ActivALEA
active e actualize a sua literacia

ASSOCIAÇÃO ENTRE VARIÁVEIS QUANTITATIVAS: O COEFICIENTE DE CORRELAÇÃO

A representação gráfica de um conjunto de dados bivariados é essencial, pois permite retirar informação sobre a forma, direcção e grau de associação entre as variáveis.

Por: Maria Eugénia Graça Martins
Departamento de Estatística e Investigação Operacional da FCUL
megm@fc.ul.pt

1 - Introdução

Quando dispomos de amostras de dados bivariados que vamos passar a representar por (x_i,y_i), i=1, …, n, onde os x_i's e os y_i's representam, respectivamente, observações das variáveis **x** e **y**, quantitativas, que constituem o par (x, y), a sua representação num *diagrama de dispersão* pode mostrar a existência de uma certa *associação linear* entre os factores x e y, que compõem os pares. Aliás, esta representação prévia dos dados bivariados, na forma de um diagrama de dispersão, é essencial, pois permite retirar informação sobre a forma, direcção e grau de associação entre as variáveis.

Se se concluir que tem sentido falar numa associação entre as variáveis, traduzida pela nuvem de pontos com a forma de uma oval, mais ou menos alongada, então passa-se a uma fase posterior, que será a medição do grau de intensidade com que as variáveis se associam, ou a construção de um modelo que permita conhecer como se reflectem numa das variáveis, as modificações processadas na outra, que são os modelos de regressão. Vamos, no que se segue, falar unicamente na forma de medir a intensidade com que as variáveis se associam.

2 - Coeficiente de correlação (amostral de Pearson)

A medida que se utiliza com mais frequência para medir o grau desta associação linear, é o *coeficiente de correlação*, que se representa por **r,** e se calcula a partir da expressão:

$$\mathbf{r} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \qquad \text{onde} \qquad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Na expressão anterior $\bar{x}$ e $\bar{y}$, representam, respectivamente, as médias dos x_i's e dos y_i's.

Na definição do coeficiente de correlação de pares de variáveis, está implícita a definição de uma medida que dá uma ideia da variabilidade conjunta existente entre as variáveis e que é a *covariância amostral*:

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

Esta medida tem o inconveniente de depender drasticamente das unidades com que se apresentam os elementos da amostra e daí o facto de normalmente não ser utilizada, passando-se à definição do coeficiente de correlação (independente das unidades utilizadas), que não é mais que a covariância das observações padronizadas $x_i^* = \frac{x_i - \bar{x}}{\sqrt{\text{var}(x)}}$ e $y_i^* = \frac{y_i - \bar{y}}{\sqrt{\text{var}(y)}}$ que, como facilmente se verifica da expressão anteriormente considerada, vem:

$$\text{Correlação (x,y)} = \frac{\text{covariância(x,y)}}{\sqrt{\text{variância(x)}}\sqrt{\text{variância(y)}}}$$

Propriedades do coeficiente de correlação:

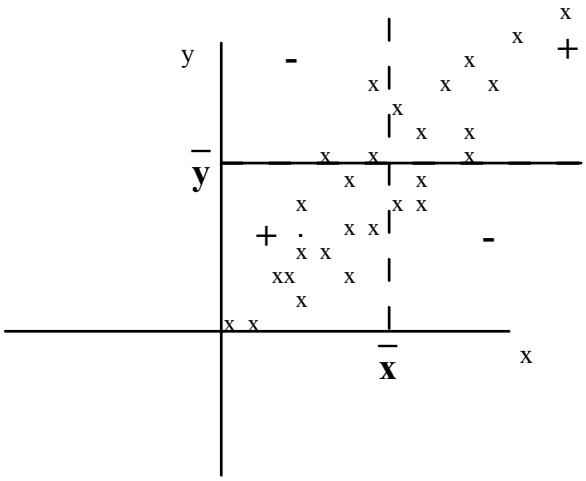
1 – O valor de **r** está no intervalo [-1,1]
2 – Quanto *maior* for o *módulo* de **r**, *maior* será a *relação linear* existente entre os x_i. e os y_i.
3 – O facto de **r** ser *positivo*, significa que a relação entre os x's e os y's é do *mesmo sentido*, isto é, a valores grandes de x, correspondem, em média, valores grandes de y e vice-versa. Quando **r** é *negativo*, a relação entre os x's e os y's é de *sentido contrário*, o que significa que a valores grandes de x, correspondem, em média, valores pequenos de y e vice-versa.

Interpretação geométrica:

1 – Se aos maiores valores de x estão associados, de um modo geral, os maiores valores de y, então r>0.
Efectivamente, quando pensamos num valor grande de x, será um valor acima da média. Por outro lado, um valor pequeno de x é um valor abaixo da média. Então se, de um modo geral, aos valores grandes de x estão associados os valores grandes de y, e aos valores pequenos de x estão associados os valores pequenos de y, os produtos

$$(x_i - \bar{x})(y_i - \bar{y})$$

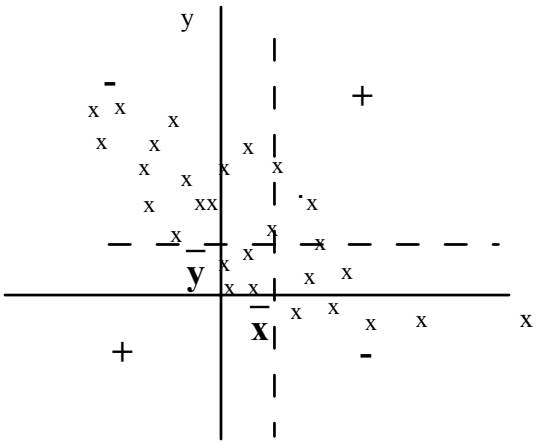
são de um modo geral positivos, já que ambos os factores são positivos ou negativos. Como o denominador da expressão do coeficiente de correlação, não depende da forma como os x's se associam com os y's, então o facto de no numerador somarmos grande número de parcelas positivas, faz com que o valor do coeficiente de correlação seja positivo e tanto maior quantas mais parcelas positivas houver.



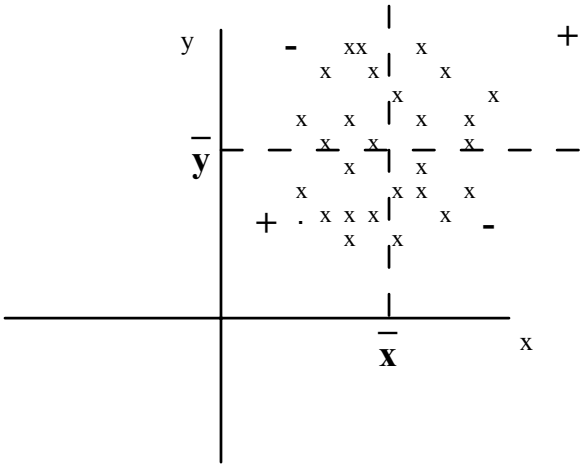
2 – Se aos maiores valores de x estão associados, de um modo geral, os menores valores de y, então r<0.
Fazendo o raciocínio como no ponto anterior, verificamos que agora as parcelas são maioritariamente negativas, já que quando x é grande (superior à média dos x's), então y é, de um modo geral, pequeno (inferior à média dos y's). Assim, os produtos

$$(x_i - \bar{x})(y_i - \bar{y})$$

são, de um modo geral, negativos.



3 – Se não existe qualquer tipo de associação linear entre os x's e os y's, então r=0.
Neste caso tanto podem surgir produtos negativos, como positivos, distribuindo-se de forma mais ou menos equitativa. Então o valor de r vem próximo de zero.

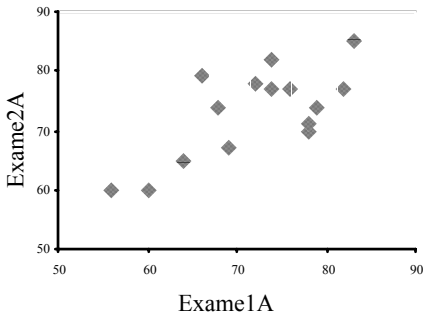


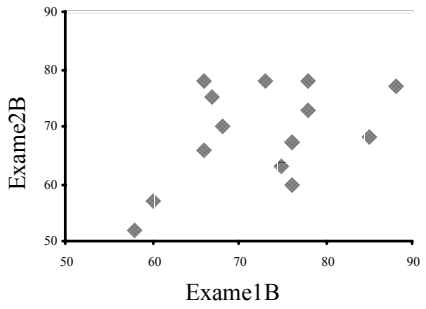
O cálculo do coeficiente de correlação deve ser objecto de alguns cuidados, como se sugere no que se segue.

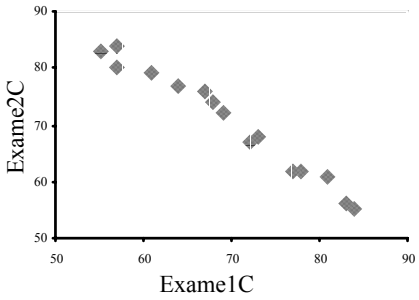
2.1 – Nem sempre o que parece, é!

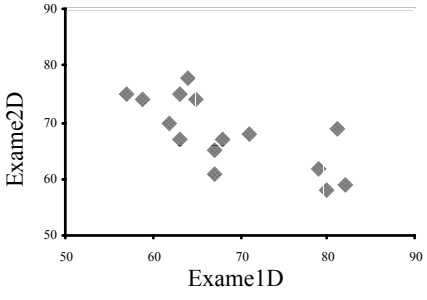
Dado um conjunto de dados, o cálculo do coeficiente de correlação, como medida de associação entre duas variáveis, pode causar-nos algumas surpresas, dando-nos informação errada sobre essa associação. Efectivamente, nem sempre o que parece, é! Vejamos o exemplo que se segue.

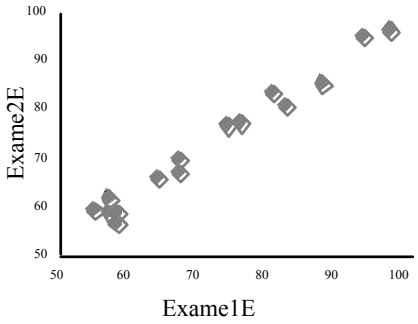
Exemplo (Rossman, 1996) **-** Considere os seguintes diagramas de dispersão correspondentes aos resultados de 2 exames de 6 classes (A-F).

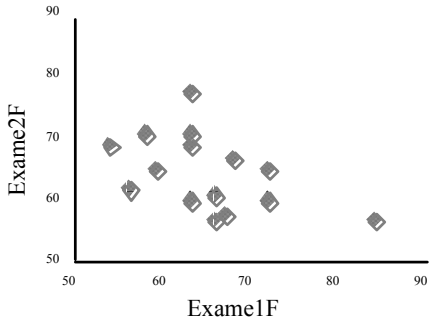












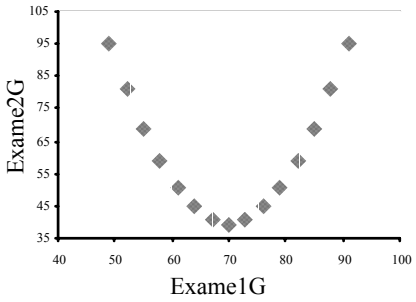
A visualização dos gráficos anteriores leva-nos a supor que entre os dois exames se possa admitir o seguinte tipo de associação:

	Forte	Moderada	Fraca
Positiva	E	A	B
Negativa	C	D	F

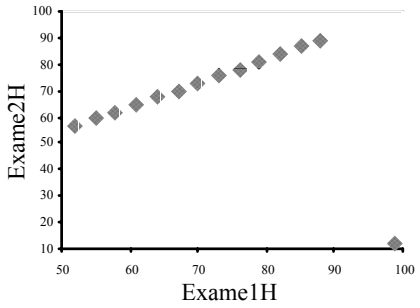
O cálculo do coeficiente de correlação, que se apresenta na tabela seguinte completa a informação da tabela anterior:

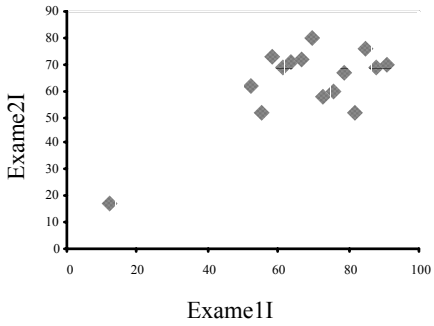
Classe	Correlação
A	0.71
B	0.47
C	-0.99
D	-0.72
E	0.99
F	-0.47

Considere agora a seguinte representação correspondente aos dados de uma classe G:



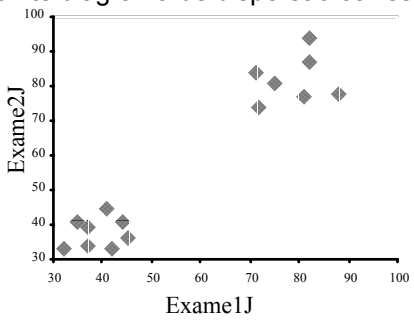
Como se verifica, existe uma forte associação entre os valores do exame 1 e os valores do exame 2. *Surpreendentemente* ao calcular o coeficiente de correlação obtemos o valor 0! Mas será assim tão surpreendente? Não, se nos lembrarmos que o que o coeficiente de correlação mede é o grau de associação linear e não outro tipo de associação, como a associação quadrática, presente nos dados da representação anterior.
Considere agora as duas representações correspondentes às notas obtidas pelas classes H e I:





O valor para o coeficiente de correlação é respectivamente 0.04 e 0.70 para as classes H e I, o que continua a ser surpreendente! Repare-se que relativamente à classe H todos os pares menos 1 seguem um padrão linear, quase perfeito, tendo-se obtido para o coeficiente de correlação um valor próximo de zero, enquanto que para a classe I, em que os valores se apresentam mais ou menos dispersos, obtivémos um valor relativamente alto. No entanto, se retirarmos a cada um dos conjuntos de dados anteriores o "outlier", que é o valor que se distingue dos restantes, em ambas as representações gráficas, já o valor do coeficiente de correlação passa para 0.9997 e 0.13, respectivamente para as classes H e I. O exemplo que acabámos de dar mostra que o coeficiente de correlação não é uma medida *resistente*, já que é muito influenciado pelos "outliers". Este facto não é de estranhar, já que no cálculo do coeficiente de correlação entramos com a média, que se sabe ser uma medida não resistente, isto é, é muito influenciada quer por valores muito pequenos, quer muito grandes, relativamente aos restantes..

Finalmente consideremos o seguinte diagrama de dispersão correspondente à classe J:



Da análise da representação anterior verificamos existirem dois grupos distintos de alunos: uns muito bons e outros muito maus. Embora para cada um dos grupos se verifique uma ligeira tendência para uma associação positiva, o facto é que o valor do coeficiente de correlação é 0.95, bem superior ao valor que seria de esperar.

Os exemplos que acabámos de ver, elucidam-nos sobre as limitações do coeficiente de correlação como medida de associação entre duas variáveis.

Antes de calcular e tentar interpretar o coeficiente de correlação entre duas variáveis, construa um diagrama de dispersão. Não esqueça que o coeficiente de correlação só mede a intensidade com que duas variáveis se associam linearmente, pelo que se a representação gráfica não mostrar evidência de associação linear, não tem sentido calculá-lo.

No caso dos exemplos apresentados, o cálculo do coeficiente de correlação das observações de um par de variáveis (x,y), sem uma visão prévia da forma como os pontos se apresentam graficamente, daria uma informação errada da forma e intensidade da associação linear entre as variáveis x e y.

Tem aqui todo o cabimento fazer referência às três regras básicas, que devemos ter presente, em qualquer análise inicial de dados (De Veaux et al, 2004):

1. *Faça uma representação gráfica*. Uma representação gráfica dos dados pode revelar informação acerca de padrões e relações existentes e escondidas nos dados, informação esta que não é visível a partir dos dados originais ou de tabelas.
2. *Faça uma representação gráfica*. Um gráfico bem escolhido realça aspectos importantes da distribuição dos dados.
3. *Faça uma representação gráfica*. A melhor forma de apresentar aos outros, o que pretende dizer sobre os seus dados, é através de uma representação grafica, bem escolhida. Não esqueça que "Um gráfico, vale mais que mil palavras". Não esqueça também que nem sempre é verdade...
Apresentamos ainda mais um exemplo que realça a importância que os gráficos têm, numa análise prévia de dados bivariados.

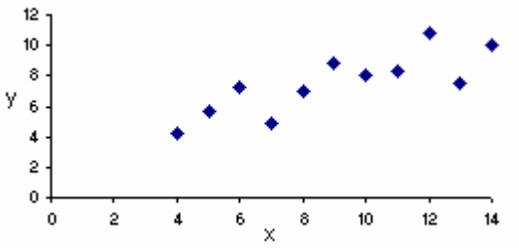
Exemplo (Adaptado de http://www.itl.nist.gov/div898/handbook/eda/section1/eda16.htm)
Este é um exemplo clássico (Anscombe) da importância que os gráficos representam no estudo de um conjunto de dados:

Dados:	x	y
	10.00	8.04
	8.00	6.95
	13.00	7.58
	9.00	8.81
	11.00	8.33
	14.00	9.96
	6.00	7.24
	4.00	4.26
	12.00	10.84
	7.00	4.82
	5.00	5.68

Estatísticas descritivas: n = 11
$\bar{x}$ = 9.0
$\bar{y}$ = 7.5
Correlação (x,y) = 0.816

A informação anterior, embora válida, dá-nos só uma informação limitada dos dados.

Diagrama de dispersão: Ao contrário, o diagrama de dispersão dos dados,



sugere o seguinte:
1. A nuvem de pontos sugere uma associação linear entre as variáveis
2. Não se justificam modelos mais complicados (por ex. Quadráticos) para descrever os dados
3. Não existem outliers.

Vejamos mais três conjuntos de dados:

x_2	y_2	x_3	y_3	x_4	y_4
10.00	9.14	10.00	7.46	8.00	6.58
8.00	8.14	8.00	6.77	8.00	5.76
13.00	8.74	13.00	12.74	8.00	7.71
9.00	8.77	9.00	7.11	8.00	8.84
11.00	9.26	11.00	7.81	8.00	8.47
14.00	8.10	14.00	8.84	8.00	7.04
6.00	6.13	6.00	6.08	8.00	5.25
4.00	3.10	4.00	5.39	8.00	12.50
12.00	9.13	12.00	8.15	8.00	5.56
7.00	7.26	7.00	6.42	8.00	7.91
5.00	4.74	5.00	5.73	8.00	6.89

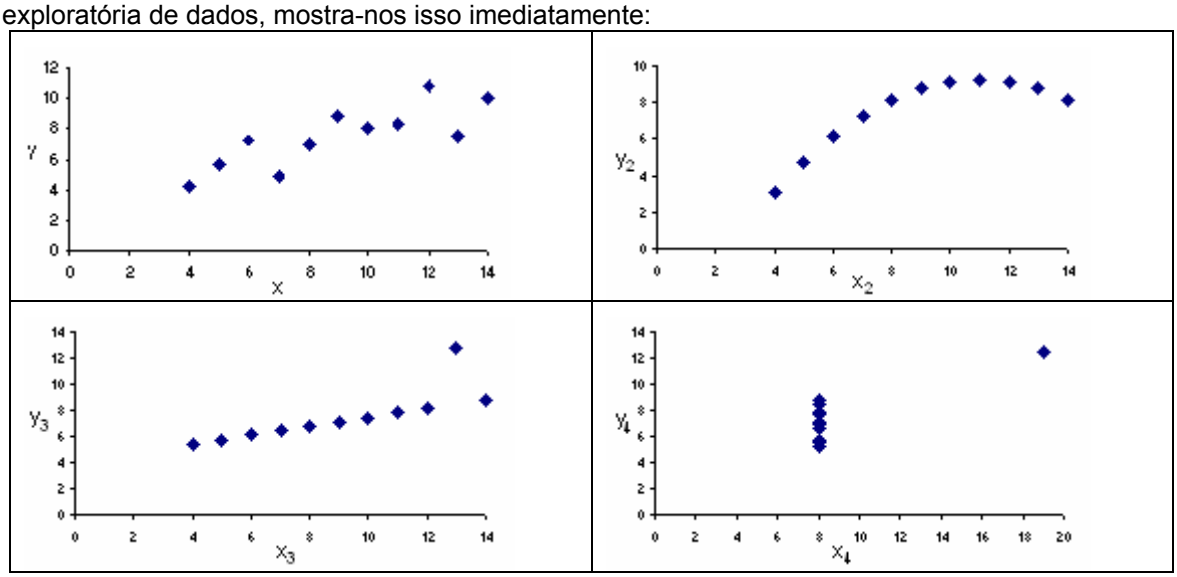
Calculando as estatísticas descritivas para os conjuntos de dados anteriores, temos:

Conjunto de dados 2: n = 11
$\bar{x}_2 = 9.0$
$\bar{y}_2 = 7.5$
Correlação (x,y) = 0.816

Conjunto de dados 3: n = 11
$\bar{x}_3 = 9.0$
$\bar{y}_3 = 7.5$
Correlação (x,y) = 0.816

Conjunto de dados 4: n = 11
$\bar{x}_4 = 9.0$
$\bar{y}_4 = 7.5$
Correlação (x,y) = 0.817

Então, tem algum sentido dizer que, sob o ponto de vista "quantitativo", os quatro conjuntos de dados são equivalentes. De facto, os quatro conjuntos de dados estão longe de serem equivalentes e uma representação gráfica, que deve ser o primeiro passo de uma análise exploratória de dados, mostra-nos isso imediatamente:



Das representações gráficas anteriores, concluímos imediatamente que:

1. O conjunto 1 apresenta uma associação claramente linear
2. O conjunto 2 apresenta uma associação quadrática
3. O conjunto 3 tem claramente um outlier
4. O conjunto 4 mostra um planeamento, eventualmente mal feito, em que um dos pontos aparece removido do conjunto dos outros.

Os exemplos anteriores mostram que as estatísticas que utilizamos para reduzir a informação contida nos dados, são úteis, mas dão uma visão muito incompleta e limitada dos dados. Elas reduzem drasticamente a informação contida nos dados, através de alguns números. Ao fazerem esta redução dos dados, omitem aspectos importantes e cruciais, pelo que, na melhor situação podemos dizer que dão informação incompleta, mas na pior situação podemos dizer mesmo, que dão informação errada.

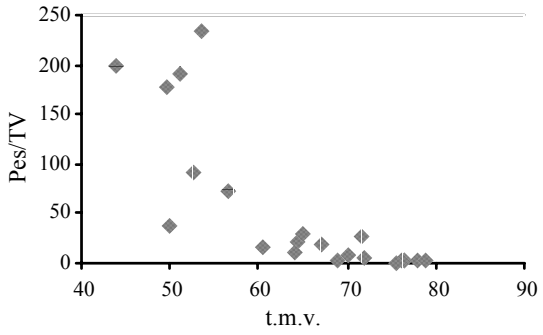
2.2 – Não confundir correlação com relação causa-efeito

Um outro aspecto que não pode deixar de ser referido quando estamos perante uma correlação forte entre duas variáveis, é que isso não significa necessariamente uma relação de causa-efeito. Vejamos o seguinte exemplo.

Exemplo (Rossman, 1996) - A seguinte tabela apresenta para um conjunto de 22 países, o tempo médio de vida (t.m.v) e o número de pessoas por aparelho de televisão (Pes/TV):

País	t.m.v.	Pes/TV	País	t.m.v.	Pes/TV
Angola	44	200	México	72	6.6
Austrália	76.5	2	Marrocos	64.5	21
Cambodja	49.5	177	Paquistão	56.5	73
Canadá	76.5	1.7	Russia	69	3.2
China	70	8	África Sul	64	11
Egipto	60.5	15	Sri Lanka	71.5	28
França	78	2.6	Uganda	51	191
Haiti	53.5	234	ReinoUnido	76	3
Iraque	67	18	EUA	75.5	1.3
Japão	79	1.8	Vietnam	65	29
Madagáscar	52.5	92	Yemen	50	38

O valor do coeficiente de correlação entre as variáveis t.m.v e Pes/TV é igual a -0.80, o que significa uma forte correlação negativa entre o tempo médio de vida e o número de pessoas por aparelho de TV, ou seja, quanto maior for o número de pessoas por aparelho de TV, menor é o tempo médio de vida. Será que então se pode aumentar o tempo médio de vida da população de um país, aumentando o número de aparelhos de TV? Seria ridículo pensar desta maneira, pois este é um exemplo em que sobressai que não se pode admitir uma relação de causa-efeito. Obviamente existem outras variáveis não observadas -*variáveis perturbadoras* - relacionadas com o nível de vida na população, que provocam alterações nas duas variáveis que estamos a estudar e que explicam a forte correlação verificada. O diagrama de dispersão das variáveis estudadas tem o seguinte aspecto:



Não confundir correlação com relação causa-efeito. Um diagrama de pontos e uma correlação não provam a existência de uma relação causa-efeito. Podem existir outras variáveis, que não são estudadas, mas influenciam as que estão a ser estudadas e que são conhecidas como "lurking variables" (temos dificuldade em arranjar uma tradução adequada, pelo que vamos utilizar o termo "variáveis perturbadoras").

Bibliografia

De Veaux, R. and Velleman, P. (2004) – *Intro Stats*, Pearson Education.
Graça Martins, M.E. (2005) – *Introdução às Probabilidades e Estatística*, Sociedade Portuguesa de Estatística
Rossman, A. (1996) – *Workshop Statistics, Discovery with data*, Springer-Verlag New York

Algumas referências úteis do ALEA:

- Dados bivariados:
  www.alea.pt/index.php?option=com_content&view=article&id=820&Itemid=1873
- Coeficiente de Correlação:
  www.alea.pt/index.php?option=com_content&view=article&id=838&Itemid=1876
- Como realizar Diagramas de dispersão e calcular do coeficiente de correlação no Excel: Dossier didáctico nº 4:
  www.alea.pt/index.php?option=com_content&view=article&id=309&Itemid=1713