UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE CIÊNCIAS NATURAIS E EXATAS
DEPARTAMENTO DE ESTATÍSTICA

APOSTILA DA DISCIPLINA
ANÁLISE DE CORRELAÇÃO E REGRESSÃO

Prof. Adriano Mendonça Souza
Profª Anaelena Bragança de Moraes
Profª Luciane Flores Jacobi
Profª Roselaine Ruviaro Zanini

1. INTRODUÇÃO

Em pesquisas, frequentemente, procura-se verificar se existe relação entre duas ou mais variáveis, isto é, saber se as alterações sofridas por uma das variáveis são acompanhadas por alterações nas outras. Por exemplo, peso vs. idade, consumo vs. renda, altura vs. peso de um indivíduo.

O termo correlação significa relação em dois sentidos (co + relação), e é usado em estatística para designar a força que mantém unidos dois conjuntos de valores. A verificação da existência e do grau de relação entre as variáveis é o objeto de estudo da correlação.

Uma vez caracterizada esta relação, procura-se descrevê-la sob forma matemática, através de uma função. A estimação dos parâmetros dessa função matemática é o objeto da regressão.

Os pares de valores das duas variáveis poderão ser colocados num diagrama cartesiano chamado "diagrama de dispersão". A vantagem de construir um diagrama de dispersão está em que, muitas vezes sua simples observação já nos dá uma idéia bastante boa de como as duas variáveis se relacionam.

Uma medida do grau e do sinal da correlação é dada pela covariância entre as duas variáveis aleatórias X e Y que é uma medida numérica de associação linear existente entre elas, e definida por:

$$\text{Cov(X, Y)} = \frac{1}{n}\left[\sum x.y - \frac{\sum x.\sum y}{n}\right].$$

É mais conveniente usar para medida de correlação, o coeficiente de correlação linear de Pearson, como estimador de ρ_{xy}, definido por:

$$r_{xy} = \frac{\text{Cov}(x, y)}{\hat{\sigma}_x \hat{\sigma}_y} = \frac{S_{xy}}{\sqrt{S_{xx}.S_{yy}}}$$

$$r_{xy} = \frac{\sum xy - \frac{\sum x.\sum y}{n}}{\left[\left[\sum x^2 - \frac{(\sum x)^2}{n}\right].\left[\sum y^2 - \frac{(\sum y)^2}{n}\right]\right]^{\frac{1}{2}}} = \frac{S_{xy}}{(S_{xx}.S_{yy})^{\frac{1}{2}}} = \sqrt{\frac{S_{xy}.S_{xy}}{S_{xx}.S_{yy}}} = \sqrt{\frac{b.S_{xy}}{S_{yy}}}$$

onde: as somas de quadrados são:

$$S_{xy} = \sum x.y - \frac{\sum x.\sum y}{n}; \qquad S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}; \qquad S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

n = número de pares das observações.

A partir de X e Y são determinadas todas as somas necessárias para este cálculo:

Y	**X**	**X^2**	**Y^2**	**X . Y**
:	**:**	**:**	**:**	**:**
:	**:**	**;**	**:**	**:**
Σy	**Σx**	**Σx^2**	**Σy^2**	**Σ(x . y)**

O coeficiente de correlação r_{xy} linear é um número puro que varia de –1 a +1 e sua interpretação dependerá do valor numérico e do sinal, como segue:

$r_{xy} = -1$ ⇨ correlação perfeita negativa
$-1 < r_{xy} < 0$ ⇨ correlação negativa
$r_{xy} = 0$ ⇨ correlação nula
$0 < r_{xy} < 1$ ⇨ correlação positiva
$r_{xy} = 1$ ⇨ correlação perfeita positiva
$0,2 < r_{xy} < 0,4$ ⇨ correlação fraca*
$0,4 < r_{xy} < 0,7$ ⇨ correlação moderada*
$0,7 < r_{xy} < 0,9$ ⇨ correlação forte*

* possui o mesmo significado para os casos negativos ou positivos.

Análise do Diagrama de Dispersão

O diagrama de dispersão mostrará que a correlação será tanto mais forte quanto mais próximo estiver o coeficiente de –1 ou +1, e será tanto mais fraca quanto mais próximo o coeficiente estiver de zero. (ver diagramas de dispersão).

a) Correlação perfeita negativa ($r_{xy} = -1$): Quando os pontos estiverem perfeitamente alinhados, mas em sentido contrário, a correlação é denominada perfeita negativa.

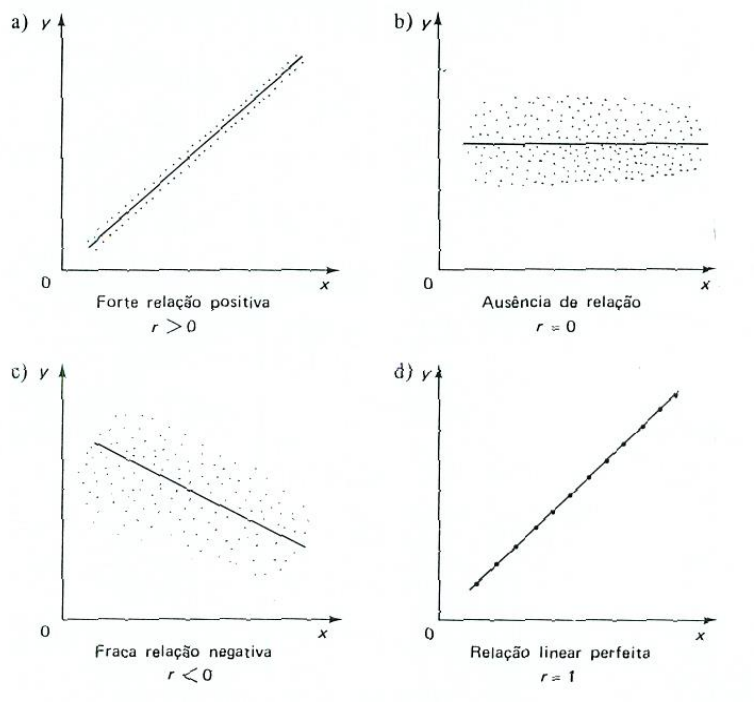
b) Correlação negativa ($-1 < r_{xy} < 0$): A correlação é considerada negativa quando valores crescentes da variável X estiverem associados a valores decrescentes da variável Y, ou valores decrescentes de X associados a valores crescentes de Y.

c) Correlação nula ($r_{xy} = 0$): Quando não houver relação entre as variáveis X e Y, ou seja, quando os valores de X e Y ocorrerem independentemente, não existe correlação entre elas.

d) Correlação positiva ($0 < r_{xy} < 1$): Será considerada positiva se os valores crescentes de X estiverem associados a valores crescentes de Y.

e) Correlação perfeita positiva ($r_{xy} = 1$): A correlação linear perfeita positiva corresponde ao caso anterior, só que os pontos (X, Y) estão perfeitamente alinhados.

f) Correlação espúria: Quando duas variáveis X e Y forem independentes, o coeficiente de correlação será nulo. Entretanto, algumas vezes, isto não ocorre, podendo, assim mesmo, o coeficiente apresentar um valor próximo de –1 ou +1. Neste caso a correlação é espúria.



OBSERVAÇÕES:

⇨ Correlação não é o mesmo que causa e efeito. Duas variáveis podem estar altamente correlacionadas e, no entanto, não haver relação de causa e efeito entre elas.

⇨ Se duas variáveis estiverem ligadas por uma relação de causa e efeito elas estarão, obrigatoriamente, correlacionadas.

⇨ O estudo de correlação pressupõe que as variáveis X e Y tenham uma distribuição normal.

⇨ A palavra simples indica que estão envolvidas somente duas variáveis na análise.

⇨ O coeficiente de correlação linear de Pearson mede a correlação sob o enfoque paramétrico.

⇨ O coeficiente de correlação de Spearman (correlação por postos) é o correspondente não-paramétrico.

⇨ Designações: $\text{Var(x)} = \frac{Sxx}{n}$; $\text{Var(y)} = \frac{Syy}{n}$; $\text{Cov(x,y)} = \frac{Sxy}{n}$.

⇨ Testar $\rho = 0$ é equivalente a testar $\beta = 0$ na equação de regressão.

1.1 TESTES DO COEFICIENTE DE CORRELAÇÃO – SIGNIFICÂNCIA DE r_{xy}

O coeficiente de correlação r_{xy} é apenas uma estimativa do coeficiente de correlação populacional ρ_{xy} e não devemos esquecer que o valor de r_{xy} é calculado com base em de "n"pares de dados.

Muitas vezes os pontos da amostra podem apresentar uma correlação e, no entanto a população não. Nesse caso, estamos diante de um problema de inferência, pois $r_{xy} \neq 0$ não é garantia de que $\rho_{xy} \neq 0$.

Podemos resolver o problema aplicando um teste de hipóteses para verificarmos se o valor de r_{xy} é coerente com o tamanho da amostra n, a um nível de significância α, que realmente existe correlação linear entre as variáveis.

H_0: $\rho = 0$ (não existe correlação entre X e Y)
H_1: $\rho \neq 0$ (existe correlação entre X e Y).

$$t_c = \frac{r_{xy} \cdot \sqrt{n-2}}{\sqrt{1-r^2_{xy}}} = \frac{r_{xy}}{S_r} \approx \text{ distribuição "t" de Student com n – 2 graus de liberdade}$$

onde: $S_r = \sqrt{\frac{1-r^2}{n-2}}$, erro padrão do coeficiente de correlação.

Exemplo 1.1: Vamos supor que 5 pessoas tenham sido questionadas a respeito das seguintes perguntas:

a) Durante quantos anos você freqüentou a escola regularmente?
b) Quantos livros você tem em sua biblioteca particular?
c) Caso exista correlação na amostra, verifique se existe também na população.

Sujeito	**A**	**B**	**C**	**D**	**E**	**Σ**
Freqüência à escola (x)	5	8	10	12	15	50
Número de livros (y)	10	30	45	50	75	210

Com base nestas respostas verifique se existe correlação entre os anos que estas pessoas freqüentaram a escola e o número de livros que possuem em sua biblioteca particular. Calcular:

$\sum x = 5 + 8 + 10 + 12 + 15 = 50$
$\sum y = 10 + 30 + 45 + 50 + 75 = 210$
$\sum x \,.\, y = 5.10 + 8.30 + 10.45 + 12.50 + 15.75 = 2.465$
$\sum x^2 = 5^2 + 8^2 + 10^2 + 12^2 + 15^2 = 558$
$\sum y^2 = 10^2 + 30^2 + 45^2 + 50^2 + 75^2 = 11.150$

$\bar{x} = \frac{\sum x}{n} = \frac{50}{5} =$ 10 anos de escola $\qquad \bar{y} = \frac{\sum y}{n} = \frac{210}{5} =$ 42 livros na sua biblioteca

$$S_{xy} = \sum x.y - \frac{\sum x.\sum y}{n} = 2465 - \frac{50.210}{5} = 365$$

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 558 - \frac{(50)^2}{5} = 58$$

$$S_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 11150 - \frac{(210)^2}{5} = 2330; \text{ logo:}$$

$$r_{xy} = \frac{Sxy}{\sqrt{Sxx.Syy}} = \frac{365}{\sqrt{58.2330}} = 0{,}9929$$ correlação forte positiva

Para efetuarmos o teste de hipótese, temos:

H_0: $\rho_{xy} = 0$ ⇨ ∄ correlação na população

H_1: $\rho_{xy} \neq 0$ ⇨ ∃ correlação na população

Utilizando-se $\alpha = 5\%$, teremos:

$$t_c = \frac{r_{xy}.\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} = \frac{0{,}99.\sqrt{5-2}}{\sqrt{1-(0{,}99)^2}} = 12{,}1382$$

$$t_{tab} = t_{\delta,\frac{\alpha}{2}} = t_{(n-2),\frac{\alpha}{2}} = t_{(5-2),\frac{5\%}{2}} = t_{3,\frac{5\%}{2}} = 3{,}182$$

Decisão: rejeita-se H_0, isto é, com 95% de confiança podemos concluir que a correlação na população não é nula, ou seja, ela é significativa.

1.3 - CORRELAÇÃO LINEAR POR POSTOS OU SPEARMAM - r_s

De todas as estatísticas baseadas em postos, o coeficiente de correlação por postos de Spearmam, foi a que surgiu primeiro, e é talvez a mais conhecida hoje. É uma medida de associação que exige que ambas as variáveis se apresentem em escala de mensuração pelo menos ordinal, de modo que os elementos em estudo possam dispor-se por postos em duas séries ordenadas.

Este teste não-paramétrico destina-se a determinar o grau de associação entre duas variáveis X e Y, dispostas em pontos ordenados, o objetivo é estudar a correlação entre duas classificações.

Resumo do Procedimento

1°) Dispor em postos as duas variáveis X e Y de 1 a n (n=número de pares de dados);
2°) Relacionar os n elementos, dar o posto de cada elemento;
3°) Determinar d_i = (posto x – posto y), d_i^2 e Σd_i^2;
4°) Se a proporção de ambas as variáveis X ou Y é grande então calcula-se r_s pela fórmula:

$$r_s = \frac{\sum x^2 + \sum y^2 - \sum d_i^2}{2.\sqrt{\sum x^2 . \sum y^2}},$$

onde:

$$\sum x^2 = \frac{n^3 - n}{12} - \sum T_x \qquad \sum y^2 = \frac{n^3 - n}{12} - \sum T_y$$

$T = \frac{t^3 - t}{12}$, onde t corresponde ao número de observações em determinado posto.

Caso contrário se aplica a fórmula: $r_s = 1 - \frac{6 \sum di^2}{n^3 - n}$.

5°) A significância de r_s é testada com $t_c = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$; com n-2 graus de liberdade, que é o mesmo teste anterior (Pearson).

Exemplo 1.2: Suponhamos que 8 professores de ciências do 1° grau tenham sido ordenados (classificados) por um júri, segundo sua capacidade técnica.

Professor	1	2	3	4	5	6	7	8	
Classificação	7^7	4^4	2^2	6^6	1^1	3^3	8^8	5^5	
Nota no curso	44^1	72^5	69^3	70^4	93^8	82^7	67^2	80^6	
d_i	6	-1	-1	2	-7	-4	6	-1	
d_i^2	36	1	1	4	49	16	36	1	$\sum d_i^2=144$

Os dados sugerem concordância entre o julgamento do júri e a nota obtida?

$r_s = 1 - \frac{6\sum d_i^2}{n^3 - n} = 1 - \frac{6(144)}{8^3 - 8} = 1 - \frac{864}{512 - 8} = 1 - \frac{864}{504} = 1 - 1{,}71$ = -0,71 9forte negativa)

Para verificar se existe correlação na população, utilizamos um teste de hipóteses.

H_0: $\rho = 0$ $\quad \alpha = 5\%$

H_1: $\rho \neq 0$ $\quad t_{tab} = t_{\delta, \frac{\alpha}{2}} = t_{(n-2), \frac{\alpha}{2}} = t_{(8-2), \frac{5\%}{2}} = t_{6, \frac{5\%}{2}} = 2{,}447$

$$t_c = \frac{r_s.\sqrt{n-2}}{\sqrt{1-r_s^{\,2}}} = \frac{-0{,}71 \cdot \sqrt{8-2}}{\sqrt{1-(-0{,}71)^2}} = \frac{-0{,}71 \cdot \sqrt{6}}{\sqrt{1-0{,}50}} = \frac{-0{,}71 \cdot 2{,}44}{\sqrt{0{,}5}} = \frac{-1{,}73}{0{,}7} = -2{,}47$$

Rejeita-se H_0, isto é, $\rho \neq 0$, logo existe correlação significativa entre as variáveis na população.

1.4 – EXERCÍCIO

Uma empresa está estudando como varia a demanda de certo medicamento em função de seu preço de venda. Para isto levantaram-se as seguintes informações:

Meses	J	F	M	A	M	J	J	A	S	O
Unidades vendidas	248	242	234	216	230	220	213	205	198	195
Preço de venda/unidade	162	167	165	173	170	176	178	180	182	187

a) Construa o diagrama de dispersão; b) Calcule o coeficiente de correlação e interprete o resultado; c) Faça o teste de significância para o coeficiente.