CORRELAÇÃO E REGRESSÃO:

COMO CONSTRUIR MODELOS EMPÍRICOS

16 aula

META

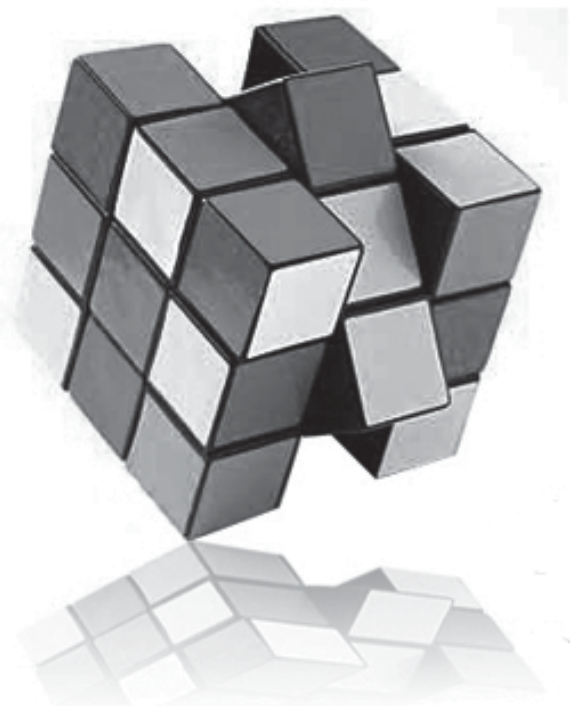
Apresentar ao aluno o método de regressão por mínimos quadrados.

OBJETIVOS

Ao final desta aula, o aluno deverá:
identificar o tipo de problema ao qual os métodos de regressão podem ser aplicados, comparando com os problemas estudados com os planejamentos fatoriais;
representar os dados em gráficos e analisar qualitativamente o comportamento;
identificar os parâmetros de um modelo linear e utilizar o método matricial;
calcular os resíduos para a avaliação da qualidade do modelo.

PRÉ-REQUISITOS

Variáveis correlacionadas.



Cubo mágico, fotomontagem, autor desconhecido (Fonte: clipperdownloads.blogspot.com)

INTRODUÇÃO

Até a última aula estudamos os planejamentos fatoriais, nos quais são estudados dois níveis distintos para as variáveis, o que nem sempre é garantia de que se tenha acesso aos dados completos a respeito do comportamento do sistema. Para situar os planejamentos fatoriais dentre os métodos quimiométricos em termos do grau de sofisticação, estes constituem apenas a etapa inicial na investigação completa. Quando temos acesso aos efeitos das variáveis sobre uma dada resposta, isto nos dá a dimensão da importância relativa de cada variável no processo. Muito bem, uma etapa posterior pode ser um estudo das variáveis mais relevantes em um número maior de níveis, para sabermos mais a respeito da chamada superfície de resposta. Neste contexto, um estudo mais aprofundado da dependência mútua entre duas variáveis importantes (uma delas sendo a própria resposta) pode nos levar à construção de um modelo empírico. Este último nada mais é do que um modelo matemático mais completo, capaz de conferir maior previsibilidade.

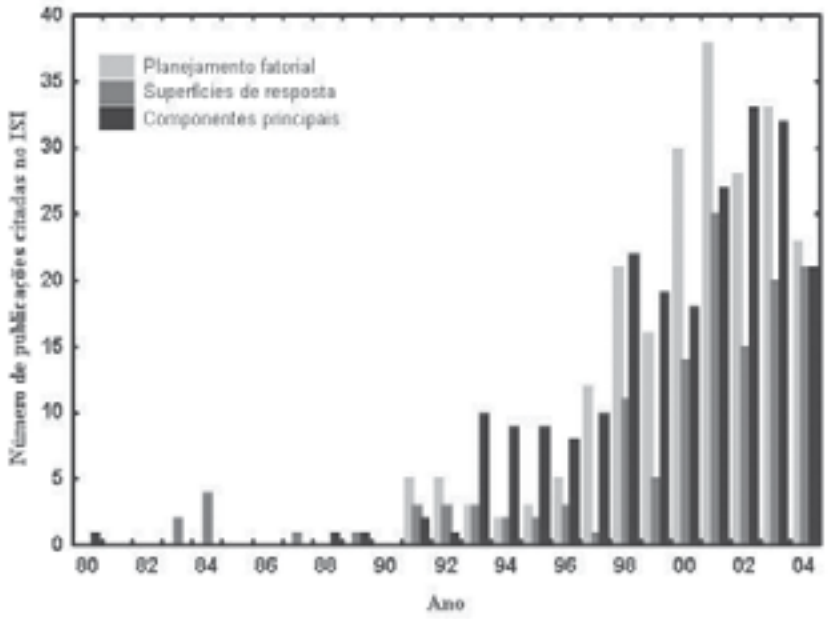


Gráfico. (Fonte: www.scielo.br).

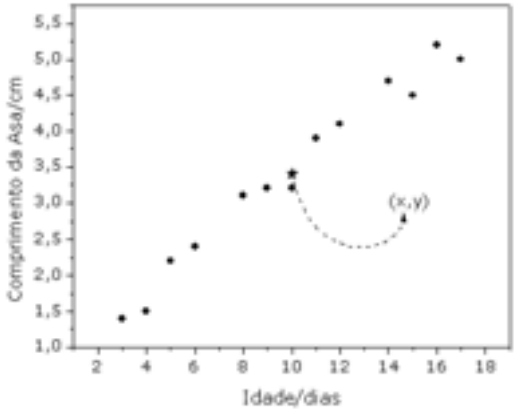
CORRELAÇÃO

Vamos examinar uma situação bastante interessante relatada por um grupo de biólogos, referente a medidas do tamanho das asas de uma determinada ave em relação à idade das mesmas. A tabela abaixo mostra os valores obtidos para o comprimento das asas e a idade das aves, bem como os valores das médias, lembrando que cada medida está sujeita a um erro aleatório.

COMPRIMENTO DE ASAS DE 12 AVES

Idade	C. Asas	Idade	C. Asas
3	1,4	11	3,9
4	1,5	12	4,1
5	2,2	14	4,5
6	2,4	15	4,7
8	3,1	16	5,0
9	3,2	17	5,2
	Media	**10**	**3,4**

Os valores foram representados em um gráfico, também mostrado a seguir, e que sugere que haja uma correlação entre as variáveis. Tomando uma régua e realizando tentativas de se traçar uma reta que englobe os pontos, ou que pelo menos os pontos fiquem próximos da reta, você poderá constatar que se trata de uma proposta razoável.



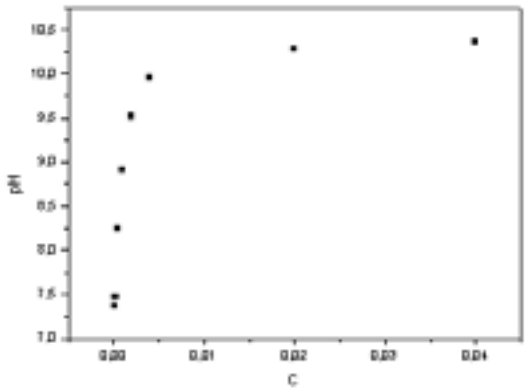
ATIVIDADES

Agora, para ilustrar a importância de se representar em um gráfico, apresentamos os dados abaixo, relativos às condutividades elétricas de soluções iônicas, para diferentes pHs. Não se preocupe com o significado destes dados, o que queremos que você responda é qual o comportamento da variação nos valores, ou seja, se quando a condutividade aumenta, o que acontece com o pH (também aumenta ou diminui?). Responda também se é possível somente olhando para os valores, prever se a dependência das variáveis segue um modelo linear.

Condutividade	pH
4E-5	7,38
2E-4	7,48
4E-4	8,25
9,99E-4	8,91
0,00199	9,52
0,00399	9,95
0,0199	10,28
0,0399	10,37

COMENTÁRIO SOBRE AS ATIVIDADES

Olhando para a tabela, fica claro que quando a condutividade aumenta o pH também aumenta. Contudo, é muito difícil afirmar se há uma reação de linearidade entre os valores. Para fazer esta avaliação, uma opção é representá-los em um gráfico, como mostrado a seguir:



Finalmente, a observação do gráfico não deixa dúvidas: não há um comportamento linear. Se você pegar uma régua e tentar posicioná-la de modo a que os pontos possam ser unidos por uma régua, certamente não irá conseguir.

Voltando ao nosso problema do comprimento das asas das aves, que pudemos atribuir um comportamento linear, neste caso, lembramos que a equação de uma reta tem a seguinte forma:

$$y = b_o + b_1 x$$

onde:

x e y: variáveis

b_o: coeficiente linear

b_1: coeficiente angular

Logo, uma expressão satisfatória para nossa reta poderia ser, levando em conta também o erro:

$$y_i = b_o + b_1 x_i + e_i$$

onde, especificamente para os dados apresentados na tabela:

y_i = comprimento da asa

x_i = idade da ave

b_o e b_1 = parâmetros do modelo

e_i = erro aleatório ao qual as medidas estão sujeitas

Na prática, ajustar o modelo, que é o que estamos interessados em fazer, significa encontrar os valores de b_o e b_1, pois os valores das variáveis x e y são os próprios valores medidos experimentalmente. Na verdade temos que resolver um sistema de equações:

$$y_1 = \beta_o + \beta_1 x_1 + \varepsilon_1$$

$$y_2 = \beta_o + \beta_1 x_2 + \varepsilon_2$$

$$y_3 = \beta_o + \beta_1 x_3 + \varepsilon_3$$

.....

$$y_i = \beta_o + \beta_1 x_i + \varepsilon_i$$

ou, mais genericamente:

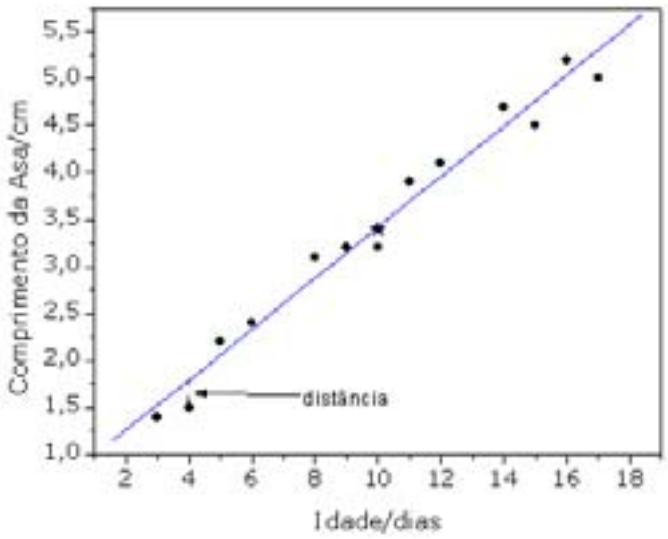
$$\mathbf{y = X\beta + \varepsilon}$$

onde **y, X, β e ε** apresentam-se como matrizes:

$$y = \begin{bmatrix} y1 \\ y2 \\ \cdots \\ yn \end{bmatrix} \quad x = \begin{bmatrix} 1 & x1 \\ 1 & x2 \\ & \cdots \\ 1 & xn \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon 1 \\ \varepsilon 2 \\ \cdots \\ \varepsilon n \end{bmatrix} \quad \beta = \begin{bmatrix} \beta o \\ \beta 1 \end{bmatrix}$$

Note que esta forma geral pode ser aplicada a qualquer número de valores diferentes para cada uma das variáveis; adicionalmente, ela se presta ao modelo linear, pois temos como parâmetros do modelo os coeficientes b_o e b_1. Na próxima aula voltaremos à tarefa específica de determinar o modelo propriamente dito. Por ora vamos refletir um pouco mais sobre a relação entre um conjunto de dados e sua obediência a um modelo.

A seguir mostramos novamente o gráfico contendo os valores representados no gráfico, no qual representamos também uma reta que passa próxima aos pontos.



Na realidade, mesmo usando uma régua, pode-se constatar que não existe uma reta que passe exatamente por todos os pontos, pois mesmo se considerarmos que estes obedecem ao modelo, certamente estaremos frente à influência de erros. Isto significa que a determinação dos valores de b_o e b_1 não é imediata. Qualquer que seja a reta escolhida (implicando escolhas de b_o e b_1) sempre algumas das observações (pontos) estarão fora da reta. No jargão estatístico, sempre haverá os chamados resíduos ou diferenças entre o valor experimental e o valor previsto pelo modelo. Estas diferenças ou resíduos podem ter sinal (+) ou (-) dependendo do posicionamento do valor experimental acima ou abaixo da reta.

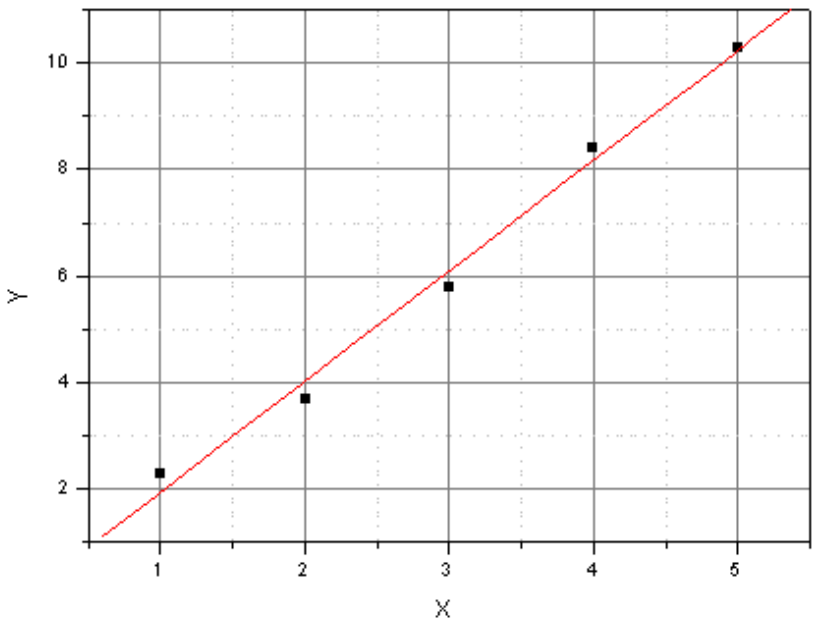
Neste ponto você deve estar pensando: bem, já que é tão difícil uma reta passar por todos os pontos e temos que "conviver" com os resíduos, haverá uma maneira de escolher qual a reta mais "correta" para representar o modelo? Na verdade há sim, e ela é definida como sendo a reta que passe "mais perto" dos pontos, ou seja, a distância global dos pontos em relação à reta deverá ser a menor possível. Na figura, esta distância é representada por um segmento vertical. Na prática, o que se faz é calcular o quadrado dos resíduos (para eliminar as diferenças de sinais) e somar estes quadrados. O nome deste método é ajuste por mínimos quadrados. A reta cuja soma seja mínima será a melhor escolha.

ATIVIDADES

Com os dados da tabela abaixo:

x	y
1	2,3
2	3,7
3	5,8
4	8,4
5	10,3

Foi construído o seguinte gráfico, onde já está representada uma reta possível (não se preocupe ainda com o fato de ser ou não esta reta a melhor escolha por mínimos quadrados):



Indique:

a) quais os pontos pelos quais a reta passa de fato.
b) Indique quais pontos apresentaram resíduo em relação à reta
c) Quais os sinais destes resíduos.

CONCLUSÃO

Concluímos a partir desta aula que para variáveis correlacionadas, ou seja, variáveis que são dependentes uma da outra, a partir de uma série de valores, é possível chegar a um modelo empírico completo na forma de uma equação de reta, por exemplo. Contudo, devido à ocorrência de erros na forma de flutuações aleatórias, concluímos que mesmo se estivermos frente a uma dependência linear legítima, dificilmente haverá uma reta que passe perfeitamente por todos os pontos. As diferenças entre os valores experimentais e os valores previstos pelo modelo (obviamente estes últimos situando-se exatamente sobre a reta) são os chamados resíduos.

Resolvendo cubo mágico (Fonte: www.permutalivre.com.br)

RESUMO

A partir desta aula, vamos nos ocupar da construção de modelos empíricos a partir de conjuntos de dados envolvendo um grande número de valores diferentes para as variáveis. Neste contexto, um estudo mais aprofundado da dependência mútua entre duas variáveis importantes (uma delas sendo a própria resposta) pode nos levar à construção de um modelo empírico. Este último nada mais é do que um modelo matemático mais completo, capaz de conferir maior previsibilidade.

Utilizamos como exemplo para a ilustração de uma regressão dados fornecidos por um grupo de biólogos, referentes a medidas do tamanho das asas de uma determinada ave em relação à idade. A representação dos dados em um gráfico mostrou que as variáveis estariam correlacionadas. Aparentemente, a dependência é linear, assim lembrando que a equação de uma reta é do tipo $y = b_o + b_1x$. O que mostramos nesta aula é como realizar um ajuste dos nossos dados a este modelo, ou seja, realizar uma série de operações que nos permitam determinar os valores de b_o e b_1, sendo que a tarefa específica de aplicar este modelo será o objeto da próxima aula.

Voltando aos dados, quando nós os representamos em um gráfico e constatamos que aparentemente eles definem uma reta, propomos a tarefa de tentar, com uma régua, desenhar uma reta que passe por todos os pontos ou que pelo menos fique o mais próximo possível deles. Em geral, se estivermos frente a dados experimentais, dificilmente iremos obter uma reta que passe exatamente por todos os pontos, pois mesmo se considerarmos que estes obedecem ao modelo, certamente estaremos frente à influência de erros. No linguajar estatístico, sempre haverá resíduos, que podem ter sinal (+) ou (-) dependendo do posicionamento do valor experimental acima ou abaixo da reta. Assim, o método que vamos estudar calcula o quadrado dos resíduos (ajuste por mínimos quadrados), sendo que melhor a reta será aquela cuja soma seja mínima.

16 aula

CONDUTIVIDADE DE SOLUÇÕES

O movimento de íons em uma solução pode ser estudado pela medida das condutividades elétricas de soluções de eletrólitos. A migração de cátions na direção de um eletrodo carregado positivamente e de ânions em direção a um eletrodo carregado negativamente provoca um fluxo de corrente elétrica através da solução, sendo, portanto, dependente das propriedades de transporte dos íons. Os íons H^+ e OH^- têm valores de condutividade anormalmente elevados. Isto acontece porque um mecanismo específico de condução pode atuar para estas duas espécies em água, pois eles são os constituintes do solvente. Outros íons são atraídos pelo campo elétrico e para se moverem têm que afastar moléculas do solvente de seu caminho. O H^+ e o OH^-, por outro lado, podem usar a própria estrutura do solvente (com as ligações de hidrogênio) para progredirem sem terem que afastar as moléculas do solvente.

PRÓXIMA AULA

Caro aluno, na próxima aula iremos aprender a realizar o método de ajuste por mínimos quadrados, para a realização de uma regressão linear.

REFERÊNCIAS

BOX, G. E. P.; HUNTER, W. G.; HUNTER, J. S. **Statistics for experimenters. An introduction to design, data analysis and model building**. New York: Wiley, 1978.

BUSSAB, W. O.; MORETIN, P. A. **Estatística básica.** São Paulo, Ed. Atual, 1985.

BARROS NETO, B.; SCARMINIO, I. E.; BRUNS. R. E. **Planejamento e otimização de experimentos**. Campinas Editora da Unicamp, 1995.