1. O que são dados multivariados?

Uma **amostra** diz-se **bivariada** quando é constituída por pares ordenados de dados. Em cada par de valores, existe sempre uma relação entre o primeiro valor e segundo valor; por exemplo, foram determinados para um mesmo indivíduo, num mesmo dia, etc…

Por outras palavras, uma amostra bivariada é composta por n pares de valores de duas variáveis X e Y, (X , Y) , sendo cada par medido sobre um mesmo indivíduo, uma mesma situação, etc.

Uma amostra bivariada de dimensão n:

$$(x_1 , y_1) ; (x_2 , y_2) ; (x_3 , y_3) ; \ldots ; (x_i , y_i) ; \ldots ; (x_n , y_n)$$

Uma **amostra** diz-se **multivariada** (**trivariada, tetravariada**, etc… , **k-variada**) quando é constituída por conjuntos de três ou mais dados ordenados. Tal como antes, em cada conjunto de valores, existe uma relação entre o primeiro valor, o segundo valor, … e o k^o valores; por exemplo, foram determinados para um mesmo indivíduo, num mesmo dia, etc…

Do mesmo modo, por outras palavras, uma amostra multivariada é composta por n conjuntos de k valores de k variáveis X, Y, Z, … (X , Y , Z , …) , sendo cada conjunto medido sobre um mesmo indivíduo, uma mesma situação, etc.

Uma amostra trivariada de dimensão n:

$$(x_1 , y_1 , z_1) ; (x_2 , y_2 , z_2) ; (x_3 , y_3 , z_3) ; \ldots ; (x_i , y_i , z_i) ; \ldots ; (x_n , y_n , z_n)$$

Neste módulo, apenas nos iremos preocupar com amostras bivariadas.

Podemos caracterizar uma amostra multivariada de várias formas, com vários objetivos:

1º - Caracterizar separadamente o conjunto dos valores amostrais de cada uma das variáveis, como sendo uma amostra univariada (isto é, caracterizar X, no que respeita a localização, dispersão, … e / ou caracterizar Y, no que respeita a localização, dispersão,… separadamente). Os conhecimentos necessários para este tratamento já foram apresentados no capítulo anterior (ver cap.2 secção 2).

2º - Verificar se existe alguma relação de associação entre as várias variáveis (X, Y, Z, …) e, em caso afirmativo, caracterizar essa relação.

É sobretudo sobre este segundo ponto que nos vamos concentrar neste capítulo.

Exemplo 1 – Seleccionaram-se 100 adultos ao acaso e registaram-se os seus altura e peso. Obtiveram-se assim 100 pares de valores (x , y), onde X = altura e Y = peso. O conjunto destes 100 pares de valores constitui uma amostra bivariada de dimensão n = 100.

Podemos:

1 - Caracterizar X separadamente, isto é, podemos tomar as 100 alturas e analisá-las como sendo uma amostra de 100 valores simples

2 - Caracterizar Y separadamente, isto é, podemos tomar os 100 pesos e analisá-los como sendo uma amostra de 100 valores simples

3 – Avaliar se há alguma relação entre X e Y (entre altura e peso) e, em caso afirmativo, caracterizar essa relação.

2. Diagrama de dispersão (para dados bivariados)

2.1 O que é um Diagrama de Dispersão? Qual é o seu interesse?

Um **diagrama de dispersão** é a forma mais simples de avaliar a relação existente entre duas variáveis e consiste na elaboração de um gráfico xy, no qual cada par ordenado de valores amostrais corresponde a um ponto.

A análise de um diagrama de dispersão apresenta a desvantagem de as conclusões dependerem do técnico que faz o estudo (análise subjectiva).

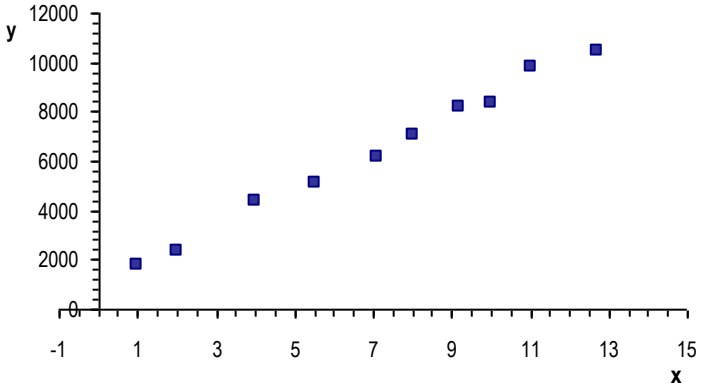
No entanto, embora haja outros métodos de caracterizar a relação existente entre X e Y, como iremos ver subsequentemente, a função de um diagrama de dispersão nunca é substituída, porque:

- fornece rapidamente um primeiro indício da existência ou não de relação e do tipo de relação;

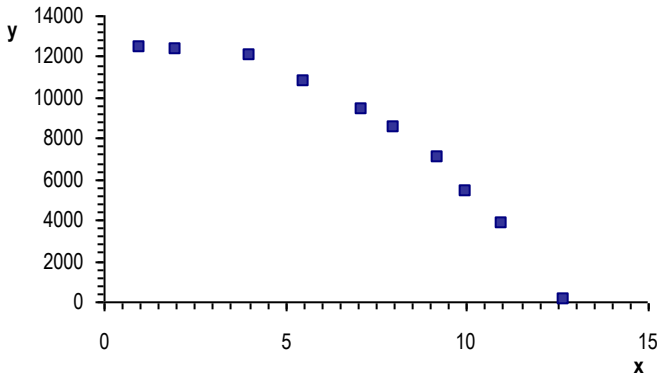
- nada substitui a análise pessoal, sobretudo em situações nas quais há logo à partida um conhecimento científico sobre o assunto.

Apresentam-se de seguida alguns exemplos.

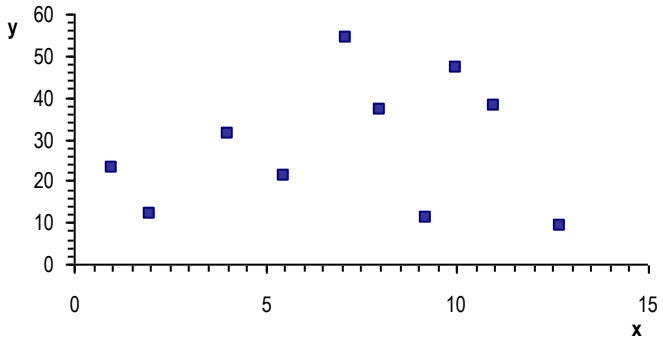
Exemplo 2 - Nesta amostra bivariada, parece haver uma relação linear entre X e Y.



Exemplo 3 – Nesta amostra bivariada, verifica-se que há uma relação entre X e Y, mas não linear – talvez possa descrita por um ramo de uma parábola…



Exemplo 4 – *Nesta amostra bivariada não parece existir qualquer relação entre X e Y, já que no respectivo diagrama de dispersão os dados apresentam-se em "nuvem". As variáveis X e Y serão portanto variáveis independentes.*
É o caso de, por exemplo, as variáveis X = temperatura do ar em Melgaço e Y = consumo diário de carne, na China.



2.2 Construção de Diagramas de Dispersão em Excel

O diagrama de dispersão pode facilmente ser construído em Excel .

Procedimento:

1 – Introduza os valores amostrais da amostra bivariada, por exemplo, em duas colunas.

2 – Seleccione as colunas dos dados e construa um gráfico do tipo ***Dispersão***, sub-tipo ***só pontos***.

Exemplo 5 – *Veja Exemplo 5, no ficheiro Capítulo 4 Exemplos.xls*

3. O coeficiente de correlação amostral

3.1 Cálculos auxiliares

Nas secções seguintes, iremos ver outras formas de caracterizar a relação existente entre as duas variáveis. Dado que estes processos envolvem a aplicação de várias expressões matemáticas, ajuda no tratamento dos dados realizar previamente os seguintes cálculos:

$$\sum_{i=1}^{n} x_i \;,\; \sum_{i=1}^{n} y_i \;,\; \sum_{i=1}^{n} x_i^2 \;,\; \sum_{i=1}^{n} y_i^2 \;,\; \sum_{i=1}^{n} x_i.y_i$$

NOTA: *Estes cálculos são mais comodamente realizados mediante a construção de uma tabela auxiliar de cálculo.*

3.2 Definição e cálculo do coeficiente de correlação amostral

A relação linear do tipo Y = a + b X é a relação mais aplicada e porventura aquela que melhor descreve, pelo menos de uma forma aproximada, muitos fenómenos naturais e sócio-culturais.

O **coeficiente de correlação amostral** , **r** , é uma medida que nos permite quantificar a possível existência de uma relação linear do tipo Y = a + b X entre as duas variáveis de estudo.

Note-se que esta medida apenas caracteriza a possibilidade de existência de uma relação linear deste tipo, nada nos dizendo acerca da possibilidade de existência de um outro tipo de relação.

Define-se **covariância amostral** entre duas variáveis X e Y, como sendo:

$$cov_{xy} = \frac{\sum (x_i - \bar{x}).(y_i - \bar{y})}{n-1} = \frac{\sum x_i . y_i - n\bar{x}.\bar{y}}{n-1}$$

A primeira é fórmula de definição ; a segunda é "mais cómoda" para efeitos de cálculo.

O **coeficiente de correlação amostral**, $\mathbf{r_{xy}}$, é dado pela seguinte expressão:

$$r_{xy} = \frac{cov_{xy}}{s_x . s_y}$$

na qual s_x e s_y são respectivamente os valores dos desvios padrão de x e de y (ver cap. 2, sec. 2).

Prova-se que $\mathbf{r_{xy}}$ também pode ser calculado por:

$$r_{xy} = \frac{\sum x_i . y_i - n\bar{x}.\bar{y}}{\sqrt{\sum x_i^2 - n.\bar{x}^2} . \sqrt{\sum y_i^2 - n.\bar{y}^2}}$$

É sempre:

$$\mathbf{-1 \leq r_{xy} \leq 1}$$

3.3 Cálculo em Excel

Há duas formas diferentes de calcular o valor de r_{xy} em Excel:

- Directamente, pela fórmula de definição

Calcule os desvios padrão e as médias das duas variáveis e depois a covariância, partindo dos somatórios ; poderá então aplicar a fórmula de cálculo de r_{xy}.

- Usando a função CORREL

Seleccione a célula de saída do resultado ; escreva = ; seleccione a função CORREL e escolha os valores das duas variáveis.

Exemplo 6 – Veja Exemplo 6, no ficheiro Capítulo 4 Exemplos.xls

3.4 Significado do coeficiente de correlação amostral

Interpretação do valor de r_{xy}:

Se $\mathbf{r_{xy} > 0}$, existe uma perfeita ou imperfeita relação linear (Y = a + b X) positiva entre X e Y. Isto significa que o declive da recta, b, que descreve a relação é positivo, ou seja, quando X aumenta, Y aumenta.

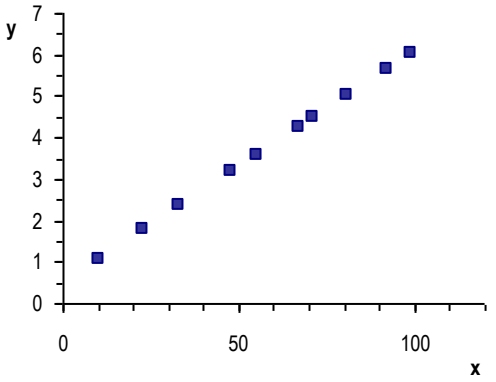
Se $\mathbf{r_{xy} < 0}$, existe uma perfeita ou imperfeita relação linear(Y = a + b X) negativa entre X e Y. Isto significa que o declive da recta, b, que descreve a relação é negativo, ou seja, quando X aumenta, Y diminui.

Se $\mathbf{r_{xy} = \pm 1}$, a relação linear é perfeita. Isso significa que no diagrama de dispersão todos os pontos se distribuem alinhados exactamente de acordo com uma recta de equação Y = a + b X . A relação é positiva (crescente) se r_{xy} for positivo e negativa (decrescente) se r_{xy} for negativo.

Se $\mathbf{-1 < r_{xy} < 0}$ ou então $\mathbf{0 < r_{xy} < 1}$, a relação linear é imperfeita ; no diagrama de dispersão os pontos não se encontram exactamente alinhados sobre uma recta. Quanto maior for o valor absoluto de r_{xy}, mais perfeita será a relação linear.
Na prática, apenas se considera que existe uma relação linear do tipo Y = a + b X de qualidade razoável se $r_{xy} < -0,9$ (relação negativa) ou então $r_{xy} > 0,9$ (relação positiva).

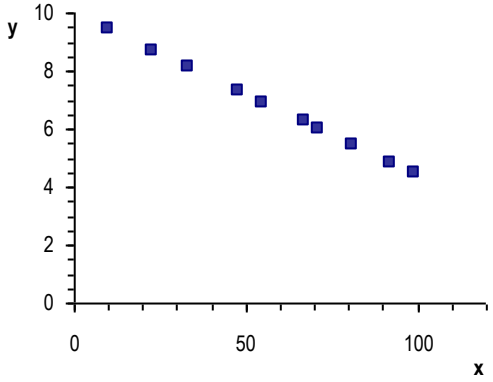
Se $\mathbf{r_{xy} = 0}$ então podemos concluir da ausência de uma relação linear do tipo Y = a + b X , o que não significa necessariamente que não exista qualquer relação entre X e Y ; ou seja, neste caso, podemos dizer que X e Y ou são variáveis independentes ou possuem entre si uma relação não linear ou linear de outro tipo. O diagrama de dispersão permitirá esclarecer melhor a situação.

Exemplo 7 -



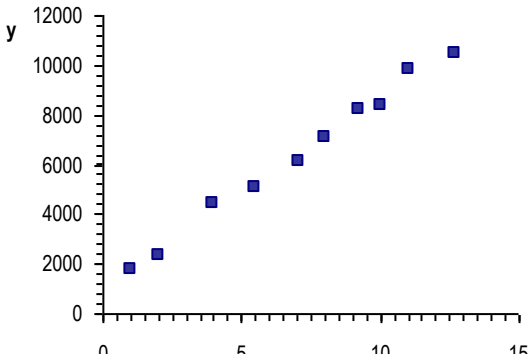
Relação linear perfeita positiva.
$r_{xy} = 1$.

Exemplo 8 -



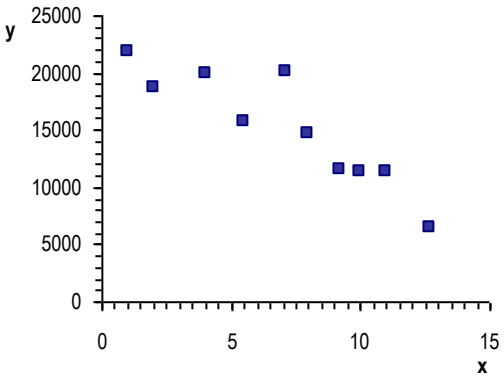
Relação linear perfeita negativa.
$r_{xy} = -1$.

Exemplo 9 -



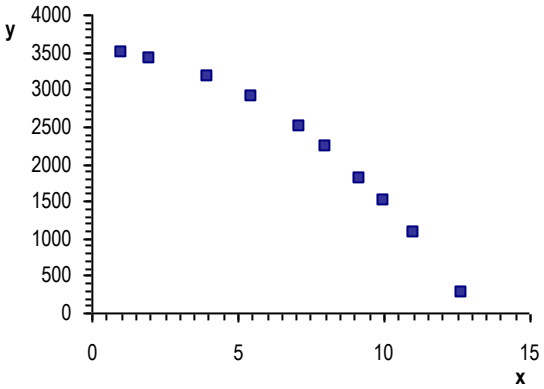
Relação linear imperfeita positiva.
r_{xy} = 0,97.

Exemplo 10 -



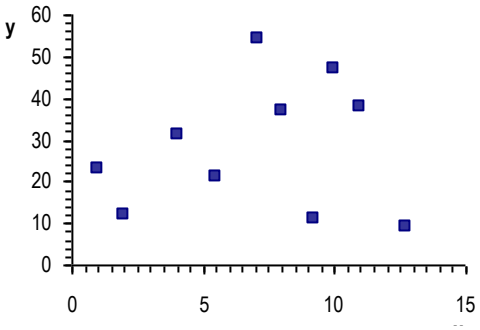
Relação linear imperfeita negativa.
r_{xy} = - 0,89.

Exemplo 11 -



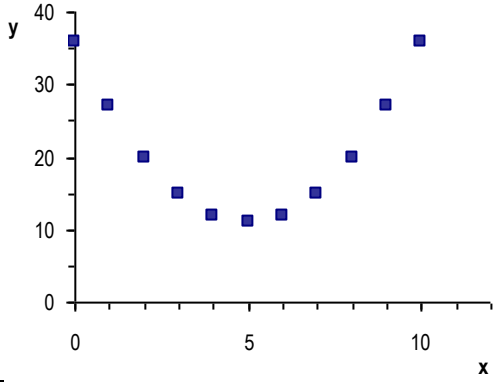
Relação linear imperfeita negativa.
r_{xy} = - 0,91.

Exemplo 12 -



Nao há qualquer relação ; X e Y são variáveis independentes.
$r_{xy} \approx 0$

Exemplo 13 -



Não há relação linear, mas há um outro tipo de relação (perfeita) entre X e Y – parece haver uma relação quadrática…

No entanto, neste caso: $r_{xy} = 0$ *!!*

4. Ajuste linear

Quando se decide que uma função linear descreve bem a relação existente entre duas variáveis X e Y, há que estabelecer – ou quantificar – essa relação; ou seja, há que fazer o chamado **ajuste linear**.

A função que se pretende será da forma:

$$\hat{Y} = \mathbf{m\,X + b}$$

pelo que, fazer o ajuste linear significa determinar os valores mais adequados para **b** (ordenada na origem) e **m** (declive) – os parâmetros da função a ajustar.

$\hat{Y}$ significa o valor de Y *estimado*, a partir do correspondente valor de X pela função de ajuste encontrada. No caso geral, $\hat{Y} \neq Y$. O que se pretende é que cada $\hat{Y}_i$ seja o mais próximo possível do correspondente verdadeiro valor Y_i.

4.1 Método a Sentimento

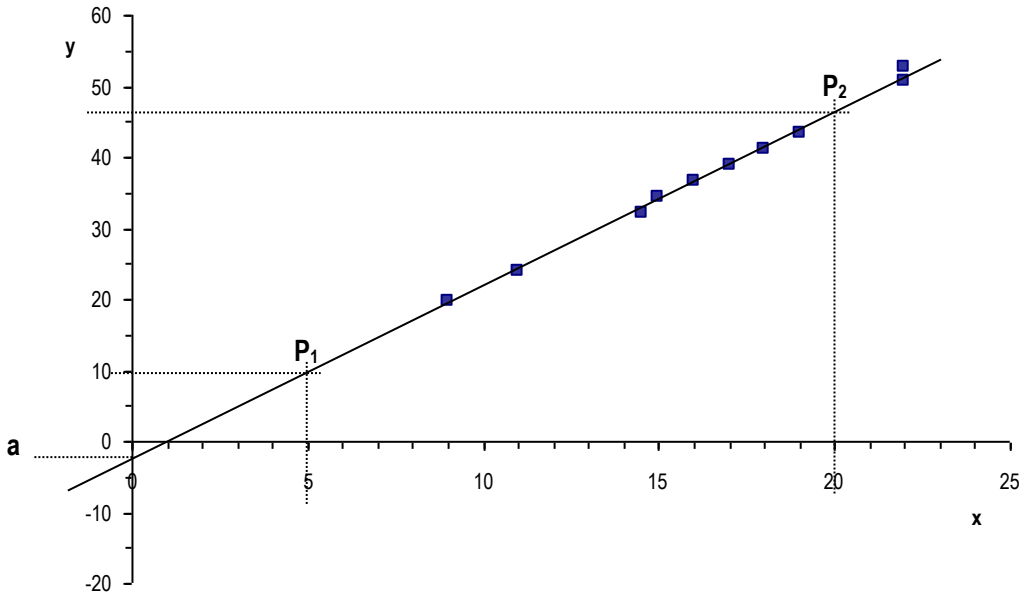
Um método que pode ser utilizado é o chamado **Método a Sentimento**. Este método consiste em elaborar o mais rigorosamente possível um diagrama de dispersão (por exemplo, em papel milimétrico) e em seguida traçar sobre o diagrama a melhor recta – aquela que, mesmo que não inclua nenhum dos pontos experimentais, se encontre simultaneamente o mais próxima possível de todos os pontos.

Prolongando-se a recta obtida até à abcissa nula; o valor da ordenada nesse ponto é o valor de **a**. Para determinar o valor do declive, é necessário primeiro fixar dois pontos quaisquer da recta ajustada – convém que sejam bastante afastados um do outro, para minimizar o erro – e determinar as suas coordenadas (x_1 , y_1) e (x_2 , y_2) . O declive, **b**, vem então dado pela seguinte expressão:

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

O método a sentimento apresenta o inconveniente de o resultado obtido depender do analista que faz o ajuste. Tem porém a vantagem de permitir opções críticas (por exemplo, se o analista acha que certo ponto experimental estará eventualmente afectado de erro de medição ou outro, poderá atribuir-lhe uma menor importância no traçado da recta de ajuste, sem no entanto o excluir completamente do seu julgamento).

Exemplo 14 - A figura abaixo ilustra a aplicação do método a sentimento na determinação de um ajuste linear a um conjunto de dados bivariados. Repare que foi dada uma menor importância ao último ponto, por se encontrar um pouco desalinhado com a recta definida pelos restantes.



Determinou-se que: $b = -2,5$

P_1*: (5,0 ; 9,5)* P_2*: (20,0 ; 46,0)*

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{46,0 - 9,5}{20,0 - 5,0} = 2,4$$

Logo, a função de ajuste linear determinada é a seguinte: $\hat{Y} = -2,5 + 2,4 \, . \, X$

4.2 Método dos Mínimos Quadrados

Outro método de determinação do ajuste linear é o chamado **Método dos Mínimos Quadrados (M.M.Q.)**. Este método baseia-se na aplicação de equações com base na minimização das distâncias de todos os pontos à recta de ajuste a estabelecer. Deduz-se que os parâmetros **m** e **b** vêm dados pelas seguintes expressões [1]:

$$m = \frac{\sum x_i \, . \, y_i - n \, . \, \bar{x} \, . \, \bar{y}}{\sum x_i^{\,2} - n \, . \, \bar{x}^{\,2}} \qquad b = \bar{y} - m \, . \, \bar{x}$$

A aplicação do Método dos Mínimos Quadrados apresenta a grande vantagem de o resultado obtido ser independente do analista que calcula o ajuste e é sem dúvida alguma o método de ajuste mais largamente usado.

[1] A dedução destas expressões encontra-se fora do âmbito deste módulo de formação.

Exemplo 15 - Pretende-se realizar um ajuste linear ao conjunto de dados tratado no exemplo 12, mas agora por aplicação do Método dos Mínimos Quadrados. Pretende-se também quantificar a qualidade do ajuste realizado.

Dados e cálculos auxiliares:

	x_i	y_i	x_i^2	y_i^2	$x_i \, . \, y_i$
	11	23,9	121	571,21	262,9
	15	34,3	225	1176,49	514,5
	17	38,8	289	1505,44	659,6
	16	36,5	256	1332,25	584
	19	43,4	361	1883,56	824,6
	22	50,7	484	2570,49	1115,4
	14,5	32	210,25	1024	464
	18	41,2	324	1697,44	741,6
	9	19,7	81	388,09	177,3
	22	52,6	484	2766,76	1157,2
SOMAS:	163,5	373,1	2835,25	14915,73	6501,1

Cálculo do coeficiente de correlação amostral:

$$r_{xy} = \frac{\sum x_i \, . \, y_i - n \, \bar{x} \, . \, \bar{y}}{\sqrt{\sum x_i^2 - n \, . \, \bar{x}^2} \, . \sqrt{\sum y_i^2 - n \, . \, \bar{y}^2}} =$$

$$= \frac{6501{,}1 - 10 \times 16{,}35 \times 37{,}31}{\sqrt{2835{,}25 - 10 \times 16{,}35^2} \times \sqrt{14915{,}73 - 10 \times 37{,}31^2}} = 0{,}9983$$

Logo, existe uma correlação linear imperfeita (bastante boa) entre X e Y.

Determinação da função linear de ajuste:

$n = 10$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{163{,}5}{10} = 16{,}35$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{373{,}1}{10} = 37{,}31$$

$$m = \frac{\sum x_i \, . \, y_i - n \, . \, \bar{x} \, . \, \bar{y}}{\sum x_i^2 - n \, . \, \bar{x}^2} = \frac{6501{,}1 - 10 \times 16{,}35 \times 37{,}31}{2835{,}25 - 10 \times 16{,}35^2} = 2{,}47$$

$$b = \bar{y} - b \, . \, \bar{x} = 37{,}31 - 2{,}47 \times 16{,}35 = -3{,}07$$

Logo, a função linear de ajuste é: $\hat{Y} = -3{,}07 + 2{,}47 \, . \, X$

4.3 Determinação do Ajuste Linear em Excel

Há três formas diferentes de calcular os valores de a e de b em Excel:

- Directamente, pela fórmula de definição

Calcule os desvios padrão e as médias das duas variáveis e depois a covariância, partindo dos somatórios ; poderá então aplicar a fórmula de cálculo de a e de b.

- Usando as funções DECLIVE e INTERCEPTAR

Seleccione a célula de saída do resultado ; escreva = ; seleccione a função desejada e escolha os valores das duas variáveis, tendo o cuidado de seleccionar correctamente a variável dependente (Y) e a variável independente (X).

A função **DECLIVE** calcula o valor de m ; a função **INTERCEPTAR** calcula o valor de b.

- Adicionando ao Diagrama de Dispersão uma Linha de Tendência

Sobre o diagrama de dispersão, seleccione os pontos (basta clicar sobre um deles)

Seleccione o Menu ***Gráfico*** e o sub-menú ***Adicionar linha de tendência***.

Seleccione ***Linear*** e, em opções escolha "***Mostrar a equação no gráfico***". Prima OK.

A linha ajustada aparece desenhada e a sua equação característica surge no gráfico.

Exemplo 16 – Veja Exemplo 16, no ficheiro Capítulo 4 Exemplos.xls

4.4 Qual é o interesse em realizar um ajuste linear (ou outro ajuste)? Previsões.

Possuir uma expressão que relacione matematicamente duas variáveis é extremamente útil porque nos permite, entre outras coisas:

- Interpretar fenómenos observados à luz de teorias científicas conhecidas.

- Desenvolver novas teorias científicas através de dados recolhidos (investigação científica).

- Modelizar fenómenos – por exemplo para programação.

- Realizar previsões dos valores de uma variável a partir de valores ainda não testados da outra. Se os novos valores forem superiores a pelo menos um valor que serviu de base ao ajuste e inferiores também a pelo menos um valor que serviu de base ao ajuste, diz-se que a previsão é obtida por **interpolação**, caso contrário, diz-se foi obtida por **extrapolação**. As previsões obtidas por interpolação são sempre mais fiáveis do que aquelas que são obtidas por extrapolação.

Exemplo 17 - Considere os resultados obtidos no exemplo 15.

Estime o valor de Y quando X = 10,00 e preveja para que valor de X será Y = 25,00.

A equação de ajuste obtida é: $\hat{Y} = -3{,}07 + 2{,}47 \, . \, X.$

. Fazendo X = 10,00 , vem: $\hat{Y} = -3{,}07 + 2{,}47 \times 10{,}00 = 21{,}63$
Este valor foi obtido por interpolação.

. Fazendo Y = 25,00 , vem: $25{,}00 = -3{,}07 + 2{,}47 \, . \, \hat{X} \Leftrightarrow \hat{X} = 11{,}36$
Este valor foi obtido por extrapolação.

4.5 Caracterização da Qualidade do Ajuste – O Coeficiente de Determinação.

Após termos realizado um ajuste (linear ou outro) é essencial quantificar a qualidade desse ajuste, isto é até que ponto a expressão encontrada descreve a relação entre as duas variáveis relacionadas.

O **Coeficiente de Determinação** , $\mathbf{R^2}$, é uma medida que nos permite quantificar uma qualquer possível relação entre as duas variáveis. No caso geral, só podemos calcular este coeficiente mediante a prévia definição para cada hipotético tipo de relação concreto.

$$R^2 = \frac{\text{Variância de Y explicada pelo ajuste}}{\text{Variância total de Y}} = \frac{\sum\left(\hat{y}_i - \bar{y}\right)^2}{\sum\left(y_i - \bar{y}\right)^2} = \frac{\sum\left(\hat{y}_i - \bar{y}\right)^2}{\sum y_i^2 - n.\bar{y}^2}$$

$\hat{y}_i$ são os valores de Y *estimados* a partir dos correspondentes valores x_i pela função de ajuste encontrada. Portanto, no caso do ajuste linear, deverão ser calculados após m e **b** por: $\hat{y}_i = \mathbf{m} \, . \, x_i + \mathbf{b}$.

Excepcionalmente, apenas no caso da análise de um ajuste linear do tipo $\hat{Y} = m\,X + b$, $\mathbf{R^2}$ também vem dado por:

$$R^2 = r_{xy}^2 = (\, r_{xy}\,)^2$$

Onde r_{xy} é o coeficiente de correlação linear entre x e y, já definido anteriormente.

O coeficiente de determinação representa a proporção da variação de Y (multiplique-se por 100 e tem-se em percentagem) que é explicada pela relação linear com X. A parte restante não está explicada e, portanto fica devida a outros factores que não estão a ser controlados e/ou a uma má escolha do modelo de ajuste. Por vezes, estes são factores verdadeiramente aleatórios, mas frequentemente, sobretudo quando R^2 é muito diferente de 1, o que acontece é que:

- ou as variáveis não estão relacionadas – são independentes,
- ou há factores que deveriam ser considerados no estudo e que não estão a ser controlados,
- ou as variáveis possuem entre si um outro tipo de relação diferente do que está a ser considerado.

Exemplo 18 - No exemplo 15, temos que:

$R^2 = r_{xy}^2 = 0{,}9983^2 = 0{,}9966 \approx 99{,}7\ \%$

Portanto, 99,7% das variações de Y estão explicadas pela relação linear com X. Porém, 0,3 % das variações ficam a dever-se a outros factores, provavelmente aleatórios. Estamos perante um bom ajuste.

Exemplo 19 - Num estudo de mercado da relação entre o grau de álcool de um vinho (X) e o seu índice de vendas (Y), perante a tentativa de um ajuste linear do tipo Y = m X + b, chegou-se à conclusão de que r = 0,54.

Admitindo esse hipotético ajuste linear, qual a percentagem de valores de consumo que fica devida a outros factores que não o grau de álcool do vinho?

Sendo r = 0,54 , $R^2 = 0,54^2 = 0,2916 \approx 29\ \%$

Logo, cerca de 29% das variações no consumo, devem-se, de acordo com uma relação linear, ao grau de álcool. No entanto, há 100% - 29% = 71 % de variações que se devem a outros factores, que não estão a ser controlados neste estudo. Parte destes factores podem ser simplesmente factores aleatórios. No entanto, perante uma tão baixa percentagem, seria ainda de considerar:

- analisar a possibilidade da existência de um outro tipo de relação diferente da linear

- a hipótese de que alguns factores de vendas (como aroma, sabor , preço…) que não estão a ser considerados no estudo, possam influenciar também o índice de vendas.

4.6 Determinação do Coeficiente de Determinação em Excel

Há três formas diferentes de calcular o valor de R^2 em Excel:

- Directamente, pela fórmula de definição

Calcule os valores de Y previstos para todos os valores de X amostrais ; calcule e some os quadrados das diferenças entre estes e a média de Y, bem como os quadrados das diferenças entre os valores de Y observados e a sua média, e depois calcule R^2.

- Directamente, partindo do cálculo de r_{xy}

Para um ajuste linear, R^2 é igual a r_{xy} elevado a 2, pelo que desta forma o seu cálculo é muito simples.

- Adicionando ao Diagrama de Dispersão uma Linha de Tendência

Sobre o diagrama de dispersão, seleccione os pontos (basta clicar sobre um deles)

Seleccione o Menu ***Gráfico*** e o sub-menú ***Adicionar linha de tendência***.

Seleccione ***Linear*** e, em opções escolha "***Mostrar o valor de R ao quadrado no gráfico***". Prima OK.

A linha ajustada aparece desenhada e o valor de R^2 surge no gráfico.

Exemplo 20 – Veja Exemplo 20, no ficheiro Capítulo 4 Exemplos.xls

5. Caso especial: relação linear do tipo Y = b x

Em muitas situações, é esperada à partida uma relação linear do tipo Y = m . X ; ou seja é esperada uma **proporcionalidade directa** entre as duas variáveis.

Exemplo 21 - A relação entre uma quantidade de leite e a quantidade de cálcio nela contida é um caso de proporcionalidade directa.
A relação entre o valor monetário da factura da electricidade e o consumo de electricidade não é um caso de proporcionalidade directa.

5.1 Ajuste

Estamos agora a falar de uma relação linear especial, na qual a ordenada na origem é obrigatoriamente nula. Portanto, nestes casos há apenas que ajustar o valor do declive, **m**.

Tal como no caso geral da relação linear, podemos empregar o **Método a Sentimento**. Para tal, traça-se sobre o diagrama de dispersão a melhor recta, tendo sempre em atenção que ela deverá obrigatoriamente passar no ponto (0, 0), e de seguida determina-se o valor de **m** por um processo semelhante ao já apresentado (ver secção 2.3.3).

Se preferirmos usar o **Método dos Mínimos Quadrados**, este indica que, para efectuar este tipo de ajuste, o valor do declive vem dado pela seguinte expressão [2]:

$$m = \frac{\sum x_i \, . \, y_i}{\sum x_i^{\,2}}$$

Estes dois métodos apresentam as vantagens e os inconvenientes já anteriormente apresentados.

5.2 Qualidade do ajuste

O coeficiente de correlação amostral, r_{xy} , apenas caracteriza a existência ou não de uma relação linear do tipo Y = m X + b, pelo que neste caso terá pouca utilidade.

Já o **coeficiente de determinação**, $\mathbf{R^2}$, caracteriza qualquer tipo de ajuste. No caso de uma relação do tipo Y = m . X, a sua definição é:

$$R^2 = \frac{\text{Variância de Y explicada pelo ajuste}}{\text{Variância total de Y}} = \frac{\sum\left(\hat{y}_i - \bar{y}\right)^2}{\sum\left(y_i - \bar{y}\right)^2} = \frac{\sum\left(\hat{y}_i - \bar{y}\right)^2}{\sum y_i^{\,2} - n \, . \, \bar{y}^{\,2}}$$

No caso deste tipo de ajuste, prova-se que: $R^2 = 1 - \dfrac{\sum y_i^{\,2} - 2m \sum x_i \, . \, y_i + m^2 \sum x_i^{\,2}}{\sum y_i^{\,2} - n \, . \, \bar{y}}$

[2] A dedução desta expressão encontra-se fora do âmbito deste módulo de formação.

Tal como antes, $\hat{y}_i$ são os valores de Y *estimados* a partir dos correspondentes valores x_i pela função de ajuste encontrada. Portanto, no caso deste ajuste, deverão ser calculados após **b** por: $\hat{y}_i = \mathbf{m} \cdot x_i$

O seu significado é sempre o mesmo que já foi apresentado anteriormente: é a proporção de variações de Y que ficam explicadas pela relação assim estabelecida com X.

Exemplo 22 - *De acordo com a Lei de Beer, a absorvância é directamente proporcional à concentração, de acordo com: A = k . C .*
Com base nos seguintes dados, encontre um valor para k.

C:	*5,00E-05*	*1,00E-04*	*2,00E-04*	*3,00E-04*	*4,00E-04*	*5,00E-04*
A:	*0,102*	*0,196*	*0,400*	*0,570*	*0,724*	*0,867*

K é o declive de uma recta com ordenada na origem obrigatoriamente nula. Portanto devemos realizar um ajuste do tipo y = m . x, com y = A , x = C e m = k.

C.A.:	*x*	*y*	*x²*	*y²*	*xy*
	5,00E-05	0,102	2,500E-09	0,0104	5,100E-06
	1,00E-04	0,196	1,000E-08	0,0384	1,960E-05
	2,00E-04	0,400	4,000E-08	0,1600	8,000E-05
	3,00E-04	0,570	9,000E-08	0,3249	1,710E-04
	4,00E-04	0,724	1,600E-07	0,5242	2,896E-04
	5,00E-04	0,867	2,500E-07	0,7517	4,335E-04
SOMA:	1,55E-03	2,859	5,525E-07	1,8096	9,988E-04

$$k = m = \frac{\sum x_i \cdot y_i}{\sum x_i^2} = \frac{9{,}988 \times 10^{-4}}{5{,}525 \times 10^{-7}} = 1808$$

Portanto: $A = 1808 \times C$.

Qualidade do ajuste:

$$R^2 = 1 - \frac{\sum y_i^2 - 2m \sum x_i \cdot y_i + m^2 \sum x_i^2}{\sum y_i^2 - n \cdot \bar{y}}$$

$$= 1 - \frac{1{,}8096 - 2 \times 1808 \times 9{,}988 \times 10^{-4} + 1808^2 \times 5{,}525 \times 10^{-7}}{1{,}8096 - 6 \times 0{,}4765^2} = 0{,}991$$

Podemos concluir que 99,1 % das variações da absorvância são explicadas pela directa proporcionalidade com a concentração, com constante igual a 1808.

5.3 Cálculos em Excel

Podemos encontrar comodamente a equação de ajuste, bem como o coeficiente de determinação deste tipo de ajuste, em Excel, do seguinte modo :

- Adicionando ao Diagrama de Dispersão uma Linha de Tendência

Sobre o diagrama de dispersão, seleccione os pontos (basta clicar sobre um deles)

Seleccione o Menu ***Gráfico*** e o sub-menú ***Adicionar linha de tendência***.

Seleccione ***Linear*** e, em opções escolha "***Definir a intersecção em 0***" , "***Mostrar a equação no gráfico***" e ainda "***Mostrar o valor de R ao quadrado no gráfico***". Prima OK.

A linha ajustada aparece desenhada e o valor de R^2 surge no gráfico.

Exemplo 23 – Veja Exemplo 23, no ficheiro Capítulo 4 Exemplos.xls

6. Regressões não lineares

Em muitas situações a dependência ou relação entre duas variáveis não assume uma forma linear, não significando isso que não exista relação. Pode haver (e muitos casos há) em que as relações assumem outras formas, outras funções matemáticas, que não a linear.

Nesta disciplina, não iremos estudar aprofundadamente a explicação de ajustes não lineares. Pretende-se porém que o Aluno seja capaz de os encontrar, utilizando o Excel, ou outra ferramenta compatível e até algumas máquinas calculadoras.

6.1 Alguns tipos de ajustes não lineares

Existem muitos tipos de ajustes não lineares (tantos quantos os formatos de funções matemáticas entre duas variáveis que consiga idealizar). Alguns dos tipos de ajustes mais comuns são:

Ajuste polinomial de grau n (n = 2 , 3 , 4 , …):

$$\hat{Y} = a_n \cdot X^n + a_{n-1} \cdot X^{n-1} + \ldots + a_2 \cdot X^2 + a_1 X + a_0$$

Ajuste logarítmico:

$$\hat{Y} = a \cdot \log (X) - b$$

Ajuste potência:

$$\hat{Y} = a \cdot X^b$$

Ajuste exponencial:

$$\hat{Y} = a \cdot e^{b \cdot X}$$

6.2 Determinação de ajustes não lineares em Excel

É possível determinar os ajustes não lineares atrás mencionados, adicionando ao diagrama de dispersão uma linha de tendência :

Sobre o diagrama de dispersão, seleccione os pontos (basta clicar sobre um deles)

Seleccione o Menu ***Gráfico*** e o sub-menú ***Adicionar linha de tendência***.

Seleccione ***<Ajuste pretendido>*** e, em opções escolha "***Mostrar a equação no gráfico***". (Caso seleccione um ajuste polinomial, indique também o grau do polinómio que pretende).

Prima OK.

A linha ajustada aparece desenhada e a sua equação característica surge no gráfico.

6.3 Caracterização da qualidade do ajuste

Como já foi referido, o **coeficiente de determinação**, $\mathbf{R^2}$, caracteriza qualquer tipo de ajuste. Para qualquer relação, a sua definição é:

$$R^2 = \frac{\text{Variância de Y explicada pelo ajuste}}{\text{Variância total de Y}} = \frac{\sum\left(\hat{y}_i - \bar{y}\right)^2}{\sum\left(y_i - \bar{y}\right)^2} = \frac{\sum\left(\hat{y}_i - \bar{y}\right)^2}{\sum y_i^2 - n.\bar{y}^2}$$

Decorre da sua definição que, quanto mais próximo o valor de R^2 for de 1, melhor o ajuste explica os dados.

Podemos encontrar comodamente o valor do coeficiente de determinação deste tipo de ajuste, em Excel, em simultâneo com a equação de ajuste, do seguinte modo :

- Quando está a adicionar ao diagrama de dispersão uma linha de tendência

Em opções escolha "***Mostrar o valor de R ao quadrado no gráfico***".

Ao mesmo tempo que a linha ajustada aparece desenhada, o valor de R^2 surge no gráfico.

- FIM do Capítulo –