Apologia da Estatística
(A Pretexto da Seriação das Escolas Secundárias)

Dinis Duarte Pestana
Centro de Estatística e Aplicações da Universidade de Lisboa

"Some people hate the very name of statistics, but I find them full of beauty and interest. Whenever they are not brutalized, but delicately handled by the higher methods, and are warily interpreted, their power of dealing with complicated phenomena is extraordinary. They are the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the Science of man."

F. Galton, *Natural Inheritance*

1. Qualidade do Ensino e Seriação das Escolas

If an unfriendly foreign power had attempted to impose on America the mediocre educational performance that exists today, we might well have viewed it as an act of war.

The National Comission on Excellence in Education

É hoje lugar comum considerar que a maior riqueza de um país é a sua população, e que o empreendimento humano de maior retorno é educação/investigação. Os países que escaparam ao subdesenvolvimento são os que nos fins do século XIX investiram na alfabetização e educação. Tal como o presente é um resultado de investimentos em fins do século XIX e princípios do século XX, é no presente que o futuro se joga, nomeadamente nos investimentos feitos em ciência e tecnologia (que, sabe-se também, são efémeros se não forem alicerçados por um investimento forte e consistente em ciência fundamental).

A qualidade do ensino é, naturalmente, uma das fundações sobre que se constrói o futuro. Há alguns anos atrás o Japão justificou as suas políticas de comércio internacional, e nomeadamente a baixa taxa de importação de bens dos EUA, referindo a sua deficiente qualidade, e invocando que a escolaridade dos alunos dos Estados Unidos é muito inferior à dos alunos do Japão, bem como os níveis de exigência em educação, do que decorre que a atitude dos profissionais japoneses face a políticas de controlo de qualidade global é mais consciente e responsável. Claro que o Japão nunca se deu ao trabalho de comentar publicamente o que pensa da educação em Portugal, mas podemos imaginar…

Os alunos - e os pais dos alunos - portugueses parecem estranhamente adormecidos para as realidades de um espaço europeu, em que há mobilidade e o preenchimento dos lugares não irá privilegiar o local do nascimento. Ocasionalmente os jornais exploram os aspectos mediáticos de situações tais como professores espanhóis serem docentes em escolas primárias portuguesas, apelando mais à rejeição (ilegal, face aos tratados que assinámos e ratificámos) do que ao despertar para a realidade que espera as gerações futuras: uma competição feroz com profissionais de uma União Europeia cada vez mais alargada, com sistemas educativos mais consolidados e controlados, e em que o desemprego é uma realidade persistente.

Saudamos por isso naturalmente todas as iniciativas tendentes a melhorar o estado das coisas. Recentemente o Ministério da Educação encomendou à Universidade Nova de Lisboa - Faculdade de Ciências Sociais e Humanas uma *Proposta de Seriação das Escolas do Ensino Secundário (Ano Lectivo de 2001/2002),* elaborada por Grácio *et al*. (2002), que se inscreve neste propósito, e que teve larga divulgação mediática. As conclusões do estudo dificilmente poderiam gerar consenso, e Grácio *et al*. (2002, p. 12-13) são os primeiros a anotar alguns aspectos controversos, decorrentes das limitações no acesso à informação, embora concluam afirmando (Grácio *et al.*, 2002, p. 13) que *"o método adoptado é mil vezes preferível a fornecer simplesmente ao público os dados em bruto, convidando à sua organização pelo seu valor facial, e à consequente imagem profundamente distorcida e injusta do trabalho das escolas e dos seus profissionais"*.

Alguns comentadores, distanciando-se naturalmente da metodologia adoptada e das conclusões, que consideraram provisórias, não deixaram de referir o salto qualitativo que é começar a estudar os problemas com metodologias científicas. Nuno Crato, identificado num telejornal como representante da Sociedade Portuguesa de Matemática, ao ser questionado sobre a oportunidade da divulgação das conclusões do estudo, voltou a pergunta ao contrário: "E porque não divulgar?", pondo a tónica na necessidade de conhecimento factual. Marcelo Rebelo de Sousa, na sua intervenção num telejornal de 2002/10/13, apontou também deficiências várias, nomeadamente modelar em bloco (isto é, calcular a mesma nota esperada para) todas as escolas de um concelho, quando por vezes - pense-se em Lisboa - a diversidade de condições sociais em sub-zonas do concelho é enorme, mas também ele referiu a necessidade de progredir de uma apreensão meramente qualitativa da realidade para uma investigação quantificada.

Posteriormente Nuno Crato, em artigo publicado no *Expresso* (2002-10-12), enunciou críticas severas ao trabalho da equipa da FCSH da UNL: inadequação e instabilidade dos modelos de *"baixos poderes explicativos, ainda por cima obtidos depois da discutível exclusão dos casos que mais se afastavam dos valores esperados pelo modelo"*, que exibem *"correlações espúrias"*, sendo o critério de seriação *"uma engenharia social paternalista* [… e] *tosca"*.

A seriação encomendada pelo Ministério da Educação baseia-se na diferença entre a classificação obtida em exame nacional e uma "classificação esperada" calculada como variável resposta a diversas variáveis sociológicas, usando regressão múltipla, uma das áreas mais usadas - e mais mal usadas - da Estatística, uma disciplina que, cada vez mais, faz parte do instrumental de qualquer trabalhador científico. Mas o deficiente uso da Estatística tem contribuído para o mau nome desta ciência, que de há muito é acusada de mentira superlativa (*"Lies, damned lies and Statistics"*, uma frase que o humorista Mark Twain atribuíu a Disraeli, certamente para exemplificar o que são *damned lies:* ainda hoje muitos pensam que a frase se deve de facto a Disraeli).

Sendo um assunto de interesse geral, mas particularmente pertinente para todos os professores, procuramos repor o bom nome da Estatística referindo os cuidados com que deve ser usada.

Optamos por ir directos ao assunto, expondo na secção 2 como se procedeu à seriação das escolas, e as críticas inevitáveis ao trabalho de Grácio *et al*. (2002), deixando para a secção 3 uma exposição elementar do que é regressão e regressão múltipla, com uma discussão cuidadosa das razões pelas quais um valor $R^2 << 1$ nos deve levar a, prudentemente, abandonar um modelo tão pouco explicativo. Esta é, afinal, a mais severa crítica *estatística* que fazemos ao trabalho daqueles autores, juntamente com um reparo sobre confundimento. No plano do mero bom senso, criticamos também Grácio *et al*. (2002) por usarem como critério de seriação a diferença entre classificações nos exames nacionais e classificações estimadas pelos modelos, um critério disparatado mesmo no caso de os modelos serem excelentes, pois tenderia a penalizar os melhores e a beneficiar os piores.

2. Seriação das Escolas do Ensino Secundário (2001/2002)

A documentação divulgada por Grácio *et al*. (2002) é interessante, mas em alguns aspectos omissa, e noutros pouco crítica.

Comecemos pela descrição do que foi feito:

• Definiram-se variáveis principais
- X_1 : indicador do poder de compra em cada concelho, com base em dados do INE em 2000;
- X_2 : taxa de não escolarização no 12º ano, quociente entre a população de 17-18 anos não escolarizada no 12º ano e a população de 17-18 anos, por concelho;
- X_3 : Número médio de anos de escolaridade da população em cada concelho, com base em dados do INE (recenseamento de 2001);
- X_4 : uma variável muda classificando o estabelecimento de ensino como público ou privado.

Consideraram-se ainda variáveis de interacção entre os factores acima,
- X_{12} : interacção entre poder de compra e taxa de escolarização;
- X_{23} : interacção entre taxa de escolarização e taxa de não escolarização

e definiu-se o modelo geral de regressão múltipla

$$\hat{Y}=a+ b_1 X_1 + b_3 X_3 + b_4 X_4 + b_{12} X_{12} + b_{23} X_{23}$$

onde a variável dependente *Y* é uma variável resposta que pretende ser a predição da nota esperada (num elenco de discipinas escolhidas, e que adiante se explicitam, e na média das disciplinas) de um aluno habitando naquele concelho.

Os coeficentes b_1 , b_3 , b_4 , b_{12} , b_{23} traduzem a influência que as correspondentes variáveis independentes têm sobre a variável resposta.

Não é para nós claro por que razões a variável de primeiro nível X_2 não aparece no modelo. Porventura a metodologia de construção do modelo em cada caso, que adiante detalhamos, nunca levou à inclusão dessa variável, e no relatório final não foi incluída, por isso, no modelo geral.

• O cálculo do modelo em cada um dos casos considerados foi feito com recurso ao *package* estatístico *SPSS*. A inclusão das variáveis foi feita passo a passo (*forward*).

O relatório é naturalmente omisso nos detalhes técnicos, mas certamente houve uma inclusão hierarquizada variável a variável - isto é, em cada passo incluiu-se a variável com maior capacidade de melhorar o modelo, avaliando-se o incremento do poder explicativo que essa inclusão trazia ao modelo. Deixou de haver inclusão de variáveis quando a última inclusão tentada não produziu melhoria significativa, eliminando-se do modelo esta última variável.

Os resultados obtidos constam da tabela abaixo. Certamente a ordem em que as diversas variáveis de regressão aparecem tem que ver com a ,ordem em que foram incluídas no modelo.

São também indicados, em cada caso, os valores de R^2.

	Modelo seleccionado	R^2
Média das disciplinas	$\hat{Y}=88.710+0.002\ X_{12} +7.326\ X_4$	0.213
Biologia	$\hat{Y}_B =76.250+0.002\ X_{12}+5.324\ X_4+1.637\ X_3$	0.216
Matemática	$\hat{Y}_M =36.255+3.606\ X_3+4.208\ X_4+0.001\ X_{12}$	0.192
Sociologia	$\hat{Y}_S =107.400 + 0.003\ X_{12}$	0.133
Des. e Geom. Desc. A	$\hat{Y}_D = 45.828+8.731\ X_3$	0.089
Química	$\hat{Y}_Q =67.983+3.837\ X_3 + 4.382\ X_4$	0.086
Português A	$\hat{Y}_P =97.660+0.002\ X_{12}$	0.055
História	$\hat{Y}_H =77.806+3.241\ X_3$	0.044
Inglês 6	$\hat{Y}_I =60.603+0.002\ X_{12} +12.261\ X_4+0.033\ X_{23}$	0.048
Filosofia	$\hat{Y}_F =107.491+0.001\ X_{12}$	0.018

A tabela acima foi transcrita da p. 15-16 de Grácio *et al*. (2002) [1]

[1] Fornecem também aqueles autores (p. 11) coeficientes estandardizados, que alguns preferem para efeitos de comparabilidade:

	Modelo seleccionado	R^2
Média das disciplinas	$\hat{Y}=88.7+0.358\ \tilde{X}_{12} +0.240\ \tilde{X}_4$	0.213
Biologia	$\hat{Y}_B =76.3 +0.302\ \tilde{X}_{12}+0.138\ \tilde{X}_4+0.140\ \tilde{X}_3$	0.216
Matemática	$\hat{Y}_M =36.3 +0.279\ \tilde{X}_3+0.098\ \tilde{X}_4+0.153\ \tilde{X}_{12}$	0.192
Sociologia	$\hat{Y}_S =107.4 + 0.365\ \tilde{X}_{12}$	0.133
Des. e Geom. Desc. A	$\hat{Y}_D = 45.8+0.299\ \tilde{X}_3$	0.089
Química	$\hat{Y}_Q =68 + 0.263\ \tilde{X}_3 + 0.091\ \tilde{X}_4$	0.086
Português A	$\hat{Y}_P =97.7+ 0.234\ \tilde{X}_{12}$	0.055
História	$\hat{Y}_H =77.8 +0.210\ \tilde{X}_3$	0.044
Inglês 6	$\hat{Y}_I =60.6 +0.126\ \tilde{X}_{12} +0.127\ \tilde{X}_4+0.104\ \tilde{X}_{23}$	0.048
Filosofia	$\hat{Y}_F =107.5 +0.136\ \tilde{X}_{12}$	0.018

• Calcula-se então a nota média Y_i (onde $i \in \{ B, M, S, \ldots \}$) obtida pelos alunos da escola, no exame nacional, em cada uma das disciplinas, e determina-se a diferença $d_i = Y_i - \hat{Y}_i$ entre a classificação média observada e a classificação esperada (isto é, postulada pelo modelo). Estas diferenças são ordenadas decrescentemente, e disso resulta a seriação das escolas no que refere cada uma das nove disciplinas consideradas.

Procede-se analogamente para calcular a diferença no que refere médias das disciplinas das escolas, $Y - \hat{Y}$, sendo

$$Y = \frac{\sum_{j=1}^{N_d} Y_j}{N_d},$$

onde N_d representa o número de disciplinas de 12º leccionadas na escola, e Y_j é a classificação média nos exames na *j*-ésima dessas disciplinas (não há indicação de que se use depois uma ponderação nas fórmulas que tenha em conta o número muito desigual de alunos que se apresentam a exame nas diversas disciplinas), e $\hat{Y}$ é definida de forma análoga, usando as notas esperadas $\hat{Y}_i$ calculadas pelos modelos de regressão ajustados a cada uma delas.

Não conseguimos explicar aos leitores a que ponto as variáveis independentes são importantes, em absoluto, na composição desta "nota esperada", porque não é explícito que valores podem assumir, qual a escala em que cada variável independente foi registada. Não há, consequentemente, informação que nos permita ter uma ideia de como comparam com a ordenada na origem, a que ponto a podem alterar.

Para ser mais concreto: se cada uma das variáveis independentes (com excepção da *dummy* X_4 , de que consideramos o estado X_4 =0) variar de -10 a 10, a nota esperada de Biologia pode ir de 76.25-0.02-16.37= 59.86 a 76.25+0.02+16.37=92.64; se variar de -5 a +5, pode ir de 76.25-(0.02+16.37)/2=68.06 a 76.25+(0.02+16.37)/2=84.455; se variar de -80 a 80, a nota esperada pode ir de 76.25-8(0.02+16.37)=-54.87 a 76.25+8(0.02 +16.37)=207.37. (Claro que nada obriga as diversas variáveis a tomarem valores nos mesmos intervalos.)

Recomendamos, naturalmente, a verificação das fontes— neste caso Grácio *et al*. (2002) - podendo assim o leitor fazer a sua própria leitura e interpretação. Até os autores manifestam algum desconforto com o estudo a que procederam, nomeadamente devido à falta de *"informação sobre o estado dos conhecimentos académicos dos alunos no âmbito de cada disciplina à entrada do 12º ano"*, com o facto de calcularem a mesma nota esperada (por disciplina e média) para todas as escolas do mesmo concelho, o que *"não repercute portanto a própria variedade socio-cultural existente no interior do mesmo concelho"*, ou haver *"escolas privadas em contrato de associação com o Ministério da Educação"*. Esta descrição detalhada foi feita no intuito de facilitar a compreensão das críticas (que não insistem nestes pontos fracos que os autores já reconheceram) que seguem:

1. O critério de seriação é inadequado

Logo à cabeça, não podemos deixar de sublinhar um ponto que nos parece contestável à luz do senso comum, crítica que ainda por cima persistirá mesmo que os modelos adoptados para atribuição de uma nota esperada venham a tornar-se muito mais sofisticados: o critério de seriação adoptado pode ser traduzido por um aforismo apropriado para um inferno mais moderno do que o de Dante: *Aqui, os bons nada podem esperar, os maus nada têm a temer!*, em vez do dantesco *"Vós que entrais, abandonai toda a Esperança"*.

Com o critério usado - desvio entre a nota média obtida pelos alunos da escola e a nota "esperada" calculada por um modelo - é obviamente mais provável que uma escola com má nota esperada seja seriada no topo (e dificilmente será seriada no fundo) do que uma escola que tenha uma nota esperada elevada, a qual provavelmente obterá um *rank* muito pouco agradável. Tomando as notas valores de 0 a 20, dificilmente uma escola em que a nota esperada é 19 poderá subir, dificilmente uma escola em que a nota esperada é 1 poderá descer.

Uma caricatura expressiva dos erros em que se incorre

com a utilização de desvios deste tipo como base de seriação é a seguinte:

- Usa-se no papel de $\hat{Y}$ a nota de ingresso do último candidato admitido num curso universitário no concurso nacional, 1ª fase, há seis anos. Suponha-se, só a título de exemplo, que
  - Na Licenciatura em Matemática da Universidade A… foi 15.9.
  - Na Licenciatura em Matemática da Universidade B… foi 5.4.
- Usa-se no papel de Y a nota média de conclusão da licenciatura dos licenciados no ano passado. Continuando o exemplo,
  - Na Licenciatura em Matemática da Universidade A…, 13.7.
  - Na Licenciatura em Matemática da Universidade B…, 12.9.
- Atribuem-se pontuações Y- $\hat{Y}$, isto é, no caso destes cursos,
  - Na Licenciatura em Matemática da Universidade A..., −2.2.
  - Na Licenciatura em Matemática da Universidade B..., +7.5.

É óbvio que a Licenciatura em Matemática da Universidade B deve ser seriada muito acima da Licenciatura em Matemática na Universidade A, não é? A Licenciatura em Matemática da Universidade B atrai alunos menos qualificados, os que terminam a licenciatura fazem-no com médias mais baixas, mas que importam esses pormenores?

2. Os modelos usados são inadequados.

Cito, do clássico de Mendenhall and Sincich (1996, 5th ed., p. 191):

> *"R^2 is a sample statistic that represents the fraction of the sample variation of the y values (measured by SS_{yy}) that is attributable to the regression model. Thus, R^2=0 implies a complete lack of fit of the model to the data, and R^2 =1 implies a perfect fit, with the model passing through every data point. In general, the closer the value of R^2 is to 1, the better the model fits the data.*
>
> *To illustrate, the value R^2 =0.9377 in the immunoglobulin example […] implies that 93.8 % of the sample variation in IgG (y) is attributable to, or explained by, the independent variable maximal oxigen uptake (x). Thus R^2 is a sample statistic that tells how well the model fits the data, and thereby represents a measure of the utility of the entire model."*

Creio que não deixa dúvidas: qualquer dos modelos propostos por Grácio *et al.* (2002) é inadequado, grosseiramente inadequado, e não sei que interpretação dar à afirmação *" No caso da média de todas as disciplinas por escola compreende-se o poder explicativo elevado do modelo"* (Grácio *et al.*, 2002, p.11) quando o valor de R^2 é apenas 0.213.

A verdade é que os valores da estatística R^2 são todos excessivamente baixos (no melhor dos casos, Biologia, o modelo não consegue explicar mais do que 22 % da variância de Y — o que quer que isto queira dizer nesta situação tão bizarra - e, em 6 dos 10 casos estudados, menos de 10 %, baixando até aos 2 %), e qualquer utilizador de Estatística, face a estes valores, deveria procurar modelos alternativos.

Nestas condições, a eliminação dos casos que mais directamente questionavam o ajustamento (os *outliers*) torna-se suspeita. Consulte-se o clássico de Barnett e Lewis (1994), e a sua discussão do que é um *outlier.*

Aliás, simples reflexão leva-nos a questionar os modelos propostos. Por exemplo, sabendo-se que os encarregados de educação de alunos do 12º ano recorrem frequentemente à contratação de explicadores de Matemática, fará sentido um modelo em que a variável X_1 - *poder de compra em cada concelho* não aparece senão através da interacção X_{12}, que apenas é incluída na terceira iteração e com uma carga 0.001?

E que pensar do modelo $\hat{Y}_S$=107.400+0.003 X_{12} para a Sociologia? A Sociologia será uma área de estudo quase imune a variáveis sociais? (Como atrás comentámos, não sabemos qual a escala de variação de X_{12}, mas a sua influência no cálculo da nota esperada parece irrelevante). Comentários idênticos, naturalmente, para os modelos $\hat{Y}_P$ =97.660+0.002 X_{12} para Português A, e $\hat{Y}_F$ =107.491+0.001 X_{12} para Filosofia.

Em alguns modelos a variável muda X_4 tem uma preponderância que leva a reflexões hamletianas: *ser ou não ser privado, eis a questão…*. Que estranho noutros casos nem sequer aparecer!

3. Justifica-se um modelo de regressão múltipla?

Não é uma questão académica. A regressão múltipla é uma área sofisticada da Estatística, com um desenvolvimento matemático rigoroso, que se apoia em pressupostos tais como gaussianidade (normalidade) dos dados, que dificilmente se aplicam quer a algumas das variáveis sociológicas usadas como regressores quer às notas dos exames nacionais (e neste caso não temos dúvidas, basta fazer um histograma). É uma questão demasiado técnica para poder ser abordada sucintamente, recomendamos aos interessados o Capítulo 6 *(Some Regression Pitfalls*) de Mendenhall and Sincich (1996).

Claro que há "Modelos Lineares Generalizados"; também neste caso há que começar por verificar se são usáveis com os dados disponíveis.

4. A seriação padece de confusão.

A meu ver, a inadequação dos modelos deveria ter levado ao seu abandono, como referido no ponto 2. Preferem Grácio *et al*. (2002, p. 4) concluir que a diferença entre a nota média obtida pelos alunos da escola e a nota calculada pelo modelo que propugnam "*é uma* aproximação *ao contributo das escolas para a aprendizagem dos alunos*.

Apesar do cuidadoso frasear (nomeadamente do termo *aproximação*, em itálico), que parece indicar que os autores têm algumas preocupações sobre *confundimento* (apesar de tal não ser arrolado nas interessantes notas críticas da Nota Final, pp.12-13), esta variável é usada para proceder à seriação das escolas.

Mas "*O alívio é irmão gémeo do desapontamento. Ambos se dizem do mesmo modo: pelo suspiro*"[2] Por isso, um suspiro não pode ser automaticamente interpretado como alívio.

Um exemplo menos poético: Se treinarmos 30 cães de um *grupo experimental* na ilha do Corvo a atravessarem a rua apenas quando um humano o faz, e não treinarmos 30 cães de um *grupo de controlo* em Lisboa, e ao fim de seis meses tiverem morrido atropelados 16 dos cães não treinados e apenas um dos cães treinados, a diferença altamente significativa não pode ser atribuída ao treino, porque a diferente localização (e concomitante tráfego rodoviário) dos dois grupos é um factor de confundimento.

Na questão da seriação das escolas, numa investigação meramente observacional como a que foi feita, não terão ficado de fora tantos factores de confundimento? Por exemplo, número de professores licenciados habitando no concelho, número de professores efectivos residindo no concelho, realização de acções de formação no concelho, variedade e abundância de locais de jogo no concelho, riqueza dos programas culturais das autarquias locais, existência de bibliotecas com ambiente agradável, locais de reunião e estudo como a *Ágora* de Lisboa.

Além disso, só planeamentos experimentais cuidadosos permitem concluir causalidade da correlação (veja-se por exemplo Schweigert, 1994, tão cuidadosa porventura por saber que o seu público alvo é de utilizadores de Estatística das áreas de Ciências Humanas). Em estudos meramente observacionais, como este, é um salto no desconhecido - autores há que lhe chamam "o pecado mortal dos maus utilizadores da Estatística" - inferir que a diferença entre notas obtidas e notas esperadas (ainda que os modelos fossem adequados, o que nem é o caso) é "causada" pela intervenção da escola e dos profissionais que nela trabalham.

Grácio *et al*. (2002) reconhecem algumas limitações do modelo que propõem. Parece-me essencial que reconheçam que o modelo que propõem, longe de *ser " mil vezes preferível a fornecer simplesmente ao público os*

[2] Mia Couto (2002). *Um Rio Chamado Tempo, Uma Casa Chamada Terra*, Caminho, Lisboa, p 137.

dados em bruto" é apenas um excelente exemplo de mau uso da Estatística (ou, mais propriamente, mau uso de um *package* estatístico).

3. Regressão - modo de usar

Usando um *package* estatístico para exprimir uma *variável resposta* como função *de variáveis controladas,* há sempre uma resposta. É aliás usual dizer que o problema reside em que quando se mete lixo no computador, dele só sai lixo.

Muitos programas - e nomeadamente o *SPSS* - são vendidos com excelente documentação, e têm informação *online*; mas em última análise, é o utilizador que tem a responsabilidade de ter uma visão crítica do que está a fazer. Áreas como regressão e análise da variância são naturalmente muito apelativas para os utilizadores, e por isso é frequente deparar com exemplos de mau uso da Estatística (alguns autores falam mesmo de *abuse of Statistics*).

Descrevemos por isso de forma rudimentar o que é a regressão linear (se o padrão não for linear, a situação é muito mais complexa, usando-se em muitos casos transformações capazes de linearizar os dados). Aconselhamos sempre uma representação prévia dos dados num diagrama de dispersão, por forma a avaliar visualmente se um padrão linear é adequado. Discutimos também o sentido do coeficiente de determinação r^2, para se perceber por que razão só se deve usar um modelo linear se $r^2 \approx 1$ (e, mesmo assim, há sempre que ser crítico, veja-se Pestana e Velosa, 2002, p. 149, onde se alerta para situações em que valores elevados de r^2 não correspondem a padrões lineares de associação, e valores de r^2 próximos de zero surgem com dependência - não linear, claro - estrita).

A regressão múltipla é uma extensão que nada tem de revolucionário, do ponto de vista conceptual, apenas já não há uma avaliação visual que nos guie - temos por isso que usar R^2 como critério da qualidade do ajustamento, devendo valores baixos serem um sinal de sentido proibido no que refere o modelo que ensaiámos.

Mas entremos na questão do porquê usar regressão, exemplificando o uso da regressão linear, a fim de simplificar a exposição:

Por vezes temos facilmente acesso a uma variável *x*, e gostaríamos de a usar para conhecer uma variável *y* que julgamos estar fortemente correlacionada com a primeira. Por exemplo, há boas razões para acreditar que há evasão fiscal, que há contribuintes que mentem sobre o montante *y* dos seus rendimentos, e seria interessante usar uma avaliação *x* dos sinais exteriores de riqueza para estimar *y.* Pretende-se por isso um *modelo de regressão* $y = \hat{y} + \varepsilon$, onde $\hat{y}_i = f(x_i)$ é a estimativa de y_i correspondente ao valor observado x_i e ε_i é o *resíduo* nesse ponto; por vezes, também se chama a $\hat{y}_i$ "sinal", e a ε "ruído".

A variável *y* é em geral denominada variável dependente ou variável resposta, e a variável *x* é a variável independente, ou variável controlada, ou preditor; não têm, obviamente, o mesmo estatuto, há uma hierarquização importante, não faz sentido inverter a função *f* para exprimir *x* como função de *y* (seria ignorar os resíduos, mas sobretudo não entender o âmago do problema, em que a referida hierarquização variável resposta/variável controlada é essencial).

Um exemplo ajuda a esclarecer: predizer o perímetro da cabeça à nascença é decerto importante, pode determinar se se deve antecipar um parto normal ou necessidade de recorrer a uma cesariana. Podemos por isso recorrer a uma amostra para estudar o problema.

Suponha-se que uma equipa de obstetras recolhe os seguintes dados, relativos ao comprimento (em cm) do biparietal, medido recorrendo a uma ecografia do feto na 34ª semana da gravidez, e ao perímetro da cabeça na altura do nascimento, também em centímetros,

x - biparietal (34ª semana)	*y* - perímetro cefálico (à nascença)	*x* - biparietal (34ª semana)	*y* - perímetro cefálico (à nascença)
8.09	33.13	7.81	31.99
9.02	36.45	8.23	34.22
8.66	35.76	9.24	37.63
9.03	35.59	9.07	36.97
8.03	32.66	7.49	30.36
8.61	34.57	8.21	33.66
8.98	36.23	9.04	35.30
8.55	35.10	8.97	36.44
8.82	35.59	8.30	35.26
8.25	33.94	8.66	35.51
8.31	33.83	8.76	34.86

sendo o objectivo exprimir a variável dependente *y* (perímetro cefálico) em função da variável independente *x* (comprimento do biparietal), *y=f(x)*.

Por outras palavras, consideramos que os valores observados y_i são flutuações amostrais em torno de um modelo $\hat{y}_i = f(x_i)$, ou seja que os valores observados podem ser escritos

$$y_i = \hat{y}_i + \varepsilon_i ,$$

sendo os ε_i resíduos desprezáveis. Claro que gostaríamos que os resíduos fossem, idealmente, nulos, um ajustamento perfeito do modelo à realidade. Mas isso é pedir demais; por isso vamos ser mais modestos, e esperar que sejam "perturbações amostrais" de 0, isto é que flutuem (moderadamente) em torno de 0, sem padrão definido.

Caso nada se diga sobre a forma analítica de *f*, o problema é naturalmente indeterminado. Em muitas circunstâncias os dados (ou transformações simples dos dados) exibem um padrão linear. Neste caso, há boas razões para esperar um padrão linear, uma vez que o perímetro cefálico à nascença *y* não deve afastar-se muito de $\pi \times$ biparietal à nascença, sendo decerto o biparietal à nascença um pouco maior do que a avaliação obtida por ecografia na 34ª semana. Um representação gráfica dos dados não suscita dúvidas sobre a adequação deste padrão linear (Figura 1).

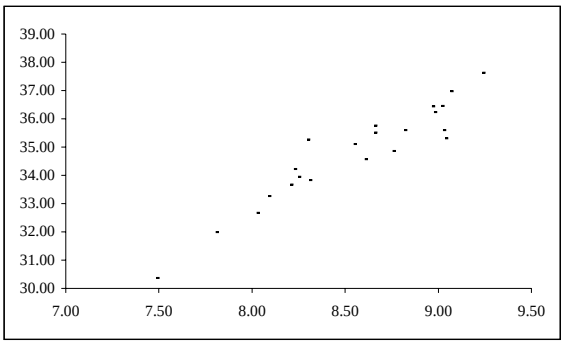

Figura 1: Gráfico de dispersão dos dados (biparietal, perímetro cefálico).

Vamos por isso usar um ajustamento linear aos dados, $\hat{y}=f(x)=ax+b$, usando um critério de aproximação adequado. O mais usual é adoptarmos o *critério dos mínimos quadrados*: vamos determinar os parâmetros (coeficientes) da função por tal forma que a soma dos quadrados dos desvios (isto é, resíduos) entre os valores observados y_i e os valores estimados $\hat{y}_i = f(x_i)$

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

seja mínima.

Note-se que neste exemplo o coeficiente de correlação é $r_{x,y}=0.95$ (o coeficiente de determinação é $r^2 \approx 0.90$) pelo que esperamos que haja uma associação linear forte entre as variáveis.

Vamos então determinar os coeficientes *a* e *b* tais que o desvio quadrático global

$$Q(a,b)= \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

seja mínimo[3].

[3] Estamos a minimizar a soma dos quadrados dos desvios *medidos na vertical* entre cada ordenada observada e a correspondente ordenada estimada pela recta. Isto complementa a observação anterior sobre a hierarquização das variáveis e a observação de que não faz sentido inverter a função *f* - para exprimir *x* como função de *y* há que resolver outro problema, que é minimizar a soma dos quadrados dos desvios *medidos na horizontal*, entre cada abcissa observada e o valor postulado como função da correspondente ordenada (Figura 2).

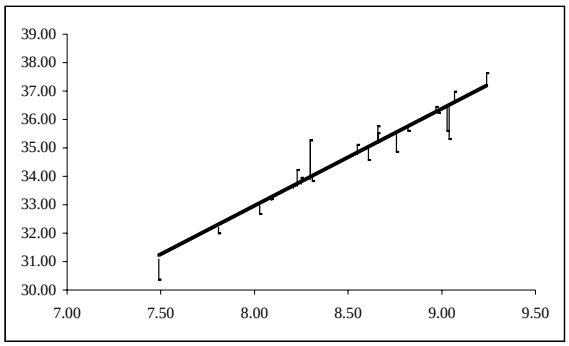

Figura 2: Desvios "verticais" dos pontos à recta.

Tomando então o sistema de estacionaridade (isto é, igualando a zero as derivadas parciais em relação às incógnitas *a* e *b*, respectivamente), obtém-se

$$\begin{cases} \dfrac{\partial}{\partial a} Q(a,b) = \displaystyle\sum_{i=1}^{n} 2(y_i - ax_i - b)(-x_i) = 0 \\ \dfrac{\partial}{\partial b} Q(a,b) = \displaystyle\sum_{i=1}^{n} 2(y_i - ax_i - b)(-1) = 0 \end{cases}$$

(a segunda daquelas equações exprime que a soma dos resíduos é nula, como tínhamos requerido).

Ficamos assim com o sistema de duas equações lineares nas duas incógnitas *a* e *b*,

$$\begin{cases} \displaystyle\sum_{i=1}^{n} x_i^2 a + \sum_{i=1}^{n} x_i b = \sum_{i=1}^{n} x_i y_i \\ \displaystyle\sum_{i=1}^{n} x_i a + nb = \sum_{i=1}^{n} y_i \end{cases}$$

cuja solução é

$$\begin{cases} a = \dfrac{n\displaystyle\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\displaystyle\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ b = \bar{y} - a\bar{x} \end{cases}$$

No exemplo, a= 3.49, b= 4.92, donde o modelo de regressão y=3.49 x+4.92+ ε, um resultado que não é inesperado: devido à forma da cabeça, estamos a calcular algo que está muito próximo do perímetro de uma circunferência com diâmetro ≈ biparietal. Assim, o coeficiente 3.49 é uma perturbação de π =3.14 (note-se que o biparietal cresce entre a trigésima quarta semana e a altura do parto, o que também explica uma ordenada na origem positiva, e que o perímetro máximo da cabeça não é exactamente o perímetro de uma circunferência).

Vejamos agora qual o sentido da recta de regressão:

Como $\displaystyle\sum_{i=1}^{n} x_i = n\bar{x}$, segue-se que

$$a = \frac{n\displaystyle\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\displaystyle\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\displaystyle\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\displaystyle\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}.$$

Notando que $\displaystyle\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = (n-1)s_x^2$, e que

$$\sum_{i=1}^{n} \left(x_i - \bar{x}\right) y_i = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = (n-1)\,\mathrm{cov}(\mathbf{x}, \mathbf{y})$$

(a primeira igualdade é consequência imediata de $\displaystyle\sum_{i=1}^{n} \left(x_i - \bar{x}\right) = 0$) deduz-se então

$$a = \frac{\mathrm{cov}(\mathbf{x}, \mathbf{y})}{s_x^2} = \frac{\mathrm{cov}(\mathbf{x}, \mathbf{y})}{s_x s_y} \frac{s_y}{s_x} = r\frac{s_y}{s_x},$$

onde r denota o coeficiente de correlação empírico entre ***x*** e ***y***.

Por conseguinte a recta de regressão pode ser reescrita

$$\hat{y} = ax + b = ax + \bar{y} - a\bar{x} \Leftrightarrow y - \bar{y} = r\frac{s_y}{s_x}(x - \bar{x}),$$

e padronizando ***x*** e ***y***

$$\frac{\hat{y} - \bar{y}}{s_y} = r\frac{x - \bar{x}}{s_x},$$

ou seja a ordenada padronizada é proporcional à abcissa padronizada, sendo $r=r_{xy}$ o coeficiente de proporcionalidade.

Em consequência, a variância da versão padronizada de $\hat{\mathbf{y}}$ é, pelo que sabemos sobre a variância de uma transformação linear, $r^2 \times$ variância da versão padronizada de $\boldsymbol{x}$,ou seja

$$\frac{\text{var}(\hat{\boldsymbol{y}})}{s_y^2} = r_{x,y}^2 \frac{\text{var}(\boldsymbol{x})}{s_x^2} = r_{x,y}^2 \Rightarrow \text{var}(\hat{\boldsymbol{y}}) = r_{x,y}^2 \, \text{var}(\boldsymbol{y})$$

e consequentemente[4], de $\mathbf{y}=\hat{\mathbf{y}}+\varepsilon$ segue-se que

$$\text{var}(\boldsymbol{y}) = \text{var}(\hat{\boldsymbol{y}}) + \text{var}(\varepsilon)$$

Quer isto dizer que

$$\text{var}(\boldsymbol{y}) = r_{x,y}^2 \, \text{var}(\boldsymbol{y}) + \text{var}(\varepsilon)$$

Por isso $r^2 = r_{x,y}^2$ é interpretado como a fracção da variância de $\boldsymbol{y}$ que é explicada pela relação linear entre $\boldsymbol{y}$ e $\boldsymbol{x}$, e denominado *coeficiente de determinação*. E, como subproduto, conclui-se que a fracção da variância de $\boldsymbol{y}$ que fica por explicar pela referida relação é $1-r^2$. Ou ainda, $\text{var}(\varepsilon)= (1- r_{x,y}^2) \, \text{var}(\mathbf{y})$. De facto, desenvolvendo $\sum_{i=1}^{n} (y_i - y_i)^2$, e recordando que $b = \bar{y} - a\bar{x}$, vem

$$\sum_{i=1}^{n} (y_i - \hat{y_i})^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2 = \sum_{i=1}^{n} \left[(y_i - \bar{y}) - a(x_i - \bar{x}) \right]^2.$$

Como $a = \dfrac{\text{cov}(\boldsymbol{x}, \boldsymbol{y})}{s_x^2}$, e $\bar{\varepsilon} = 0$, segue-se que

$$(n-1)\, s_\varepsilon^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 =$$

$$= \sum_{i=1}^{n} (y_i - \bar{y})^2 - (n-1) \left[\frac{\text{cov}(\boldsymbol{x}, \boldsymbol{y})}{s_x} \right]^2 =$$

$$= (n-1)\, s_y^2 - (n-1) \left[\frac{\text{cov}(\boldsymbol{x}, \boldsymbol{y})}{s_x} \right]^2 =$$

$$= (n-1)\, s_y^2 \left[1 - \left(\frac{\text{cov}(\boldsymbol{x}, \boldsymbol{y})}{s_x s_y} \right)^2 \right],$$

concluindo-se então que

$$s_\varepsilon^2 = \left(1 - r^2\right) s_y^2.$$

Naturalmente, se $r^2 << 1$ consideramos que o modelo de regressão não é adequado. Consulte-se Pestana e Velosa (2002, Capítulo 2), de que o exemplo acima é extraído, onde muitos gráficos expressivos ilustram os cuidados que há a ter na interpretação de correlações.

Da expressão $\dfrac{\hat{\mathbf{y}} - \bar{y}}{s_y} = r_{x,y} \left(\dfrac{\boldsymbol{x} - \bar{x}}{s_x} \right)$ tira-se

$$\hat{\boldsymbol{y}} = \bar{y} + r_{x,y} \frac{s_y}{s_x} (\boldsymbol{x} - \bar{x}).$$

Por outras palavras, para predizer *y* fazemos uma correcção ao seu valor médio $\bar{y}$, correcção essa que é proporcional ao desvio estandardizado do valor observado do preditor *x* relativamente à sua média $\bar{x}$ - directamente proporcional em termos da correlação entre as variáveis e do desvio padrão de *y*. (Esta expressão paraleliza um importante resultado sobre valores médios condicionais em pares aleatórios gaussianos; a fundamentação teórica da adaptação de uma *recta* de regressão é a admissão implícita de que o modelo populacional é "multinormal"[5]).

Por exemplo, suponha-se que a correlação entre peso e altura de jovens portugueses na classe etária dos 18-25 anos (uma classe bem conhecida no que respeita muitas características físicas, por causa da inspecção militar) é 0.95, que sabemos que a altura média é 1.72 m, que o desvio padrão é 0.07 m, que o peso médio é 68 kg, e o desvio padrão é 4 kg.

A melhor ideia que podemos ter sobre o peso de um indivíduo escolhido ao acaso é o peso médio, 68 kg.

---

[4] Aqui permitimo-nos alguma falta de rigor: estamos a admitir que resíduos e modelo são "independentes", e que por isso a variância da soma é a soma das variâncias. Mas é irresistível tentar mostrar que há ideias simples e interessantes motivadoras do que estamos a fazer, e que não é apenas uma colecção de fórmulas complicadas e sem sentido. Por outro lado, deduzimos já de seguida que $s_\varepsilon^2 = \left(1 - r^2\right) s_y^2$.

[5] Lukacs (1956) caracterizou os modelos populacionais em que a regressão é linear.

Porém, se quiserem ter uma ideia melhor do peso, e conseguirem avaliar que a altura é 1.86 m - um desvio de 0.14 em relação à altura média, ou seja um desvio padronizado 0.14/0.07=2, parece sensato fazer uma correcção à estimativa inicial peso, correcção essa que é $0.95 \times 4 \times 2$=7.6 kg, avaliando-se agora o peso em 75.6 kg.

Claro que procurar relações que já se conhecem à partida, como no exemplo perímetro cefálico à nascença = 3.49 x biparietal na 34ª semana +4.92, só tem interesse na exposição inicial do método. Obter dados como os descritos nem é fácil (muito antes da 34ª semana de gravidez já a maioria dos bebés tem a cabeça voltada para baixo, numa posição que não permite boa medição do biparietal numa ecografia). O que em geral vem registado numa ecografia é o comprimento do fémur (Hadlock *et al.*, 1983). A correlação entre o comprimento do fémur e o perímetro da cabeça não é com certeza tão forte quanto a correlação entre o biparietal e o perímetro da cabeça - mas é suficientemente elevada para permitir predizer linearmente usando o método dos mínimos quadrados, obtendo intervalos de predição com um grau de confiança elevado.

Detemo-nos um pouco mais neste ponto. Queremos predizer o perímetro cefálico à nascença, e seria natural, em termos geométricos, usar a medição ecográfica do biparietal na 34ª semana. Mas por razões biológicas que todos entendemos, nem sempre esta medição está acessível. Mas podemos aceder facilmente a outras medições - batimento cardíaco, comprimento do fémur, comprimento do polegar, *pH* do líquido amniótico, perímetro do pescoço, comprimento da coluna. O natural é, de entre as várias candidatas, escolher a que tem a melhor relação custo/benefício, sendo naturalmente o custo a dificuldade em medir essa variável, e o benefício um coeficiente de correlação entre preditor e variável resposta com valor absoluto próximo de 1. Teremos assim um coeficiente de determinação elevado, por outras palavras a dispersão da variável resposta fica quase totalmente explicada pela variabilidade da variável controlada.

A capacidade preditiva do biparietal é excelente (Kurtz *et al.*, 1980), decerto melhor do que a do fémur. Mas na prática é o comprimento do fémur esquerdo que se usa, porque é fácil de medir, e o coeficiente de determinação é suficientemente elevado para nos permitir predizer de forma útil (Hadlock *et al.*, 1983). Os outros candidatos a regressor referidos levam a coeficientes de determinação mais baixos, e por isso são preteridos.

Convém também anotar que maior comprimento do fémur não "causa" maior perímetro craniano. Trata-se de uma associação estatística, e não de causalidade. Estudos observacionais como o descrito não permitem aceder a causalidades, apenas estudos experimentais, em que deliberadamente se altera o "tratamento" de um grupo experimental para comparar a diferença de efeitos médios face a um grupo "de controlo" que não foi alterado, permitem estabelecer relações de causa a efeito.

Note ainda que num sentido estrito os dados acima deveriam ser analisados numa perspectiva de correlação, e não numa perspectiva de regressão. No referido estudo, nunca poderíamos considerar o biparietal *x* uma "variável controlada", e o perímetro cefálico *y* uma "variável resposta". Mas pareceu-nos didacticamente interessante tratá-los deste modo[6], pela interpretação intuitiva que fornecem para os coeficientes da recta de regressão. E na perspectiva em que nos colocámos, embora *x* não seja de facto uma variável controlada, é anterior (34ª semana) a *y*, e serve decerto como preditor.

Tudo o que foi exposto faz sentido, insistimos, se houver, de facto, um padrão linear; $r^2 \approx 1$ não garante a existência desse padrão (veja-se Pestana e Velosa, 2002, p.149),

[6] Apesar de haver diferenças conceptuais importantes entre estudos em que se usa a regressão e estudos em que se usa a correlação, na prática muitas vezes os dados são usados "como se" servissem para uma e outra abordagem.

Para além dos aspectos meramente técnicos, é importante para os utilizadores de Estatística saberem se estão a fazer um estudo experimental ou um estudo observacional, se podem supor que uma das variáveis é controlada, e que condicionalmente a cada valor dessa variável é possível calcular uma resposta média (mesmo que seja apenas de uma observação!) que possa ser encarada como valor observado do valor médio condicional, que é o conceito teórico de regressão.

mas $r^2 << 1$ garante que não existe. É um ponto importante em que os utilizadores da Estatística devem atentar: o valor da Estatística está na sua capacidade de arredar ideias erradas, preconceitos, e é nesse papel que é um garante da Ciência. Já no século XIX Carlyle (*Chartism*) observou argutamente:

> *A witty statesman said you might prove anything by figures, but a judicious man looks at statistics, not to get knowledge, but to save himself from having ignorance foisted on him.*

A Estatística é um instrumento que nos ajuda a transformar informação em conhecimento. Se houver mais informação disponível, por exemplo diversas variáveis controladas $x_1, x_2, \ldots, x_n$ disponíveis, podemos decerto estimar melhor a variável resposta y - desde que, evidentemente, haja uma correlação apreciável entre x_k e y. Podem os interessados ver em Shepard *et al.* (1982), por exemplo, a excelente predição que se pode fazer do peso de um bebé à nascença, usando como variáveis regressoras o comprimento e o perímetro abdominal do feto, medidos em ecografias.

Não há assim novidades conceptuais em considerar um modelo de *regressão múltipla*

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n.$$

Neste caso, obviamente, perde-se a capacidade de avaliação visual da (in)existência de padrão linear. Os cálculos são obviamente mais complexos, e raramente alguém os faz sem recurso a um *package* adequado. Claro que pode haver informação redundante, e nessa altura a utilização de algumas das variáveis controladas pode não trazer mais-valias. Por isso os algoritmos são em geral *stepwise,* ou vão incluindo as variáveis uma a uma, escolhendo em cada passo a que pode trazer mais poder explicativo ao modelo, parando quando não há melhoria significativa (*forward*), ou partem de um modelo considerando todas as variáveis controladas disponíveis, e em cada passo elimina-se a que é menos explicativa, enquanto essa eliminação não causa uma degradação qualitativa significativa *(backwards*).

A avaliação da qualidade do modelo é feita em termos da estatística R^2, uma extensão do r^2 que usámos em regressão linear simples: $R^2 \approx 1$ não garante que o modelo de regressão linear múltipla seja adequado - há correlações espúrias[7] -, mas permite-nos ter esperança de que não seja grosseiramente inadequado. Mas $R^2 << 1$ denuncia a inadequação do modelo.

O leitor decerto não estranhará que consideremos indefensáveis os modelos propostos por Grácio *et al.* (2002) para a seriação das escolas.

4. Uso e Mau Uso da Estatística, Informação e Conhecimento

Não é raro aparecerem trabalhos científicos fazendo mau uso da Estatística, uma situação que decorre de os próprios avaliadores de revistas de outras áreas do conhecimento não dominarem apropriadamente as metodologias estatísticas. Em particular em Ciências Humanas, a tentação de propor modelos simples para fenómenos complexos tem levado a polémicas como as que envolveram a publicação de Herrnstein and Murray (1994). Quiseram estes autores usar o Quociente de Inteligência como preditor do sucesso do indivíduo na sociedade - com corolários perversos como a irrelevância de gastos públicos em educação.

A comunidade científica (Devlin *et al.*, 1997) cerrou fileiras a condenar o mau uso da Estatística feito por aqueles autores[8] e S. Jay Gould (1997) publicou um livro com o sa-

[7] 100% das pessoas que morrem de cancro praticaram relações sexuais ou são filhas de pessoas que praticaram relações sexuais, mas daí não se deve inferir que a prática de relações sexuais explica a preocupante prevalência da doença, uma das principais causas de morte.
Este exemplo é inspirado num conto de *Empresta-nos o Seu Marido ? - e Outras Comédias da Vida Sexual*, de Graham Greene.

[8] Não deixa de ser curioso anotar que em geral é o número de citações que seria os investigadores, para as agências financiadoras. Herrnstein e Murray estão no topo, de tão citados que são pelas exemplares asneiras que cometeram no mau uso da Estatística!
Mais um exemplo de como a utilização acrítica de bases de dados, modelos e algoritmos de cálculo é um itinerário para o disparate, e pertinente o conselho "antes de ligar o computador ligue o cérebro".

boroso título *The Mismeasure of Man*, em que a "má medida" é o QI, que ele arrola, justificadamente, como um instrumento sofisticado de exploração do homem pelo homem.

As Ciências Exactas têm tendência a ser mais prudentes, e a incorrer menos no fascínio que as Ciências Humanas parecem ter pelos números, esquecendo por vezes o necessário recuo na crítica e avaliação dos modelos.

As estações meteorológicas registam diariamente os valores de centenas de variáveis. No entanto, até hoje não foi adoptado nenhum modelo de previsão do estado do tempo a médio ou a longo prazo, apesar da relevância que essa descoberta teria para áreas tão diversas como aviação, protecção civil, agricultura e turismo. Porventura porque neste caso os erros do modelo seriam imediatamente detectados, e poderiam sair muito caros à Sociedade e, por reflexo, aos autores. Um excelente exemplo de contenção, que recomendamos a todas as áreas do saber.

A utilização da Estatística é indispensável na investigação em qualquer área. Não foi por acaso que o governo britânico, preocupado com a estreiteza de vistas que os novos doutorados revelavam, escreveu o Livro Branco *Realizing our Potential*, em que pedia às universidades que atendessem à necessidade de uma formação pósgraduada prévia em metodologias da investigação científica, em que a Estatística tem papel protagonista (Greenfield, 2002; Graziano and Raulin, 1997). Mas a Estatística é uma disciplina em que confluem raciocínio indutivo (nomeadamente nas aplicações) e dedutivo (na criação matemática dos modelos, sua caracterização e condições de aplicabilidade), os resultados são válidos sob hipóteses que por vezes não admitem relaxação, e por isso é um instrumento que tem que ser usado com cuidado. A citação de Galton que usámos na abertura não poderia ser mais eloquente. E basta consultar Milliken and Johnson (1989, 1997, 2001) para nos darmos conta dos desenvolvimentos conceptuais que a análise de dados "problemáticos" trouxe recentemente à Estatística.

A Estatística, infelizmente, tem mau nome pelo mau uso que dela ocasionalmente se faz. É uma questão que preocupa muitos profissionais - é de facto tão injusto quanto vilipendiar a Medicina pelas desgraças que acontecem a quem recorre a curandeiros.

Não queremos dizer com isto que os estudiosos de outras áreas não devem usar a Estatística, longe disso. Mas com bom senso (quando está em risco a nossa saúde, é bom alguma clarividência: tomarmos por iniciativa nossa uma aspirina, ou ir ao médico?).

É muito citada uma frase de Laplace, afirmando que a Teoria da Probabilidade não é mais do que bom senso sob forma matematizada. Com a sua enorme autoridade, que pena não ter deixado também para reflexão dos vindouros uma frase sobre o bom senso que deve presidir ao uso de qualquer instrumento de transformação da informação em conhecimento, de que a Estatística é porventura o exemplo mais saliente.

Agradecimentos

Agradeço à Drª Ivone Dias Ferreira, assessora de imprensa do Sr. Ministro da Educação, que me enviou prontamente toda a documentação necessária.

Bibliografia

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data,* 3rd ed., Wiley, New York.

Crato, N. (2002). As limitações da Estatística. *Expresso,* 2002—10—12, p. 15.

Devlin, B., Fienberg, S. E., Resnick, D. P. and Roeder, K. (eds.) (1997). *Intelligence, Genes and Success — Scientists Respond to The Bell Curve*, Springer, New York.

Gigerenzer, G. (2002). *Calculated Risks. How To Know When Numbers Deceive You,* Simon and Schusteer, New York.

Gould, S. J. (1997). *The Mismeasure of Man,* Penguin, London.

Grácio, S., Franco, L., Velho, S., Sanches, E. e Rijo, S. (2002). *Proposta de Seriação das Escolas Secundárias Segundo os Resultados Obtidos nos Exames Nacionais de 12º Ano em 2001/2002*, FCSH, Universidade Nova de Lisboa.

Graziano, A. M. and Raulin, M. L. (1997). *Research Methods. A Procedure of Enquiry*, Longman, New York.

Greenfield, T. (2002). *Research Methods. Guidance for Postgraduates*, Arnold, London.

Hadlock, F. P., Deter, R. L., Harrist, R. B., Roecker, E., Park, S. K. (1983). A date independent predictor of intrauterine growth retardation: femur length/abdominal circunference ratio, *Appl. J. Radiology* **141**, 979—984.

Herrnstein, R. J. and Murray, C. (1994). *The Bell Curve: Intelligence and Class Structure in American Life*, The Free Press, New York.

Huff, D. (1991). *How to Lie With Statistics*, Penguin, London.

Kurtz, A. B., Wapner, R. J., and Kurtz, R. J. (1980). Analysis of biparietal diameter as an indicator of gestational age, *J. Clin. Ultrasound* **8**, 319—326.

Lukacs, E.(1956). Characterizations of populations by properties of suitable statistics. *Proc. Third Berkeley Symp. Math. Statist. Probab.*, vol. 2, 195—214.

Mendenhall, W. and Sincich, T. (1996*). A Second Course in Statistics — Regression Analysis*, Prentice Hall, Upper Saddle River.

Milliken, G. A. and Johnson, D. E. (1989). *Analysis of Messy Data. Nonreplicated Experiments*. Chapmann and Hall, London.

Milliken, G. A. and Johnson, D. E. (1997). *Analysis of Messy Data. Designed Experiments*. Chapmann and Hall, London.

Milliken, G. A. and Johnson, D. E. (2001). *Analysis of Messy Data. Analysis of Covariance*. Chapmann and Hall, London.

Pestana, D. D. e Velosa, S. F. (2002). *Introdução à Probabilidade e à Estatística*, Fundação Calouste Gulbenkian, Lisboa.

Shepard, M. U., Richards, V. A., and Berkowitz, R. L. (1982). An evaluation of two equations for predicting fetal weight by ultrasound, *Am. J. Obstet. Gynecol.* **142**, 48.

Schweigert, W. A. (1994). *Research Methods and Statistics for Psychology*, Brooks/Cole Publ. Co., Pacific Grove, Cal.

Bartoon

Luís Afonso, *Público, 28-05-*2002
(Publicação gentilmente autorizada pelo autor)