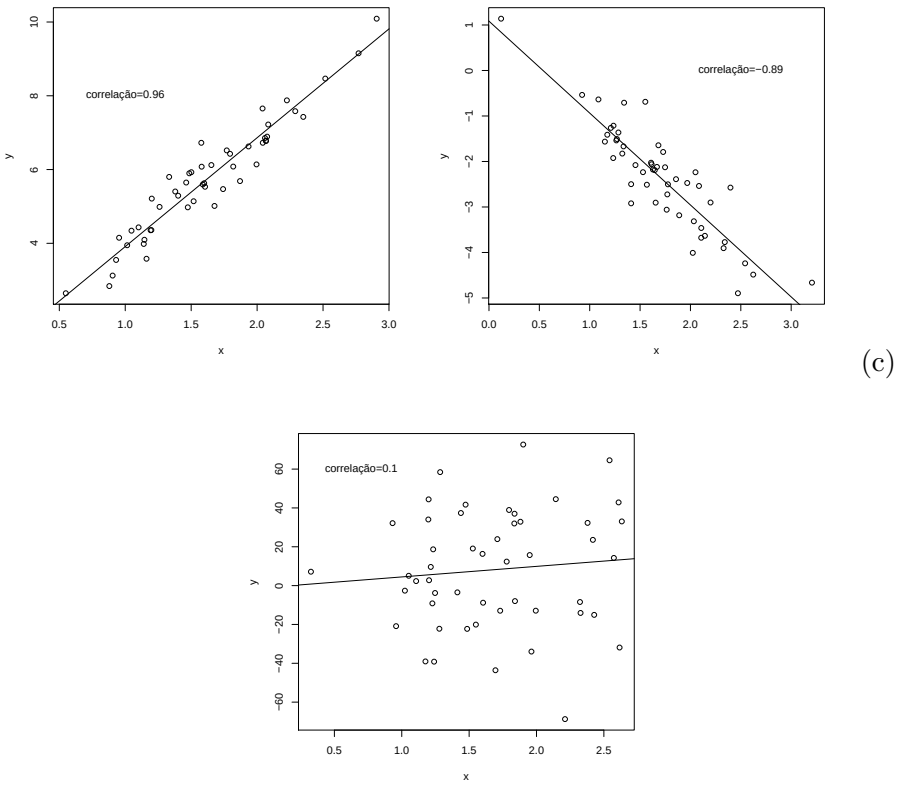
Associação entre Variáveis Quantitativas

Como visto nas aulas de probabilidade, quando dispomos de informações de duas ou mais variáveis, um dos principais objetivos é descrever a associação entre elas.

Para variáveis quantitativas, uma forma inicial de explorar essa associação é utilizando um diagrama de dispersão. Cada par ordenado (x, y) observado é representado por um ponto em um gráfico cartesiano.

Abaixo podemos ver alguns exemplos das possíveis formas que um diagrama de dispersão pode apresentar:

(a) (b)



(c)

Figura 1: Diagrama de dispersão: (a) indica correlação positiva, (b) correlação negativa e (c) ausência de correlação linear.

A linha exibida nos gráficos é a linha que melhor descreve o comportamento linear entre as variáveis. Veremos algumas medidas que indicam o grau das relações lineares entre as variáveis

quantitativas. Sejam duas variáveis X e Y, as medidas de covariância e o coeficiente de correlação medem a linearidade da relação entre as variáveis consideradas.

Covariância: Dados n pares de valores $(x_1, y_2), \ldots, (x_n, y_n)$, chamaremos de covariância entre X e Y a medida:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}.$$

Essa medida de covariância pode assumir valores no intervalo $(-\infty, \infty)$, e por isso é de difícil interpretação pois depende muito da escala das variáveis consideradas. Uma padronização na medida de covariância permite limitar o intervalo de valores possíveis, simplificando a interpretação dos resultados.

Coeficiente de Correlação:

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

ou

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{V}(X)}\sqrt{\text{V}(Y)}}.$$

Uma propriedade importante é que $-1 \leq \text{corr}(X, Y) \leq 1$, sendo que corr=0 quando não há correlação linear entre as variáveis X e Y. Quanto mais próximo de -1 ou de 1 mais forte é a dependência linear entre as variáveis. Se a relação entre as variáveis for linear na forma $Y = a + bX$, então corr $= 1$ e no caso $Y = a - bX$, temos corr $= -1$.

Se X e Y forem independentes $\text{cov}(X, Y) = \text{corr}(X, Y) = 0$, entretanto $\text{cov}(X, Y) = \text{corr}(X, Y) = 0$ não implicam independência, mas apenas falta de linearidade na relação entre as variáveis.

Exemplo

Consideremos os dados de pesquisa feita com 10 famílias, em que foram avaliados a renda bruta mensal (expressa em número de salários mínimos), e a porcentagem da renda bruta anual gasta com assistência médica.

Familia	X	Y
A	12	7,2
B	16	7,4
C	18	7,0
D	20	6,5
E	28	6,6
F	30	6,7
G	40	6,0
H	48	5,6
I	50	6,0
J	54	5,5

Para calcular a covariância e correlação precisamos calcular $\bar{x} = 31,6$, $\bar{y} = 6,45$, $V(x) = 214,24$ e $V(y) = 0,3885$ e assim temos:

$$\begin{aligned} Cov(x,y) &= \frac{(12-31,6)(7,2-6,45)+(16-31,6)(7,4-6,45)+\ldots+(54-31,6)(5,5-6,45)}{n} \\ Cov(x,y) &= -8,58 \end{aligned}$$

$$Corr(x,y) = \frac{Cov(x,y)}{\sqrt{V(x)V(y)}} = \frac{-8,58}{\sqrt{214,240,3885}} = -0,9404625 \quad (1)$$

O gráfico de dispersão das variáveis observadas tem a seguinte forma: