ANÁLISE DO DESEMPENHO DO MODELO EM REGRESSÕES

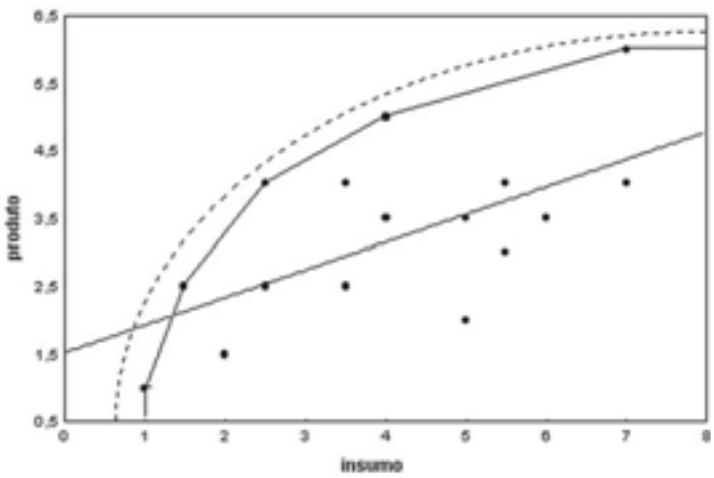
18 aula

META

Fazer com que o aluno seja capaz de realizar os procedimentos existentes para a avaliação da qualidade dos ajustes aos modelos.

OBJETIVOS

Ao final desta aula o aluno deverá:
calcular os desvios em torno da média, os desvios devidos à regressão e os desvios residuais e localizá-los graficamente;
calcular as somas quadráticas, graus de liberdade e médias quadráticas;
aplicar a análise da variância e o coeficiente de determinação;
aplicar a análise da variância;
construir e avaliar os gráficos de resíduos.



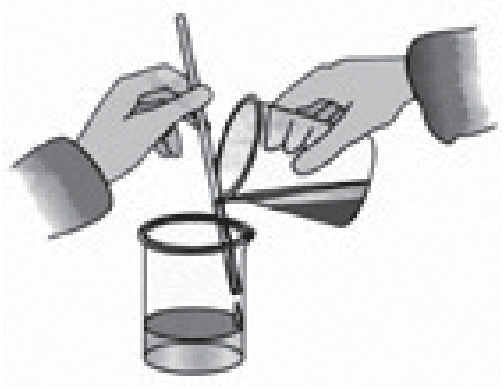
Análise do desempenho em regressões - reprodução - desconhecido (Fonte: http://www.eps.ufsc.br)

PRÉ-REQUISITOS

Conhecimento do método de ajuste linear por mínimos quadrados.

INTRODUÇÃO

Descobrimos na última aula o que fazer com dados experimentais para ajustar um modelo matemático pelo método de mínimos quadrados. Mas nosso problema não para por aí. A qualidade do modelo ajustado a partir dos métodos de regressão deve ser analisada para que este modelo possa ser considerado válido, ou seja, para validar o modelo. Para a avaliação da qualidade do ajuste fazemos uso dos resíduos deixados pelo modelo, com base na suposição de que um modelo ideal não deixaria resíduo algum. Mesmo considerando que isto é uma situação ideal e que praticamente todos os ajustes vão deixar algum resíduo, existe um limite para isto e um modelo que envolva resíduos significativos será considerado um modelo ruim. Na prática, como os resíduos são calculados pela diferença entre os valores previstos pelo modelo e os valores observados, um modelo com resíduos significativos apresenta uma previsibilidade muito baixa e não pode ser validado. É destas questões que vamos tratar nesta aula.

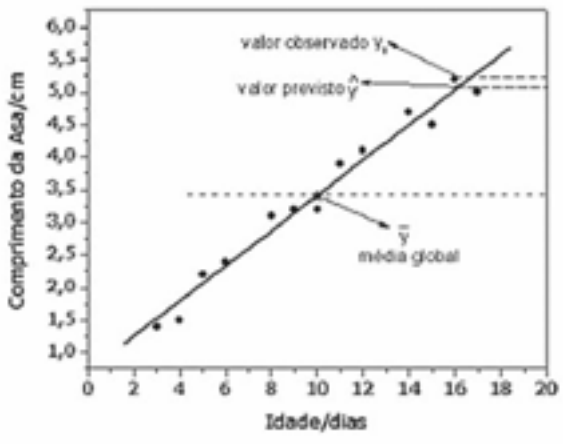


Análise (Fonte: www.erau.edu).

REGRESSÃO

O ponto de partida para a avaliação da qualidade de um ajuste é a análise dos desvios dos valores observados em relação à média global. No gráfico abaixo, está representada a média global (sobre a reta ajustada), os valores observados y_i (ligeiramente fora da reta) e os valores previstos, que também se situam sobre a reta (um deles foi marcado com um sinal +). Note o seguinte: a distância (vertical) entre os pontos observados e a média global pode ser dividida em duas partes: a distância entre o valor observado e o valor previsto e a distância entre o valor previsto e a média global. Desta observação iremos extrair a expressão para a variância do modelo. Em um modelo bem ajustado, em que os valores previstos são muito próximos aos valores observados, a distância entre ambos e a média global é muito pequena.

Devemos notar também o seguinte: quando medimos os tamanhos das asas e relacionamos com a idade das aves, é perfeitamente natural que haja variação na resposta medida, pois a ave está crescendo e se desenvolvendo. Portanto, apesar de podermos calcular uma media global para a resposta, os desvios dos valores observados em relação à média global são naturais e esperados. Se nosso modelo foi bem ajustado, os desvios dos valores previstos em relação à média global são explicados pela regressão. Por outro lado, parte dos desvios dos valores observados em relação à média global é explicada pelos resíduos. Quanto menores os resíduos, mais bem ajustado será o modelo.



Podemos representar matematicamente esta divisão da distância entre os pontos observados e a média global pela seguinte expressão:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

Desvio em torno da média	=	Desvio devido a regressão	+	Desvio residual

Contudo, como os desvios podem ser positivos ou negativos, se não utilizarmos nenhum artifício para uniformizar os sinais, teremos uma estimativa irreal. Por isso, fazemos a somatória dos quadrados destes desvios e chamamos estes parâmetros de somas quadráticas (SQ):

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

S.Q em torno da média	=	S.Q. devido a regressão	+	S.Q. residual

ou:

$$\mathbf{SQ_T = SQ_R + SQ_r}$$

ANÁLISE DA VARIÂNCIA: COEFICIENTE DE DETERMINAÇÃO

Como destacamos anteriormente, a regressão explica parte dos desvios das observações em relação à media, ficando o restante por conta dos resíduos. Quanto maior a fração descrita pela regressão, melhor o ajuste. O parâmetro que quantifica a qualidade do ajuste é o coeficiente de determinação R^2:

$$R^2 = \frac{SQ_R}{SQ_T}$$

Quando o ajuste é o melhor possível, os valores observados são iguais aos previstos, a regressão explica todos os desvios. Assim, a soma quadrática devido à regressão (SQ_R) é igual à soma quadrática total (SQ_T) e o coeficiente de determinação, portanto seria igual a 1. Este é o maior valor possível, mas valores por exemplo, de 0,999 são considerados excelentes.

A análise da variância normalmente é feita a partir de uma tabela conhecida como tabela ANOVA (acrônimo do inglês *ANalysis Of VAriance*), na qual aparecem médias quadráticas ao invés de somas quadráticas. Estas médias quadráticas são calculadas, por sua vez, dividindo-se as somas quadráticas por um número conhecido como grau de liberdade (ver nota explicativa).

Voltando às nossas somas quadráticas, cada uma delas tem um certo grau de liberdade. Para a soma quadrática dos N desvios em relação à média, neste caso também terá N-1 graus de liberdade. O número de graus de liberdade de SQ_R é igual a 1 pois se lembrarmos que esta é a somatória das diferenças entre os valores previstos e a média global, os valores previstos dependem apenas do parâmetro b_1 do modelo (lembre-se de que os valores observados nos experimentos são independentes mas os previstos não). No caso da soma quadrática residual, observando a equação que relaciona todas as SQs ($SQ_T = SQ_R + SQ_r$), se o número de graus de liberdade de SQ_T é 1, o número para SQ_R é N-1, o número de graus de liberdade de SQ_r só pode ser N-2. Logo:

GRAU DE LIBERDADE:

$$\nu_T = \nu_R + \nu_r$$

Como dissemos anteriormente, a tabela de análise da variância contém as médias quadráticas, que são o resultado da divisão de cada soma quadrática pelo seu número de graus de liberdade:

ANÁLISE DA VARIÂNCIA

Fonte de Variação	Soma Quadrática	Nº de g.l.	Média Quadrática
Regressão	$\sum_i (\hat{y}_i - \bar{y})^2$	1	$MQ_R = SQ_R$
Resíduo	$\sum_i (y_i - \hat{y}_i)^2$	n-2	$MQ_r = SQ_r/(n-2)$
Total	$\sum_i (y_i - \bar{y})^2$	n-1	

Utilizando os dados referentes ao ajuste que fizemos na aula passada:

ANÁLISE DA VARIÂNCIA (TABELA ANOVA)

Fonte de Variação	Soma Quadratica	Nº de g.l.	Média Quadratica
Regressão	25,881	1	25,881
Resíduo	0,9406	4	0,2352
Total	26,8334	5	5,3667

$$\mathbf{R^2 = SQ_R/SQ_T = 25,881/26,8334 = 0,965}$$

É um valor próximo de 1, o que indica que a regressão de fato explica a maior parte da variância, ou seja, 96,5%.

Significância estatística da regressão

A significância estatística da regressão é testada fazendo-se a razão MQ_R/MQ_r e comparando-se o valor resultante com o valor encontrado na tabela de distribuição $F_{n1,n2}$ onde n_1 e n_2 são os graus de liberdade das médias quadráticas MQ_R e MQ_r, respectivamente, em um determinado nível de confiança. Por exemplo, se fizermos a razão MQ_R/MQ_r para os valores obtidos para nosso exemplo, obteremos uma razão de 110,038. O número de graus de liberdade de MQ_R foi 1 e de MQ_r foi 4. Assim, para um nível de confiança de

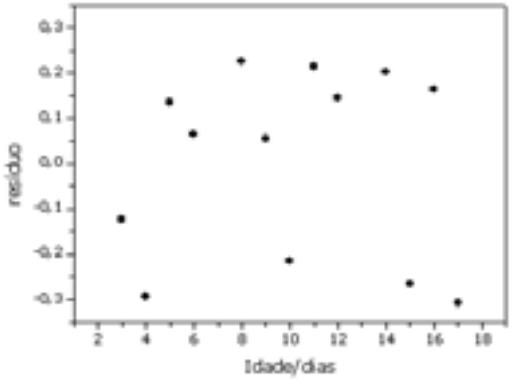
95% (que é um valor muito usado), o valor previsto na tabela de distribuição F a 5% (que é 100 % - 95 %) é $F_{1,4} = 7,71$. Esta tabela pode ser encontrada em livros de estatística.

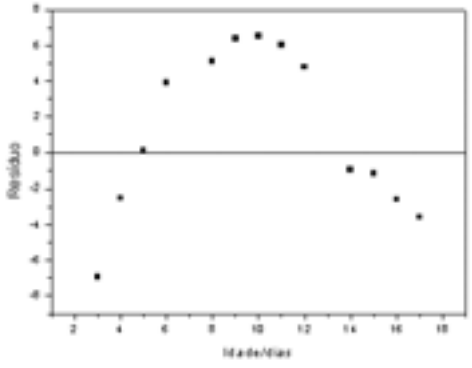
$\mathbf{MQ_R/MQ_r} = 25,881/0,2352 = 110,038 >>> F_{1,4} = 7,71$

Para que nossa regressão tenha significância estatística, a razão MQ_R/MQ_r deve ser maior ou igual ao valor contido na tabela de distribuição F. Como o valor 110,038 é muito maior do que 7,71, nossa regressão tem uma significância estatística bastante satisfatória.

GRÁFICO DOS RESÍDUOS

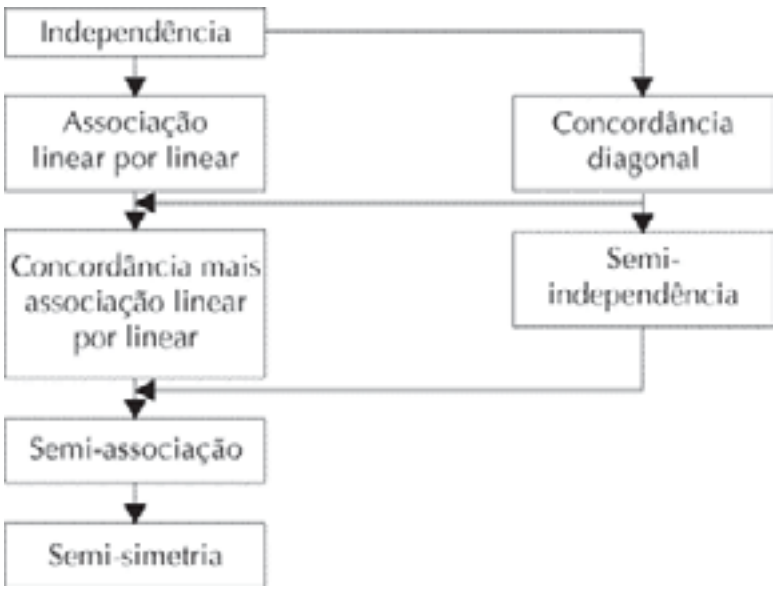
Finalmente, uma das maneiras mais práticas de avaliarmos o desempenho do modelo obtido é a observação do chamado gráfico dos resíduos. Neste gráfico representamos os resíduos em função do valor da variável *x* e, caso o modelo esteja bem ajustado, os valores dos resíduos devem se distribuir aleatoriamente (sem apresentar um padrão geométrico). Além disso, deve haver uma distribuição razoavelmente homogênea de valores positivos e negativos. Mostramos abaixo um exemplo de gráfico de resíduos em que estas condições são satisfeitas, sugerindo um modelo bem ajustado.





CONCLUSÃO

Concluímos, pelo que vimos nesta aula, que após o ajuste de um conjunto de dados experimentais a um determinado modelo, no nosso caso um modelo linear, é sempre necessário realizar uma análise do desempenho deste ajuste. Muitas pessoas simplesmente observam o valor do chamado coeficiente de correlação ou de determinação, mas vimos que isto não é suficiente. Na realidade, um modelo só pode ser considerado bem ajustado, de modo inequívoco, se for feita a análise da variância e o gráfico dos resíduos resultar em uma distribuição aleatória.



Sequencia de modelos log-lineares herárquicos embutidos – reprodução (Fonte: http://www.scielo.br)

RESUMO

Vimos na última aula o que fazer com dados experimentais para ajustar um modelo matemático pelo método de mínimos quadrados, mas o fato de podermos calcular os parâmetros de um modelo não basta para que este seja validado. Em linhas gerais, para a avaliação da qualidade do ajuste, fazemos uso de vários recursos, como a chamada análise da variância e a análise dos resíduos deixados pelo modelo.

O ponto de partida para a avaliação da qualidade de um ajuste é a análise dos desvios dos valores observados em relação à média global que podem ser divididos em duas partes: a distância entre o valor observado e o valor previsto e a distância entre o valor previsto e a média global. Se nosso modelo for bem ajustado, os desvios dos valores previstos em relação à média global são explicados pela regressão. Por outro lado, parte dos desvios dos valores observados em relação à média global é explicado pelos resíduos. Quanto menores os resíduos, mais bem ajustado será o modelo. Contudo, como os desvios podem ser positivos ou negativos, se não utilizarmos nenhum artifício para uniformizar os sinais, teremos uma estimativa irreal. Por isso, fazemos a somatória dos quadrados destes desvios e chamamos estes parâmetros de somas quadráticas (SQ): $SQ_T = SQ_R + SQ_r$. Quanto maior a fração descrita pela regressão, melhor o ajuste. Posteriormente, calculam-se as médias quadráticas: $MQ_R=SQ_R$ e $MQ_r=SQ_r/(n-2)$.A significância estatística da regressão é testada fazendo-se a razão MQ_R/MQ_r e comparando-se o valor resultante com o valor encontrado na tabela de distribuição $F_{n1,n2}$ onde n_1 e n_2 são os graus de liberdade das médias quadráticas MQ_R e MQ_r, respectivamente, em um determinado nível de confiança.

Por outro lado, outra forma de avaliarmos o desempenho do modelo obtido é a observação do chamado gráfico dos resíduos. Neste gráfico representamos os resíduos em função do valor da variável x e, caso o modelo esteja bem ajustado, os valores dos resíduos devem se distribuir aleatoriamente.

GRAUS DE LIBERDADE

Os graus de liberdade de um parâmetro representam quantos valores independentes envolvendo as n observações y_1, y_2, y_3 são necessários para determinar este parâmetro. Por exemplo, no cálculo de uma média (lembre-se de que ela é calculada por: $(1/N)Ó x_i$), são necessários todos os N valores e a média tem, portanto, N graus de liberdade. Para os desvios (calculados por: $d_i = x_i - x_{médio}$) a situação é diferente, pois a soma de todos os desvios N sempre deve dar zero. Faça um teste rápido: some quatro números quaisquer e faça a média, por exemplo, 3,4,5,e,6. A média dá 4,5. Agora faça a soma de todos os desvios: (3-4,5+4-4,5+5-4,5+6-4,5= 0). Se conhecermos o valor de três dos desvios, o quarto valor não é independente: é obrigatoriamente o valor que tornará a soma nula. Em outras palavras, a soma dos desvios tem N-1 graus de liberdade, sendo N-1 valores independentes uns dos outros e 1 valor "amarrado" à soma de todos os outros.

PRÓXIMA AULA

Caro aluno, na próxima aula, finalmente veremos uma aplicação de ajuste de modelos a um problema típico de química experimental.

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.