Regressão e correlação linear simples na Geografia

BARBARA-CHRISTINE NENTWIG SILVA (x)

Este trabalho corresponde a uma continuação do artigo "Métodos Quantitativos Aplicados em Geografia: uma introdução", (Geografia, 3(6), 1978). Ao invés de apresentarmos de uma só vez o numeroso conjunto de métodos quantitativos avançados aplicados em Geografia, optamos, nesta oportunidade, por discutir esses métodos separadamente, evitando a elaboração de trabalho demasiadamente extenso. Entretanto, os objetivos e a maneira de apresentação são os mesmos do artigo anterior, quais sejam, os de contribuir para o desenvolvimento do ensino e da pesquisa geográfica no Brasil, através de análise acessível da metodologia quantitativa.

As análises de correção e regressão simples são técnicas importantes para a interpretação dos dados e fenômenos geográficos envolvendo, ao mesmo tempo, duas variáveis ao invés de uma só, objeto de nossa preocupação em artigo anterior (Nentwig Silva, 1978). Além disto, o conhecimento destas análises é necessário para o emprego de outras técnicas mais avançadas, como, por exemplo, a análise fatorial.

1. *CONCEITO DE REGRESSÃO*

A nossa pergunta, na análise de regressão, é se é possível, saindo de uma variável, predizer a outra, ou seja, predizer que valor de uma variável Y corresponde a um valor dado de uma variável X. Normalmente X é a variável independente, Y a variável dependente. Segundo Sokal e Rohlf (1969), em regressão o objetivo é estimar o relacionamento de uma variável com a outra, exprimindo uma em termos de uma função linear (ou mais complexa) da outra.

Podemos chegar a esse objetivo através dos seguintes passos: depois de ter colecionado os dados para as duas variáveis sobre

(x) Professor-Adjunto do Instituto de Geociências da Universidade Federal da Bahia (Depto. de Geografia).

as quais queremos testar o relacionamento, podemos, como segundo passo, representar graficamente as duas variáveis utilizando o sistema de coordenadas cartesianas. Cada par de valores X_i, Y_i é indicado através de um ponto. Muitas vezes, como é o caso no gráfico 1, o resultado, denominado diagrama de dispersão, mostra na Geografia uma quantidade de pontos indicando que, com maiores valores de X, os valores de Y aumentam também. A impressão visual dá uma primeira indicação da relação, no nosso caso, uma relação linear entre as duas variáveis que deve agora ser expressa sob forma matemática.

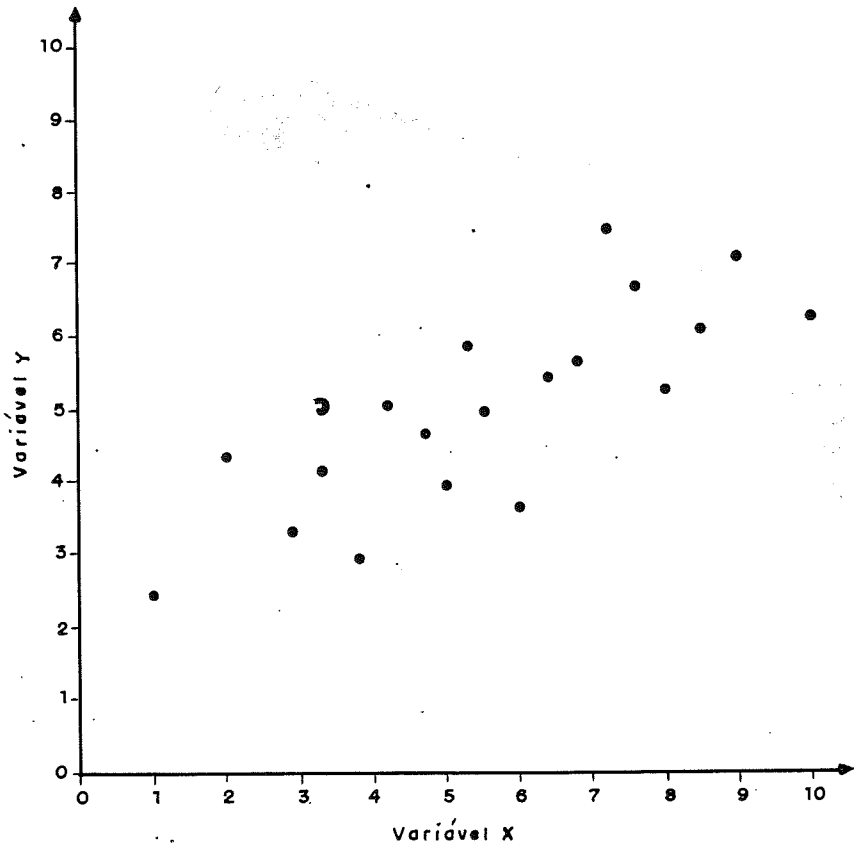

Figura 1 — Diagrama de dispersão (dados da Tab. 1).

Queremos traçar através dos pontos marcados uma linha reta de tal maneira que saindo de *X* conseguimos os melhores valores de predição de *Y*. Seria muito subjetivo traçar esta linha de ajustamento à mão livre. Diversos pesquisadores fariam linhas retas diferentes e, consequentemente, teriam equações também diferentes. A linha reta traçada com a equação $Y = a + bX$

deve ter a característica de ser a melhor reta de ajustamento, onde, como é definido, a soma dos quadrados de todos os desvios verticais dos valores reais da linha reta é um mínimo $(\Sigma (Y_i - Y_i)^2 = \text{mínimo}$. Trabalhamos com $\Sigma (Y_i - Y_i)^2$ porque a soma dos desvios $\Sigma (Y_i - Y_i)$ seria zero. A reta resultante é a chamada reta de mínimo quadrado (fig. 2).

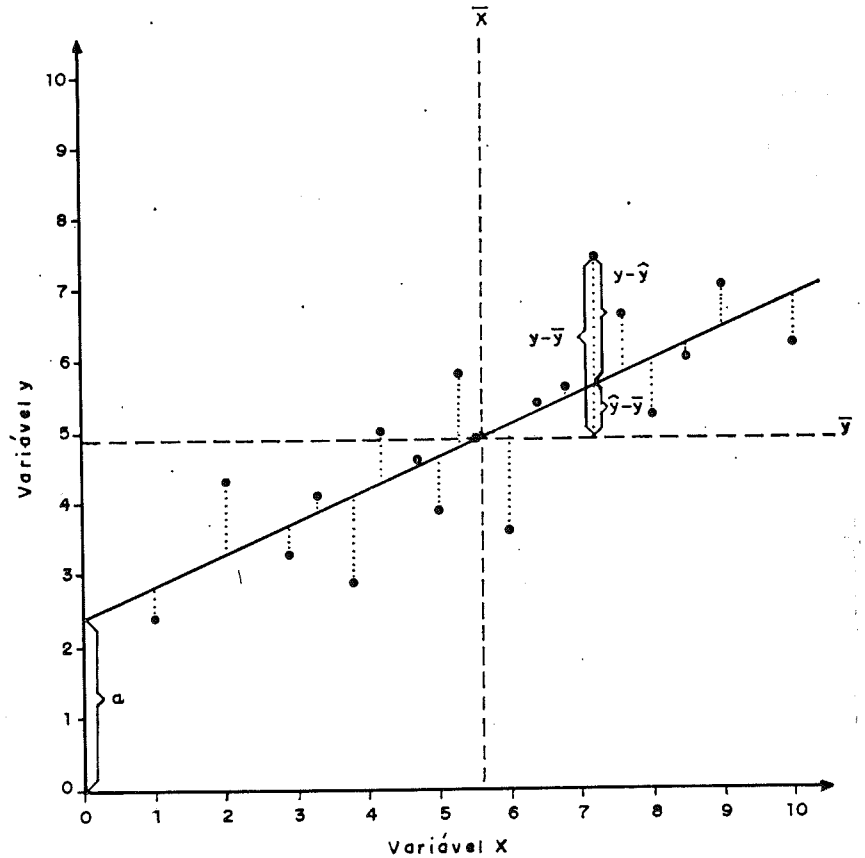

Figura 2 — Desvios verticais da reta de regressão (dados da Tab. 1).

Na função linear, *a* é o interceptor sobre o eixo *Y*, ou seja, o valor de *Y* quando *X* é zero, sendo que *a* pode ser positivo, como na figura 2, negativo ou zero. Por sua vez, *b* é responsável pela inclinação da linha reta, podendo ser positivo ou negativo. É chamado de coeficiente de regressão.

Na figura 2 desenhamos todos os desvios verticais de uma linha. Os desvios abaixo da linha são negativos, os acima, positivos. Seria, teoricamente, possível desenhar uma quantidade de

retas, fazendo cada vez o mesmo cálculo para achar $(Y_i - Y_i)^2$ um mínimo.

Temos um outro meio através da determinação dos parâmetros *a* e *b* segundo as equações normais:

$$na + b\,X_i = \Sigma^2\, Y_i \qquad (1)$$
$$a\ \Sigma\, X_i + b\ \Sigma\, X_i = \Sigma\, X_i\, Y_i \qquad (2)$$

A primeira equação normal (1) é calculada multiplicando cada equação de observação pelo coeficiente de *a* (que é 1) e somando as equações (tab. 1 e 2). Na segunda equação normal (2) multiplica-se os membros de cada equação de observação pelo respectivo coeficiente de *b*, somando as equações em seguida (tab. 3).

TABELA 1

X	*Y*
1,0	2,4
2,0	4,3
2,9	3,3
3,3	4,1
3,8	2,9
4,2	5,0
4,7	4,6
5,0	3,9
5,3	5,8
5,6	4,9
6,0	3,6
6,4	5,4
6,8	5,6
7,2	7,4
7,6	6,6
8,0	5,2
8,5	6,0
9,0	7,0
10,0	6,2
107,3	94,2

TABELA 2

2,4	=	1a	+	b (1,0)
4,3	=	1a	+	b (2,0)
3,3	=	1a	+	b (2,9)
4,1	=	1a	+	b (3,3)
2,9	=	1a	+	b (3,8)
5,0	=	1a	+	b (4,2)
4,6	=	1a	+	b (4,7)
3,9	=	1a	+	b (5,0)
5,8	=	1a	+	b (5,3)
4,9	=	1a	+	b (5,6)
3,6	=	1a	+	b (6,0)
5,4	=	1a	+	b (6,4)
5,6	=	1a	+	b (6,8)
7,4	=	1a	+	b (7,2)
6,6	=	1a	+	b (7,6)
5,2	=	1a	+	b (8,0)
6,0	=	1a	+	b (8,5)
7,0	=	1a	+	b (9,0)
6,2	=	1a	+	b (10,0)
94,2	=	19a	+	b(107,3)

$$\Sigma Y_i = na + b\,\Sigma X_i$$

TABELA 3

2,40	=	1,0a	+	b(1,00)
8,60	=	2,0a	+	b(4,00)
9,57	=	2,9a	+	b(8,41)
13,53	=	3,3a	+	b(10,89)
11,02	=	3,8a	+	b(14,44)
21,00	=	4,2a	+	b(17,64)
21,62	=	4,7a	+	b(22,09)
19,50	=	5,0a	+	b(25,00)
30,74	=	5,3a	+	b(28,09)
27,44	=	5,6a	+	b(31,36)
21,60	=	6,0a	+	b(36,00)
34,56	=	6,4a	+	b(40,96)
38,08	=	6,8a	+	b(46,24)
53,28	=	7,2a	+	b(51,84)
50,16	=	7,6a	+	b(57,76)
41,60	=	8,0a	+	b(64,00)
51,00	=	8,5a	+	b(72,25)
63,00	=	9,0a	+	b(81,00)
62,00	=	10,0a	+	b(100,00)
580,70	=	107,0a	+	b(712,97)

$$\Sigma X_i Y_i = a\,\Sigma X_i + b\ \Sigma X_i^2$$

Segundo a equação (2) podemos escrever:

$$a = \frac{\Sigma X_i Y_i - b \Sigma X_i^2}{\Sigma X_i}$$

e substituir na equação (1).

$$n\ \frac{\Sigma X_i Y_i - b\,\Sigma X_i^2}{\Sigma X_i} + b\,\Sigma X_i = \Sigma Y_i$$

$$\frac{n\ \Sigma X_i Y_i}{\Sigma X_i} - \frac{n\ b\ \Sigma X_i^2}{\Sigma X_i} + b\ \Sigma X_i = \Sigma Y_i$$

$$\frac{n\ \Sigma X_i Y_i}{\Sigma X_i} - \Sigma Y_i = \frac{n\ b \Sigma X_i^2}{\Sigma X_i} - b\,\Sigma X_i$$

$$n\ \Sigma X_i Y_i - \Sigma Y_i \Sigma X_i = n\ b\ \Sigma X_i^2 - b\,(\Sigma X_i)^2$$

$$n\ \Sigma X_i Y_i - \Sigma Y_i \Sigma X_i = b\ (n \Sigma X_i^2 - (\Sigma X_i)^2)$$

$$\frac{n\ \Sigma X_i Y_i - \Sigma X_i \Sigma Y_i}{n\ \Sigma X_i^2 - (\Sigma X_i)^2} = b$$

Assim, as equações (1) e (2) podem ser reescritas da seguinte forma:

$$b = \frac{\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i / n}{\Sigma - (X_i \Sigma X_i)^2 / n} \qquad (3)$$

$$a = \overline{Y} - b\overline{X} \qquad (4)$$

Com a determinação de *a* e *b* podemos achar $\hat{Y} = a + bX$, ou seja, a linha do mínimo quadrado.

Na Geografia, as observações raramente se colocam de forma exata sobre uma linha de regressão, ou seja, é raro que elas tenham um relacionamento linear perfeito. Normalmente *Y* é só parcialmente explicado através de *X*. Isto, segundo Norcliffe (1977), ocorre em função de duas razões principais: *a*) os fenômenos que o geógrafo estuda são geralmente de caráter multivariado, ou seja, uma dada variável é influenciada através de muitas outras variáveis de tal maneira que uma variável independente, X, é só responsável para uma parte da variação em Y. Nessa situação podemos aplicar a regressão múltipla e a variável é substituída através de um vetor de variáveis, X_i, de tal maneira que $Y \rightarrow X_1, X_2, X_3 \ldots\ldots, X_n$.

b) Por outro lado, embora alguma variação em um fenômeno possa ser logicamente atribuída a um conjunto de variáveis explanatórias, sobra um componente impredizível de forma inerente que é atribuído a acontecimentos acidentais como sejam enchentes, abaixamento da temperatura ou mortes inesperadas numa família.

Devemos ressaltar que é fundamental saber e determinar que variável numa determinada pesquisa é a variável independente e qual a dependente, porque a regressão de *Y* para *X* não é idêntica com a de *X* para *Y*. A determinação da variável dependente ou independente deve ser decidida individualmente, na pesquisa. Em Geografia temos poucos exemplos onde as variáveis independentes e dependentes poderiam ser trocadas. A dependência da área de pastagens artificiais com relação a quantidade bovinos/ha ou a dependência de bovinos/ha com relação a pastagens artificiais seria um exemplo de variáveis que podem ser dependentes ou independentes.

Se *X* for a variável dependente e *Y* a independente deveríamos fazer a minimização da soma dos quadrados dos desvios horizontais em vez dos verticais, o que equivale à uma troca dos eixos *X* e *Y* (fig. 3). Só no caso raríssimo de se tratar de uma relação perfeita entre as duas variáveis, as retas do mínimo quadrado coincidiriam; nos outros casos teríamos retas diferentes (fig. 4).

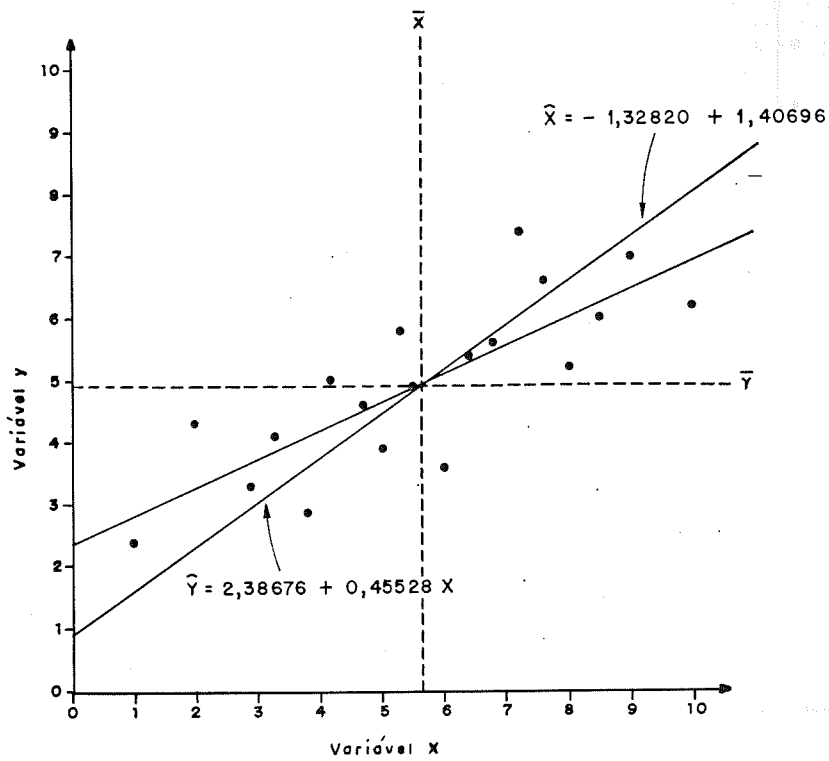

Figura 3 — Regressão de x para y com desvios horizonatis da linha de regressão (dados da Tab. 1).

Na regressão de *Y* para *X* e de *X* para *Y* as retas passam pelo ponto $\overline{X}$, $\overline{Y}$. Isto pode ser provado se escrevemos a fórmula (1) da seguinte maneira:

$$b \frac{\Sigma X_i}{n} + a = \frac{\Sigma Y_i}{n}$$

Assim, temos a indicação que o ponto $\frac{\Sigma X_i}{n}$ e $\frac{\Sigma Y_i}{n}$ se coloca sobre a linha de regressão e que, por consequência, a linha passa pelo ponto $\overline{X}$ e $\overline{Y}$.

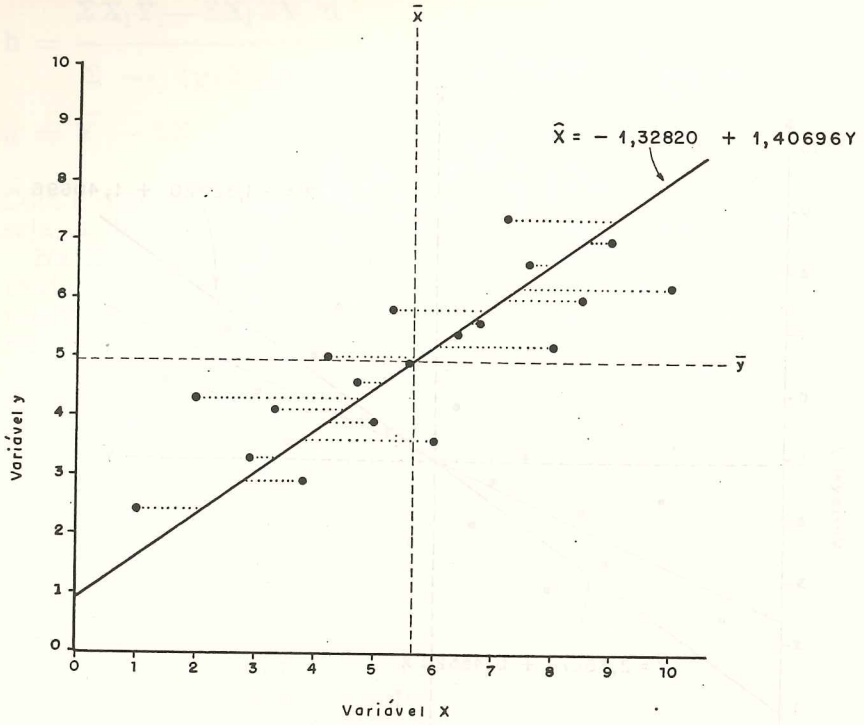

Figura 4 — Retas de regressão de y para x e de x para y (dados da Tab. 1).

Se queremos determinar, numa pesquisa, segundo as fórmulas para *a* e *b*, a regressão de Y para X, precisamos calcular ΣX_i $\Sigma\Sigma Y_i$, ΣX_i, Y_i, ΣX^2, $(\Sigma X_i)^2$, $\bar{X}$ e $\bar{Y}$. A melhor maneira de proceder é construir uma tabela na forma expressa na tabela 4.

TABELA 4 — CÁLCULO DA REGRESSÃO LINEAR SIMPLES PARA OS DADOS DA TABELA 1

X	Y	NY	X^2
1,0	2,4	2,40	1,00
2,0	4,3	8,60	4,00
2,9	3,3	9,57	8,41
3,3	4,1	13,53	10,89
3,8	2,9	11,02	14,44
4,2	5,0	21,00	17,64
4,7	4,6	21,62	22,09
5,0	3,9	19,50	25,00
5,3	5,8	30,74	28,09
5,6	4,9	27,44	31,36
6,0	3,6	21,60	36,00
6,4	5,4	34,56	40,96
6,8	5,6	38,08	46,24
7,2	7,4	53,28	51,84
7,6	6,6	50,16	57,76
8,0	5,2	41,60	64,00
8,5	6,0	51,00	72,25
9,0	7,0	63,00	81,00
10,0	6,2	62,00	100,00
107,3	94,2	580,70	712,97

$n = 19$
$\bar{X} = 5{,}64737$
$\bar{Y} = 4{,}95789$
$(\Sigma X)^2 = 11513{,}29000$

Segundo as fórmulas (3) e (4) temos:

$$b = \frac{580{,}70 - 107{,}3 \cdot 94{,}2/19}{712{,}97 - 11513{,}29/19} = 0{,}45528$$

$$a = 4{,}95789 - 0{,}45528 \cdot 5{,}64737 = 2{,}38676$$

Assim, para o exemplo em questão a reta do mínimo quadrado tem a equação $Y = 2{,}38676 + 0{,}45528X$

Observa-se a esta altura um fato muito importante, mas pouco mencionado, qual seja, a necessidade de realizar os cálculos com grande número de algarismos para achar a reta do mínimo quadrado, mesmo sabendo que do ponto de vista matemático este procedimento não é desejável. A necessidade apontada decorre da constatação de que, somente calculando com o número de algarismos recomendáveis, erros de arredondamento influenciariam fortemente o valor dos parâmetros *a* e *b*.

Depois de determinada a equação de regressão devemos pensar como construir os limites de confiança em torno da reta. Quase sem exceção, a bibliografia existente refere-se a limites de confiança constantes, mostrados a seguir.

Devemos ressaltar de novo que os valores para Y, dados através da reta de regressão, são só as melhores estimativas. Assim, é desejável calcular o erro padrão da estimativa de Y para X para poder indicar até que ponto os valores observados diferem provavelmente da estimativa da linha de regressão. O cálculo é feito através da seguinte maneira: depois de ter sido determinado para cada valor de X o valor estimado de Y, através da equação de regressão, subtraímos o valor estimado do valor observado. recebemos os assim chamados resíduos, que, no gráfico, são as distâncias verticais entre cada observação e a linha de regressão. O desvio padrão dos resíduos mostra o desvio dos valores de Y em torno da linha de regressão. A fórmula seria:

$$s_{y.x} = \sqrt{\frac{\Sigma\ (Y_i - Y_i)^2}{n - 2}} \qquad (5)$$

Dividimos por n—2, porque dois graus de liberdade são perdidos com a estimativa de a e b. Calculamos, assim, o erro padrão das estimativas.

No nosso caso, aplicando a fórmula acima, encontramos como resultado o seguinte valor:

$$s_{y.x} = \sqrt{\frac{12{,}44621}{17}} = 0{,}85565$$

Com este valor de $s_{y.x}$ podemos colocar limites de confiança e desenhá-los no gráfico. As retas paralelas à reta de regressão de Y para X podem ser construídas com as respectivas distâncias de $1s_{y.x}$, $2s_{y.x}$, $3s_{y.x}$. Podemos determinar, se temos muitos valores com uma distribuição mais ou menos normal, que com 68,26% de probabilidade os valores observadores não são mais distantes da linha de regressão que $\pm\ 1s_{y.x}$, com 95,44% de probabilidade os valores não diferem de mais de $\pm\ 2s_{y.x}$ e com 99,74% de probabilidade não mais de $\pm\ 3s_{y.x}$. Ou, em outras palavras, recebemos limites de confiança em relação aos valores estimados através da linha de regressão.

Estes limites de confiança em forma de linhas paralelas em torno da reta de regressão são mostrados na maioria dos trabalhos geográficos (Gregory, 1968; King, 1969; Toyne e Newby, 1971; Yeates, 1974; Taylor, 1977). Entretanto, alguns autores (por exemplo, Haworth e Vincent, 1974; Bahrenberg e Ciese, 1975; Norcliffe, 1977) propõem, para definir o erro padrão de um valor estimado ao longo da reta de regressão, intervalos de confiança que, com mais distância de $\bar{X}$, são maiores, isto é, que a exatidão da estimativa através da reta de regressão diminua com mais distância de $\bar{X}$. Os limites de confiança formam limites de confiança de forma hiperbólica quando os erros de amostragem nas estimativas dos parâmetros de regressão são levados em consideração.

Os limites de confiança de forma hiperbólica tem um efeito importante no cálculo de tendências. Neste caso os valores de X representam o tempo. Isto ocorre, por exemplo, quando queremos fazer predições sobre o futuro desenvolvimento de uma população utilizando a analise de regressão que se baseia sobre um determinado período colocado sobre o eixo X. Se queremos fazer predições para um futuro mais distante, menos exatidão podemos esperar e, por consequência, menos util é a nossa predição para fins de análise e planejamento.

O erro padrão da estimativa de $\hat{Y}$ é:

$$s_{\hat{y}} = s_{y.x}\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma (X_i - \bar{X})^2}} \qquad (6)$$

É preciso determinar s_y para muitos valores de X. Para poder colocar os limites de confiança para μ_y, o valor paramétrico que corresponde ao valor estimado $\hat{Y}_i$, ao nível de confiança de, por exemplo 95%, s_y é multiplicado com $t_{.05}$ (encontrado nas tabelas estatísticas sobre valores críticos da distribuição t de Student, correspondendo a um teste bilateral), com n—2 graus de liberdade. No nosso exemplo da tabela 1 o valor crítico de t, com 17 graus de liberdade, no nível de confiança de 95% é 2,110.

Os limites do intervalo de predição para uma nova observação são dados por:

$$Y \pm t_{(.05;n-2)} s_{y.x}\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma (X_i - \bar{X})^2}} \qquad (7)$$

A tabela 5 mostra para o nosso exemplo os respectivos valores de $\hat{Y}$, do intervalo de confiança e do intervalo de predição. Eles foram colocados no gráfico 5.

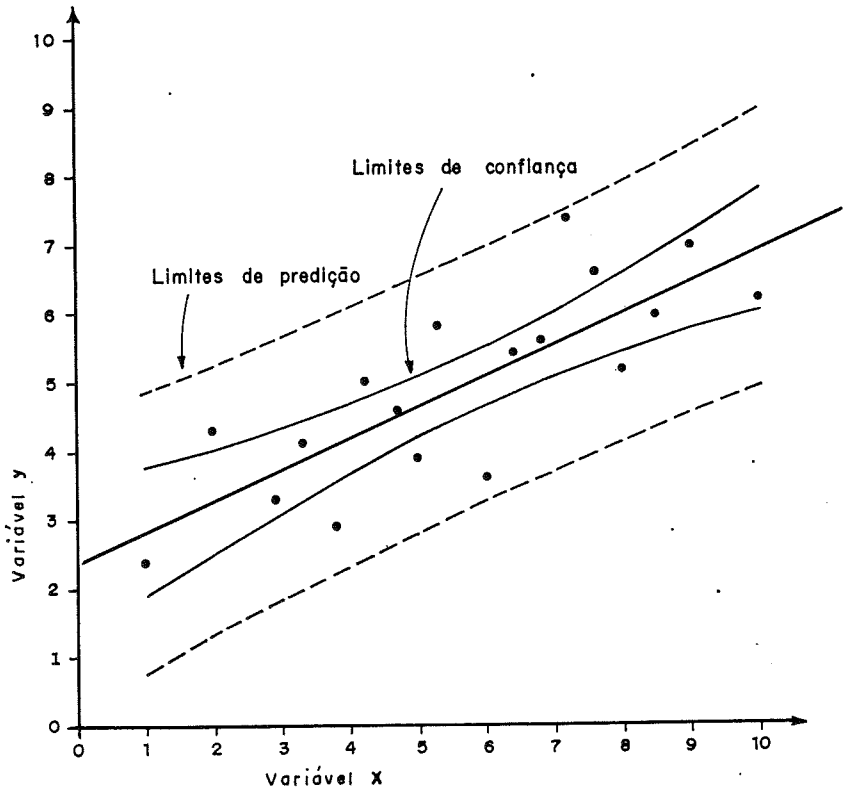

Figura 5 — Reta de regressão com o intervalo de confiança e o intervalo de predição, ao nível de confiança de 95% (dados da Tab. 1).

		limites de confiança ao nível de 95%		*limites de predição ao nível de 95%*	
X	*Y*				
1,0	2,84204	3,75277	1,93131	4,86417	0,81991
2,0	3,29732	4,05678	2,53786	5,25597	1,33867
2,9	3,70707	4,34069	3,07345	5,62045	1,79369
3,3	4,11682	4,64171	3,59193	5,99699	2,23665
3,8	4,11682	4,64171	3,59193	5,99699	2,23665
4,2	4,29894	4,78408	3,81380	6,16841	2,42947
4,7	4,52658	4,97255	4,08061	6,38627	2,66689
5,0	4,66316	5,09248	4,23384	6,51892	2,80740
5,3	4,79974	5,21834	4,38114	6,65305	2,94643
5,6	4,93633	5,35060	4,52206	6,78867	3,08399
6,0	5,11844	5,53717	4,69971	6,97178	3,26510
6,4	5,30055	5,73507	4,86603	7,15752	3,44358
6,8	5,48266	5,94312	5,02220	7,34587	3,61945
7,2	5,66478	6,15974	5,16882	7,53682	2,79274
7,6	5,84689	6,38326	5,31052	7,73030	3,96348
8,0	6,02900	6,61222	5,44578	7,92629	4,13171
8,5	6,25664	6,90427	5,60901	8,17470	4,33858
9,0	6,48428	7,20117	5,76739	8,42683	4,54173
10,0	6,93956	7,80480	6,07432	8,94161	4,93751

3. *ANÁLISES DOS RESÍDUOS*

Podemos, como mencionamos anteriormente, verificar cada ponto de observação da variável dependente em relação ao valor predito, indicado através da linha de regressão, e calcular os resíduos, $Y_{res.} = (Y_i - Y_i)$.

Considerando que, na Geografia, os valores são muitas vezes relacionados a áreas, é interessante mapear os resíduos e destacar as áreas com valores reais acima da predição ou abaixo da predição. Pequenos resíduos indicam que existe grande correspondência entre o valor predito e o valor observado. Se temos uma correlação perfeita, não temos resíduos. A carta mostraria claramente onde temos valores reais acima da predição, abaixo ou onde a predição corresponde à realidade. O pesquisador deve tentar explicar este fenômeno e tentar 'encontrar possíveis outras variáveis desconhecidas que influenciam a variável dependente. A análise dos resíduos dá início a uma outra parte da pesquisa, tornando necessária a formulação de novas hipóteses que nesta fase devem ser testadas para, finalmente, serem aceitas ou rejeitadas.

De interesse particular para a Geografia são os resíduos que caem fora do intervalo de confiança de, por exemplo, 95%.

4. *CONCEITO DE CORRELAÇÃO*

Depois de ter sido determinada a reta de regressão, podemos, através da análise de correlação, medir o grau de associação entre as duas variáveis. Contrariamente à análise de regressão, não mais exprimimos uma como função linear da outra. Na análise de correlação não existe mais esta distinção entre a variável dependente e a independente. Falamos da correlação entre X e Y e examinamos particularmente até que grau duas variáveis são interdependentes ou covariam, isto é, variam juntas, e determinamos a direção dessa covariação.

A nossa preocupação, agora, é a de como medir a intensidade da relação entre duas variáveis. Existem muitos coeficientes de correlação na estatística, sendo que o coeficiente de correlação produto-momento (product moment correlation coefficient) de Karl Pearson, conhecido como *r*, é o mais utilizado. Ele é uma medida em forma de índice para indicar o grau de associação linear entre duas variáveis, com dados na escala de intervalo ou de razão.

Como sabemos, cada ponto num gráfico pode ser determinado pelas suas coordenadas, mas, por outro lado, podemos também determiná-lo em termos de seus desvios de $\overline{X}$ e $\overline{Y}$, ou seja, em termos de $(-\overline{X})$ e $(X_iY_i-\overline{Y})$. Assim, se temos duas variáveis,

X e Y, a chamada covariação de X e Y é

$\Sigma (X_i - \bar{X}) . (Y_i - \bar{Y})$ e a covariância é $\frac{\Sigma (X_i - \bar{X}) . (Y_i - \bar{Y})}{n}$

A covariância tem muita semelhança com a variância. A primeira é medida absoluta e pode ser positiva, ou, contrariamente à variância, negativa. Sendo o coeficiente de correlação uma medida padronizada, isto é, independente da escala original de mensuração, devemos dividir a covariância pelo desvio padrão da variável *X* e *Y*. Em termos matemáticos escrevemos:

$$r = \frac{\Sigma (X_i - \bar{X}) . (Y_i - \bar{Y})/n}{s_x . s_y} \qquad (8)$$

ou

$$r = \frac{\Sigma (X_i - \bar{X}) . (Y_i - \bar{Y})}{\sqrt{[\Sigma (X_i - \bar{X})^2] [\Sigma (Y_i - \bar{Y})^2]}} \qquad (9)$$

O coeficiente de correlação é medida relativa da correlação entre as duas variáveis. Para o uso do computador é preferível utilizar a fórmula (10), conseguida de maneira análoga à explicada por nós para a variância e o desvio padrão (Nentwig Silva, 1978:

$$r = \frac{n \Sigma X_i Y_i - (\Sigma X_i)(\Sigma Y_i)}{\sqrt{[n \Sigma X_i^2 - (\Sigma X_i)^2] [n \Sigma Y_i^2 - (\Sigma Y_i)^2}} \qquad (10)$$

O coeficiente de correlação pode variar entre +1 e —1. Ele é positivo se com valores crescentes de *X* os valores de *Y* aumentam; é negativo se com crescentes valores de X os valores de Y diminuem (fig. 6c e 6d). Assim, r = +1 indica perfeita associação positiva (fig. 6a); r = —1 perfeita associação negativa. Se r =0 não temos correlação entre as duas variáveis (fig. 6b). O gráfico mostra claramente como o coeficiente de correlação muda segundo os diagramas de dispersão diferentes. Segundo Toyne e Newby (1971), pode-se falar, em termos gerais, de alto grau de correlação se temos um índice de ± 0,7 até 1,0, de correlação substancial, tendo um índice de ± 0,4 até 0,7, de baixo grau de correlação se o índice é entre ± 0,2 até 0,4, e abaixo ± 0,2 a correlação é negligenciável.

Se r =0 as duas retas de regressão (de *Y* para *X* e de *X* para *Y*) cortam-se com ângulo de 90°. Se r=1, as duas retas coincidem e o ângulo torna-se zero. Maior o valor de r, menor o ângulo entre as duas linhas de regressão.

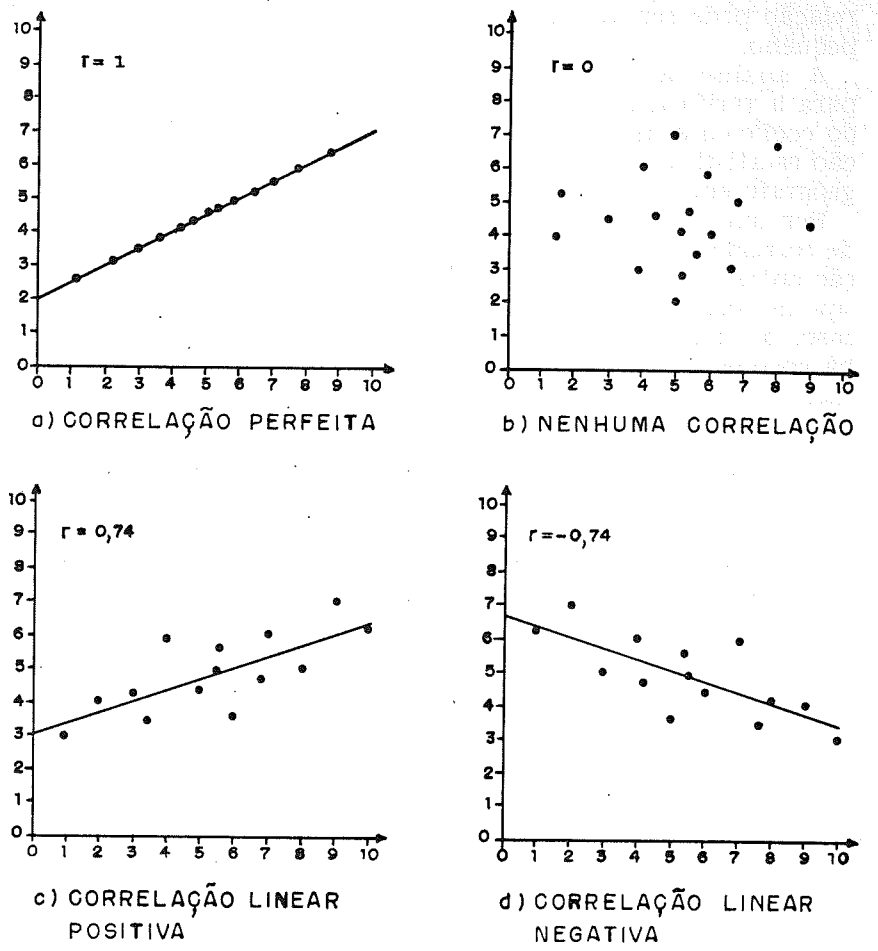

Figura 6 — Exemplos de diagramas de correlação.

Como mencionamos anteriormente, praticamente não se encontra na Geografia uma associação perfeita resultando em um coeficiente de correlação r=1. Por outro lado, devemos tomar cuidado na interpretação do coeficiente. Um alto valor de r não significa necessariamente que a relação indicada seja real, do ponto de vista geográfico. Pode-se tratar da chamada falsa correlação: as duas variáveis podem aparecer correlacionadas por

acaso e não porque existe um associação entre elas. Por outro lado, pode acontecer que as duas variáveis dependem de uma terceira, sendo que as duas em questão não possuem relação entre elas. Como exemplo, Fliri (1969), cita a frequência de nascimentos e o aparecimento de cegonhas. O coeficiente de correlação pode ser também alto porque o tamanho da amostra é pequeno.

A análise de correlação deve ser utilizada particularmente para a verificação quantitativa das prováveis relações. O valor do coeficiente de correlação indica unicamente o grau de relação estatística e não indica unicamente o grau de trabalho do geógrafo encontrar a explicação, a causa do fenômeno.

Por outro lado, devemos ainda destacar que um coeficiente de correlação zero não indica necessariamente que não há relação entre as duas variáveis. É possível que se trate de um outro tipo de correlação que não seja linear, como a parabólica. Neste caso, o coeficiente de correlação indica unicamente que não há correlação linear. Se fosse construído um gráfico de dispersão, como aconselhamos anteriormente, poderíamos rapidamente ver se se trata de uma correlação não linear ou de nenhuma correlação.

Voltando para o nosso exemplo da tabela 1, calculamos o coeficiente de correlação segundo a fórmula (10). Tendo já determinado na tabela 4 $\Sigma X_i = 107,3$; $\Sigma Y_i = 94,2$; $\Sigma X_i^2 = 712,97$; $\Sigma X_i Y_i = 580,70$; $X_i^2 \Sigma = 11513,29$, falta só calcular ΣY_i^2 que é 501,66 e $(\Sigma Y_i)^2 = 8873,64$. Assim, para o nosso exemplo da tabela 1 poderemos escrever:

$$r = \frac{(19).(580,70)-(107,3).(94,2)}{[(19).(712,97)-11513,29].[(19).(501,66)-8873,64]}$$

$$r = \frac{11033,30-10107,66}{[13546,43-11513,29].[9531,54-8873,64]}$$

$$r = \frac{825,64}{[2033,14].[657,90]}$$

$$r = \frac{925,64}{1156,55}$$

$$r = 0,80$$

Para este exemplo o coeficiente de correlação é de 0,80, indicando estatisticamente um alto grau de correlação positiva entre as duas variáveis.

Queremos ainda destacar que o coeficiente de correlação é o mesmo, não importando que a variável X seja designada na análise de regressão como a variável independente e a variável Y a dependente, ou o contrário.

O quadrado do coeficiente de correlação chama-se coeficiente de determinação (r^2 e é expresso em porcentagem. Ele varia entre 0% e 100% devendo sem sempre positivo. O coeficiente indica a proporção da variação de Y explicada pela regressão. 0% indica que nenhuma variação em Y é associada com X e 100% indica que toda a variação em Y é associada com a variação em X. Para o nosso exemplo da tabela 1, $r^2 = 64$, o que quer dizer que 64% da variação em Y é associada com X.

Podemos definir o coeficiente de determinação também como o quociente entre a parte da variação explicada através da reta de regressão e a variação total.

$r^2 = \dfrac{\Sigma (Y_i - Y)^2}{\Sigma (Y_i - \overline{Y})^2}$, sendo que a raiz quadrada de r^2 é o coeficiente de correlação.

A pergunta agora é: quando é que o coeficiente de correlação é ainda significativo? Formulamos a hipótese de nulidade, H_0, que diz que o coeficiente observado aconteceu por acaso. A hipótese alternativa é a de que o coeficiente é maior do que se poderia esperar caso acontecesse por acaso. Só uma das duas hipóteses é certa. Fazemos o teste através da distribuição t de Student com n—2 graus de liberdade:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

No nosso exemplo da tabela 1 temos:

$$t = \frac{0,80\sqrt{17}}{\sqrt{1-0,64}}$$

$$t = \frac{3,30}{0,60}$$

$$t = 5,50$$

O valor calculado de t é comparado com os valores críticos da distribuição de t já tabelados para diversos graus de liberdade e níveis de significância (Taylor, 1977, p. 346; Norcliffe, 1977, p. 191). Para rejeitar a hipótese H_0 os valores computados devem ser maiores do que os valores indicados na tabela. Podemos rejeitar para o nosso exemplo a hipótese H_0 ao nível de 0,1% e concluir que o coeficiente de correlação é altamente significativo.

Temos uma outra possibilidade mais rápida para testar a significância da correlação. Foram já tabelados os valores críticos do coeficiente de correlação para diversos graus de liberdade e níveis de significância. Para ser significante, ao nível da significância escolhida, o valor absoluto de *r* calculado deve ser igual ou maior do que o valor tabelado (Bahrenberg e Giese, 1975, p. 293).

Devemos ainda considerar um fato importante. O coeficiente de correlação foi desenvolvido de uma distribuição normal bidimensional. Se temos outras distribuições ele não é claramente interpretável e a interpretação deve ser feita com cuidado.

A pressuposição da normalidade é necessária se queremos fazer inferência estatística. Em caso de acentuado desvio da normalidade podemos transformar os dados, por exemplo, através de logarítmos, para conseguir a pressuposição da normalidade.

Concluindo, as análises de regressão e de correlação apresentam-se como métodos de pesquisa de inegável valor por possibilitar não somente a verificação de relações entre variáveis e para testar hipóteses, mas particularmente pelo seu valor preditivo, contribuindo para a obtenção de resultados objetivos. É preciso destacar também, com relação à análise de regressão, que a mesma não deve se constituir em um fim em si mesma mas levar o pesquisador, especialmente através da análise e mapeamento de resíduos, a formular ciclicamente novas hipótese a serem testadas com o objetivo de tentar explicar a totalidade do fenômeno.

Deve-se observar que tratamos neste artigo da análise de regressão e de correlação linear simples, onde são só utilizadas duas variáveis. Se queremos pesquisar como uma variável depende de duas ou mais variáveis, e examinar a relação entre um número de variáveis, entramos no assunto de regressão e de correlação múltipla, a ser tratado separadamente.

BIBLIOGRAFIA

Bahrenberg, G. e Giese, E. (1975). *Statistische Mehoden und ihre Anwendung in der Geographie.* Stuttgart:Teubner.

Fliri, F. (1969). *Statistik und Diagramm.* Braunschweing:Westermann.

Gregory, S. (1968). *Statistical methods and the geographer.* 2.ª ed. London: Longman.

Haworth, J. e Vincent, P. (1974). Calculation of prediction limits in linear regression. *Area,* 6(2):113-116.

Hoffmann, R. e Vieira, S. (1977). *Análise de regressão: uma introdução à Econometria.* São Paulo:HUCITEC e EDUSP.

King, L. J. (1969). *Sattistical analysis in geopraphy.* Englewood Cliffs: Prentice-Hall.

Mark, D. M. e Peucker, Th. K. (1978). Regression analysis and geographic models. *Le Géographe Canadien,* 22(1):51-64.

Nentwig Silva, B. C. (1978). Métodos quantitativos aplicados em Geografia: uma introdução. *Geografia,* 3(6):33-73.

Norcliffe, G. B. (1977). *Inferential statistics for geographers: an introduction.* London: Hutchinson.

Sokaal, R. R. e Rohlf, F. J. (1969). *Biometry: The principles and practice of statistics in biological research.* San Francisco: Freeman.

Spiegel, M. R. (1974). *Estatística.* Rio de Janeiro: McGraw-Hill.

Taylor, P. J. (1977). *Quantitative methods in geography: an introduction to spatial analysis.* Boston: Houghton Mifflin.

Toyne, P. e Newby, P. T. (1971). *Techniques in human geography.* Basingstoke: Macmillan.

Yeates, M. (1974). *An introduction to quantitative analysis in human geography.* New York: MacGraw-Hill.

ABSTRACT

Simple linear regression and correlation analysis. This work analyses the use of simple regression and correlation in Geography as a continuation of the author's effort in discussing the quantitative methods in his publication. First of all, the concept of regression is analysed, followed by the construction of the confidence limits, the study of the residuals and, finally, the concept of correlation. Using a theoretical example, the author shows the utilization and the importance of such methods in geographical research.