REGRESSÃO COM VARIÁVEIS SUJEITAS À IMPRECISÃO: UMA ABORDAGEM PELO MÉTODO DE MONTE CARLO

G.Amaral[1]

RESUMO

O presente trabalho propõe uma abordagem pelo método de Monte Carlo, para o cálculo de linhas de regressão quando os pontos são afetados por erros ou incertezas. O método consiste na definição de um espaço de incerteza para cada ponto, baseada nos erros absolutos em cada dimensão, e amostragem aleatória dentro dele. O conjunto de pontos assim obtidos é utilizado para o cálculo dos parâmetros de regressão, os quais são armazenados pelo programa. Esse procedimento é repetido um número significativo de vezes (tipicamente 1000) e, ao final, os resultados armazenados são descritos estatisticamente. Os resultados assim obtidos foram comparados com aqueles de outros métodos, utilizados rotineiramente para o cálculo de isócronas Rb/Sr, sem discrepâncias significativas. As principais vantagens do método ora proposto são sua simplicidade matemática e possibilidade de visualização e análise estatística dos resultados.

ABSTRACT

A Monte Carlo approach is presented for the problem of least squares fitting of points with correlated errors. The method consists of the definition of an uncertainity space based on absolute error and random sampling for each data point. This procedure is repeated a large number of times (typically 1000 times) and a regression is obtained for each set of data. The results are stored and described statistically at the end of the process. Comparisons were made with results obtained by conventional methods applied to Rb/Sr isochrons, without any significant discrepancies. The main advantage of the method now proposed is its simplicity and the possibility of graphical representation of variability.

INTRODUÇÃO

Um problema ainda não resolvido de modo satisfatório, é o da obtenção de coeficientes em modelos de regressão quando as variáveis envolvidas estão sujeitas a imprecisões. Nestes casos, qualquer que seja o modelo escolhido (duas ou mais variáveis ou grau das funções), ao invés de pontos teremos áreas ou volumes e o problema se resumirá em obter uma função que melhor satisfaça essas áreas ou volumes.

Diversos métodos têm sido propostos para esses casos, como por exemplo YORK (1966,

[1]Departamento de Metalogênese Geoquímica, Instituto de Geociências/UNICAMP, Campinas, e Departamento de Paleontologia e Estratigrafia, Instituto de Geociências/USP, São Paulo.

1967 e 1969), McINTYRE et al. (1966), WILLIAMSON (1968), BROOKS et al. (1972), e outros, para o caso das isócronas Rb/Sr. Tendo trabalhado nesse problema, o autor (AMARAL, 1978) testou uma abordagem pelo método de Monte Carlo, a qual é aqui discutida.

A análise dos trabalhos acima citados, mostra que as abordagens utilizadas são complexas do ponto de vista matemático e geralmente com resultados não muito satisfatórios ou de difícil interpretação. No caso da regressão linear simples, onde temos diversos pares de pontos (x_i, y_i) e desejamos os coeficientes de uma função do tipo:

$$y = a + bx,$$

o problema resume-se na obtenção de a e b. No caso particular da isócrona Rb/Sr, $x = Rb^{87}/Sr^{86}$, $y = Sr^{87}/Sr^{86}$, $a = R_0$ (razão inicial) e b = inclinação da reta, proporcional à idade. Como em qualquer regressão, desejamos obter uma estimativa das incertezas em a e b para que possamos utilizar corretamente as razões iniciais e idades obtidas. Os métodos propostos anteriormente, não possibilitam um exame claro das incertezas, o que nos levou a desenvolver uma abordagem do tipo simulação, pelo método de Monte Carlo, o qual é um processo para a resolução de problemas matemáticos através de amostragem aleatória (SOBOL, 1975).

REGRESSÃO PELO MÉTODO DE MONTE CARLO

A idéia básica do método é a obtenção de um número representativo de regressões, para pontos escolhidos aleatoriamente dentro dos espaços de incerteza. Em outras palavras, no caso da regressão linear, $x_i \pm e_{xi}$ e $y_i \pm e_{yi}$, definem um retângulo de incerteza para cada ponto de observação ou medida, para cada qual é escolhido um ponto ao acaso e calculada a regressão correspondente. Armazenando-se um número significativo de coeficientes (a e b, no caso da regressão linear simples) e parâmetros de ajuste (coeficiente de determinação e erro padrão), estes podem ser tratados estatisticamente e suas incertezas examinadas de modo mais compreensível. A Figura 1 esquematiza o procedimento adotado, para uma das regressões.

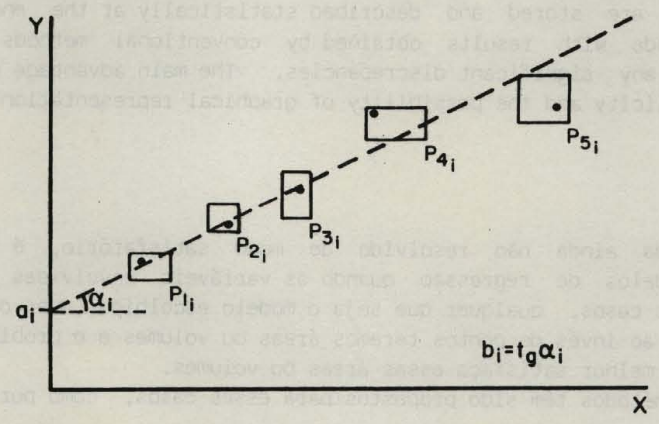

Figura 1 - Esquema de operação do método proposto para uma regressão. Os retângulos indicam o valor central dos pontos e suas incertezas para x e y. O programa, no passo i, gera aleatoriamente um ponto dentro de cada ponto P_k e calcula a regressão correspondente, obtendo a_i e b_i que são armazenados. Após n interações o programa calcula os parâmetros estatísticos e desenha os histogramas respectivos.

De modo a testar o método, desenvolvemos inicialmente um programa em FORTRAN para o sistema Burroughs 6700 da USP, que resultou em nossa primeira comunicação (AMARAL, 1978). Mais recentemente, com a implantação do Laboratório de Informática no Instituto de Geociências da USP, desenvolvemos um novo programa, em BASIC, para microcomputadores compatíveis com IBM-PCXT. O algorítmo usado é listado a seguir:

1. Lê x(i), erro de x(i), y(i), erro de y (i), i=1,n;
2. selecione a opção de regressão, x em y, y em x ou perpendicular à reta;
3. defina o número de regressões;
4. o programa define os limites inferior e superior, para x e y, de cada ponto de observação (define os retângulos de incerteza);
5. utilizando-se a função de números aleatórios, o programa seleciona um ponto em cada retângulo de incerteza e calcula a regressão para cada conjunto, armazenando os coeficientes e parâmetros de ajuste encontrados;
6. repetir o passo 5 para o número de regressões definido em 3;
7. descreve estatisticamente os coeficientes e parâmetros de ajuste, para o conjunto de regressões efetuadas;
8. imprime os resultados.

O programa acima foi desenvolvido para o caso particular da regressão linear simples. Todavia, com pequenas modificações, ele poderia ser ampliado para outros tipos de regressão. Mais ainda, no caso do presente programa o usuário define os erros em termos absolutos. Caso se disponha de um número significativo de réplicas para cada ponto, os erros poderão ser tratados em termos de distribuições multidimensionais, com esquemas próprios de amostragem aleatória. Nestes casos, ao invés de retângulos, teremos elípses cujos eixos serão proporcionais à variância das incertezas ou flutuações aleatórias (BEERS, 1957).

RESULTADOS E DISCUSSÃO

De modo a testar o método, selecionamos conjuntos de resultados de isócronas Rb-Sr, para os quais são disponíveis estimativas das incertezas nos valores Rb^{87}/Sr^{86} e Sr^{87}/Sr^{86}, como por exemplo aqueles apresentados por AMARAL & KAWASHITA (1967) para folhelhos do Grupo Bambuí na região de Vazante, MG, abaixo listados:

Rb^{87}/Sr^{86}(x)	Incerteza em x	Sr^{87}/Sr^{86}(y)	Incerteza em y
22.8	1.1	1.025	0.004
162.0	8.0	2.170	0.020
9.3	0.5	0.876	0.010
60.0	3.0	1.390	0.020
68.0	4.0	1.440	0.030

Para este conjunto, obtivemos os seguintes resultados:

Intersecção .8415±.0246 (2 sigma)
Inclinação .0084±.0004 "
Coeficiente de Correlação .9966±.0026

Idade 589±40 Ma (2 sigma) $\lambda = 1{,}42.10^{-11}.ano^{-1}$

De modo a comparar o método aqui proposto com aqueles de YORK (1966 e 1968), WILLIAMSON (1986) e WENDT (In: BROOKS et al., 1972), utilizamo-nos dos dados referentes aos Gnaisses Jeribá, de SIGA Jr. (1968), discutidos em outro trabalho do presente Boletim.

O procedimento aqui proposto, tanto para regressão y em x, como normal à reta, forneceu uma idade de 583±25 Ma (2 sigma), com razão inicial .7188 ± .0004 (2 sigma). A Tabela I resume as comparações efetuadas entre o método de Monte Carlo e aqueles acima mencionados. A Figura 2 mostra os histogramas dos resultados para a intersecção (a) e inclinação (b) para 1000 regressões. A distribuição quase normal daqueles valores facilita sua análise estatística:

Intersecção máxima = .7212
Intersecção mínima = .7161
Intersecção média = .7188
Desvio padrão = .0002

Inclinação máxima = $8.7969 \cdot 10^{-3}$
Inclinação mínima = $7.8280 \cdot 10^{-3}$
Inclinação média = $8.2722 \cdot 10^{-3}$
Desvio padrão = $0.1816 \cdot 10^{-3}$

O exame dos resultados apresentados no parágrafo precedente, permite afirmar que os resultados obtidos com o método ora proposto são aproximadamente coincidentes com aqueles obtidos pelos métodos rotineiros usados no Centro de Pesquisas Geocronológicas do IG/USP. A principal vantagem do método de Monte Carlo é a sua simplicidade matemática e a possibilidade de visualização das variabilidades sob a forma de histogramas, fatores da maior importância para a correta interpretação dos resultados.

TABELA I

Método	Idade (Ma)	R_0
Monte Carlo	583±25	.7188±.0004
York (1966)	586±26	.7183±.0010
Williamson (1968)	586±38	.7183±.0028
York (1968)	584±42	.7185±.0032
Wendt	584±36	.7185±.0026

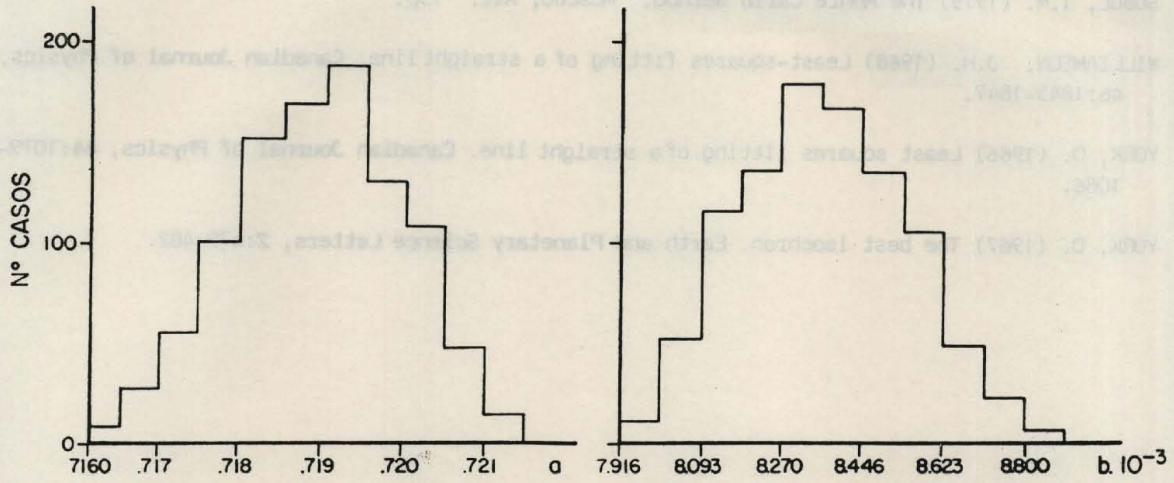

Figura 2 - Histogramas obtidos para a intersecção (a) e inclinação (b), no caso dos Gnaisses Jeribá. Foram efetuadas 1000 regressões. Notar a distribuição claramente normal dos resultados.

CONCLUSÃO

O método de regressão linear pela técnica de amostragem aleatória (Monte Carlo), proposto no presente trabalho, aplicado a dados de isócronas Rb/Sr, forneceu resultados comparáveis àqueles obtidos por outros métodos. A grande vantagem, especialmente para nós que não temos uma profunda formação matemática, é a simplicidade, fator indispensável para a correta interpretação dos resultados.

REFERÊNCIAS BIBLIOGRÁFICAS

AMARAL, G. & KAWASHITA, K. (1967) Determinação da idade do Grupo Bambuí pelo método Rb-Sr. In: CONGRESSO BRASILEIRO DE GEOLOGIA, 21., Curitiba, 1967. **Anais**. Curitiba, SBG. p.214-217.

AMARAL, G. (1978) Cálculo da isócrona Rb-Sr pelo método de Monte Carlo e significado dos parâmetros estatísticos associados. In: CONGRESSO BRASILEIRO DE GEOLOGIA, 30., Recife, 1978. **Boletim de Resumos**. Recife, SBG. p.145.

BEERS, Y. (1962) **Introduction to the theory of error**. Reading, Addison Wesley. 66p.

BROOKS, C.; HART, S.R.; WENDT, I. (1972) Realistic use of two-error regression treatments as applied to rubidium-strontium data. **Reviews of Geophysics and Space Physics, 10**(2):551-577.

SIGA Jr., O. (1986) **A evolução geotectônica da porção nordeste de Minas Gerais, com base em interpretações geocronológicas. São Paulo.** 140p. (Tese de Mestrado, Instituto de Geociências/USP).

SOBOL, I.M. (1975) **The Monte Carlo method**. Moscou, Mir. 73p.

WILLIAMSON, J.H. (1968) Least-squares fitting of a straight line. **Canadian Journal of Physics**, **46**:1845-1847.

YORK, D. (1966) Least squares litting of a straight line. **Canadian Journal of Physics**, **44**:1079-1086.

YORK, D. (1967) The best isochron. **Earth and Planetary Science Letters**, **2**:479-482.

Recebido para publicação em 27/10/1989