

# Regressão não-linear utilizando a ferramenta Solver<sup>®</sup> do Microsoft Excel<sup>®</sup>

Eduardo Esteves

9 de Julho de 2010

Instituto Superior de Engenharia, Universidade do Algarve, Campus da Penha, 8005-139 Faro e Centro de Ciências do Mar do Algarve, CIMAR - Laboratório Associado, Campus de Gambelas, 8005-139 Faro; E-mail: eesteves@ualg.pt URL <http://w3.ualg.pt/~eesteves>

## Resumo

Em vários domínios do conhecimento científico são usados modelos matemáticos para descrever um conjunto de dados empíricos. Se no caso dos modelos lineares, a análise de regressão para obter os parâmetros dos modelos é “simples” e vulgarizada, é mais difícil ajustar aos dados funções (matematicamente) mais “complicadas”, p.ex. modelos não-lineares. Neste artigo<sup>1</sup>, pretende-se 1) apresentar os conceitos fundamentais da análise de regressão não-linear, 2) descrever, usando um exemplo, a utilização da ferramenta Solver<sup>®</sup> do Microsoft Excel<sup>®</sup> para analisar problemas cujo objectivo é “descrever” relações estatísticas (não-lineares) entre variáveis, e 3) compilar os “problemas” identificados com a utilização do Excel<sup>®</sup> (incluindo o Solver<sup>®</sup>) como ferramenta de análise estatística.

Palavras chave: Regressão não-linear; Solver<sup>®</sup>; Microsoft Excel<sup>®</sup>

## Introdução

Em vários domínios do conhecimento científico, e.g. biologia, física, química, engenharia, economia, etc., são usados modelos matemáticos para descrever um conjunto de dados empíricos, genericamente  $y = f(x)$ , em que  $y$  é a variável dependente,  $x$  é a variável “independente” – controlada pelo investigador – e  $f(x)$  é uma função que pode incluir um ou mais parâmetros  $\theta$ . Quanto melhor  $f(x)$  se ajustar aos dados, mais “rigorosamente” descreverá aquela relação Brown (2001). No caso dos modelos lineares, do tipo  $y = a + bx$ , a análise de regressão é “simples” e vulgarizada, pela utilização de “calculadoras científicas” e/ou computadores pessoais e porque constitui tópico de estudo da maioria das disciplinas de Estatística do(s) primeiro(s) dos cursos de ensino superior. Pelo contrário, é mais difícil ajustar aos dados funções (matematicamente) mais “complicadas”, e.g.  $y = aexp(bx)$  – modelos não-lineares Brown (2001). A propósito, modelos do tipo  $y = a + bx + cx^2 + dx^3$  (polinomial cúbica), que graficamente correspondem a curvas, são, de facto, lineares nos termos (coeficientes), *i.e.* porque nenhum parâmetro aparece como expoente ou a multiplicar/dividir por outro parâmetro, e podem, por isso, ser estudados através de análise de regressão linear múltipla

---

<sup>1</sup>Citação recomendada: Esteves, E. (2009) Regressão não-linear utilizando a ferramenta Solver<sup>®</sup> do Microsoft Excel<sup>®</sup>. Instituto Superior de Engenharia da Universidade do Algarve, Faro, 13 p. [disponível em <https://sapiencia.ualg.pt/>]

RLM (neste caso,  $\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$ ), fazendo  $x_1 = x$ ,  $x_2 = x^2$  e  $x_3 = x^3$  (cf. Neter et al. (1996); Motulsky & Christopoulos (2004); Cabral & Guimarães (1997); Fonseca (2001) para RLM). A padronização dos dados poderá evitar problemas de (multi)colinearidade (i.e. correlação) entre os termos  $x_1$ ,  $x_2$ , etc.

Tradicionalmente, transformam-se as variáveis de alguns modelos não-lineares de forma a linearizar a relação e a permitir a sua análise através da regressão linear. Por exemplo, se se logaritimizarem (através de  $\ln$ ) ambos os termos em  $y = aexp(bx)$ , obtém-se  $\ln(y) = \ln(a) + bx$ , ou seja uma “equação da recta”:  $y' = a' + bx$ . Actualmente, a capacidade de processamento dos PC e a disponibilidade de software (sejam folhas de cálculo, e.g. Microsoft Excel®), ou programas comerciais dedicados, e.g. SPSS®) possibilita o ajuste de funções não-lineares directamente aos dados - análise de regressão não-linear.

Neste artigo pretende-se 1) apresentar os conceitos fundamentais da análise de regressão não-linear (cf. Motulsky & Christopoulos, 2004; Motulsky & Ransnas, 1987; Smyth, 2001, para descrições pormenorizadas deste tópico), 2) descrever, usando exemplos, a utilização da ferramenta Solver® do Microsoft Excel® para analisar problemas cujo objectivo é “descrever” relações estatísticas (não-lineares) entre variáveis, e 3) compilar os “problemas” identificados com a utilização do Excel® (incluindo o Solver®) como ferramenta de análise estatística.

## Regressão não-linear

Neste artigo, será abordada a situação mais comum/simples, i.e. casos com uma única variável independente,  $X$ , controlada pelo investigador e cujos resultados foram obtidos sem erro (ou com erro negligenciável) e uma variável dependente,  $Y$ , obtida experimentalmente e com distribuição (de probabilidades) normal. A equação que descreve a relação (estatística) entre essas variáveis pode ser generalizada para

$$y = f(x, \theta) + \varepsilon \quad (1)$$

em que  $f(x, \theta)$  é uma função com um ou mais parâmetros  $\theta$ , e  $\varepsilon$  são os erros aleatórios, independentes e com distribuição normal. Outra formulação, equivalente, é  $\hat{y} = f(x)$  (em que  $\hat{y}$  se lê valor esperado, ou estimado, de  $y$ ). Pretende-se ajustar a função  $f(x)$  aos dados empíricos de forma a minimizar os erros  $\varepsilon = (y_i - \hat{y}_i)$ . De facto, o objectivo é estimar o(s) parâmetro(s) da função  $f(x)$  de modo a minimizar a soma dos quadrados dos erros, SQE,

$$SQE = \sum (y_i - \hat{y}_i)^2 \quad (2)$$

ou seja, minimizar as distâncias verticais entre os pontos (dados) e a curva (modelo) (Figura 1). Este procedimento é designado método dos mínimos quadrados (MMQ).

No caso de funções (ou modelos) não-lineares, não é possível obter as estimativas dos parâmetros num único passo, como no caso de regressões lineares<sup>2</sup>. Sendo assim, a SQE é minimizada através dum processo iterativo (cíclico) utilizando um algoritmo<sup>3</sup> apropriado que necessita dos valores

<sup>2</sup>No caso da regressão linear, é possível obter a solução analítica para o sistema de equações (ditas) normais cujo objectivo é estimar os parâmetros do modelo de regressão (ou seja, recorrendo ao cálculo, trigonometria ou outras técnicas matemáticas é possível estudar o comportamento de uma dada função e obter a(s) soluções dessa função para determinadas condições). Pelo contrário, os problemas de natureza não-linear são, na maior parte dos casos, demasiado complexos para serem estudados analiticamente e, por isso, utilizam-se métodos numéricos (e.g. algoritmos) para obter soluções (muito) aproximadas.

<sup>3</sup>Um algoritmo é uma sequência não-ambígua de instruções que é executada até que determinada condição se

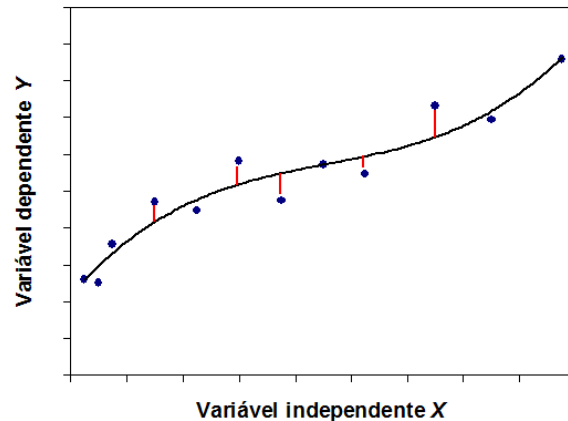


Figura 1: Ilustração do conceito de desvio/erro/resíduo (sensu método dos mínimos quadrados) – alguns erros estão assinalados com linhas vermelhas verticais.

iniciais dos parâmetros  $\theta_0$  (Bowen & Jerman, 1995).

## Escolha do modelo matemático

A equação que relaciona  $Y$  e  $X$ ,  $f(x, \theta)$ , tem de ser escolhida pelo investigador de acordo com a teoria, isto é, porque descreve uma determinada hipótese física, química, molecular, biológica, ecológica, etc. (Motuslky & Ransnas, 1987). Em contraponto a esta perspectiva determinística da análise de regressão não-linear, existem metodologias (estatísticas) para descrever a relação entre variáveis sem ser necessário admitir um dado modelo subjacente aos dados (por exemplo, funções polinomiais ou funções *spline*) – perspectiva empirista – que por razões de espaço e complexidade não se abordarão neste artigo. Deve, ainda, considerar-se a hipótese de estabelecer restrições às estimativas dos parâmetros (e.g.  $\theta_1 < 0$  e  $\theta_2 \geq 0$ ) ou de se ponderarem os resultados uma vez que, por exemplo, a sua variabilidade é proporcional à respectiva magnitude (Bowen & Jerman, 1995).

## Estimativas iniciais

As estimativas iniciais do(s) parâmetro(s), podem ou devem ser especificadas tendo em consideração a experiência do investigador, eventuais análises preliminares ou simplesmente com base num palpite. O conhecimento do (significado dos parâmetros do) modelo que se pretende ajustar facilita a escolha dos valores iniciais do(s) respectivo(s) parâmetros. Uma escolha “infeliz” desses valores pode ter várias consequências, nomeadamente 1) aumentar, desnecessariamente, o tempo de computação e/ou cálculo, 2) impedir a convergência do algoritmo e, por conseguinte, a obtenção duma solução, ou 3) dar origem a uma solução errada, uma vez que o algoritmo convergiu num valor mínimo local e não geral. A escolha dos valores iniciais é (mais) influente e, por isso, mais importante nos casos em que os modelos incluem muitos parâmetros (Motuslky & Ransnas, 1987).

verifique. Mais especificamente, em matemática constitui o conjunto de processos (e símbolos que os representam) para efectuar um cálculo (Wikipédia, 2009). Donald Knuth propõe que "An algorithm is a set of rules for getting a specific output from a specific input. Each step must be so precisely defined that it can be translated into a computer language and executed by machine". Um algoritmo não tem que ser necessariamente executado por um computador, pode ser executado "à mão" por uma pessoa (Fernando Lobo, Departamento de Engenharia Electrónica e Informática da Faculdade de Ciências e Tecnologia da Universidade do Algarve, <http://www.deei.fct.ualg.pt/PI/> em 28 de Junho de 2007).

## Escolha do Algoritmo

À excepção do método Simplex, todos os algoritmos usados para análise de regressão não-linear (e.g. Gauss–Newton, Marquardt–Levenberg, etc.) calculam, repetidamente (em cada iteração), a derivada (“declive”) de  $Y$  em relação a todos os parâmetros (Motulsky & Ransnas, 1987). Felizmente, muitas aplicações informáticas de uso generalizado integram funções de optimização poderosas e relativamente flexíveis que podem ser programadas para minimizar a SQE num processo iterativo (Bowen & Jerman, 1995). Por exemplo, a ferramenta de optimização Solver<sup>®</sup> do Microsoft Excel<sup>®</sup> – cuja utilização se pormenoriza mais adiante neste artigo – utiliza um algoritmo “original”, o *Generalized Reduced Gradient (GRG2) nonlinear optimization code*<sup>4</sup>. No entanto, todos os métodos referidos possuem propriedades similares, designadamente a introdução de valores iniciais para os parâmetros, e deveriam providenciar resultados iguais quando usados para dado um conjunto de dados (Brown, 2001).

## Avaliação da bondade do ajuste

Depois de ajustar um modelo a um conjunto de dados, deve avaliar-se a “qualidade” do ajuste. A bondade do ajuste relaciona a variabilidade dos pontos (dados) em relação à curva (modelo) e a variância dos resíduos.

O modo mais simples de “testar” a bondade do ajuste é representar graficamente os dados e o modelo ajustado, de modo a verificar (visualmente) se os parâmetros obtidos numericamente descrevem, de facto, a relação entre variáveis. Por outro lado, o gráfico (eventualmente ligado de forma dinâmica aos dados e às soluções) permite descortinar se o modelo escolhido é adequado e/ou se os resultados da análise (soluções propostas pelo algoritmo) são mínimos locais ou gerais.

Os testes estatísticos para avaliar a qualidade de ajustamento servem para complementar a inspecção visual e permitem decidir, formalmente, se se deve aceitar ou rejeitar um determinado modelo. Infelizmente, as funções de optimização incluídas nas folhas-de-cálculo não providenciam estatísticas acerca dos valores finais do(s) parâmetro(s), por exemplo o erro-padrão das estimativas ou o “grau” de correlação entre os parâmetros. Esta limitação não é partilhada pelas aplicações “dedicadas” (e.g. R)<sup>5</sup> ou pelos pacotes comerciais (e.g. SPSS<sup>®</sup>). Ainda assim, de acordo com (Bowen & Jerman, 1995; Motulsky & Ransnas, 1987) podem analisar-se os resultados do ajustamento com folhas-de-cálculo através de: 1) um diagrama dos resíduos *vs.* os valores esperados de  $Y$  (se o modelo for adequado, os resíduos representam apenas o erro experimental e não apresentam tendências ou padrões); 2) gráfico de probabilidade normal dos resíduos (ou “normal probability plot of residuals”) para verificar se os resíduos se distribuem “normalmente” – também se poderia testar a normalidade dos resíduos através do teste de Kolmogorov-Smirnov (*sensu* Lillefors)<sup>6</sup>; 3) um teste de sequências (ver adiante); e 4) do coeficiente de determinação. Mais ainda, Brown

---

<sup>4</sup>©Frontline System Inc., <http://www.frontsys.com>, desenvolvido por Leon Lasdon, da University of Texas at Austin, e Allan Waren, da Cleveland State University.

<sup>5</sup>O R (R Development Core Team, 2006) é ao mesmo tempo uma linguagem de programação e um ambiente para computação estatística e gráfica. Trata-se de uma linguagem de programação especializada em computação com dados. Uma das suas principais características é o seu carácter gratuito e a sua disponibilidade para uma gama bastante variada de sistemas operativos (*vd.* <http://www.r-project.org>). Apesar de gratuito, o R é uma ferramenta bastante poderosa com boas capacidades ao nível da programação e um conjunto bastante vasto (e em constante crescimento) de *packages* que acrescentam bastantes potencialidades à já poderosa versão base do R (Torgo, 2006).

<sup>6</sup>Para informação adicional sobre este teste consultar Esteves & Sousa (2007) [disponível no URL do autor, em <http://w3.ualg.pt/~eesteves>]

(2001), concretizando uma sugestão de Motuslky & Ransnas (1987), propõe que, a partir do erro-padrão de  $Y$ ,  $se(Y)$ , se obtenha o intervalo de confiança do valor esperado de  $Y$ ,  $\hat{y}$ , e que esta informação seja representada em conjunto com os dados e com o modelo ajustado. O erro-padrão de  $Y$  calcula-se através de

$$se(Y) = \sqrt{\frac{\sum(y_i - \hat{y})^2}{n - p}} \quad (3)$$

em que  $n$  é o nº de pontos e  $p$  o nº de parâmetros do modelo. Então, o intervalo de confiança do valor esperado de  $Y$  para um dado  $X$ ,  $x_0$ , é definido como

$$\hat{y} \pm t \cdot se(Y) \quad (4)$$

em que  $t$  se obtém da distribuição de probabilidade de t-Student (distribuição de probabilidades derivada da distribuição normal) para  $n - p$  graus de liberdade (g.l.) e uma probabilidade  $1 - \alpha/2$  (no Excel® em Português, usar =INVT( $\alpha$ ,g.l.)). Poderá afirmar-se com  $(1 - \alpha) \times 100\%$  de confiança que o “verdadeiro” valor esperado de  $Y$  para  $x_0$  se encontra no intervalo de confiança obtido.

De acordo com Motuslky & Ransnas (1987) o teste de sequências (ou “runs test” nos manuais anglófonos) permite testar, de forma simples e robusta, se os pontos (dados) diferem sistematicamente da curva ajustada (modelo) complementando a informação do diagrama dos resíduos *vs.* valores observados de  $X$ . Contudo, Conover (1999) alerta para a reduzida potência deste teste. Uma sequência é definida como um conjunto de resultados (resíduos, neste caso) cujo sinal é idêntico (negativo ou positivo). Por exemplo, no conjunto de  $n=17$  resíduos cujos sinais são  $+ + - - - - - + + - + +$  existem 5 sequências ( $r=5$ ). No caso de  $n \geq 20$ , é possível usar uma estatística de teste (e.t.),  $Z_R$ , para testar a hipótese de aleatoriedade da distribuição dos pontos em relação à curva:

$$Z_R = (r - \mu_R) / \sigma_R \quad (5)$$

em que  $r$  é o nº de sequências,  $\mu_R = 1 - 2n_A n_B / 2$ ,  $\sigma_R = \sqrt{[2n_A n_B (2n_A n_B - n)] / [n^2 (n - 1)]}$ , e  $n = n_A + n_B$  (por convenção  $n_B \leq n_A$ ; para o exemplo anterior,  $n_A$  é o nº de resultados negativos (-) e  $n_B$  é o nº de resultados positivos (+)). Se  $|Z_R| > Z$  (ou então  $\text{valor-p} < \alpha$ )<sup>7</sup>, poderá afirmar-se com  $(1 - \alpha) \times 100\%$  de confiança que a curva não se desvia sistematicamente dos pontos (para obter  $Z$  no Excel® em português, usar =INV.NORMP( $\alpha$ )). Nos casos em que  $n < 20$ , podem utilizar-se tabelas apropriadas (e.g. Tabela 1) para decidir acerca da aleatoriedade directamente a partir de  $r$  (no caso, uma vez que  $r = 5$ , para  $n_A = 9$  e  $n_B = 8$ , e os valores na tabela são 5 e 14; i.e. não se rejeita a hipótese (nula) de que a curva não se desvia sistematicamente dos pontos).

O coeficiente de determinação é muito utilizado no caso da análise de regressão linear e representa a fracção da variação total (de  $Y$ ) que é explicada por  $X$  de acordo com o modelo matemático ajustado aos dados. O coeficiente calcula-se através de

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \quad (6)$$

---

<sup>7</sup>O valor de prova ou probabilidade de significância do teste, *valor-p* ou *p-value* (em inglês), constitui uma medida do grau com que os dados amostrais contradizem a hipótese nula. De facto, corresponde à probabilidade da estatística de teste tomar um valor igual ou mais extremo do que aquele que é observado. Como é evidente, quanto menor for o *valor-p* maior será o grau com que a hipótese nula é contradita. Nestes termos pode-se pensar que quando se utiliza o *valor-p* não é necessário especificar a região de rejeição (nem fixar previamente o valor de  $\alpha$ ) num teste de hipóteses.

Tabela 1: Valores críticos do Teste de Sequências para um nível de significância de 5%. Rejeitar a hipótese nula (aleatoriedade dos resultados) se o  $n^o$  de sequências,  $r$ , for menor ou igual ao que o 1<sup>o</sup> valor ou igual ou maior do que o 2<sup>o</sup> valor na célula da tabela. Por exemplo, se  $n_B = 10$ ,  $n_A = 11$  e  $r=8$ , obtêm-se os  $n.º$  6 e 17. Uma vez que 8 está nesse intervalo de valores, não se rejeita a hipótese nula (adaptado de Roger Bove, West Chester University of Pennsylvania (EUA), acedido em 4 de Junho de 2007, <http://courses.wcupa.edu/rbove/eco252/252runstable.doc>).

$n$	$A$																			
$B$	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
2											2	2	2	2	2	2	2	2	2	
3					2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	
4				2	2	2	3	3	3	3	3	3	3	3	3	4	4	4	4	
5		2	2	2	3	3	3	3	3	4	4	4	4	4	4	4	4	5	5	
6		2	2	3	3	3	3	4	4	4	4	4	5	5	5	5	5	6	6	
7		2	2	3	3	3	4	4	4	5	5	5	5	5	6	6	6	6	6	
8		2	3	3	3	4	4	5	5	5	6	6	6	6	6	6	7	7	7	
9		2	3	3	4	4	4	5	5	5	6	6	6	7	7	7	7	8	8	
10		2	3	3	4	5	5	5	6	6	6	7	7	7	8	8	8	8	9	
11		2	3	4	4	5	5	6	6	7	7	7	8	8	8	9	9	9	9	
12	2	2	3	4	4	5	6	6	7	7	7	8	8	8	9	9	9	10	10	
13	2	2	3	4	5	5	6	6	7	7	8	8	9	9	9	10	10	10	10	
14	2	2	3	4	5	5	6	7	7	8	8	9	9	9	10	10	10	11	11	
15	2	3	3	4	5	6	6	7	7	8	8	9	9	10	10	11	11	11	12	
16	2	3	4	4	5	6	6	7	8	8	9	9	10	10	11	11	11	12	12	
17	2	3	4	4	5	6	7	7	8	9	9	10	10	11	11	11	12	12	13	
18	2	3	4	5	5	6	7	8	8	9	9	10	2	11	11	12	12	13	13	
19	2	3	4	5	6	6	7	8	8	9	10	10	11	11	12	12	13	13	13	
20	2	3	4	5	2	6	7	8	9	9	10	10	11	12	12	13	13	13	14	

Tradicionalmente,  $r^2$  aplica-se apenas a relações lineares (Motuslky & Ransnas, 1987). Embora seja possível calcular  $R^2$  após o ajuste dum modelo não-linear (cf. Cameron & Windmeijer, 1997) alguns autores (Douglas Bates, University of Wisconsin (E.U.A.), e Bill Venables, University of Adelaide (Austrália), comunicação pessoal) advertem para as dificuldades com o cumprimento de alguns pressupostos subjacentes àquele coeficiente. Sendo assim, o resultado de  $R^2$  deve ser usado com cuidado<sup>8</sup>.

Com objectivo de (rápida e facilmente) verificar os pressupostos da análise regressão, propõe-se a representação gráfica de informação relativa aos resíduos<sup>9</sup> do modelo ajustado,  $e_i = y_i - \hat{y}_i$  (em que  $\hat{y}_i = f(x_i, \theta)$ ), designadamente:

- A representação dos resíduos num gráfico cartesiano ou “NORMAL PROBABILITY PLOT OF RESIDUALS”;
- Representação gráfica dos RESÍDUOS *versus* VALORES AJUSTADOS (ESPERADOS), ou *versus* a variável independente.

A representação dos resíduos estandardizados,  $e'_i = e_i/s$  (em que  $s = \sqrt{SQ_E/(n-p)}$ ) num gráfico (cartesiano) de probabilidade normal (*vulgo* “normal probability plot of residuals” - *npp*) permite avaliar a normalidade dos resíduos. Sendo o modelo correcto, os  $n$  resíduos deverão representar  $n$  observações de uma variável aleatória com  $N(0, \sigma^2)$  e a sua representação (em papel de probabilidade normal, i.e. através de um *npp*) deverá corresponder, aproximadamente, a uma linha recta. O *npp* pode obter-se no Microsoft Excel<sup>®</sup> (ou noutra folha-de-cálculo) como se descreve mais adiante no Exemplo. No gráfico dos resíduos estandardizados  $e'_i$  *vs* os valores ajustados (ou estimados pelo modelo)  $\hat{Y}$ , ou então, contra a variável independente  $X$ , representam-se, respectivamente, os pares ordenados  $(\hat{y}_i, e'_i)$  ou  $(x_i, e'_i)$ ,  $i = 1, \dots, n$ . A forma da “nuvem” de pontos representada fornece, de modo empírico, informação (complementar) específica sobre o modelo ajustado. Vejam-se, em separado, essas situações em Esteves (2009).

## Comparação de dois modelos ajustados a um conjunto de dados

É possível comparar o ajuste de dois modelos distintos a um dado conjunto de dados através das respectivas SQE. Essa comparação é válida desde que não se tenham transformado as variáveis  $Y$  e  $X$ . A comparação entre modelos não deve ser exclusivamente estatística, considerações (teóricas) acerca da consistência com os dados e da adequação ao problema dos modelos também são muito importantes (Motuslky & Ransnas, 1987).

De acordo com Motuslky & Ransnas (1987), para comparar dois modelos, pode-se usar um teste de  $F$  cuja e.t. é

$$F_0 = \frac{(SQ_1 - SQ_2)/(gl_1 - gl_2)}{SQ_2/gl_2} \quad (7)$$

em que  $SQ$  se refere à SQE,  $gl$  são o  $n^o$  de g.l. (i.e.  $n^o$  pontos menos o  $n^o$  de parâmetros) e os índices 1 e 2 dizem respeito ao modelo com menor e maior  $n^o$  de parâmetros, respectivamente. O

<sup>8</sup>Decidi distinguir o coeficiente de determinação que se determina para a regressão linear daquele que se obtém para regressão não-linear através de  $r^2$  e  $R^2$ , respectivamente.

<sup>9</sup>Vários autores, e.g. Neter et al. (1996), propõem a utilização de uma das “versões” estandardizadas dos resíduos, por exemplo  $e'_i = e_i/s$  em que  $s = \sqrt{SQ_E/(n-2)}$  (este é o “tipo” de resíduos que se utiliza no Exemplo que se apresenta mais adiante).

valor- $p$  (para a e.t.  $F_0$ ) obtém-se a partir da distribuição de probabilidades  $F$  de Snedecor para uma probabilidade  $(1 - \alpha)$  e para  $(gl_1 - gl_2)$  e  $gl_2$  graus de liberdade (no Excel<sup>®</sup> em português, usar =DISTF( $F_0;gl_1-gl_2;gl_2$ )). Se valor- $p < \alpha$ , então o modelo mais complexo (com maior  $n^o$  de parâmetros) ajusta-se significativamente melhor aos dados do que o modelo mais simples.

## Exemplo

Para concretizar o que se abordou nas secções anteriores apresenta-se, de seguida, um exemplo (relacionado com a prática da Engenharia Alimentar) da utilização do Microsoft Excel<sup>®</sup> 2003 Solver<sup>®</sup> para análise de regressão não-linear. Oportunamente, outros exemplos poderão ser acesados no sítio electrónico do autor. O que se apresenta a seguir também se aplicará (com pequenas adaptações) à ferramenta *Optimization Solver* em desenvolvimento para o OpenOffice Calc<sup>10</sup>.

Em Análise Sensorial, "disciplina da Ciência usada para evocar, medir, analisar e interpretar as reacções às características dos alimentos e materiais tal como são percebidas pelos sentidos da visão, olfacto, paladar, tacto e audição", é útil determinar a intensidade dum estímulo que é detectável/reconhecível pelos consumidores (dum produto alimentar), i.e. limiar. O "método intermédio", proposto pela American Society for Testing and Materials (ASTM E1432) e pela International Organization for Standardization (ISO 13301), recorre à repetição de provas sensoriais simples em que se solicita aos consumidores (provedores) que provem amostras de produtos alimentares cuja concentração/intensidade duma característica sensorial é crescente. Para cada provedor  $i$ , a relação entre a proporção do  $n^o$  de respostas em que detectou aquele atributo sensorial ("respostas correctas/positivas") e a intensidade do estímulo (concentração) é descrita (teoricamente) por um modelo "dose-resposta" (logístico modificado):

$$P_i = \frac{\frac{1}{3} + \exp(B(T - \log(x)))}{1 + \exp(B(T - \log(x)))} \quad (8)$$

em que  $P_i$  é a proporção de respostas positivas do provedor  $i$ ,  $\log(x)$  é o logaritmo do valor do estímulo  $x$ ,  $B$  é o "declive" e  $T$  é o limiar para o provedor  $i$  (em  $\log(x)$ ). A partir das estimativas de  $T$  para um conjunto de  $n$  provedores pode obter-se o limiar do grupo através de  $L_g = 10^G$ , em que  $G = \frac{1}{n} \sum T$ .

Numa folha-de-cálculo do Excel<sup>®</sup> (Figura 2) podem incluir-se os dados experimentais, neste caso de um trabalho em que se pretendia determinar o limiar de detecção da gaseificação duma bebida (conteúdo  $CO_2$ , em ppm) usando o "método intermédio" descrito anteriormente. Das quatro provas sensoriais (repetições) realizadas para cada uma das nove concentrações de  $CO_2$  testadas, obteve-se a proporção de "respostas correctas/positivas" (Prop). Nas células B3:B10 e C3:C10 encontram-se os valores experimentais da concentração de  $CO_2$  (C02) e da proporção de respostas correctas (Prop), respectivamente.

Considerando que o modelo logístico (equação 8) é aplicável, obtém-se os valores esperados da proporção de respostas correctas (Prop<sup>^</sup>, nas células E3:E10) usando a equação descrita na barra de fórmulas do Excel<sup>®</sup> como =(1/3+EXP(\$I\$3\*((\$I\$4-LOG(B3)))))/(1+EXP(\$I\$3\*((\$I\$4-LOG(B3)))) com os valores iniciais dos parâmetros  $B$  e  $T$ , respectivamente -10,00 e 2,50 (inseridos nas células I3 e I4).

A partir dos resíduos (nas células F3:F10, e.g. =C3-E3), calcula-se: a *SQE* (célula I7, através da equação 2 usando {=SOMA((C\$3:C\$10-E\$3:E\$10)^2)}); o erro-padrão de  $Y$  (na célula I9,

<sup>10</sup>vd. <http://code.google.com/p/scsolver> (consultado em 26 de Março de 2009).

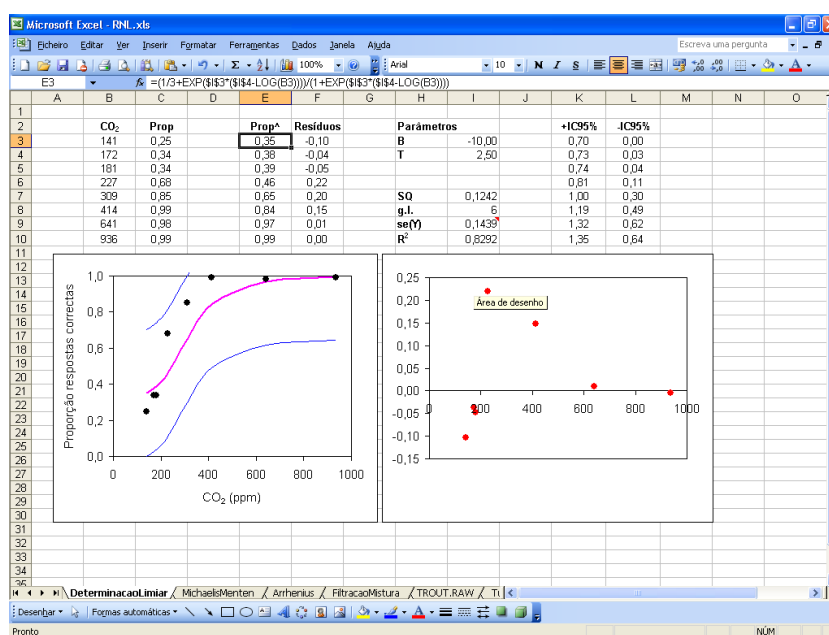


Figura 2: Folha-de-cálculo do Excel<sup>®</sup> preparada para a análise de regressão não-linear. Nas células B3:B10 e C3:C10 encontram-se os valores experimentais da concentração de  $CO_2$  ( $C_{O2}$ , em ppm) e da proporção de respostas correctas (Prop), respectivamente. Na coluna E (células E3:E10) temos os valores esperados ( $Prop^{\wedge}$ ) obtidos pelo modelo logístico (equação 8), cujos valores iniciais dos parâmetros  $B$  e  $T$  foram inseridos nas células I3 e I4, respectivamente -10,00 e 2,50. Também se incluem, os resíduos, a soma dos quadrados dos erros (SQ), n<sup>o</sup> de g.l. (g.l.), erro-padrão de  $Y$  (i.e.  $se(Y)$ ) e o coeficiente de determinação ( $R^2$ ) respectivamente. Nas colunas K e L encontram-se limites (superior e inferior) dos intervalos de confiança de  $Y$  (+IC95% e -IC95%). Os diagramas (gráficos) de dispersão (ligados dinamicamente aos dados e às soluções) dizem respeito aos dados, modelo ajustado e respectivo intervalo de confiança (à esquerda) e aos resíduos *vs.* valores observados de  $CO_2$ , *vs.* valores esperados e gráfico de probabilidade normal (em baixo).

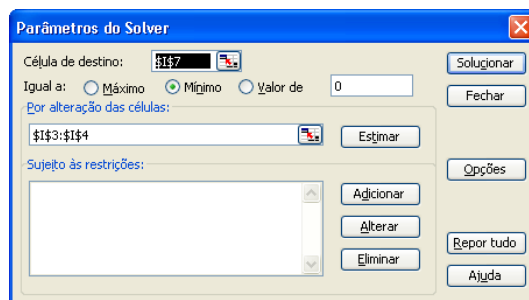


Figura 3: Caixa de diálogo inicial da ferramenta Solver®. É necessário indicar a **Célula de destino**, o “tipo” de otimização (em **Igual a:**) e a «localização» (endereço ou referência das células) dos parâmetros que se farão variar durante o processo de otimização (neste caso,  $B$  e  $T$ , no campo **Por alteração das células**). É possível, e por vezes desejável, sujeitar as estimativas/soluções a restrições (no campo **Sujeito às restrições**).

através de  $\{=RAIZQ(SOMA((\$C\$3:\$C\$10-\$E\$3:\$E\$10)^2)/\$I\$8)\}$  que reproduz a equação 3); e o coeficiente de determinação ( $R^2$ , na célula I10, que se obtém com  $\{=1-SOMA((C3:C10-E3:E10)^2)/(SOMA((C3:C10-MÉDIA(C3:C10))^2))\}$ , i.e. equação 6). Nestas fórmulas, as chavetas  $\{\}$  indicam uma “fórmula de matriz” e obtêm-se fazendo **Control+Shift+Enter** em vez de simplesmente fazer **Enter** após introduzir a fórmula.

Nas colunas K e L encontram-se os limites (superior e inferior) dos intervalos de confiança de  $Y$  (+IC95% e -IC95%) que, por exemplo, se determinam para  $C02=141$  (*vide* célula E3) através de  $=E3+INVT(0,05; \$I\$8)*\$I\$9$  e de  $=E3-INVT(0,05; \$I\$8)*\$I\$9$ .

Os diagramas (gráficos) de dispersão (ligados dinamicamente aos dados e às soluções) dizem respeito aos dados, modelo ajustado e respectivo intervalo de confiança (à esquerda) e aos resíduos *vs.* valores observados de  $C02$  e *vs.* valores esperados de **Prop** (à direita) e o gráfico de probabilidade normal (em baixo). Os gráficos apresentados na Fig. 2 foram editados/formatados.

Depois do procedimento que se descreve acima, pode-se iniciar o processo de estimação dos parâmetros do modelo de regressão e estatísticas associadas. A ferramenta Solver® é acedida pelo menu **Ferramentas** do Excel®. Caso não conste da lista de ferramentas, poderá ser instalada seleccionando **Suplemento Solver** na caixa de diálogo que se abre através de **Ferramentas>Suplementos...** Se a opção **Suplemento Solver** não surgir nesta caixa de diálogo, então será necessário instalar software recorrendo ao CD de instalação do Microsoft Office®.

Nas Figuras 3 a 5 apresentam-se e descrevem-se os principais passos (e opções) do procedimento de análise de regressão não-linear usando o Solver®. Necessariamente, devem indicar-se: a **célula destino** (neste caso, a expressão a minimizar é a SQE que está na célula I7), o “tipo” de otimização (maximização, minimização ou atingir determinado valor-alvo, nas opções de **Igual a:**) e as células que contêm os valores (iniciais) dos parâmetros a estimar,  $B$  e  $T$ , e que se farão variar durante o processo de otimização (neste caso, I3:I4 no campo **Por alteração das células**) (Figura 3).

De entre as opções de otimização disponíveis (através da caixa de diálogo **Opções**), destacam-se: a **Precisão** (por defeito  $1 \times 10^{-6}$ , controla a precisão das soluções tomando em consideração as diferenças das soluções relativamente às restrições impostas); a **Tolerância** (diferença admissível, em %, entre o valor na célula destino, neste caso SQE, e a solução óptima; apenas aplicável nos casos que envolvem “restrições inteiras”); a **Convergência** (valor que controla o processo de iteração; se, nas últimas 5 iterações, a alteração relativa na célula destino for inferior ao valor

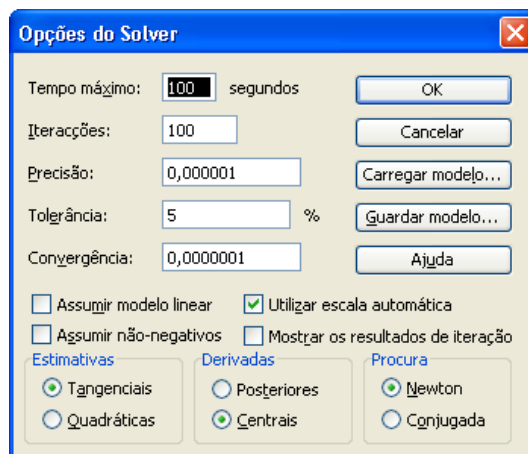


Figura 4: Caixa de diálogo Opções da ferramenta Solver® permite que o utilizador defina algumas opções relativas ao processo de optimização.

indicado terminam as iterações); **Assumir modelo linear** (marcar nos casos que envolvem modelos lineares); **Utilizar escala automática** (marcar esta opção se existirem grandes diferenças de magnitude entre  $y$  e  $x$ ); **Estimativas** (tangenciais ou quadráticas, i.e. modo de obter estimativas dos coeficientes/parâmetros do modelo: extrapolação linear do vector tangente ou a partir do mínimo (ou máximo) duma função quadrática ajustada à estimativa actual, respectivamente); **Derivadas** (posteriores ou centrais, ou seja, o método de diferenciação usado pelo algoritmo para obter as derivadas parciais dos objectivos e das funções/restrições); e **Procura** (especifica o algoritmo usado nas iterações/optimização, **Newton** – por defeito – ou **conjugada** – usando o *GRG2*). McCullough & Wilson (2005) aconselham a utilização de escala automática, a definição da convergência igual a  $1E-7$  (i.e.  $1 \times 10^{-7}$  ou 0,0000001), a obtenção de estimativas tangenciais (mais “lentas” mas mais rigorosas) e o cálculo de derivadas centrais (na caixa de diálogo Opções, Figura 4).

Caso o nº de iterações máximo (que por defeito é igual a 100) seja excedido sem que se tenha chegado a uma solução, surge um caixa de diálogo de aviso (Figura 5, em cima) que permite parar ou continuar o processo iterativo. Logo após a convergência do processo de optimização, uma caixa de diálogo surge no ecrã (Figura 5, em baixo). Neste caso, é possível **Aceitar a solução proposta pelo Solver** ou **Repor valores originais**, e obter relatórios mais completos das soluções encontradas pelo Solver® (seleccionando as opções da lista à direita).

No final do processo de optimização, e após aceitar a solução proposta pelo Solver® (Figura 5), as estimativas dos parâmetros são  $B=-19,56$  e  $T=2,37$  ( $SQE=0,0259$ ,  $se(Y)=0,0657$  e  $R^2 = 0,9644$ ) (Figura 6). Estas estimativas são coincidentes com as que se obtêm com o SPSS® (pelo algoritmo de Levenberg-Marquardt) ou com o R (Gauss-Newton), partindo dos mesmos valores iniciais. Não é possível usar o teste de sequências em virtude do reduzido nº de resultados, embora nos casos (aproximados) tabelados o limite  $r=2$  dê indicação de que o modelo não parece desviar-se significativamente dos pontos. Os gráficos dos resíduos (standardizados)  $e'_i$  vs. valores observados/esperados de  $Y$  (Fig. 6) parecem indicar que o pressuposto da homogeneidade das variâncias não se verifica. Por outro lado, o gráfico de probabilidade normal dos resíduos standardizados (Fig. 6) revela que os resíduos não se distribuem normalmente. Este último gráfico obteve-se, conforme Frees (1996), a partir dos resíduos standardizados,  $e'_i = e/s$  (em que  $s^2 = SQE/(n-p)$ ), determinando os respectivos nº de ordem (ou “ranks”)  $R_i$  - ou seja, para um dado resíduo  $e'_i$ ,  $R_i/n$  é a fracção

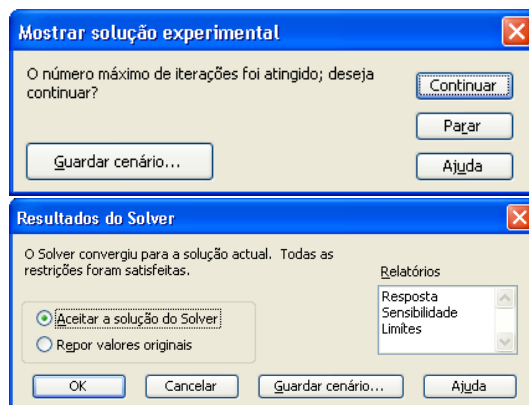


Figura 5: Caixas de diálogo que surgem caso o nº de iterações máximo (por defeito é igual a 100) não tenha permitido chegar a uma solução (em cima) ou logo após a convergência do processo de optimização (em baixo). Neste caso, é possível Aceitar a solução proposta pelo Solver® ou Repor valores originais, e obter relatórios mais completos das soluções encontradas pelo Solver® (seleccionando as opções da lista à direita).

de valores menor ou igual a  $e'_i$ ; por exemplo para o 1º valor através de `=ORDEM(N3;$N$3:$N$10;1)-`, depois determinando o correspondente valor  $z_i$  (“normal score” ou “z score”) para a probabilidade  $p_i = (R_i - 3/8)/(n + 1/4)$ , i.e. - através de `=INV.NORMP((03-0,375)/(8+0,25))` para o 1º valor, e por fim, produzindo o diagrama de dispersão de  $z_i$  vs.  $e'_i$ .

Comparando o modelo “dose-resposta” (equação 8) obtido com um modelo logístico de formulação mais generalizada,  $\hat{y} = C/[1 + Aexp(-Bx)]$  (em que  $\hat{y} = Prop$  e  $x = CO_2$ ), cujos parâmetros estimados foram  $A=66,039$ ,  $B=0,021$  e  $C=0,984$  ( $SQE=0,0089$ ,  $se(Y)=0,0422$  e  $R^2 = 0,9877$ ), verifica-se que este último se ajusta significativamente melhor aos dados ( $F_0 = 9,493$  com  $valor-p=0,0129$ ) - aliás, até a análise gráfica dos resíduos do modelo logístico generalizado (gráficos não apresentados) são “mais favoráveis”.

## Considerações finais

À facilidade de utilização e disseminação do Excel® da Microsoft (e da respectiva ferramenta Solver®) estão associadas algumas dificuldades. De acordo com McCullough & Wilson (2005), o algoritmo usado pelo Solver®, *GRG2*, é «deficiente» uma vez que, em testes com conjuntos de dados-padrão (providenciados pelo American National Institute for Standards and Technology, NIST, em <http://www.itl.nist.gov/div898/strd>) o rigor das soluções foi de zero dígitos em 14 dos 27 problemas analisados. Parte dessa má performance pode ser atribuída às prioridades “esquisitas” nas operações matemáticas, que pode ser ultrapassável em muitos casos (vide Berger, 2007). Heiser (2007), num completo e minucioso artigo *on-line*<sup>11</sup>, identifica as falhas e os problemas do Excel® assim como providencia soluções alternativas e/ou correcções para essas dificuldades. Burns (2007), estende as considerações às folhas-de-cálculo em geral<sup>12</sup>. A ferramenta Solver® do Excel® é aplicável e útil para efeitos pedagógicos ou em casos “simples” (como no exemplo descrito aqui) mas em situações “profissionais” deve usar-se software dedicado, por exemplo SPSS® ou R (a

<sup>11</sup>David Heiser, *Microsoft Excel 2000 and 2003 faults, problems, workarounds and fixes* em <http://www.daheiser.info/excel/frontpage.html>, consultado em 1 de Junho de 2007.

<sup>12</sup>Patrick Burns, *Spreadsheet addiction*, em [http://www.burns-stat.com/pages/Tutor/spreadsheet\\_addiction.html](http://www.burns-stat.com/pages/Tutor/spreadsheet_addiction.html), consultado em 25 de Setembro de 2007.

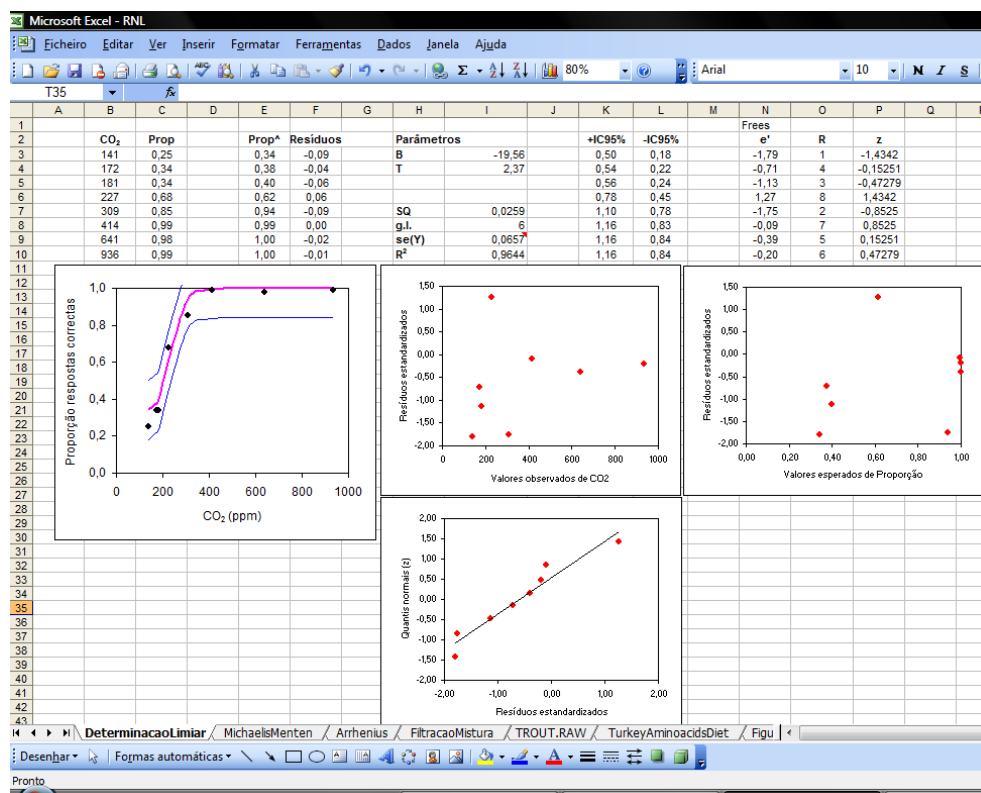


Figura 6: Aspecto da folha-de-cálculo do Excel® no final do processo de otimização (i.e. depois de aceitar a solução proposta pelo Solver®).

utilização desta linguagem de programação e/ou ambiente de computação estatística e gráfica poderá, eventualmente, ser abordada noutro artigo).

## Agradecimentos

Este artigo muito beneficiou dos comentários, correções e sugestões dos colegas Doutora Paula Cabral e Doutor Jaime Aníbal.

## Referências

- Berger, R. (2007). Nonstandard operator precedence in excel. *Computational Statistics and Data Analysis*, 51, 2788–2791.
- Bowen, W. & Jerman, J. (1995). Nonlinear regression using spreadsheets. *Trends in Pharmacological Sciences*, 16, 413–417.
- Brown, A. M. (2001). A step-by-step guide to non-linear regression analysis of experimental data using a microsoft excel spreadsheet. *Computer methods and Programs in Biomedicine*, 65, 191–200.
- Cabral, J. & Guimarães, R. (1997). *Estatística*. Lisboa: McGrawHill.

- Cameron, A. & Windmeijer, F. (1997). An r-squared measure of goodness of fit for some common non-linear regression models. *Journal of Econometrics*, 77, 329–342.
- Conover, W. (1999). *Practical nonparametric statistics*. New York: John Wiley & Sons. Inc.
- Esteves, E. (2009). Análise de regressão linear (simples). Instituto Superior de Engenharia, Universidade do Algarve, Faro, Portugal [disponível em <http://w3.ualg.pt/~eesteves/>].
- Esteves, E. & Sousa, C. (2007). Análise de dados e planeamento experimental (apontamentos da disciplina). Escola Superior de Tecnologia, Universidade do Algarve, Faro.
- Fonseca, J. (2001). *Estatística Matemática. Vols. I e II*. Lisboa: Edições Silabo Lda.
- Frees, E. (1996). *Data analysis using regression models. The business perspective*. New Jersey: Prentice-Hall.
- McCullough, B. & Wilson, B. (2005). On the accuracy of statistical procedures in microsoft excel 2003. *Computational Statistics and Data Analysis*, 49(4), 1244–1252.
- Motulsky, H. & Christopoulos, A. (2004). *Fitting models to biological data using linear and nonlinear regression: a practical guide to curve fitting*. U.S.A.: Oxford University Press. (preview available at <http://books.google.com/books?id=g1FO9pquF3kC&printsec=frontcover&hl=pt-PT> ).
- Motulsky, H. & Ransnas, L. (1987). Fitting curves to data using nonlinear regression: a practical and nonmathematical review. *FASEB J*, 1, 365–374.
- Neter, J., Kutner, M., Nachtsheim, C., & Wasserman, W. (1996). *Applied linear regression models*. Chicago: McGraw-Hill Inc. and Irwin Inc., 3rd ed. edition.
- Smyth, G. K. (2001). *Encyclopedia of Environmetrics. Volume 3*, chapter Nonlinear regression, (pp. 1405–1411). John Wiley & Sons, Chichester.
- Wikipédia (2009). Algoritmo — wikipédia, a enciclopédia livre. [Online; em 26 Março 2009].