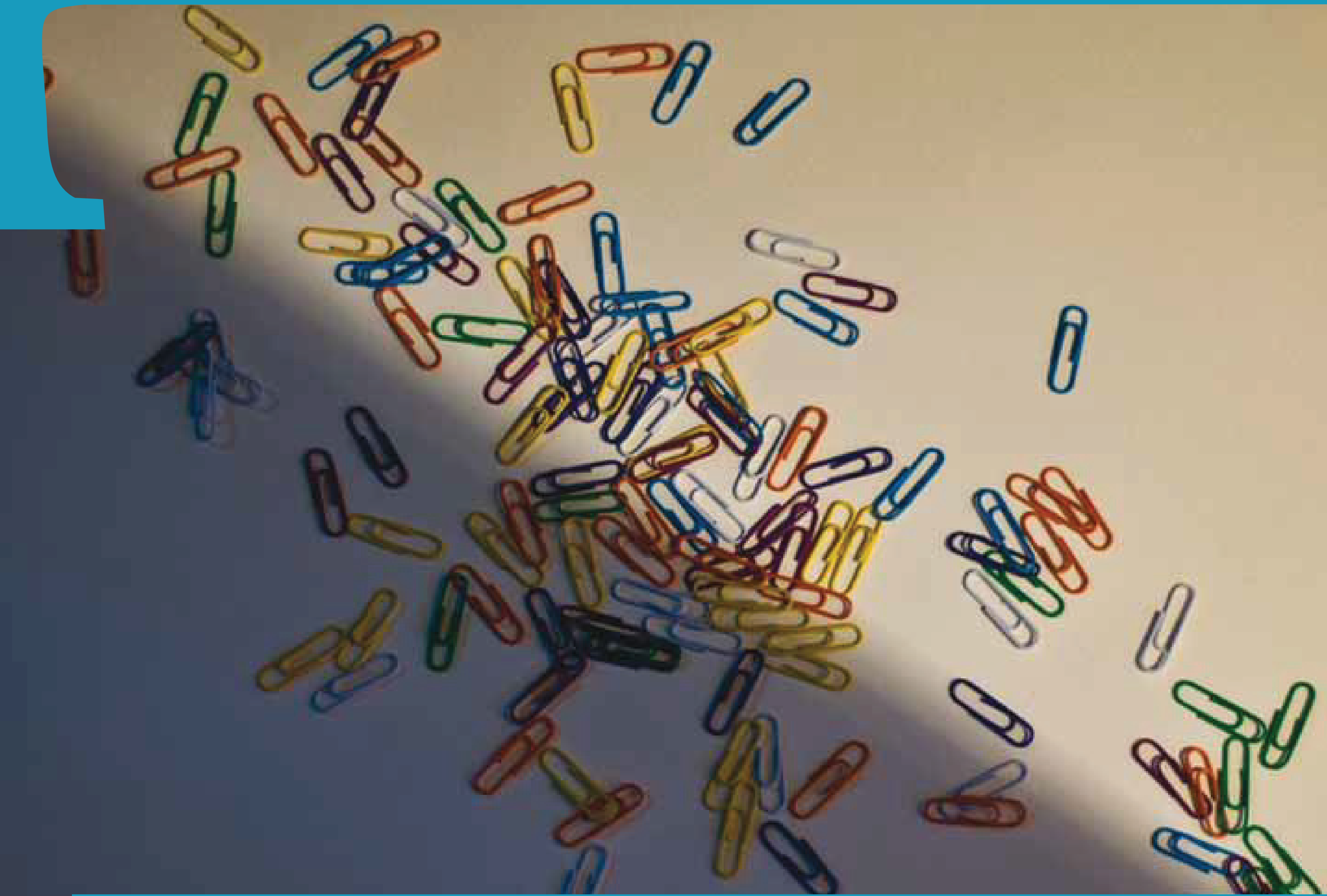

A Regressão Linear Simples no Ensino Secundário

Helena Ribeiro[a], Maria Alice Martins[b] e Rui Santos[c]

[a] helena.ribeiro@ipleiria.pt, Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria

[b] martinsalice69@gmail.com, Agrupamento de Escolas Artur Gonçalves de Torres Novas

[c] rui.santos@ipleiria.pt, Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria

Neste artigo iremos focar a atenção no erro mais comum que detetamos em manuais por nós consultados, o da utilização da mesma reta de regressão, obtida pelo método dos mínimos quadrados, para estimar um valor de x condicionado a um dado valor de y bem como para estimar um valor de y condicionado a um valor de x quando, corretamente, dever-se-iam utilizar duas retas distintas (exceto em alguns casos muito particulares onde as duas retas são análogas). Este erro será exemplificado utilizando um *software* (que é *freeware*) frequentemente utilizado no ensino da geometria no ensino básico e secundário, o *GeoGebra*, e que é uma potencial ferramenta no ensino da regressão linear, bem como de outros temas da Estatística.

Apesar de a Estatística desempenhar, cada vez mais, um lugar de destaque na disciplina de Matemática, quer ao nível do ensino básico quer ao nível do ensino secundário, os materiais disponíveis para o ensino e a compreensão dos conceitos mais elementares desta área nem sempre são os mais apropriados (apesar de contributos extremamente bem sucedidos pela literacia estatística, como ilustra o premiado projeto ALEA – Acção Local de Estatística Aplicada – que pode ser consultado em *http://www.alea.pt*).

1. REGRESSÃO LINEAR NO PROGRAMA DO ENSINO SECUNDÁRIO

No programa de Matemática A do 10.º ano (ensino secundário) a terceira unidade a ser lecionada é a Estatística (cf. Silva *et al.*, 2001). Nesta fase os alunos já possuem algumas noções que adquiriram no 3.º ciclo do ensino básico e já realizaram pequenos trabalhos, no entanto, é a primeira vez que se fará referência a distribuições bidimensionais, sendo efetuada uma abordagem gráfica e intuitiva. Deste modo, os tópicos que fazem parte do programa são:

- Diagrama de dispersão, explorando o conceito de dependência estatística;
- Ideia intuitiva de correlação: exemplos gráficos de correlação positiva, negativa ou nula;
- Coeficiente de correlação e sua variação;
- Definição de centro de gravidade de um conjunto finito de pontos, sua interpretação física;
- Ideia intuitiva de reta de regressão, sua interpretação e limitações.

O programa sugere as seguintes indicações metodológicas: introdução ao estudo dos dados bivariados, generalizando o estudo de uma única variável e destacando a representação gráfica sob a forma de diagrama de dispersão ou diagrama de pontos. Com base nesta representação, pretende-se identificar a existência de uma associação linear entre as duas variáveis em estudo, utilizando uma medida que quantifica o grau de associação – o coeficiente de correlação, assim como apresentando um modelo matemático que permitirá, conhecido o valor de uma das variáveis, obter uma estimativa para o valor da outra variável. Os recursos propostos a utilizar no ensino desta temática são:

- "Papel e lápis" para a construção do diagrama de dispersão (nuvens de pontos);
- Calculadora gráfica ou computador para visualizar o diagrama de dispersão, desenhar a reta de regressão, calcular o coeficiente de correlação e encontrar o modelo matemático (função polinomial de grau 1) que caracterize a relação entre as duas variáveis.

Desta forma, no 10.º ano de escolaridade é transmitida uma ideia intuitiva de reta de regressão, tentando explorar a sua interpretação e as suas limitações. Apesar de não ser objetivo deste nível de ensino explicar formalmente a reta obtida, é transmitida a ideia pela qual ela é determinada – corresponde à reta que faz com que a soma dos quadrados das distâncias de cada ponto da nuvem à reta seja mínima **(método dos mínimos quadrados**), sendo esta reta unicamente determinada recorrendo a uma calculadora.

2. O MÉTODO DOS MÍNIMOS QUADRADOS

Consideremos um conjunto de dados (x_i, y_i), com $i = 1, \ldots, n$, para o qual se pretende ajustar a reta

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{1}$$

onde y_i representa a variável dependente ou endógena (que o modelo pretende explicar o comportamento), x_i a variável independente ou exógena (que será uma variável explicativa para a modelação de y_i), ε_i a variável erro (que é uma variável aleatória com algumas características fundamentais para a fiabilidade da inferência estatística associada à regressão) e β_0 e β_1 os parâmetros da regressão. A reta estimada será da forma

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \tag{2}$$

onde $\widehat{y}_i$, $\widehat{\beta}_0$ e $\widehat{\beta}_1$ representam, respetivamente, os estimadores (ou estimativas, para as quais, ao longo deste texto, iremos utilizar igual representação) de $\widehat{y}_i$, $\widehat{\beta}_0$ e $\widehat{\beta}_1$. Desta forma, os erros ε_i correspondem à diferença entre os valores observados para y_i e os valores estimados, *i.e.*

$$\varepsilon_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i. \tag{3}$$

O método habitualmente utilizado para estimar os parâmetros da reta é o **método dos mínimos quadrados** que determina os estimadores $\widehat{\beta}_0$ e e $\widehat{\beta}_1$ que minimizam a soma dos quadrados dos erros, *i.e.* que minimizam

$$f\left(\widehat{\beta}_0, \widehat{\beta}_1\right) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2. \tag{4}$$

Para determinar o mínimo de $f\left(\widehat{\beta}_0, \widehat{\beta}_1\right)$ teremos

$$\begin{cases} \frac{\partial f(\widehat{\beta}_0, \widehat{\beta}_1)}{\partial \widehat{\beta}_0} & = & 0 \\ \frac{\partial f(\widehat{\beta}_0, \widehat{\beta}_1)}{\partial \widehat{\beta}_1} & = & 0 \end{cases} \Leftrightarrow$$

$$\Leftrightarrow \begin{cases} \widehat{\beta}_0 & = & \overline{y} - \widehat{\beta}_1 \overline{x} \\ \widehat{\beta}_1 & = & \dfrac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2} = \dfrac{s_{xy}}{s_x^2} = \dfrac{s_y}{s_x} r_{xy} \end{cases}, \tag{5}$$

onde:

- $\overline{y}$ e $\overline{x}$ representam as médias de y e x;
- s_y e s_x representam os desvios padrão de y e x;
- s_{xy} representa a covariância entre x e y;
- r_{xy} o coeficiente de correlação entre x e y.

O determinante da matriz Hessiana é igual a $4n\sum_{i=1}^{n} x_i^2 - 4\left(\sum_{i=1}^{n} x_i\right)^2 = 4n^2 s_x^2$ que é positivo (desde que x_i não assuma sempre o mesmo valor). Assim sendo, a matriz Hessiana é definida positiva e os valores determinados em (5) correspondem efetivamente ao mínimo pretendido.

Desta forma, nas equações presentes em (5) temos as fórmulas dos estimadores dos mínimos quadrados de β_0 e β_1 com os quais podemos, recorrendo à reta de regressão (2), obter estimativas para y conhecendo um valor específico de x. As propriedades da inferência estatística resultante desta aplicação dependem das características dos resíduos (variável ε). Contudo, uma vez que esta abordagem não é efetuada no ensino secundário, sublinhamos apenas que os estimadores assim obtidos gozam de excelentes propriedades (não enviesamento, eficiência e consistência) caso a variável ε satisfaça determinadas características (nomeadamente, normalidade, independência e homocedasticidade). Para mais informações acerca das propriedades dos estimadores, pode ser consultado, por exemplo, Montgomery *et al.* (2006).

Algumas propriedades da regressão são exploradas no ensino secundário, como, por exemplo, o facto de a **soma dos erros ser nula**, $\sum_{i=1}^{n} \varepsilon_i = 0$, pois

$$\sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right) = n\left(\overline{y} - \widehat{\beta}_0 - \widehat{\beta}_1 \overline{x}\right) = 0, \tag{6}$$

tendo em conta que $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$ (cf. equação (5)), e a reta estimada **passar sempre pelo centro de gravidade dos dados** $(\overline{x}, \overline{y})$ uma vez que, utilizando (2) e (5), se obtém

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{x} = \overline{y} - \widehat{\beta}_1 \overline{x} + \widehat{\beta}_1 \overline{x} = \overline{y}, \tag{7}$$

logo, quando $x = \overline{x}$, teremos $y = \overline{y}$. Estas e outras propriedades da regressão linear simples podem ser analisadas em Murteira (1993), Reis (1998), Martins (2005), Pestana e Velosa (2009), entre outros.

Contudo, salientemos que a reta obtida pelo método dos mínimos quadrados para estimar y em função de x e a reta obtida para estimar x em função de y **não são, em geral, idênticas.**

3. A REGRESSÃO LINEAR INVERSA

Podemos utilizar fórmulas análogas às apresentadas em (5) para efetuarmos a regressão de x em função de y obtendo

$$\widehat{x}_i = \widehat{\alpha}_0 + \widehat{\alpha}_1 y_i, \tag{8}$$

onde $\widehat{\alpha}_0 = \overline{x} - \widehat{\alpha}_1 \overline{y}$ e $\widehat{\alpha}_1 = \frac{s_x}{s_y} r_{xy}$. Invertendo a reta (i.e. resolvendo em função de y_i e trocando os papéis desempenhados pelas variáveis), obtemos

$$\widetilde{y}_i = \widetilde{\beta}_0 + \widetilde{\beta}_1 x_i. \tag{9}$$

Os estimadores de β_0 e β_1 assim obtidos, supondo $r_{xy} \neq 0$ (pois caso $r_{xy} \approx 0$ a reta de regressão não terá qualquer sentido), serão dados por:

$$\widetilde{\beta}_0 = -\frac{\widehat{\alpha}_0}{\widehat{\alpha}_1} = -\frac{\overline{x} - \widehat{\alpha}_1 \overline{y}}{\widehat{\alpha}_1} = \overline{y} - \widetilde{\beta}_1 \overline{x} \;\text{ e }\; \widetilde{\beta}_1 = \frac{1}{\widehat{\alpha}_1} = \frac{s_y}{s_x} r_{xy}^{-1}. \tag{10}$$

As retas (2) e (9) são coincidentes se e só se $\widehat{\beta}_0 = \widetilde{\beta}_0$ e $\widehat{\beta}_1 = \widetilde{\beta}_1$ que, recorrendo às equações (5) e (10) só se verifica quando

$$r_{xy} = 1 \quad \vee \quad r_{xy} = -1. \tag{11}$$

Assim sendo, ambas as retas passam pelo mesmo ponto $(\overline{x}, \overline{y})$, mas só se obtém a mesma reta (retas coincidentes) utilizando os dois métodos se o módulo da correlação for unitário (sendo os erros, nestes casos, todos nulos, uma vez que a reta passa precisamente por todos os pontos). Deste modo, a reta de regressão de y em função de x, determinando os parâmetros da reta $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$ que minimizam $\sum_i (y_i - \widehat{y}_i)^2$, será distinta (exceto se $\left| r_{xy} \right| = 1$) da reta obtida quando efetuamos uma regressão de x em função de y, determinando os parâmetros da reta $\widehat{x}_i = \widehat{\alpha}_0 + \widehat{\alpha}_1 y_i$ que minimizam $\sum_i (x_i - \widehat{x}_i)^2$. Esta diferença resulta da forma como definimos os erros nas duas regressões, pois enquanto na primeira os erros são medidos paralelamente ao eixo das ordenadas (o erro é definido pela diferença entre o valor observado de y e o seu valor estimado condicionalmente a x, $\varepsilon_i = y_i - \widehat{y}_i$, cf., ilustra o primeiro gráfico da figura 1), na segunda os erros são medidos paralelamente ao eixo das abcissas (o erro é definido pela diferença entre o valor observado x e o seu valor estimado pela regressão em função de y, $\varepsilon_i = x_i - \widehat{x}_i$, cf. segundo gráfico da figura 1). Desta forma, será erróneo utilizar a regressão de y em função de x para efetuar previsões para x quando conhecemos um determinado valor para y e, apesar de em algumas aplicações a diferença das duas retas poder ser diminuta, existem outras situações em que o erro pode assumir valores elevados. Há, contudo, determinadas situações específicas para as quais se justifica a necessidade de utilização de regressão inversa, como ilustram alguns modelos de calibração (*cf.* Brown, 1993).

Os modelos de calibração estatística (ou, por vezes, denominados por previsão inversa) são muito utilizados em química, engenharia, bioestatística e podem ser extremamente úteis em algumas aplicações. Contudo, estas aplicações são caracterizadas por contextos distintos dos que previamente referimos. Consideremos, então, que a variável y é medida através de processo complicado (muitas vezes fora do nosso alcance), dispendioso (em termos de tempo e/ou monetariamente) mas que os seus resultados são extremamente precisos (sem erros ou com erros negligenciáveis em relação aos erros de medição da variável x). A variável x é medida por um processo simples, rápido, e económico mas que tem pouca precisão (obtemos valores aproximados). O objetivo é estimar novos valores de y para alguns valores de x conhecidos. Neste contexto, os erros a ter em conta no método de regres-

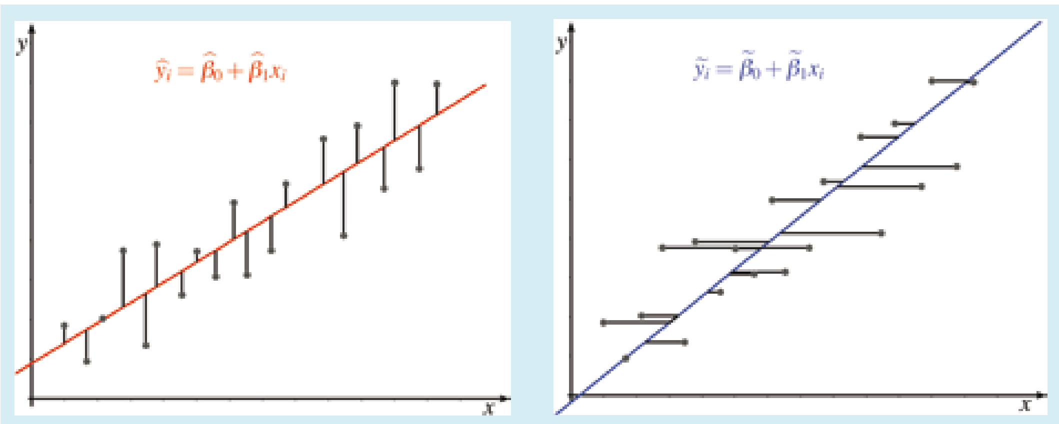

Figura 1: Definição dos erros na regressão linear

são linear deverão ser $\varepsilon_i = x_i - \widehat{x}_i$ (pois $y_i - \widehat{y}_i \approx 0$) estimando-se a regressão $\widetilde{y}_i = \widetilde{\beta}_0 + \widetilde{\beta}_1 x_i$ utilizando os estimadores apresentados na equação (10). Com esta regressão podemos estimar novos valores para y condicionados a valores conhecidos de x. Os estimadores assim obtidos, sob determinadas condições, gozam igualmente de boas propriedades. Osborne (1991) faz uma apresentação histórica da evolução destes métodos, incluindo alguma discussão sobre a problemática inerente à sua utilização.

Contudo, estes contextos específicos nos quais a utilização da calibração é frequentemente utilizada, não são abordados no ensino secundário.

4. ESTIMAÇÃO DE *Y* CONDICIONADO A *X* = *X₀*

Consideremos que se pretende estimar o valor de y quando x assume o valor x_0. Recorrendo à equação (2), regressão de y em função de x, obtemos

$$\begin{aligned}\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_0 &= \overline{y} - \frac{s_y}{s_x} r_{xy} \overline{x} + \frac{s_y}{s_x} r_{xy} x_0 \\ &= \overline{y} + \frac{s_y}{s_x} r_{xy} \left(x_0 - \overline{x}\right), \qquad (12)\end{aligned}$$

e recorrendo à equação (9), função inversa da regressão de x em função de y, obtemos

$$\begin{aligned}\widetilde{y} = \widetilde{\beta}_0 + \widetilde{\beta}_1 x_0 &= \overline{y} - \frac{s_y}{s_x} r_{xy}^{-1} \overline{x} + \frac{s_y}{s_x} r_{xy}^{-1} x_0 \\ &= \overline{y} + \frac{s_y}{s_x} r_{xy}^{-1} \left(x_0 - \overline{x}\right). \qquad (13)\end{aligned}$$

A distância entre as duas estimativas obtidas é dada por

$$|\widehat{y} - \widetilde{y}| = \frac{s_y}{s_x} |x_0 - \overline{x}| \left| r_{xy} - r_{xy}^{-1} \right|, \qquad (14)$$

que depende do quociente entre o desvio padrão de y e o de x, da distância entre x_0 e $\overline{x}$ [o que era espectável uma vez que ambas as retas passam no ponto $(\overline{x}, \overline{y})$] e da distância entre r_{xy} e r_{xy}^{-1} (que será nula se $|r_{xy}| = 1$, isto é, quando a relação linear é perfeita e as duas retas de regressão são coincidentes).

Conforme claramente ilustram os quatro exemplos retratados na figura 2, onde são apresentadas as duas retas obtidas utilizando conjuntos distintos de 200 observações (recorrendo ao *software GeoGebra*), com coeficiente de correlação igual a 0.95, 0.90, 0.80 e 0.70, podemos constatar a distinção entre as duas retas, a diferença entre o valor estimado de y obtido pelas duas retas quando x assume o valor x_0, bem como o aumento desta distância com a diminuição da correlação entre x e y e/ou o afastamento de x_0 relativamente a $\overline{x}$.

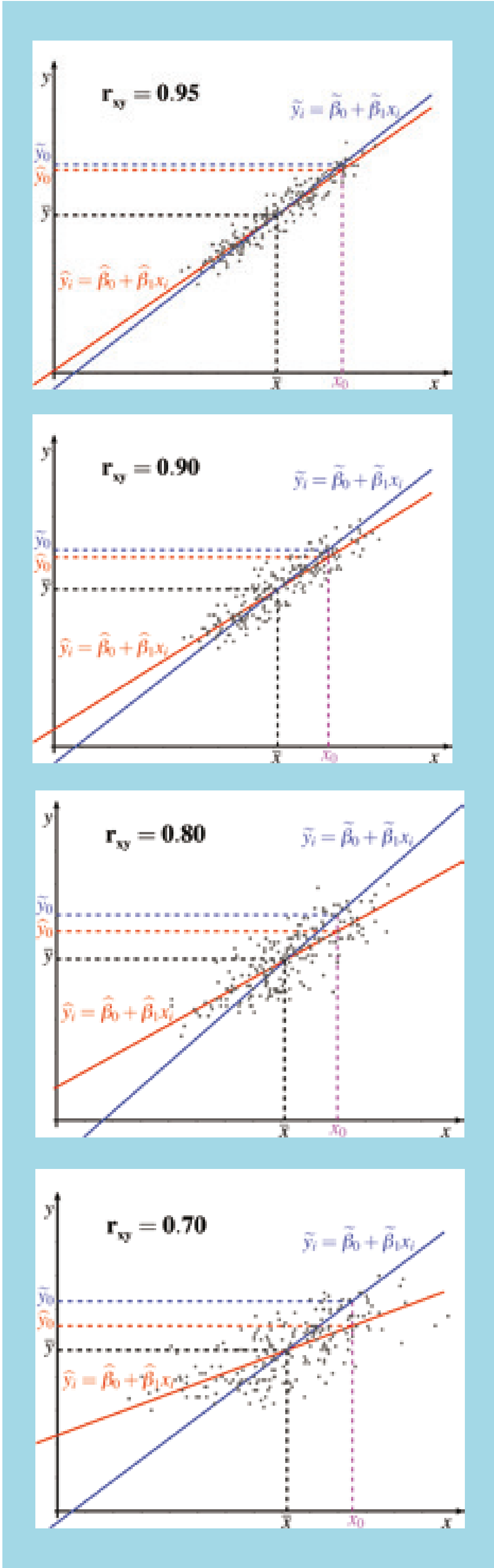

Figura 2: Regressão de *y* em função de *x* *versus* regressão de *x* em função de *y*

5. O *GEOGEBRA* NO ENSINO DA REGRESSÃO LINEAR

Gráficos análogos aos representados na figura 2 podem ser construídos em sala de aula com recurso ao *software GeoGebra*, tendo a vantagem de poderem ser dinâmicos, isto é, ao alterarmos um ponto (ou um conjunto de pontos) visualizarmos imediatamente as consequentes alterações no coeficiente de correlação e nas retas estimadas, bem como ao mudarmos a coordenada x_0 percebermos as consequentes implicações no valor de y estimado. Para este fim, no *GeoGebra*, podemos exibir a Folha de Cálculo e utilizar as duas primeiras colunas (A e B) para definir as coordenadas dos pontos a utilizar para a regressão linear. Com os pontos definidos podemos criar uma lista de pontos (bastará selecionar as coordenadas e, utilizando o botão do lado direito do rato, selecionar "Criar lista de pontos"). Desta forma será criada a lista1 (nome atribuído automaticamente pelo *GeoGebra*, mas que pode ser facilmente alterado nas propriedades da lista de pontos) que contém os n pontos que correspondem ao nosso conjunto de dados (x_i, y_i). Podemos de seguida determinar o coeficiente de correlação, através do comando CoeficienteCorrelação[lista1]; a reta de regressão de y em função de x de acordo com a equação (2), recorrendo ao comando RetadeRegressão[lista1] e, no mesmo gráfico, representar a reta de regressão de x em função de y [cf. equação (9)] utilizando o comando RegressãoLinearX[lista1]. As duas retas podem ser representadas por cores diferenciadas de forma a podermos distingui-las facilmente. Assim, ao ser alterado um ou mais pontos da lista1, quer o coeficiente de correlação quer as retas serão automaticamente ajustadas à nova nuvem de pontos, o que permitirá visualizar facilmente a diferença entre as duas retas, que a sua interseção se verifica sempre no ponto $(\overline{x}, \overline{y})$, bem como observar as alterações no coeficiente de correlação. Em sala de aula, pode explorar-se a sensibilidade do coeficiente de correlação, bem como da reta ajustada, a alterações das coordenadas dos pontos, em particular à existência de pontos mais afastados (*outliers* e/ou observações influentes). Caso pretendamos analisar igualmente os valores estimados para y para um dado valor de x, bastará definir um ponto A [de coordenadas $(x_0, 0)$] no eixo das abcissas e construir os segmentos de reta paralelos ao eixo das ordenadas que liguem o ponto A a cada uma das retas estimadas (sejam P_1 e P_2 os pontos das retas com abcissa x_0). As ordenadas dos pontos P_1 e P_2 correspondem às estimativas obtidas com cada uma das retas (valores $\widehat{y}_0$ e $\widetilde{y}_0$ representados nos gráficos da figura 2). Para melhor visualizar o valor das estimativas obtidas podemos criar segmentos de retas paralelos ao eixos das abcissas que liguem P_1 e P_2 ao eixo das ordenadas (de forma análoga à apresentada nos gráficos da figura 2). Com esta construção, caso alteremos a abcissa x_0 do ponto A, as estimativas obtidas serão automaticamente ajustadas, permitindo, desta forma, observar as suas diferenças quando utilizamos valores para x_0 mais próximos ou mais afastados de $\overline{x}$.

De referir que, no ensino da Estatística, a utilidade do *GeoGebra* não se restringe à regressão linear, pois este *software* pode igualmente ser utilizado no ensino de outros temas da Estatística, uma vez que permite determinar a média, mediana, moda, variância, desvio padrão e os quartis, bem como construir tabelas de frequências e representar alguns gráficos, tais como o diagrama de caule e folhas, o gráfico de barras, o gráfico de pontos, o histograma e o diagrama de extremos e quartis. Por este motivo, o *GeoGebra* é uma ferramenta que pode perfeitamente ser utilizada com sucesso no ensino dos conceitos elementares de Estatística, quer no ensino secundário quer no ensino básico. Por outro lado, o facto de este *software* possibilitar efetuar uma construção dinâmica, tais construções permitem a exploração da sensibilidade de algumas medidas (como, por exemplo, das medidas de tendência central ou das medidas de dispersão) a variações do valor de uma ou mais observações.

6. MANUAIS ADOPTADOS NO ENSINO SECUNDÁRIO

Em diversos manuais adotados no ensino secundário é utilizada a mesma reta de regressão para estimar o valor de y quando conhecemos um valor de x (condicionado a $x = x_0$) e para estimar o valor de x em função de um valor específico da variável y ($y = y_0$), o que não deveria ocorrer pelas razões previamente apontadas (não iremos incluir estes manuais nas referências bibliográficas uma vez que é nosso objetivo contactar diretamente com os autores desses manuais). Esta confusão parece derivar do desconhecimento de que as retas são distintas. Sublinhemos que, dos manuais por nós consultados, há um que apresenta corretamente este tema e, partindo de um exemplo simples, explica aos alunos a razão pela qual não devem usar a mesma reta de regressão em ambos os casos, referindo:

"A recta de regressão, de equação $y = 0,797x + 121,282$, foi construída para, dado o peso x, em kg, de um jogador prever a altura y, em cm, do mesmo (. . .). Os erros cometidos, relativamente aos valores de y medidos e os valores previstos são os comprimentos dos segmentos de recta assinalados na figura (. . .). A equação da recta de regressão é determinada, utilizando ferramentas matemáticas, de forma a minimizar a acumulação destes comprimentos (...). Percebe-se assim, que é possível utilizar a equação desta recta para prever valores da altura dado o valor do peso, mas, no entanto não se pode utilizar esta equação para prever o valor do peso dado o valor da altura"

[Negra & Martinho, 2010, pp.181-182]

Mesmo quando no desenvolvimento teórico nada é referido acerca da distinção entre estas duas retas, diversos manuais apresentam exercícios para estimar um valor de x condicionado a um valor de $y = y_0$ utilizando a regressão de y em função de x como se as duas regressões fornecessem o mesmo resultado (de facto, tudo leva a crer que alguns dos autores de manuais escolares desconheçam as diferenças entre as duas retas obtidas). É, por este motivo, igualmente do desconhecimento de alguns professores do ensino secundário a existência das referidas diferenças, razão pela qual é importante divulgá-las de forma a serem efetuadas as correções necessárias nos manuais, bem como alertar e clarificar o corpo docente para esta situação.

Por outro lado, pretendemos igualmente salientar que a exploração de exemplos dinâmicos em *GeoGebra*, *software* já utilizado por muitos docentes em diversos contextos da Matemática, torna intuitiva a diferença entre os dois métodos. O cálculo das duas retas é automático e, alterando os valores da nossa amostra (presentes na folha de cálculo do *GeoGebra*), permite visualizar a consequente alteração (sensibilidade) das retas, bem como dos valores estimados pelos dois modelos, razão pela qual nos parece uma ferramenta adequada para o ensino da regressão linear, permitindo atingir com sucesso todos os objetivos estipulados no programa de Matemática do ensino secundário.

Agradecimentos:

Este trabalho foi financiado por Fundos Nacionais através da FCT — Fundação para a Ciência e a Tecnologia, no âmbito do projeto PEst- OE/MAT/UI0006/2011 e pelo Instituto Politécnico de Leiria.

REFERÊNCIAS

Brown, P.J. (1993). *Measurement, Regression, and Calibration*, Oxford University Press.

Martins, M.E.G. (2005). *Introdução à Probabilidade e à Estatística*, Sociedade Portuguesa de Estatística, Lisboa.

Montgomery, D.C., Peck, E.A. and Vining, G.G. (2006). *Introduction to Linear Regression Analysis*, 4th Ed., Wiley Series in Probability and Statistics, John Wiley & Sons.

Murteira, B. (1993). *Análise exploratória de dados - Estatística Descritiva*, McGraw-Hill, Lisboa.

Negra, C. e Martinho, E. (2010). *Matemática A – 10.º ano Volume 2*, Santillana, Carnaxide.

Osborne, C. (1991). "Statistical Calibration: A Review", *International Statistical Review* 59, n.º 3, pp. 309–336.

Pestana, D.D. e Velosa, S.F. (2009). *Introdução à Probabilidade e à Estatística*, Vol. 1, 3.ª ed., Fundação Calouste Gulbenkian, Lisboa.

Reis, E. (1998). *Estatística Descritiva*, Edições Sílabo, Lisboa.

Silva, J.C., Fonseca, M.G., Martins, A.A., Fonseca, C.M.C. e Lopes, I.M.C. (2001). *Programa de Matemática A - Ensino Secundário*, ME-DES, Lisboa.

PÁGINAS DA INTERNET

ALEA — *www.alea.pt*

SOBRE OS AUTORES

Alice Martins é licenciada em Matemática (ramo educacional) pela FCT-UC e mestre em Educação e Tecnologia em Matemática no IPLeiria. É docente de Matemática no Agrupamento de Escolas Artur Gonçalves de Torres Novas.

Helena Ribeiro é licenciada em Matemática (ramo educacional) pela FCT-UC e doutorada em Matemática pelo IST-UTL. É docente no Departamento de Matemática da Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria e membro do CEMAT — Centro de Matemática e Aplicações.

Rui Santos é licenciado em Matemática Aplicada à Economia e Gestão pelo ISEG-UTL e doutorado em Probabilidades e Estatística pela FCUL. É docente no Departamento de Matemática da Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria e membro do CEAUL — Centro de Estatística e Aplicações da Universidade de Lisboa.