Probabilidades e Estatística

Regressão Linear Simples

Prof. Caldeira Duarte e Prof.ª Anabela Pereira

Actualizado pelo Prof. António Sardinha (Fevereiro de 2013)

Departamento de Matemática

IPS Instituto Politécnico de Setúbal
Escola Superior de Tecnologia de Setúbal

7 REGRESSÃO LINEAR SIMPLES

7.1 Dados Bivariados

Por vezes certos fenómenos em estudo não se descrevem apenas através de uma variável, sendo necessária a observação de duas (ou mais) variáveis para termos uma visão global do problema. Quando tal ocorre, cada unidade estatística pode contribuir com um conjunto de dois valores passando a trabalhar-se com dados bivariados. Exemplos de dados bivariados são: a altura e peso da população portuguesa, o rendimento mensal de um agregado familiar e o respectivo montante de despesas mensais, as horas de estudo de um aluno e notas obtidas nas disciplinas, etc.

7.2 Representação de Dados Bivariados

A informação da população que se pretende estudar aparece sob a forma de pares de valores da amostra, isto é, cada unidade estatística contribui com um conjunto de dois valores. Surge então o problema de como estudar a existência ou não de relações entre essas variáveis observadas.

Como ponto de partida para o estudo da existência (ou não) de relação estatística (correlação) entre duas variáveis ou características de uma amostra podemos representá-las graficamente através de um **Diagrama de Dispersão** ou **Nuvem de Pontos.** Esta representação gráfica para os dados bivariados consiste em marcarmos os valores das observações realizadas, x_i e y_i, num sistema de eixos cartesianos e obtermos os pontos correspondentes aos pares ordenados (x_i, y_i).

Exemplo 7.1. Considerando as idades de 16 conjuges na data dos seus casamentos representadas na tabela seguinte (em que X representa a idade do marido e Y a idade da mulher):

X	18	20	21	21	22	23	23	23	24	25	25	26	26	26	28	28
Y	17	20	20	22	22	21	22	23	23	24	25	23	24	27	26	27

Estes dados podem representar-se no Diagrama de Dispersão ou Nuvem de Pontos da Figura 7.1. Este diagrama, de forma intuitiva, sugere-nos a existência de uma relação linear entre as duas variáveis em estudo, isto é, uma relação que se pode traduzir geometricamente através de uma recta. ■

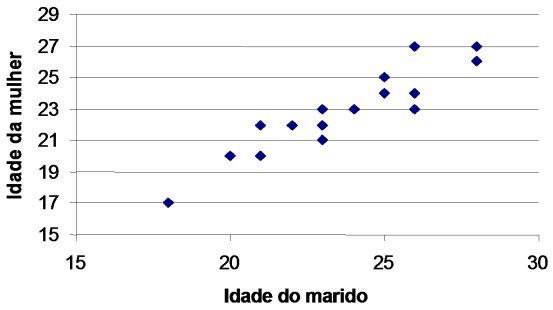

Figura 7.1: Diagrama de Dispersão ou Nuvem de Pontos.

Através da simples observação do diagrama de dispersão ou nuvem de pontos podemos concluir acerca da existência ou não de correlação linear entre duas variáveis X e Y.

Exemplo 7.2. Os gráficos das Figuras 7.2 e 7.3 ilustram vários tipos de correlações lineares entre duas variáveis.

Embora o Diagrama de Dispersão seja um método simples de detecção de relação linear é, no entanto, insuficiente para quantificar a correlação, assim como, quando há observações que se repetem, o diagrama não realça a sua frequência.

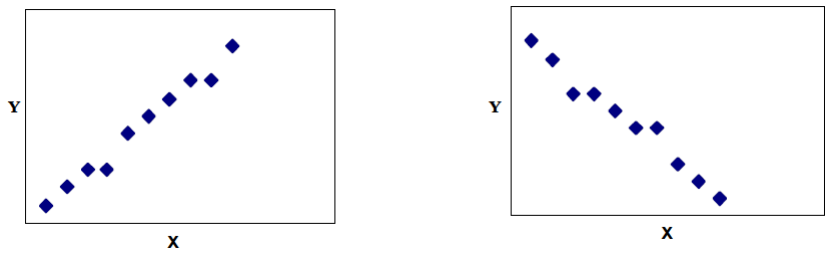

Figura 7.2: Correlação Linear Positiva (forte) à esquerda e Negativa (forte) à direita.

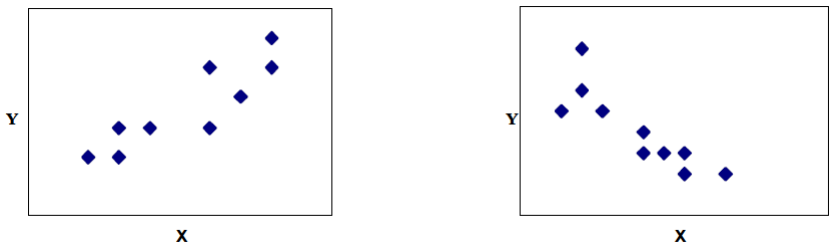

Figura 7.3: Correlação Linear Positiva (fraca) à esquerda e Negativa (fraca) à direita.

■

7.3 Coeficiente de Correlação Linear Empírico

O **Coeficiente de Correlação Linear Empírico** (ou Amostral), r_{XY}, mede o grau de associação linear entre dados bivaridos, sendo calculado através da expressão:

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

em que

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\sum_{i=1}^{n} x_i y_i - \frac{n}{n-1}\bar{x}\bar{y}$$

se denomina de **covariância amostral**, sendo uma medida de variabilidade conjunta entre as variáveis X e Y; S_X e S_Y são os desvios padrões amostrais de X e Y respectivamente.

Observação 7.3. A covariância amostral e o coeficiente de correlação linear empírico são estatísticas respectivamente da covariância e do coeficiente de correlação linear da população.

Deste modo podemos reescrever o coeficiente de correlação linear empírico como:

$$r_{XY} = \frac{\text{covariância}_{XY}}{\sqrt{\text{variância}_X \times \text{variância}_Y}}$$

O coeficiente de correlação linear empírico é um número do intervalo $[-1, 1]$. O sinal do mesmo indica se uma variável aumenta à medida que a outra também aumenta ($r_{XY} > 0$) ou se uma variável aumenta à medida que a outra diminui ($r_{XY} < 0$). A magnitude indica a proximidade dos pontos em relação a uma linha recta, isto é, quanto mais próximo r_{XY} estiver dos extremos do intervalo $[-1, 1]$, maior é o grau de associação linear; em particular se $r_{XY} = \pm 1$ existe uma correlação linear perfeita estando todos os pontos situados na recta; se $r_{XY} = 0$ a correlação linear é nula (embora possa existir uma relação não linear entre X e Y).

O valor de r_{XY} só é válido dentro da amplitude de valores x e y da amostra. Não se pode inferir que este coeficiente terá o mesmo valor quando se consideram valores de x e y mais extremos do que os constantes na amostra.

É possível trocar a variável dependente e independente sem alterar o valor de r_{XY}.

A existência de um "bom"[1] coeficiente de correlação linear empírico entre X e Y, por si só, não implica necessariamente uma relação de "causa e efeito". Como tal, este coeficiente deve ser sempre acompanhado pelo diagrama de dispersão. Na Figura 7.4 temos exemplos de situações em que r_{XY} tem um valor próximo dos extremos do intervalo $[-1, 1]$ e, no entanto, não são adequados os modelos lineares; conclui-se deste modo que o simples cálculo r_{XY} é, por vezes, insuficiente.

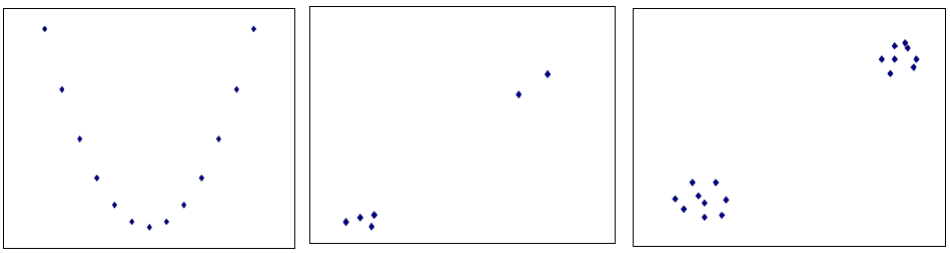

Figura 7.4: À esquerda temos uma relação quadrática; ao centro temos observações isoladas; à direita temos dados que compreendem subgrupos.

O exemplo seguinte ilustra uma situação deste género, com um caso concreto.

Exemplo 7.4. Considere o conjunto de observações da tabela

X	1	1.5	1.6	8	8.25	1.9	9.1	8.9	2	8.75	1	8.1	8.5	1.5
Y	3	3.75	3	10.5	11.5	2.6	11	11.5	3.1	10	2.5	10	10.75	2.35

Vamos verificar que o simples cálculo do coeficiente de correlação linear empírico é insufici-

[1] Vamos considerar como "bom" um coeficiente de correlação linear empírico que se situe no intervalo $[-1, -0.8] \cup [0.8, 1]$. Este intervalo, no entanto, depende dos objectivos e dos dados da pesquisa; como tal, deve ser entendido como um intervalo indicativo e não fixo.

ente para concluir se existe associação linear entre X e Y.

$$r_{XY} = \frac{\frac{1}{131}\sum_{i=1}^{13}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{13}\sum_{i=1}^{13}(x_i - \bar{x})^2 \times \frac{1}{13}\sum_{i=1}^{13}(y_i - \bar{y})^2}} = 0.989.$$

Pela simples leitura de r_{XY} seríamos levados a concluir que existiria uma boa associação linear entre X e Y. No entanto tal é falso como podemos verificar pelo diagrama de dispersão da Figura 7.5, onde é nitida a existência de dois subgrupos nas observações em análise. ■

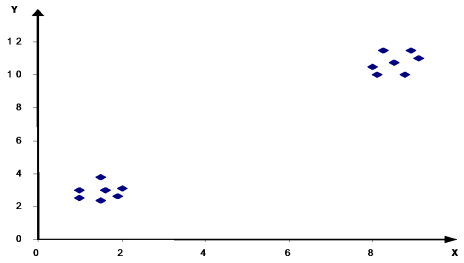

Figura 7.5: Observações com subgrupos.

7.4 Recta de Regressão

Tem-se por objectivo a construção de um modelo matemático que expresse a relação de tipo linear existente entre duas variáveis, com base nos correspondentes valores amostrais.

Considera-se, em geral, X a **variável independente** (explicativa ou explanatória) e Y a **variável dependente** (explicada ou resposta). O modelo matemático que relaciona as duas variáveis permite efectuar previsões para Y.

A recta de regressão pode calcular-se quando no

. Diagrama de Dispersão se averiguar a existência de uma relação linear entre as variáveis e no

. Coeficiente de Correlação Linear Empírico se obtiver um valor considerado "bom".

Quando se verifica uma forte correlação linear entre as variáveis sob observação podemos descrever a relação entre X e Y, traçando na nuvem de pontos uma recta que seja (segundo algum critério) a que melhor se ajusta aos dados.

Um dos métodos mais conhecidos de ajustar uma recta a um conjunto de dados, é o Método dos Mínimos Quadrados (MMQ), que consiste em determinar a recta que minimiza a soma dos quadrados das distâncias verticais entre os valores observados e a recta (denominadas por erros ou resíduos)

$$e_i^2 = (y_i - \hat{y}_i)^2$$

tal como é ilustrado na Figura 7.6.

O modelo matemático que expressa a relação linear de Y sobre X é a recta de regressão

$$\hat{y} = a + bx$$

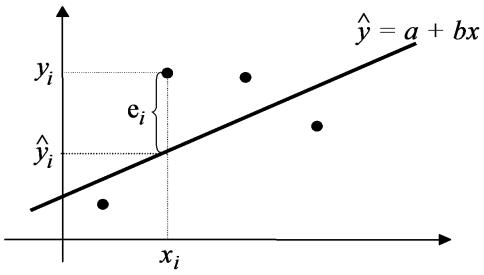

Figura 7.6: Ajustamento da recta de regressão.

obtida de tal modo que os desvios ou resíduos quadráticos das observações em relação à recta sejam mínimos,

$$\min \sum_{i=1}^{n} e_i^2 = \min \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \min \sum_{i=1}^{n} \left[y_i - (a + bx_i)\right]^2 .$$

Como tal, é necessário calcular os pontos de estacionariedade através das primeiras derivadas:

$$\left\{ \begin{array}{l} \frac{\partial}{\partial a} \sum\limits_{i=1}^{n} \left(y_i - a - bx_i\right)^2 = 0 \\ \frac{\partial}{\partial b} \sum\limits_{i=1}^{n} \left(y_i - a - bx_i\right)^2 = 0 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} -2 \sum\limits_{i=1}^{n} \left(y_i - a - bx_i\right) = 0 \\ -2 \sum\limits_{i=1}^{n} \left(y_i - a - bx_i\right) x_i = 0 \end{array} \right. \Leftrightarrow$$

$$\Leftrightarrow \left\{ \begin{array}{l} \sum\limits_{i=1}^{n} y_i - na - b \sum\limits_{i=1}^{n} y_i = 0 \\ \sum\limits_{i=1}^{n} x_i y_i - a \sum\limits_{i=1}^{n} x_i - b \sum\limits_{i=1}^{n} x_i^2 = 0 \end{array} \right. \Leftrightarrow$$

$$\Leftrightarrow \left\{ \begin{array}{l} a = \frac{\sum\limits_{i=1}^{n} y_i}{n} - b \frac{\sum\limits_{i=1}^{n} x_i}{n} \\ \sum\limits_{i=1}^{n} x_i y_i - \left(\frac{\sum\limits_{i=1}^{n} y_i}{n} - \frac{b \sum\limits_{i=1}^{n} x_i}{n} \right) \sum\limits_{i=1}^{n} x_i - b \sum\limits_{i=1}^{n} x_i^2 = 0 \end{array} \right. \Leftrightarrow$$

$$\Leftrightarrow \left\{ \begin{array}{l} a = \bar{y} - b\bar{x} \\ n \sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} y_i \sum\limits_{i=1}^{n} x_i + b \left(\sum\limits_{i=1}^{n} x_i \right)^2 - nb \sum\limits_{i=1}^{n} x_i^2 = 0 \end{array} \right. \Leftrightarrow$$

$$\Leftrightarrow \left\{ \begin{array}{l} a = \bar{y} - b\bar{x} \\ b \left(n \sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i \right)^2 \right) = n \sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} y_i \sum\limits_{i=1}^{n} x_i \end{array} \right. \Leftrightarrow$$

$$\Leftrightarrow \left\{ \begin{array}{l} a = \bar{y} - b\bar{x} \\ b = \frac{n \sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{n \sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i \right)^2} \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} a = \bar{y} - b\bar{x} \\ b = \frac{\sum\limits_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} a = \bar{y} - b\bar{x} \\ b = \frac{S_{XY}}{S_X^2} \end{array} \right.$$

Com base nas segundas derivadas obtem-se a matriz hessiana,

$$H = \begin{bmatrix} \frac{\partial^2}{\partial a^2} \sum_{i=1}^{n} (y_i - a - bx_i)^2 & \frac{\partial^2}{\partial a \partial b} \sum_{i=1}^{n} (y_i - a - bx_i)^2 \\ \frac{\partial^2}{\partial b \partial a} \sum_{i=1}^{n} (y_i - a - bx_i)^2 & \frac{\partial^2}{\partial b^2} \sum_{i=1}^{n} (y_i - a - bx_i)^2 \end{bmatrix} =$$

$$= \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

que tem uma forma quadrática definida positiva[2], isto é, os pontos de estacionaridade obtidos, a (ordenada na origem) e b (declive da recta), conduzem a desvios quadráticos mínimos.

7.5 Análise Elementar de Resíduos

Uma das formas de verificar se o modelo linear ajustado é adequado, é através da análise dos resíduos.

7.5.1 Diagrama de Dispersão dos Resíduos

Uma forma simples de visualizar os resíduos (e_i) é através de um diagrama de dispersão, representando os pontos (x_i, e_i). Num modelo bem ajustado os resíduos não podem ser "muito grandes" e devem apresentar-se de forma aleatória sem nenhum padrão particular definido.

Exemplos de resíduos com padrões típicos de ajustamentos inadequados são ilustrados na Figura 7.7.

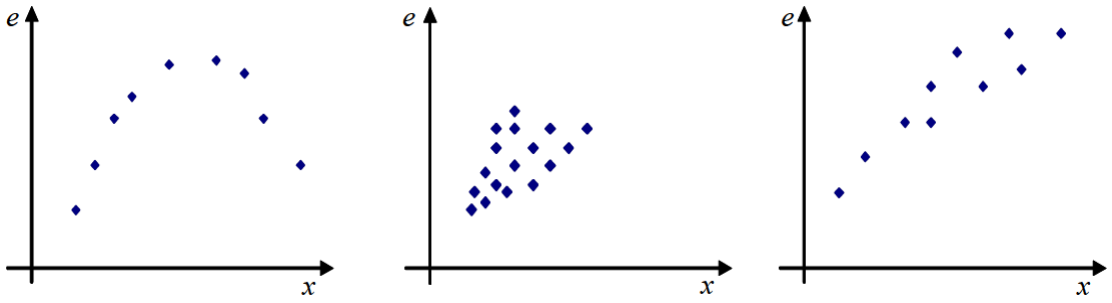

Figura 7.7: Diagramas de dispersão de resíduos.

[2]Esta hesseana é definida positiva pois,
$m_1 = |n| = n > 0$ e

$$m_2 = \begin{vmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{vmatrix} = n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 > 0.$$

Exemplo 7.5. Admita-se que X e Y representam, respectivamente, a altura e o peso de 12 estudantes seleccionados ao acaso entre os alunos de uma escola estando os dados representados na tabela seguinte,

Altura (cm)	Peso (kg)
155	70
150	63
180	72
135	60
156	66
168	70
178	74
160	65
132	62
145	67
139	67
152	68

Vamos começar por analisar estas duas variáveis através do diagrama de dispersão da Figura 7.8 e do coeficiente de correlação linear empírico.

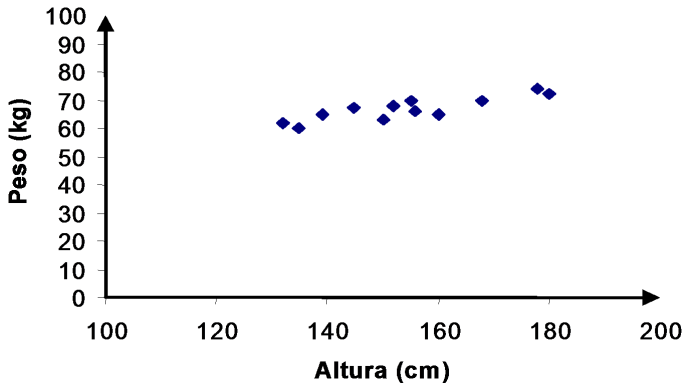

Figura 7.8: Diagrama de dispersão das alturas e pesos.

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{11}\sum_{i=1}^{12} x_i y_i - \frac{12}{11}\bar{x}\bar{y}}{\sqrt{\frac{1}{11}\sum_{i=1}^{12} x_i^2 - \frac{12}{11}\bar{x}^2}\sqrt{\frac{1}{11}\sum_{i=1}^{12} y_i^2 - \frac{12}{11}\bar{y}^2}} = 0.863.$$

Conclui-se que, tanto através do diagrama de dispersão como do coeficiente de correlação linear empírico, é favorável o ajustamento de uma recta de regressão linear. Vamos então

proceder ao seu cálculo com base nos valores da tabela seguinte:

	x	x^2	y	y^2	xy
	155	24025	70	4900	11550
	150	22500	63	3969	9450
	180	32400	72	5184	12960
	135	18225	60	3600	8100
	156	24336	66	4356	102960
	168	28224	70	4900	11760
	178	31684	74	5476	13172
	160	26500	65	4225	10400
	132	17424	62	3844	8184
	145	21025	67	4489	9715
	139	19321	65	4225	9035
	152	23104	68	4624	10336
Total	1850	287868	802	53792	124258

$$b = \frac{s_{xy}}{s_x^2} = \frac{12\sum_{i=1}^{12} x_i y_i - \sum_{i=1}^{12} x_i \sum_{i=1}^{12} y_i}{12\sum_{i=1}^{12} x_i^2 - \left(\sum_{i=1}^{12} x_i\right)^2} =$$
$$= \frac{12 \times 124258 - 1850 \times 802}{12 \times 287868 - (1850)^2} = 0.231733$$

$$a = \bar{y} - b\bar{x} = \frac{802}{12} - 0.231733 \times \frac{1850}{12} = 31.10778$$

Logo, a recta de regressão é

$$\hat{y} = 31.10778 + 0.231733x.$$

Graficamente, na Figura 7.9, está ajustada a recta de regressão à nuvem de pontos:

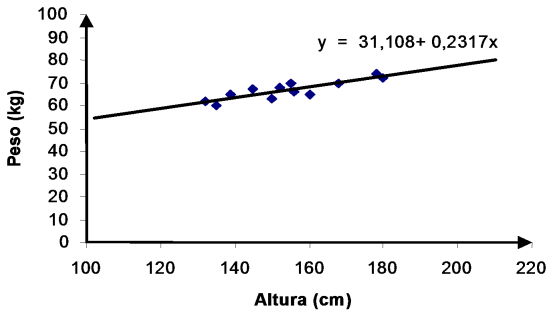

Figura 7.9: Ajustamento da recta de regressão à nuvem de pontos.

A análise da qualidade do ajustamento pode ainda fazer-se através da análise de resíduos.

Procedendo ao cálculo dos mesmos temos:

	x	y	$\hat{y}$	Resíduos ($e = y - \hat{y}$)
	155	70	67.03	2.97
	150	63	65.87	-2.87
	180	72	72.82	-0.82
	135	60	62.39	-2.39
	156	66	67.26	-1.26
	168	70	70.04	-0.04
	178	74	72.36	1.64
	160	65	68.19	-3.19
	132	62	61.70	0.30
	145	67	64.71	2.29
	139	65	63.32	1.68
	152	68	66.33	1.67
Total	1850	802	802.00	0.00

Podemos representar estes desvios graficamente através do diagrama de dispersão dos resíduos representado na Figura 7.10.

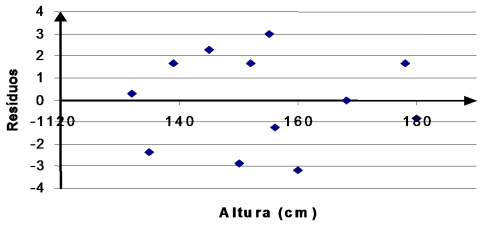

Figura 7.10: Diagrama de dispersão dos resíduos.

Este diagrama tem desvios pequenos (inferiores a 4 kgs) e exibe um padrão aleatório, concluindo-se que o modelo é adequado aos dados. ■

7.6 Outliers

Designa-se por *outlier* uma observação que se destaca das restantes. Os *outliers* podem existir devido a erros de recolha ou registo de dados ou pelo simples facto dos dados em análise possuirem observações com comportamentos distintos em relação às restantes. Observações deste tipo podem, de uma forma sumária, dividir-se em duas classes:

- *outliers* não influentes, em que a sua existência não altera o modelo linear ajustado;
- *outliers* influentes, em que a sua existência altera o modelo linear ajustado. Este tipo de outliers deve ser examinado e omitir-se quando se conclui que decorre de um erro; caso contrário deve ser estudado cuidadosamente.

Exemplo 7.6. No primeiro diagrama de dispersão da Figura 7.11 estamos perante um outlier não influente, pois o facto de este ser considerado, ou não, não altera o modelo linear ajustado.

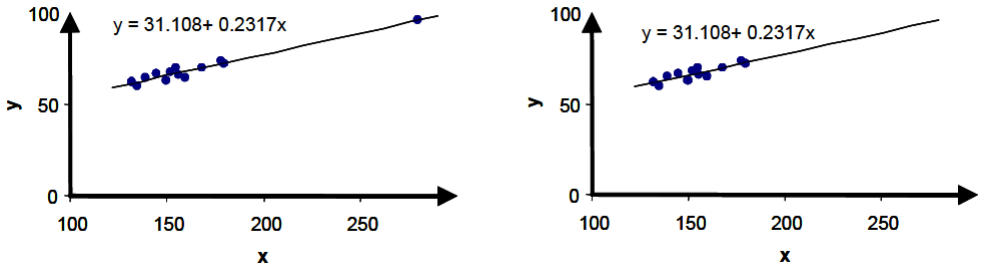

Figura 7.11: Diagrama de dispersão com outlier (esquerda) e sem outlier (direita).

No primeiro diagrama de dispersão da Figura 7.12 estamos perante um outlier influente, pois o facto de este ser considerado, ou não, altera completamente o modelo linear ajustado.

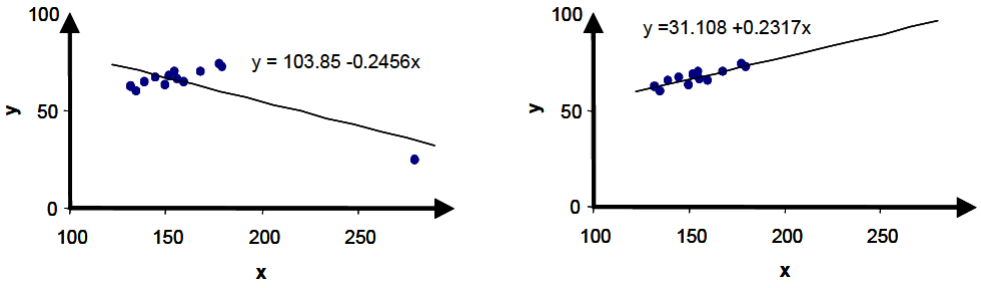

Figura 7.12: Diagrama de dispersão com outlier (esquerda) e sem outlier (direita).

■

Referências

[1] FISZ, M., *Probability Theory and Mathematical Statistics*, Jonh Wiley & Sons, Inc., New York, 1963.

[2] GUIMARÃES, R.C. e CABRAL, J.A.S., *Estatística*, McGraw-Hill de Portugal, Lisboa, 1997.

[3] MURTEIRA, B.J.F., *Probabilidades e Estatística*, McGraw-Hill de Portugal, Lisboa, 1979.

[4] SPIEGEL, M.R., *Probabilidade e Estatística*, Colecção Schaum, McGraw-Hill do Brasil, São Paulo, 1978

[5] OLIVEIRA, J.T., *Probabilidades e Estatística*, vol. I, Escolar Editora, Lisboa,1967.

[6] MELLO, F., *Introdução aos Métodos Estatísticos*, vol. I e II, Cadernos do Instituto de Orientação Profissional, Lisboa, 1973.

[7] MONTGOMERY, D.; RUNGER, G., *Applied Statistics and Probability for Engineers*, Jonh Wiley & Sons, Inc., New York, 2003.