

REGRESSÃO E CORRELAÇÃO

CAPÍTULO 1

INTRODUÇÃO À ESTATÍSTICA MULTIVARIADA

1. Introdução

A Estatística Multivariada consiste em um conjunto de métodos estatísticos utilizados em situações nas quais várias variáveis são medidas simultaneamente, em cada elemento amostral. Em geral, as variáveis são correlacionadas entre si e quanto maior o número de variáveis, mais complexa torna-se a análise por métodos comuns de estatística univariada. Embora historicamente o uso de métodos multivariados esteja relacionado com trabalhos na Psicologia, Ciências Sociais e Biológicas, mais recentemente eles têm sido aplicados em um grande universo de áreas diferentes, como: Educação, Geologia, Química, Física, Engenharia, Ergonomia etc. Esta expansão na aplicação dessas técnicas somente foi possível graças ao grande avanço da tecnologia computacional e ao grande número de *softwares* estatísticos com módulos de análise multivariada implementados.

Basicamente, a Estatística Multivariada se divide em dois grupos: um primeiro, consistindo em técnicas exploratórias de sintetização (ou simplificação) da estrutura de variabilidade dos dados, e um segundo, consistindo em técnicas de inferência estatística. Fazem parte do primeiro grupo métodos como a análise de componentes principais, análise fatorial, análise de correlações canônicas, análise de agrupamentos, análise discriminante e análise de correspondência. Esses métodos têm um apelo prático muito interessante, pois, na sua grande maioria, independem do conhecimento da forma matemática da distribuição de probabilidades geradora dos dados amostrais. No segundo grupo, encontram-se os métodos de estimação de parâmetros, testes de hipóteses, análise de variância, de covariância e de regressão multivariadas. Em termos de literatura, muito se evoluiu ao longo dos anos, sendo que hoje existem livros de excelente qualidade tratando apropriadamente dos dois grupos de técnicas.

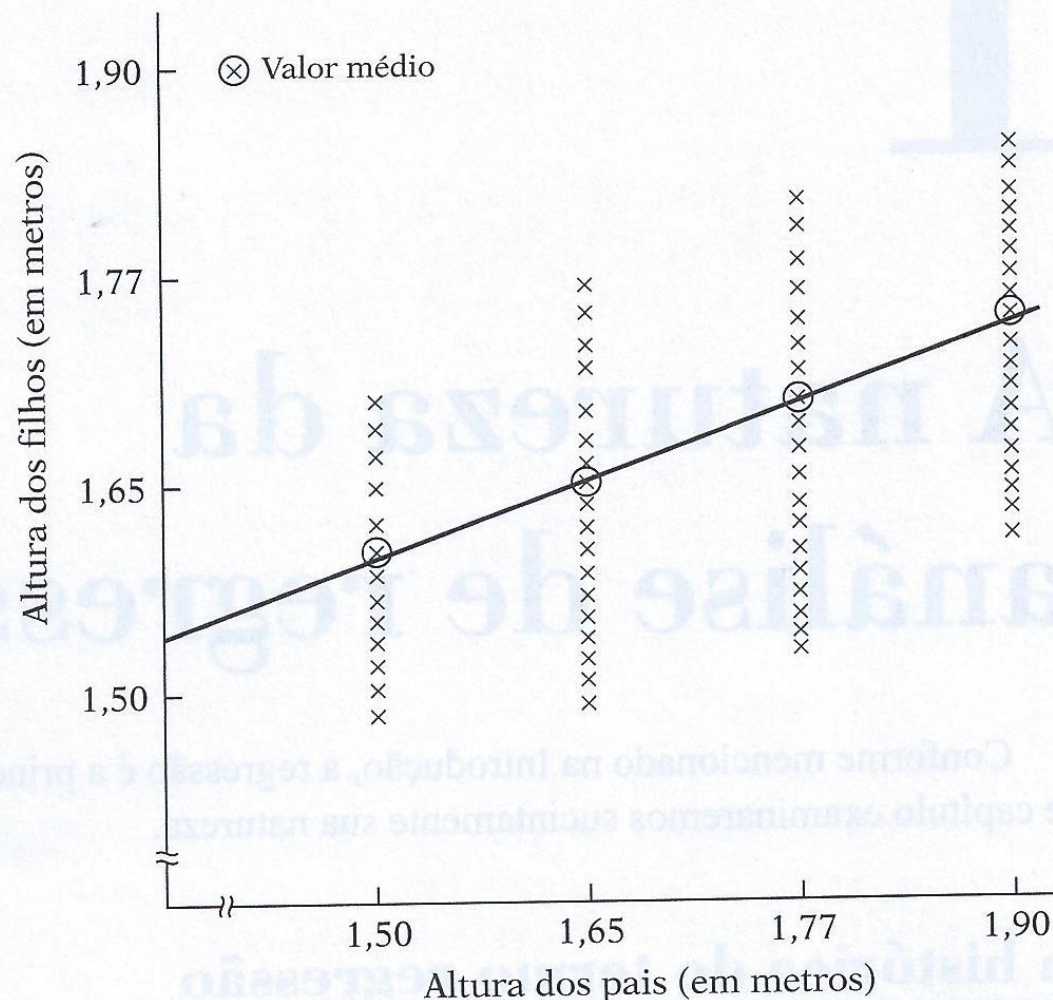
Basicamente, a Estatística Multivariada se divide em dois grupos: um primeiro, consistindo em técnicas exploratórias de sintetização (ou simplificação) da estrutura de variabilidade dos dados; e um segundo, consistindo em técnicas de inferência estatística. Fazem parte do primeiro grupo métodos como a análise de componentes principais, análise fatorial, análise de correlações canônicas, análise de agrupamentos, análise discriminante e análise de correspondência. Esses métodos têm um apelo prático muito interessante, pois, na sua grande maioria, independem do conhecimento da forma matemática da distribuição de probabilidades geradora dos dados amostrais. No segundo grupo, encontram-se os métodos de estimação de parâmetros, testes de hipóteses, análise de variância, de covariância e de regressão multivariadas. Em termos de literatura, muito se evoluiu ao longo dos anos, sendo que hoje existem livros de excelente qualidade tratando apropriadamente dos dois grupos de técnicas.

A interpretação moderna da regressão

- A análise de regressão diz respeito ao estudo da dependência de uma variável, a variável dependente, em relação a uma ou mais variáveis explanatórias, visando estimar e/ou prever o valor médio (da população) da primeira em termos dos valores conhecidos ou fixados (em amostragens repetidas) das segundas.

FIGURA 1.1

Distribuição hipotética das alturas dos filhos em relação à altura dos pais.



Galton, F. Family likeness in stature. **Proceedings of Royal Society**, Londres, v.40, p. 42 – 72, 1886.

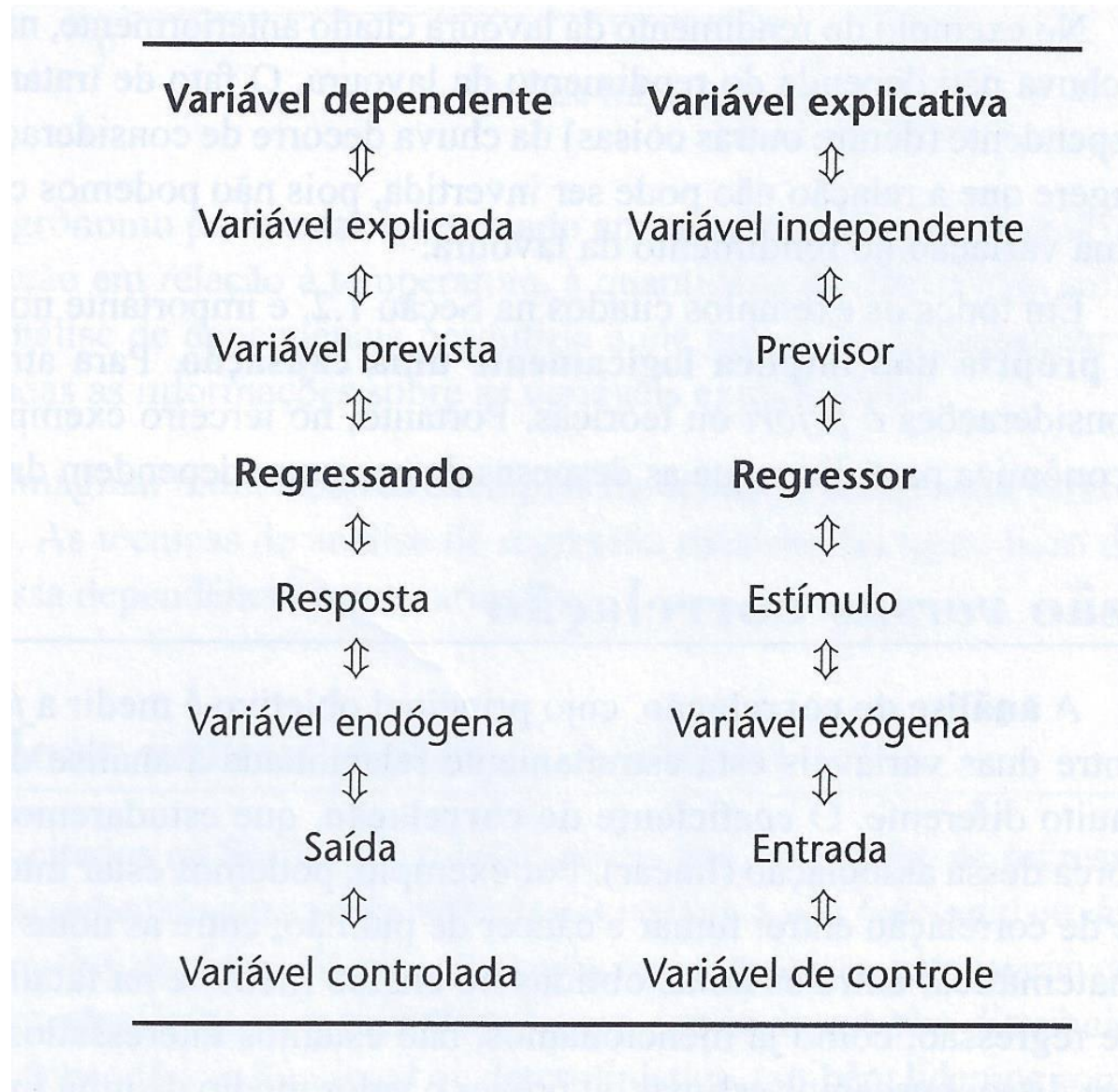
Regressão *versus* Correlação

- O coeficiente de correlação mede a força da associação linear entre duas variáveis:
 - Podemos estar interessados em determinar a correlação entre as notas de matemática e de estatística, p.ex. (vice-versa)
- A análise de regressão mede e descreve a forma de relacionamento entre duas variáveis e permite a estimativa dos valores de uma variável em relação aos valores das demais.
- Busca-se estimar ou prever o valor médio de uma variável com base em valores fixos de outras variáveis:
 - É possível prever a nota média de uma prova de estatística conhecendo-se as notas do estudante na prova de matemática?

Regressão *versus* Correlação

- Na análise de correlação não há distinção entre as variáveis dependentes e explanatórias.
- A teoria da **correlação** baseia-se, especialmente, na premissa da **aleatoriedade das variáveis**, enquanto boa parte da teoria da **regressão** está condicionada à premissa de que **a variável dependente é estocástica, mas as variáveis explanatórias são fixas.**

Terminologia e notação



Regressão linear simples

- Equação

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

- Y e X representam as variáveis dependentes e independentes, respectivamente;
- β é o coeficiente de regressão na população amostrada;
- α é o valor previsto para Y na população quando X é igual a zero
- i indica o i th par de dados X e Y na amostra

Equação de regressão múltipla

- Em algumas situações, Y pode ser considerado dependente em relação a mais do que uma variável.

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j}$$

- Neste caso, Y é linearmente dependente em relação à variável X_1 e também em relação à variável X_2 . Aqui i denota a i th variável independente, e X_{ij} a j th observação da variável i .

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j}$$

- Neste caso, há uma variável dependente e duas variáveis independentes.
- β_1 indica quanto varia Y com a mudança de uma unidade em X_1 , se X_2 for mantido constante. Isto é, β_1 é uma medida do relacionamento de Y com X_1 após controlar o efeito de X_2 .
- Similarmente, β_2 descreve a taxa de mudança em Y à medida que X_2 varia, com X_1 permanecendo constante.
- Os dois parâmetros da população, β_1 e β_2 , são chamados de coeficientes de regressão parcial porque cada um expressa somente parte do relacionamento de dependência.

- A equação de regressão linear simples representa uma linha em duas dimensões

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

- A equação de regressão múltipla define um plano (neste caso, três dimensões)

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j}$$

- Uma regressão com m variáveis independentes define uma superfície m -dimensional, algumas vezes referidas como “superfície de resposta” ou hiperplano

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \varepsilon_j$$

- ε_j é o erro ou resíduo e representa a diferença entre o valor mensurado e o previsto com a equação.
- Teoricamente, em regressão múltipla não existe limite para m , porém, o número de variáveis independentes deve ser ($n \geq m+2$).
- **n é o número de pares de dados (observações)**
- **m = número de variáveis independentes**

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \cdots + \beta_m X_{mj} + \varepsilon_j$$

$$Y_j = \alpha + \sum_{i=1}^m \beta_i X_{ij} + \varepsilon_j$$

- Modelo **linear** múltiplo: arranjo aditivo dos parâmetros (α e β_j).
- Importante:
 - Pelo menos $m + 2$ pares de dados são requeridos para realizar a análise de regressão múltipla.

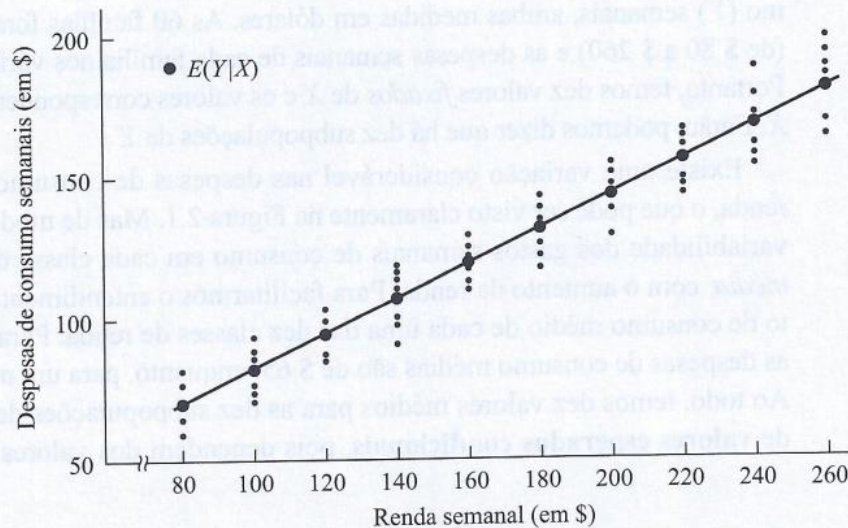
TABELA 2.1

Renda familiar
semanal, X , em \$

$Y \downarrow \quad X \rightarrow$	80	100	120	140	160	180	200	220	240	260
Despesas de consumo semanais das famílias, Y , em \$	55 60 65 70 75 – –	65 70 74 80 85 88 –	79 84 90 94 98 – –	80 93 95 103 108 113 115	102 107 110 116 118 125 –	110 115 120 130 135 140 –	120 136 140 144 145 – –	135 137 140 152 157 160 162	137 145 155 165 175 189 –	150 152 175 178 180 185 191
Total	325	462	445	707	678	750	685	1043	966	1211
Médias condicionais de Y , $E(Y X)$	65	77	89	101	113	125	137	149	161	173

FIGURA 2.1

Distribuição
condicional das
despesas para vários
níveis de renda (dados
da Tabela 2.1).

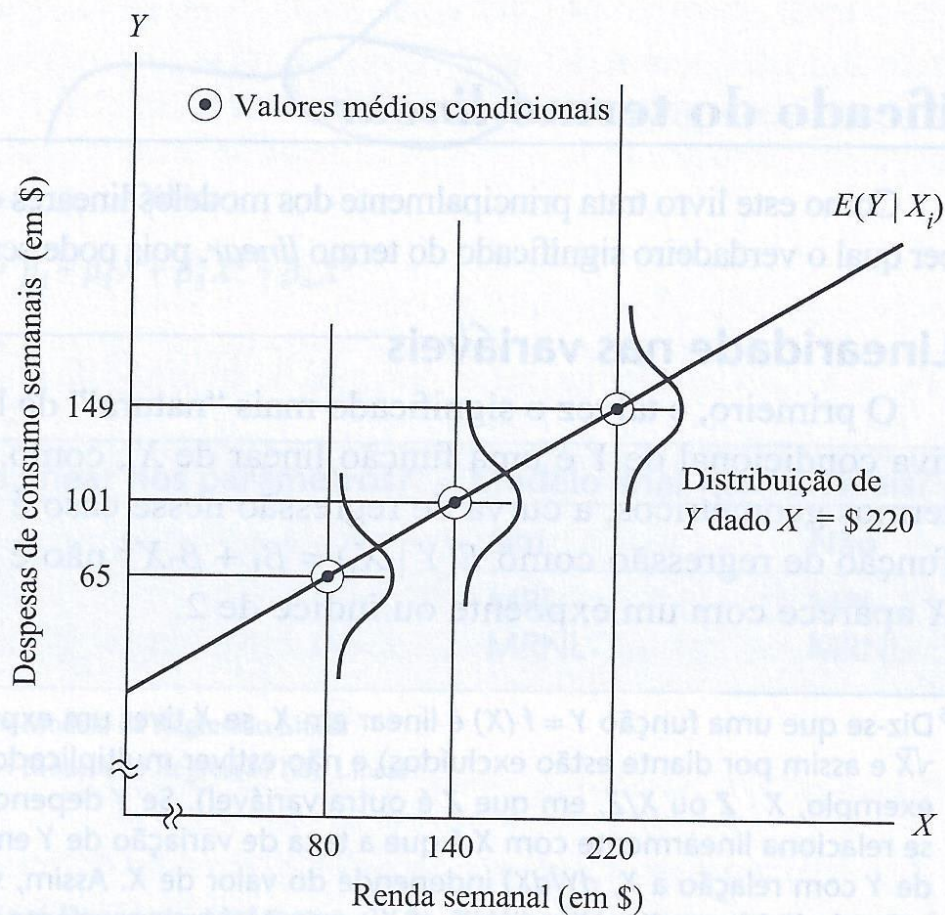


³ Conforme mostrado no **Apêndice A**, em geral, os valores das médias condicionais e incondicionais são diferentes.

⁴ Devo a James Davidson esta perspectiva. Veja DAVIDSON, James. *Econometric theory*. Oxford, R. U.: Blackwell Publishers, 2000. p. 11.

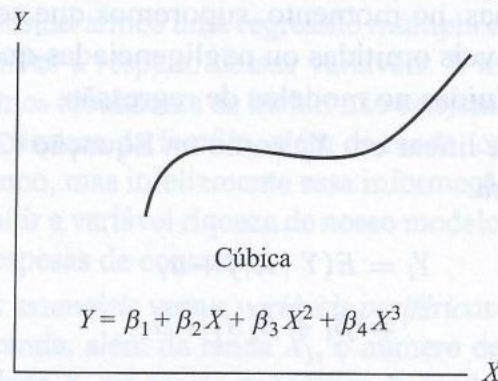
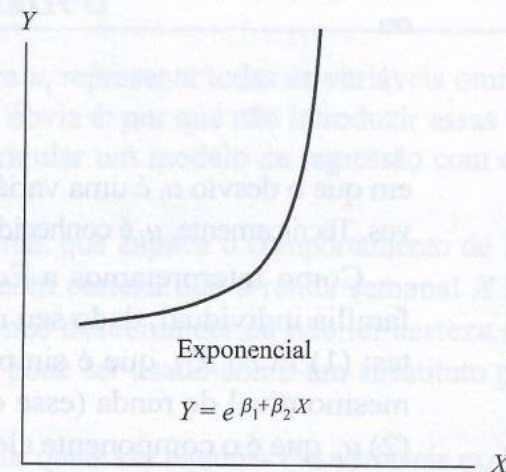
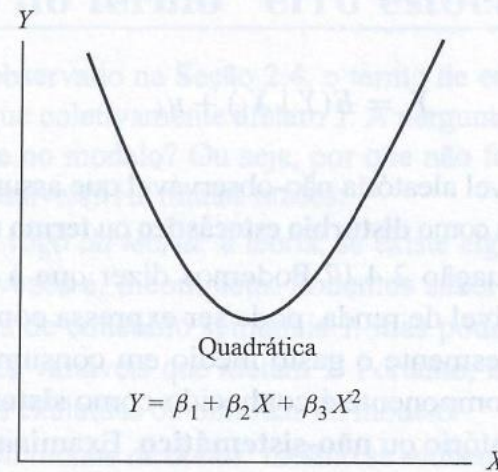
FIGURA 2.2

Linha de regressão
populacional (dados
da Tabela 2.1).



⁵ No presente exemplo, a LRP é uma linha reta, mas também poderia ser uma curva (veja a Figura 2.3).

Funções lineares nos parâmetros



Modelos de regressão linear

Modelo linear nos parâmetros?	Modelo linear nas variáveis?	
	Sim	Não
Sim	MRL	MRL
Não	MRNL	MRNL

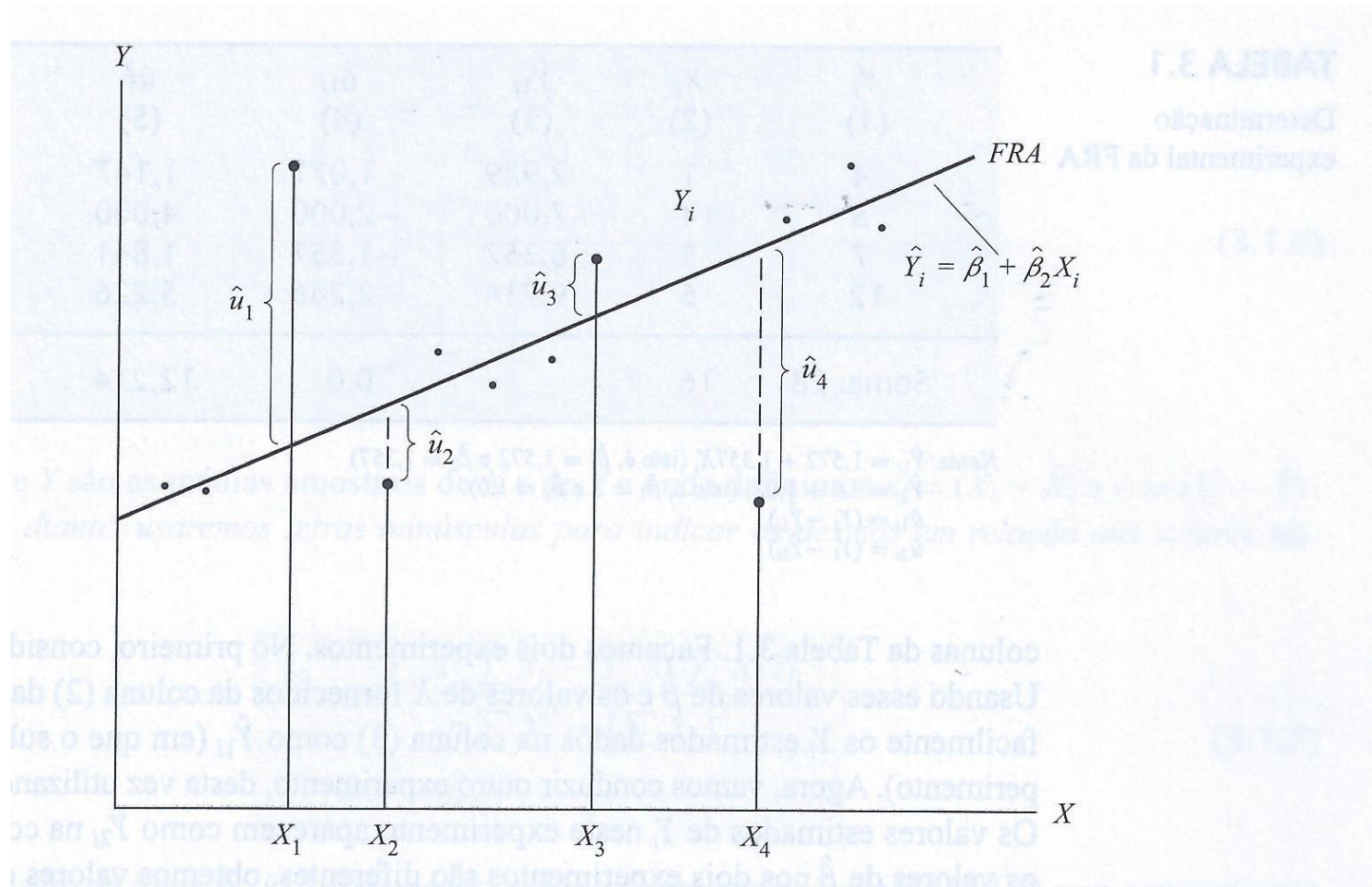
Nota: MRL = Modelo de Regressão Linear

MRNL = Modelo de Regressão Não Linear

⁷ Diz-se que uma função é linear no parâmetro β_1 , se β_1 só aparece com um expoente 1 e não está multiplicado ou dividido por nenhum outro parâmetro (por exemplo, $\beta_1\beta_2$, β_2/β_1 e assim por diante).

Método dos mínimos quadrados ordinários

Critério dos mínimos quadrados

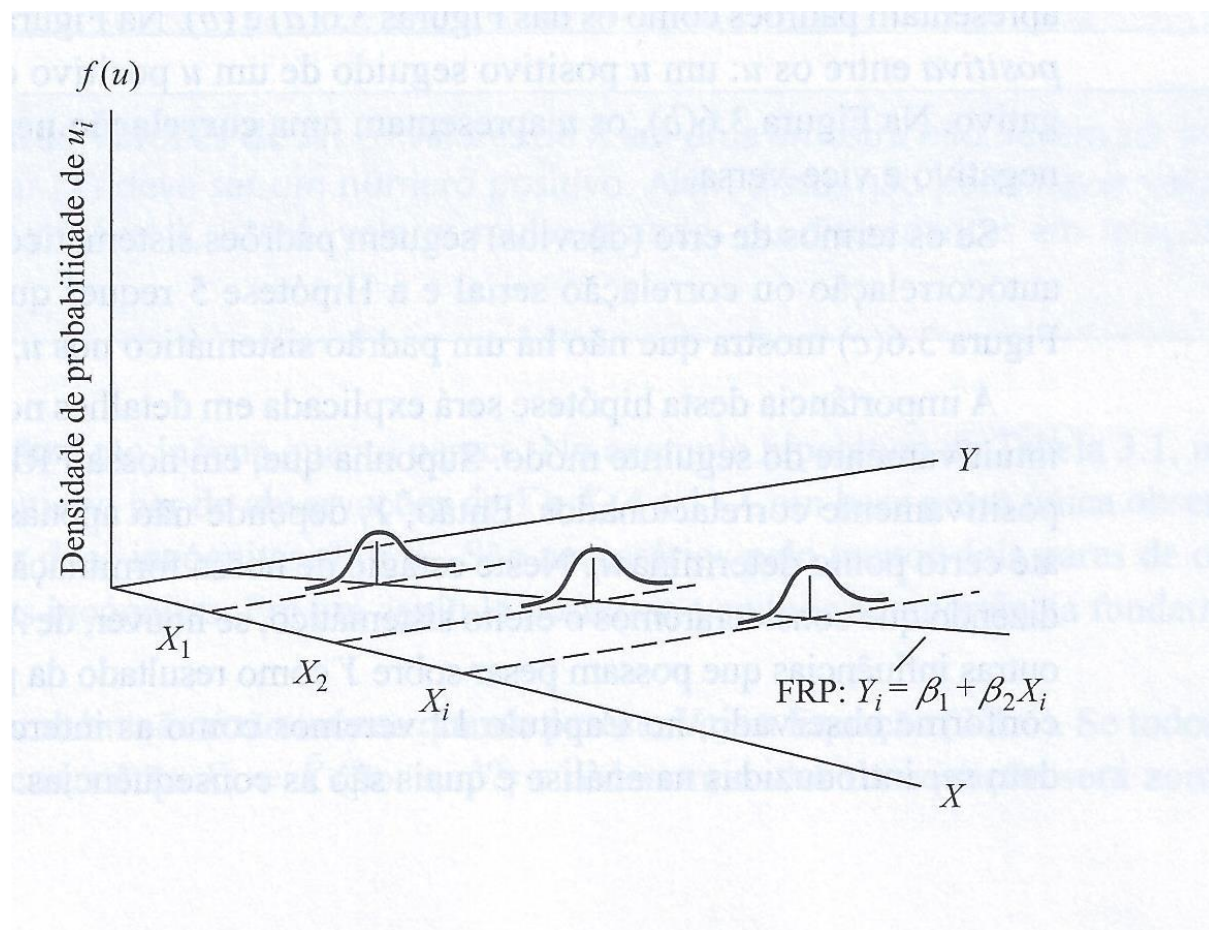


Mínimos Quadrados Ordinários (MQO)

- O mínimo valor de

$$\sum_{j=1}^n (Y_j - \hat{Y}_j)^2$$

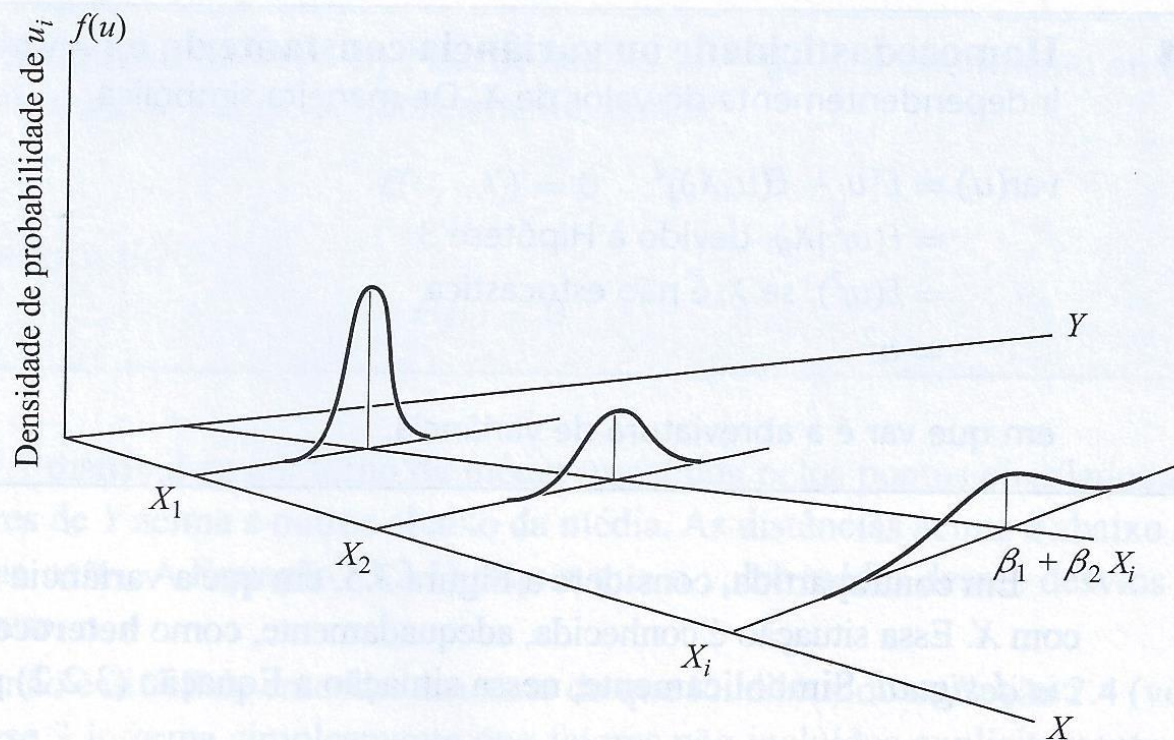
Homocedasticidade



Heterocedasticidade

FIGURA 3.5

Heterocedasticidade.



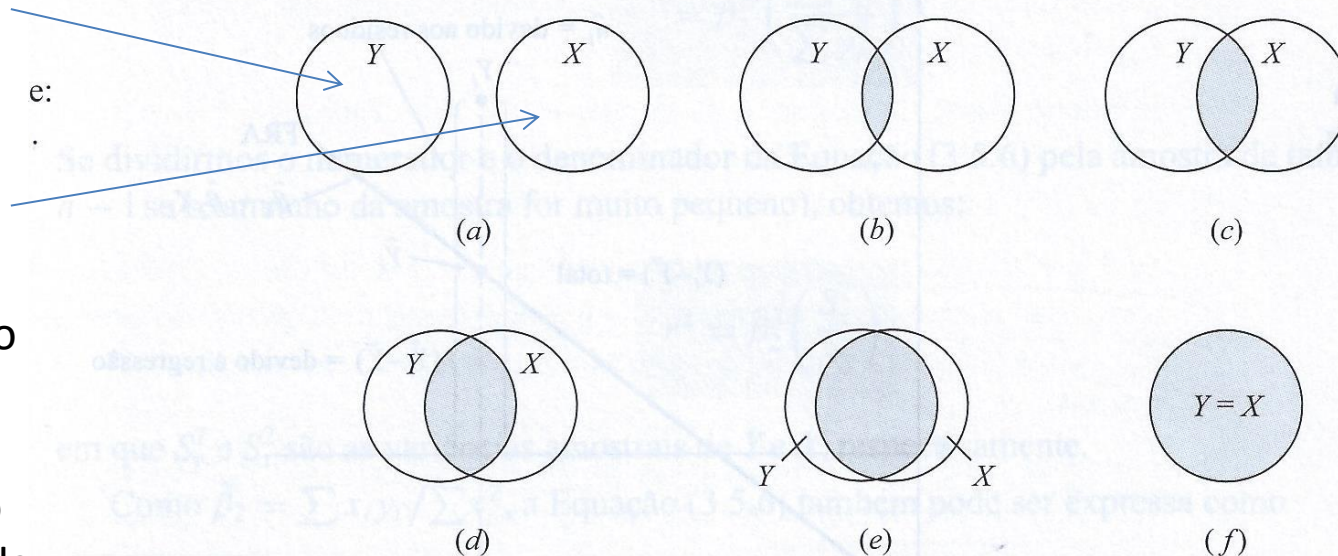
r^2 : “uma medida da qualidade do ajustamento”

$$Y_i = \hat{Y}_i + \hat{u}_i \quad (2.6.3)$$

Variação da
variável Y

e:

Variação da
variável X



A sobreposição
indica a
extensão em
que a variação
de Y é explicada
pela variação em
X

²⁰ Veja KENNEDY, Peter. “Ballentine: a graphical aid for econometrics.” *Australian Economics Papers*, 1981. v. 20, p. 414-416. O nome Ballentine faz referência aos círculos da logomarca de uma famosa cerveja com esse nome.

²¹ Os termos *variação* e *variância* indicam coisas diferentes. A *variação* é a soma dos quadrados dos desvios de uma variável de seu valor médio. A *variância* é a soma dos quadrados dividida pelos graus de liberdade adequados. Em resumo, *variância* = *variação*/gl.

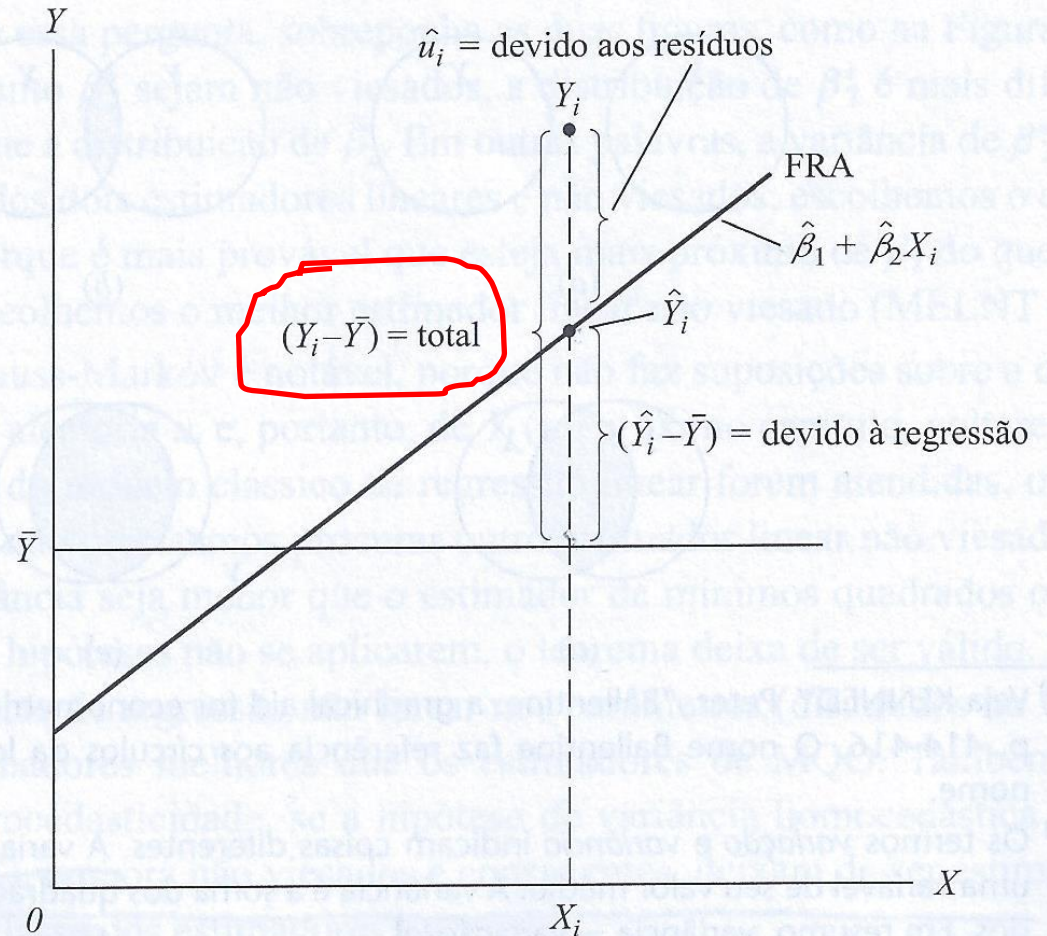
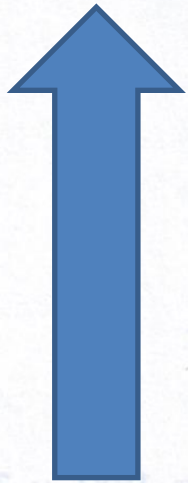
$a \neq f$?

$$1 = \frac{SQE}{STQ} + \frac{SQR}{STQ}$$

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

FIGURA 3.9

Separação da
variação de Y_i em
dois componentes.



$$r^2 = \frac{[\sum(Y_i - \bar{Y})(\hat{Y}_i - \bar{\hat{Y}})]^2}{\sum(Y_i - \bar{Y})^2 \sum(\hat{Y}_i - \bar{\hat{Y}})^2}$$

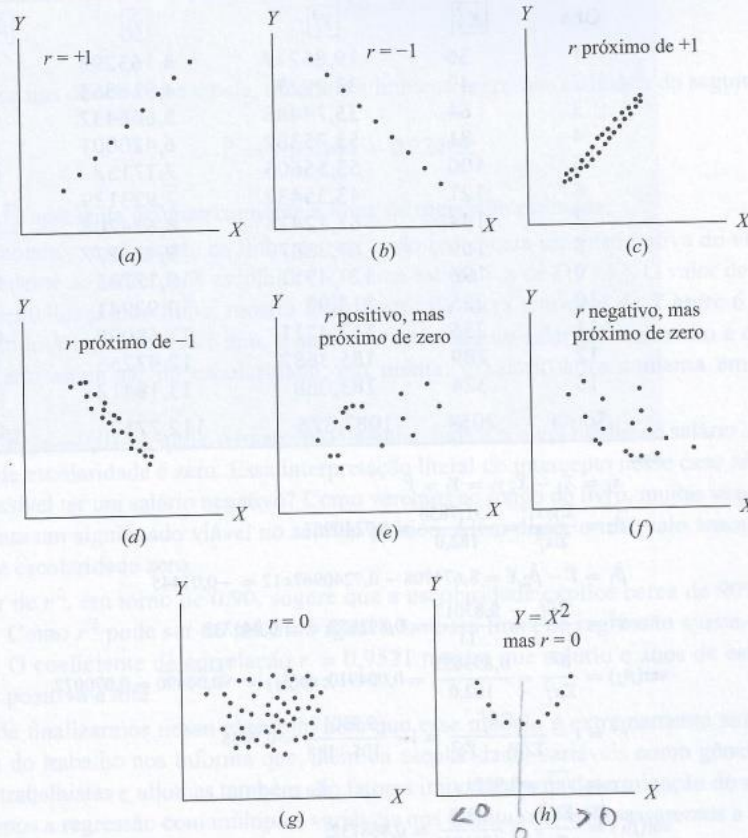
Isto é,

$$r^2 = \frac{(\sum y_i \hat{y}_i)^2}{(\sum y_i^2)(\sum \hat{y}_i^2)} \quad (3.5.14)$$

em que $Y_i = Y$ observado, $\hat{Y}_i = Y$ estimado, e $\bar{Y} = \bar{\hat{Y}}$ = média de Y . Para uma demonstração, veja o Exercício 3.15. A Equação (3.5.14) justifica a descrição de r^2 como uma medida de qualidade de ajustamento, pois nos diz quanto os valores estimados de Y estão próximos de seus valores observados.

FIGURA 3.10

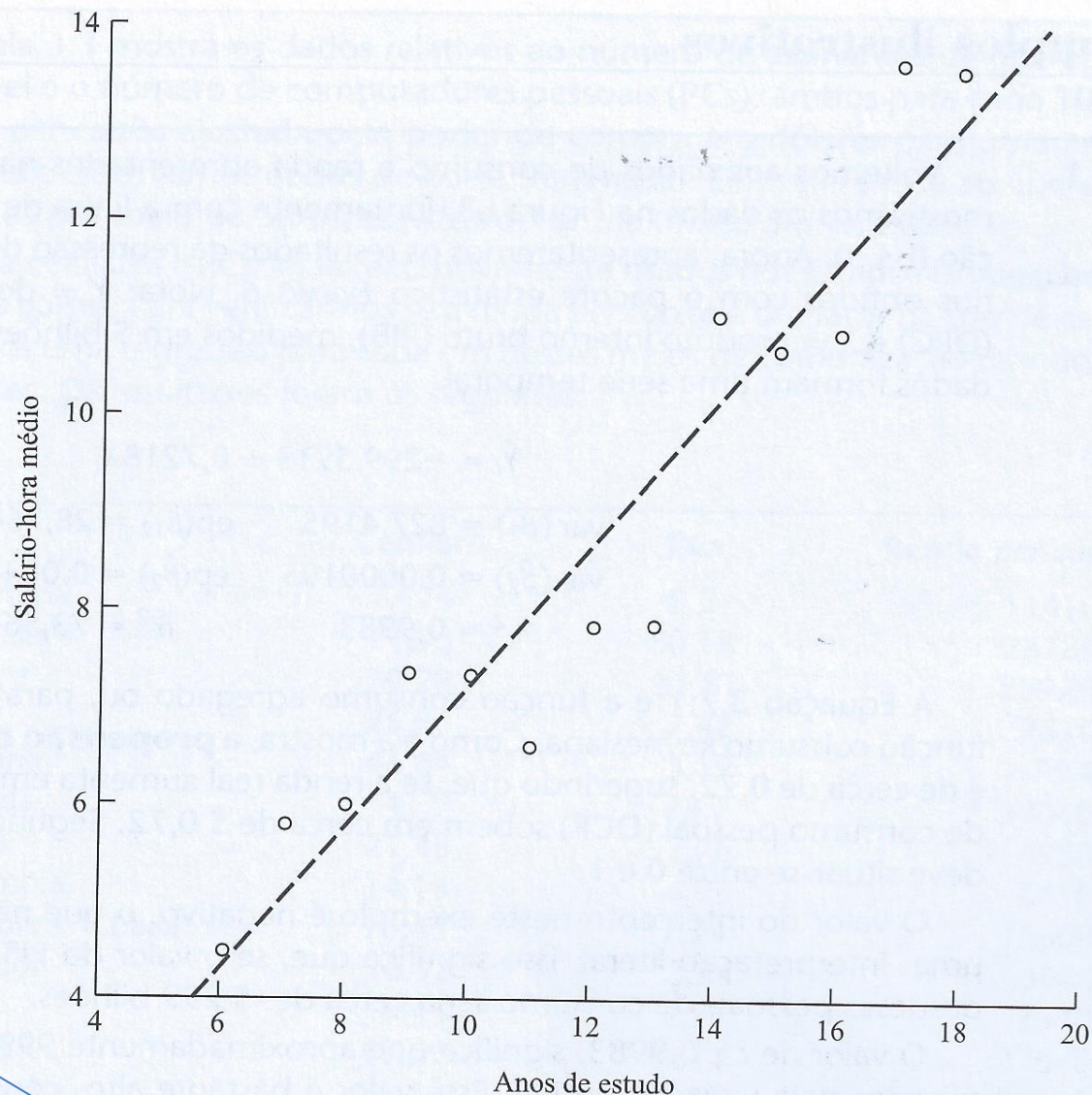
Padrões de correlação
adaptado de Theil,
*Econ. Introduction to
econometrics*,
Englewood Cliffs,
N.J.: Prentice Hall,
1971, p. 86.)



²³ No modelo da regressão, a teoria subjacente indica a direção de causalidade entre Y e X , que, no contexto dos modelos de uma única equação, é em geral de X para Y .

Linha de regressão estimada

Interprete matematicamente a equação



Com base nos dados dessa tabela, obtemos a linha de regressão estimada do seguinte modo:

$$\hat{Y}_i = -0,0144 + 0,7240X_i$$

(3.6.1)

EXAMPLE 20.1a The n X M Data Matrix for a Hypothetical
Multiple Regression or Correlation(n = 33; M = 5)

	Variable(<i>i</i>)				
J	1 °C	2 cm	3 mm	4 min	5 ml
1	6	9,9	5,7	1,6	2,12
2	1	9,3	6,4	3	3,39
3	-2	9,4	5,7	3,4	3,61
4	11	9,1	6,1	3,4	1,72
5	-1	6,9	6	3	1,8
6	2	9,3	5,7	4,4	3,21
7	5	7,9	5,9	2,2	2,59
8	1	7,4	6,2	2,2	3,25
9	1	7,3	5,5	1,9	2,86
10	3	8,8	5,2	0,2	2,32
11	11	9,8	5,7	4,2	1,57
12	9	10,5	6,1	2,4	1,5
13	5	9,1	6,4	3,4	2,69
14	-3	10,1	5,5	3	4,06
15	1	7,2	5,5	0,2	1,98
16	8	11,7	6	3,9	2,29
17	-2	8,7	5,5	2,2	3,55
18	3	7,6	6,2	4,4	3,31
19	6	8,6	5,9	0,2	1,83
20	10	10,9	5,6	2,4	1,69
21	4	7,6	5,8	2,4	2,42
22	5	7,3	5,8	4,4	2,98
23	5	9,2	5,2	1,6	1,84
24	3	7	6	1,9	2,48
25	8	7,2	5,5	1,6	2,83
26	8	7	6,4	4,1	2,41
27	6	8,8	6,2	1,9	1,78
28	6	10,1	5,4	2,2	2,22
29	3	12,1	5,4	4,1	2,72
30	5	7,7	6,2	1,6	2,36
31	1	7,8	6,8	2,4	2,81
32	8	11,5	6,2	1,9	1,64
33	10	10,4	6,4	2,2	1,82

EXAMPLE 20.1c A Computer Fit of a Multiple-Regression Equation to the Data of Example 20.1a, Where Variable 5 Is the Dependent Variable

Regression model:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

For each i (where $i = 1, 2, 3, 4$),

$$H_0: \beta_i = 0$$

$$H_A: \beta_i \neq 0$$

Variable i	b_i	s_{b_i}	t	ν	b'_i
X_1	-0.12932	0.021287	-6.075	28	-0.73176
X_2	-0.018785	0.056278	-0.334	28	-0.41108
X_3	-0.046215	0.20727	-0.223	28	-0.26664
X_4	0.20876	0.067034	3.114	28	0.36451

Y intercept: $a = 2.9583$

EXAMPLE 20.1c A Computer Fit of a Multiple-Regression Equation to the Data of Example 20.1a, Where Variable 5 Is the Dependent Variable

Regression model:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

For each i (where $i = 1, 2, 3, 4$),

$$H_0: \beta_i = 0$$

$$H_A: \beta_i \neq 0$$

Variable					
i	b_i	s_{b_i}	t	ν	b'_i
X_1	-0.12932	0.021287	-6.075	28	-0.73176
X_2	-0.018785	0.056278	-0.334	28	-0.41108
X_3	-0.046215	0.20727	-0.223	28	-0.26664
X_4	0.20876	0.067034	3.114	28	0.36451

Y intercept: $a = 2.9583$

- Notar que um aumento de 1°C na variável 1 está associado a uma diminuição de 0,129 ml na variável 5 (Y), se não houver mudança nas demais variáveis.

Análise de Variância da regressão e correlação múltipla

- hipótese nula:
 - todos os coeficientes de regressão parcial são iguais a zero
- hipótese alternativa:
 - pelo menos um deles é diferente de zero.

- A soma de quadrados total é uma medida da variabilidade total entre os valores Y

$$\sum(Y_j - \bar{Y})^2;$$

- A soma de quadrados da regressão expressa a variabilidade entre os valores Y que é atribuída à regressão

$$\sum(\hat{Y}_j - \bar{Y}_j)^2.$$

- A soma de quadrados residual é a variabilidade total entre o Y's que permanece após ajustar a regressão

$$\sum(Y_j - \hat{Y}_j)^2$$

TABLE 20.1: Definitions of the Appropriate Sums of Squares, Degrees of Freedom, and Mean Squares Used in Multiple Regression or Multiple Correlation Analysis of Variance

Source of variation	Sum of squares (SS)	DF*	Mean square (MS)
Total	$\sum (Y_j - \bar{Y})^2$	$n - 1$	
Regression	$\sum (\hat{Y}_j - \bar{Y}_j)^2$	m	$\frac{\text{regression SS}}{\text{regression DF}}$
Residual	$\sum (Y_j - \hat{Y}_j)^2$	$n - m - 1$	$\frac{\text{residual SS}}{\text{residual DF}}$

* n = total number of data points (i.e., total number of Y values); m = number of independent variables in the regression model.

EXAMPLE 20.1d A Computer Analysis of Variance for the Multiple Regression Data of Example 20.1a

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$H_A: \beta_1 \text{ and/or } \beta_2 \text{ and/or } \beta_3 \text{ and/or } \beta_4 \neq 0$$

<i>Source of Variation</i>	SS	DF	MS
Total	14.747	32	
Multiple regression	9.7174	4	2.4294
Residual	5.0299	28	0.17964

$F = 13.5$, with DF of 4 and 28

$F_{0.05(1),4,28} = 2.71$, so reject H_0 .

$P \ll 0.0005$ [$P = 2.99 \times 10^{-6}$]

Coefficient of determination: $R^2 = 0.65893$

Adjusted coefficient of determination: $R_a^2 = 0.61021$

Multiple correlation coefficient: $R = 0.81175$

Standard error of estimate: $s_{Y \cdot 1,2,3,4} = 0.42384$

$$F = \frac{\textit{regression MS}}{\textit{residual MS}}$$

- Assumindo que Y é dependente de cada um dos X 's, significa que trata-se de **regressão múltipla**.
- Se nenhuma dependência está implícita, qualquer uma das variáveis ($M = m + 1$) poderia ser designada como Y (para propósitos computacionais), e, nesse caso, trata-se de **correlação múltipla**.
- Nas duas situações pode-se testar a hipótese de que não há inter-relacionamento entre as variáveis – teste F .

Multiple Regression and Correlation

The numerator and denominator degrees of freedom for this variance ratio are the regression DF and the residual DF, respectively. For multiple regression, this F tests

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_M = 0,$$

which may be written as

$$H_0: \beta_i = 0 \text{ for all } i\text{'s},$$

against

$$H_A: \beta_i \neq 0 \text{ for one or more } i\text{'s}.$$

The ratio,

$$R^2 = \frac{\text{regression SS}}{\text{total SS}} \quad (20.19)$$

or, equivalently,

$$R^2 = 1 - \frac{\text{residual SS}}{\text{total SS}} \quad (20.20)$$

R^2

- A taxa R^2 é o coeficiente de determinação para a regressão ou correlação múltipla:
 - Coeficiente de determinação múltipla – mede a força do relacionamento linear
- **Regressão**
 - Expressa a proporção da variabilidade total em Y que é atribuída à dependência de Y sobre todos os X's.
- **Correlação**
 - Expressa a variabilidade total em qualquer uma das M variáveis que é representada pela correlação com todas as outras $M - 1$ variáveis.

Suposições sobre a Análise de Regressão Múltipla

1. Os valores de Y vêm de uma amostra aleatória da população e são independentes entre si.
2. Para qualquer combinação de valores de X_i 's na população, existe uma distribuição normal de Y . Portanto, uma distribuição normal de ε 's.
3. Homogeneidade de variâncias: as variâncias dos valores Y para todas as combinações de X 's são iguais. O Quadrado Médio do Resíduo é a estimativa desta variância comum.

Hipóteses sobre os coeficientes de regressão parcial

- Se $H_0: \beta_1 = \beta_2 = \dots = \beta_m = 0$ é rejeitada
 - Deve-se concluir que pelo menos um β_i é diferente de zero e o X_i associado contribui para explicar Y .
 - Neste caso, cada um dos coeficientes de regressão parcial podem ser submetidos a um novo teste:

$$t = \frac{b_i}{s_{b_i}}$$

Testando: $H_0: \beta_i = 0$

Multicolinearidade

- Significa a correlação entre variáveis independentes.
- Se existe correlação somente entre duas das variáveis independentes, denomina-se **colinearidade**.
- Se as variáveis independentes X_1 e X_2 são altamente correlacionadas entre si, os coeficientes de regressão parcial associados a elas (b_1 e b_2) podem não refletir a dependência de Y sobre X_1 ou Y sobre X_2 que existe na população.
- As interpretações dos efeitos dos X_i 's sobre Y podem ser irreais ou ambíguos.
- Os erros-padrão dos coeficientes de regressão parcial dos X_i 's correlacionados serão grandes (assim como o IC será amplo), significando que os coeficientes são imprecisos nas estimativas dos relacionamentos na população;

Como suspeitar de multicolinearidade

- Se existem t 's significativos mesmo sem o F ser significativo;
- Se algum coeficiente de correlação entre os pares de X_i 's são muito altos ($>0,80$ ou $>0,90$)
- Se R^2 é muito maior do que $\sum_{i=1}^m r^2_{Y_i}$, onde $r^2_{Y_i}$ representa a correlação simples entre a variável dependente Y e cada variável independente X_i .

Coeficiente de correlação parcial padronizado – b'_i

- A padronização ocorre pelo desvio-padrão de Y (S_Y) e X_i (S_{X_i}).

$$b'_i = b_i \left(\frac{S_{X_i}}{S_Y} \right)$$

- Este coeficiente é adimensional e, portanto, especialmente útil quando os X_i 's estão em diferentes unidades de medida.

Seleção de variáveis independentes

- Embora seja possível obter os coeficientes de regressão parcial para todas as variáveis estudadas, o desafio é concluir sobre quais delas têm efeito significativo sobre Y .
- Isso é feito para simplificar a interpretação da equação de regressão resultante;
- Um menor número de variáveis pode aumentar a precisão das estimativas de Y pelo modelo.
- Porém, se a meta é descrever e entender relacionamentos biológicos, alguma cautela deve ser tomada com relação às regras para escolha das variáveis significativas.

Procedimentos para encontrar o melhor ajuste de regressão múltipla

- *Fitting All Possible Equations*
- *Backward Elimination of Variables*
- *Forward Addition of Variables*
- *Stepwise Regression*

Fitting All Possible Equations

- Considerando o exemplo com 4 variáveis independentes e uma variável dependente:
- Procedimento:
 1. Ajustar uma equação usando todos os quatro X_i 's;
 2. Ajustar as quatro possíveis equações contendo três das quatro variáveis independentes;
 3. Ajustar uma equação com cada uma das seis possíveis combinações de duas variáveis independentes;
 4. Ajustar uma regressão simples usando cada uma das quatro variáveis independentes.

Fitting All Possible Equations

- Após ajustar as 15 equações possíveis, pode-se decidir pela melhor em relação ao menor quadrado médio residual ou maior R^2 ajustado.
- R^2 ajustado é útil para comprar equações que apresentam diferentes número de variáveis independentes:

$$R^2_a = 1 - \frac{\text{residual } MS}{\text{total } MS}$$

Dificuldades do Método

- O número de equações possíveis é calculado da seguinte forma $2^m - 1$.
- Como são muitas as equações fica difícil avaliar todas. Ex. com $m = 10$, haverá $2^{10} - 1 = 1.023$ equações.
- Como determinar se uma equação é significativamente melhor do que a segunda melhor?

Backward Elimination of Variables

- Este procedimento envolve eliminar as variáveis não significativas da regressão.
- A hipótese $H_0: \beta_i = 0$ pode ser examinada para cada um dos m coeficientes de regressão parcial. Se todos os m testados são rejeitados, pode-se concluir que todos os X_i 's tem efeito significativo sobre Y e que nenhum deles deve ser retirado da análise.
- Porém, se qualquer valor $|t|$ for menor do que o valor crítico ($t_{\alpha(2),v}$), onde v é GL resíduo ($n - m - 1$), a variável independente associada ao menor t -absoluto é retirada do modelo e uma nova equação é ajustada com $m-1$ variáveis independentes.....

Forward Addition of Variables

- Este procedimento inicia-se com o menor modelo de regressão possível, com somente uma variável independente
- Escolhe-se o melhor modelo de regressão simples avaliando o maior valor do $|t|$ para os coeficientes de regressão.
- Se nenhum dos coeficientes for significativo, conclui-se que não há relacionamento entre Y e as variáveis independentes. O processo para aí.
- Se pelo menos um coeficiente é significativo, o ajuste será efetuado para cada regressão que já contém o X selecionado e um dos outros X_i 's, escolhendo-se a equação com o maior $|t|$ associado com o outros X 'is.

Problema

- Como o relacionamento entre as variáveis muda na medida em que cada variável é adicionada, não há garantia de que a importância de cada variável está relacionada à sequência a qual ela foi adicionada ao modelo de regressão.

EXAMPLE 20.1e Backward Elimination of Variables in Multiple-Regression Analysis, Using the Data from Example 20.1a

As shown in Example 20.1c, the multiple regression analysis for the model $\hat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$ yields the following statistics:

Variable	b_i	s_{b_i}	t	ν
X_1	-0.12932	0.021287	-6.075	28
X_2	-0.018785	0.056278	-0.334	28
X_3	-0.046215	0.20727	-0.223	28
X_4	0.20876	0.067034	3.114	28
$a = 2.9583$				

The critical value for testing $H_0: \beta_j = 0$ against $H_A: \beta_j \neq 0$ is $t_{0.05(2),28} = 2.048$. Therefore, H_0 would be rejected for β_1 and β_4 , but not for β_2 or β_3 . Of the t tests for the latter two, the t for testing the significance of β_3 has the smaller absolute value. Therefore, $\beta_3 X_3$ is deleted from the model, leaving $\hat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4$. The data are then subjected to a multiple-regression analysis using this model with three independent variables, and the following statistics are obtained:

Variable	b_i	s_{b_i}	t	ν
X_1	-0.13047	0.020312	-6.423	29
X_2	-0.015424	0.053325	-0.289	29
X_4	0.20450	0.063203	3.236	29
$a = 2.6725$				

The critical value for testing the significance of these partial regression coefficients is $t_{0.05(2),29} = 2.045$. Therefore, $H_0: \beta_j = 0$ would be rejected for β_1 and for β_4 , but not for β_2 . Therefore, $\beta_2 X_2$ is deleted from the regression model, leaving $\hat{Y} = \alpha + \beta_1 X_1 + \beta_4 X_4$. The analysis of the data using this model, with two independent variables, yields the following statistics:

Variable	b_i	s_{b_i}	t	ν
X_1	-0.13238	0.018913	-6.999	30
X_4	0.20134	0.061291	3.285	30
$a = 2.5520$				

The critical value for testing $H_0: \beta_j = 0$ against $H_A: \beta_j \neq 0$ is $t_{0.05(2),30} = 2.042$. Therefore, both β_1 and β_4 are concluded to be different from zero, and $\hat{Y} = 2.552 - 0.132X_1 + 0.201X_4$ is the final model.

Stepwise Regression

- Envolve a adição e a eliminação de variáveis independentes
- O procedimento inicia-se como no Forward, mas sempre que um X é adicionado, o β associado a cada X_i no modelo é examinado checando a significância do t . Se algum t é não significativo, a variável com o menor $|t|$ é eliminada.

Correlação parcial

- Quando o interesse está no relacionamento entre todas as M variáveis, onde nenhuma delas é considerada dependente das outras, usa-se o coeficiente de correlações parciais múltiplas, R .
- Considerando três variáveis, o R (coeficiente de correlação parcial de primeira ordem), será:

$$r_{ik.l}$$

e refere-se à correlação entre as variáveis i e k , considerando que a variável l não muda de valor.

Correlação parcial

- Para quatro variáveis, o coeficiente de correlação parcial (segunda ordem) será:

$$r_{ik.lp}$$

- E expressa a correlação entre i e k mantendo-se l e p constantes.
- Em geral, a correlação parcial pode ser expressa como

$$r_{ik...}$$

EXAMPLE 20.2 A Matrix of Partial Correlation Coefficients, as It Might Appear as Computer Output (from the Data of Example 20.1a)

	1	2	3	4	5
1	1.00000	0.19426	0.12716	0.33929	−0.75406
2	0.19426	1.00000	−0.26977	0.23500	−0.06296
3	0.12716	−0.26977	1.00000	0.26630	−0.04210
4	0.33929	0.23500	0.26630	1.00000	0.50720
5	−0.75406	−0.06296	−0.04210	0.50720	1.00000

Variável “Dummy”

- É utilizada para incluir variáveis com escala nominal nos modelos de regressão linear múltiplos.

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j}$$

- X3 pode ser uma variável Dummy: sexo (M ou F).
- Y a pressão do sangue, X1 a idade, X2 o peso.
- Neste caso, X3 = 1 para M; X3 = 0 para F.
- Testa-se a significância de X3 em determinar a pressão do sangue.

Variável “Dummy”

- Em geral, quando L níveis de uma variável a ser representada por uma variável dummy, $L - 1$ variáveis dummy são requeridas para representá-la.
- Ex. se existem três subespécies sendo testadas em um modelo ($L = 3$ subespécies), então 2 variáveis dummy devem ser utilizadas. Cada uma será sempre 0 ou 1 para cada Y (0 e 0 para subespécie 1, 0 e 1 para subsp. 2, 1 e 0 para subsp. 3).
- Para qualquer Y , a soma dos 0 e 1 não pode exceder 1. Portanto, a combinação 1 e 1 não é aceita.

Estudo Dirigido IV

1. O volume de um cubo de 1m de aresta é 1m^3 . Qual a aresta de um cubo cujo volume é 2m^3 ? Teste a hipótese de que o valor da aresta é um número racional.
2. Defina erro padrão e número de graus de liberdade.