Biometria

Regressão e Correlação

(Leitura complementar ao capítulo 7)

Sumário:

Definições

Diz-se que existe *correlação* entre duas ou mais variáveis quando as alterações sofridas por uma delas são acompanhadas por modificações nas outras. Ou seja, no caso de duas variáveis *x* e *y* os aumentos (ou diminuições) em *x* correspondem a aumentos (ou diminuições) em *y*.

Assim, a correlação revela se existe uma *relação funcional* entre uma variável e as restantes..

Note-se que a palavra *regressão* em Estatística corresponde à palavra *função* em Matemática. Ou seja, enquanto o matemático diz que *y* é função de *x*, o estatístico fala em regressão de *y* sobre *x*.

Reta de regressão

Uma função muito interessante é a que representa a *linha reta*, cuja expressão matemática é

$y = a + bx$ em que	
y =	variável dependente
x =	variável independente
a =	constante = intercepto (ponto em que a reta corta o eixo dos y)
b =	constante = coeficiente de regressão

sendo que o *intercepto a* pode ser calculado a partir de:

$$a = \bar{y} - b.\bar{x}$$

Ressalte-se que necessariamente o ponto determinado pela média das variáveis está contido na reta.

A melhor reta que descreve a regressão

(Se desejar mais detalhes sobre como criar gráficos de retas, clique aqui).

Supondo uma amostra em que um caráter métrico tenha a seguinte distribuição de idades e larguras de um órgão:

Idade (x)	*Largura (y)*	
1	30	Em que:
2	40	total de larguras = 520
3	50	total de idades = 36
4	60	
5	70	média de larguras = 65
6	80	média de idades = 4,5
7	90	
8	100	Supondo a = 20 e b = 10

Quando se deseja desenhar uma reta , para facilitar, atribui-se 2 valores de x próximos aos extremos dos dados. Depois, usa-se esses valores na equação:

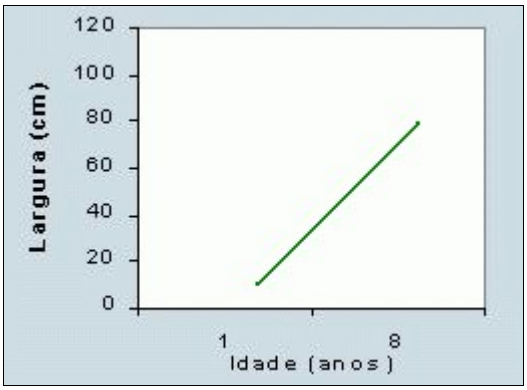
$$y = \overline{y} + b.(x - \overline{x})$$

Portanto,

para a idade x = 1 ano, largura: y = 65 + 10 (1 - 4,5) = 30

para a idade x = 8 anos, largura: y = 65 + 10 (8 - 4,5) = 100

E chega-se ao seguinte gráfico:



Essa reta, que passa pelos *pontos médios* dos valores de *x* e *y* é a *melhor reta* que descreve a regressão.

Evidentemente, pode-se usar o mesmo processo em gráficos feitos em programas computacionais. (No Calc veja como criar gráficos clicando aqui.)

Proporcionalidade: Direta e Inversa

Quando se observa o coeficiente de regressão *b* e o sentido da reta pode-se concluir se existe correlação entre as variáveis e qual é o sentido da correlação.

Nesse caso, verifica-se que a aumentos na variável Idade (*x*) correspondem aumentos na variável Largura do órgão (*y*). Assim sendo, elas têm o *mesmo sentido* de variação. Essa é uma

correlação *positiva*.

Evidentemente, uma correlação será *negativa* quando a aumentos na variável *x* corresponderem diminuições na variável *y*. Nesse caso, as variáveis estudadas variam em sentidos *opostos*.

Paralelamente, percebe-se que quando a reta de regressão em *y* é paralela ao eixo dos *x* (*b* = 0) não há correlação. Portanto, para que exista correlação é necessário que a reta corte o eixo dos *x* em algum ponto (b ≠ 0). Assim, quando há correlação, a reta de regressão em *y* não é paralela ao eixo dos *x*.

Existe correlação?

Para se decidir sobre a existência de correlação e o sentido da variação da reta de regressão, calcula-se b e o erro de b.

Depois efetua-se um teste *t*, testando as seguintes hipóteses:

H0: b = 0, ou seja, *H. Nula*: a reta de regressão em *y* é paralela ao eixo dos x

H0: b ≠ 0, isto é, *H. Alternativa*: a reta de regressão em *y* não é paralela ao eixo dos x.

Como calcular

Recordando que as somatórias de quadrados (SQ) e de produtos (SP) são calculadas por:

$$SQx = \sum x^2 - [(\sum x)^2 / n]$$

$$SQy = \sum y^2 - [(\sum y)^2 / n]$$

$$SP = \sum (x.y) - n \bar{x} . \bar{y}$$

O coeficiente de regressão, *b*, pode ser calculado a partir de várias fórmulas:

$$b = \sum [(x - \bar{x}) (y - \bar{y})] / \sum (x - \bar{x})^2$$

ou

$$b = ((\sum (x.y) - n. \bar{x} . \bar{y}) / \sum x^2 - [(\sum x)^2 / n]$$

ou

$$b = SP / SQx$$

O erro de b também pode ser calculado de maneiras diferentes:

$$s_b = raiz\ (s_{yx} / SQy)\ ou$$

$$s_b = raiz\ \{(SQy - b.SP) / [SQx\ (n - 2)]\}$$

Para se testar a significância de b, ou seja, para testar se pode ser considerado ou não como significativamente diferente de zero, calcula-se *t*, com *GL = n - 2,* sendo:

$$t = b / s_b$$

Para encontrar o t crítico, consulta-se a tabela de *t*, e obedece-se o seguinte critério:

$t < t_c$ *t não é* significativo *b* não é significativamente diferente de 0 (a reta é paralela ao eixo dos x)	t_c	$t > t_c$ *t é* significativo *b* é significativamente diferente de 0 (a reta não é paralela ao eixo dos x)

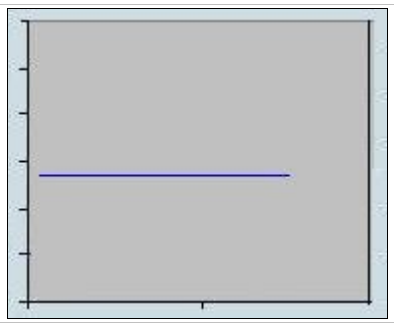
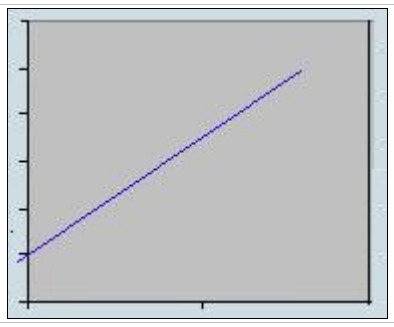
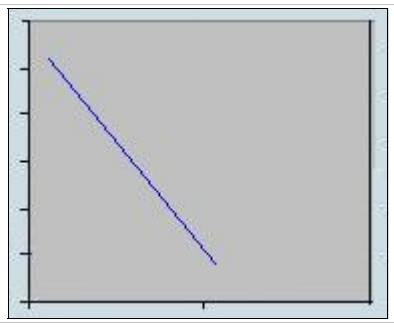
Portanto:

1. Se *t* não for significativo os caracteres não estão correlacionados: ($t = 0$)

   Se *t* for significativo os caracteres estão correlacionados: ($t \neq 0$)

2. Sendo $t \neq 0$, se $b < 0$ a correlação é negativa. Os caracteres variam em sentidos opostos.

   Sendo $t \neq 0$, se $b > 0$ a correlação é positiva. Os caracteres variam no mesmo sentido.

		
ausência de correlação	*correlação positiva*	*correlação negativa*
t = 0, qualquer b	$t \neq 0$, $b > 0$	$t \neq 0$, $b < 0$
Não há sentido de variação	As variáveis variam no *mesmo* sentido	As variáveis variam em sentidos *opostos*

Exemplo: Os seguintes dados foram obtidos amostrando dimensões do mesmo órgão de 10 indivíduos.

comprimento	*x*	40	25	65	75	65	40	50	40	15	25
largura	*y*	25	15	50	65	50	25	40	40	15	15

que geraram os seguintes valores:

$\sum$	440		$\sum y$	340		n	10
$\bar{x}$	44		$\bar{y}$	34		$\sum(x.y)$	17950
$\sum x^2$	22850		y^2	14350		n. $\bar{x}$ $\bar{y}$	14960
$\sum x^2 / n$	19360		$\sum y^2 / n$	11560		SP	2990
SQx	3490		SQy	2790		SP^2	8940100
s^2_x	387,78		s^2_y	310			

Exercício: Confira os cálculos abaixo e *complete* as seguintes frases:

1. Existe correlação entre os caracteres da amostra? Porque?

b = SP / SQx = 2990 / 3490 = 0,86

s_b = *raiz* (SQy - b.SP) / [SQx (n -2)]

= *raiz* (2790 - 0,86 2990) */* [3490 (10 -2)] = 0,09

t = b / s_b = 0,86 / 0,09 = 9,556.

Consulta-se a tabela de *t*

Sendo que: G.L. = _____ t_c = _____ P = 0,001

Resposta: Sendo t = _____ sua probabilidade é _____ . Como t é _____ (maior - menor) que t_c (t_c = _____), conclui-se que t _____ (é - não é) significativo, portanto, _____ (há - não há) correlação entre as variáveis x e y.

Como b é _____ (igual a - diferente de) zero, a reta será _____ (paralela - não paralela) ao eixo dos x e _____ (ascendente - descendente), já que b é _____ (positivo – negativo).

2. Qual o sentido da variação desses caracteres?

A correlação é _____ (positiva - negativa) , pois b (_____) é _____ (positivo - negativo). Portanto, o comprimento e a largura desse órgão variam _____ (no mesmo sentido - em sentidos postos), ou seja são _____ (diretamente - inversamente) proporcionais.

3. Qual a reta de regressão que melhor se ajusta aos dados da amostra?

Atribui-se 2 valores extremos de x, e substitui-se em y = $\bar{y}$ + b.(x - $\bar{x}$) . Por exemplo:

para *x = 10*, y = 34 + 0,86.(10 - 44) = 4,8 e

para *x = 80*, y = 34 + 0,86.(80 - 44) = 65,0

Com esses valores crie o melhor gráfico que representa esses dados. (Veja como clicando aqui).

Para facilitar os cálculos utilize uma planilha especial:

Regressão e Correlação
Copie a planilha comprimida em formato livre ods

http://www.cultura.ufpa.br/dicas/biome/biozip/regre01.zip

Coeficiente de correlação linear de Pearson (r)

Pode ser obtido a partir de diferentes fórmulas:

r = n $\sum(x.y) - (\sum x)\text{-}\sum y$ / *raiz* $[n.\sum x^2 - (\sum x)^2]\ [\ n.\sum y^2 - (\sum y)^2\ ]$

$r = (\sum(x.y) - n.\bar{x}.\bar{y}) / [(n - 1).\sigma x.\sigma y]$

r = *raiz* (b.SP / SQy)

r = b.(σx / σy)

Observando as duas últimas fórmulas rapidamente percebe-se que se não houver correlação entre x e y, ou seja, se *r = 0*, então *b = 0* e a reta será paralela ao eixo dos x.

O coeficiente ***r*** varia entre -1 e +1. Portanto, a correlação pode ser:

-1	-0,95	-0,50	-0,10	*0*	0,10	0,50	+0,95	+1
neg	neg	neg	neg	*ausência*	pos	pos	pos	pos
perfeita	forte	moderada	fraca		fraca	moderada	forte	perfeita

Para testar a significância usamos um *teste t*. Estabelecemos as hipóteses:

H_0: r = 0, ou seja, *H. Nula*: Não há correlação entre as variáveis x e y.

H_a: r ≠ 0, isto é, *H. Alternativa:* Há correlação entre as variáveis x e y.

Calcula-se *t*, com GL = n-2, por meio da seguinte fórmula:

$$t = r \cdot \text{raiz}\,[(N - 2) / (1 - r^2)]$$

Coeficiente de determinação

O coeficiente de determinação é simbolizado por r^2 e indica *quanto* da variação total é comum aos elementos que constituem os pares analisados.

Assim, a *qualidade* da regressão é indicada por este coeficiente.

$$r^2 = \text{Variação } \textit{explicada} \text{ de Y / Variação } \textit{total} \text{ de Y}$$

É importante notar que r^2 varia entre 0 (zero) e 1 (um).

Evidentemente, quanto mais próximo da unidade for o coeficiente de Determinação, tanto maior será a validade da regressão.

Exemplo 1:

Supondo que numa certa amostra tivessem sido obtidos os seguintes valores:

$$b = 0,86;\ SP = 2990;\ SQy = 2790$$

Estima-se r = *raiz* (b.SP / SQy), r = *raiz* (0,86.2990 / 2790), r = 0,96

Portanto, $r^2 = 0,92$

1 - 0,92 = 0,08, ou seja, 8%

Assim, pode-se dizer que apenas 8% da variância da regressão não depende das variáveis estudadas.

Exemplo 2:

Dados obtidos de 7 pares de pai-filho, amostrando o número de anos de escola cursados pelo pai (x) e o número de anos de escola cursados pelo filho (y). Qual é o valor do coeficiente de correlação entre esses dados? Qual é o seu significado?

x	x^2	y	y^2	$x.y$
12	144	12	144	144
10	100	8	64	80
6	36	6	36	36
16	256	11	121	176
8	64	10	100	80
9	81	8	64	72
12	144	11	121	132
x = 73	$\sum x^2 = 825$	$\sum y = 66$	$\sum y^2 = 650$	$\sum (x.y) = 720$

$r = N .\sum xy - (\sum x) (\sum y) / raiz [N.\sum x^2 - (\sum x)^2] [N.\sum y^2 - (\sum y)^2]$

$r = 7 . 720 - 73 . 66 / raiz [7 . 825 - (73)^2] [7 . 650 - (66)^2]$

$r = + 0,754$

Para testar a significância usamos um teste t. Estabelecemos as hipóteses:

$$\mathbf{H_0: r = 0} \quad \text{e} \quad \mathbf{H_a: r \neq 0}$$

$t = r . raiz [(N - 2) / (1 - r^2)]$

$t = [+ 0,754. raiz[(7-2)] / (1 - 0,754^2)]$, portanto, $t = 2,581$

Verificando a tabela de *t*, com GL = 5 e a = 5%, $t_5 = 2,571$

Conclui-se que como *t* calculado é maior que t_c, pode-se rejeitar a hipótese nula ($r = 0$) e aceitar a hipótese alternativa em que $\mathbf{r \neq 0}$, admitindo-se que o número de anos de escola cursados pelo pai está positivamente correlacionado (r = + 0,754) ao número de anos de escola cursados pelo filho nesta amostra.

Como $r^2 = 0,5685$ e 1 - 0,5685 = 0,4315, pode-se dizer que nessa amostra, o número de anos de escola cursados pelo pai explica 56,85% da variância do número de anos de escola cursados pelo filho. Assim, 43,15% da variância da regressão depende de outras variáveis, não estudadas aqui.

Coeficiente de associação

Para verificar se dois caracteres qualitativos são interdependentes pode-se:

- empregar um teste de χ^2

- calcular o coeficiente de associação.

Yule propôs esse coeficiente e o chamou de Q , para homenagear um pioneiro da Estatística, Lambert A. J. Quételet (1796-1874).

Monta-se uma tabela 2 x 2 e designa-se as células pelas letras a, b, c e d, ficando a-d e b-c nas diagonais.

a	b
c	d

Obtém-se o coeficiente de associação Q por meio de:

$$Q = (ad - bc) / (ad + bc)$$

O desvio padrão de Q é obtido por:

$$s = (1 - Q^2) / 2 \; raiz \; (1/a + 1/b + 1/c + 1/d)$$

O intervalo de confiança de 95% de Q é obtido por:

$$Q \pm t.s$$

Exemplo:

Supondo que a distribuição de 200 pacientes adultos (92 homens e 108 mulheres) segundo as formas maligna e benigna de uma doença foi:

Forma / Sexo	Homens	Mulheres	Total
Maligna	60 *a*	40 *b*	100
Benigna	32 *c*	68 *d*	100
Total	92	108	200

Q = (ad - bc) / (ad + bc) = (60 x 68) - (40 x 32) / (60 x 68) + (40 x 32)
Q = (4080 - 1280) / (4080 + 1280) = 2800 / 5360
Q = 0,5224

O *desvio padrão* de Q é obtido por:
s = (1 - Q^2) / 2 . *raiz* (1/a + 1/b + 1/c +1/d)
s = (1 - $0,5224^2$) / 2 . *raiz* (1/60 + 1/40 + 1/32 +1/68)
s = 0,3635 . *raiz* (0,0167 + 0,0250 + 0,0312 + 0,01470)
s = 0,3635 . *raiz* 0,0876 = 0,3635 . 0,2960 = 0,1076

O intervalo de confiança de 95% de Q é obtido por:

$$Q \pm t.s = 0,5224 \pm 1,96 \times 0,1076$$

Portanto, o valor mínimo é 0,3115 e o valor máximo é 0,7333

Como o valor calculado de Q (0,5224) se encontra entre esses 2 valores (0,3115 e 0,7333), conclui-se que existe associação entre o sexo e as formas da doença, estando o sexo masculino associado à forma maligna, pois nesse sexo há maior frequência dessa forma.

Regressão múltipla

Quando se quer investigar se uma variável está correlacionada concomitantemente a várias outras, considera-se a primeira como variável dependente e as outras como variáveis independentes, e aplica-se aos dados a seguinte fórmula:

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + ... + b_nx_n$$

em que:

y = é a estimativa da variável dependente

x = variável independente

a = constante = intercepto múltiplo

b = constante = coeficientes de regressão

A análise de regressão múltipla é *trabalhosa* pois envolve a construção e multiplicação de matrizes tanto maiores quanto maior for o número de variáveis independentes analisadas. Assim, é necessário realizar tal análise em computadores. Portanto, aqui nos preocupamos com a interpretação de resultados de análise de regressão múltipla.

Exemplo

Em uma amostra de 36 hansenianos de sexo masculino tentou-se verificar se a quantidade de um certo medicamento presente no sangue 6 hs após a sua ingestão (variável dependente) está correlacionada com idade, peso corporal, duração da doença, anos de sulfonoterapia, valor do hematócrito, taxa de hemoglobina, nível de globulinas e nível de albumina (variáveis independentes).

	Quantidade do medicamento no sangue, após 6 hs de ingestão	*b*	s_b	$t_{(27)}$	*P*
x_1	idade	-0,0586	0,0542	-1,081	> 0,20
x_2	peso corporal	-0,0145	0,0374	-0,388	> 0,60
x_3	duração da doença	-0,0115	0,0468	0,246	> 0,80
x_4	anos de sulfonoterapia	-0,0894	0,0520	1,719	> 0,05
x_5	valor do hematócrito	-0,2317	0,0990	-2,340	*< 0,05*
x_6	taxa de hemoglobina	0,00005	0,0318	0,002	> 0,90
x_7	nível de globulinas	0,0695	0,0876	0,793	> 0,40
x_8	nível de albumina	-0,0079	0,0601	-0,131	> 0,80

que GL = N -1 - número de variáveis = 36 -1 - 8 = 27

Conclui-se que o nível sanguíneo desse medicamento, após 6 hs de ingestão *depende apenas* da variável $x_{5,}$ valor do hematócrito, pois entre todos os coeficientes de regressão calculados somente o b (-0,2317) dessa variável é significativamente diferente de zero (pois $t_{(27)} =$ -2,340), que determina uma probabilidade menor que 0,05.

Um cuidado a ser tomado *antes* de se realizar uma análise de regressão múltipla é calcular os coeficientes de correlação de todas as variáveis tomadas aos pares. Sabe-se que se houver duas ou mais variáveis com coeficientes de correlação muito altos (*r* igual ou superior a 0,95) elas interferirão nos cálculos de regressão múltipla. Se forem encontradas 2 ou mais variáveis nessa condição deve-se escolher apenas uma delas para o processamento da análise de regressão múltipla.

Regressão múltipla escalonada

É um modelo de regressão que permite selecionar as variáveis independentes por ordem decrescente de intensidade de correlação com a variável dependente. Matematicamente se chega à formula do coeficiente de determinação r^2, que mede o componente da regressão que decorre da variação concomitante das variáveis estudadas. (Como já foi visto, a expressão 1 - r^2 indica o quanto da variância não depende dessas variáveis em estudo).

Nessa análise se ordena as variáveis independentes de acordo com o valor de bSP. E, depois desse ordenamento se faz a análise de regressão simples da variável dependente sobre a independente que apresentou o maior valor de bSP. Finalmente,inicia-se a análise de regressão múltipla introduzindo as outras variáveis independentes pela ordem de grandeza decrescente do valor de bSP.

Ao final, verifica-se se o acréscimo de r^2 é significativo ou não por meio de um teste *t* :

$$t = (b / s_b)$$

A tabela que se segue mostra o resultado da análise de regressão múltipla escalonada aplicada aos mesmos dados que foram usados para a tabela anterior.

	Qtdd do medicamento no sangue após6 hs de ingestão	r^2	*Acrés-cimo*	*b*	s_b	$t_{(27)}$	*P*
x_5	valor do hematócrito	0,1750	---------	-0,2317	0,0990	*-2,340*	*< 0,05*
x_4	anos de sulfonoterapia	0,3133	0,1383	-0,0894	0,0520	1,719	> 0,05
x_3	duração da doença	0,3155	0,0022	-0,0115	0,0468	0,246	> 0,80
x_7	nível de globulinas	0,3472	0,0317	0,0695	0,0876	0,793	> 0,40
x_2	peso corporal	0,3613	0,0141	-0,0145	0,0374	-0,388	> 0,60
x_8	nível de albumina	0,3615	0,0002	-0,0079	0,0601	-0,131	> 0,80
x_6	taxa de hemoglobina	0,3517	0,0002	0,00005	0,0318	0,002	> 0,90
x_1	idade	0,3882	0,0265	-0,0586	0,0542	-1,081	> 0,20

———

Este "site", destinado prioritariamente aos alunos de Fátima Conti, pretende auxiliar quem esteja começando a se interessar por Bioestatística, computadores e programas, estando em permanente construção. Sugestões e comentários são bem vindos. Agradeço antecipadamente.

———

Endereço dessa página:

HTML: http://www.cultura.ufpa.br/dicas/biome/bioreg.htm

PDF: http://www.cultura.ufpa.br/dicas/pdf/bioreg.pdf

Última alteração: 4 nov 2009 (Solicito conferir datas. Pode haver atualização só em HTML)