Serviço Geológico do Brasil

XII SIMPÓSIO DE RECURSOS HÍDRICOS DO NORDESTE

4 a 7 / novembro / 2014 ★ Natal ★ RN

MODELAGEM HIDROLÓGICA UTILIZANDO REGRESSÃO: CONCEITUAÇÃO MATEMÁTICA, LOGARITMIZAÇÃO E AVALIAÇÃO

Francisco F. N. Marcuzzo

CPRM/SGB - Porto Alegre/RS

Introdução

São vários os recursos matemáticos disponíveis para o hidrologista no seu trabalho diário e no desenvolvimento da pesquisa hidrológica, contudo, a utilização de gráficos de dispersão de dados hidrológicos ocupa uma posição de importância dentre as inúmeras ferramentas disponíveis. Equacionar esta relação cartesiana entre dados hidrológicos e/ou dados físicos e/ou geomorfológicos, que os gráficos de dispersão destas variáveis fornecem é de suma importância para o hidrologista compreender o fenômeno presente e, se possível, estimar o mesmo fenômeno para áreas sem dados coletados em campo ou, até mesmo, prever fenômenos futuros.

Objetivo

O presente trabalho objetivou estudar e discutir, para a hidrologia, a conceituação de regressão matemática, logaritmização (linearização) de modelos de regressão com uma ou duas variáveis explicativas, as possíveis diferenças entre modelos logaritmizados e "naturais", a aplicação e avaliação de modelos.

Conceituações, Equações, Linearização, Aplicação e Avaliação

Gráfico de dispersão entre os valores explicados e os explicativos

O processo básico consiste, primeiramente, na construção de um gráfico de dispersão entre fenômenos geomorfológicos e/ou hidrológicos explicativos (área de drenagem e precipitação, por exemplo) com outros fenômenos hidrológicos que se deseja explicar (vazão, por exemplo). No eixo *X* (das abcissas, eixo horizontal) deve ser plotada a variável independente hidrológica (área de drenagem, por exemplo) associada à variável hidrológica que, ao modificar, assume valores que não dependem dos valores da outra grandeza hidrológica (vazão, por exemplo). No eixo *Y* (das ordenadas, eixo vertical) deve ser registrada a variável dependente (vazão, por exemplo) associada à grandeza hidrológica que, para variar, depende de como varia a outra grandeza hidrológica (área de drenagem, por exemplo).

Regressão matemática simples e múltipla

Em hidrologia, os modelos matemáticos que utilizam a regressão relacionam o desempenho de uma variável dependente (como por exemplo, a vazão) com outra variável, que deve ser independente (como por exemplo, a área de drenagem). Exemplificando matematicamente, quando a função f que relaciona duas variáveis é do tipo:

$$f\ (Xi) = Y = \ a + b\ Xi + \varepsilon i$$

$$f\ (Xi) = Y = \ a + b1\ Xi + b2\ X2i + \cdots + bN\ XNi + \varepsilon i$$

$$a = \ \bar{Y} - b\bar{X}$$

$$b = \ \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - \left(\sum X\right)^2}$$

em que, *f (Xi) = Y* é o valor estimado para um específico valor *Xi*; *b* é a inclinação da reta (coeficiente angular), ou seja, o acréscimo ou decréscimo do valor de *Yi* em relação à *Xi*; a localiza o ponto de interseção da reta em relação ao sistema de coordenada retangulares, ou seja, é ponto no qual uma linha irá interceptar o eixo *Y* usando valores de *X* e *Y* existentes na regressão;   representa uma perturbação aleatória na função, ou seja, nem todos os pontos tocam a reta, e essa diferença é o erro de aproximação; *e N* é o tamanho da amostra; $\bar{Y}$ é a média dos valores explicados; $\bar{X}$ é a média dos valores explicativos.

Observa-se na primeira equação acima um modelo de regressão linear simples. Sabe-se que a variável *X* (vazão) é a variável independente da equação, enquanto *Y* (área de drenagem) é a variável dependente (ou seja, *Y = f (X)*), das variações que ocorrem em *X*. Observa-se na segunda equação acima um modelo de regressão linear múltipla.

Construção da equação de regressão pelo método dos mínimos quadrados

O método dos mínimos quadrados é o mais utilizado para a construção de equações (retas, curvas) de regressão. Este método definirá uma reta que minimizará a soma das distâncias ao quadrado entre os pontos plotados explicativos (*X*), como por exemplo, área de drenagem, e explicados (*Y*), como por exemplo, vazão, e a reta, ou seja, os valores obtidos pela equação de regressão (*X'*,*Y'*). Pelo método dos mínimos quadrados calculam-se os parâmetros "*a*" e "*b*" da reta que minimiza estas distâncias ou as diferenças (denominado também de desvios ou erros) entre *Y* e *Y'*. O método consiste nos seguintes passos:

Calcular a diferença entre o valor observado em campo (explicados, *Y*), como por exemplo, vazão, e o mesmo valor estimado *Y'* pela equação de regressão:

$$Erro\ = \ E\ = \ (Y - \ Y')$$

O objetivo do modelo de regressão é fornecer uma reta (curva) com r o somatório dos erros menor possível:

$$E1^2\ + \ E2^2\ + \ E3^2 + \ldots + En^2\ = \ Mínimo$$

A hipótese de trabalho é a seguinte:

$$Erro\ Total\ = \ \boldsymbol{\Sigma}(Y - Y')^2$$

Equação de regressão (da reta) que minimiza o erro (ou os desvios):

$$Y' = \ a + b\ X$$

Em seguida substitui a equação de regressão na equação de hipótese de trabalho:

$$\Sigma(Y - a - b\ X)^2$$

Objetivando-se que a soma dos quadrados dos desvios (erros dos valores explicados coletados em campo e os estimados pela equação de regressão) tenha um valor mínimo possível, devem-se aplicar os conceitos de cálculo diferencial com derivadas parciais. Como as incógnitas do problema são os coeficientes "*a*" e "*b*" estrutura-se um sistema de duas equações. Portanto, aplicando os conceitos descritos acima, arma-se o sistema de equações normais que consentirá determinar os valores dos coeficientes "*a*" e "*b*", sendo *N* o tamanho da amostra:

$$\frac{\partial e}{\partial a} = \ -2\ \Sigma\ (Y - \ a\ - \ bX)$$

$$-2\ \Sigma\ Y + 2\ \Sigma\ a\ + 2\ \Sigma\ bX$$

$$\Sigma\ Y = \Sigma\ a + \Sigma\ bX$$

$$\Sigma Y = Na\ \ + b\ \Sigma\ X$$

$$a = \frac{\Sigma\ Y + b\ \Sigma\ X}{N}$$

$$\Sigma\ XY = a\Sigma\ X + b\Sigma X^2$$

$$b = \ \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - \left(\sum X\right)^2}$$

Os valores a e b acima correspondem aos parâmetros da equação de regressão que minimiza as diferenças entre os valores de *Y* (levantados) e os de *Y'* (estimados pela regressão). Portanto, o problema de "*fitting*" (ajustar) uma reta que melhor se adeque à nuvem de dados se reduz em calcular os parâmetros a e b da equação de regressão.

Transformações de equações naturais em equações linearizadas

Por heurística entende-se que a relação entre duas grandezas hidrológicas quase sempre não é linear, e é essencial descobrir de que tipo é e quais são os parâmetros que a caracterizam. Numa relação linear é muito simples o processo de se definir os parâmetros envolvidos da equação, logo, quando se observa pelo gráfico de dispersão que os valores envolvidos não formam uma reta, pode-se linearizá-lo utilizando uma mudança de variáveis, transformando em equações de retas mesmo regressões que formam curvas visivelmente complexas. Este método, de transformar uma equação de regressão que gera uma curva em uma equação que gera uma reta, denomina-se linearização ou logaritimização.

Sugere-se um conhecimento mediano com as representações gráficas das principais funções matemáticas utilizadas em regressão, já que se deve ter uma base teórica sobre que tipo de função matemática de regressão poderia gerar uma curva parecida a indicada pela sequência de pontos formados por variáveis explicativas e explicadas no gráfico de dispersão.

O Quadro 1 mostra um resumo com os diferentes tipos de equações mais utilizadas em estudos hidrológicos e suas respectivas transformações lineares por logaritmos.

Quadro 1. Resumo dos tipos de equações e suas respectivas transformações lineares.

Tipo de Equação		Equação	Linearização (logaritmização)	Variáveis Observadas	
				X	Y
1	Linear	Y' = a + bX	Y' = a + bX	X	Y
2	Exponencial	$Y' = a\ e^{bX}$	ln Y' = ln a + bX	X	ln Y
3	Logarítmica	Y' = a + b ln X	Y' = a + b ln X	ln X	Y
4	Potencial	$Y' = a\ X^b$	ln Y' = ln a + b ln X	ln X	ln Y

Avaliação da utilidade das equações de regressão

A equação de regressão que se obtém, seja ela linearizada ou não linearizada, utilizando o método dos mínimos quadrados é apenas uma aproximação dos valores reais. A reta produzida é um modo útil para indicar a tendência dos dados explicados por meio dos dados explicativos. Contudo deve-se avaliar a equação produzida de modo a se ter um ou mais parâmetros qualitativos da curva de regressão para se aprovar ou não, ou, até mesmo, escolher a que representa melhor os dados explicados por meio de comparação com outros tipos de regressão, conjuntos de dados, etc.

Existem critérios qualitativos que podem recomendar o quanto útil ou aproximado dos dados reais coletados em campo é a curva produzida pela regressão executada. Destacam-se os seguintes critérios:

- a. Desvio (erro) dos valores estimados (*Se*): O desvio padrão (*Se*) dos valores estimados, ou erro padrão, mede o desvio médio entre os valores reais de *Y* e os valores estimados *Y'*. Este dado qualitativo informa de modo aproximado a extensão do desvio (erro) entre os valores obtidos das estimativas explicadas e os valores de *Y* captados em campo. O desvio (erro) é medido na unidade da ordenada *Y*. O objetivo é perseguir o menor valor possível de Se.
- b. Coeficiente de determinação (*r*²): O coeficiente de determinação indica o quanto a curva de regressão explica o ajuste da curva. Após a estimativa dos coeficientes da reta de regressão, é necessário verificar se os dados amostrais são descritos pelo modelo das equações, além disso, determinar a parcela da variabilidade amostral que foi, de fato, explicada pela reta de regressão.
- c. Coeficiente de correlação (*r*): O coeficiente de determinação é sempre positivo, variando de 0 a 1, enquanto que o coeficiente de correlação (*r*) pode admitir valores negativos e positivos, variando de –1 a 1, ou seja, -1 ≥ r ≤ +1. Valores de r igual ou próximos de 1 ou –1 indica que determina uma intensa relação entre as variáveis: no primeiro caso a relação é direta, enquanto que no segundo a relação é inversa. Valores iguais ou próximos de zero, significa que existe nenhuma ou pouca relação entre as variáveis explicativas e explicadas.

Considerações Finais

No ajuste do modelo da equação de regressão, se obtém uma melhor performance da explicação do conjunto de dados hidrológicos recolhidos. A equação de regressão que se obtém, seja ela linearizada ou não linearizada, utilizando o método dos mínimos quadrados é apenas uma aproximação dos valores reais. Deve-se avaliar a equação produzida com um ou mais parâmetros qualitativos da curva de regressão para se aprovar ou não, ou, escolher a que representa melhor os dados explicados por meio de comparação com outros tipos de regressão, conjuntos de dados, etc.

Agradecimentos

O autor agradece a CPRM/SGB (Companhia de Pesquisa Recursos Minerais / Serviço Geológico do Brasil - Empresa Pública do Ministério de Minas e Energia) pelo fomento que possibilitou o desenvolvimento deste estudo.

CPRM Serviço Geológico do Brasil | Secretaria de Geologia, Mineração e Transformação Mineral | Ministério de Minas e Energia