

HEP0170-Estatísticas de Saúde

Análise de Regressão

Gleice M S Conceição



UNIVERSIDADE DE SÃO PAULO
FACULDADE DE SAÚDE PÚBLICA
DEPARTAMENTO DE EPIDEMIOLOGIA



Análise de Regressão

- Duas variáveis quantitativas
- Descrever a relação entre elas
- Eventualmente, prever o valor de uma delas para um determinado indivíduo quando só conhecemos o valor da outra



Exemplos

- Volume do plasma sanguíneo e peso
- Tempo de reação a um estímulo e idade
- Perda de peso e concentração de uma determinada substância
- Número de óbitos e concentração de um determinado poluente



Análise de Regressão

✓Variável resposta, dependente ou preditiva

A variável que está sendo afetada pela outra ou outras, que acreditamos depender das outras, que pode ser explicada ou prevista pelas outras.

✓Variável explicativa, independente ou preditora

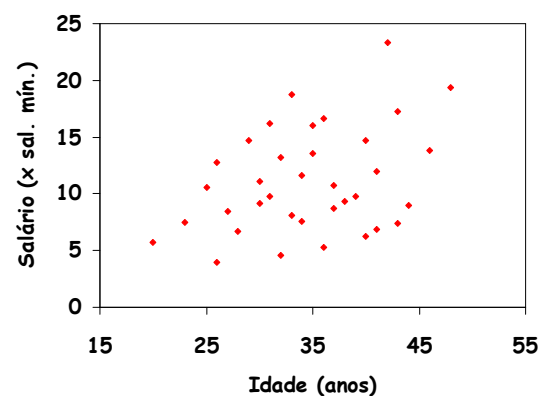
A variável que afeta a outra, que pode ajudar a explicar a variabilidade da outra e a prever a outra.

Regressão Linear Simples

Se o estudo envolver apenas uma variável explicativa, o método será chamado de Regressão Linear Simples.

Medidas descritivas: Gráfico de dispersão

Gráfico de dispersão para as variáveis idade e salário



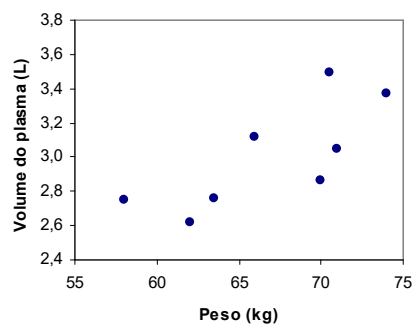
Exemplo

Volume do plasma sanguíneo (L) e peso (Kg)

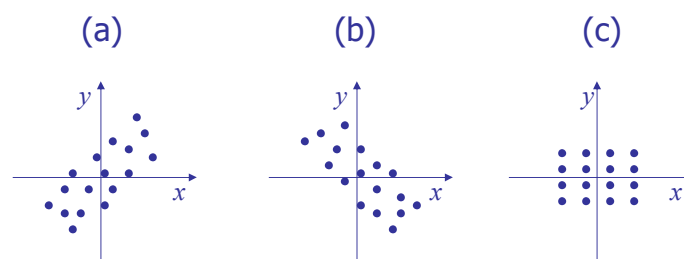
Indivíduo	Volume do Plasma (L)	Peso (kg)
1	2,75	58,0
2	2,86	70,0
3	3,37	74,0
4	2,76	63,5
5	2,62	62,0
6	3,49	70,5
7	3,05	71,0
8	3,12	66,0
média	3,0	66,9

Peso e volume do plasma em 8 indivíduos saudáveis

Indivíduo	Volume do Plasma (L)	Peso (kg)
1	2,75	58,0
2	2,86	70,0
3	3,37	74,0
4	2,76	63,5
5	2,62	62,0
6	3,49	70,5
7	3,05	71,0
8	3,12	66,0
média	3,0	66,9



Tipos de associação entre variáveis



Por exemplo, $x \cdot y$ ou $\frac{x \cdot y}{n}$

Desenvolvendo uma medida de associação

Indivíduo	Peso (kg) x_i	Volume (L) y_i
1	58,0	2,75
2	70,0	2,86
3	74,0	3,37
4	63,5	2,76
5	62,0	2,62
6	70,5	3,49
7	71,0	3,05
8	66,0	3,12

	Peso (kg)	Volume (L)
média	66,8750	3,0025
DP*	5,0667	0,2911

onde

$$\left\{ \begin{array}{l} DP^*(X) = \sum_{i=1}^n \frac{(x_i - \bar{X})^2}{n} \\ DP^*(Y) = \sum_{i=1}^n \frac{(y_i - \bar{Y})^2}{n} \end{array} \right.$$



Desenvolvendo uma medida de associação

Individuo	Peso (kg) x_i	Volume (L) y_i	$x_i - \bar{X}$	$y_i - \bar{Y}$	$\frac{x_i - \bar{X}}{dp(X)} = z_x$	$\frac{y_i - \bar{Y}}{dp(Y)} = z_y$	$z_x \cdot z_y$
1	58,0	2,75	-8,9	-0,3	-1,75	-0,87	1,52
2	70,0	2,86	3,1	-0,1	0,62	-0,49	-0,30
3	74,0	3,37	7,1	0,4	1,41	1,26	1,78
4	63,5	2,76	-3,4	-0,2	-0,67	-0,83	0,55
5	62,0	2,62	-4,9	-0,4	-0,96	-1,31	1,26
6	70,5	3,49	3,6	0,5	0,72	1,67	1,20
7	71,0	3,05	4,1	0,0	0,81	0,16	0,13
8	66,0	3,12	-0,9	0,1	-0,17	0,40	-0,07
			0,0	0,0	0,0	0,0	6,07

	Peso (kg)	Volume (L)
média	66,8750	3,0025
DP*	5,0667	0,2911

onde

$$\begin{cases} DP^*(X) = \sum_{i=1}^n \frac{(x_i - \bar{X})^2}{n} \\ DP^*(Y) = \sum_{i=1}^n \frac{(y_i - \bar{Y})^2}{n} \end{cases}$$

$$corr(X, Y) = \frac{6,07}{8} = 0,76$$



Coeficiente de correlação

$$corr(X, Y) = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{X}}{DP^*(X)} \right) \left(\frac{y_i - \bar{Y}}{DP^*(Y)} \right)$$

onde $DP^*(X) = \sum_{i=1}^n \frac{(x_i - \bar{X})^2}{n}$

$$DP^*(Y) = \sum_{i=1}^n \frac{(y_i - \bar{Y})^2}{n}$$

isto é, a média dos produtos dos valores padronizados das variáveis X e Y.



Coeficiente de correlação

ou

$$\text{corr}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{X}}{DP(X)} \right) \left(\frac{y_i - \bar{Y}}{DP(Y)} \right)$$

onde $DP(X) = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1}}$

$$DP(Y) = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{Y})^2}{n-1}}$$

isto é, a média dos produtos dos valores padronizados das variáveis X e Y.



Coeficiente de correlação

Formas alternativas de cálculo:

$$\text{corr}(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{X})^2 \sum_{i=1}^n (y_i - \bar{Y})^2}}$$



Coeficiente de correlação

Formas alternativas de cálculo:

$$\text{corr}(X, Y) = \frac{\sum_{i=1}^n x_i y_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - n\bar{X}^2\right)\left(\sum_{i=1}^n y_i^2 - n\bar{Y}^2\right)}}$$



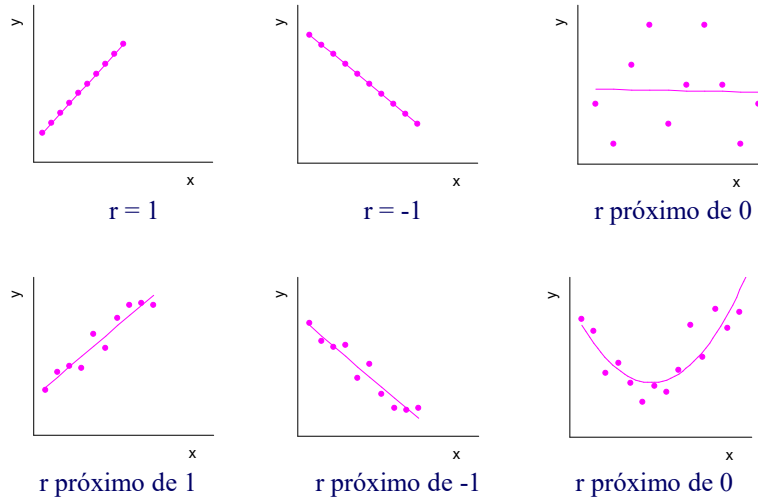
Coeficiente de correlação

- $-1 \leq \text{corr}(X, Y) \leq 1$
- Valores próximos de 1 ou -1 indicam uma associação forte
- Valores próximos de zero quando não existe associação

O coeficiente de correlação mede:

- Presença de associação linear
- Força de uma associação linear

Coeficiente de correlação



Coeficiente de correlação

Alguns autores sugerem avaliar a presença de associação linear a partir do coeficiente de correlação do seguinte modo:

- de 0,10 a 0,39 - **fraca**
- de 0,40 a 0,69 - **moderada**
- de 0,70 até 1 - **forte**

Mas não há, de fato, uma norma rígida sobre isto.

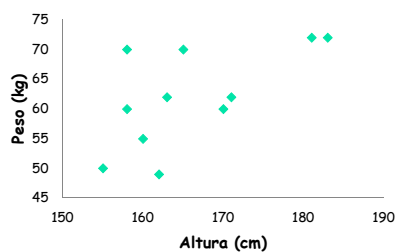
Deve-se levar em conta o contexto e sempre avaliar a associação observando conjuntamente o coeficiente de correlação e o diagrama de dispersão.

Exemplo

Tabela 1. Peso e altura de 11 alunos da Farmácia

Aluno (i)	Altura (cm)	Peso (kg)
1	171	62
2	160	55
3	158	70
4	158	60
5	155	50
6	163	62
7	162	49
8	165	70
9	170	60
10	181	72
11	183	72

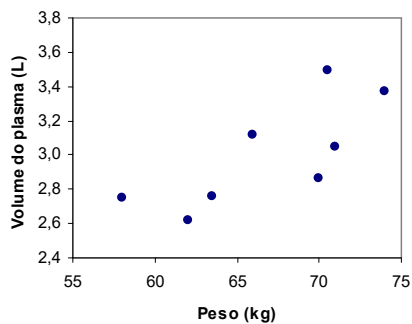
Figura 1. Diagrama de dispersão do peso em função da altura para 11 alunos da Farmácia



$$r=0.62 \quad r^*=0.28$$

Peso e volume do plasma em 8 indivíduos saudáveis

	y	x
Indivíduo	Volume do Plasma (L)	Peso (kg)
1	2,75	58,0
2	2,86	70,0
3	3,37	74,0
4	2,76	63,5
5	2,62	62,0
6	3,49	70,5
7	3,05	71,0
8	3,12	66,0
média	3,0	66,9

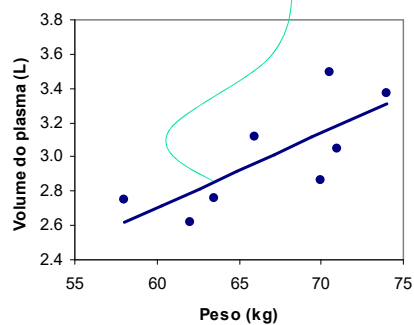


$$\text{corr}(X, Y) = 0,76$$

Peso e volume do plasma em 8 indivíduos saudáveis

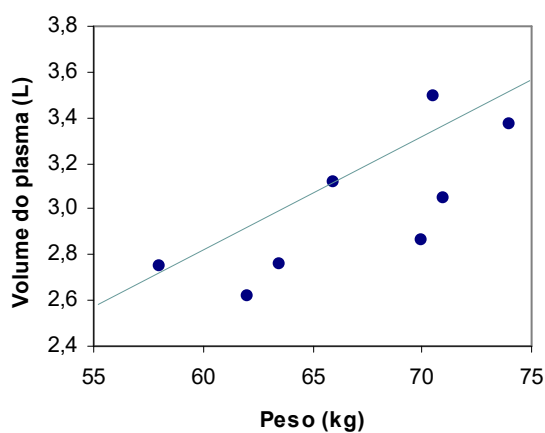
	y	x
Indivíduo	Volume do Plasma (L)	Peso (kg)
1	2,75	58,0
2	2,86	70,0
3	3,37	74,0
4	2,76	63,5
5	2,62	62,0
6	3,49	70,5
7	3,05	71,0
8	3,12	66,0
média	3,0	66,9

Reta de Regressão

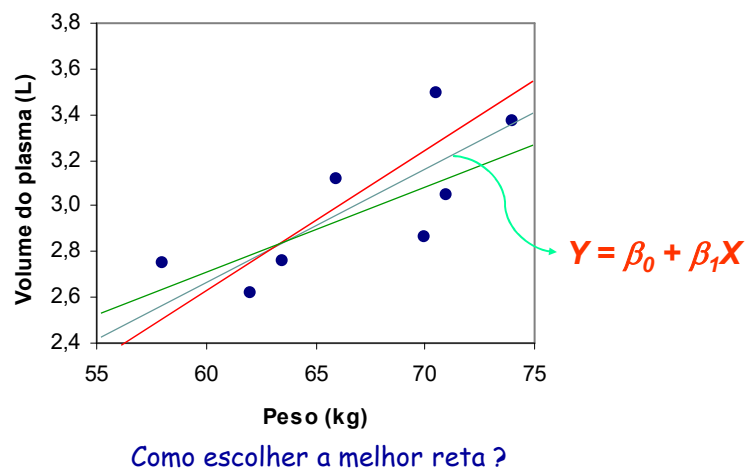


$$\text{corr}(X, Y) = 0,76$$

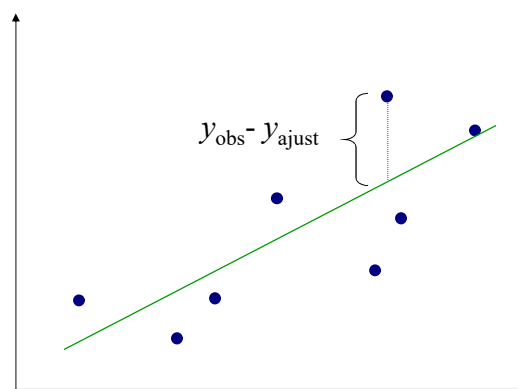
Regressão Linear Simples



Regressão Linear Simples



Erro



Escolha da "melhor" reta de regressão

Para cada observação, calculo o erro:

$$e = y_{obs} - y_{ajust}$$

$$\sum e^2 = \sum \{y_{obs} - y_{ajust}\}^2 = Q$$

$$Q = \sum \{y_{obs} - (\beta_0 + \beta_1 X)\}^2$$

$$Q = \sum \{y_{obs} - \beta_0 - \beta_1 X\}^2$$

Escolha da "melhor" reta de regressão

Para obter o ponto de mínimo da função Q:

$$\frac{\partial Q}{\partial \beta_0} = 0 \quad e \quad \frac{\partial Q}{\partial \beta_1} = 0$$

E obtemos as estimativas para b_0 e b_1 :

$$\Rightarrow \begin{cases} b_1 = \frac{\sum (x - \bar{X})(y - \bar{Y})}{\sum (x - \bar{X})^2} \\ b_0 = \bar{Y} - b_1 \bar{X} \end{cases}$$

Este método é chamado de
método de mínimos quadrados

Peso e volume do plasma em 8 indivíduos saudáveis

de outra forma:

$$b_1 = \frac{\sum xy - n\bar{X}\bar{Y}}{\sum x^2 - n\bar{X}^2}$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

Peso e volume do plasma em 8 indivíduos saudáveis

	y	x
Indivíduo	Volume do Plasma (L)	Peso (kg)
1	2,75	58,0
2	2,86	70,0
3	3,37	74,0
4	2,76	63,5
5	2,62	62,0
6	3,49	70,5
7	3,05	71,0
8	3,12	66,0
média	3,0	66,9

$$b_1 = \frac{\sum xy - n\bar{X}\bar{Y}}{\sum x^2 - n\bar{X}^2}$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

$$\sum xy = 1615,3 \quad \bar{X} = \frac{\sum x}{8} = 66,9 \quad \bar{Y} = \frac{\sum y}{8} = 3,0 \quad \sum x^2 = 35983,5$$

Peso e volume do plasma em 8 indivíduos saudáveis

Indivíduo	Volume do Plasma (L)	Peso (kg)
1	2,75	58,0
2	2,86	70,0
3	3,37	74,0
4	2,76	63,5
5	2,62	62,0
6	3,49	70,5
7	3,05	71,0
8	3,12	66,0
média	3,0025	66,8750

$$b_1 = \frac{\sum xy - n\bar{X}\bar{Y}}{\sum x^2 - n\bar{X}^2}$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

$$\sum xy = 1615,2950 \quad \bar{X} = \frac{\sum x}{8} = 66,8750 \quad \bar{Y} = \frac{\sum y}{8} = 3,0025 \quad \sum x^2 = 35983,500$$

Peso e volume do plasma em 8 indivíduos saudáveis

$$b_1 = \frac{\sum xy - n\bar{X}\bar{Y}}{\sum x^2 - n\bar{X}^2} = \frac{1615,3 - 8 * 66,9 * 3,0}{35983,5 - 8 * 66,9^2} = 0,0436$$

$$b_0 = \bar{Y} - b_1\bar{X} = 3,0 - b_1 * 66,9 = 3,0 - 0,0436 * 66,9$$

$$b_0 = 0,0857$$

$$\hat{y} = 0,0857 + 0,0436x$$

Peso e volume do plasma em 8 indivíduos saudáveis

$$b_1 = \frac{\sum xy - n\bar{X}\bar{Y}}{\sum x^2 - n\bar{X}^2} = \frac{1615,295 - 8 * 66,8750 * 3,0025}{35983,5000 - 8 * 66,8750^2} = 0,043615$$

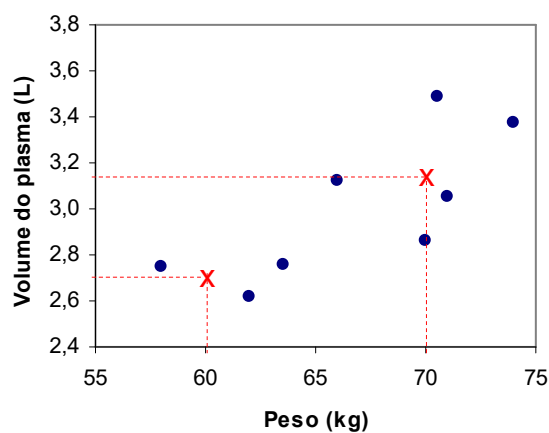
$$= \frac{1615,2950 - 1606,3375}{35983,5000 - 35778,1250} = \frac{8,9575}{205,3750} = 0,043615$$

$$b_0 = \bar{Y} - b_1\bar{X} = 3,0025 - b_1 * 66,8750$$

$$= 3,0025 - 0,043615 * 66,8750 = 0,0857$$

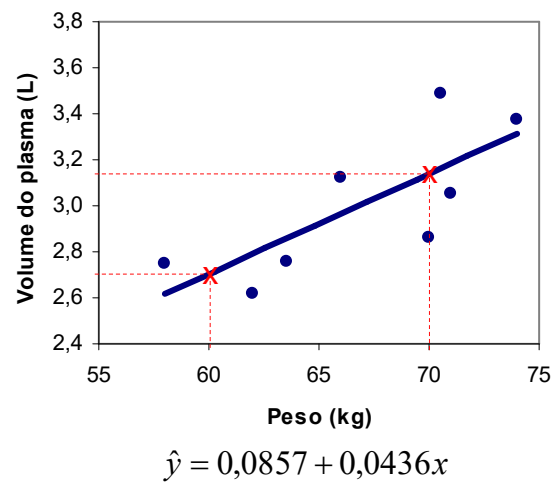
$$\hat{y} = 0,0857 + 0,0436x$$

Peso e volume do plasma em 8 indivíduos saudáveis



$$\hat{y} = 0,0857 + 0,0436x$$

Peso e volume do plasma em 8 indivíduos saudáveis

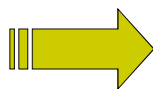


Interpretação dos coeficientes

$$\hat{y} = b_0 + b_1X \quad (1)$$

Se $X = 0$ em (1):

$$y = b_0$$



b_0 é o valor esperado de y quando $X=0$,
é o intercepto da reta de regressão

Interpretação dos coeficientes

$$\hat{y} = b_0 + b_1 X$$

Se aumentarmos X em uma unidade

$$\begin{aligned}\hat{y}_{\text{novo}} &= b_0 + b_1(X + 1) \\ &= b_0 + b_1 X + b_1 \\ &= \hat{y} + b_1\end{aligned}$$



b_1 é o aumento esperado em y quando aumentamos X de 1 unidade, é o "efeito" de X em y .

Interpretação dos coeficientes

$$\hat{y} = 0,0857 + 0,0436x$$

Um aumento de um 1kg no peso corresponderá a um aumento de 0,04 litros de plasma