BIOESTATÍSTICA

GLEICE M S CONCEIÇÃO
MARIA DO ROSÁRIO D D LATORRE
FSP USP

Coeficiente de correlação de Pearson e Modelos de Regressão linear Simples

Como analisar a relação de duas variáveis são quantitativas?

Estratégia

- ✓ Diagrama de dispersão
- ✓ Coeficiente de correlação
- ✓ Regressão Linear Simples

Exercício

- Para esses exercícios utilizaremos a base de dados DC.

- Vamos assumir que minha variável de interesse (dependente) é a PA e as outras (IMC e idade) são as variáveis independentes

Só para lembrar....

parâmetro	Pressão Sistólica (mmHg)	IMC (kg/m2)	Idade (anos)
N	40	40	40
média	164	30,1	48
mediana	162	30	44,8
desvio padrão	35,5	6,32	15,4
valor mínimo	90,2	18,7	23,3
valor máximo	239	39,9	85,6

O diagrama de dispersão

- É um gráfico no qual são representadas 2 variáveis quantitativas.
- Se o interesse é fazer um modelo de regressão linear, no eixo vertical (Y) será representada a variável dependente e, no eixo horizontal (X), a variável independente.
- Cada indivíduo da amostra é representado, ao mesmo tempo pelas 2 variáveis (x_i, y_i).

Diagrama de dispersão de pressão sistólica em função do índice de massa corpórea dos pacientes. Araraquara, 2016.

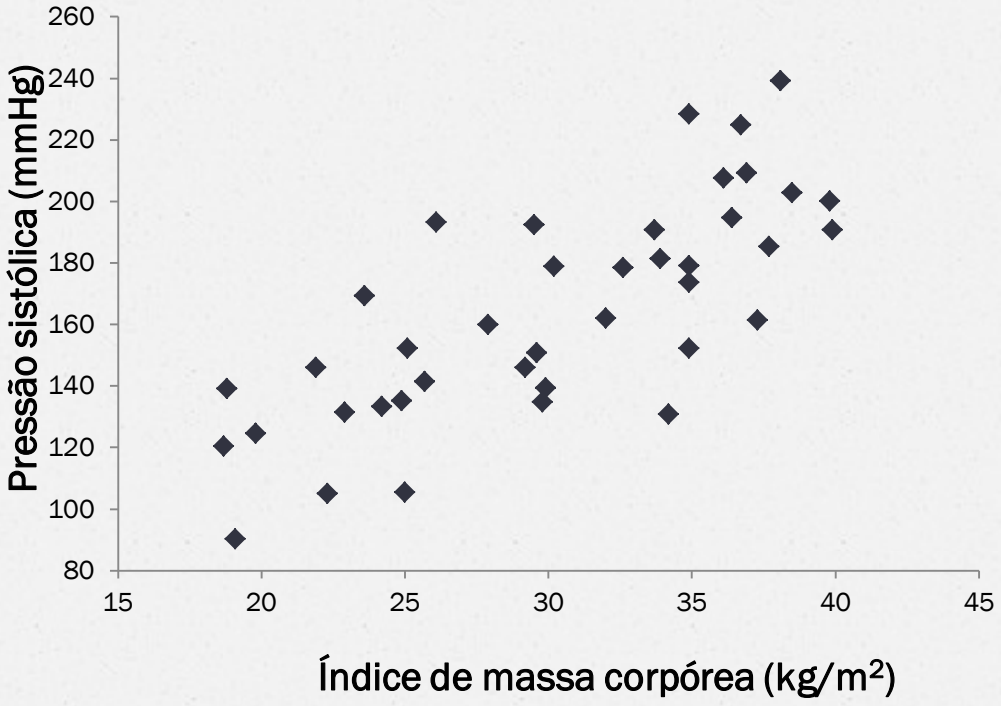


Diagrama de dispersão de pressão sistólica em função do índice de massa corpórea dos pacientes. Araraquara, 2016.

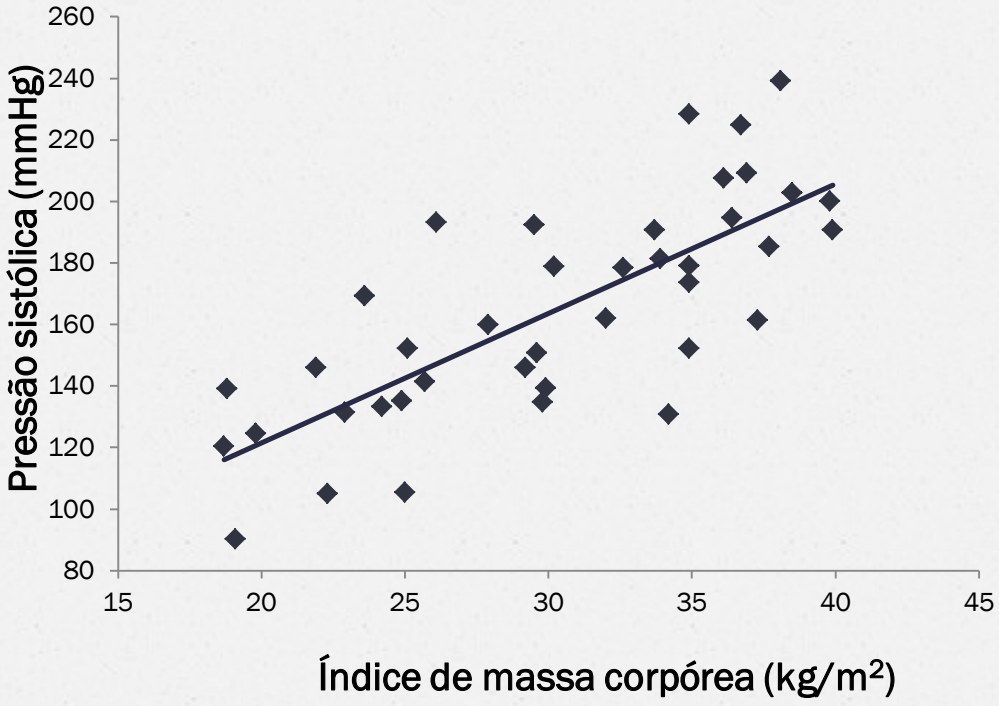
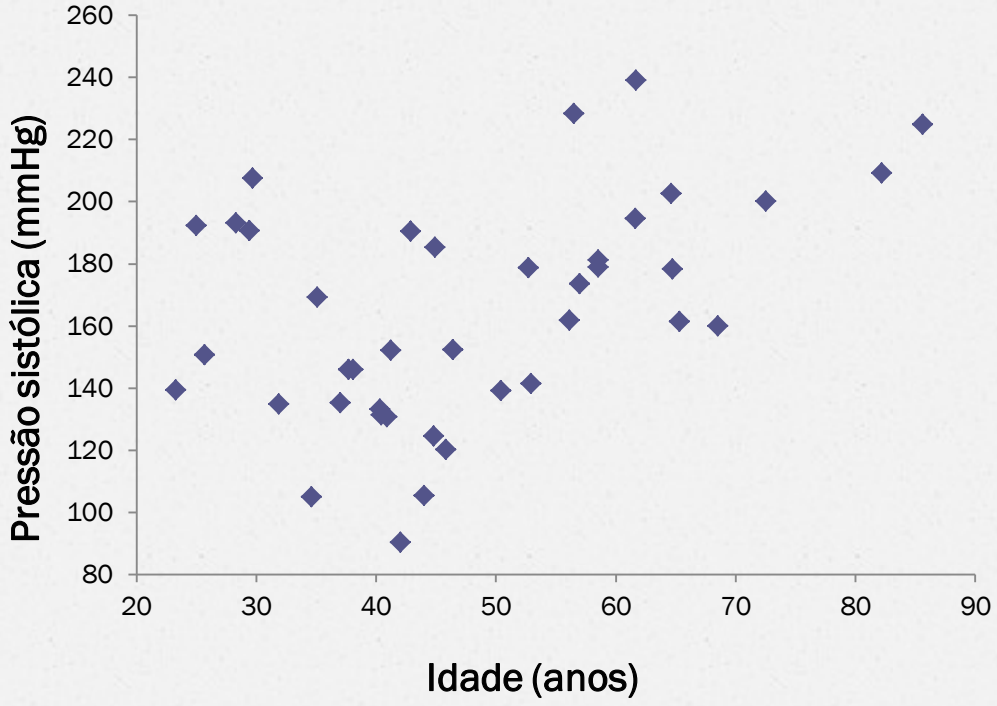
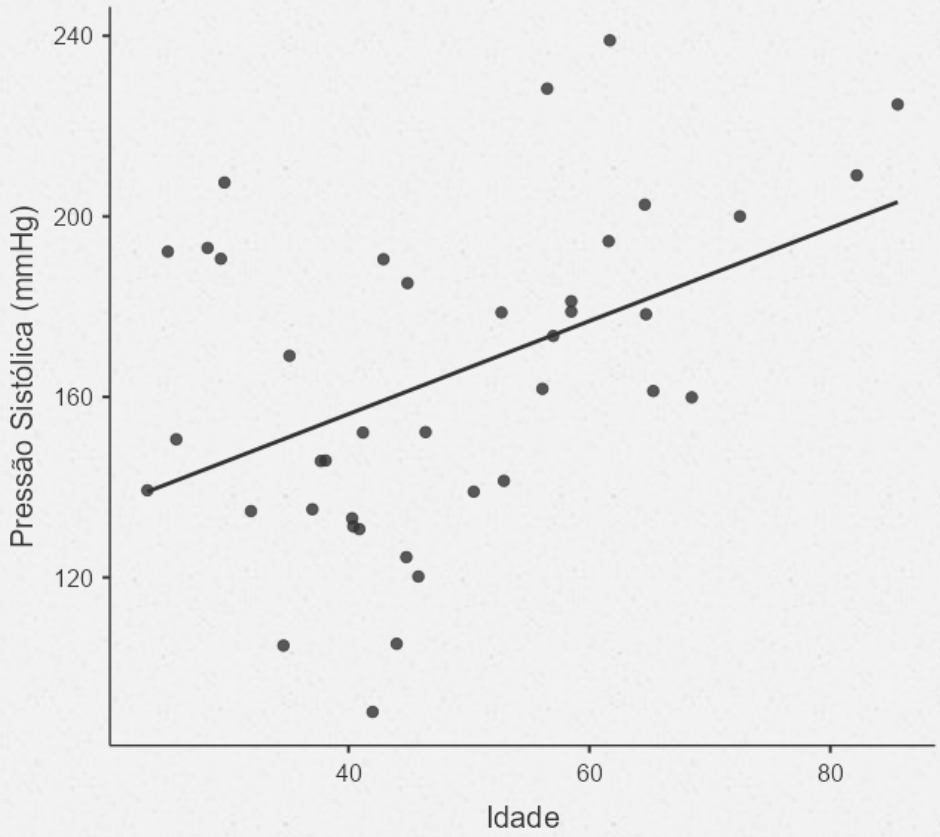


Diagrama de dispersão de pressão sistólica em função do idade dos pacientes. Araraquara, 2016.




240
200
160
120
Pressão Sistólica (mmHg)
40
60
80
Idade

Como medir o grau de linearidade dessa relação?

O coeficiente de Correlação de Pearson:

mede o grau de linearidade de 2 variáveis quantitativas.

COVARIÂNCIA

É o valor médio do produto dos desvios de X e Y, em relação às suas respectivas médias.

$$S_{XY} = \frac{\sum \left(X_i - \bar{X} \right) \left(Y_i - \bar{Y} \right)}{n-1} = \frac{n \sum X_i Y_i - \left(\sum X_i \right) \left(\sum Y_i \right)}{n-1} = \sum X_i Y_i - n \bar{X} \bar{Y}$$

COEFICIENTE DE CORRELAÇÃO (ρ)

Mede o grau de linearidade entre 2 variáveis X e Y.

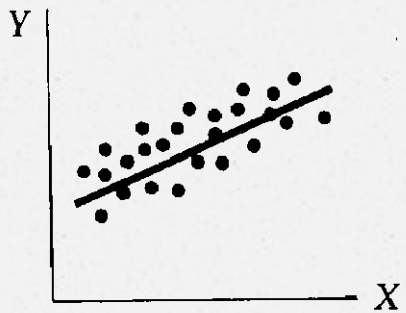
DEFINIÇÃO : $$r = \frac{S_{XY}}{S_X \; S_Y} = \frac{\mathrm{Cov}(X, Y)}{S_X \; S_Y}$$

$$r = \frac{\sum \left(X_i - \bar{X} \right) \left(Y_i - \bar{Y} \right)}{\sqrt{\left[\sum \left(X_i - \bar{X} \right)^2 \sum \left(Y_i - \bar{Y} \right)^2 \right]}} = \frac{\sum X_i Y_i - \frac{\left(\sum X_i \right) \left(\sum Y_i \right)}{n}}{\sqrt{\left[\sum X_i^2 - \frac{\left(\sum X_i \right)^2}{n} \right] \left[\sum Y_i^2 - \frac{\left(\sum Y_i \right)^2}{n} \right]}}$$

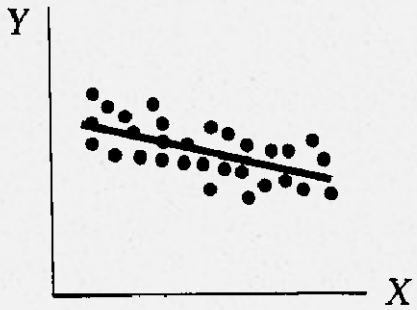
PROPRIEDADES :

a) $-1 \leq r \leq +1$

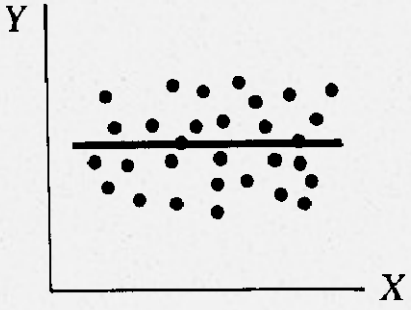
b) r não possui dimensão, ié, não depende das unidades de X e Y



(a) r close to 1: strong positive association



(b) r close to -1: strong negative association



(c) r close to 0: no association

resultados

- r (PA;IMC)= +0,76
- r(PA; idade)=+0,45
- r(IMC e idade)=+0,41

Exercício:

Vamos supor que a variável dependente (Y) seja DT e as demais as variáveis independentes. Calcular o coeficiente de correlação entre as despesas totais com saúde/habitante e o número de nascidos vivos/habitante, o total da população/100.000, a porcentagem de despesas com pessoal/despesa total e a porcentagem de não alfabetizados.

Indicadores Municipais, segundo UF. Brasil.2001

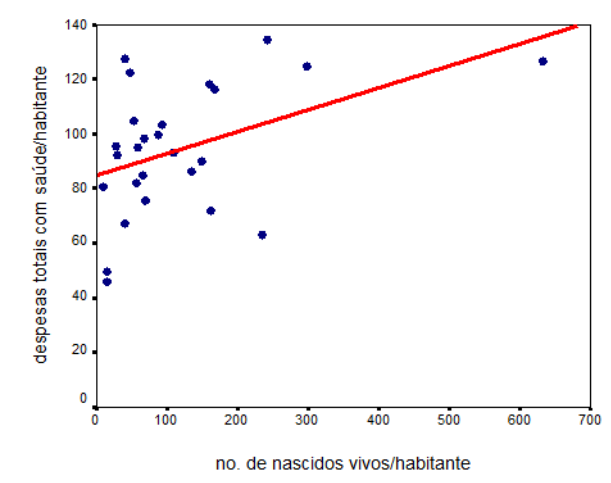
UF	1.DT	2.nasc.vivos	3. Pop.	4.Porc.DP	% Não alfab.
Rondônia	92,39	28,24	12,19	53,65	25,57
Acre	45,72	14,90	4,76	38,96	38,85
Amazonas	75,41	68,42	24,56	55,69	32,87
Roraima	80,51	9,61	3,13	36,54	28,84
Pará	86,44	134,74	54,46	30,56	32,77
Amapá	49,83	14,64	4,85	28,48	29,70
Tocantins	95,44	26,64	8,97	39,52	31,43
Maranhão	93,27	108,61	44,81	23,35	40,33
Piauí	95,16	58,65	25,53	25,54	40,19
Ceará	90,00	149,09	75,04	35,73	36,97
Rio Grande do Norte	104,76	53,53	25,88	28,13	35,09
Paraíba	84,83	64,95	33,43	30,81	37,73
Pernambuco	71,77	162,09	76,48	39,67	34,41
Alagoas	98,46	66,79	28,32	22,18	44,08
Sergipe	67,07	39,84	18,05	42,23	35,90
Bahia	62,85	234,75	120,66	28,23	33,33
Minas Gerais	124,59	297,76	176,17	36,41	21,83
Espírito Santo	81,89	57,12	29,57	47,36	21,99
Rio de Janeiro	134,51	242,36	144,22	33,81	16,97
São Paulo	126,72	632,54	374,68	51,26	17,35
Paraná	116,38	167,29	94,90	34,49	19,77
Santa Catarina	99,44	87,88	54,08	39,48	17,04
Rio Grande do Sul	118,40	160,60	103,10	32,96	16,99
Mato Grosso do Sul	127,67	40,07	21,11	35,85	22,10
Mato Grosso	122,57	47,58	24,94	41,86	23,59
Goiás	103,31	92,70	49,72	31,39	22,10
Total	107,74	3106,53	1633,61	38,98	24,76

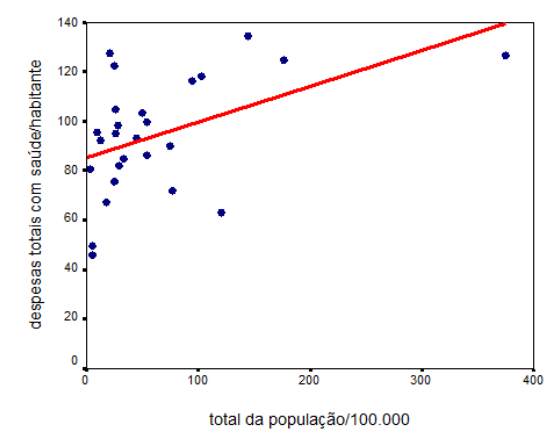
Fonte: http://tabnet.datasus.gov.br/cgi/sinasc/nvmap.htm ; 05/03/2004

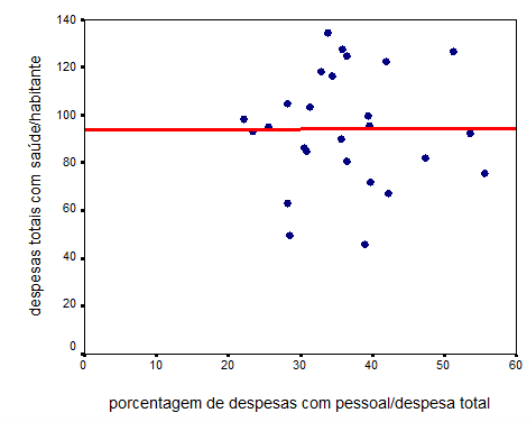
1. Despesas totais com saúde/habitante
2. Número de nascidos vivos (por 1000)
3. Total da população (por 100.000 habitantes)
4. % de despesas com pessoal/Despesas Totais
5. Porcentagem de não alfabetizados

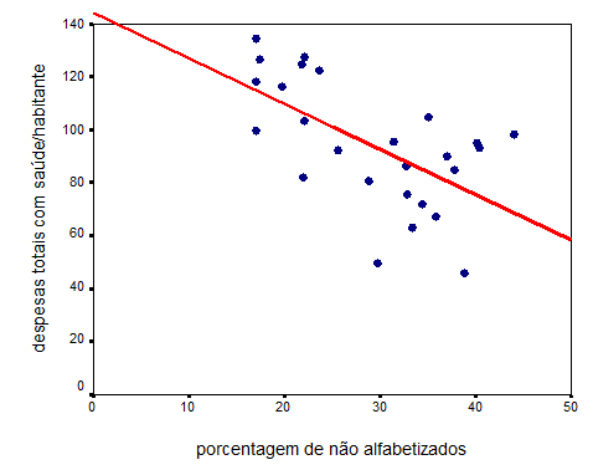
Tabela. Coeficientes de correlação de Pearson (r) entre as despesas totais com saúde/habitante e as demais variáveis de estudo.

variável	r (p)
número de nascidos vivos/habitante	0,44 (p=0,025)
total da população/100.000	0,048 (0,014)
porcentagem de despesas com pessoal/despesa total	0,01 (p=0,970)
porcentagem de não alfabetizados	-0,60 (p=0,001)


despesas totais com saúde/habitante
0
20
40
60
80
100
120
140
0
100
200
300
400
500
600
700
no. de nascidos vivos/habitante


despesas totais com saúde/habitante
0
20
40
60
80
100
120
140
0
100
200
300
400
total da população/100.000


despesas totais com saúde/habitante
0
20
40
60
80
100
120
140
0
10
20
30
40
50
60
porcentagem de despesas com pessoal/despesa total


despesas totais com saúde/habitante
0
20
40
60
80
100
120
140
0
10
20
30
40
50
porcentagem de não alfabetizados

O MODELO DE REGRESSÃO LINEAR SIMPLES

Quando utilizar um modelo de regressão?

1. Quando se deseja caracterizar a relação entre uma variável dependente (Y) e uma ou mais variáveis independentes (Xi), ié, avaliar a extensão, direção e força da relação (associação).

2. Procurar uma função matemática ou equação para descrever a variável dependente (Y) como função da variáveis independentes (Xi), ié, predizer Y em função dos Xi; determinando o melhor modelo estatístico que descreva essa relação.

3. Descrever quantitativa e/ou qualitativamente a relação entre os Xi e Y, controlando o efeito de outras variáveis (Ci).

4. Verificar o efeito interativo de 2 ou mais variáveis independentes às quais se relacionam com a variável dependente.

5. Determinar quais das muitas variáveis independentes são importantes para descrever ou predizer a variável dependente. Ordenar as variáveis independentes em sua ordem de importância em relação à variável dependente.

O MODELO DE REGRESSÃO LINEAR SIMPLES

A função que determina uma reta é: $Y = \beta_0 + \beta_1 X$.

Porém, como se deseja fazer uma estimativa, a reta de regressão estimada pode ser escrita da seguinte maneira:

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$, e $Y = \beta_0 + \beta_1 X + \varepsilon$ ①, onde ε = erro $= Y - \hat{Y}$

$\hat{\beta}_0$ e $\hat{\beta}_1$ são estimados pelo Método dos Mínimos Quadrados.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Como já definido, a variável dependente (Y) será a PA e as demais serão as variáveis independentes.

Modelo 1: Y=f(IMC)

Modelo 2: Y=f(idade)

respostas

- **Modelo 1:**
- Y=36,6 +4,3 (IMC)

- **Modelo 2**
- Y=115,0+ 1,03 (idade)