explanatórias são  $\Delta X_{it} = X_{it} - X_{i,t-1}$  ( $i = 0, \dots, k$ ) e o número de observações se reduz a  $n - 1$ . Note que, se houver um termo constante no modelo original ( $X_{0t} = 1$  para todo  $t$ ), ele desaparece na equação (7.14). O modelo (7.14) terá um termo constante somente se uma das variáveis explanatórias do modelo original (7.11) for igual a  $t$ .

É fácil verificar que, para  $\rho = -1$ , obteríamos o modelo

$$Y_t + Y_{t-1} = \sum_{i=0}^k \beta_i (X_{it} + X_{i,t-1}) + \varepsilon_t$$

Consideremos, agora, que  $-1 < \rho < 1$ , ou seja,  $|\rho| < 1$ .

Utilizando (7.12) sucessivamente, obtemos

$$\begin{aligned} u_t &= \rho u_{t-1} + \varepsilon_t = \\ &= \rho(\rho u_{t-2} + \varepsilon_{t-1}) + \varepsilon_t = \\ &= \rho^2 u_{t-2} + \rho \varepsilon_{t-1} + \varepsilon_t = \\ &= \rho^3 u_{t-3} + \rho^2 \varepsilon_{t-2} + \rho \varepsilon_{t-1} + \varepsilon_t = \\ &= \dots = \\ &= \varepsilon_t + \rho \varepsilon_{t-1} + \rho^2 \varepsilon_{t-2} + \rho^3 \varepsilon_{t-3} + \dots = \\ &= \sum_{r=0}^{\infty} \rho^r \varepsilon_{t-r} \end{aligned}$$

Então

$$E(u_t) = 0, \quad (7.15)$$

$$\begin{aligned} E(u_t^2) &= (1 + \rho^2 + \rho^4 + \rho^6 + \dots) \sigma_\varepsilon^2 = \\ &= \frac{\sigma_\varepsilon^2}{1 - \rho^2} = \sigma_u^2 \end{aligned} \quad (7.16)$$

e, com  $h \geq 0$ ,

$$\begin{aligned}
E(u_t u_{t-h}) &= \\
&= E[(\varepsilon_t + \rho \varepsilon_{t-1} + \rho^2 \varepsilon_{t-2} + \dots)(\varepsilon_{t-h} + \rho \varepsilon_{t-h-1} + \rho^2 \varepsilon_{t-h-2} + \dots)] = \\
&= \rho^h \sigma_\varepsilon^2 + \rho^{h+2} \sigma_\varepsilon^2 + \rho^{h+4} \sigma_\varepsilon^2 + \dots = \\
&= \rho^h \frac{\sigma_\varepsilon^2}{1-\rho^2} = \rho^h \sigma_u^2
\end{aligned} \tag{7.17}$$

De (7.15), (7.16) e (7.17) concluímos que  $E(\mathbf{u}) = \mathbf{0}$  e  $E(\mathbf{u}\mathbf{u}') = \mathbf{V}\sigma_\varepsilon^2$ , com

$$\mathbf{V} = \frac{1}{1-\rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{bmatrix} \tag{7.18}$$

Verifica-se que

$$\mathbf{V}^{-1} = \begin{bmatrix} 1 & -\rho & 0 & \dots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \dots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \dots & -\rho & 1 \end{bmatrix}$$

e que  $\mathbf{V}^{-1} = \boldsymbol{\Lambda}' \boldsymbol{\Lambda}$  com

$$\boldsymbol{\Lambda} = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \dots & 0 & 0 \\ -\rho & 1 & 0 & \dots & 0 & 0 \\ 0 & -\rho & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & -\rho & 1 \end{bmatrix}$$

Sabemos que o método de mínimos quadrados generalizados corresponde a aplicar mínimos quadrados ordinários ao modelo transformado

$$\mathbf{Ay} = \boldsymbol{\Lambda}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{u} \tag{7.19}$$

Essa relação matricial representa um sistema de  $n$  equações. A primeira equação, para  $t = 1$ , é

$$\left(\sqrt{1-\rho^2}\right)Y_1 = \sum_{i=0}^k \beta_i \left(\sqrt{1-\rho^2}\right)X_{i1} + \left(\sqrt{1-\rho^2}\right)u_1 \quad (7.20)$$

e as  $n - 1$  equações restantes, para  $t = 2, \dots, n$ , são

$$Y_t - \rho Y_{t-1} = \sum_{i=0}^k \beta_i (X_{it} - \rho X_{i,t-1}) + u_t - \rho u_{t-1}$$

Uma vez que, de acordo com (7.12),  $u_t = \rho u_{t-1} + \varepsilon_t$ , temos

$$Y_t - \rho Y_{t-1} = \sum_{i=0}^k \beta_i (X_{it} - \rho X_{i,t-1}) + \varepsilon_t \quad (7.21)$$

Nos casos em que o valor de  $\rho$  é desconhecido, o que comumente acontece, podemos adotar ou o procedimento recomendado por Theil (1971, p. 254) ou o procedimento recomendado por Johnston (1972, p. 260-265). De acordo com Theil (1971) ajustamos, inicialmente, uma regressão de  $Y_i$  contra  $X_{it}$  ( $i = 0, \dots, k$ ) pelo método dos mínimos quadrados ordinários. A partir dos desvios dessa regressão, indicados por  $e_t$ , calculamos a estimativa de  $\rho$ :

$$\hat{\rho} = \frac{\frac{1}{n-1} \sum_{t=2}^n e_t e_{t-1}}{\frac{1}{n-p} \sum_{t=1}^n e_t^2} = \frac{\sum_{t=2}^n e_t e_{t-1}}{(n-1)(Q.M.Res.)}$$

De posse dessa estimativa, aplicamos o método dos mínimos quadrados generalizados, usando  $\hat{\rho}$  em lugar de  $\rho$ .

Johnston (1972) recomenda o procedimento descrito a seguir.

De (7.21) obtemos

$$Y_t = \rho Y_{t-1} + \sum_{i=0}^k \beta_i X_{it} - \sum_{i=0}^k \rho \beta_i X_{i,t-1} + \varepsilon_t \quad (t = 2, \dots, n)$$

Essa expressão sugere que uma estimativa de  $\rho$  seria a estimativa do coeficiente de regressão de  $Y_{t-1}$  numa regressão de  $Y_t$  contra  $Y_{t-1}$ , os  $X_{it}$  e os  $X_{i,t-1}$  ( $i = 1, \dots, k$ ), ajustada pelo método de mínimos quadrados ordinários.

A seguir, a estimativa de  $\rho$ , assim obtida, é usada para aplicar o método dos mínimos quadrados generalizados.

Vejamos, a seguir, o que ocorre se, erroneamente, aplicarmos mínimos quadrados ordinários ao modelo (7.11).

Uma vez que  $E(u_t) = 0$ , o estimador

$$\mathbf{b}_* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

é não-tendencioso; ele não é, entretanto, eficiente.

Para comparar a variância incorreta, obtida de  $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$ , com variância correta, dada, de acordo com (7.8), por  $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$ , consideremos o modelo

$$Y_t = \beta X_t + u_t$$

com os  $u_t$  sendo gerados pelo processo auto-regressivo de 1ª ordem

$$u_t = \rho u_{t-1} + \varepsilon_t$$

Neste caso  $(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{\sum X_t^2}$  e o estimador de mínimos quadrados ordinários para

$$\beta \text{ é } b_* = \frac{\sum X_t Y_t}{\sum X_t^2}$$

A variância incorreta ficaria

$$V_*(b_*) = (\mathbf{X}'\mathbf{X})^{-1}\sigma_u^2 = \frac{\sigma_u^2}{\sum_{t=1}^n X_t^2} \quad (7.22)$$

Lembrando (7.18), verificamos que a variância correta de  $b_*$  é

$$\begin{aligned} V(b_*) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\sigma_\varepsilon^2 = \\ &= \frac{\sigma_\varepsilon^2}{1-\rho^2} \cdot \frac{1}{\left(\sum_{t=1}^n X_t^2\right)^2} \left( \sum_{t=1}^n X_t^2 + 2\rho \sum_{t=2}^n X_t X_{t-1} + 2\rho^2 \sum_{t=3}^n X_t X_{t-2} + \dots + 2\rho^{n-1} X_n X_1 \right) \\ &= \frac{\sigma_u^2}{\sum_{t=1}^n X_t^2} \left( 1 + 2\rho \frac{\sum_{t=2}^n X_t X_{t-1}}{\sum_{t=1}^n X_t^2} + 2\rho^2 \frac{\sum_{t=3}^n X_t X_{t-2}}{\sum_{t=1}^n X_t^2} + \dots + 2\rho^{n-1} \frac{X_n X_1}{\sum_{t=1}^n X_t^2} \right) \end{aligned} \quad (7.23)$$

Se  $\rho$  é positivo e se, como é comum, os valores de  $X_t$  são positivamente autocorrelacionados, o valor da expressão entre parênteses em (7.23) será maior do que 1.

A comparação de (7.22) e (7.23) mostra que, neste caso, a expressão  $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$  leva a uma subestimativa da variância de  $b_*$  (o estimador de mínimos quadrados ordinários).

De acordo com (7.10) temos

$$\begin{aligned} E(\mathbf{y}'\mathbf{y} - \mathbf{b}_*\mathbf{X}'\mathbf{y}) &= \sigma_\varepsilon^2 \{ \text{tr}(\mathbf{V}) - \text{tr}[\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \} = \\ &= \frac{\sigma_\varepsilon^2}{1-\rho^2} \left[ n - \left( 1 + 2\rho \frac{\sum_{t=2}^n X_t X_{t-1}}{\sum_{t=1}^n X_t^2} + \dots + 2\rho^{n-1} \frac{X_n X_1}{\sum_{t=1}^n X_t^2} \right) \right] = \\ &= \sigma_u^2 \left[ n - \left( 1 + 2\rho \frac{\sum_{t=2}^n X_t X_{t-1}}{\sum_{t=1}^n X_t^2} + \dots \right) \right] \end{aligned} \quad (7.24)$$

O pesquisador que estivesse, erroneamente, aplicando mínimos quadrados ordinários estimaria  $\sigma_u^2$  por meio de

$$s_*^2 = \frac{\mathbf{y}'\mathbf{y} - \mathbf{b}_*\mathbf{X}'\mathbf{y}}{n-1} \quad (7.25)$$

Se  $\rho > 0$  e se os valores  $X_t$  são positivamente autocorrelacionados, a expressão entre parênteses em (7.24) é maior do que 1. Neste caso, (7.25) subestima  $\sigma_u^2$ .

Ao estimar a variância de  $b_*$  através de  $(\mathbf{X}'\mathbf{X})^{-1}s_*^2$ , com  $s_*^2 = (\mathbf{y}'\mathbf{y} - \mathbf{b}_*\mathbf{X}'\mathbf{y})/(n-1)$ , o pesquisador estaria, portanto, cometendo erro de subestimação por duas razões: primeiro, porque  $(\mathbf{X}'\mathbf{X})^{-1}\sigma_u^2$  subestima as variâncias verdadeiras e, segundo, porque (7.25) tende a subestimar  $\sigma_u^2$ .

### 7.3. O teste de Durbin-Watson

Para verificar a existência de autocorrelação nos resíduos da regressão utilizamos, freqüentemente, o teste de Durbin-Watson, baseado no valor

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2} \quad (7.26)$$

onde os  $e_t$  são os desvios da regressão ajustada pelo método de mínimos quadrados ordinários.

De (7.26) obtemos

$$d = \frac{\sum_{t=2}^n e_t^2}{\sum_{t=1}^n e_t^2} + \frac{\sum_{t=2}^n e_{t-1}^2}{\sum_{t=1}^n e_t^2} - 2 \frac{\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2}$$

Para  $n$  bastante grande temos  $d$  aproximadamente igual a  $1 + 1 - 2r = 2(1 - r)$ , onde  $r$  é o coeficiente de correlação entre  $e_t$  e  $e_{t-1}$ . Então, o valor de  $d$  varia entre zero (se  $r = 1$ ) e quatro (se  $r = -1$ ). Um valor de  $d$  perto de zero indica a existência de autocorrelação positiva nos erros e um valor de  $d$  próximo de 4 indica que os erros estão negativamente autocorrelacionados.

A distribuição de  $d$  depende do tamanho ( $n$ ) da amostra, do número ( $p$ ) de parâmetros estimados e também da matriz  $\mathbf{X}$ . Para tornar mais simples a maneira de efetuar o teste, foram tabelados, para diferentes valores de  $n$  e de  $p$ , aos níveis de significância de 1% e 5% (unilaterais), valores críticos  $d_L$  e  $d_U$  que permitem tomar uma decisão independentemente da matriz  $\mathbf{X}$ .

Para testar  $H_0 : \rho = 0$  contra  $H_A : \rho > 0$ , o valor de  $d$  é comparado com  $d_L$  e  $d_U$ . Se  $d < d_L$ , o resultado é significativo, rejeitando-se  $H_0$  em favor de  $H_A$ . Se  $d > d_U$ , o resultado é não-significativo, isto é, não se rejeita  $H_0$ . Se  $d_L < d < d_U$ , o resultado é inconclusivo.

Para testar  $H_0 : \rho = 0$  contra  $H_A : \rho < 0$ , o valor de comparado com  $4 - d_L$  e  $4 - d_U$ . O resultado é significativo se  $d > 4 - d_L$ , e é não significativo se  $d > 4 - d_U$ . Se  $4 - d_U < d < 4 - d_L$ , o resultado é inconclusivo. Obviamente, o resultado será o mesmo se compararmos  $4 - d$  com  $d_L$  e  $d_U$ .

Atualmente já existem programas de computador que fornecem a probabilidade caudal associada ao valor calculado do teste de Durbin-Watson. Neste caso basta comparar a probabilidade caudal com o nível de significância adotado para decidir se o resultado é ou não é significativo, evitando-se o problema do resultado inconclusivo.

A validade do teste depende de os erros terem distribuição normal com média zero e variância constante e das variáveis explanatórias não serem aleatórias. Devemos ressaltar

que não se deve aplicar o teste de Durbin-Watson quando há variáveis explanatórias aleatórias, como é o caso de modelos onde valores de  $Y$  defasados aparecem entre as variáveis explanatórias. Nestes casos, outros testes devem ser usados. (Ver Johnston, 1972, p. 309-313).

Draper e Smith (1966, p. 95-99) recomendam o uso de um teste não-paramétrico, baseado no agrupamento dos sinais dos desvios, para analisar os resíduos da regressão. Para uma apresentação do teste do agrupamento dos sinais, ou teste da ordenação casual, ver, também, Hoel (1968, p. 220-223) ou Hoffmann (2006, seção 13.4).

É interessante notar que podemos obter um teste significativo, indicando a existência de autocorrelação positiva nos resíduos, quando existe erro na especificação do modelo. Consideremos, por exemplo, que as variáveis  $Y$  e  $X$  estão relacionadas de acordo com o modelo  $Y_t = \alpha + \beta X_t + \gamma X_t^2 + u_t$ , onde os  $u_t$  são erros independentes com média zero e variância constante. Consideremos, ainda, que, dada uma amostra de  $n$  pares de valores dessas variáveis, foi ajustada uma regressão linear simples, isto é, em lugar da parábola, ajustamos uma reta. Se antes de estimar a reta de regressão, as observações tiverem sido ordenados conforme valores crescentes de  $X_t$ , é fácil perceber que os desvios tenderão a apresentar autocorrelação positiva.

### Exercícios

7.1. Admite-se que as variáveis  $X_t$  e  $Y_t$  estão relacionadas de acordo com o modelo

$$Y_t = \alpha + \beta X_t + u_t,$$

onde  $u_t = u_{t-1} + \varepsilon_t$ , sendo  $\varepsilon_t$  um ruído branco.

É dada uma amostra de 4 pares de valores das variáveis:

$X_t$	$Y_t$
5	10
7	14
11	30
17	46

- a) Estime  $\beta$ .
- b) Teste, ao nível de significância de 5%, a hipótese de que  $\beta = 0$ .

7.2. São dados os 4 pares de valores observados em uma amostra aleatória:

$X_t$	$Y_t$
2	19
7	50
12	75
7	40

Admite-se que  $Y$  e  $X$  estão relacionados de acordo com o modelo

$$Y_t = \alpha + \beta X_t + u_t,$$

onde  $u_t = 0,8u_{t-1} + \varepsilon_t$ ,  $E(\varepsilon_t) = 0$ ,  $E(\varepsilon_t^2) = \sigma_\varepsilon^2$  e  $E(\varepsilon_t \varepsilon_{t+h}) = 0$  para  $h \neq 0$ .

- Determine as estimativas lineares não-tendenciosas de variância mínima de  $\alpha$  e  $\beta$ .
- Obtenha a estimativa não-tendenciosa de  $\sigma_\varepsilon^2$ .
- Teste, ao nível de significância de 5%, a hipótese  $H_0 : \beta = 0$ .

7.3. A tabela ao lado mostra os 6 valores consecutivos de  $X_t$  e  $Y_t$  observados ao longo de 6 anos.

Admite-se que essas variáveis estão relacionadas de acordo com o modelo

$$Y_t = \alpha + \beta X_t + u_t,$$

com  $u_t = u_{t-1} + \varepsilon_t$ , sendo  $\varepsilon_t$  um ruído branco.

$X_t$	$Y_t$
4	12
2	4
7	28
4	20
4	16
6	20

- Obtenha a estimativa linear não-tendenciosa de variância mínima para  $\beta$ .
- Teste, ao nível de significância de 1%, a hipótese  $H_0 : \beta = 0$  contra a hipótese alternativa  $H_A : \beta > 0$ .

7.4. São dados os valores de  $Y_t$  para 4 semestres consecutivos:

Ano	Semestre	$Y_t$
1	1º	23
	2º	8
2	1º	31
	2º	10

Admite-se que essa variável tem variações cíclicas estacionais e que ela *não* tem tendência (crescimento ou decréscimo monotônico no tempo). Admite-se, também,

que o termo aleatório ( $u_t$ ) apresenta autocorrelação de 1<sup>a</sup> ordem com  $\rho = 0,5$ , isto é  $u_t = 0,5u_{t-1} + \varepsilon_t$ , sendo  $\varepsilon_t$  um ruído branco.

- a) considerando o modelo  $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ , onde  $\mathbf{y}$  é o vetor-coluna com 4 valores de  $Y_t$ , apresente uma matriz  $\mathbf{X}$  apropriada para captar a variação estacional de  $Y_t$  e obtenha a estimativa linear não-tendenciosa de variância mínima para o vetor  $\beta$ .
- b) Verifique se a variação estacional é estatisticamente significativa, adotando um nível de significância de 5%.

7.5. Dada uma série temporal de pares de valores  $X_t$ ,  $Y_t$ , com  $t = 1, 2, \dots, n$ , considere o modelo

$$Y_t = \alpha + \beta X_t + u_t,$$

onde  $u_t = \rho u_{t-2} + \varepsilon_t$ ,

$$0 < \rho < 1, E(\varepsilon_t) = 0, E(\varepsilon_t^2) = \sigma_\varepsilon^2 \text{ e}$$

$$E(\varepsilon_t \varepsilon_{t-h}) = 0 \text{ para } h \neq 0.$$

Note que o erro de uma observação está correlacionado com o erro da observação defasada de *dois* períodos. Isso pode ocorrer se os dados são semestrais e o valor de  $u_t$  em um semestre é afetado pelo valor do erro no mesmo semestre do ano anterior.

Sendo  $\mathbf{u}$  um vetor coluna cujos elementos são os  $u_t$  ( $t = 1, \dots, n$ ), determine:

- a)  $E(\mathbf{u})$
- b)  $E(\mathbf{u}\mathbf{u}')$ , em função de  $\rho$  e  $\sigma_\varepsilon^2$

7.6. Admite-se que as variáveis  $Y$ ,  $X_1$  e  $X_2$  estão relacionadas de acordo com o modelo

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + u_j \quad (j = 1, \dots, n)$$

com  $E(\mathbf{u}) = \mathbf{0}$  e  $E(\mathbf{u}\mathbf{u}') = \mathbf{V}\sigma^2$ , onde

$$\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \quad \text{e} \quad \mathbf{V} = \begin{bmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \vdots & \vdots & \vdots & & \vdots \\ \rho & \rho & \rho & \dots & 1 \end{bmatrix}$$

Os cálculos, entretanto, são feitos tendo em vista o modelo com as variáveis centradas:

$$y_j = \beta_1 x_{1j} + \beta_2 x_{2j} + u_j - \bar{u},$$

onde  $\bar{u} = \frac{1}{n} \sum u_j$ , ou, em notação matricial,

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} - \bar{\mathbf{u}}$$

onde

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ \vdots & \vdots \\ x_{1n} & x_{2n} \end{bmatrix} \quad \text{e} \quad \bar{\mathbf{u}} = \begin{bmatrix} \bar{u} \\ \bar{u} \\ \vdots \\ \bar{u} \end{bmatrix}$$

Demonstre que  $\mathbf{b} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$  é um estimador não-tendencioso e que a correspondente matriz de variâncias e covariâncias é  $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\sigma^2$ .

Verifique, preliminarmente, que

$$\mathbf{V}^{-1} = \begin{bmatrix} d & f & f & \dots & f \\ f & d & f & \dots & f \\ f & f & d & \dots & f \\ \vdots & \vdots & \vdots & & \vdots \\ f & f & f & \dots & d \end{bmatrix}$$

$$\text{com } d = \frac{1}{1-\rho} \left[ 1 - \frac{\rho}{1+\rho(n-1)} \right] \quad \text{e} \quad f = -\frac{1}{1-\rho} \cdot \frac{\rho}{1+\rho(n-1)}$$

7.7. São dados os valores de  $Y$ ,  $X_1$  e  $X_2$  em uma amostra com  $n = 4$  observações.

$Y_j$	$X_{1j}$	$X_{2j}$
-4	0	3
5	1	1
4	2	2
11	3	0

Admite-se que essas variáveis estejam relacionadas de acordo com o modelo

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + u_j \quad (j = 1, \dots, 4),$$

com  $E(\mathbf{u}) = \mathbf{0}$  e  $E(\mathbf{u}\mathbf{u}') = \mathbf{V}\sigma^2$ , onde

$$\mathbf{V} = \begin{bmatrix} 1 & 0,5 & 0,5 & 0,5 \\ 0,5 & 1 & 0,5 & 0,5 \\ 0,5 & 0,5 & 1 & 0,5 \\ 0,5 & 0,5 & 0,5 & 1 \end{bmatrix}$$

Obtenha as estimativas de  $\beta_1$ , de  $\beta_2$  e das respectivas variâncias e covariâncias, de acordo com o método de mínimos quadrados generalizados.

Verifique, preliminarmente, que

$$\mathbf{V}^{-1} = \begin{bmatrix} 1,6 & -0,4 & -0,4 & -0,4 \\ -0,4 & 1,6 & -0,4 & -0,4 \\ -0,4 & -0,4 & 1,6 & -0,4 \\ -0,4 & -0,4 & -0,4 & 1,6 \end{bmatrix}$$

### Respostas

7.1. a)  $b = 3$

b)  $t = 6,481$ , significativo ( $t_0 = 4,303$ )

7.2. a)  $a = 3$  e  $b = 6$

b)  $s_\varepsilon^2 = 25$

c)  $t = 9,650$ , não-significativo ( $t_0 = 4,303$ )

7.3. a)  $b = 4$

b)  $t = 6,481$ , significativo ( $t_0 = 3,747$ )

7.4. a)  $\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$  e  $\mathbf{b} = \begin{bmatrix} 27 \\ -19 \end{bmatrix}$  ou  $\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$  e  $\mathbf{b} = \begin{bmatrix} 27 \\ 8 \end{bmatrix}$

b)  $t = -6,504$ , significativo ( $t_0 = 4,303$ )

7.5. a)  $E(\mathbf{u}) = \mathbf{0}$

b) Temos  $E(u_t^2) = \frac{\sigma_\varepsilon^2}{1-\rho^2}$ ,

$$E(u_t u_{t-h}) = 0 \text{ se } h \text{ é um número ímpar e}$$

$$E(u_t u_{t-h}) = \frac{\sigma_\varepsilon^2 \rho^{0.5h}}{1-\rho^2} \text{ se } h \text{ é um número par positivo.}$$

Então

$$E(\mathbf{u}\mathbf{u}') = \frac{\sigma_e^2}{1-\rho^2} \begin{bmatrix} 1 & 0 & \rho & 0 & \rho^2 & 0 & \dots \\ 0 & 1 & 0 & \rho & 0 & \rho^2 & \dots \\ \rho & 0 & 1 & 0 & \rho & 0 & \dots \\ 0 & \rho & 0 & 1 & 0 & \rho & \dots \\ \rho^2 & 0 & \rho & 0 & 1 & 0 & \dots \\ 0 & \rho^2 & 0 & \rho & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \dots \end{bmatrix}$$

7.7.  $b_1 = 2$  e  $b_2 = -3$

$$\hat{V}(b_1) = 5/9, \quad \hat{V}(b_2) = 5/9 \quad \text{e} \quad \text{côv}(b_1, b_2) = 4/9$$

Observação: neste exercício temos um caso de equicorelação, isto é,  $E(u_i u_j) = \rho$ , com  $i \neq j$ , para todo  $i$ . Aplica-se, então, o teorema de McElroy (1967) (Ver Theil, 1971, p. 241-243). De acordo com esse teorema, em caso de equicorrelação, e desde que o modelo tenha um termo constante, as estimativas dos parâmetros obtidas pelo método de mínimos quadrados generalizados são iguais às estimativas obtidas aplicando mínimos quadrados ordinários. Com exceção da variância da estimativa do termo constante, serão, também, iguais, as estimativas das variâncias e covariâncias das estimativas dos parâmetros, obtidas pelos dois métodos. Compare os resultados obtidos aplicando mínimos quadrados generalizados com os obtidos aplicando mínimos quadrados ordinários (Ver resultados do exercício 4.11).

## 8. VARIÁVEIS INSTRUMENTAIS E ERROS NAS VARIÁVEIS EXPLANATÓRIAS

### 8.1. Introdução

Lembremos, inicialmente, as pressuposições do modelo “ordinário” de regressão linear. Temos

$$Y_j = \alpha + \sum_{i=1}^k \beta_i X_{ij} + u_j \quad (j = 1, \dots, n)$$

ou, em notação matricial,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \quad (8.1)$$

e admitimos que  $E(\mathbf{u}) = \mathbf{0}$ ,  $E(\mathbf{u}\mathbf{u}') = \mathbf{I}\sigma^2$  e  $\mathbf{X}$  é uma matriz com característica  $p = k + 1 < n$ , cujos elementos são valores fixados.

Lembremos também que, se as variáveis explanatórias ( $X_{ij}$ ) são aleatórias (com valores variando de amostra para amostra), o método de mínimos quadrados continua válido se

- a distribuição dos  $X_{ij}$  não depende de  $\alpha$ , dos  $\beta_i$  ( $i = 1, \dots, k$ ) ou de  $\sigma^2$ ;
- os erros ( $u_j$ ) são independentes dos valores de  $X_{ij}$ , isto é,

$$E[(X_{ij} - \mu_i)u_j] = 0 \quad (i = 1, \dots, k),$$

onde  $\mu_i = E(X_{ij})$

Nessas condições, aplicando as fórmulas de mínimos quadrados ordinários, obtemos estimativas não-tendenciosas dos parâmetros e, se os  $u_j$  têm distribuição normal, os intervalos de confiança, determinados da maneira indicada, e o procedimento para os testes de hipótese continuam válidos.

### 8.2. A consistência dos estimadores de mínimos quadrados ordinários

Na seção 1.10 foi analisado o conceito de estimador consistente e foram dadas algumas propriedades da convergência em probabilidade. Esse conceito e essas propriedades são facilmente estendidos para o caso de matrizes. Por definição, o limite em probabilidade de uma matriz, quando  $n$  tende a infinito, é igual à matriz constituída

pelos limites em probabilidade de cada um de seus elementos, desde que as dimensões da matriz não dependam de  $n$ . Assim, por exemplo, se

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad \text{plim } \mathbf{A} = \begin{bmatrix} \text{plim } a_{11} & \text{plim } a_{12} \\ \text{plim } a_{21} & \text{plim } a_{22} \end{bmatrix}$$

Conhecidos os limites em probabilidade de várias matrizes, não há dificuldade em determinar o limite em probabilidade de qualquer expressão envolvendo tais matrizes. Assim,

$$\text{plim } (\mathbf{AB}) = (\text{plim } \mathbf{A})(\text{plim } \mathbf{B})$$

e

$$(\text{plim } \mathbf{A}^{-1}) = (\text{plim } \mathbf{A})^{-1}$$

Consideremos o modelo (8.1), acrescentando a pressuposição de que

$$\lim_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{X}' \mathbf{X} \right) = \mathbf{Q}, \quad (8.2)$$

sendo  $\mathbf{Q}$  uma matriz não-singular.

Suponhamos, para exemplificar, que dispomos de uma amostra de dados experimentais, relativos a um ensaio. Suponhamos ainda que amostras maiores são obtidas repetindo-se, sucessivamente, esse ensaio básico. Seja  $\mathbf{X}_0$ , de dimensões  $n_0 \times p$ , a matriz de valores das variáveis independentes para um ensaio. Então, para  $m$  ensaios temos  $n = mn_0$ ,

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_0 \\ \mathbf{X}_0 \\ \vdots \\ \mathbf{X}_0 \end{bmatrix} \quad \text{e} \quad \mathbf{X}' \mathbf{X} = m \mathbf{X}'_0 \mathbf{X}_0$$

Segue-se que  $\frac{1}{n} \mathbf{X}' \mathbf{X} = \frac{1}{n_0} \mathbf{X}'_0 \mathbf{X}_0$ . Concluímos então que

$$\lim_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{X}' \mathbf{X} \right) = \frac{1}{n_0} \mathbf{X}'_0 \mathbf{X}_0 = \mathbf{Q}$$

Isto é, a pressuposição (8.2) é válida neste caso.

Passemos à demonstração de que os estimadores de mínimos quadrados ordinários são consistentes. De acordo com (4.10), temos

$$\mathbf{b} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} = \boldsymbol{\beta} + \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \frac{1}{n}\mathbf{X}'\mathbf{u}$$

Considerando a pressuposição (8.2), obtemos

$$\text{plim } \mathbf{b} = \boldsymbol{\beta} + \mathbf{Q}^{-1} \text{plim} \left( \frac{1}{n} \mathbf{X}'\mathbf{u} \right) \quad (8.3)$$

Se  $\mathbf{X}$  é uma matriz cujos elementos são fixados,  $E(\mathbf{u}) = \mathbf{0}$  e  $E(\mathbf{u}\mathbf{u}') = \mathbf{I}\sigma^2$ , temos

$$E\left(\frac{1}{n}\mathbf{X}'\mathbf{u}\right) = \mathbf{0} \quad (8.4)$$

e a matriz de variâncias e covariâncias do vetor  $\frac{1}{n}\mathbf{X}'\mathbf{u}$  é

$$V\left(\frac{1}{n}\mathbf{X}'\mathbf{u}\right) = \frac{1}{n^2} E(\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}) = \frac{\sigma^2}{n^2} \mathbf{X}'\mathbf{X}$$

Então, lembrando (8.2), obtemos

$$\lim_{n \rightarrow \infty} V\left(\frac{1}{n}\mathbf{X}'\mathbf{u}\right) = \left( \lim_{n \rightarrow \infty} \frac{\sigma^2}{n} \right) \left( \lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{X}'\mathbf{X} \right) = 0 \quad (8.5)$$

De (8.4) e (8.5) concluímos que  $\frac{1}{n}\mathbf{X}'\mathbf{u}$  converge em média quadrática para uma matriz nula e, consequentemente, que

$$\text{plim}\left(\frac{1}{n}\mathbf{X}'\mathbf{u}\right) = \mathbf{0} \quad (8.6)$$

De (8.3) e (8.6) obtemos

$$\text{plim } \mathbf{b} = \boldsymbol{\beta}, \quad (8.7)$$

Isto é,  $\mathbf{b}$  é um estimador consistente de  $\boldsymbol{\beta}$

Se as variáveis explanatórias forem aleatórias, e admitirmos que

$$\text{plim}\left(\frac{1}{n} \mathbf{X}' \mathbf{X}\right) = \mathbf{Q},$$

é fácil ver que a relação (8.3) continua válida, isto é,

$$\text{plim } \mathbf{b} = \boldsymbol{\beta} + \mathbf{Q}^{-1} \text{plim}\left(\frac{1}{n} \mathbf{X}' \mathbf{u}\right) \quad (8.8)$$

Então  $\mathbf{b}$  é um estimador consistente de  $\boldsymbol{\beta}$  se

$$\text{plim}\left(\frac{1}{n} \mathbf{X}' \mathbf{u}\right) = \mathbf{0} \quad (8.9)$$

Se pressupomos que  $X_{ij}$  ( $i = 1, \dots, k$ ) são variáveis aleatórias independentes de  $u_j$ , temos  $\text{cov}(X_{ij}, u_j) = 0$  e a condição (8.9) é obedecida. Alternativamente, podemos considerar (8.9) como pressuposição do modelo. Essa pressuposição pode ser expressa dizendo que as variáveis  $X_{ij}$  são *assintoticamente não-correlacionadas* com  $u_j$ .

### **8.3. A inconsistência dos estimadores de mínimos quadrados quando os erros estão assintoticamente correlacionados com uma ou mais das variáveis explanatórias**

Consideremos o modelo  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ , com  $E(\mathbf{u}) = \mathbf{0}$  e  $E(\mathbf{u}\mathbf{u}') = \mathbf{I}\sigma^2$ . Admitamos que as variáveis explanatórias são aleatórias e que

$$\text{plim}\left(\frac{1}{n} \mathbf{X}' \mathbf{X}\right) = \mathbf{Q}$$

Se houver pelo menos uma variável explanatória assintoticamente correlacionada com o erro, temos

$$\text{plim}\left(\frac{1}{n} \mathbf{X}' \mathbf{u}\right) \neq \mathbf{0}$$

Então, de acordo com (8.3), temos  $\text{plim } \mathbf{b} \neq \boldsymbol{\beta}$ , isto é, o estimador de mínimos quadrados é inconsistente.

É importante notar que a covariância assintótica não-nula entre  $X_{hj}$  e  $u_j$  não torna inconsistente apenas a estimativa de mínimos quadrados de  $\beta_h$ . Note-se, na expressão (8.3), que se houver um único elemento não-nulo no vetor  $\text{plim}\left(\frac{1}{n}\mathbf{X}'\mathbf{u}\right)$ , isso pode tornar inconsistentes *todos* os elementos de  $\mathbf{b}$ , devido à pré-multiplicação pela matriz  $\mathbf{Q}^{-1}$ .

#### 8.4. O uso de variáveis instrumentais para obter estimativas consistentes

Por simplicidade, consideremos o modelo

$$Y_j = \beta X_j + u_j, \quad (8.10)$$

pressupondo que  $E(u_j) = 0$ ,  $E(u_j^2) = \sigma^2$ ,  $E(u_j u_h) = 0$  se  $h \neq j$ ,

$$\lim_{n \rightarrow \infty} \left( \frac{1}{n} \sum X_j^2 \right) = Q \quad \text{e} \quad \text{plim} \left( \frac{1}{n} \sum X_j u_j \right) \neq 0.$$

O estimador de mínimos quadrados de  $\beta$  é

$$b = \frac{\sum X_j Y_j}{\sum X_j^2}$$

De acordo com (8.8), temos

$$\text{plim } b - \beta = Q^{-1} \text{ plim} \left( \frac{1}{n} \sum X_j u_j \right)$$

Se  $X_j$  e  $u_j$  apresentam correlação assintótica positiva,  $b$  tende a superestimar o valor de  $\beta$ , pois  $Q > 0$ , por tratar-se de uma soma de quadrados. Por outro lado, se  $X_j$  e  $u_j$  apresentam correlação assintótica negativa,  $b$  tende a subestimar o valor de  $\beta$ .

Admitamos que sejam conhecidos, para as observações da amostra, os valores de uma variável  $Z_j$  tal que

$$\text{plim} \left( \frac{1}{n} \sum Z_j X_j \right) \neq 0 \quad (8.11)$$

e

$$\text{plim} \left( \frac{1}{n} \sum Z_j u_j \right) = 0 \quad (8.12)$$

A condição (8.12) significa que  $Z_j$  e  $u_j$  devem ser assintoticamente não-correlacionados. Uma variável ( $Z_j$ ) com tais propriedades é denominada variável instrumental.

A seguir demonstraremos que

$$\hat{\beta} = \frac{\sum Z_j Y_j}{\sum Z_j X_j},$$

denominado estimador de variável instrumental, é consistente.

De (8.10) e (8.13) obtemos

$$\hat{\beta} = \beta + \frac{\frac{1}{n} \sum Z_j Y_j}{\frac{1}{n} \sum Z_j X_j}$$

Então

$$\text{plim } \hat{\beta} = \beta + \frac{\text{plim} \left( \frac{1}{n} \sum Z_j Y_j \right)}{\text{plim} \left( \frac{1}{n} \sum Z_j X_j \right)}$$

Considerando (8.11) e (8.12), segue-se que

$$\text{plim } \hat{\beta} = \beta$$

isto é,  $\hat{\beta}$  é um estimador consistente de  $\beta$ .

Generalizando, consideremos o modelo

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \quad (8.14)$$

pressupondo que  $E(\mathbf{u}) = \mathbf{0}$ ,  $E(\mathbf{u}\mathbf{u}') = \mathbf{I}\sigma^2$  e  $\text{plim} \left( \frac{1}{n} \mathbf{X}'\mathbf{u} \right) \neq \mathbf{0}$ .

Sabemos que nestas condições o estimador  $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$  é inconsistente.

Admitamos que, sendo  $p$  o número de parâmetros em  $\beta$ , é conhecida uma matriz  $\mathbf{Z}$ , de dimensões  $n \times p$ , com as seguintes propriedades:

$$a) \text{plim} \left( \frac{1}{n} \mathbf{Z}'\mathbf{u} \right) = \mathbf{0} \quad (8.15)$$

b) a matriz

$$\text{plim} \left( \frac{1}{n} \mathbf{Z}' \mathbf{X} \right) = \boldsymbol{\Omega} \quad (8.16)$$

existe e não é singular.

c) existe a matriz

$$\text{plim} \left( \frac{1}{n} \mathbf{Z}' \mathbf{Z} \right) = \boldsymbol{\Theta} \quad (8.17)$$

Nestas condições,  $\mathbf{Z}$  é denominada matriz de variáveis instrumentais. Se admitirmos que algumas das variáveis explanatórias ( $X_{ij}$ ) são assintoticamente não-correlacionadas com o erro, tais variáveis podem ser utilizadas como variáveis instrumentais, isto é, podem constituir colunas da matriz  $\mathbf{Z}$ . entretanto, será necessário dispor das observações de uma variável instrumental adicional para cada variável explanatória que admitirmos correlacionada com o erro.

O vetor das estimativas dos parâmetros, de acordo com o método das variáveis instrumentais, é

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}' \mathbf{X})^{-1} \mathbf{Z}' \mathbf{y} \quad (8.18)$$

Note que, se todas as variáveis explanatórias forem assintoticamente não-correlacionadas com o erro, a própria matriz  $\mathbf{X}$  pode ser usada como matriz de variáveis instrumentais e então o estimador (8.18) coincide com o estimador de mínimos quadrados.

A seguir demonstraremos que, obedecidas as condições (8.15) e (8.16), o estimador (8.18) é consistente. De (8.14) e (8.18) obtemos

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{Z}' \mathbf{X})^{-1} \mathbf{Z}' \mathbf{u}$$

Então

$$\text{plim} \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left[ \text{plim} \left( \frac{1}{n} \mathbf{Z}' \mathbf{X} \right) \right]^{-1} \text{plim} \left( \frac{1}{n} \mathbf{Z}' \mathbf{u} \right)$$

De acordo com (8.15) e (8.16) segue-se que  $\text{plim} \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ , c.q.d.

Pode-se demonstrar que a matriz de variâncias e covariâncias assintóticas de  $\hat{\boldsymbol{\beta}}$  é  $n^{-1} \boldsymbol{\Omega}^{-1} \boldsymbol{\Theta} \boldsymbol{\Omega}^{-1} \sigma^2$ . As correspondentes estimativas são dadas por,<sup>13</sup>

$$(\mathbf{Z}' \mathbf{X})^{-1} (\mathbf{Z}' \mathbf{Z}) (\mathbf{X}' \mathbf{Z})^{-1} s^2 \quad (8.19)$$

onde

---

<sup>13</sup> Ver Johnston (1972, p. 280).

$$s^2 = (\mathbf{y} - \hat{\mathbf{X}}\hat{\beta})'(\mathbf{y} - \hat{\mathbf{X}}\hat{\beta})/(n - p)$$

Note que (8.19) se transforma em  $(\mathbf{X}'\mathbf{X})^{-1}s^2$  se  $\mathbf{Z} = \mathbf{X}$ .

### 8.5. Regressão linear simples com as duas variáveis sujeitas a erros de medida

Admitamos que as variáveis  $\chi_j$  e  $\psi_j$  estão relacionadas de acordo com o modelo

$$\psi_j = \alpha + \beta\chi_j + \varepsilon_j, \quad (8.20)$$

onde os  $\varepsilon_j$  são erros aleatórios independentes de média zero e  $\text{plim}\left(\frac{1}{n}\sum\varepsilon_j^2\right) = \sigma_\varepsilon^2$

Admitamos também que as observações disponíveis incluem erros de medida nas duas variáveis. Por isso não temos os valores de  $\chi_j$  e  $\psi_j$ , mas apenas os valores de

$$X_j = \chi_j + v_j \quad (8.21)$$

e

$$Y_j = \psi_j + w_j, \quad (8.22)$$

onde  $v_j$  e  $w_j$  são erros aleatórios independentes com média zero,

$$\text{plim}\left(\frac{1}{n}\sum v_j^2\right) = \sigma_v^2 \quad \text{e} \quad \text{plim}\left(\frac{1}{n}\sum w_j^2\right) = \sigma_w^2.$$

De (8.20), (8.21) e (8.22), obtemos

$$Y_j - w_j = \alpha + \beta(X_j - v_j) + \varepsilon_j$$

ou

$$Y_j = \alpha + \beta X_j + \omega_j - \beta v_j \quad (8.23)$$

onde  $\omega_j = \varepsilon_j + w_j$ . Se  $\varepsilon_j$  e  $w_j$  forem assintoticamente não-correlacionados, temos

$$\text{plim}\left(\frac{1}{n} \sum \omega_j^2\right) = \sigma_\varepsilon^2 + \sigma_w^2$$

De (8.23), fazendo

$$u_j = \omega_j - \beta v_j \quad (8.24)$$

obtemos

$$Y_j = \alpha + \beta X_j + u_j \quad (8.25)$$

Admitimos que as variáveis  $v_j$  e  $\omega_j$  não são assintoticamente correlacionadas entre si, nem são assintoticamente correlacionadas com  $\chi_j$ , isto é,

$$\text{plim}\left(\frac{1}{n} \sum v_j \omega_j\right) = 0, \quad (8.26)$$

$$\text{plim}\left(\frac{1}{n} \sum v_j \chi_j\right) = 0, \quad (8.27)$$

e

$$\text{plim}\left(\frac{1}{n} \sum \omega_j \chi_j\right) = 0, \quad (8.28)$$

De (8.21) e (8.24) obtemos

$$\frac{1}{n} \sum X_j u_j = \frac{1}{n} \sum \omega_j \chi_j + \frac{1}{n} \sum v_j \omega_j - \frac{\beta}{n} \sum v_j \chi_j - \frac{\beta}{n} \sum v_j^2$$

Considerando (8.26), (8.27) e (8.28), segue-se que

$$\text{plim}\left(\frac{1}{n} \sum X_j u_j\right) = -\beta \text{plim}\left(\frac{1}{n} \sum v_j^2\right)$$

Como  $\text{plim}\left(\frac{1}{n} \sum v_j^2\right) = \sigma_v^2$ , obtemos

$$\text{plim} \left( \frac{1}{n} \sum X_j u_j \right) = -\beta \sigma_v^2 \quad (8.29)$$

Esse resultado mostra que se  $\beta > 0$ ,  $X_j$  e  $u_j$  apresentam correlação assintótica negativa. De acordo com o que foi visto na seção 8.3, sabemos que o estimador de mínimos quadrados não é consistente.

A seguir determinaremos o limite em probabilidade do estimador de mínimos quadrados,

$$b = \frac{\sum x_j Y_j}{\sum x_j^2} \quad (8.30)$$

Substituindo (8.25) em (8.30), obtemos

$$b = \beta + \frac{\sum x_j Y_j}{\sum x_j^2}$$

Então

$$\text{plim } b = \beta + \frac{\text{plim} \left( \frac{1}{n} \sum x_j Y_j \right)}{\text{plim} \left( \frac{1}{n} \sum x_j^2 \right)} \quad (8.31)$$

Pode-se demonstrar que, de acordo com (8.29),

$$\text{plim} \left( \frac{1}{n} \sum x_j Y_j \right) = -\beta \sigma_v^2 \quad (8.32)$$

e que

$$\text{plim} \left( \frac{1}{n} \sum x_j^2 \right) = \sigma_\chi^2 + \sigma_v^2, \quad (8.33)$$

onde  $\sigma_\chi^2 = \text{plim} \left[ \frac{1}{n} \sum (\chi - \bar{\chi})^2 \right]$

Substituindo (8.32) e (8.33) em (8.31), obtemos

$$\text{plim } b = \beta - \frac{\beta \sigma_v^2}{\sigma_x^2 + \sigma_v^2}$$

ou

$$\text{plim } b = \frac{\beta}{1 + \frac{\sigma_v^2}{\sigma_x^2}} \quad (8.34)$$

Portanto, o estimador de mínimos quadrados ordinários tende a subestimar o valor absoluto do parâmetro  $\beta$  quando há erros de medida na variável explanatória.

Nas próximas seções examinaremos diferentes métodos para obter uma estimativa consistente de  $\beta$  quando há erros de medida na variável explanatória.

## 8.6. O método da variável instrumental

Note que a inconsistência do estimador de mínimos quadrados, quando há erros de medida na variável explanatória, decorre da existência de correlação assintótica entre  $X_j$  e  $u_j$ , como mostra (8.29). Portanto, este é um caso especial do problema analisado na seção 8.3.

De acordo com o que vimos na seção 8.4, o método da variável insrumental nos fornece um estimador consistente de  $\beta$ , no modelo (8.25). Para isso precisamos dispor de uma variável instrumental  $Z_j$ . Podemos, então, constituir a matriz

$$\mathbf{Z} = \begin{bmatrix} 1 & Z_1 \\ 1 & Z_2 \\ \vdots & \vdots \\ 1 & Z_n \end{bmatrix}$$

Para o modelo (8.25) temos

$$\mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$$

De acordo com (8.15) e (8.16), a variável  $Z_j$  deve ser tal que:

a)  $\text{plim}\left(\frac{1}{n}\mathbf{Z}'\mathbf{u}\right) = \mathbf{0}$

b) a matriz

$$\text{plim}\left(\frac{1}{n}\mathbf{Z}'\mathbf{X}\right) = \boldsymbol{\Omega},$$

existe e é não-singular, o que implica que

$$\text{plim}\left(\frac{1}{n}\sum z_j x_j\right) \neq 0$$

Vimos que, nestas condições,  $\hat{\beta} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$  é um estimador consistente. Para o modelo (8.25) obtemos

$$\hat{\beta} = \frac{\sum z_j y_j}{\sum z_j x_j} \quad \text{e} \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

Entretanto, nem sempre dispomos de uma variável instrumental, obtida dos dados observados. Vejamos, então, uma forma especial de obter uma variável instrumental. Admitamos, sem perda de generalidade, que as observações estão ordenadas de acordo com os valores de  $X_j$ , em ordem crescente. Se  $n$  for par, estabelecemos  $Z_j = -1$  para as primeiras  $n/2$  observações e  $Z_j = 1$  para as  $n/2$  últimas observações. Se  $n$  for ímpar, estabelecemos  $Z_j = -1$  para  $j = 1, \dots, (n-1)/2$ ,  $Z_j = 0$  para  $j = (n+1)/2$  e  $Z_j = 1$  para  $j = (n+3)/2, \dots, n$ .

É fácil verificar que o estimador de  $\beta$  de acordo com (8.18) é, neste caso,

$$\hat{\beta} = \frac{\bar{Y}_2 - \bar{Y}_1}{\bar{X}_2 - \bar{X}_1} \tag{8.35}$$

onde  $\bar{X}_1$  e  $\bar{Y}_1$  são as médias dos valores de  $X_j$  e  $Y_j$ , respectivamente, para as primeiras  $n/2$  ou  $(n - 1)/2$  observações, e  $\bar{X}_2$  e  $\bar{Y}_2$  são as médias dos valores de  $X_j$  e  $Y_j$ , respectivamente, para as últimas  $n/2$  ou  $(n - 1)/2$  observações.

O estimador (8.35) foi proposto por Wald (1940). Uma vez que esse estimador é obtido a partir das médias de  $X$  e de  $Y$  para dois conjuntos de observações, o método é denominado método do agrupamento das observações.

Bartlett (1949) mostrou que a eficiência do estimador aumenta se dividirmos as observações, ordenadas de acordo com os valores crescentes de  $X_j$ , em 3 grupos, com aproximadamente o mesmo número de observações em cada um, e estabelecermos  $Z_j = -1$  para as observações do 1º grupo,  $Z_j = 0$  para as observações do 2º grupo e  $Z_j = 1$  para as observações do 3º grupo. Então

$$\hat{\beta} = \frac{\bar{Y}_3 - \bar{Y}_1}{\bar{X}_3 - \bar{X}_1}$$

onde  $\bar{X}_1$  e  $\bar{Y}_1$  são as médias dos valores de  $X_j$  e  $Y_j$ , respectivamente, para as observações do 1º grupo e  $\bar{X}_3$  e  $\bar{Y}_3$  são as médias para as observações do 3º grupo.

## 8.7. Outro método

Consideremos, novamente, o modelo de regressão linear simples analisado na seção 8.5.

O estimador de  $\beta$ , de acordo com o método de mínimos quadrados ordinários, é

$$b = \frac{\frac{1}{n} \sum x_j Y_j}{\frac{1}{n} \sum x_j^2} \quad (8.36)$$

Com  $Y_j = \alpha + \beta X_j + u_j$ , obtemos

$$\frac{1}{n} \sum x_j Y_j = \beta \frac{1}{n} \sum x_j^2 + \frac{1}{n} \sum x_j u_j$$

Lembrando (8.32) e (8.33), concluímos que o numerador de (8.36) converge em probabilidade para

$$\text{plim}\left(\frac{1}{n} \sum x_j Y_j\right) = \beta \sigma_x^2$$

e que o denominador de (8.36) converge em probabilidade para

$$\text{plim}\left(\frac{1}{n} \sum x^2\right) = \sigma_x^2 + \sigma_v^2$$

Segue-se que o estimador

$$\hat{\beta} = \frac{\frac{1}{n} \sum x_j Y_j}{\frac{1}{n} \sum x_j^2 - \sigma_v^2} \quad (8.37)$$

converge em probabilidade para  $\beta$ .

O método pode ser estendido para o caso de regressões múltiplas, como consta em Johnston (1972, p. 289-290).

É claro que, dada uma amostra de valores de  $X_j$  e  $Y_j$ , para que se possa calcular o valor do estimador consistente (8.37) é necessário conhecer a variância ( $\sigma_v^2$ ) do erro de medida da variável independente.

Esse método foi utilizado por Perez (1973) em um estudo da elasticidade-renda do consumo de alimentos em Piracicaba, S.P. A variável independente, nas regressões ajustadas, era o logaritmo da renda mensal *per capita*. Perez admitiu que os dados sobre a renda mensal *per capita* apresentaram erros de medida e que esse erro era, em 95% dos casos, menor do que 20% do valor dessa renda. Uma alteração de 20% em um valor equivale a multiplicá-lo ou dividi-lo por 1,2. Como a variável independente era o logaritmo da renda, segue-se que seu erro é, em 95% dos casos, inferior ao log 1,2, em módulo. Lembrando que o intervalo entre  $-2$  e  $+2$  comprehende pouco mais de 95% da distribuição normal reduzida, temos aproximadamente, que

$$2\sigma_v = \log 1,2$$

Donde

$$\sigma_v^2 = \left( \frac{1}{2} \log 1,2 \right)^2$$

### Exercícios

- 8.1. Admite-se que há uma relação linear entre os verdadeiros valores das variáveis  $X$  e  $Y$ , com coeficiente angular  $\beta$ . Sabe-se, entretanto, que os valores observados das duas variáveis têm erros de medida. É fornecida uma amostra de 5 pares de valores apresentados na tabela ao lado.

$X$	$Y$
11	25
15	33
19	37
11	21

- a) Obtenha uma estimativa consistente de  $\beta$  admitindo que a variância de medida em  $X$  é igual a 0,8.
- b) Neste caso o estimador  $b = \sum x_i Y_i / \sum x_i^2$ , com  $x_i = X_i - \bar{X}$ , tende a superestimar  $\beta$ ? Quais são, neste caso, as propriedades do estimador  $b$ ?

- 8.2. Na tabela ao lado estão os valores de  $Y$  e  $X$  obtidos de uma amostra com 8 observações.

$X$	$Y$
2	14
6	18
1	12
4	18
5	16
2	10
5	20

- a) Determine a equação de regressão de  $Y$  contra  $X$  de acordo com o método de mínimos quadrados ordinários.
- b) Admitindo que  $X$  inclui um erro de medida, determine as estimativas dos parâmetros da regressão de acordo com o método de Wald (1940), dividindo as observações em dois grupos.
- c) Determine as estimativas dos parâmetros admitindo que a variância de medida em  $X$  é igual a 0,5.
- d) Admitindo que haja erro de medida em  $X$ , quais são as propriedades estatísticas das estimativas do coeficiente regressão obtidas em (a), (b) e (c)? Se o estimador não for consistente, diga se ele tende a subestimar ou superestimar o valor verdadeiro.

- 8.3. É dada a seguinte amostra de valores das variáveis  $X$ ,  $Y$  e  $Z$ .

$X$	$Y$	$Z$
-4	-6	-1
-2	-4	-5
-2	0	1
0	-2	-3
0	4	-1
2	2	3
2	6	1
4	8	5

Sabe-se que as variáveis  $X$  e  $Y$  apresentam erros de medida e admite-se que  $X$  e  $Y$  estão relacionados de acordo com o modelo

$$Y_j = \alpha + \beta X_j + u_j$$

Devido aos erros de medida em  $X_j$ , o erro  $u_j$  é assintoticamente correlacionado com  $X_j$ .

Admite-se, também, que a variável  $Z_j$  não é assintoticamente correlacionada com  $u_j$ .

- Determine a estimativa de  $\beta$  de acordo com o método de mínimos quadrados ordinários.
- Se  $\beta > 0$ , qual o sinal da tendenciosidade assintótica do estimador de mínimos quadrados ordinários de  $\beta$ ?
- Determine a estimativa de  $\beta$  utilizando  $Z_j$  como variável instrumental.

Obtenha, também, a estimativa do respectivo desvio padrão.

- Determine a estimativa consistente de  $\beta$  admitindo que a variância do erro de medida de  $X$  é  $\sigma_v^2 = 1$ .

- 8.4. Considere o modelo

$$Y_j = \beta X_j + u_j,$$

onde os  $u_j$  são erros aleatórios indepedentes, identicamente distribuídos, com média zero e  $E(u_j^2) = \text{plim} \left( \frac{1}{n} \sum u_j^2 \right) = \sigma^2$ . Admite-se que os  $X_j$  são fixos e que

$$\lim_{n \rightarrow \infty} \left( \frac{1}{n} \sum X_j^2 \right) = Q.$$

a) Demonstre que o limite em probabilidade de

$$\hat{\beta} = \frac{\sum Y_j^2}{\sum X_j Y_j}$$

é

$$\beta + \frac{\sigma^2}{\beta Q}$$

### Respostas

8.1. a)  $\hat{\beta} = 2$

b) O estimador de mínimos quadrados ordinários (b) é inconsistente, com tendenciosidade assintótica *negativa* (tende a subestimar  $\beta$ ).

8.2. a)  $\hat{Y} = 9,273 + 1,636X$

b) ou c)  $\hat{Y} = 8 + 2X$

8.3. a) 5/3

- b) negativo
- c) 2
- d) 2

Observação: O fato dos itens (c) e (d) terem a mesma resposta é uma coincidência, devida ao caráter artificial dos dados desse exercício. Em geral os vários métodos de estimação darão resultados diferentes.

## 9. EQUAÇÕES SIMULTÂNEAS

### 9.1. Introdução

Para iniciar a análise dos problemas econométricos relacionados com sistemas de equações simultâneas, consideremos um modelo muito simples de determinação da renda nacional, constituído pelas duas equações dadas a seguir:

$$\begin{cases} C_t = \alpha + \beta Y_t + u_t \\ Y_t = C_t + Z_t \end{cases} \quad (9.1)$$

$$\quad (9.2)$$

onde  $C_t$  é a despesa de consumo no  $t$ -ésimo período ( $t = 1, 2, \dots, n$ ),  $Y_t$  é a renda,  $Z_t$  é a despesa de investimento e  $u_t$  é o erro aleatório,

$$\text{com } E(u_t) = 0, \quad E(u_t^2) = \text{plim} \left( \frac{1}{n} \sum u_t^2 \right) = \sigma^2 \quad \text{e} \quad E(u_t u_{t+h}) = 0 \text{ para } h \neq 0.$$

Uma vez que  $\beta$  representa a propensão marginal a consumir, devemos ter  $0 < \beta < 1$ .

A variável  $Z_t$  é *exógena*, ou seja, é uma variável cujos valores são determinados por um processo independente do descrito pelo sistema de equações em análise.

Pressupomos que as variáveis exógenas são não-correlacionadas com os erros do sistema de equações simultâneas.

No caso do exemplo que estamos considerando, pressupomos, pois, que

$$E\{[Z_t - E(Z_t)]u_t\} = 0$$

ou

$$E(Z_t u_t) = 0 \quad (9.3)$$

Neste exemplo  $C_t$  e  $Y_t$  são *variáveis endógenas*, isto é, são variáveis determinadas conjunta e simultaneamente, da maneira indicada pelo sistema de equações pela(s) variável(eis) exógena(s) e o(s) erro(s).

Diz-se que um sistema é *completo* quando o número de equações é igual ao número de variáveis endógenas, de maneira que o sistema pode ser resolvido para essas variáveis. A solução é chamada *forma reduzida* do sistema. Uma equação na forma reduzida mostra como uma variável endógena varia em função das variáveis exógenas e dos erros aleatórios. As equações originais são chamadas *equações estruturais*.

No caso do exemplo que estamos considerando, a partir das equações estruturais (9.1) e (9.2) obtemos as equações de forma reduzida:

$$C_t = \frac{\alpha}{1-\beta} + \frac{\beta}{1-\beta} Z_t + \frac{1}{1-\beta} u_t \quad (9.4)$$

e

$$Y_t = \frac{\alpha}{1-\beta} + \frac{1}{1-\beta} Z_t + \frac{1}{1-\beta} u_t \quad (9.5)$$

Vamos verificar, preliminarmente, que na equação (9.1) o erro  $u_t$  e a variável explanatória  $Y_t$  estão positivamente correlacionados.

De (9.5) segue-se que

$$E(Y_t) = \frac{\alpha}{1-\beta} + \frac{1}{1-\beta} Z_t$$

e que

$$Y_t - E(Y_t) = \frac{1}{1-\beta} u_t$$

Então

$$\text{cov}(Y_t, u_t) = E\{(Y_t - E(Y_t))u_t\} =$$

$$= E\left(\frac{u_t^2}{1-\beta}\right) = \frac{\sigma^2}{1-\beta}$$

Como  $0 < \beta < 1$ , temos que  $\text{cov}(Y_t, u_t) > 0$ , isto é, na relação (9.1) o resíduo e a variável explanatória estão positivamente correlacionados.

O estimador de mínimos quadrados ordinários de  $\beta$  em (9.1) é

$$b = \frac{\sum y_t C_t}{\sum y_t^2} \quad (9.6)$$

De acordo com o que foi visto na seção 8.3, a existência de covariância entre  $Y_t$  e  $u_t$  faz com que o estimador de mínimos quadrados seja inconsistente. Aplicando a relação (8.3) a esse caso, verificamos que  $b$  tende a superestimar o valor de  $\beta$ .

Determinemos o limite em probabilidade de  $b$ . Substituindo (9.1) em (9.6), obtemos

$$b = \beta + \frac{\sum y_t u_t}{\sum y_t^2}$$

Então

$$\text{plim } b = \beta + \frac{\text{plim} \left( \frac{1}{n} \sum y_t u_t \right)}{\text{plim} \left( \frac{1}{n} \sum y_t^2 \right)} \quad (9.7)$$

De (9.5) segue-se que

$$y_t = \frac{1}{1-\beta} z_t + \frac{1}{1-\beta} (u_t - \bar{u}) \quad (9.8)$$

Uma vez que  $u_t$  e  $Z_t$  são não-correlacionados, temos

$$\text{plim} \left( \frac{1}{n} \sum z_t u_t \right) = 0$$

Então, de (9.8), obtemos

$$\text{plim} \left( \frac{1}{n} \sum y_t u_t \right) = \frac{\sigma^2}{1-\beta} \quad (9.9)$$

e

$$\text{plim} \left( \frac{1}{n} \sum y_t^2 \right) = \frac{Q}{(1-\beta)^2} + \frac{\sigma^2}{(1-\beta)^2} \quad , \quad (9.10)$$

onde  $Q = \text{plim} \left( \frac{1}{n} \sum z_t^2 \right)$

Substituindo (9.9) e (9.10) em (9.7), obtemos

$$\text{plim } b = \beta + \frac{(1-\beta)\sigma^2}{Q+\sigma^2} \quad (9.11)$$

De acordo com o esperado, a tendenciosidade assintótica é positiva, uma vez que  $0 < \beta < 1$ .

## 9.2. Um exemplo numérico

Consideremos os valores de  $Z_t$ ,  $C_t$  e  $Y_t$  da tabela 9.1. Verifica-se que  $\bar{Z} = 20$ ,  $\bar{C} = 30$  e  $\bar{Y} = 150$ .

TABELA 9.1. Amostra de 6 valores de  $Z_t$ ,  $C_t$  e  $Y_t$

$Z_t$	$z_t = Z_t - \bar{Z}$	$C_t$	$c_t = C_t - \bar{C}$	$Y_t$	$y_t = Y_t - \bar{Y}$
16	-4	119	-11	135	-15
14	-6	126	-4	140	-10
18	-2	132	2	150	0
20	0	125	-5	145	-5
24	4	131	1	155	5
28	8	147	17	175	25

Para construir esse exemplo numérico escolhemos, inicialmente, os valores de  $Z_t$  apresentados na tabela 9.1. Depois obtivemos 6 valores de  $u_t$  com média zero e de maneira que a correlação entre  $Z_t$  e  $u_t$  fosse igual a zero. Em seguida estabelecemos que  $\alpha = 40$  e  $\beta = 0,6$  e utilizamos as equações (9.4) e (9.5) para calcular os valores de

$C_t$  e  $Y_t$ . Dessa maneira, nesse exemplo artificial, os métodos consistentes de estimação irão reproduzir os valores “verdadeiros” dos parâmetros, que são  $\alpha = 40$  e  $\beta = 0,6$ .

Da tabela 9.1 obtemos  $\sum c_t y_t = 660$  e  $\sum y_t^2 = 1000$ . Então o estimador de mínimos quadrados ordinários para  $\beta$  é

$$\frac{\sum c_t y_t}{\sum y_t^2} = \frac{660}{1000} = 0,66$$

que, como esperávamos, superestima o valor de  $\beta$ .

### 9.3. O estimador de variável instrumental

De acordo com o que foi visto na seção 8.4, o método das variáveis instrumentais pode ser usado para obter estimativas consistentes dos parâmetros quando uma variável explanatória está correlacionada com o erro do modelo.

De acordo com (8.18), utilizando  $Z_t$  como variável instrumental para  $Y_t$ , as estimativas dos parâmetros da equação estrutural (9.1) são dadas por

$$\hat{\beta} = \frac{\sum z_t c_t}{\sum z_t y_t} \quad (9.12)$$

e

$$\hat{\alpha} = \bar{C} - \hat{\beta} \bar{Y} \quad (9.13)$$

No caso do exemplo numérico apresentado temos  $\sum z_t c_t = 204$  e  $\sum z_t y_t = 340$ .

Então  $\hat{\beta} = \frac{204}{340} = 0,6$  e  $\hat{\alpha} = 130 - 0,6 \cdot 150 = 40$ .

### 9.4. Mínimos quadrados indiretos

A equação (9.4), da forma reduzida, pode ser escrita como segue:

$$C_t = A + BZ_t + \varepsilon_t$$

com

$$A = \frac{\alpha}{1 - \beta}, \quad (9.14)$$

$$B = \frac{\beta}{1 - \phi} \quad (9.15)$$

e

$$\varepsilon_t = \frac{1}{1-\beta} u_t$$

Como  $\text{cov}(u_t, Z_t) = 0$ , temos  $\text{cov}(\varepsilon_t, Z_t) = 0$ . Então o método de mínimos quadrados ordinários fornecerá estimadores lineares não-tendenciosos de variância mínima e consistentes dos parâmetros  $A$  e  $B$ . Tais estimadores são

$$\hat{B} = \frac{\sum z_t c_t}{\sum z_t^2} \quad (9.16)$$

e

$$\hat{A} = \bar{C} - \hat{B} \bar{Z} \quad (9.17)$$

De (9.14) e (9.15) obtemos

$$\beta = \frac{B}{1+B} \quad (9.18)$$

e

$$\alpha = (1-\beta)A \quad (9.19)$$

Substituindo, em (9.18) e (9.19),  $A$  e  $B$  pelas suas estimativas consistentes, dadas por (9.16) e (9.17), obtemos estimativas consistentes de  $\alpha$  e  $\beta$ . Apesar de  $\hat{A}$  e  $\hat{B}$  serem estimadores não-tendenciosos de  $A$  e  $B$ , as estimativas de  $\alpha$  e  $\beta$ , obtidas da maneira descrita, serão tendenciosas porque a não-tendenciosidade só é preservada em transformações lineares.

De acordo com (9.16) e (9.18) o estimador de  $\beta$  é

$$\frac{\frac{\sum z_t c_t}{\sum z_t^2}}{1 + \frac{\sum z_t c_t}{\sum z_t^2}} = \frac{\sum z_t c_t}{\sum z_t^2 + \sum z_t c_t}$$

Como  $Y_t = C_t + Z_t$ , temos  $\sum z_t y_t = \sum z_t c_t + \sum z_t^2$ . Então o estimador de  $\beta$  fica

$$\hat{\beta} = \frac{\sum z_t c_t}{\sum z_t y_t},$$

isto é, obtemos, através de mínimos quadrados indiretos, o mesmo estimador anteriormente obtido utilizando  $Z_t$  como variável instrumental.

De acordo com (9.17) e (9.19) o estimador de  $\alpha$ , obtido através de mínimos quadrados indiretos, é

$$\begin{aligned}
(1 - \hat{\beta})(\bar{C} - \hat{B}\bar{Z}) &= \bar{C} - \hat{\beta}\bar{C} - (1 - \hat{\beta})\hat{B}\bar{Z} = \\
&= \bar{C} - \hat{\beta}\bar{C} - (1 - \hat{\beta})\frac{\hat{\beta}}{1 - \hat{\beta}}\bar{Z} = \\
&= \bar{C} - \hat{\beta}(\bar{C} + \bar{Z}) = \\
&= \bar{C} - \hat{\beta}\bar{Y},
\end{aligned}$$

ou seja,

$$\hat{\alpha} = \bar{C} - \hat{\beta}\bar{Y},$$

que é o estimador anteriormente obtido pelo método da variável instrumental.

O aluno deve verificar que, se em lugar de utilizarmos a equação (9.4), utilizarmos a equação (9.5) da foram reduzida, os estimadores de  $\alpha$  e  $\beta$  obtidos através do método de mínimos quadrados indiretos serão os mesmos, isto é, serão iguais a

$$\hat{\beta} = \frac{\sum z_t c_t}{\sum z_t y_t}$$

e

$$\hat{\alpha} = \bar{C} - \hat{\beta}\bar{Y}$$

Aplicemos o método de mínimos quadrados indiretos ao exemplo numérico apresentado.

Para a equação de regressão, derivada de (9.4),

$$C_t = A + BZ_t + \varepsilon_t,$$

obtemos

$$\hat{B} = \frac{\sum z_t c_t}{\sum z_t^2} = \frac{204}{136} = 1,5 \quad \text{e} \quad \hat{A} = \bar{C} - \hat{B}\bar{Z} = 130 - 1,5 \cdot 20 = 100$$

De acordo com (9.18) temos

$$\hat{\beta} = \frac{\hat{B}}{1 + \hat{B}} = 0,6$$

De acordo com (9.19) temos

$$\hat{\alpha} = (1 - \hat{\beta})\hat{A} = 0,4 \cdot 100 = 40$$

Como exercício o aluno deve repetir os cálculos considerando a equação (9.5), em lugar da equação (9.4), verificando que as estimativas de  $\alpha$  e  $\beta$  obtidas serão as mesmas.

## 9.5. Mínimos quadrados em dois estágios

Para estimar os parâmetros da equação (9.1) através do método de mínimos quadrados em dois estágios fazemos, inicialmente, a regressão linear simples de  $Y_t$  (a variável endógena que aparece no segundo membro dessa equação) em relação a  $Z_t$  (a variável exógena do sistema). Sejam  $\hat{\pi}_1$  e  $\hat{\pi}_2$  as estimativas dos parâmetros dessa regressão. Sabemos que

$$\hat{\pi}_2 = \frac{\sum z_t y_t}{\sum z_t^2}$$

e

$$\hat{\pi}_1 = \bar{Y} - \hat{\pi}_2 \bar{Z}$$

Podemos então calcular os valores de

$$\hat{Y}_t = \hat{\pi}_1 + \hat{\pi}_2 Z_t \quad (9.20)$$

Em (9.1), substituindo  $Y_t$  por  $\hat{Y}_t + e_t$ , onde  $e_t$  são os desvios da regressão de  $Y_t$  contra  $Z_t$ , obtemos

$$C_t = \alpha + \beta(\hat{Y}_t + e_t) + u_t$$

ou

$$C_t = \alpha + \beta \hat{Y}_t + (u_t + \beta e_t) \quad (9.21)$$

De (9.3) e (9.20) obtemos

$$\text{cov}(\hat{Y}_t, u_t) = \hat{\pi}_2 \text{cov}(Z_t, u_t) = 0$$

Sabemos que  $\hat{Y}_t$  não está correlacionado com  $e_t$  (Esta é uma propriedade do método de mínimos quadrados). Concluímos, então, que, na equação (9.21),  $\hat{Y}_t$  não está

correlacionado com o erro  $(u_t + \beta e_t)$ . Podemos, portanto, aplicar a essa questão o método de mínimos quadrados ordinários, obtendo o seguinte estimador para  $\beta$ :

$$\frac{\sum c_t y_t}{\sum \hat{y}_t^2}$$

$$\text{Como } \hat{y}_t = \hat{\pi}_2 z_t \quad \text{e} \quad \hat{\pi}_2 = \frac{\sum z_t y_t}{\sum z_t^2},$$

esse estimador fica

$$\frac{\sum c_t \hat{y}_t}{\sum \hat{y}_t^2} = \frac{\hat{\pi}_2 \sum z_t c_t}{\hat{\pi}_2^2 \sum z_t^2} = \frac{\sum z_t c_t}{\hat{\pi}_2 \sum z_t^2} = \frac{\sum z_t c_t}{\sum z_t y_t}$$

Obtemos, pois, o mesmo estimador consistente já obtido pelo método de mínimos quadrados indiretos e pelo uso de  $Z_t$  como variável instrumental:

$$\hat{\beta} = \frac{\sum z_t c_t}{\sum z_t y_t}$$

Vamos aplicar o método de mínimos quadrados em dois estágios ao exemplo numérico apresentado.

No primeiro estágio, fazemos a regressão de  $Y_t$  contra  $Z_t$ .

Temos

$$\hat{\pi}_2 = \frac{\sum z_t y_t}{\sum z_t^2} = \frac{340}{136} = 2,5$$

e

$$\hat{\pi}_1 = \bar{Y} - 2,5 \bar{Z} = 150 - 2,5 \cdot 20 = 100$$

Em seguida, calculamos os valores de  $\hat{Y}_t = 100 + 2,5 Z_t$ , que são apresentados na tabela 9.2.

Tabela 9.2. Os valores de  $\hat{Y}_t$  obtidos no 1º estágio

$Z_t$	$\hat{Y}_t$	$\hat{y}_t$
16	140	-10
14	135	-15
18	145	-5
20	150	0
24	160	10
28	170	20

Temos  $\sum \hat{Y}_t = \sum Y_t = 900$ ,

$$\frac{\sum \hat{Y}_t}{n} = \bar{Y} = 150$$

e

$$\sum \hat{y}_t^2 = 850$$

No segundo estágio fazemos a regressão de  $C_t$  contra  $\hat{Y}_t$ , obtendo

$$\hat{\beta} = \frac{\sum c_t \hat{y}_t}{\sum \hat{y}_t^2} = \frac{510}{850} = 0,6$$

e

$$\hat{\alpha} = \bar{C} - \hat{\beta} \bar{Y} = 130 - 0,6 \cdot 150 = 40$$

## 9.6. Variáveis conjuntamente determinadas e variáveis predeterminadas

Já vimos a distinção entre variáveis endógenas e variáveis exógenas. Se, no sistema de equações simultâneas, além das variáveis correntes (referentes ao período  $t$ ), há variáveis defasadas (referentes ao período  $t - h$ , com  $h > 0$ ), podemos distinguir os seguintes tipos de variáveis:

- a) variáveis endógenas correntes
- b) variáveis exógenas correntes
- c) variáveis endógenas defasadas
- d) variáveis exógenas defasadas

As variáveis endógenas correntes são denominadas *variáveis conjuntamente determinadas*. As outras são chamadas *variáveis predeterminadas*.

No caso mais geral, então, a forma reduzida descreve o comportamento das variáveis conjuntamente determinadas em termos das variáveis predeterminadas e dos erros.

## 9.7. Notação geral

Consideremos um sistema de equações simultâneas completo, com  $L$  equações e  $L$  variáveis endógenas (ou  $L$  variáveis conjuntamente determinadas, já que não consideraremos, no que se segue, a possibilidade de existirem variáveis endógenas defasadas).

A  $j$ -ésima equação estrutural poderia ser representada da seguinte maneira:

$$\sum_{h=1}^L \gamma_{hj} Y_{ht} + \sum_{k=1}^K \beta_{kj} X_{kt} = \varepsilon_{jt} \quad (j = 1, \dots, L \text{ e } t = 1, \dots, n),$$

onde  $Y_{ht}$  é a  $t$ -ésima observação da  $h$ -ésima variável endógena e  $X_{kt}$  é a  $t$ -ésima observação da  $k$ -ésima variável predeterminada ( $k = 1, \dots, K$ , sendo  $K$  o número de variáveis predeterminadas do sistema de equações em análise).

O modelo especificará, normalmente, que parte dos coeficientes  $\gamma_{hj}$  ( $h = 1, \dots, L$ ) e  $\beta_{kj}$  ( $k = 1, \dots, K$ ) são iguais a zero, indicando as variáveis excluídas das  $j$ -ésima equação.

Em notação matricial temos

$$\mathbf{Y}\Gamma + \mathbf{XB} = \mathbf{E} \quad (9.22)$$

onde  $\mathbf{Y}$  tem dimensões  $n \times L$ ,  $\Gamma$  tem dimensões  $L \times L$ ,  $\mathbf{X}$  tem dimensões  $n \times K$ ,  $\mathbf{B}$  tem dimensões  $K \times L$  e  $\mathbf{E}$  tem dimensões  $n \times L$ .

Cada coluna da matriz  $\mathbf{E}$  é constituída de valores dos erros para uma das equações do sistema:

$$\mathbf{E} = [\boldsymbol{\varepsilon}_1 \quad \boldsymbol{\varepsilon}_2 \quad \dots \quad \boldsymbol{\varepsilon}_L]$$

Admite-se que

$$E(\boldsymbol{\varepsilon}_j \boldsymbol{\varepsilon}'_j) = \mathbf{I} \sigma_j^2$$

e que

$$E(\boldsymbol{\varepsilon}_j \boldsymbol{\varepsilon}'_h) = \mathbf{I} \sigma_{jh},$$

isto é, o erro de uma equação em determinado período é não-correlacionado com os erros referentes a outros períodos, mas pode ser correlacionado com os erros das outras equações no mesmo período.

De (9.22), que representa as equações estruturais, obtemos a forma reduzida

$$\mathbf{Y} = -\mathbf{X}\mathbf{B}\boldsymbol{\Gamma}^{-1} + \mathbf{E}\boldsymbol{\Gamma}^{-1} \quad (9.23)$$

Essa notação mais geral não é cômoda para a análise dos métodos de estimação que vamos considerar a seguir.

Em princípio cada uma das equações estruturais pode ser colocada na forma

$$\mathbf{y}_j = \mathbf{Z}_j \boldsymbol{\delta}_j + \boldsymbol{\varepsilon}_j \quad (9.24)$$

onde:

- a)  $\mathbf{y}_j$  é um vetor-coluna com os  $n$  valores da variável endógena que aparece no primeiro membro da  $j$ -ésima equação estrutural;
- b)  $\mathbf{Z}_j$  é a matriz  $n \times N_j$ , onde  $N_j$  é o número de variáveis no segundo membro da  $j$ -ésima equação, incluindo variáveis endógenas e exógenas (Se a equação tem um termo constante, vamos considerar que existe uma variável exógena fictícia cujas valores são todos iguais a 1, associada a esse termo);
- c)  $\boldsymbol{\delta}_j$  é um vetor com os  $N_j$  parâmetros a serem estimados e
- d)  $\boldsymbol{\varepsilon}_j$  é um vetor-coluna com os  $n$  valores do erro da  $j$ -ésima equação.

O estimador de mínimos quadrados ordinários para  $\boldsymbol{\delta}_j$  é

$$(\mathbf{Z}'_j \mathbf{Z}_j)^{-1} \mathbf{Z}'_j \mathbf{y}_j$$

Entretanto, se  $\mathbf{Z}_j$  inclui variáveis conjuntamente determinadas (variáveis endógenas correntes), esse estimador não é consistente.

## 9.8. Variáveis instrumentais

Para relembrar a idéia do uso de variáveis instrumentais na obtenção de estimativas consistentes dos parâmetros, consideremos o modelo básico de regressão linear múltipla, com uma única equação,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

Premultiplicando por  $\mathbf{X}'$  obtemos

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \mathbf{X}'\mathbf{u}$$

O valor de  $\mathbf{X}'\mathbf{u}$  é desconhecido, mas se  $\text{plim}\left(\frac{1}{n}\mathbf{X}'\mathbf{u}\right) = \mathbf{0}$  e  $\text{plim}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)$  é uma matriz não-singular, podemos, para uma amostra suficientemente grande, desprezar  $\mathbf{X}'\mathbf{u}$  obtendo  $\mathbf{X}'\mathbf{y} = \mathbf{X}'\hat{\mathbf{X}}\hat{\beta}$ .

Segue-se que  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$  é um estimador consistente de  $\beta$ .

Se  $\text{plim}\left(\frac{1}{n}\mathbf{X}'\mathbf{u}\right) \neq \mathbf{0}$ , mas dispusermos de uma matriz de variáveis instrumentais  $\mathbf{W}$ , com  $n$  linhas, sendo que  $\text{plim}\left(\frac{1}{n}\mathbf{W}'\mathbf{u}\right) = \mathbf{0}$  e  $\text{plim}\left(\frac{1}{n}\mathbf{W}'\mathbf{X}\right)$  é uma matriz não-singular, obteríamos, analogamente, o sistema de equações

$$\mathbf{W}'\mathbf{y} = \mathbf{W}'\mathbf{X}\hat{\beta},$$

que poderá ser resolvido para  $\hat{\beta}$ , se existir  $(\mathbf{W}'\mathbf{X})^{-1}$ . Obter-se-á, dessa maneira, o estimador consistente  $\hat{\beta} = (\mathbf{W}'\mathbf{X})^{-1}\mathbf{W}'\mathbf{y}$ .

Vamos aplicar o método das variáveis instrumentais para estimar os parâmetros de (9.24).

A matriz  $\mathbf{X}$ , com todas as  $K$  variáveis predeterminadas do sistema, é uma matriz de variáveis instrumentais apropriada se pudermos admitir que<sup>14</sup>

$$\text{plim}\left(\frac{1}{n}\mathbf{X}'\boldsymbol{\varepsilon}_j\right) = \mathbf{0} \quad (9.25)$$

De (9.24), considerando  $\mathbf{X}$  como matriz de variáveis instrumentais, obtemos

$$\mathbf{X}'\mathbf{y}_j = \mathbf{X}'\mathbf{Z}_j\hat{\delta}_j + \mathbf{X}'\boldsymbol{\varepsilon}_j \quad (9.26)$$

e

$$\mathbf{X}'\mathbf{y}_j = \mathbf{X}'\mathbf{Z}_j\hat{\delta}_j \quad (9.27)$$

Se  $N_j = K$ , a matriz  $\mathbf{X}'\mathbf{Z}_j$  é uma matriz quadrada de dimensões  $K \times K$ .

Se, além disso,  $\mathbf{X}'\mathbf{Z}_j$  for não-singular, existe  $(\mathbf{X}'\mathbf{Z}_j)^{-1}$  e de (9.27) obtemos

$$\hat{\delta}_j = (\mathbf{X}'\mathbf{Z}_j)^{-1}\mathbf{X}'\mathbf{y}_j \quad (9.28)$$

---

<sup>14</sup> Se  $\mathbf{X}$  incluir variáveis endógenas defasadas, só podemos admitir a validade de (9.25) se os erros não forem autocorrelacionados. Uma exposição didática do problema de ajuste de equações simultâneas quando há variáveis endógenas defasadas e autocorrelação nos erros pode ser encontrada em Kelejian e Oates (1978, p. 321-325).

## 9.9. Identificação

A relação (9.27) representa um sistema de  $K$  equações com  $N_j$  incógnitas (as estimativas dos  $N_j$  parâmetros, cada um correspondendo a uma das  $N_j$  variáveis que aparecem no segundo membro da  $j$ -ésima equação do sistema de equações simultâneas). Para que exista uma solução, isto é, para que a equação (9.24) seja identificável, é necessário que o número de equações em (9.27) seja pelo menos igual ao número de incógnitas, ou seja, devemos ter

$$K \geq N_j \quad (9.29)$$

Portanto, se as equações do modelo estão na forma (9.24), para que a  $j$ -ésima equação estrutural seja identificável, a condição necessária, mas não suficiente, é que o número ( $K$ ) de variáveis predeterminadas do sistema de equações simultâneas em estudo seja pelo menos igual ao número ( $N_j$ ) de variáveis que aparecem no segundo membro da  $j$ -ésima equação.

Se representarmos por  $K_j$  e  $L_j$ , respectivamente, o número de variáveis predeterminadas e o número de variáveis endógenas que aparecem no segundo membro da  $j$ -ésima equação, temos

$$N_j = K_j + L_j$$

Então, a condição  $K \geq N_j$  fica

$$K \geq K_j + L_j$$

ou

$$K - K_j \geq L_j \quad (9.30)$$

Portanto, para que a  $j$ -ésima equação seja identificável, é necessário que o número de variáveis predeterminadas do sistema que não aparecem nessa equação seja igual ou maior do que o número de variáveis endógenas correntes no segundo membro da equação.

Se  $K < N_j$  (o que implica  $K - K_j < L_j$ ) o sistema (9.27) não tem solução. Dizemos, então, que a  $j$ -ésima equação estrutural não é identificável, isto é, há subidentificação.

Se o sistema de equações (9.27) for possível e determinado, isto é, se existir apenas uma solução para esse sistema, dizemos que a  $j$ -ésima equação estrutural é exatamente identificável. Isso ocorre, por exemplo, se  $N_j = K$  e  $\mathbf{X}'\mathbf{Z}_j$  for uma matriz não-singular. Nesse caso, a solução do sistema é dada por (9.28).

Se o número de equações independentes em (9.27) for maior do que o número de incógnitas ( $N_j$ ), temos superidentificação.

Note que tanto no caso de identificação exata como no caso de superidentificação dizemos que a equação é identificável.

Antes de mostrar como podemos obter estimativas dos parâmetros em caso de superidentificação, vamos examinar o conceito de identificação sob outros ângulos.

Para ilustrar o problema de subidentificação, consideremos o sistema (não identificável) constituído pelas equações de oferta e demanda de um produto:

$$\begin{cases} q_t = \beta_0 + \beta_1 p_t + u_t & (\text{demanda}) \\ q_t = \gamma_0 + \gamma_1 p_t + \varepsilon_t & (\text{oferta}) \end{cases}$$

onde  $q_t$  e  $p_t$  são, respectivamente, a quantidade transacionada e o preço do produto no  $t$ -ésimo período. Aqui, o fato de utilizarmos letras minúsculas não significa que as variáveis sejam centradas.

Neste caso, para qualquer uma das 2 equações, temos  $N_j=2$  (considerando uma variável fictícia  $x_0 = 1$ , associada aos parâmetros  $\beta_0$  e  $\gamma_0$ ). Como  $K = 1$  (só há uma variável predeterminada que é  $x_0 = 1$ ), temos um caso de subidentificação. Nenhuma das duas equações é identificável já que para ambas  $K < N_j$ . A relação (9.27), isto é,

$$\mathbf{X}'\mathbf{y}_j = \mathbf{X}'\mathbf{Z}\hat{\boldsymbol{\delta}}_j,$$

consiste, tanto no caso da demanda ( $j = 1$ ) como no caso da oferta ( $j = 2$ ), de uma única equação com 2 incógnitas ( $\hat{\beta}_0$  e  $\hat{\beta}_1$  no caso da demanda e  $\hat{\gamma}_0$  e  $\hat{\gamma}_1$  no caso da oferta).

Podemos mostrar, graficamente, o fato de, nestas condições, as equações de demanda e oferta não serem identificáveis. Na figura 9.1 representamos a função de demanda e a função de oferta. Os pontos correspondentes aos pares de valores  $(p_j, q_j)$  estarão em redor do ponto de intersecção das 2 funções (o ponto de equilíbrio). Tais pontos não nos permitem estimar nenhuma das duas funções.

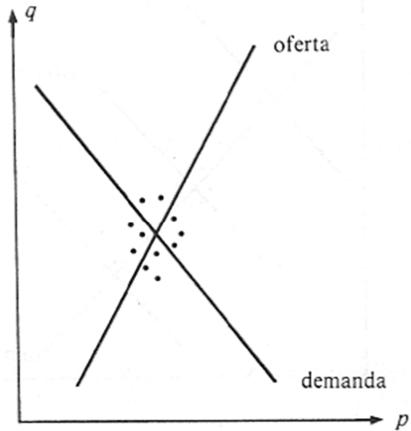


Figura 9.1

Consideremos, agora, que a oferta depende, também, do preço ( $x_{1t}$ ) de uma matéria-prima, e que este preço é uma variável exógena que não afeta a demanda

$$\begin{cases} q_t = \beta_0 + \beta_1 p_t + u_t & (\text{demanda}) \\ q_t = \gamma_0 + \gamma_1 p_t + \gamma_2 x_{1t} + \varepsilon_t & (\text{oferta}) \end{cases} \quad (9.31)$$

Esse sistema tem  $K = 2$  variáveis exógenas (a variável fictícia  $x_0 = 1$  e  $x_1$ ).

Para a equação de oferta temos  $N_2 = 3 > K$ . Portanto, essa equação não é identificável.

A equação de demanda, entretanto, é exatamente identificável, já que  $N_1 = 2 = K$ .

Para ilustrar graficamente o problema, consideremos um sistema de dois eixos cartesianos ortogonais, onde são lidos os valores de  $p_t$  e  $q_t$ . A função de oferta estará em diferentes posições, dependendo do valor de  $x_{1t}$ . Com isso, os pontos correspondentes aos pares de valores  $(p_t, q_t)$ , que estão ao redor dos vários pontos de equilíbrio, se distribuirão ao longo da função de demanda, como mostra a figura 9.2. Tais pontos poderão, portanto, ser utilizados para obter uma estimativa da função de demanda, isto é, essa função é, neste caso, identificável.

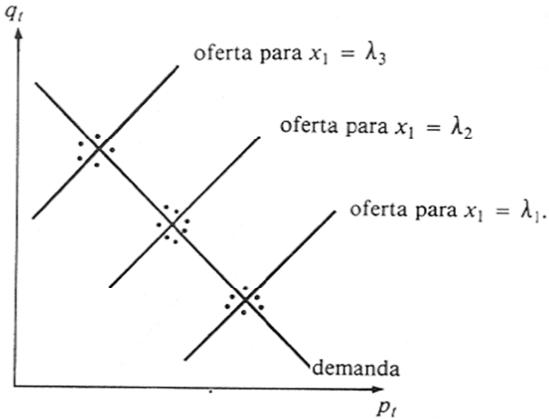


Figura 9.2

Consideremos, agora, um outro tipo de análise para mostrar que no sistema (9.31) a demanda é identificável, mas a oferta não o é. Multiplicando a equação de demanda por uma constante  $\theta$  e somando à equação de oferta, obtemos

$$(1 + \theta)q_t = \gamma_0 + \theta\beta_0 + (\gamma_1 + \theta\beta_1)p_t + \gamma_2x_{1t} + \varepsilon_t + \theta u_t$$

ou

$$q_t = \frac{\gamma_0 + \theta\beta_0}{1 + \theta} + \frac{\gamma_1 + \theta\beta_1}{1 + \theta}p_t + \frac{\gamma_2}{1 + \theta}x_{1t} + \frac{\varepsilon_t + \theta u_t}{1 + \theta} \quad (9.32)$$

Essa equação tem exatamente as mesmas variáveis que a equação de oferta. Não podemos, portanto, distinguir as duas equações sem conhecer os valores dos parâmetros. Com  $\theta \neq 0$ , a equação (9.32) é, na realidade, uma mistura das equações de oferta e de demanda. Isso mostra que a equação de oferta, nesse modelo, não é identificável. Por outro lado, a equação de demanda em (9.31) se distingue de (9.32) pelo fato de não apresentar um termo em  $x_{1t}$ . Concluímos, então, que a equação de demanda é identificável.

A condição  $K \geq N_j$  (ou  $K - K_j \geq L_j$ ) é denominada *condição de ordem* para identificação e é, apenas, uma condição necessária.

Para mostrar que a condição  $K \geq N_j$  não é suficiente, consideremos o sistema

$$\left. \begin{aligned} y_{1t} &= \alpha_1 y_{2t} + \alpha_2 y_{3t} + \alpha_3 x_{1t} & + \varepsilon_{1t} \\ y_{1t} &= & + \beta_2 y_{3t} + \beta_3 x_{1t} + \beta_4 x_{2t} + \beta_5 x_{3t} + \varepsilon_{2t} \\ y_{1t} &= \gamma_1 y_{2t} & + \gamma_3 x_{1t} + \gamma_4 x_{2t} + \gamma_5 x_{3t} + \varepsilon_{3t} \end{aligned} \right\} \quad (9.33)$$

admitindo que se sabe, de acordo com a teoria que serviu de base para a construção do modelo, que

$$\frac{\beta_4}{\gamma_4} = \frac{\beta_5}{\gamma_5} = \theta \quad (9.34)$$

Nesse sistema  $y_{1t}, y_{2t}$  e  $y_{3t}$  são variáveis endógenas e  $x_{1t}, x_{2t}$  e  $x_{3t}$  são variáveis exógenas. Temos, portanto,  $K = 3$  variáveis (exógenas) predeterminadas no sistema.

Para a primeira equação temos  $N_1 = 3 = K$ , ou seja, é satisfeita a condição de ordem para identificação dessa equação.

Entretanto, subtraindo da segunda equação do sistema a terceira multiplicada por  $\theta$  e considerando (9.34), obtemos

$$(1 - \theta)y_{1t} = -\theta\gamma_1 y_{2t} + \beta_2 y_{3t} + (\beta_3 - \theta\gamma_3)x_{1t} + \varepsilon_{2t} - \theta\varepsilon_{3t}$$

ou

$$y_{1t} = \frac{-\theta\gamma_1}{1-\theta} y_{2t} + \frac{\beta_2}{1-\theta} y_{3t} + \frac{\beta_3 - \theta\gamma_3}{1-\theta} x_{1t} + \frac{\varepsilon_{2t} - \theta\varepsilon_{3t}}{1-\theta}$$

Essa equação não se distingue da primeira equação do sistema, uma vez que ambas têm as mesmas variáveis. A primeira equação do sistema não é, portanto, identificável, apesar de ser obedecida a condição de ordem  $K \geq N_j$ .

A seguir vamos apresentar, sem demonstração, a condição necessária e suficiente para identificação de uma equação de um sistema de equações simultâneas. Cabe ressaltar que estaremos considerando essencialmente apenas a identificação baseada em restrições de exclusão de variáveis, isto é, baseada no conhecimento prévio de que os coeficientes de algumas variáveis são iguais a zero.

Lembrando a notação usada na expressão (9.22), consideremos a matriz

$$\mathbf{A} = [\boldsymbol{\Gamma}' \quad \mathbf{B}'] \quad (9.35)$$

Nessa matriz estão todos os coeficientes do sistema, cada linha correspondendo a uma das  $L$  equações. Cada uma das  $L + K$  colunas da matriz  $\mathbf{A}$  mostra os coeficientes de determinada variável (endógena ou exógena) nas  $L$  equações.

Sem perda de generalidade, vamos determinar a condição de identificação da primeira equação. Para isso verificamos onde há zeros na 1ª linha de  $\mathbf{A}$  e formamos uma matriz com os elementos da matriz que ficam abaixo desses zeros, representando essa nova matriz por

$$\mathbf{A}_0 = [\boldsymbol{\Gamma}'_0 \quad \mathbf{B}'_0] \quad (9.36)$$

Essa matriz terá  $L - 1$  linhas e um número de colunas igual ao número de zeros na 1<sup>a</sup> linha de  $\mathbf{A}$ .

A condição necessária e suficiente para identificação da 1<sup>a</sup> equação do sistema (por meio de restrições de exclusão) é que a característica da matriz  $\mathbf{A}_0$  seja igual a  $L - 1$ . Como essa matriz tem  $L - 1$  linhas, a condição de identificação é que essas  $L - 1$  linhas sejam linearmente independentes.

Como a condição (9.29) ou (9.30), denominada condição de ordem para identificação, é uma condição necessária, sendo verificado que ela não é obedecida para determinada equação, podemos concluir que a equação não é identificável, sendo dispensável examinar a condição necessária e suficiente baseada na característica da matriz  $\mathbf{A}_0 = [\Gamma'_0 \quad \mathbf{B}'_0]$ . Por outro lado, se a condição de ordem para identificação for satisfeita, isso não garante que a equação seja efetivamente identificável, devendo-se verificar se é atendida a condição necessária e suficiente.

Para exemplificar, consideremos a 1<sup>a</sup> equação do sistema (9.33). Já vimos que ela atende à condição de ordem para identificação. Obtemos

$$\mathbf{A}_0 = \begin{bmatrix} \beta_4 & \beta_5 \\ \gamma_4 & \gamma_5 \end{bmatrix}$$

À primeira vista essa matriz tem característica 2, sendo atendida a condição suficiente para identificação da 1<sup>a</sup> equação. Mas, dada a relação (9.34), concluímos que essa matriz  $\mathbf{A}_0$  tem característica igual a 1 e que a 1<sup>a</sup> equação do sistema (9.33) não é identificável.

O leitor pode verificar que para todos os demais exemplos apresentados, quando a condição de ordem indica que uma equação é identificável, isso é confirmado pelo exame da condição necessária e suficiente.

Vejamos, finalmente, um exemplo em que o exame da condição necessária e suficiente altera radicalmente o que é sugerido pela condição de ordem (Phillips e Wickens, 1978, p. 283):

$$\left. \begin{array}{l} Y_{1t} = \gamma_{12}Y_{2t} + \gamma_{13}Y_{3t} + \beta_{11}X_{1t} + \beta_{12}X_{2t} + \varepsilon_{1t} \\ Y_{2t} = \gamma_{21}Y_{1t} + \gamma_{23}Y_{3t} + \varepsilon_{2t} \\ Y_{3t} = \gamma_{31}Y_{1t} + \gamma_{32}Y_{2t} + \varepsilon_{3t} \end{array} \right\} \quad (9.37)$$

O sistema tem  $K = 2$  variáveis exógenas ( $X_{1t}$  e  $X_{2t}$ ). Para a 1<sup>a</sup> equação temos  $N_1 = 4 > K$ , concluindo-se que ela não é identificável. A condição de ordem indica que

tanto a 2<sup>a</sup> como a 3<sup>a</sup> equação são exatamente identificáveis (pois  $N_2 = N_3 = 2 = K$ ).

Vejamos a condição necessária e suficiente. Passando todos os termos com parâmetros para o primeiro membro, a matriz com todos os coeficientes fica

$$\mathbf{A} = [\mathbf{\Gamma}' \quad \mathbf{B}'] = \begin{bmatrix} 1 & -\gamma_{12} & -\gamma_{13} & -\beta_{11} & -\beta_{12} \\ -\gamma_{21} & 1 & -\gamma_{23} & 0 & 0 \\ -\gamma_{31} & -\gamma_{32} & 1 & 0 & 0 \end{bmatrix}$$

Para a 2<sup>a</sup> equação obtemos

$$\mathbf{A}_0 = \begin{bmatrix} -\beta_{11} & -\beta_{12} \\ 0 & 0 \end{bmatrix}$$

com característica igual a 1 (menor do que  $L - 1 = 2$ ), concluindo-se que a 2<sup>a</sup> equação não é identificável.

Para a 3<sup>a</sup> equação a matriz  $\mathbf{A}_0$  é a mesma, concluindo-se que a 3<sup>a</sup> equação também não é identificável.

## 9.10. Estimação dos parâmetros em caso de superidentificação

Em caso de superidentificação ( $K > N_j$ ), o sistema (9.27), isto é,

$$\mathbf{X}'\mathbf{y}_j = \mathbf{X}'\mathbf{Z}_j \hat{\boldsymbol{\delta}}_j,$$

tem mais equações do que incógnitas. A matriz  $\mathbf{X}'\mathbf{Z}_j$  não é quadrada e a solução (9.28) não se aplica.

Encaremos a relação (9.26), isto é,

$$\mathbf{X}'\mathbf{y}_j = \mathbf{X}'\mathbf{Z}_j \boldsymbol{\delta}_j + \mathbf{X}'\boldsymbol{\varepsilon}_j,$$

como o modelo de uma regressão linear múltipla de  $\mathbf{X}'\mathbf{y}_j$  contra  $\mathbf{X}'\mathbf{Z}_j$ . O vetor de erros desse modelo de regressão linear múltipla é  $\mathbf{X}'\boldsymbol{\varepsilon}_j$ , que é um vetor-coluna com  $K$  elementos. Desde que a matriz  $\mathbf{X}$  não contenha variáveis aleatórias, temos

$$E(\mathbf{X}'\boldsymbol{\varepsilon}_j \boldsymbol{\varepsilon}_j' \mathbf{X}) = \mathbf{X}'\mathbf{X}\sigma_j^2$$

O estimador de  $\boldsymbol{\delta}_j$  deve ser obtido, então, aplicando o método de mínimos quadrados generalizados. Obtemos

$$\hat{\delta}_j = [\mathbf{Z}'_j \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{Z}_j]^{-1} \mathbf{Z}'_j \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{y}_j \quad (9.38)$$

com matriz de variâncias e covariâncias assintóticas

$$[\mathbf{Z}'_j \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{Z}_j]^{-1} \sigma_j^2 \quad (9.39)$$

Se  $\mathbf{X}' \mathbf{Z}_j$  for uma matriz quadrada ( $K = N_j$ ) e não-singular, de (9.38) obtemos

$$\hat{\delta}_j = (\mathbf{X}' \mathbf{Z}_j)^{-1} \mathbf{X}' \mathbf{X} (\mathbf{Z}'_j \mathbf{X})^{-1} \mathbf{Z}'_j \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}_j = (\mathbf{X}' \mathbf{Z}_j)^{-1} \mathbf{X}' \mathbf{y}_j$$

Isso mostra que (9.28) pode ser encarado, simplesmente, como um caso particular de (9.38).

Note que o estimador (9.38) não apresenta todas as propriedades desejáveis de um estimador de mínimos quadrados generalizados, porque  $\mathbf{Z}_j$  (em  $\mathbf{X}' \mathbf{Z}_j$ ) inclui, normalmente, variáveis conjuntamente determinadas.

Entretanto, pode-se demonstrar que (9.38) é um estimador consistente e que

$$[\mathbf{Z}'_j \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{Z}_j]^{-1} s_j^2 \quad (9.40)$$

com

$$s_j^2 = \frac{(\mathbf{y}_j - \mathbf{Z}_j \hat{\delta}_j)'(\mathbf{y}_j - \mathbf{Z}_j \hat{\delta}_j)}{n - N_j} \quad (9.41)$$

fornecendo estimativas de variâncias e covariâncias das estimativas dos parâmetros assintoticamente válidas (Validade aproximada para amostras suficientemente grandes).

O estimador (9.38) é um estimador de mínimos quadrados em dois estágios (ou estimador de Theil-Basmann).

Se a matriz  $(\mathbf{X}' \mathbf{Z}_j)$  for quadrada e não-singular, a expressão (9.40) pode ser escrita

$$(\mathbf{X}' \mathbf{Z}_j)^{-1} \mathbf{X}' \mathbf{X} (\mathbf{X}' \mathbf{Z}_j)^{-1} s_j^2 \quad (9.42)$$

## 9.11. Outras maneiras de obter o estimador de mínimos quadrados em dois estágios

As estimativas dos parâmetros de (9.24), isto é, de

$$\mathbf{y}_j = \mathbf{Z}_j \boldsymbol{\delta}_j + \boldsymbol{\epsilon}_j,$$

também podem ser obtidas através de um processo em dois estágios.

No primeiro estágio fazemos a regressão de cada uma das variáveis em  $\mathbf{Z}_j$  contra a matriz de variáveis predeterminadas  $\mathbf{X}$ , obtendo

$$\hat{\mathbf{Z}}_j = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_j \quad (9.43)$$

No segundo estágio temos duas variantes:

a) Usar  $\hat{\mathbf{Z}}_j$  como matriz de variáveis instrumentais em  $\mathbf{y}_j = \mathbf{Z}_j \boldsymbol{\delta}_j + \boldsymbol{\epsilon}_j$ , obtendo

$$\hat{\mathbf{Z}}'_j \mathbf{y}_j = \hat{\mathbf{Z}}'_j \mathbf{Z}_j \hat{\boldsymbol{\delta}}_j^*$$

e

$$\hat{\boldsymbol{\delta}}_j^* = (\hat{\mathbf{Z}}'_j \mathbf{Z}_j)^{-1} \hat{\mathbf{Z}}'_j \mathbf{y}_j \quad (9.44)$$

b) Substituir  $\mathbf{Z}_j$  por  $\hat{\mathbf{Z}}_j$  em (9.24) e aplicar mínimos quadrados ordinários, obtendo

$$\hat{\boldsymbol{\delta}}_j^A = (\hat{\mathbf{Z}}'_j \hat{\mathbf{Z}}_j)^{-1} \hat{\mathbf{Z}}'_j \mathbf{y}_j \quad (9.45)$$

Agora, substituindo (9.43) em (9.44) e (9.45), e lembrando (9.38), verifica-se que

$$\hat{\boldsymbol{\delta}}_j^A = \hat{\boldsymbol{\delta}}_j^* = \hat{\boldsymbol{\delta}}_j,$$

isto é, qualquer um dos três caminhos leva ao estimador de mínimos quadrados em dois estágios.

Lembrando (9.43), verifica-se facilmente que a matriz de variâncias e covariâncias assintóticas das estimativas dos parâmetros, dada por (9.39), pode ser escrita

$$(\hat{\mathbf{Z}}'_j \hat{\mathbf{Z}}_j)^{-1} \sigma_j^2$$

## 9.12. Um exemplo numérico

Para ilustrar a aplicação de algumas das fórmulas obtidas consideremos, novamente, o exemplo apresentado nas seções 9.1 e 9.2. Mudando a notação, para adaptá-la à notação geral utilizada, o sistema de equações (9.1) e (9.2) pode ser reescrito da seguinte maneira:

$$\begin{cases} Y_{1t} = \alpha X_{1t} + \beta Y_{2t} + u_t \\ Y_{2t} = Y_{1t} + X_{2t} \end{cases}$$

onde  $X_{1t}$  é uma variável fictícia ( $X_{1t} = 1$ ),  $X_{2t}$  é o valor do investimento,  $Y_{1t}$  é a despesa de consumo e  $Y_{2t}$  é a renda.

Os dados numéricos são reproduzidos na tabela 9.3.

TABELA 9.3. Valores de  $X_{1t}$ ,  $X_{2t}$ ,  $Y_{1t}$  e  $Y_{2t}$

$X_{1t}$	$X_{2t}$	$Y_{1t}$	$Y_{2t}$
1	16	119	135
1	14	126	140
1	18	132	150
1	20	125	145
1	24	131	155
1	28	147	175

O sistema tem  $K = 2$  variáveis predeterminadas ( $X_{1t}$  e  $X_{2t}$ ) e duas variáveis endógenas ( $Y_{1t}$  e  $Y_{2t}$ ).

Sendo  $\mathbf{y}_1$  um vetor-coluna com os valores de  $Y_{1t}$ ,  $\mathbf{Z}_1$  a matriz com as variáveis ( $X_{1t}$  e  $X_{2t}$ ) que aparecem no segundo membro da primeira equação e  $\boldsymbol{\delta}_1$  o vetor-coluna com os parâmetros dessa equação, isto é,

$$\boldsymbol{\delta}_1 = \begin{bmatrix} \alpha \\ \beta \end{bmatrix},$$

a primeira equação fica

$$\mathbf{y}_1 = \mathbf{Z}_1 \boldsymbol{\delta}_1 + \mathbf{u},$$

onde  $\mathbf{u}$  é o vetor dos erros.

Temos

$$\mathbf{y}_1 = \begin{bmatrix} 119 \\ 126 \\ 132 \\ 125 \\ 131 \\ 147 \end{bmatrix} \quad \text{e} \quad \mathbf{Z}_1 = \begin{bmatrix} 1 & 135 \\ 1 & 140 \\ 1 & 150 \\ 1 & 145 \\ 1 & 155 \\ 1 & 175 \end{bmatrix}$$

A matriz  $\mathbf{X}$  com as variáveis predeterminadas é

$$\mathbf{X} = \begin{bmatrix} 1 & 16 \\ 1 & 14 \\ 1 & 18 \\ 1 & 20 \\ 1 & 24 \\ 1 & 28 \end{bmatrix}$$

Obtemos

$$\mathbf{X}'\mathbf{Z}_1 = \begin{bmatrix} 6 & 900 \\ 120 & 18340 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y}_1 = \begin{bmatrix} 780 \\ 15804 \end{bmatrix} \text{ e } (\mathbf{X}'\mathbf{Z}_1)^{-1} = \frac{1}{1020} \begin{bmatrix} 9170 & -450 \\ -60 & 3 \end{bmatrix}$$

Neste exemplo temos identificação exata, com matriz  $\mathbf{X}'\mathbf{Z}_1$  quadrada e não-singular.

De acordo com (9.28) obtemos

$$\hat{\boldsymbol{\delta}}_1 = (\mathbf{X}'\mathbf{Z}_1)^{-1}\mathbf{X}'\mathbf{y}_1 = \begin{bmatrix} 40 \\ 0,6 \end{bmatrix}$$

ou seja,  $\hat{\alpha} = 40$  e  $\hat{\beta} = 0,6$ .

Temos o seguinte vetor de desvios

$$\mathbf{y}_1 - \mathbf{Z}_1 \hat{\boldsymbol{\delta}}_1 = \begin{bmatrix} -2 \\ 2 \\ 2 \\ -2 \\ -2 \\ 2 \end{bmatrix}$$

Então, de acordo com (9.41),

$$s_1^2 = \frac{24}{4} = 6$$

De acordo com (9.42), a matriz das estimativas das variâncias e covariâncias assintóticas das estimativas dos parâmetros é  $(\mathbf{X}'\mathbf{Z}_j)^{-1}\mathbf{X}'\mathbf{X}(\mathbf{Z}'_j\mathbf{X})^{-1}s_j^2$ .

$$\text{Como } \mathbf{X}'\mathbf{X} = \begin{bmatrix} 6 & 120 \\ 120 & 2536 \end{bmatrix}$$

pode-se verificar que a matriz das estimativas das variâncias e covariâncias assintóticas fica

$$\frac{1}{2550} \begin{bmatrix} 67925 & -450 \\ -450 & 3 \end{bmatrix} s_1^2$$

ou

$$\frac{1}{425} \begin{bmatrix} 67925 & -450 \\ -450 & 3 \end{bmatrix}$$

Segue-se que  $\hat{V}(\hat{\alpha}) = 159,82$ ,  $\hat{V}(\hat{\beta}) = 0,00706$  e  $\text{côv}(\hat{\alpha}, \hat{\beta}) = -1,0588$ , lembrando que se trata das variâncias e covariâncias *assintóticas*.

Os testes  $t$  e  $F$  não são estritamente válidos quando há uma ou mais variáveis endógenas no segundo membro da equação estimada. Mas, por falta de alternativa melhor, e tendo em vista que eles são assintoticamente válidos, utilizamos valores de  $t$  e  $F$  para testar hipóteses. No exemplo analisado, para testar  $H_0 : \beta = 0$ , calculamos

$$t = \frac{0,6}{\sqrt{0,00706}} = \frac{0,6}{0,0840} = 7,14$$

Resultados de computador indicam que, na distribuição de  $t$  com 4 graus de liberdade, a probabilidade de o valor absoluto de  $t$  ser maior do que 7,24 é 0,002, mostrando que o teste é significativo ao nível de 1%.

Verificamos, no caso desse exemplo numérico, que os três métodos de estimação analisados (uso de variáveis instrumentais, mínimos quadrados indiretos e mínimos quadrados em dois estágios) levam ao mesmo resultado.

O uso de variáveis instrumentais é sempre equivalente ao método de mínimos quadrados indiretos. Estes dois métodos, entretanto, só são aplicáveis nos casos de identificação exata, conduzindo, então, a resultados idênticos aos do método de mínimos quadrados em dois estágios. Este último método é aplicável, também, em casos de superidentificação.

Assinale-se que os métodos de estimação analisados podem ser aplicados a uma equação se essa equação for exatamente identificável, mesmo que outras equações do sistema sejam superidentificadas ou não identificáveis.

### 9.13. Um segundo exemplo numérico

Consideremos o sistema

$$\begin{cases} y_{1t} = \gamma_1 y_{2t} + \beta_1 x_{1t} + \varepsilon_{1t} \\ y_{1t} = \gamma_2 y_{2t} + \beta_2 x_{2t} + \varepsilon_{2t} \end{cases}$$

onde  $y_{1t}$  e  $y_{2t}$  são variáveis endógenas,  $x_{1t}$  e  $x_{2t}$  são variáveis exógenas, e  $\varepsilon_{1t}$  e  $\varepsilon_{2t}$  são erros aleatórios com média zero e variância constante, sendo que o erro de um período é independente dos erros em outros períodos.

O sistema tem  $K = 2$  variáveis predeterminadas ( $x_{1t}$  e  $x_{2t}$ ) e duas variáveis endógenas ( $y_{1t}$  e  $y_{2t}$ ).

Os valores observados em uma amostra hipotética estão na tabela 9.4.

TABELA 9.4. Amostra de 8 valores para as variáveis  $x_{1t}$ ,  $x_{2t}$ ,  $y_{1t}$  e  $y_{2t}$

$x_{1t}$	$x_{2t}$	$y_{1t}$	$y_{2t}$
-3	-1	-1	-2
-3	-1	-2	0
-1	1	1	-2
-1	1	0	0
1	1	0	0
1	1	2	0
3	-1	-1	2
3	-1	1	2

Note que todas as variáveis estão centradas.

Temos

$$\begin{array}{lll} \sum x_{1t}^2 = 40 & \sum x_{1t} y_{1t} = 10 & \sum x_{1t} y_{2t} = 20 \\ \sum x_{2t}^2 = 8 & \sum x_{2t} y_{1t} = 6 & \sum x_{2t} y_{2t} = -4 \\ \sum x_{1t} x_{2t} = 0 & \sum y_{1t} y_{2t} = 0 & \end{array}$$

Neste exemplo, as duas equações são exatamente identificáveis ( $N_j = K = 2$ ).

Podemos, então, obter as estimativas dos parâmetros por qualquer um dos três métodos analisados (uso de variáveis instrumentais, mínimos quadrados indiretos e mínimos quadrados em dois estágios).

Deixamos para o aluno verificar, como exercício, que qualquer um dos três métodos levará às seguintes estimativas:

$$\hat{\beta}_1 = 1, \hat{\beta}_2 = 1, \hat{\gamma}_1 = -1,5 \text{ e } \hat{\gamma}_2 = 0,5.$$

### 9.14. Terceiro exemplo

Consideremos o seguinte sistema, constituído pelas equações de demanda e de oferta de um produto agrícola

$$\begin{cases} Y_{1t} = \alpha_0 + \alpha_1 Y_{2t} + \alpha_2 X_{1t} + \alpha_3 X_{2t} + \varepsilon_{1t} & (\text{demanda}) \\ Y_{1t} = \beta_0 + \beta_1 Y_{2t} + \beta_2 X_{1t} + \varepsilon_{2t} & (\text{oferta}) \end{cases}$$

onde  $Y_{1t}$  é a quantidade transacionada,  $Y_{2t}$  é o preço,  $X_{1t}$  é o tempo (em anos ou meses),  $X_{2t}$  é a renda *per capita* e  $X_{3t}$  é a pluviosidade.

As variáveis  $Y_{1t}$  e  $Y_{2t}$  são endógenas e as variáveis  $X_{1t}$ ,  $X_{2t}$  e  $X_{3t}$  são exógenas e independentes dos erros  $\varepsilon_{1t}$  e  $\varepsilon_{2t}$ .

Na tabela 9.5 são dados os valores observados em uma amostra. Trata-se de dados artificiais criados tendo em vista fazer com que os cálculos sejam relativamente simples, não se pretendendo que os valores apresentados sejam nem mesmo aproximações de valores reais das variáveis.

TABELA 9.5. Amostra de 8 valores para as variáveis  $Y_{1t}$ ,  $Y_{2t}$ ,  $X_{1t}$ ,  $X_{2t}$  e  $X_{3t}$

$Y_{1t}$	$Y_{2t}$	$X_{1t}$	$X_{2t}$	$X_{3t}$
7	0	-3	1	2
9	0	-3	1	2
10	1	-1	1	2
10	3	-1	1	2
11	6	1	2	1
11	8	1	2	1
16	5	3	2	3
18	5	3	2	3

Obtemos

$$\begin{array}{lllll}
 \sum Y_{1t} = 92 & \sum Y_{2t} = 28 & \sum X_{1t} = 0 & \sum X_{2t} = 12 & \sum X_{3t} = 16 \\
 \bar{Y}_1 = 11,5 & \bar{Y}_2 = 3,5 & \bar{X}_1 = 0 & \bar{X}_2 = 1,5 & \bar{X}_3 = 2 \\
 \sum Y_{1t}^2 = 1152 & \sum Y_{2t}^2 = 160 & \sum X_{1t}^2 = 40 & \sum X_{2t}^2 = 20 & \sum X_{3t}^2 = 36 \\
 \sum y_{1t}^2 = 94 & \sum y_{2t}^2 = 62 & \sum x_{1t}^2 = 40 & \sum x_{2t}^2 = 2 & \sum x_{3t}^2 = 4 \\
 \sum X_{1t}Y_{1t} = 56 & \sum X_{2t}Y_{1t} = 148 & \sum X_{3t}Y_{1t} = 196 & & \\
 \sum x_{1t}y_{1t} = 56 & \sum x_{2t}y_{1t} = 10 & \sum x_{3t}y_{1t} = 12 & & \\
 \sum X_{1t}Y_{2t} = 40 & \sum X_{2t}Y_{2t} = 52 & \sum X_{3t}Y_{2t} = 52 & & \\
 \sum x_{1t}y_{2t} = 40 & \sum x_{2t}y_{2t} = 10 & \sum x_{3t}y_{2t} = -4 & & \\
 \sum X_{1t}X_{2t} = 8 & \sum X_{2t}X_{3t} = 4 & \sum X_{2t}X_{3t} = 24 & & \\
 \sum x_{1t}x_{2t} = 8 & \sum x_{2t}x_{3t} = 4 & \sum x_{2t}x_{3t} = 0 & &
 \end{array}$$

Considerando que existe uma variável fictícia  $X_0 = 1$ , associada aos parâmetros  $\alpha_0$  e  $\beta_0$ , o sistema tem  $K = 4$  variáveis predeterminadas.

No caso da equação de demanda temos, incluindo a variável  $X_0 = 1$ ,  $N_1 = 4$  variáveis no segundo membro da equação. A equação de demanda é exatamente identificável e as estimativas dos parâmetros podem ser obtidas através de qualquer um dos três métodos analisados (variáveis instrumentais, mínimos quadrados em dois estágios ou mínimos quadrados indiretos).

No caso da equação de oferta temos, incluindo a variável  $X_0 = 1$ ,  $N_2 = 3$  variáveis no segundo membro da equação. Como  $N_2 < K = 4$ , temos um caso de superidentificação. Os parâmetros dessa equação serão estimados pelo método de mínimos quadrados em dois estágios.

Seja  $\mathbf{y}_1$  o vetor-coluna com os  $n = 8$  valores de  $y_{1t}$ . Analogamente, sejam  $\mathbf{y}_2$ ,  $\mathbf{x}_1$ ,  $\mathbf{x}_2$  e  $\mathbf{x}_3$  os vetores-coluna com os valores das variáveis centradas  $y_{2t}$ ,  $x_{1t}$ ,  $x_{2t}$  e  $x_{3t}$ , respectivamente.

Se trabalharmos com todas as variáveis centradas ficam, provisoriamente, eliminados os termos constantes das equações do sistema.

A matriz de variáveis predeterminadas é, neste caso

$$\mathbf{X} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{x}_3]$$

Além de  $\mathbf{y}_1$  e  $\mathbf{X}$ , utilizaremos adiante as matrizes:

- a)  $\mathbf{Z}_1 = [\mathbf{y}_2 \quad \mathbf{x}_1 \quad \mathbf{x}_2]$ , que é a matriz com as variáveis que aparecem no segundo membro da primeira equação (a demanda). Uma vez que todas as variáveis são centradas, deixa de existir a variável  $X_0$ .
- b)  $\mathbf{Z}_2 = [\mathbf{y}_2 \quad \mathbf{x}_3]$ , que é a matriz com as variáveis que aparecem no segundo membro da segunda equação (a oferta).
- c)  $\hat{\boldsymbol{\delta}}_1 = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\alpha}_3 \end{bmatrix}$  e  $\hat{\boldsymbol{\delta}}_2 = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$ , que são os vetores com as estimativas dos parâmetros da equação de demanda ( $\hat{\boldsymbol{\delta}}_1$ ) e da equação de oferta ( $\hat{\boldsymbol{\delta}}_2$ ).

Aplicando o método das variáveis instrumentais para estimar os parâmetros da primeira equação (a demanda) obtemos, de acordo com (9.27), o sistema de equações

$$\mathbf{X}'\mathbf{y}_1 = \mathbf{X}'\mathbf{Z}_1\hat{\boldsymbol{\delta}}_1$$

ou

$$\begin{cases} \sum x_{1t} y_{1t} = \hat{\alpha}_1 \sum x_{1t} y_{2t} + \hat{\alpha}_2 \sum x_{1t}^2 + \hat{\alpha}_3 \sum x_{1t} x_{2t} \\ \sum x_{2t} y_{1t} = \hat{\alpha}_1 \sum x_{2t} y_{2t} + \hat{\alpha}_2 \sum x_{1t} x_{2t} + \hat{\alpha}_3 \sum x_{2t}^2 \\ \sum x_{3t} y_{1t} = \hat{\alpha}_1 \sum x_{3t} y_{2t} + \hat{\alpha}_2 \sum x_{1t} x_{3t} + \hat{\alpha}_3 \sum x_{2t} x_{3t} \end{cases}$$

Considerando os valores numéricos dados, o sistema fica

$$\begin{cases} 56 = 40\hat{\alpha}_1 + 40\hat{\alpha}_2 + 8\hat{\alpha}_3 \\ 10 = 10\hat{\alpha}_1 + 8\hat{\alpha}_2 + 2\hat{\alpha}_3 \\ 12 = -4\hat{\alpha}_1 + 4\hat{\alpha}_2 \end{cases}$$

Resolvendo, obtemos  $\hat{\alpha}_1 = -1$ ,  $\hat{\alpha}_2 = 2$  e  $\hat{\alpha}_3 = 2$ .

É evidente que o mesmo resultado seria obtido se utilizássemos a relação (9.28).

A equação de demanda estimada é

$$\hat{y}_{1t} = -y_{2t} + 2x_{1t} + 2x_{2t}$$

ou

$$\hat{Y}_{1t} - 11,5 = -(Y_{2t} - 3,5) + 2(X_{1t} - 0) + 2(X_{2t} - 1,5)$$

Então

$$\hat{Y}_{1t} = 12 - Y_{2t} + 2X_{1t} + 2X_{2t}$$

Verifica-se que  $\hat{\alpha}_0 = 12$ .

Vejamos, a seguir, como as estimativas dos parâmetros da equação de demanda poderiam ser obtidas pelo método de mínimos quadrados em dois estágios.

No primeiro estágio, obtemos a regressão de  $y_{2t}$  (a variável endógena que aparece no segundo membro da equação) contra  $x_{1t}, x_{2t}$  e  $x_{3t}$ , que são as variáveis predeterminadas do sistema. Seja  $\hat{\theta}$  o vetor das estimativas dos parâmetros dessa regressão.

Temos

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \sum x_{1t}^2 & \sum x_{1t}x_{2t} & \sum x_{1t}x_{3t} \\ \sum x_{1t}x_{2t} & \sum x_{2t}^2 & \sum x_{2t}x_{3t} \\ \sum x_{1t}x_{3t} & \sum x_{2t}x_{3t} & \sum x_{3t}^2 \end{bmatrix} = \begin{bmatrix} 40 & 8 & 4 \\ 8 & 2 & 0 \\ 4 & 0 & 4 \end{bmatrix}$$

,

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{4} \begin{bmatrix} 1 & -4 & -1 \\ -4 & 18 & 4 \\ -1 & 4 & 2 \end{bmatrix} \text{ e } \mathbf{X}'\mathbf{y}_2 = \begin{bmatrix} \sum x_{1t}y_{2t} \\ \sum x_{2t}y_{2t} \\ \sum x_{3t}y_{2t} \end{bmatrix} = \begin{bmatrix} 40 \\ 10 \\ -4 \end{bmatrix}$$

A seguir obtemos

$$\hat{\theta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}_2 = \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix}$$

Então  $\hat{y}_{2t} = x_{1t} + x_{2t} - 2x_{3t}$

$$\text{ou } \hat{\mathbf{y}}_2 = \mathbf{X}\hat{\theta} = \begin{bmatrix} -3,5 \\ -3,5 \\ -1,5 \\ -1,5 \\ 3,5 \\ 3,5 \\ 1,5 \\ 1,5 \end{bmatrix}$$

No segundo estágio fazemos a regressão de  $y_{1t}$  contra  $\hat{y}_{2t}$ ,  $x_{1t}$  e  $x_{2t}$ , obtendo  $\hat{\delta}_1$ .

Sendo  $\hat{\mathbf{Z}}_1 = [\hat{\mathbf{y}}_2 \quad \mathbf{x}_1 \quad \mathbf{x}_2]$ , temos

$$\hat{\delta}_1 = (\hat{\mathbf{Z}}_1' \hat{\mathbf{Z}}_1)^{-1} \hat{\mathbf{Z}}_1' \mathbf{y}_1$$

Obtemos

$$\begin{aligned}\hat{\mathbf{Z}}_1' \hat{\mathbf{Z}}_1 &= \begin{bmatrix} \sum \hat{y}_{2t}^2 & \sum x_{1t} \hat{y}_{2t} & \sum x_{2t} \hat{y}_{2t} \\ \sum x_{1t} \hat{y}_{2t} & \sum x_{1t}^2 & \sum x_{1t} x_{2t} \\ \sum x_{2t} \hat{y}_{2t} & \sum x_{1t} x_{2t} & \sum x_{2t}^2 \end{bmatrix} = \begin{bmatrix} 58 & 40 & 10 \\ 40 & 40 & 8 \\ 10 & 8 & 2 \end{bmatrix}, \\ (\hat{\mathbf{Z}}_1' \hat{\mathbf{Z}}_1)^{-1} &= \frac{1}{8} \begin{bmatrix} 1 & 0 & -5 \\ 0 & 1 & -4 \\ -5 & -4 & 45 \end{bmatrix} \text{ e } \hat{\mathbf{Z}}_1' \mathbf{y}_1 = \begin{bmatrix} \sum y_{1t} \hat{y}_{2t} \\ \sum x_{1t} y_{1t} \\ \sum x_{2t} y_{1t} \end{bmatrix} = \begin{bmatrix} 42 \\ 56 \\ 10 \end{bmatrix}\end{aligned}$$

Então

$$\hat{\delta}_1 = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\alpha}_3 \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \\ 2 \end{bmatrix}$$

A matriz de variâncias e covariâncias assintóticas dessas estimativas é

$$(\hat{\mathbf{Z}}_1' \hat{\mathbf{Z}}_1)^{-1} \sigma_1^2 = \frac{1}{8} \begin{bmatrix} 1 & 0 & -5 \\ 0 & 1 & -4 \\ -5 & -4 & 45 \end{bmatrix} \sigma_1^2$$

De acordo com (9.41) o estimador de  $\sigma_1^2$  é

$$s_1^2 = \frac{(\mathbf{y}_1 - \mathbf{Z}_1 \hat{\delta}_1)' (\mathbf{y}_1 - \mathbf{Z}_1 \hat{\delta}_1)}{n - N_1}$$

Temos  $n = 8$ ,  $N_1 = 4$  e

$$\mathbf{y}_1 - \mathbf{Z}_1 \hat{\delta}_1 = \begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \\ -1 \\ 1 \\ -1 \\ 1 \end{bmatrix}$$

Então

$$s_1^2 = \frac{8}{4} = 2$$

e a matriz das estimativas das variâncias e covariâncias assintóticas de  $\hat{\alpha}_1$ ,  $\hat{\alpha}_2$  e  $\hat{\alpha}_3$  é

$$\begin{bmatrix} \frac{1}{4} & 0 & -\frac{5}{4} \\ 0 & \frac{1}{4} & -1 \\ -\frac{5}{4} & -1 & \frac{45}{4} \end{bmatrix}$$

Passemos à equação de oferta. Como, para essa equação, temos superidentificação, dos três métodos de estimação apresentados, só o método de mínimos quadrados em dois estágios é aplicável.

Então, de acordo com esse método, fazemos, no primeiro estágio, uma regressão de  $y_{2t}$  (a variável endógena que aparece no segundo membro da equação de oferta) contra  $x_{1t}$ ,  $x_{2t}$  e  $x_{3t}$ , que são as variáveis predeterminadas do sistema. Neste exemplo, este estágio coincide com o primeiro estágio da aplicação do mesmo método à equação de demanda.

No segundo estágio fazemos a regressão de  $y_{1t}$  contra  $\hat{y}_{2t}$  e  $x_{3t}$ , obtendo  $\hat{\delta}_2$ .

Sendo  $\hat{\mathbf{Z}}_2 = [\hat{\mathbf{y}}_2 \quad \mathbf{x}_3]$ , temos

$$\hat{\delta}_2 = (\hat{\mathbf{Z}}_2' \hat{\mathbf{Z}}_2)^{-1} \hat{\mathbf{Z}}_2' \mathbf{y}_1$$

Obtemos

$$\hat{\mathbf{Z}}_2' \hat{\mathbf{Z}}_2 = \begin{bmatrix} \sum \hat{y}_{2t}^2 & \sum x_3 \hat{y}_{2t} \\ \sum x_3 \hat{y}_{2t} & \sum x_{3t}^2 \end{bmatrix} = \begin{bmatrix} 58 & -4 \\ -4 & 4 \end{bmatrix},$$

$$(\hat{\mathbf{Z}}_2' \hat{\mathbf{Z}}_2)^{-1} = \frac{1}{108} \begin{bmatrix} 2 & 2 \\ 2 & 29 \end{bmatrix} \text{ e } \hat{\mathbf{Z}}_2' \mathbf{y}_1 = \begin{bmatrix} \sum y_{1t} \hat{y}_{2t} \\ \sum x_{3t} y_{1t} \end{bmatrix} = \begin{bmatrix} 42 \\ 12 \end{bmatrix}$$

Então

$$\hat{\delta}_2 = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \end{bmatrix}$$

A equação de oferta estimada é

$$\hat{y}_{1t} = y_{2t} + 4x_{3t}$$

Então

$$\hat{Y}_{1t} - 11,5 = Y_{2t} - 3,5 + 4(X_{3t} - 2)$$

e

$$\hat{Y}_{1t} = Y_{2t} + 4X_{3t}$$

Verifica-se que  $\hat{\beta}_0 = 0$

A matriz de variâncias e covariâncias assintóticas de  $\hat{\beta}_1$  e  $\hat{\beta}_2$  é

$$(\hat{\mathbf{Z}}_2' \hat{\mathbf{Z}}_2)^{-1} \sigma^2 = \frac{1}{108} \begin{bmatrix} 2 & 2 \\ 2 & 29 \end{bmatrix} \sigma^2$$

De acordo com (9.41) o estimador de  $\sigma^2$  é

$$s^2 = \frac{(\mathbf{y}_1 - \mathbf{Z}_2 \hat{\boldsymbol{\delta}}_2)' (\mathbf{y}_1 - \mathbf{Z}_2 \hat{\boldsymbol{\delta}}_2)}{n - N_2}$$

Temos  $n = 8$ ,  $N_2 = 3$  e

$$\mathbf{y}_1 - \mathbf{Z}_2 \hat{\boldsymbol{\delta}}_2 = \begin{bmatrix} -1 \\ 1 \\ 1 \\ -1 \\ 1 \\ -1 \\ -1 \\ 1 \end{bmatrix}$$

Então  $s^2 = 8/5$  e a matriz das estimativas das variâncias e covariâncias assintóticas de  $\hat{\beta}_1$  e  $\hat{\beta}_2$  é

$$\begin{bmatrix} \frac{4}{135} & \frac{4}{135} \\ \frac{4}{135} & \frac{58}{135} \end{bmatrix}$$

### 9.15. Uma visão global

Para o modelo básico  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ , com, entre outras pressuposições,

$$E(\mathbf{u}) = \mathbf{0}$$

$$E(\mathbf{u}\mathbf{u}') = \mathbf{I}\sigma^2$$

$$E(\mathbf{X}'\mathbf{u}) = \mathbf{0},$$

obtemos  $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ , que é um estimador linear, não-tendencioso, de variância mínima e consistente.

Entretanto, freqüentemente uma ou mais dessas pressuposições não são obedecidas. Consideremos 2 casos gerais:

(I) O vetor de erros  $\mathbf{u}$  tem matriz de variâncias e covariâncias

$$E(\mathbf{u}\mathbf{u}') = \mathbf{V}\sigma^2, \text{ com } \mathbf{V} \neq \mathbf{I}$$

Então o melhor estimador de  $\boldsymbol{\beta}$  é o de mínimos quadrados generalizados, dado por

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y},$$

com matriz de variâncias e covariâncias

$$(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\sigma^2$$

O estimador de mínimos quadrados ordinários é, neste caso, não-tendencioso, mas ineficiente. A expressão usual  $(\mathbf{X}'\mathbf{X})^{-1}s^2$  dará estimativas incorretas, tendendo freqüentemente a subestimar as variâncias e covariâncias verdadeiras.

Heterocedasticia e autocorrelação nos resíduos são casos particulares desse problema.

(II) elementos de  $\mathbf{X}$  são aleatórios e correlacionados com elementos de  $\mathbf{u}$ .

O estimador de mínimos quadrados ordinários, neste caso, é tendencioso.

É interessante distinguir dois subcasos:

(II-a) Os valores de  $\mathbf{X}$  correspondentes à  $t$ -ésima observação (à  $t$ -ésima linha de  $\mathbf{X}$ ) são correlacionados com algum valor do erro ( $u_h, h \neq t$ ), mas não com o erro da mesma observação ( $u_t$ ).

Exemplo disso é um modelo com variáveis defasadas

$$Y_t = \alpha + \beta X_t + \gamma Y_{t-1} + u_t,$$

onde os  $u_t$  não são autocorrelacionados. Então  $Y_{t-1}$  está correlacionado com  $u_{t-h}$  ( $h = 1, 2, \dots$ ), mas não está correlacionado com  $u_t$  ou  $u_{t+h}$  ( $h = 1, 2, \dots$ ).

Pode-se demonstrar que, neste caso, o estimador de mínimos quadrados ordinários é consistente, embora seja tendencioso.

(II-b) Os valores de  $\mathbf{X}$  correspondentes à  $t$ -ésima observação são correlacionados com  $u_t$ .

Nessa situação o estimador de mínimos quadrados ordinários é inconsistente.

Isto ocorre, por exemplo, quando variáveis importantes não são incluídas no modelo, quando há erros de observação nas variáveis independentes e no caso de sistemas de equações simultâneas.

## Exercícios

- 9.1. Considere um modelo muito simples de determinação da renda nacional, constituído por apenas duas equações simultâneas:

$$\begin{cases} C_t = \alpha + \beta Y_t + u_t \\ Y_t = C_t + Z_t \end{cases}$$

As variáveis endógenas são a renda nacional ( $Y_t$ ) e a despesa com consumo ( $C_t$ ). Os erros  $u_t$  são variáveis aleatórias independentes com média zero e variância constante. O valor dos investimentos ( $Z_t$ ) é uma variável exógena (não correlacionada com  $u_t$ ).

É dada uma amostra de 4 valores de  $C_t$ ,  $Y_t$  e  $Z_t$ :

$C_t$	$Y_t$	$Z_t$
14	15	1
12	15	3
20	25	5
38	45	7

Verifica-se que  $\sum C_t = 84$ ,  $\sum Y_t = 100$  e  $\sum Z_t = 16$ .

- Obtenha as estimativas de  $\alpha$  e  $\beta$  por mínimos quadrados ordinários, ajustando a regressão linear simples de  $C_t$  contra  $Y_t$ , como sugere a equação (1), isoladamente.
- Considerando o sistema de equações simultâneas, obtenha estimativas consistentes de  $\alpha$  e  $\beta$ . Como é denominado o método que você usou?
- O que você pode dizer dos estimadores usados no item (a)?

- 9.2. Na tabela ao lado estão os valores de  $Y_1$ ,  $Y_2$  e  $X$  em uma amostra com 4 observações. Admite-se que essas variáveis estão relacionadas de acordo com o sistema

$Y_1$	$Y_2$	$X$
10	22	2
21	14	5
28	4	8
17	16	5

$$\begin{cases} Y_{2t} = \alpha + \beta Y_{1t} + u_t \\ Y_{1t} = \gamma + \delta X_t + \varepsilon_t \end{cases}$$

As variáveis endógenas são  $Y_1$ ,  $Y_2$ . Pressupõe-se que os erros  $u_t$  e  $\varepsilon_t$  têm média zero e variância constante, não apresentam autocorrelação e não são correlacionados com  $X_t$ . Admite-se, entretanto, que  $u_t$  pode ter correlação com  $\varepsilon_t$ .

- a) Analise a identificação de cada equação.
- b) Quando a equação for identificável, obtenha estimativas (apropriadas) dos seus parâmetros.
- c) Teste, ao nível de significância de 5%, a hipótese de que  $\delta = 0$ , contra a hipótese alternativa de que  $\delta > 0$ .

9.3. Seja o sistema de equações

$$\begin{cases} Y_t = \alpha + \beta X_t + u_t \\ X_t = \gamma + \delta Z_t + \varepsilon_t \end{cases}$$

onde  $u_t$  e  $\varepsilon_t$  são variáveis aleatórias com média zero e variância constante.

Admite-se que  $u_t$  e  $\varepsilon_t$  apresentam correlação assintótica negativa, isto é,  $\text{plim}\left(\frac{1}{n} \sum u_t \varepsilon_t\right) < 0$ . Admite-se, também que  $u_t$  e  $\varepsilon_t$  não são assintoticamente correlacionados com  $Z_t$  e que o erro de uma observação ( $u_t$  ou  $\varepsilon_t$ ) é independente dos erros das demais observações ( $u_{t+h}$  e  $\varepsilon_{t+h}$ ).

É dada a seguinte amostra de valores de  $Y_t$ ,  $X_t$  e  $Z_t$  (extraída de KONG CHU, *Principles of Econometrics*, International Textbook, 1968, p. 110):

$Y_t$	$X_t$	$Z_t$
10	5	2
19	8	4
16	6	3
22	7	5
33	10	8
45	12	11
48	13	12
61	15	15
70	20	14
76	24	16

Pode-se verificar que

$$\begin{array}{lll}
 \sum Y_t = 400 & \sum X_t = 120 & \sum Z_t = 90 \\
 \bar{Y} = 40 & \bar{X} = 12 & \bar{Z} = 9 \\
 \sum Y_t^2 = 21016 & \sum X_t^2 = 1788 & \sum Z_t^2 = 1060 \\
 \sum y_t^2 = 5016 & \sum x_t^2 = 348 & \sum z_t^2 = 250 \\
 \sum X_t Y_t = 6085 & \sum X_t Z_t = 1352 & \sum Z_t Y_t = 4700 \\
 \sum x_t y_t = 1285 & \sum x_t z_t = 272 & \sum z_t y_t = 1100
 \end{array}$$

- a) Mostre que o estimador de mínimos quadrados ordinários,  $b = \sum xy / \sum x^2$ , tende a subestimar  $\beta$ , isto é, mostre que  $\text{plim } b < \beta$ .
  - b) Obtenha estimativas dos parâmetros  $\alpha$ ,  $\beta$ ,  $\gamma$  e  $\delta$  e dê as qualidades dos estimadores utilizados.
- 9.4. Considere o seguinte sistema, constituído pelas equações de demanda e de oferta de um produto, em certo mercado:

$$\begin{cases} Y_{1t} = \alpha_1 + \gamma_1 Y_{2t} + \beta_1 X_{1t} + \varepsilon_{1t} & (\text{demanda}) \\ Y_{1t} = \alpha_2 + \gamma_2 Y_{2t} + \beta_2 X_{2t} + \varepsilon_{2t} & (\text{oferta}) \end{cases}$$

As variáveis endógenas  $Y_{1t}$  e  $Y_{2t}$  são a quantidade transacionada e o preço do produto, respectivamente, no período  $t$ . As variáveis exógenas  $X_{1t}$  e  $X_{2t}$  são a renda *per capita* e o preço de um subproduto, respectivamente.

Admite-se que  $\varepsilon_{1t}$  ou  $\varepsilon_{2t}$  são erros aleatórios com média zero e variância constante, e que o erro de um período ( $\varepsilon_{1t}$  ou  $\varepsilon_{2t}$ ) é independente dos erros em outros períodos.

É dada a seguinte amostra de valores  $X_{1t}$ ,  $X_{2t}$ ,  $Y_{1t}$  e  $Y_{2t}$ :

$X_{1t}$	$X_{2t}$	$Y_{1t}$	$Y_{2t}$
1	1	2	1
1	1	1	3
3	3	4	1
3	3	3	3
5	3	3	3
5	3	5	3
7	1	2	5
7	1	4	5

- a) É satisfeita a condição necessária para que as duas equações sejam identificáveis?
- b) Obtenha estimativas consistentes dos parâmetros da equação de demanda e determine a matriz das estimativas das respectivas variâncias e covariâncias assintóticas.
- c) Idem, para a equação de oferta.
- d) Para a equação de demanda, faça os cálculos sem centrar e centrando as variáveis e verifique que os resultados são idênticos.

Note que se todas as variáveis forem centradas, esse exercício se confunde com o segundo exemplo dado no texto (seção 9.13).

9.5. Considere o seguinte sistema de equações simultâneas:

$$Y_{1t} = \alpha + \beta X_t + u_{1t}$$

$$Y_{2t} = \gamma + \delta Y_{1t} + u_{2t}$$

onde  $X_t$  é uma variável exógena,  $Y_{1t}$  e  $Y_{2t}$  são variáveis endógenas e os erros  $u_{1t}$  e  $u_{2t}$  são ruídos brancos, podendo haver correlação entre  $u_{1t}$  e  $u_{2t}$ .

Obtenha estimativas consistentes de  $\gamma$  e  $\delta$  com base na seguinte amostra:

$X_t$	$Y_{1t}$	$Y_{2t}$
10	26	120
10	22	96
2	6	35
10	18	81

9.6. Considere o seguinte modelo (extremamente simples) de determinação da renda nacional ( $Y_{1t}$ ) e do consumo ( $Y_{2t}$ ), com duas equações:

$$Y_{2t} = \alpha + \beta Y_{1t} + u_t \quad (1)$$

$$Y_{1t} = Y_{2t} + Z_t \quad (2)$$

Admite-se que o investimento ( $Z_t$ ) é uma variável exógena, não correlacionada com o erro  $u_t$ .

São dados os valores de  $Z_t$ ,  $Y_{1t}$  e  $Y_{2t}$  em uma amostra com 5 observações:

$Z_t$	$Y_{1t}$	$Y_{2t}$
10	60	50
12	68	56
10	64	54
11	76	65
12	72	60

- a) Estime  $\beta$  ajustando a equação (1) por mínimos quadrados ordinários.
- b) Quais são, neste caso, as propriedades desse estimador?
- c) Obtenha estimativas consistentes de  $\beta$  e de  $\alpha$ .
- d) Determine a estimativa do desvio padrão do estimador de  $\beta$  utilizado no item anterior.

9.7. Na tabela ao lado estão os valores de  $X$ ,  $Y_1$ ,  $Y_2$  em uma amostra com 5 observações. Admite-se que estas variáveis estão relacionadas de acordo com o sistema

$$\begin{cases} Y_{2t} = \alpha + \beta Y_{1t} + u_t \\ Y_{1t} = \gamma + \delta X_t + \varepsilon_t \end{cases}$$

$X$	$Y_1$	$Y_2$
4	16	87
2	4	3
5	20	102
1	6	24
3	9	39

As variáveis endógenas são  $Y_1$  e  $Y_2$ . Pressupõe-se que os erros  $u_t$  e  $\varepsilon_t$  tem média zero e variância constante, não apresentam autocorrelação e não são correlacionados com  $X_t$ . Admite-se, entretanto, que  $u_t$  pode ter correlação com  $\varepsilon_t$ .

- a) Analise a identificação de cada equação.
- b) Quando a equação for identificável, obtenha estimativas consistentes dos seus parâmetros.
- c) Se possível, obtenha uma estimativa do desvio padrão da estimativa de  $\beta$ .

9.8. Considere o seguinte sistema de equações simultâneas, no qual  $Y_1$  e  $Y_2$  são variáveis endógenas e  $X_1$  e  $X_2$  são variáveis exógenas:

$$\begin{aligned} Y_{2t} &= \alpha + \beta Y_{1t} + u_{2t} \\ Y_{1t} &= \theta_0 + \theta_1 X_{1t} + \theta_2 X_{2t} + u_{1t} \end{aligned}$$

É dada a seguinte amostra de valores das 4 variáveis:

$X_{1t}$	$X_{2t}$	$Y_{1t}$	$Y_{2t}$
11	1	22	37
9	1	21	37
9	7	45	85
7	4	26	51
5	1	17	29
5	7	41	77
3	7	38	69

Sabe-se que a estimativa da segunda equação é  $\hat{Y}_{1t} = 7 + X_{1t} + 4X_{2t}$

- a) Analise a identificação da primeira equação.
- b) Obtenha estimativas consistentes de  $\alpha$  e  $\beta$
- c) Teste, ao nível de significância de 1%, a hipótese de que  $\beta = 0$ .

9.9. Considere o seguinte sistema, constituído pelas equações de demanda e de oferta de um produto, em certo mercado:

$$\begin{cases} Y_{1t} = \alpha_0 + \alpha_1 Y_{2t} + \alpha_2 X_{1t} + \varepsilon_{1t} & \text{(demanda)} \\ Y_{1t} = \beta_0 + \beta_1 Y_{2t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \varepsilon_{2t} & \text{(oferta)} \end{cases}$$

As variáveis endógenas  $Y_{1t}$  e  $Y_{2t}$  são a quantidade transacionada e o preço do produto, respectivamente, no período  $t$ .

As variáveis exógenas  $X_{1t}$ ,  $X_{2t}$  e  $X_{3t}$  são, respectivamente, a renda *per capita*, o montante de subsídios recebidos pelos produtores e o preço da matéria-prima.

Admite-se que  $\varepsilon_{1t}$  e  $\varepsilon_{2t}$  são erros aleatórios com média zero e variância constante, e que os erros de um período, embora correlacionados entre si, são não-correlacionados com os erros relativos a outros períodos.

É dada a seguinte amostra de valores  $X_{1t}$ ,  $X_{2t}$ ,  $X_{3t}$ ,  $Y_{1t}$  e  $Y_{2t}$ :

$X_{1t}$	$X_{2t}$	$X_{3t}$	$Y_{1t}$	$Y_{2t}$
5	130	30	18,5	7,5
6	110	20	20,0	8,0
7	120	26	19,5	7,5
7	120	26	18,5	8,5
8	110	30	19,5	12,5
9	130	24	24,0	10,0

- a) A equação de demanda é identificável?
- b) A equação de oferta é identificável?
- c) Obtenha estimativas consistentes dos parâmetros  $\alpha_1$  e  $\alpha_2$  da equação de demanda e determine a matriz das estimativas das respectivas variâncias e covariâncias assintóticas.
- d) Idem, para os parâmetros  $\beta_1$ ,  $\beta_2$  e  $\beta_3$  da equação de oferta.
- e) Determine as estimativas da elasticidade-preço da demanda, da elasticidade-renda da demanda e da elasticidade-preço da oferta para os valores médios de  $Y_{1t}$ ,  $Y_{2t}$  e  $X_{1t}$ .

9.10. Considere o modelo

$$\begin{cases} \mathbf{y}_1 = \gamma \mathbf{y}_2 + \mathbf{v} \\ \mathbf{y}_2 = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \end{cases}$$

onde

$$\mathbf{y}_1 = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n} \end{bmatrix}, \quad \mathbf{y}_2 = \begin{bmatrix} y_{21} \\ y_{22} \\ \vdots \\ y_{2n} \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix},$$

$$\mathbf{X} = \begin{bmatrix} X_{01} & X_{11} & \dots & X_{k1} \\ X_{02} & X_{12} & \dots & X_{k2} \\ \vdots & \vdots & & \vdots \\ X_{0n} & X_{1n} & \dots & X_{kn} \end{bmatrix} \text{ e } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

Admite-se que:

a)  $E(v_t) = E(u_t) = E(v_t v_{t+h}) = E(u_t u_{t+h}) = 0$  para  $h \neq 0$ .

b)  $E(v_t^2) = \text{plim} \left( \frac{1}{n} \sum v_t^2 \right) = \sigma_v^2$

c)  $E(u_t^2) = \text{plim} \left( \frac{1}{n} \sum u_t^2 \right) = \sigma_u^2$

d)  $E(u_t v_t) = \text{plim} \left( \frac{1}{n} \sum u_t v_t \right) = \sigma_{uv}$

e)  $\text{plim} \left( \frac{1}{n} \mathbf{X}' \mathbf{u} \right) = \text{plim} \left( \frac{1}{n} \mathbf{X}' \mathbf{v} \right) = \mathbf{0}$

f)  $\text{plim} \left( \frac{1}{n} \mathbf{X}' \mathbf{X} \right) = \mathbf{Q}$

Seja  $c = (\mathbf{y}_1' \mathbf{y}_2) / (\mathbf{y}_2' \mathbf{y}_2)$  o estimador de mínimos quadrados ordinários de  $\gamma$ .

Mostre que

$$c = \gamma + \frac{\frac{1}{n} \boldsymbol{\beta}' \mathbf{X}' \mathbf{v} + \frac{1}{n} \mathbf{u}' \mathbf{v}}{\frac{1}{n} \boldsymbol{\beta}' \mathbf{X}' \mathbf{X} \boldsymbol{\beta} + \frac{2}{n} \boldsymbol{\beta}' \mathbf{X}' \mathbf{u} + \frac{1}{n} \mathbf{u}' \mathbf{u}}$$

e que

$$\text{plim } c = \gamma + \frac{\sigma_{uv}^2}{\boldsymbol{\beta}' \mathbf{Q} \boldsymbol{\beta} + \sigma_u^2}$$

Note que o estimador de mínimos quadrados ordinários é consistente se  $\sigma_{uv}^2 = 0$ ; neste caso, o modelo dado constituiria um sistema recursivo (Ver Johnston, 1972, p. 377-380 ou Wonnacott e Wonnacott, 1976, p. 180-182).

Verifique que o estimador de mínimos quadrados em dois estágio de  $\gamma$  é

$$\hat{\gamma} = \frac{\mathbf{y}_2' \mathbf{N} \mathbf{y}_1}{\mathbf{y}_2' \mathbf{N} \mathbf{y}_2}, \text{ onde } \mathbf{N} = \mathbf{X}(\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'$$

Mostre que

$$\hat{\gamma} = \gamma + \frac{\boldsymbol{\beta}' \left( \frac{1}{n} \mathbf{X}' \mathbf{v} \right) + \left( \frac{1}{n} \mathbf{u}' \mathbf{X} \right) \left( \frac{1}{n} \mathbf{X}' \mathbf{X} \right)^{-1} \left( \frac{1}{n} \mathbf{X}' \mathbf{v} \right)}{\boldsymbol{\beta}' \left( \frac{1}{n} \mathbf{X}' \mathbf{X} \right) \boldsymbol{\beta} + 2 \boldsymbol{\beta}' \left( \frac{1}{n} \mathbf{X}' \mathbf{u} \right) + \left( \frac{1}{n} \mathbf{u}' \mathbf{X} \right) \left( \frac{1}{n} \mathbf{X}' \mathbf{X} \right)^{-1} \left( \frac{1}{n} \mathbf{X}' \mathbf{u} \right)}$$

e que  $\text{plim } \hat{\gamma} = \gamma$ , se  $\boldsymbol{\beta}' \mathbf{Q}' \boldsymbol{\beta} \neq 0$ .

## Respostas

9.1. a)  $b = 5/6 = 0,833$  e  $a = 1/6 = 0,167$

b)  $\hat{\beta} = 0,8$  e  $\hat{\alpha} = 1$

- c) São estimadores inconsistentes, com  $b$  tendendo a superestimar  $\beta$ .
- 9.2. a) A 1<sup>a</sup> equação é exatamente identificável. A segunda equação não tem variável endógena no segundo membro e, consequentemente, é identificável e pode ser estimada por mínimos quadrados ordinários.
- b)  $\hat{Y}_2 = 33 - Y_1$  e  $\hat{Y}_1 = 4 + 3X$
- c)  $t = 6,364$ , significativo (a região de rejeição é  $t \geq 2,920$  ).
- 9.3. b)  $\hat{\alpha} = -8,529$  e  $\hat{\beta} = 4,044$  são estimativas consistentes.
- $\hat{\gamma} = 2,208$  e  $\hat{\delta} = 1,088$  são estimativas consistentes, não-tendenciosas e de variância mínima.
- 9.4. a) Sim.
- b)  $\hat{\alpha}_1 = 3,5$ ;  $\hat{\gamma}_1 = -1,5$ ;  $\hat{\beta}_1 = 1$
- $s_1^2 = 1,6$
- Matriz das estimativas das variâncias e covariâncias assintóticas de  $\hat{\alpha}_1$ ,  $\hat{\gamma}_1$  e  $\hat{\beta}_1$ :
- $$\begin{bmatrix} 1,64 & -0,80 & 0,24 \\ -0,80 & 0,80 & -0,40 \\ 0,24 & -0,40 & 0,24 \end{bmatrix}$$
- c)  $\hat{\alpha}_2 = -0,5$ ;  $\hat{\gamma}_2 = 0,5$ ;  $\hat{\beta}_2 = 1$
- $s_1^2 = 1,6$
- Matriz das estimativas das variâncias e covariâncias assintóticas de  $\hat{\alpha}_2$ ,  $\hat{\gamma}_2$  e  $\hat{\beta}_2$ :
- $$\begin{bmatrix} 3,56 & -0,64 & 0,72 \\ -0,64 & 0,16 & 0,08 \\ -0,72 & 0,08 & 0,24 \end{bmatrix}$$
- 9.5.  $c = 11$  e  $d = 4$
- 9.6. a)  $b = 0,9$                   b) Estimador inconsistente, tendendo a superestimar  $\beta$ .
- c)  $\hat{\beta} = 0,75$  e  $\hat{\alpha} = 6$                   s)  $s(\hat{\beta}) = 0,1768$
- 9.7. a) A primeira equação é exatamente identificável. A segunda equação é obviamente identificável, pois não há variável endógena no segundo membro.
- b)  $\hat{\gamma} = -1$ ,  $\hat{\delta} = 4$ ,  $\hat{\alpha} = -15$  e  $\hat{\beta} = 6$
- c)  $s(\hat{\beta}) = 0,433$
- 9.8. a) Ocorre superidentificação na 1<sup>a</sup> equação. A 2<sup>as</sup> equação é obviamente identificável, pois não há variável endógena no 2<sup>o</sup> membro.

- b)  $\hat{\alpha} = -5$  e  $\hat{\beta} = 2$
- c)  $t = 24,495$ , significativo ( $t_0 = 4,032$ )
- 9.9. a) Há superidentificação se  $\beta_2 \neq 0$  e  $\beta_3 \neq 0$ .
- b) É exatamente identificável se  $\alpha_2 \neq 0$ .
- c) Equação de demanda estimada

$$\hat{Y}_1 = 15 - Y_2 + 2X_1$$

$$s_1^2 = 4$$

Matriz das estimativas das variâncias e covariâncias assintóticas de  $\hat{\alpha}_1$  e  $\hat{\alpha}_2$ :

$$\frac{1}{259} \begin{bmatrix} 160 & -152 \\ -152 & 248 \end{bmatrix}$$

d) Equação de oferta estimada:

$$\hat{Y}_1 = Y_2 + 0,2X_2 - 0,5X_3$$

$$s_2^2 = 1$$

Matriz das estimativas das variâncias e covariâncias assintóticas de  $\hat{\beta}_1$ ,  $\hat{\beta}_2$  e  $\hat{\beta}_3$ :

$$\frac{1}{67600} \begin{bmatrix} 6800 & 660 & -1500 \\ 660 & 243 & -245 \\ -1500 & -245 & 1325 \end{bmatrix}$$

e) -0,45, 0,7 e 0,45

## 10. SÉRIES TEMPORAIS

### 10.1. Processos estocásticos

Uma *série temporal* é um conjunto de valores de uma variável ordenados no tempo. Exemplos: a série de temperaturas máximas diárias em determinado posto meteorológico, a série dos preços médios mensais de milho em determinado mercado e a série de valores anuais do PIB brasileiro.

Para uma conceituação mais formal, é necessário considerar o *processo estocástico* subjacente. Dado um conjunto  $T$ , um processo estocástico é uma família  $Y = \{Y(t), t \in T\}$ , tal que, para cada  $t \in T$ ,  $Y(t)$  é uma variável aleatória. O conjunto  $T$  é, normalmente, o conjunto dos números inteiros  $\{0, \pm 1, \pm 2, \dots\}$  ou o conjunto dos números reais. A figura abaixo ilustra esta interpretação de um processo estocástico (Moretin e Toloi, p. 17):

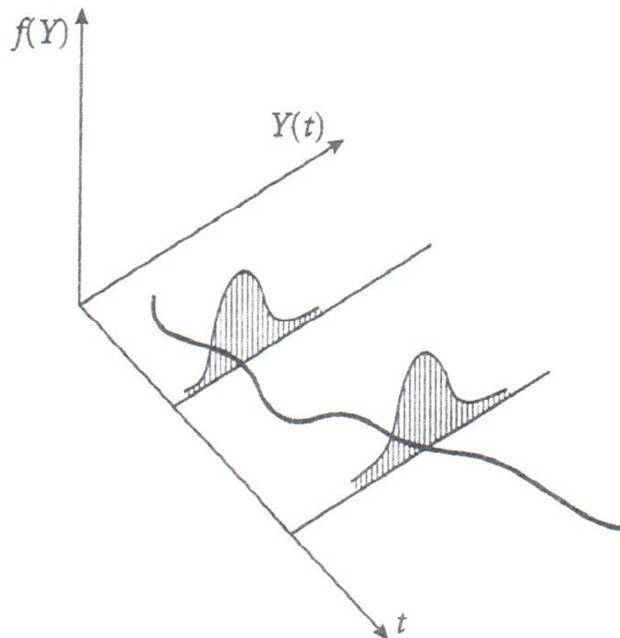


Figura 10.1. Um processo estocástico.

Alternativamente, um processo estocástico pode ser interpretado como uma família de trajetórias ou realizações do processo, como ilustra a figura a seguir:

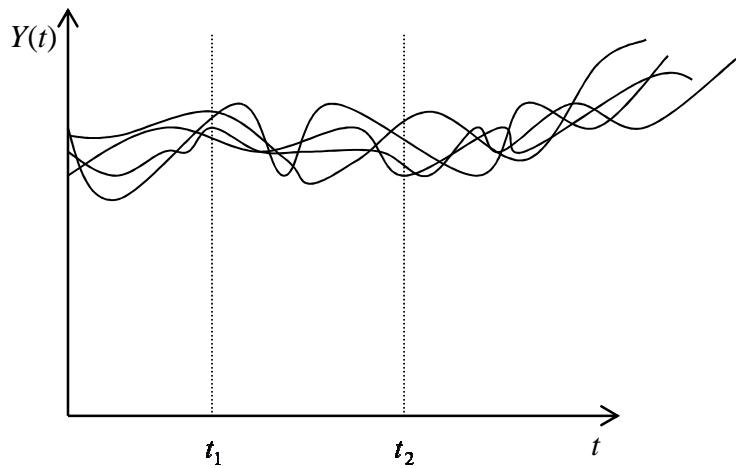


Figura 10.2. Conjunto de trajetórias

Imaginando um conjunto infinito de trajetórias, um “corte” no instante  $t$  permitiria obter a distribuição de  $Y(t)$  naquele instante.

Uma série temporal é uma trajetória ou realização de um processo estocástico. Aqui serão analisadas apenas séries temporais de valores  $Y_t$  equiespaçados no tempo (séries completas de dados anuais ou mensais, por exemplo).

Um exemplo de processo estocástico é um conjunto infinito de séries formadas pelos resultados de 100 lançamentos consecutivos de um dado.

Note-se que o critério de ordenação dos  $Y(t)$ , na definição de um processo estocástico, não precisa ser o tempo. Mas é certo que na grande maioria das aplicações  $t$  representa o tempo.

Um processo estocástico é caracterizado pelas distribuições conjuntas de  $Y(t_1)$ ,  $Y(t_2)$ , ...,  $Y(t_n)$  para qualquer conjunto finito de valores de  $t$  ( $t_1, t_2, \dots, t_n$ ) pertencentes a  $T$ . Na prática, a análise usualmente fica limitada a três tipos de parâmetros:

- a) as médias  $\mu(t) = E[Y(t)]$
- b) as variâncias  $\sigma_t^2 = V[Y(t)] = E[Y(t) - \mu(t)]^2$
- c) as covariâncias  $\gamma(t_1, t_2) = \text{cov}[Y(t_1), Y(t_2)]$

Note que  $\sigma_t^2 = \gamma(t, t)$

Um processo estocástico é *estritamente estacionário* se as distribuições conjuntas de  $Y(t_1)$ ,  $Y(t_2)$ , ...,  $Y(t_n)$  não são afetadas por translações no tempo, isto é,

se a distribuição conjunta de  $Y(t_1)$ ,  $Y(t_2)$ , ...,  $Y(t_n)$  é idêntica à distribuição conjunta de  $Y(t_1 + k)$ ,  $Y(t_2 + k)$ , ...,  $Y(t_n + k)$ .

No caso de um processo estacionário, a escolha da origem no eixo  $t$  não afeta nenhuma característica do processo.

Na prática é impossível conhecer todas as distribuições conjuntas de  $Y(t_1)$ ,  $Y(t_2)$ , ...,  $Y(t_n)$  e ficamos restritos ao conceito de processo *fracamente estacionário* (ou estacionário de segunda ordem), que é aquele que obedece às seguintes condições:

- a) a média de  $Y(t)$  é constante, isto é,  $\mu(t) = \mu$  para todo  $t$ .
- b) a variância de  $Y(t)$  é constante, isto é,  $V[Y(t)] = \sigma^2$  para todo  $t$ .
- c)  $\gamma(t_1, t_2)$  é uma função de  $t_2 - t_1$ , isto é, a covariância entre dois  $Y(t)$  depende apenas da *defasagem* entre eles.

Essa última condição permite que, no caso de um processo estacionário, a covariância entre  $Y(t)$  e  $Y(t+k)$  seja indicada por  $\gamma_k$ . Consequentemente, a variância de  $Y(t)$  pode ser indicada por  $\gamma_0$ .

## 10.2. Ruído branco

Por simplicidade, a partir desse ponto passamos a representar os valores da série temporal por  $Y_t$ , em lugar de  $Y(t)$ .

Denomina-se *ruído branco* uma série temporal  $(a_t)$  com média igual a zero, variância constante e sem covariância entre valores referentes a dois momentos distintos, isto é,

$$E(a_t) = 0$$

$$E(a_t^2) = \sigma_a^2$$

$$\gamma_k = \text{cov}(a_t, a_{t+k}) = 0 \quad \text{para } k \neq 0$$

Sejam  $X_t$  os valores obtidos em lançamentos consecutivos de um dado não-chumbado. Então a série de valores de  $Y_t = X_t - 3,5$  é um exemplo de ruído branco.

### 10.3. Modelos de regressão

Um possível modelo para uma série temporal é  $Y_t = \varphi(t) + a_t$ ,  $t = 1, \dots, n$ , onde  $a_t$  é um ruído branco não correlacionado com  $\varphi(t)$ , que é uma determinada função do tempo (a parcela sistemática ou determinística de  $Y_t$ ).

Um exemplo bastante simples é o modelo de uma regressão linear simples

$$Y_t = \alpha + \beta t + a_t$$

Se  $Y_t$  é uma série de dados mensais e incluirmos variáveis binárias para captar uma possível variação estacional, o modelo fica

$$Y_t = \alpha + \beta t + \sum_{h=2}^{12} \gamma_h Z_h + a_t,$$

com  $Z_h = 1$  no  $h$ -ésimo mês do ano e  $Z_h = 0$  nos demais meses do ano.

### 10.4. Modelos de decomposição

O modelo clássico de decomposição de uma série temporal baseia-se na pressuposição de que ela pode ser separada em parcelas. Tipicamente são considerados 3 componentes: a tendência ( $D_t$ ), a variação cíclica sazonal ( $S_t$ ) e a parte aleatória ( $a_t$ ). Então

$$Y_t = D_t + S_t + a_t$$

Se os componentes forem multiplicativos o modelo fica

$$Y_t = D_t S_t a_t$$

Neste caso usam-se os logaritmos para obter um modelo aditivo.

Uma exposição da técnica usada para separar o componente estacional pode ser encontrada no capítulo 20 do livro “Estatística para Economistas” (Hoffmann, 2006).

### 10.5. Modelos ARMA

Em comparação com os modelos de regressão e a técnica de decomposição das séries temporais, os modelos ARMA, descritos nesta seção e nas seguintes, tiveram desenvolvimento relativamente recente. O trabalho pioneiro foi o livro de Box e

Jenkins, intitulado “Time Series Analysis: Forecasting and Control”, cuja 1<sup>a</sup> edição foi publicada em 1970. Há uma 3<sup>a</sup> edição, com Reinsel como novo co-autor, publicada em 1994.

O modelo de um processo auto-regressivo de primeira ordem, indicado por AR(1), é

$$Y_t = \alpha + \phi Y_{t-1} + a_t, \quad (10.1)$$

em que  $a_t$  é um ruído branco e  $\alpha$  e  $\phi$  são parâmetros.

O modelo de um AR( $p$ ) (processo auto-regressivo de ordem  $p$ ) é

$$Y_t = \alpha + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + a_t$$

ou

$$Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - \dots - \phi_p Y_{t-p} = \alpha + a_t$$

Usando o operador de defasagem  $B$ , tal que  $B^k(Y_t) = Y_{t-k}$ , o modelo pode ser escrito

$$(1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p)Y_t = \alpha + a_t \quad (10.2)$$

ou, sinteticamente,

$$\phi(B)Y_t = \alpha + a_t$$

Por definição, o modelo de um processo de médias móveis de primeira ordem, indicado por MA(1) (devido à expressão em inglês “moving average”), é

$$Y_t = \alpha + a_t - \theta a_{t-1} \quad (10.3)$$

O modelo de um MA( $q$ ) (processo de médias móveis de ordem  $q$ ) fica

$$Y_t = \alpha + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_q a_{t-q}$$

ou

$$Y_t = \alpha + (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q)a_t \quad (10.4)$$

ou

$$Y_t = \alpha + \theta(B)a_t$$

O nome “médias móveis” está consagrado, embora não seja estritamente correto, uma vez que não se trata de uma média móvel dos  $a_t$  (pois os coeficientes  $1, -\theta_1, -\theta_2, \dots, -\theta_q$  não precisam ser positivos nem ter soma igual a 1).

Combinando (10.2) e (10.4) obtemos o modelo de um ARMA( $p, q$ ):

$$(1 - \phi_1 B - \dots - \phi_p B^p)Y_t = \alpha + (1 - \theta_1 B - \dots - \theta_q B^q)a_t \quad (10.5)$$

## 10.6. Análise do AR(1)

De acordo com (10.1) temos que

$$Y_{t-1} = \alpha + \phi Y_{t-2} + a_{t-1}$$

Substituindo essa expressão em (10.1), obtemos

$$Y_t = \alpha + \phi\alpha + \phi^2 Y_{t-2} + a_t + \phi a_{t-1}$$

Mas  $Y_{t-2}$  pode ser substituído por  $\alpha + \phi Y_{t-3} + a_{t-2}$ , obtendo-se

$$Y_t = \alpha + \phi\alpha + \phi^2\alpha + \phi^3 Y_{t-3} + a_t + \phi a_{t-1} + \phi^2 a_{t-2}$$

Após sucessivas substituições desse tipo, obtemos

$$Y_t = \alpha + \alpha\phi + \alpha\phi^2 + \dots + \alpha\phi^m + \phi^{m+1} Y_{t-m-1} + a_t + \phi a_{t-1} + \phi^2 a_{t-2} + \dots + \phi^m a_{t-m} \quad (10.6)$$

com  $m$  arbitrariamente grande.

Se  $|\phi| < 1$ , podemos desprezar o termo em  $Y_{t-m-1}$  e obtemos (no limite, para  $m \rightarrow \infty$ )

$$Y_t = \frac{\alpha}{1-\phi} + a_t + \phi a_{t-1} + \phi^2 a_{t-2} + \dots \quad (10.7)$$

Como  $a_t$  é um ruído branco, obtemos

$$E(Y_t) = \frac{\alpha}{1-\phi} \quad (10.8)$$

e

$$V(Y_t) = (1 + \phi^2 + \phi^4 + \dots) \sigma_a^2$$

ou

$$V(Y_t) = \frac{1}{1-\phi^2} \sigma_a^2 \quad (10.9)$$

Pode-se verificar que

$$\text{cov}(Y_t, Y_{t-k}) = \frac{\phi^k}{1-\phi^2} \sigma_a^2 = \gamma_k \quad (10.10)$$

As expressões (10.8), (10.9) e (10.10) mostram que um AR(1) com  $|\phi| < 1$  tem média constante, variância finita e constante e que a covariância entre  $Y_t$  e  $Y_{t-k}$  depende da defasagem  $k$  mas não depende de  $t$ . Conclui-se que um AR(1) com  $|\phi| < 1$  é (fracamente) estacionário.

As expressões (10.1) ou (10.6) mostram que um AR(1) com  $\phi > 1$  é “explosivo”. Tanto os valores absolutos de  $Y_t$  como sua variância crescem ilimitadamente. O processo não é estacionário.

### 10.7. O passeio aleatório com deslocamento

No caso particular em que  $\phi = 1$  o modelo AR(1) fica

$$Y_t = Y_{t-1} + \alpha + a_t \quad (10.11)$$

Esse processo é denominado *passeio aleatório com deslocamento*. Em cada período o valor da variável é acrescido de  $\alpha$  e de um elemento aleatório  $a_t$ .

O *passeio aleatório* é o caso particular em que  $\alpha = 0$ :

$$Y_t = Y_{t-1} + a_t \quad (10.12)$$

Para um passeio aleatório com deslocamento, partindo de um valor inicial  $Y_0$  e aplicando sucessivamente a relação (10.11), obtemos

$$Y_1 = Y_0 + \alpha + a_1$$

$$Y_2 = Y_0 + 2\alpha + a_1 + a_2$$

e, generalizando,

$$Y_t = Y_0 + \alpha t + \sum_{j=1}^t a_j \quad (10.13)$$

O mesmo resultado pode ser obtido de (10.6) fazendo  $\phi = 1$  e  $m = t - 1$ .

A expressão (10.13) mostra que  $E(Y_t)$  cresce linearmente com  $t$  e que a variância da parcela aleatória também cresce sempre com  $t$ , mostrando que um AR(1) com  $\phi = 1$  não é estacionário.

Se substituirmos  $\alpha$  por  $\beta$ , o modelo do passeio aleatório com deslocamento fica

$$Y_t = Y_{t-1} + \beta + a_t \quad (10.14)$$

Se, além disso, definirmos  $u_t = \sum_{j=1}^t a_j$ , a expressão (10.13) fica

$$Y_t = Y_0 + \beta t + u_t \quad (10.15)$$

Esse modelo é enganadoramente semelhante ao modelo

$$Y_t = \alpha + \beta t + a_t \quad (10.16)$$

apresentado na seção 3. A diferença está na matriz de variâncias e covariâncias dos erros. Para o modelo (10.16) essa matriz é  $\mathbf{I}\sigma_a^2$ , ao passo que no caso do modelo (10.15), admitindo que os valores observados sejam  $Y_1, Y_2, \dots, Y_n$ , essa matriz é  $\mathbf{W}\sigma_a^2$  com

$$\mathbf{W} = \begin{bmatrix} 1 & 1 & 1 & 1 & \dots & 1 \\ 1 & 2 & 2 & 2 & \dots & 2 \\ 1 & 2 & 3 & 3 & \dots & 3 \\ 1 & 2 & 3 & 4 & \dots & 4 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & 2 & 3 & 4 & \dots & n \end{bmatrix} \quad (10.17)$$

Observa-se que no modelo (10.15) há heterocedasticidade e covariância entre os erros. Se formos estimar o parâmetro  $\beta$  em (10.15) por meio de uma análise de regressão, é necessário utilizar mínimos quadrados generalizados. Mas há uma maneira muito mais fácil de estimar  $\beta$ . De (10.14) segue-se que

$$\Delta Y_t = Y_t - Y_{t-1} = \beta + a_t \quad (10.18)$$

Dada a série de  $n$  valores de  $Y_t$ , podemos calcular  $Z_t = \Delta Y_t$  e a estimativa de  $\beta$  é, simplesmente,

$$b = \bar{Z} = \frac{1}{n-1} \sum_{t=2}^n Z_t \quad (10.19)$$

Pode-se verificar que

$$\sum_{t=2}^n Z_t = \sum_{t=2}^n (Y_t - Y_{t-1}) = Y_n - Y_1$$

Então

$$b = \frac{Y_n - Y_1}{n-1} \quad (10.20)$$

Essa equação mostra que a estimativa de  $\beta$  depende apenas dos valores inicial e final de  $Y_t$ . Isso ocorre porque todas as variações de  $Y_t$  em períodos intermediários foram sendo acumuladas, sem nenhuma perda, e estão contidas no valor final  $Y_n$ .

Se a série de valores de  $Y_t$  for analisada fazendo uma regressão de  $Y_t$  contra  $t$  usando as fórmulas de mínimos quadrados ordinários, será obtida uma estimativa não-tendenciosa mas ineficiente de  $\beta$ . O problema principal é que serão utilizadas fórmulas erradas para estimar as variâncias, fazendo com que os testes de hipótese não sejam válidos. Ao testar  $H_0 : \beta = 0$ , há uma grande probabilidade de obter um valor de  $t$  ou  $F$  significativo mesmo que a hipótese seja verdadeira.

O modelo (10.15), obtido de (10.14), é um processo estacionário nas diferenças (“difference stationary process” – DSP), pois as diferenças  $Z_t = \Delta Y_t$  constituem uma série estacionária, como mostra (10.18).

O modelo (10.16) gera um processo estacionário depois de eliminada a tendência (determinística)  $\alpha + \beta t$  (é um “trend stationary process” – TSP).

Um exemplo numérico simples permite ressaltar as diferenças entre os procedimentos estatísticos apropriados no caso dos modelos (10.15) e (10.16). Vamos considerar uma série com apenas 4 valores consecutivos de  $Y_t$ , dados na tabela a seguir:

TABELA 10.1. Série de 4 valores de  $Y_t$ .

$t = X$	$Y_t$
1	10
2	11
3	17
4	19

Admitindo que (10.16) seja o modelo que gerou os dados, a estimativa de  $\beta$  é

$$b = \frac{\sum xY}{\sum x^2} = \frac{16,5}{5} = 3,3 \quad (10.21)$$

O quadrado médio do resíduo é

$$s^2 = \frac{58,75 - 54,45}{2} = 2,15 \quad (10.22)$$

Para testar a hipótese  $H_0 : \beta = 0$  calculamos

$$t = \frac{3,3}{\sqrt{\frac{2,15}{5}}} = 5,032 \quad (10.23)$$

Ao nível de significância de 5%, o valor crítico de  $t$  é 4,303. O resultado é significativo. A esse nível de significância, rejeita-se  $H_0 : \beta = 0$ .

Vamos admitir, agora, que a série de 4 valores de  $Y_t$  foi gerada pelo modelo (10.14). Um procedimento apropriado para obter a estimativa de  $\beta$  é calcular os valores de  $Z_t = \Delta Y_t$  (que são 1, 6 e 2) e, de acordo com (10.19), obter sua média:

$$b = \bar{Z} = \frac{9}{3} = 3 \quad (10.24)$$

Pode-se verificar que o mesmo resultado é obtido utilizando a expressão (10.20).

A estimativa de  $\sigma_a^2$  é dada por

$$s^2 = \frac{\sum z^2}{2} = \frac{14}{2} = 7 \quad (10.25)$$

Lembrando a fórmula para variância de uma média, obtemos

$$\hat{V}(b) = \frac{7}{3} \quad (10.26)$$

Para testar  $H_0 : \beta = 0$ , calculamos

$$t = \frac{3}{\sqrt{\frac{7}{3}}} = 1,964 \quad (10.27)$$

O resultado não é significativo. Não se rejeita  $H_0 : \beta = 0$ .

É interessante verificar que o mesmo resultado é obtido ajustando uma regressão linear simples de  $Y_t$  contra  $t$ , mas tomando o cuidado de utilizar as fórmulas de mínimos quadrados generalizados, respeitando a estrutura da matriz de variâncias e covariâncias dos erros  $u_t$  do modelo (10.15) [Ver (10.17)]. Temos

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \quad \mathbf{W} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix}$$

Verifica-se que

$$\mathbf{W}^{-1} = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{W}^{-1}\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix} (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1} = \frac{1}{3} \begin{bmatrix} 4 & -1 \\ -1 & 1 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{W}^{-1}\mathbf{y} = \begin{bmatrix} 10 \\ 19 \end{bmatrix} \mathbf{b} = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1} \mathbf{X}'\mathbf{W}^{-1}\mathbf{y} = \begin{bmatrix} 7 \\ 3 \end{bmatrix}$$

$$s^2 = \frac{1}{n-2} (\mathbf{y}'\mathbf{W}^{-1}\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{W}^{-1}\mathbf{y}) = \frac{1}{2} (141 - 127) = 7$$

A estimativa de  $\beta$  é igual a 3, reproduzindo o resultado obtido em (10.24), e a estimativa da respectiva variância é 7/3, reproduzindo o resultado obtido em (10.26). É óbvio que para testar  $H_0 : \beta = 0$  seria obtido o mesmo valor já calculado em (10.27).

É claro que a estimativa de  $\beta$  por mínimos quadrados generalizados é uma complicação desnecessária. O método anterior, baseado nas diferenças  $\Delta Y_t$ , é muito mais simples. A finalidade da exposição foi deixar claro que quando se estima o parâmetro  $\beta$  no modelo (10.15) fazendo uma regressão por mínimos quadrados ordinários o erro consiste em utilizar um procedimento inapropriado à estrutura dos erros  $u_t$ , não havendo nenhuma incompatibilidade entre a análise de séries temporais e o método de mínimos quadrados ou outros procedimentos “clássicos” da econometria.

### 10.8. Transformando modelos AR em modelos MA e vice-versa

O modelo de um processo auto-regressivo de primeira ordem pode ser escrito como

$$(1 - \phi B)Y_t = \alpha + a_t \tag{10.28}$$

Se esse processo for estacionário, isto é, se  $|\phi| < 1$ , de acordo com a expressão (10.7) ele pode ser escrito como

$$Y_t = \frac{\alpha}{1-\phi} + a_t + \phi a_{t-1} + \phi^2 a_{t-2} + \dots \quad (10.29)$$

Pode-se verificar que isso é o modelo de um MA de ordem infinita, com coeficientes que são todos potências de  $\phi$ .

A passagem de (10.28) para (10.29) foi feita anteriormente por meio de sucessivas substituições e manipulações algébricas. Ela pode ser feita mais rapidamente considerando que podemos dividir os dois membros de (10.28) por  $1 - \phi B$  e que

$$\frac{1}{1 - \phi B} = 1 + \phi B + \phi^2 B^2 + \phi^3 B^3 + \dots \quad (10.30)$$

Note-se que essa expressão é análoga à fórmula do limite da soma de uma progressão geométrica com razão menor do que 1:

$$\frac{a}{1 - q} = a + aq + aq^2 + aq^3 + \dots \quad (10.31)$$

Cabe ressaltar que a relação (10.31) *não* permite concluir que a relação (10.30) é correta, pois  $B$  é um *operador*, e não uma grandeza algébrica. Pode-se verificar, entretanto, que, a relação (10.30) é válida sempre que  $|\phi| < 1$ .

Sabemos que um MA(1) pode ser escrito como

$$Y_t = \alpha + (1 - \theta B) a_t \quad (10.32)$$

Se  $|\theta| < 1$ , podemos dividir todos os termos por  $1 - \theta B$ , obtendo

$$(1 + \theta B + \theta^2 B^2 + \dots) Y_t = \frac{\alpha}{1 - \theta} + a_t \quad (10.33)$$

Esse resultado mostra que um MA(1) é um AR de ordem infinita cujos coeficientes são todos potências de  $\theta$ .

Um MA(1) com  $|\theta| < 1$  é denominado *invertível*. Há uma clara simetria formal entre a condição para estacionariedade de um AR(1) e a condição de invertibilidade de um MA(1).

## 10.9. Raiz unitária e modelos ARIMA

Consideremos, novamente, o modelo de um AR(1):

$$(1 - \phi B)Y_t = \alpha + a_t$$

Vamos considerar a expressão que multiplica  $Y_t$  como um polinômio em  $B$  e definir a *equação característica*

$$1 - \phi B = 0$$

A raiz dessa equação é

$$B = \frac{1}{\phi}$$

Sabemos que o AR(1) é estacionário apenas quando  $|\phi| < 1$ . Em outras palavras, um AR(1) é estacionário se, e somente se, a raiz da respectiva equação característica é, em módulo, maior do que 1.

Para um AR(2) a equação característica é

$$1 - \phi_1 B - \phi_2 B^2 = 0$$

Neste caso as raízes podem ser números complexos e pode-se provar que a condição de estacionariedade é que as duas raízes estejam fora do círculo unitário (na representação geométrica dos números complexos  $h \pm vi$  em que  $h$  e  $v$  são, respectivamente, a abscissa e a ordenada do ponto em um sistema de eixos cartesianos ortogonais).

Genericamente, a condição de estacionariedade de um AR( $p$ ) é que todas as raízes da equação característica  $\phi(B) = 0$  estejam fora do círculo unitário.

Analogamente, um MA( $q$ ) é invertível se, e somente se, todas as raízes da equação característica  $\theta(B) = 0$  estiverem fora do círculo unitário.

Um caso especial de grande interesse é a existência de uma raiz unitária. O AR(1) com raiz unitária é o passeio aleatório com deslocamento analisado na seção anterior. Vimos que esse modelo exige procedimentos estatísticos específicos, passando-se a utilizar as diferenças  $\Delta Y_t = Y_t - Y_{t-1}$ . Sempre que uma série  $Y_t$  for não-

estacionária mas as diferenças  $Z_t = \Delta Y_t$  formarem uma série estacionária, dizemos que a série  $Y_t$  é integrada de primeira ordem ou I(1). Se a série  $Z_t = \Delta Y_t$  também for não-estacionária e a série  $\Delta Z_t = \Delta^2 Y_t$  for estacionária, dizemos que a série  $Y_t$  é integrada de segunda ordem ou I(2), e assim por diante. Uma série estacionária é I(0).

Se após  $d$  diferenças obtemos uma série que é um ARMA( $p, q$ ), a série original é um ARIMA ( $p, d, q$ ).

O modelo de um ARIMA (1, 1, 1), por exemplo, é

$$(1 - \phi B) \Delta Y_t = \alpha + (1 - \theta B) a_t$$

ou

$$(1 - \phi B)(1 - B) Y_t = \alpha + (1 - \theta B) a_t \quad (10.34)$$

### 10.10. Função de autocorrelação

Dada uma série temporal  $Y_t$ , sua autocorrelação com defasagem  $k$  é

$$\rho_k = \frac{\text{cov}(Y_t, Y_{t-k})}{\sqrt{V(Y_t)V(Y_{t-k})}} \quad (10.35)$$

Para uma série estacionária a covariância entre  $Y_t$  e  $Y_{t-k}$  não depende de  $t$  e será indicada por  $\gamma_k$ , e a variância, que é constante, será indicada por  $\gamma_0$ . Então a autocorrelação com defasagem  $k$  fica

$$\rho_k = \frac{\gamma_k}{\gamma_0} \quad (10.36)$$

A função de autocorrelação da série mostra como  $\rho_k$  varia com a defasagem  $k$ .

Para um processo AR(1) estacionário, substituindo (10.9) e (10.10) em (10.36), obtemos

$$\rho_k = \phi^k , \quad (10.37)$$

mostrando que neste caso o valor absoluto da autocorrelação cai exponencialmente com a defasagem .

Para um processo MA(1)

$$Y_t = \alpha + a_t - \theta a_{t-1}$$

verifica-se que  $\gamma_0 = (1 + \theta^2)\sigma_a^2$ ,  $\gamma_1 = -\theta\sigma_a^2$  e  $\gamma_k = 0$  para  $k > 1$ . Então

$$\rho_1 = \frac{-\theta}{1 + \theta^2} \text{ e } \rho_k = 0 \text{ para } k > 1 \quad (10.38)$$

Generalizando, pode-se demonstrar que para um MA( $q$ ) apenas as  $q$  primeiras autocorrelações são diferentes de zero.

Na análise de uma série temporal observada, é usual calcular as estimativas das autocorrelações. Indicando a estimativa de  $\gamma_k$  por  $c_k$  e a estimativa de  $\rho_k$  por  $r_k$ , temos

$$r_k = \frac{c_k}{c_0} , \quad (10.39)$$

com

$$c_k = \frac{1}{n} \sum_{t=k+1}^n (Y_t - \bar{Y})(Y_{t-k} - \bar{Y}) \quad (10.40)$$

onde  $\bar{Y}$  é a média dos valores de  $Y_t$  na amostra.

As autocorrelações estimadas podem ser utilizadas para identificar o processo que deu origem à série observada. Se, por exemplo, apenas  $r_1$  for claramente diferente de zero, o processo deve ser um MA(1).

Se os valores de  $r_k$  diminuem muito lentamente com  $k$ , isso indica que deve haver uma raiz unitária. Entretanto, quanto mais curta a série disponível, mais rapidamente diminuem os valores de  $r_k$ , mesmo que a série tenha uma raiz unitária. Na prática, pode ser impossível distinguir um passeio aleatório de um AR(1) cujo parâmetro  $\phi$  é menor, mas próximo de 1.

## 10.11. Os testes de Dickey-Fuller

O AR(1) definido por (10.1) pode ser escrito como

$$\Delta Y_t = \alpha + (\phi - 1)Y_{t-1} + a_t$$

Fazendo  $\phi - 1 = \delta$ , obtemos

$$\Delta Y_t = \alpha + \delta Y_{t-1} + a_t \quad (10.41)$$

Então a hipótese de que  $\phi = 1$  corresponde à hipótese de que  $\delta = 0$ , e  $\phi < 1$  corresponde a  $\delta < 0$ . O teste de Dickey-Fuller consiste em fazer uma regressão de  $\Delta Y_t$  contra  $Y_{t-1}$  e calcular, da maneira usual, o valor de  $t$  referente à hipótese  $H_0 : \delta = 0$ . Mas, dadas as características especiais da variável  $Y_t$  quando  $\phi = 1$ , os valores críticos para esse teste não são os da distribuição de  $t$  de Student. Para distinguir esse teste do  $t$  usual, é comum indicá-lo com a letra grega  $\tau$ .

Há três versões do teste de Dickey-Fuller:

- a) Teste  $\tau$ , quando se considera que a relação entre  $\Delta Y_t$  e  $Y_{t-1}$  não tem termo constante:

$$\Delta Y_t = \delta Y_{t-1} + a_t \quad (10.42)$$

- b) Teste  $\tau_\mu$ , quando se considera a relação (10.41).

- c) Teste  $\tau_\tau$ , quando é incluída uma tendência linear no tempo:

$$\Delta Y_t = \alpha + \beta t + \delta Y_{t-1} + a_t \quad (10.43)$$

Os valores críticos (com sinal negativo, para lembrar que se trata de teste unilateral à esquerda) são apresentados na tabela 10.2

TABELA 10.2. Valores críticos do teste de Dickey-Fuller

Modelo	Nível de significância		
	10%	5%	1%
$\Delta Y_t = \delta Y_{t-1} + a_t$	-1,62	-1,94	-2,56
$\Delta Y_t = \alpha + \delta Y_{t-1} + a_t$	-2,57	-2,86	-3,43
$\Delta Y_t = \alpha + \beta t + \delta Y_{t-1} + a_t$	-3,13	-3,41	-3,96

Fonte: Davidson e Mackinnon (1993).

Vamos admitir que o modelo subjacente seja um AR(2):

$$(1 - \phi_1 B - \phi_2 B^2)Y_t = \alpha + a_t \quad (10.44)$$

ou

$$Y_t = \alpha + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + a_t \quad (10.45)$$

Se existir uma raiz unitária teremos  $1 - \phi_1 - \phi_2 = 0$ .

A equação (10.45) pode ser escrita como

$$\Delta Y_t = \alpha + (\phi_1 + \phi_2 - 1)Y_{t-1} + \phi_2 \Delta Y_{t-1} + a_t$$

ou

$$\Delta Y_t = \alpha + \delta Y_{t-1} + \phi_2 \Delta Y_{t-1} + a_t \quad (10.46)$$

Quando a equação ajustada inclui termos com valores defasados de  $\Delta Y_t$  no segundo membro, o teste da hipótese de que  $\delta = 0$  é denominado teste de Dickey-Fuller aumentado (Augmented Dickey-Fuller test, ou ADF). Os valores críticos são os mesmos já apresentados na tabela 10.2).

## 10.12. Modelo de correção de erro e co-integração

Vamos considerar que temos duas variáveis econômicas,  $Y_{1t}$  e  $Y_{2t}$ , estreitamente associadas entre si, sendo que ambas são I(1), isto é, são integradas de primeira ordem. É interessante imaginar que  $Y_{1t}$  e  $Y_{2t}$  sejam os logaritmos dos preços de dois produtos substitutos próximos no consumo, como o preço, no varejo, da lata de óleo de milho e da lata de óleo de arroz. Outra alternativa é imaginar que se trata dos logaritmos dos preços de um mesmo produto em dois mercados próximos, como, por exemplo, o preço do feijão em Curitiba e em São Paulo. Nestes casos, embora tanto  $Y_{1t}$  como  $Y_{2t}$  sejam séries não-estacionárias, a diferença entre elas está limitada pela relação econômica entre ambas. Se uma delas subir ela “puxa” a outra. A imagem física seria de um elástico ou uma mola ligando as duas variáveis.

Generalizando um pouco, vamos considerar duas variáveis I(1) relacionadas pela equação

$$Y_{2t} = \alpha + \beta Y_{1t} + \varepsilon_t \quad (10.47)$$

Se o erro

$$\varepsilon_t = Y_{2t} - \alpha Y_{1t} \quad (10.48)$$

for elevado, indicando que o valor de  $Y_{2t}$  está relativamente elevado,  $Y_{2t}$  tende a diminuir e  $Y_{1t}$  tende a aumentar. Considerando que esses efeitos se manifestam no próximo período, temos

$$\Delta Y_{2t} = \phi_2 (Y_{2,t-1} - \alpha - \beta Y_{1,t-1}) + u_{2t} \quad (10.49)$$

e

$$\Delta Y_{1t} = \phi_1 (Y_{2,t-1} - \alpha - \beta Y_{1,t-1}) + u_{1t} \quad (10.50)$$

com  $\phi_2 < 0$  e  $\phi_1 > 0$ .

As equações (10.47), (10.49) e (10.50) constituem um exemplo simples de *modelo de correção de erro* envolvendo duas variáveis.

Como admitimos que  $Y_{1t}$  e  $Y_{2t}$  são variáveis I(1),  $\Delta Y_{1t}$  e  $\Delta Y_{2t}$  são variáveis estacionárias. Se admitirmos, ainda, que  $u_{2t}$  e  $u_{1t}$  são ruídos brancos, podemos afirmar que  $\phi_2 \varepsilon_t$  em (10.49) e  $\phi_1 \varepsilon_t$  em (10.50) são estacionários, pois uma combinação linear de variáveis estacionárias é sempre estacionária. Finalmente, se  $\phi_1 \neq 0$  ou  $\phi_2 \neq 0$  podemos concluir que  $\varepsilon_t$  é estacionário, ou seja, é uma variável I(0).

Verifica-se, portanto, que o modelo de correção de erros descrito implica a existência de uma combinação linear entre variáveis I(1) [a equação (10.48)] que é I(0). Dizemos, então, que  $Y_{1t}$  e  $Y_{2t}$  são variáveis *co-integradas* e que (10.48) [ou (10.47)] é a relação de co-integração.

Cabe ressaltar que a inclusão ou não do termo constante  $\alpha$  não afeta em nada a análise da estacionariedade das séries.

Podemos admitir que os valores de  $\Delta Y_{2t}$  e  $\Delta Y_{1t}$  também sejam afetados pelos valores prévios dessas diferenças. Então as equações (10.49) e (10.50) seriam substituídas por

$$\Delta Y_{2t} = \phi_2 (Y_{2,t-1} - \alpha - \beta Y_{1,t-1}) + \theta_{21} \Delta Y_{1,t-1} + \theta_{22} \Delta Y_{2,t-1} + u_{2t} \quad (10.51)$$

e

$$\Delta Y_{1t} = \phi_1(Y_{2,t-1} - \alpha - \beta Y_{1,t-1}) + \theta_{11} \Delta Y_{1,t-1} + \theta_{12} \Delta Y_{2,t-1} + u_{1t} \quad (10.52)$$

Como as diferenças defasadas também são estacionárias, sua inclusão nas equações não altera a conclusão anterior sobre a estacionariedade de  $\varepsilon_t$ .

O exemplo analisado ilustra um caso particular de co-integração. O conceito mais geral e sua relação com os modelos de correção de erro foi rigorosamente apresentado em artigo clássico de Engle e Granger (1987). Consideremos um processo envolvendo  $k$  variáveis ( $Y_{1t}, Y_{2t}, \dots, Y_{kt}$ ) e seja  $\mathbf{y}_t$  o vetor-coluna com os valores dessas variáveis no tempo  $t$ . Essas  $k$  variáveis são integradas de ordem  $d$ , se todas as  $k$  variáveis são  $I(d)$  e existe um vetor-coluna  $\boldsymbol{\beta}$ , com  $\boldsymbol{\beta} \neq \mathbf{0}$ , tal que  $\mathbf{y}'_t \boldsymbol{\beta}$  é  $I(d-c)$ , com  $c > 0$ . Em outras palavras, a ordem de integração da combinação linear  $\mathbf{y}'_t \boldsymbol{\beta}$  é menor do que a ordem de integração ( $d$ ) das variáveis em  $\mathbf{y}_t$ . O vetor  $\boldsymbol{\beta}$  é denominado *vetor de co-integração*. No caso do modelo de correção de erro analisado anteriormente temos  $d=1$  e  $c=1$ .

Quando duas variáveis são co-integradas, a relação de co-integração pode ser estimada pelo método de mínimos quadrados ordinários. Para testar se as variáveis são co-integradas devemos, inicialmente, verificar sua ordem de integração. Vamos admitir que testes de Dickey-Fuller tenham mostrado que as variáveis  $Y_{1t}$  e  $Y_{2t}$  são  $I(1)$ . Então elas são co-integradas se existe uma relação de co-integração cujo erro ( $\varepsilon_t$ ) é estacionário. Como não dispomos dos valores desse erro, o teste é feito com base nos resíduos da regressão de  $Y_{2t}$  contra  $Y_{1t}$ , que passamos a indicar por  $\hat{\varepsilon}_t$ . Da mesma maneira que no teste  $\tau$  de Dickey-Fuller, fazemos a regressão de  $\Delta \hat{\varepsilon}_t$  contra  $\hat{\varepsilon}_{t-1}$  e calculamos, da maneira usual, o valor de  $t$  referente à hipótese de que o parâmetro de  $\hat{\varepsilon}_{t-1}$  é igual a zero. O resultado deve ser comparado com os valores críticos apresentados na tabela 10.3. Note-se que esses valores críticos são diferentes daqueles que estão nas duas últimas linhas da tabela 10.2, pois neste caso a variável utilizada para fazer o teste já é o resíduo de uma regressão.

TABELA 10.3. Valores críticos para o teste da estacionariedade do erro de uma equação de co-integração de duas variáveis, incluindo termo constante.

Equação incluindo	Nível de significância		
	10%	5%	1%
Constante	-3,04	-3,34	-3,90
Constante e tendência	-3,50	-3,78	-4,32

Fonte: Davidson e Mackinnon (1993).

O fato de a variável utilizada na equação estimada para obter o valor do teste ser o resíduo da possível equação de co-integração faz com que, para a realização do teste, seja indiferente (assintoticamente) que o termo constante tenha sido incluído na primeira ou na segunda equação, o mesmo valendo para a inclusão de uma tendência.

Para exemplificar, vamos considerar os valores fictícios de  $Y_{1t}$  e  $Y_{2t}$  apresentados na tabela 10.4.

TABELA 10.4. Séries de 24 valores consecutivos de  $Y_{1t}$  e  $Y_{2t}$ .

$t$	$Y_{1t}$	$Y_{2t}$	$t$	$Y_{1t}$	$Y_{2t}$
1	0,77	0,62	13	0,86	0,87
2	0,79	0,71	14	0,98	0,77
3	0,81	0,68	15	1,02	0,81
4	0,84	0,61	16	1,11	0,84
5	0,89	0,71	17	1,07	0,91
6	0,95	0,73	18	1,13	0,87
7	0,98	0,72	19	1,22	1,01
8	0,95	0,92	20	1,30	1,09
9	1,04	0,78	21	1,34	1,21
10	0,96	0,80	22	1,45	1,10
11	0,99	0,64	23	1,47	1,14
12	0,87	0,83	24	1,40	1,16

Para a série  $Y_{1t}$  os testes  $\tau$ ,  $\tau_\mu$  e  $\tau_\tau$  de Dickey-Fuller são 2,00, -0,32 e -1,64, respectivamente. Trata-se de resultados claramente não-significativos (ver tabela 10.2), não se podendo rejeitar a hipótese de que a série dos  $Y_{1t}$  tem raiz unitária. Repetindo os testes para a série dos  $\Delta Y_{1t}$  obtemos  $\tau = -3,88$ ,  $\tau_\mu = -4,52$  e  $\tau_\tau = -4,42$ , significativos

ao nível de 1%, rejeitando-se a hipótese de raiz unitária na série das diferenças. Conclui-se que  $Y_{1t}$  é uma série I(1).

Para a série  $Y_{2t}$  obtemos  $\tau = 0,91$ ,  $\tau_\mu = -1,10$  e  $\tau_\tau = -3,11$ , todos não-significativos a 10%. Já para a série dos  $\Delta Y_{2t}$  obtemos  $\tau = -6,62$ ,  $\tau_\mu = -6,96$  e  $\tau_\tau = -6,90$ , todos significativos a 1%. Concluímos que a série  $Y_{2t}$  também é I(1).

Admitindo que haja co-integração entre as duas variáveis, estimamos, pelo método de mínimos quadrados ordinários, a seguinte equação (testes  $t$  entre parênteses, abaixo do coeficiente):

$$\hat{Y}_{2t} = 0,06289 - 0,75508 Y_{1t}$$

Para a série de 24 desvios ( $\hat{e}_t$ ) dessa regressão obtemos  $\tau = -5,66$ , significativo ao nível de 1% (ver tabela 10.3), o que permite concluir que esses desvios são estacionários. Tendo em vista que já havíamos verificado que as séries  $Y_{1t}$  e  $Y_{2t}$  são I(1), concluímos que  $Y_{1t}$  e  $Y_{2t}$  são co-integradas.

Considerando a versão mais simples do modelo de correção de erro descrito no início dessa seção (mas incluindo um termo constante nas equações), estimamos as seguintes equações:

$$\hat{\Delta}Y_{1t} = 0,028 + 0,486 \hat{\varepsilon}_t$$

(2,76)            (3,78)

e

$$\hat{\Delta}Y_{2t} = 0,022 - 0,822 \hat{\varepsilon}_t$$

(1,46)      (-4,33)

Cabe ressaltar que a natureza artificial dos dados desse exemplo fez com que os diversos testes conduzissem claramente à construção de um modelo de correção de erros. Na prática, os resultados geralmente não são tão evidentes e a construção de um modelo apropriado vai depender bastante do discernimento do pesquisador, combinando o conhecimento de técnicas estatísticas, de teoria econômica e das qualidades e limitações dos dados utilizados.

## Exercícios

10.1. Admite-se que a variável  $Y_t$  é um passeio aleatório com deslocamento:

$$Y_t = Y_{t-1} + \beta + \varepsilon_t,$$

sendo  $\beta$  uma constante e  $\varepsilon_t$  um ruído branco. É fornecida uma série de 9 valores consecutivos de  $Y_t$ : 6, 5, 10, 18, 22, 22, 31, 36 e 46.

- a) Teste, ao nível de significância de 5%, a hipótese de que  $\beta = 0$ .
- b) Teste, ao nível de significância de 5%, a hipótese de que  $\beta = 3$ .

10.2. É dada uma série de 5 valores consecutivos de  $Y_t$ : 17, 20, 32, 38 e 41.

Admite-se que esses valores foram gerados pelo modelo

$$Y_t = Y_{t-1} + \beta + \varepsilon_t,$$

sendo  $\varepsilon_t$  um ruído branco com variância  $\sigma^2$ .

- a) Estime  $\beta$ .
- b) Estime  $\sigma^2$ .
- c) Teste, ao nível de significância de 5%, a hipótese de que  $\beta = 0$ , contra a hipótese alternativa de que  $\beta > 0$ .

10.3. Admite-se que a variável  $Y_t$  é um passeio aleatório com deslocamento:

$$Y_t = Y_{t-1} + \beta + \varepsilon_t, \tag{1}$$

onde  $\beta$  é uma constante e  $\varepsilon_t$  é um ruído branco.

O modelo (1) permite deduzir que  $Y_t$  varia no tempo de acordo com a equação

$$Y_t = \alpha + \beta t + u_t$$

Se  $\sigma^2$  é a variância de  $\varepsilon_t$  e  $\mathbf{u}$  é o vetor-coluna dos valores de  $u_t$ , tem-se

$$E(\mathbf{u}\mathbf{u}') = \mathbf{V}\sigma^2.$$

É dada uma série de 9 valores consecutivos de  $Y_t$  (para  $t = 1, 2, \dots, 9$ ): 12, 12, 18, 27, 32, 33, 43, 49 e 60.

- a) Quais os valores dos elementos  $v_{22}$  e  $v_{48}$  da matriz  $\mathbf{V}$ ?
- b) Determine a estimativa linear não-tendenciosa de variância mínima de  $\beta$ .
- c) Teste, ao nível de significância de 1% a hipótese de que  $\beta = 0$ .
- d) Teste, ao nível de significância de 5% a hipótese de que  $\beta = 5$ .

10.4. Sendo  $a_t$  um ruído branco, verifique, para cada um dos modelos a seguir, se ele gera um processo estacionário ou não, justificando sumariamente sua resposta:

- a)  $Y_t = 18 - 2,3Y_{t-1} + a_t$
- b)  $Y_t = 144 + a_t + 2a_{t-1}$
- c)  $Y_t = 0,9Y_{t-1} + a_t$
- d)  $Y_t = 7 + Y_{t-1} + a_t$
- e)  $Y_t = 11 + 1,8Y_{t-1} - 0,8Y_{t-2} + a_t$

Mostre, inicialmente, que essa equação pode ser escrita como

$$(1 - 0,8B)(1 - B)Y_t = 11 + a_t$$

10.5. São dadas as séries de 24 valores consecutivos das variáveis  $Y_{1t}$  e  $Y_{2t}$ :

$t$	$Y_{1t}$	$Y_{2t}$	$t$	$Y_{1t}$	$Y_{2t}$
1	44	99	13	47	95
2	56	90	14	26	96
3	60	103	15	41	81
4	64	105	16	38	105
5	53	108	17	51	103
6	71	108	18	57	119
7	60	100	19	51	131
8	59	98	20	60	116
9	37	107	21	73	115
10	54	99	22	61	124
11	49	106	23	61	136
12	48	104	24	71	132

- a) Usando um programa para computador apropriado, faça os testes de Dickey-Fuller para verificar se a série  $Y_{1t}$  têm uma raiz unitária.
- b) Idem, para a série  $\Delta Y_{1t}$ .
- c) Idem, para as séries  $Y_{2t}$  e  $\Delta Y_{2t}$ .
- d) O que se pode concluir sobre a ordem de integração das séries  $Y_{1t}$  e  $Y_{2t}$ ?
- e) Estime a equação  $Y_{2t} = \alpha + \gamma t + \beta Y_{1t} + \varepsilon_t$  e verifique se esta é uma relação de co-integração.

## Respostas

10.1 a)  $b = 5$ ,  $s^2(b) = 2$ ,  $t = 3,536$ , significativo ( $t_0 = 2,365$ )

b)  $t = 1,414$ , não-significativo ( $t_0 = 2,365$ )

10.2. a)  $b = 6$

b)  $s^2 = 18$

c)  $t = 2,828$ , significativo (região de rejeição:  $t \geq 2,353$ )

10.3. a)  $\nu_{22} = 2$  e  $\nu_{48} = 4$

b)  $b = 6$

c)  $t = 4,243$ , significativo ( $t_0 = 3,499$ )

d)  $t = 0,707$ , não-significativo ( $t_0 = 1,895$ )

10.4. a) Não-estacionário: AR(1) com  $|\phi| > 1$ .

b) Estacionário: MA(1).

c) Estacionário: AR(1) com  $|\phi| < 1$ .

d) Não-estacionário: passeio aleatório com deslocamento ou AR(1) com  $\phi = 1$ .

e) Não-estacionário: tem raiz unitária.

10.5.a)  $\tau = 0,03$ ,  $\tau_\mu = -2,53$  e  $\tau_\tau = -2,48$ , não-significativos.

b)  $\tau = -7,32$ ,  $\tau_\mu = -7,20$  e  $\tau_\tau = -7,08$ , todos significativos a 1%.

c) Para  $Y_{2t}$ :  $\tau = 0,51$ ,  $\tau_\mu = -1,44$  e  $\tau_\tau = -2,38$ , não-significativos

Para  $\Delta Y_{2t}$ :  $\tau = -6,10$ ,  $\tau_\mu = -6,24$  e  $\tau_\tau = -6,12$ , significativos a 1%

d) Tanto  $Y_{1t}$  como  $Y_{2t}$  são I(1).

e)  $\hat{Y}_{2t} = 69,09 + 1,128t + 0,4517Y_{1t}$ , com  $R^2 = 0,562$ .

Para a série dos desvios dessa regressão obtém-se  $\tau = -5,05$ . Como o valor crítico ao nível de 1%, na tabela 10.3, é  $-4,32$ , conclui-se que  $Y_{1t}$  e  $Y_{2t}$  são séries co-integradas.

## APÊNDICE

TABELA I. Distribuição de  $t$  de Student. Valor crítico  $t_0$  tal que

$$P(t > t_0) = P(t < -t_0) = \alpha / 2$$

Número de Graus de Liberdade	Nível de significância para o teste bilateral ( $\alpha$ )					
	0,20	0,10	0,05	0,02	0,01	0,005
1	3,078	6,314	12,706	31,821	63,657	127,32
2	1,886	2,920	4,303	6,965	9,925	14,089
3	1,638	2,353	3,182	4,541	5,841	7,453
4	1,533	2,132	2,776	3,747	4,604	5,598
5	1,476	2,015	2,571	3,365	4,032	4,773
6	1,440	1,943	2,447	3,143	3,707	4,317
7	1,415	1,895	2,365	2,998	3,499	4,029
8	1,397	1,860	2,306	2,896	3,355	3,832
9	1,383	1,833	2,262	2,821	3,250	3,690
10	1,372	1,812	2,228	2,764	3,169	3,581
11	1,363	1,796	2,201	2,718	3,106	3,497
12	1,356	1,782	2,179	2,681	3,055	3,428
13	1,350	1,771	2,160	2,650	3,012	3,372
14	1,345	1,761	2,145	2,624	2,977	3,326
15	1,341	1,753	2,131	2,602	2,947	3,286
16	1,337	1,746	2,120	2,583	2,921	3,252
17	1,333	1,740	2,110	2,567	2,898	3,222
18	1,330	1,734	2,101	2,552	2,878	3,197
19	1,328	1,729	2,093	2,539	2,861	3,174
20	1,325	1,725	2,086	2,528	2,845	3,153
21	1,323	1,721	2,080	2,518	2,831	3,135
22	1,321	1,717	2,074	2,508	2,819	3,119
23	1,319	1,714	2,069	2,500	2,807	3,104
24	1,318	1,711	2,064	2,492	2,797	3,090
25	1,316	1,708	2,060	2,485	2,787	3,078
26	1,315	1,706	2,056	2,479	2,779	3,067
27	1,314	1,703	2,052	2,473	2,771	3,056
28	1,313	1,701	2,048	2,467	2,763	3,047
29	1,311	1,699	2,045	2,462	2,756	3,038
30	1,310	1,697	2,042	2,457	2,750	3,030
40	1,303	1,684	2,021	2,423	2,704	2,971
60	1,296	1,671	2,000	2,390	2,660	2,915
120	1,289	1,658	1,980	2,358	2,617	2,860
$\infty$	1,282	1,645	1,960	2,326	2,576	2,807

Interpolações devem ser feitas com base nos recíprocos dos graus de liberdade (interpolação harmônica).

Fonte: Theil (1971), p. 717, e Hoel (1968), p. 295.

TABELA II. Distribuição de qui-quadrado. Valor crítico  $\chi_0^2$  tal que  
 $P(\chi_k^2 > \chi_0^2) = \alpha$

Número de Graus de Liberdade ( $k$ )	$\alpha$					
	0,995	0,975	0,050	0,025	0,010	0,005
1	$3927 \cdot 10^{-8}$	$9821 \cdot 10^{-7}$	3,841	5,024	6,635	7,879
2	0,010025	0,05064	5,991	7,378	9,210	10,60
3	0,07172	0,2158	7,815	9,348	11,34	12,84
4	0,2070	0,4844	9,488	11,14	13,28	14,86
5	0,4117	0,8312	11,07	12,83	15,09	16,75
6	0,6757	1,237	12,59	14,45	16,81	18,55
7	0,9893	1,690	14,07	16,01	18,48	20,28
8	1,344	2,180	15,51	17,53	20,09	21,96
9	1,735	2,700	16,92	19,02	21,67	23,59
10	2,156	3,247	18,31	20,48	23,21	25,19
11	2,603	3,816	19,68	21,92	24,72	26,76
12	3,074	4,404	21,03	23,34	26,22	28,30
13	3,565	5,009	22,36	24,74	27,69	29,82
14	4,075	5,629	23,68	26,12	29,14	31,32
15	4,601	6,262	25,00	27,49	30,58	32,80
16	5,142	6,908	26,30	28,85	32,00	34,27
17	5,697	7,564	27,59	30,19	33,41	35,72
18	6,265	8,231	28,87	31,53	34,81	37,16
19	6,844	8,907	30,14	32,85	36,19	38,58
20	7,434	9,591	31,41	34,17	37,57	40,00
21	8,034	10,28	32,67	35,48	38,93	41,40
22	8,643	10,98	33,92	36,78	40,29	42,80
23	9,260	11,69	35,17	38,08	41,64	44,18
24	9,886	12,40	36,42	39,36	42,98	45,56
25	10,52	13,12	37,65	40,65	44,31	46,93
26	11,16	13,84	38,89	41,92	45,64	48,29
27	11,81	14,57	40,11	43,19	46,96	49,64
28	12,46	15,31	41,34	44,46	48,28	50,99
29	13,12	16,05	42,56	45,72	49,59	52,34
30	13,79	16,79	43,77	46,98	50,89	53,67
40	20,71	24,43	55,76	59,34	63,69	66,77
50	27,99	32,36	67,50	71,42	76,15	79,49
60	35,53	40,48	79,08	83,30	88,38	91,95
70	43,28	48,76	90,53	95,02	100,4	104,2
80	51,17	57,15	101,9	106,6	112,3	116,3
90	59,20	65,65	113,1	118,1	124,1	128,3
100	67,33	74,22	124,3	129,6	135,8	140,2

Fonte: Theil (1971), p. 718-719.

TABELA III – Distribuição de  $F$ . Valor crítico  $F_0$  tal que  $P(F > F_0) = 0,01$ .

Nº de graus de liberdade do denominador	Número de graus de liberdade do numerador																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	$\infty$
1	4052	5000	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,05	26,87	26,69	26,50	26,41	26,32	26,22	26,13	
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	4,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
$\infty$	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

Interpolações devem ser feitas com base nos recíprocos dos graus de liberdade (interpolação harmônica).

Fonte: Christ (1966, p. 671) e Pimentel Gomes (1966, p. 408-409).

TABELA IV – Distribuição de  $F$ . Valor crítico  $F_0$  tal que  $P(F > F_0) = 0,05$ .

Nº de graus de liberdade do denominador	Número de graus de liberdade do numerador																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	$\infty$
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	2,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
$\infty$	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

Interpolações devem ser feitas com base nos recíprocos dos graus de liberdade (interpolação harmônica).

Fonte: Christ (1966, p. 670) e Pimentel Gomes (1966, p. 406-407).

Tabela V. Distribuição de  $F$ . Valor crítico  $F_0$  tal que  $P(F > F_0) = 0,10$ .

Nº de graus de liberdade do denominador	Número de graus de liberdade do numerador																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	$\infty$
1	39,9	49,5	53,6	55,8	57,2	58,2	58,9	59,4	59,9	60,2	60,7	61,2	61,7	62,0	62,3	62,5	62,8	63,1	63,3
2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,41	9,42	9,44	9,45	9,46	9,47	9,47	9,48	9,49
3	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,22	5,20	5,18	5,18	5,17	5,16	5,15	5,14	5,13
4	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,90	3,87	3,84	3,83	3,82	3,80	3,79	3,78	3,76
5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,27	3,24	3,21	3,19	3,17	3,16	3,14	3,12	3,10
6	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94	2,90	2,87	2,84	2,82	2,80	2,78	2,76	2,74	2,72
7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,67	2,63	2,59	2,58	2,56	2,54	2,51	2,49	2,47
8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,50	2,46	2,42	2,40	2,38	2,36	2,34	2,32	2,29
9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,38	2,34	2,30	2,28	2,25	2,23	2,21	2,18	2,16
10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,28	2,24	2,20	2,18	2,16	2,13	2,11	2,08	2,06
11	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25	2,21	2,17	2,12	2,10	2,08	2,05	2,03	2,00	1,97
12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,15	2,10	2,06	2,04	2,01	1,99	1,96	1,93	1,90
13	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14	2,10	2,05	2,01	1,98	1,96	1,93	1,90	1,88	1,85
14	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10	2,05	2,01	1,96	1,94	1,91	1,89	1,86	1,83	1,80
15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06	2,02	1,97	1,92	1,90	1,87	1,85	1,82	1,79	1,76
16	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03	1,99	1,94	1,89	1,87	1,84	1,81	1,78	1,75	1,72
17	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00	1,96	1,91	1,86	1,84	1,81	1,78	1,75	1,72	1,69
18	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00	1,98	1,93	1,89	1,84	1,81	1,78	1,75	1,72	1,69	1,66
19	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98	1,96	1,91	1,86	1,81	1,79	1,76	1,73	1,70	1,67	1,63
20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94	1,89	1,84	1,79	1,77	1,74	1,71	1,68	1,64	1,61
21	2,96	2,57	2,36	2,23	2,14	2,08	2,02	1,98	1,95	1,92	1,88	1,83	1,78	1,75	1,72	1,69	1,66	1,62	1,59
22	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93	1,90	1,86	1,81	1,76	1,73	1,70	1,67	1,64	1,60	1,57
23	2,94	2,55	2,34	2,21	2,11	2,05	1,99	1,95	1,92	1,89	1,84	1,80	1,74	1,72	1,69	1,66	1,62	1,59	1,55
24	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91	1,88	1,83	1,78	1,73	1,70	1,67	1,64	1,61	1,57	1,53
25	2,92	2,53	2,32	2,18	2,09	2,02	1,97	1,93	1,89	1,87	1,82	1,77	1,72	1,69	1,66	1,63	1,59	1,56	1,52
26	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88	1,86	1,81	1,76	1,71	1,68	1,65	1,61	1,58	1,54	1,50
27	2,90	2,51	2,30	2,17	2,07	2,00	1,95	1,91	1,87	1,85	1,80	1,75	1,70	1,67	1,64	1,60	1,57	1,53	1,49
28	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87	1,84	1,79	1,74	1,69	1,66	1,63	1,59	1,56	1,52	1,48
29	2,89	2,50	2,28	2,15	2,06	1,99	1,93	1,89	1,86	1,83	1,78	1,73	1,68	1,65	1,62	1,58	1,55	1,51	1,47
30	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82	1,77	1,72	1,67	1,64	1,61	1,57	1,54	1,50	1,46
40	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76	1,71	1,66	1,61	1,57	1,54	1,51	1,47	1,42	1,38
60	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74	1,71	1,66	1,60	1,54	1,51	1,48	1,44	1,40	1,35	1,29
120	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,68	1,65	1,60	1,55	1,48	1,45	1,41	1,37	1,32	1,26	1,19
$\infty$	2,71	2,30	2,08	1,94	1,85	1,77	1,72	1,67	1,63	1,60	1,55	1,49	1,42	1,38	1,34	1,30	1,24	1,17	1,00

Interpolações devem ser feitas com base nos recíprocos dos graus de liberdade (interpolação harmônica).

Fonte: Scheffé (1959), p. 424-425.

Tabela VI. Valores críticos do teste de Durbin-Watson para o nível de significância de 5%.

Nº de observações (n)	Número de variáveis explanatórias									
	k = 1		k = 2		k = 3		k = 4		k = 5	
	$d_L$	$d_U$	$d_L$	$d_U$	$d_L$	$d_U$	$d_L$	$d_U$	$d_L$	$d_U$
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,88
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
31	1,36	1,50	1,30	1,57	1,23	1,65	1,16	1,74	1,09	1,83
32	1,37	1,50	1,31	1,57	1,24	1,65	1,18	1,73	1,11	1,82
33	1,38	1,51	1,32	1,58	1,26	1,65	1,19	1,73	1,13	1,81
34	1,39	1,51	1,33	1,58	1,27	1,65	1,21	1,73	1,15	1,81
35	1,40	1,52	1,34	1,58	1,28	1,65	1,22	1,73	1,16	1,80
36	1,41	1,52	1,35	1,59	1,29	1,65	1,24	1,73	1,18	1,80
37	1,42	1,53	1,36	1,59	1,31	1,66	1,25	1,72	1,19	1,80
38	1,43	1,54	1,37	1,59	1,32	1,66	1,26	1,72	1,21	1,79
39	1,43	1,54	1,38	1,60	1,33	1,66	1,27	1,72	1,22	1,79
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79
45	1,48	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,78
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77
55	1,53	1,60	1,49	1,64	1,45	1,68	1,41	1,72	1,38	1,77
60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77
65	1,57	1,63	1,54	1,66	1,50	1,70	1,47	1,73	1,44	1,77
70	1,58	1,64	1,55	1,67	1,52	1,70	1,49	1,74	1,46	1,77
75	1,60	1,65	1,57	1,68	1,54	1,71	1,51	1,74	1,49	1,77
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77
85	1,62	1,67	1,60	1,70	1,57	1,72	1,55	1,75	1,52	1,77
90	1,63	1,68	1,61	1,70	1,59	1,73	1,57	1,75	1,54	1,78
95	1,64	1,69	1,62	1,71	1,60	1,73	1,58	1,75	1,56	1,78
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78

Fonte: Johnston (1972), p. 430.

Tabela VII. Valores críticos do teste de Durbin-Watson para o nível de significância de 1%.

Nº de observações (n)	Número de variáveis explanatórias									
	k = 1		k = 2		k = 3		k = 4		k = 5	
	$d_L$	$d_U$	$d_L$	$d_U$	$d_L$	$d_U$	$d_L$	$d_U$	$d_L$	$d_U$
15	0,81	1,07	0,70	1,25	0,59	1,46	0,49	1,70	0,39	1,96
16	0,84	1,09	0,74	1,25	0,63	1,44	0,53	1,66	0,44	1,90
17	0,87	1,10	0,77	1,25	0,67	1,43	0,57	1,63	0,48	1,85
18	0,90	1,12	0,80	1,26	0,71	1,42	0,61	1,60	0,52	1,80
19	0,93	1,13	0,83	1,26	0,74	1,41	0,65	1,58	0,56	1,77
20	0,95	1,15	0,86	1,27	0,77	1,41	0,68	1,57	0,60	1,74
21	0,97	1,16	0,89	1,27	0,80	1,41	0,72	1,55	0,63	1,71
22	1,00	1,17	0,91	1,28	0,83	1,40	0,75	1,54	0,66	1,69
23	1,02	1,19	0,94	1,29	0,86	1,40	0,77	1,53	0,70	1,67
24	1,04	1,20	0,96	1,30	0,88	1,41	0,80	1,53	0,72	1,66
25	1,05	1,21	0,98	1,30	0,90	1,41	0,83	1,52	0,75	1,65
26	1,07	1,22	1,00	1,31	0,93	1,41	0,85	1,52	0,78	1,64
27	1,09	1,23	1,02	1,32	0,95	1,41	0,88	1,51	0,81	1,63
28	1,10	1,24	1,04	1,32	0,97	1,41	0,90	1,51	0,83	1,62
29	1,12	1,25	1,05	1,33	0,99	1,42	0,92	1,51	0,85	1,61
30	1,13	1,26	1,07	1,34	1,01	1,42	0,94	1,51	0,88	1,61
31	1,15	1,27	1,08	1,34	1,02	1,42	0,96	1,51	0,90	1,60
32	1,16	1,28	1,10	1,35	1,04	1,43	0,98	1,51	0,92	1,60
33	1,17	1,29	1,11	1,36	1,05	1,43	1,00	1,51	0,94	1,59
34	1,18	1,30	1,13	1,36	1,07	1,43	1,01	1,51	0,95	1,59
35	1,19	1,31	1,14	1,37	1,08	1,44	1,03	1,51	0,97	1,59
36	1,21	1,32	1,15	1,38	1,10	1,44	1,04	1,51	0,99	1,59
37	1,22	1,32	1,16	1,38	1,11	1,45	1,06	1,51	1,00	1,59
38	1,23	1,33	1,18	1,39	1,12	1,45	1,07	1,52	1,02	1,58
39	1,24	1,34	1,19	1,39	1,14	1,45	1,09	1,52	1,03	1,58
40	1,25	1,34	1,20	1,40	1,15	1,46	1,10	1,52	1,05	1,58
45	1,29	1,38	1,24	1,42	1,20	1,48	1,16	1,53	1,11	1,58
50	1,32	1,40	1,28	1,45	1,24	1,49	1,20	1,54	1,16	1,59
55	1,36	1,43	1,32	1,47	1,28	1,51	1,25	1,55	1,21	1,59
60	1,38	1,45	1,35	1,48	1,32	1,52	1,28	1,56	1,25	1,60
65	1,41	1,47	1,38	1,50	1,35	1,53	1,31	1,57	1,28	1,61
70	1,43	1,49	1,40	1,52	1,37	1,55	1,34	1,58	1,31	1,61
75	1,45	1,50	1,42	1,53	1,39	1,56	1,37	1,59	1,34	1,62
80	1,47	1,52	1,44	1,54	1,42	1,57	1,39	1,60	1,36	1,62
85	1,48	1,53	1,46	1,55	1,43	1,58	1,41	1,60	1,39	1,63
90	1,50	1,54	1,47	1,56	1,45	1,59	1,43	1,61	1,41	1,64
95	1,51	1,55	1,49	1,57	1,47	1,60	1,45	1,62	1,42	1,64
100	1,52	1,56	1,50	1,58	1,48	1,60	1,46	1,63	1,44	1,65

Fonte: Johnston (1972), p. 431.

## BIBLIOGRAFIA

- AIGNER, D.J. (1971). *Basic econometrics*. New Jersey, Prentice-Hall.
- ANGRIST, J.D.; PISCHKE, J.S. (2009). *Mostly harmless econometrics: an empiricist's companion*. Princeton University Press.
- BARTLETT, M.S. (1949). Fitting a Straight Line when Both Variables are subject to Error. *Biometrics*, 5: 207-242.
- BOX, G.E.P.; JENKINS, G. M. e REINSEL, G.C. (1994) *Time series analysis - Forecasting and Control*. 3<sup>a</sup> ed. New York, Prentice-Hall.
- CARTER, H.O. e HARTLEY, H.O. (1958). A Variance Formula for Marginal Productivity Estimates Using the Cobb-Douglas Function. *Econometrica*, 26: 306-313.
- CASTRO, J. de (1961). *Geopolítica da Fome: ensaio sobre os problemas de alimentação e de população do mundo*. 6<sup>a</sup> ed. São Paulo, Editora Brasiliense.
- COCHRAN, W.G. (1965). *Técnicas de Amostragem*. Rio de Janeiro, Fundo de Cultura.
- CHRIST, C.F. (1966). *Econometric Models and Methods*. New York, John Wiley.
- CROXTON, F.E. e COWDEN, D.J. (1955). *Applied General Statistics*. New York, Prentice-Hall.
- DAVIDSON, R e MACKINNON, J. (1993). *Estimation and Inference in Econometrics*. New York, Oxford University Press.
- DRAPER, N. e SMITH, H. (1966). *Applied Regression Analysis*. New York, John Wiley.
- ENDERS, W. (1995). *Applied Econometric Time Series*. New York, John Wiley.
- ENGLE, R.F. e GRANGER, W.J. (1987). Co-integration and Error Correction: representation, estimation and testing. *Econometrica*, 55(2): 251-276.
- GLEJSER, H. (1969). A New Test for Heteroscedasticity. *J. Am. Statist. Assoc.*, 64: 316-323.
- GOLDFELD, S.M. e QUANDT, R.E. (1965). Some Tests for Homoscedasticity. *J. Am. Statist. Assoc.*, 60: 539-547.

- GREENE, W.H. (2000). *Econometric Analysis*. 4<sup>a</sup> ed. Prentice-Hall.
- HADLEY, G. (1967). *Introduction to Probability and Statistical Decision Theory*, San Francisco, Holden-Day.
- HEADY, E.O. e DILLON, J.L. (1961). *Agricultural Production Functions*. Ames, Iowa State University Press.
- HILL, R.C.; GRIFFITHS, W.E. e JUDGE, G.G. (2003). *Econometria*. 2<sup>a</sup> ed. São Paulo, Editora Saraiva.
- HOEL, P.G. (1962). *Introduction to Mathematical Statistics*, 3<sup>a</sup>ed. New York, John Wiley.
- HOEL, P.G. (1968). *Estatística Elementar*, 2<sup>a</sup> ed. Rio de Janeiro, Fundo de Cultura.
- HOFFMANN, R. (2006) *Estatística para economistas*. 4<sup>a</sup> ed. rev. e ampl.. São Paulo, Pioneira Thomson Learning.
- HOOD, W.C. e KOOPMANS, T.C., ed. (1953). *Studies in Econometric Method*. New Haven, Yale University Press.
- HUANG, D.S. (1970). *Regression and Econometric Methods*. New York, John Wiley.
- JOHNSTON, J. (1971). *Métodos Econométricos*. São Paulo, Atlas.
- JOHNSTON, J. (1972). *Econometric Methods*, 2<sup>a</sup> ed. New York, McGraw-Hill.
- JOHNSTON, J. e DINARDO, J. (1997). *Econometric Methods*. 4<sup>a</sup> ed. McGraw-Hill.
- KELEJIAN, H.H. e OATES, W.E. (1978). *Introdução à Econometria: Princípios e aplicações*. Rio de Janeiro, Campus.
- KMENTA, J. (1971). *Elements of Econometrics*. New York, McMillan.
- LANGE, O. (1967). *Introdução à Econometria*, 2<sup>a</sup> ed. Rio de Janeiro, Fundo de Cultura.
- LÜTKEPOHL, H. (1991). *Introduction to Multiple Time Series Analysis*. Berlin, Springer-Verlag.
- MADDALA, G.S. (1988). *Introduction to Econometrics*. New York, MacMillan.
- MALINVAUD, E. (1970). *Statistical Methods of Econometrics*, 2<sup>a</sup> ed. Amsterdam, North-Holland.

- MARINHO, E.; ARAÚJO, J. (2010). Pobreza e o sistema de seguridade social rural no Brasil. *Revista Brasileira de Economia* 64(2): 161-174.
- MARINHO, E.; LINHARES, F.; CAMPELO, G. (2011). Os programas de transferência de renda do governo impactam a pobreza no Brasil? *Revista Brasileira de Economia* 65(3): 267-288.
- MORETTIN, P.A. E TOLOI, C.M.C. (1985). *Previsão de Séries Temporais*. São Paulo, Atual Editora.
- NERLOVE, M. (1958). *The Dynamics of Supply: Estimation of Farmers' Response to Price*. Baltimore, John Hopkins.
- PEREZ, M.C.R.C. (1973). *Contribuição ao Estudo da Elasticidade-renda do Consumo de Alimentos*. Piracicaba, ESALQ-USP (Dissertação de Mestrado).
- PHILLIPS, P.C.B. e WICKENS, M.R. (1978). *Exercises in Econometrics*, volume two. Oxford, Philip Allan Publishers.
- PIMENTEL GOMES, F. (1966). *Curso de Estatística Experimental*. 3<sup>a</sup> ed. Piracicaba, ESALQ-USP.
- PIMENTEL GOMES, F. e NOGUEIRA, I.R. (1964). *Rregressão e Covariância*. Piracicaba, ESALQ-USP (mimeografado).
- SCHEFFÉ, H. (1959). *The Analysis of Variance*. New York, John Wiley.
- SILVA LEME, R.A. da (1965). *Curso de Estatística*. 2<sup>a</sup> ed. Rio de Janeiro, Livro Técnico.
- THEIL, H. (1971). *Principles of Econometrics*. New York, John Wiley.
- WALD, A. (1940). The Fitting of Straight Lines in both Variables are Subject to Error. *Ann. Math. Statist.*, 11:284-300.
- WALLIS, K.F. (1967). Lagged Dependent Variables and Serially Correlated Errors: A Reappraisal of three-Pass Least Squares. *Rev. Economics and Statistics*, 49:555-567.
- WHITE, H. (1980). A Heteroskedasticity-consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica* 48(4): 817-838,

WONNACOTT, R.J. e WONNACOTT, T.H. (1970). *Econometrics*. New York, John Wiley.

WONNACOTT, R.J. e WONNACOTT, T.H. (1976). *Econometria*. Rio de Janeiro, Livros Técnicos e Científicos.

YAMANE, T. (1967). *Statistics, an Introductory Analysis*. 2<sup>a</sup> ed. New York, Harper and Row.

YULE, G.U. e KENDALL, M.G. (1940). *An introduction to the Theory of Statistics*. 12<sup>a</sup> ed. London, Griffin.

## ÍNDICE ANALÍTICO

### A

- ADF, 368
- Anamorfose, 77-79, 173
- AR ( $p$ ), 356
- ARIMA ( $p, d, q$ ), 365
- ARMA ( $p, q$ ), 357
- Auto-regressivo, 356
- Autocorrelação, 46, 278

### B

- Bartlett, 303, 383
- Basmann, 328
- Binária (ver Variável binária)
- Box, 355, 383
- Breusch-Pagan/Godfrey, 265

### C

- Chebyshev, 28
- Christ, 378, 379, 383
- Cochran, 13, 383
- Coeficiente de correlação
  - parcial, 146-153
  - simples, 103-104, 186
- Coeficiente de determinação, 61, 68, 105, 129-130
- Coeficiente de determinação parcial, 150-153
- Coeficiente de variação, 68-69
- Consistente (ver Estimador consistente)
- Co-integração, 368-371
- Convergência em média quadrática, 28

Convergência em probabilidade, 26

Covariância

Definição, 6

Propriedades, 10

Cramér-Rao, 32-34

## D

Davidson, 268, 367, 371, 383

Desigualdade de Chebyshev, 28

Draper, 232, 285, 383

DSP, 360

Dummy variable (ver Variável binária)

Durbin-Watson, 284-286, 381, 382

Dickey-Fuller, 367-368

## E

Econometria, 1

Eficiência relativa, 16

Engle, 370, 383

Equações estruturais, 309, 318-319

Erro tipo I, 34

Erro tipo II, 34

Especificação, 77-79, 168-171

Esperança matemática

Definição, 5

Propriedades, 5

Estimador

assintoticamente eficiente, 25-26

assintoticamente não-tendencioso, 25

consistente, 27, 31, 291-294

de máxima verossimilhança, 21-24, 32-34, 80

de mínimos quadrados, 19-21

de variância mínima, 15-18

de Theil-Basmann, 328

de White, 265-269

eficiente, 16, 34

imparcial, 10

não-tendencioso, 10-12

não-viesado, 10

Extrapolação, 76-77, 142

## F

Falta de ajustamento, 236-239

Forma reduzida (de um sistema de equações simultâneas), 309, 319

## G

Gauss-Markov, 60, 121

Glejser, 264-265

Goldfeld, 263-266

Granger, 370, 383

## H

Heterocedasticia, 46, 254

Hipótese (ver teste de hipóteses)

Hoel, 262, 285, 376, 384

Hoffmann, 4, 285, 355, 384

Homocedasticia, 44

Homocedásticos, 44, 120

## I

Identificação, 321-327

Intervalo de confiança, 71, 73, 75-76, 129, 140-141

Intervalo de previsão, 73-76, 141-142

**J**

Jenkins, 356, 383

Johnston, 281, 285, 297, 304, 381, 382, 384

**K**

Kendall, 108, 386

**L**

Limite em probabilidade, 26

Limite inferior de Cramér-Rao, 32-34

**M**

Mackinnon, 268, 367, 371, 383

MA ( $q$ ), 356

Matriz de variâncias e covariâncias, 124

Médias móveis, 356

Método

- das variáveis instrumentais (ver Variável instrumental)
- de máxima verossimilhança, 19, 21-24
- dos mínimos quadrados, 19-21, 47, 121-123
- dos mínimos quadrados em dois estágios, 315-317, 328-329
- dos mínimos quadrados generalizados, 275-278
- dos mínimos quadrados indiretos, 312-313
- dos mínimos quadrados ordinários, 256, 257-261, 282-283
- dos mínimos quadrados ponderados, 46, 256-258

Modelo de correção de erros, 368-370

Modelo

Estatístico, 2-3

Matemático, 1-2

de uma regressão linear simples, 44-47  
de uma regressão linear múltipla, 120-121  
restrito, 178

Mudança estrutural, 230-232

Multicolinearidade, 174-178, 191

## N

Nível de significância, 34-40

## O

Ortogonalidade, 173, 182-183

## P

Passeio aleatório, 358

Perez, 304, 385

Pimentel Gomes, 378, 385

Poder do teste, 34

Poligonal, 226-229

Probabilidade caudal do teste, 67-71, 137, 138, 284

Processo estacionário, 353-354

Processo estocástico, 352-353

## Q

Quandt, 263-266

## R

Raiz unitária, 364

Regressão linear múltipla, 120-121

Regressão linear simples, 44-47

Ruído branco, 354

**S**

Scheffé, 380, 385

Smith, 232, 285, 383

**T**

Teorema de Gauss-Markov (ver Gauss-Markov)

Teste de Breusch-Pagan/Godfrey, 265

Teste de Chow, 232

Teste de Dickey-Fuller, 367-368

Teste de Durbin-Watson, 284-286

Teste de Glejser, 264

Teste de Goldfeld e Quandt, 263-266

Teste de hipóteses (conceitos básicos), 34-40

Teste de hipóteses no modelo linear, 157, 178-181

Teste de White, 266

Teste para “falta de ajustamento” (Ver Falta de ajustamento)

Teste para homocedasticia, 261-267

Theil, 33, 34, 263, 267, 281, 328, 376, 377, 385

Trajetória (de um processo estocástico), 352-353

TSP, 360

**V**

Variância

Definição, 5-6

Propriedades, 6-7

Variância assintótica, 25

Variáveis conjuntamente determinadas, 317

Variável

aleatória, 4-5

binária, 219-240

- defasada, 285, 317
- endógena, 308
- exógena, 308
- instrumental, 295-297, 301-302, 312, 319-320
- predeterminada, 317

**W**

- Wald, 303, 385
- White, 266-268, 385
- Wonnacott, 191, 385

**Y**

- Yule, 108, 386