Análise de regressão: aplicação em biologia

CLAUDIO LUIZ MELO DE SOUZA

Doutor em Proteção de Plantas.

Professor do Instituto Superior de Tecnologia em Ciências Agrárias - FAETEC

Professor de Estatística no Curso de Pós-Graduação de Ciências Ambientais e da Saúde - ISECENSA

Resumo

Este artigo enfatiza a aplicação da estatística como ferramenta fundamental aos diversos estudos científicos. Observa-se, atualmente, considerável acréscimo de publicações sobre as aplicações da análise de regressão, principalmente para estudos biológicos. É discutido, detalhadamente, um exemplo sobre a análise da regressão de y sobre x a partir da soma de quadrado dos desvios calculados para análise da variância. Nesse exemplo, foi calculado o número de graus de liberdade, um termo usual em estatística, que será posteriormente comentado em detalhe. O quadrado médio para regressão foi testado contra o quadrado médio do erro por meio do teste F (tabelado) e os coeficientes de regressão foram calculados pelo método da soma dos mínimos quadrados. Muitos irão preferir métodos mais elaborados, mas esse se mostra altamente suficiente e fácil. Finalmente, comenta-se sobre a confiabilidade dos coeficientes de regressão estimados. Com esse artigo, espera-se contribuir para o uso do ajustamento de curvas pelos pesquisadores, com o propósito de economizar tempo e trabalho.

Palavras-chave:

Análise de regressão, estatística.

Correspondência:
Rua Salvador Correa, 139 - Centro
28035-310 - Campos dos Goytacazes - RJ
Telefone: +55 (22) 2726.2727
Fax: +55 (22) 2726.2720
www.isecensa.edu.br
e-mail: isecensa@isecensa.edu.br

Regression analysis: application in biology

CLAUDIO LUIZ MELO DE SOUZA
D.Sc. in Plant Protection
Master of the Superior Institute of Agrarians Sciences Technology – FAETEC
Master of statistic in Pos-Graduate Course of Environments Sciences and Health - ISECENSA

Abstract

This article emphasizes the application of the statistic like fundamental instrument to several scientifics studies. Actually, considerable literature has grown up around the application of regression analysis, especially to biologics studies. It is discussed in detail an example about analysis regression of Y on X from sum of squared deviations calculated to analysis of variance. In this example, the number of degrees of freedom, a term usual in textbooks, was calculated and to be referred to tater in more detail. Mean square of regression was tested by F test against mean square of error and, the regression coefficients were calculated by sum squares minimums method. Many will prefer the more elaborate methods, but the ones shown are highly sufficient and simple. Finally, it is referred about of the accuracy of regression coefficients estimated. It hopes to contribute for the use of fitting equations to data by researchers to save time and work.

Key works:

Regression analysis, statistic.

Correspondence:
Rua Salvador Correa, 139 - Centro
28035-310 - Campos dos Goytacazes - RJ
Phone number: +55 (22) 2726.2727
Fax: +55 (22) 2726.2720
www.isecensa.edu.br
e-mail: isecensa@isecensa.edu.br

A análise de regressão e suas aplicações

Chamamos ajustamento de curvas o fato de utilizarmos dados observados em uma pesquisa para chegar a uma equação matemática que descreva a relação entre duas variáveis. O ajustamento de curvas para eventos biológicos é muito útil e simples, ainda que muitos pesquisadores relutem em utilizá-lo por desconhecimento (GUNST & MASON, 1980).

Diversos fenômenos biológicos podem ser expressos por equações matemáticas, facilitando o entendimento das relações entre grandezas conhecidas e aquelas que queremos estimar. Isso ocorre com freqüência e exemplificar todas as possibilidades enumeraria uma página inteira. Por exemplo, o ajustamento de curvas é um instrumento imprescindível, quando sabemos que a medida cefálica de um animal arisco (ave, morcego, peixe, etc.) apresenta uma afinada relação com outras medidas corporais (peso total, comprimento da asa, altura, etc), que na prática, representam dificuldades de execução, como a fuga do animal da balança ou a necessidade de anestesiá-lo, colocando-o em risco de vida. Nesse caso, o ajustamento de curva entre a medida cefálica e outras medidas corporais economiza tempo e trabalho, pois a medida cefálica é executada e as demais medidas corporais são estimadas por meio de equações matemáticas com segurança e precisão. Nesse artigo, visando à simplificação, vamos concentrar-nos principalmente em equações lineares com duas incógnitas, do tipo: $y = a + bx$, onde a incógnita a é o intercepto de y, (ou seja, o valor de y para o qual x = 0, ou ainda, o ponto da reta que toca o eixo de x) e b é o coeficiente angular da reta (ou seja, a variação de y que acompanha um aumento de uma unidade em x, ou ainda, a intensidade com que x interfere em y).

As equações lineares são úteis porque muitas relações biológicas têm efetivamente esta forma, pois quando grafadas constituem uma reta, além de representarem aproximações de relações com alta precisão. Na prática, os valores de a e b são calculados com base em dados observados e, uma vez estimados, podemos introduzir, na equação, valores de x, calculando os correspondentes valores preditos de y. Quando assim fazemos, encontramos três tipos de problemas: 1) decidir se existe regressão linear entre as grandezas investigadas, 2) calcular a equação que melhor descreve o evento e 3) investigar problemas relativos ao mérito (confiabilidade) da equação.

Vamos resolver um estudo de caso para exemplificar a solução desses três problemas: Em um levantamento da fauna de uma caverna, um biólogo deseja estimar o peso dos morcegos (y) capturados, utilizando o diâmetro da cabeça (x) dos mesmos e para tanto, captura 10 espécimes e obtém as seguintes medidas contidas na Tabela 1.

Diâmetro cefálico (cm)		Peso corporal (g)	Quadrados e produtos de x e y		
x		y	x^2	y^2	xy
3		57	9	3.249	171
4		78	16	6.084	312
4		72	16	5.184	288
2		58	4	3.364	116
5		89	25	7.921	445
3		63	9	3.969	189
4		73	16	5.329	292
5		84	25	7.056	420
3		75	9	5.625	225
2		48	4	2.304	96
Somatórios:	35	697	133	50.085	2554

Fonte: adaptado de FREUND, 2000.

As calculadoras científicas, em função estatística, processam as somas que podem ser acumuladas diretamente, não havendo necessidade de percorrer todos os detalhes do cálculo. Fazendo os somatórios:

$\Sigma x = 35$, $\Sigma = 697$, $\Sigma x^2 = 133$, $\Sigma y^2 = 50.085$, $\Sigma xy = 2.554$, $n = 10$, $\bar{x} = \Sigma x/n = 3,5$ e $\bar{y} = \Sigma y/n = 6,97$

O primeiro problema é facilmente resolvido calculando-se o quadro abaixo da análise da variância para efeito de regressão, considerando-se as seguintes fontes de variação (FV):

QUADRO 1. Resultados dos cálculos para análise da variância da regressão baseados nos dados da Tabela 1.

				Teste F		
FV	GL	SQ	QM	calculado	tabelado	Sig.
Efeito de regressão	1	1248,60	1248,60	39,09	5,32	*
Resíduo	8	255,50	31,94			
Total	9	1504,10				

Como calcular o quadro acima?:

GL (Grau de Liberdade). Esse artifício matemático considera a perda de um evento observado (n-1) para efeito de cálculos. Com isso, aumenta o rigor dos testes de hipóteses, como o Teste F que veremos mais adiante (FREUND, 2000, PIMENTEL-GOMES, 2000). Fazendo:

GL para regressão, com duas incógnitas: (n-1) = (2 – 1) = 1, onde n = nº de incógnitas;

GL total com 10 dados: (n-1) = (10 – 1) = 9,

onde n = número de dados observados e

GL residual é obtido por diferença entre os anteriores: 9 – 1 = 8.

SQ (Soma de Quadrados). Esse artifício matemático permite o cálculo da soma de todos os desvios em torno da média geral e da equação de regressão estimada, sem que o resultado final seja nulo, pois desvios negativos elevados ao quadrado tornam-se positivos evitando-se a subtração de desvios (FREUND, 2000). Para esclarecer, veja na Figura 1, o exemplo para apenas um ponto observado com as coordenadas (x,y). Observe que a média geral é a melhor estatística para representar todos os dados, caso não haja efeito de regressão, ou seja, se acréscimos em x não alterarem os valores de y. Então para qualquer dado de x, a média geral de y será assumida como a melhor medida. A conseqüência disso é uma reta paralela ao eixo de x (reta tracejada). Agora, observe que, a equação apresentada (y = 31,53 + 10,90x) e que mais à frente iremos calcular, está representada pela reta contínua. Essa reta não passa pelo ponto observado na coordenada x,y. Uma equação perfeita, geralmente improvável para eventos biológicos, geraria uma reta que passaria sobre todos os dados observados. Portanto, nessa equação há imperfeições (erros ou resíduos devidos ao acaso), o que por enquanto não a desclassifica como útil.

Bem, quanto ao SQtotal = $\sum(yO-yM)^2$, ou seja, a soma dos quadrados das distâncias entre os pontos observados e o ponto médio (PIMENTEL-GOMES, 2000). Esse termo mede a variação total dos valores de y em relação à média geral. Para fazermos o somatório dos desvios ao quadrado, utilizaremos uma fórmula mais prática:

$$SQtotal \quad ou \quad S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n},$$

fazendo: $S_{yy} = 50.085 - (697)^2/10 = 1504,10$

$SQregressão=(S_{xy})^2/S_{xx}$:

Onde, $S_{xy}=\sum xy^2 - \frac{(\sum x)(\sum y)}{n}$,

fazendo: $S_{xy}= 2554–(35x697)/10=114,50$

e $S_{xx}= \sum x^2 - \frac{(\sum x)^2}{n}$,

fazendo: $S_{xx}= 133 –(35)^2/10 = 10,50$

$SQregressão= (S_{xy})^2/S_{xx}$,

substituindo: $(114,5)^2/10,50 = 1248,60$

SQresíduo é obtido pela diferença entre SQtotal e SQregressão = 1504,10 - 1248,5 = 255,5

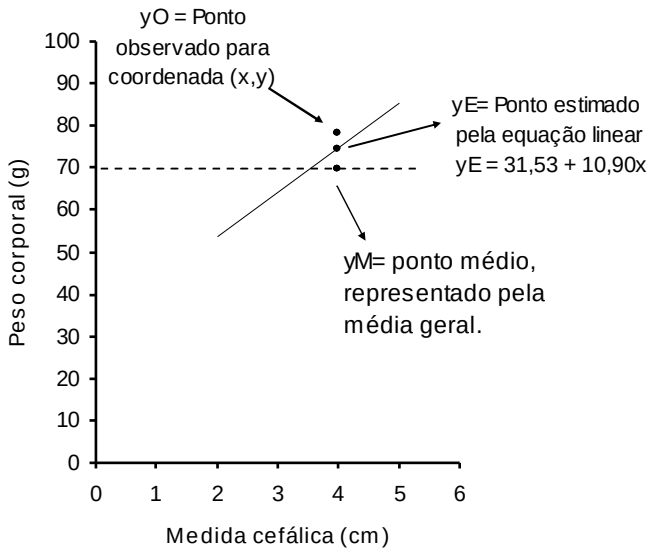

FIGURA 1. Exemplo de medida cefálica (x) e peso corporal (y) de morcegos para explicar as relações matemáticas entre as somas de quadrados dos desvios observados em relação à média [SQtotal =$\sum(yO-yM)^2$], dos desvios estimados pela equação de regressão linear em relação à média [SQregressão = =$\sum(yE-yM)^2$] e aos resíduos devidos ao acaso, que configuram o erro entre os pontos observados e estimados pela equação [SQresíduo=$\sum(yO-yE)^2$]. Fonte: Adaptado de CHARTERJEE & PRICE, 1991 e FREUND, 2000.

QM (Quadrados Médios). Refere-se à variância do efeito em estudo, ou seja, variância devido ao efeito de regressão e devido ao acaso (PIMENTEL-GOMES, 2000). A variância da regressão deve ser maior do que o efeito ao acaso para assumirmos a existência de correlação entre as grandezas x e y. Para calcularmos, dividimos as SQ pelos respectivos graus de liberdade, fazendo:
QMregressão = 1248,6/1 = 1248,6 e QMresíduo = 255,5/8 = 31,94.

O Teste F. Trata-se da relação entre variâncias estimadas para o efeito de regressão e do resíduo (Fcalculado=QMregressão/QMresíduo) (PIMENTEL-GOMES, 2000). Admitindo-se a hipótese de nulidade, isto é, não existe efeito de regressão e que as estimativas são independentes e referem-se ao mesmo parâmetro, não deveriam diferir a não ser devido ao acaso. Para compará-las usamos o Teste F (calculado e tabelado). O Ftabelado é um valor crítico de refutação da hipótese de nulidade e pode ser obtido em tabelas estatísticas com proba-

bilidade de erro de 5% ou 1% (respectivamente, α=0,05 ou α=0,01). Essas probabilidades de erros são convencionalmente mais utilizadas para os eventos biológicos. Para encontrar o Ftabelado utilizamos o GLregressão (numerador) e GLresíduo (denominador) como indicadores da rigorosidade do teste. Assim, nesse exemplo, ao utilizarmos uma tabela de valores críticos de F (α=0,05), e alcançarmos o cruzamento do GLnumerador = 1 e GLdenominador = 8, encontraremos o valor de Ftabelado = 5,32. Sendo o valor calculado maior do que o valor tabelado, refuta-se a hipótese de nulidade e aceita-se que existe efeito de regressão entre x e y, ou seja, alterações no diâmetro cefálico dos morcegos que ocasionam diferenças no peso corporal. Cabe-nos responder sobre a intensidade dessa relação calculando os valores de a e b na equação y = a + bx, objeto do segundo problema.

O segundo problema é facilmente resolvido por sistemas de equações normais ou pelo método dos mínimos quadrados (FREUND, 2000). Nas equações normais, obtemos:

$$\Sigma y = na + b(\Sigma x)$$
$$\Sigma xy = a(\Sigma x) + b(\Sigma x)^2$$

Substituindo:

$$697 = 10a + 35b$$
$$2.554 = 35a + 133b$$

Resolvendo essas duas equações simultâneas pelo chamado método da eliminação (dentre os vários métodos existentes), obtemos a = 31,55 e b = 10,90. Como alternativa, apresentamos a seguir, diversas fórmulas de grande utilidade para cálculos que teremos de efetuar em problemas de mínimos quadrados:

$b = S_{xy} / S_{xx}$, substituindo: 114,5/10,5 = 10,90

$a = \bar{y} - b\bar{x}$, substituindo: 69,7 - 10,90 x 3,5 = 31,55

Há diversos programas de computador para o ajuste de retas de mínimos quadrados. Uma vez determinada a equação de uma reta de mínimos quadrados, podemos aplicá-la para fazer predições. Assim, poderá o biólogo a partir desse exemplo, medir o diâmetro cefálico e estimar o peso corporal dos morcegos economizando tempo e esforços, mas com qual confiabilidade?

Esse é o terceiro problema que pode ser facilmente resolvido pelo cálculo do coeficiente de determinação da reta (r^2). O valor de r^2 estabelece a relação entre as somas de quadrados dos desvios da regressão e do efeito total (r^2 = SQregressão/SQtotal). Anteriormente, afirmamos que uma equação perfeita seria capaz de gerar uma reta que passaria sobre todos os dados observados. Conseqüentemente, essa equação não teria erros ou resíduos devidos ao acaso, pois SQregressão seria igual a SQtotal. Nesse caso, r^2 seria 1, ou seja, 100% dos dados observados entre x e y foram ajustados pela equação estimada. No exemplo que calculamos, r^2 = 1.248,6/1.504,1 = 0,83, ou seja, 83% da relação entre a medida cefálica e o peso corporal será explicada pela equação y = 31,55 + 10,90x. Caso o pesquisador julgue que esse valor atenda às necessidades de suas avaliações, estará resolvido o problema. Se não, terá que observar grandezas de maior afinidade para atingir seus objetivos.

A precisão dos coeficientes de regressão estimados (a e b) pode ser obtida calculando-se o erro-padrão da estimativa (Se) pela fórmula:

$$Se = \text{raiz quarada de } \{[S_{yy} - (S_{xy}^2/S_{xx})]/(n-2)\} = 5,65$$

Nesse exemplo, temos n = 10; S_{yy} = 1.504,10; S_{xy}^2 = 114,50 e S_{xx} = 10,50.

Admitindo-se que corresponda ao desvio-padrão verdadeiro, podemos fazer inferências sobre os coeficientes de regressão, estabelecendo-se limites de confiança, ou seja, intervalos de confiança (IC) que permitiriam a comparação dos

coeficientes de regressão oriundos de dois estudos, tratamentos ou outra comparação desejada pelo pesquisador. Este será o assunto do tópico seguinte.

Limites de confiança para os coeficientes de regressão a e b (CHARTERJJE & PRICE, 1991) – esses limites são calculados assumindo-se que o erro-padrão estimado (Se) corresponde ao desvio-padrão verdadeiro e estão associados à probabilidades de erro alfa (α) por meio de tabelas que determinam os valores críticos de distribuição t indicados por número de graus de liberdade. Portanto, podem ser definidos como o intervalo que contém os valores verdadeiros de a e b. Podem ser obtidos pelas fórmulas:

$$ICa = a \pm t_{(a=0,05/2)} . Se . \text{raiz quadrada de } [(1/n) + (M^2/S_{xx})]$$

$$[$$

$$ICb = b \pm t_{(a=0,05/2)} . Se / \text{raiz quadrada de } S_{xx}$$

Onde:

-O valor $t_{(a=0,05/2)}$ é obtido em tabelas de t, comuns nos livros de estatística. Para consultar a tabela admite-se 5% de probabilidade de erro alfa (= 0,05) e o número de dados observados – 2 graus de liberdade. Preste atenção: para tabelas com distribuição t bilateral deve-se dividir por 2 a probabilidade de erro e portanto consultar tabelas bilaterais ao nível de 0,025. Assim, para esse exemplo, o valor t = 2,306, pode ser encontrado em tabela bilateral para valores críticos de t com alfa de 0,025 e 8 graus de liberdade;

-O valor M^2 corresponde ao quadrado da média geral, que no exemplo é o somatório do diâmetro cefálico (35) dividido por 10 e elevado ao quadrado = $(35/10)^2 = 12,25$;

-O valor de Se foi calculado no tópico anterior e

-O valor de S_{xx} foi anteriormente calculado como 10,5 (veja no cálculo de SQtotal).

Substituindo:

$ICa = 31,55 \pm 2,306 . 5,651 . \text{raiz quadrada de } [(1/10) + (12,25/10,5)]$;

$ICa = 31,55 \pm 14,67$; e

$ICb = 10,90 \pm 2,306 . (5,651/\text{raiz quadrada de } 10,5)$;

$ICb = 10,90 \pm 4,02$

Na comparação dos coeficientes de regressão (a e b) de duas equações, pode-se inferir que quando não há sombreamento, entre os intervalos de confiança dos coeficientes, essas equações começam em pontos distintos e quando não há sombreamento dos intervalos de b, essas equações possuem inclinações diferentes. Caso os intervalos de confiança de a e b, sejam diferentes nas duas equações, podemos inferir que essas equações provêm de fenômenos ou tratamentos distintos.

Desta forma, concluímos que por meio de cálculos relativamente simples, principalmente com os atuais recursos computacionais, podemos enriquecer a apresentação dos dados de uma pesquisa, além de economizarmos tempo e esforços. A estatística é uma importante ferramenta suporte para pesquisas científicas. A partir dela, podemos nos apaixonar, ainda mais, por aquilo que nos propomos a investigar. Nela não está contida apenas a matemática, mas a confirmação daquilo que vivenciamos em nossos estudos biológicos. Sua aplicabilidade é capaz de transformar a matemática em algo prazeroso de aprender e utiliza.

Referências bibliográficas

CHARTERJEE, S., AND PRICE, B., *Regression Analysis by Example*, 2ª ed. New York: John Wiley & Sons Inc., 1991.

FREUND, JOHN E. *Estatística aplicada: economia, administração e contabilidade*. John E. Freund e Gary A. Simon; trad. Alfredo Alves de Farias. 9ª ed – Porto Alegre: Bookman, 2000.

GUNST, R.F., AND MASON, R.L., *Análise de regressão and its applications. A Data-oriented Approach.* New York: Marcel Dekker, Inc., 1980.

PIMENTEL-GOMES, F. *Estatística Experimental*, 14ª ed. Piracicaba:Degaspar ed., 2000.