

Correlação e Regressão

Prof.: Wagner Pinheiro
wagner2235@gmail.com

Sumário

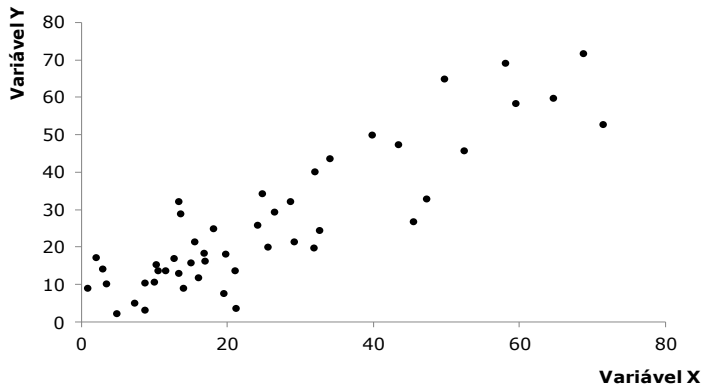
- 1 Coeficiente de correlação linear de Person
 - Teste de hipótese para correlação linear
- 2 Regressão Linear Simples
 - Modelo Estatístico
- 3 Coeficiente de determinação simples

Introdução

A correlação refere-se à relação ou associação entre duas variáveis, para o caso simples.

Na análise de correlação, procura-se determinar o grau de associação linear entre duas variáveis aleatórias.

Diagrama ou gráfico de Dispersão



Cálculo populacional do coeficiente de correlação

É uma medida da relação linear entre duas variáveis. Para duas variáveis populacionais X e Y , o coeficiente de correlação populacional (ρ) é definido por:

$$\rho_{XY} = \frac{COV(X, Y)}{\sqrt{V(X) V(Y)}}$$

Cálculo amostral do coeficiente de correlação

Para uma amostra de n pares de valores (x_i, y_i) , com $i = 1, 2, 3, \dots, n$. Pode-se obter a correlação amostral de Pearson r_{xy} a partir da expressão:

$$r_{xy} = \frac{c\hat{ov}(x, y)}{\sqrt{\hat{V}(x) \hat{V}(y)}}$$

$$-1 \leq r_{xy} \leq 1$$


$$r_{xy} = \frac{\text{côv}(x, y)}{\sqrt{\hat{V}(x) \hat{V}(y)}}$$

$$\text{côv}(x, y) = \sum_i xy - \frac{\left(\sum_i x_i\right)\left(\sum_i y_i\right)}{n} = SPD_{xy}$$

$$\hat{V}(x) = \sum_i x_i^2 - \frac{\left(\sum_i x_i\right)^2}{n} = SQD_x$$

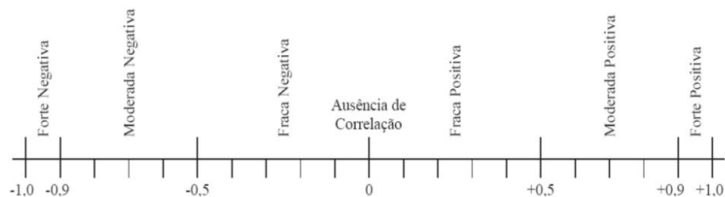
$$\hat{V}(y) = \sum_i y_i^2 - \frac{\left(\sum_i y_i\right)^2}{n} = SQD_y$$

$$r_{xy} = \frac{\text{côv}(x, y)}{\sqrt{\hat{V}(x) \hat{V}(y)}}$$


$$r_{xy} = \frac{\sum_i xy - \frac{\left(\sum_i x_i\right)\left(\sum_i y_i\right)}{n}}{\sqrt{\left(\sum_i x_i^2 - \frac{\left(\sum_i x_i\right)^2}{n}\right) \left(\sum_i y_i^2 - \frac{\left(\sum_i y_i\right)^2}{n}\right)}}$$

$$r_{xy} = \frac{SPD_{xy}}{\sqrt{SQD_x SQD_y}}$$

Escala de correlação linear entre variáveis

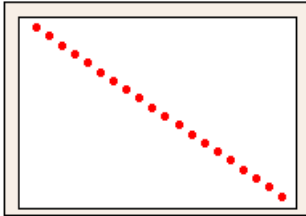


Importante

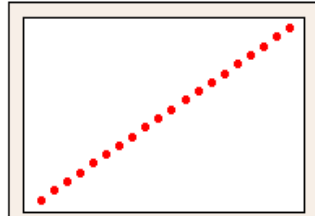
Importante

- 1 Se não observada a correlação linear entre o par X e Y , **não significa que não há relação** e sim que pode haver outro tipo de relação não linear.
- 2 Qual quer que seja a correlação observada, **não significa causalidade**.

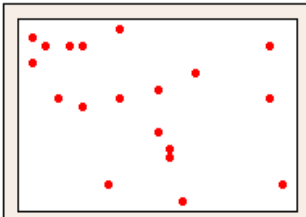
Tipos de relações



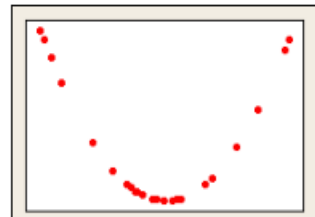
(a)



(b)



(c)



(d)

Estatística do teste

Testes de significância ou hipótese com respeito aos vários valores de ρ requerem o conhecimento das distribuições amostrais de r . Para $\rho = 0$ esta distribuição é simétrica, e a estatística envolvendo a distribuição t de Student, com $n - 2$ graus de liberdade, pode ser utilizada e sua formulação é dada por

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

As hipóteses testadas são:

$$H_0 : \rho = 0 \quad \text{versus} \quad H_a : \rho \neq 0$$

Exemplo

X	Y
4	1
8	5
3	2
9	14
7	3
5	11

Introdução

A regressão é o método para a estimação de valores de uma variável (variável de resultado/consequência ou dependente) com base nos valores de uma outra ou mais variáveis independentes ou prognósticas.

Obter o modelo de regressão é o processo no qual são utilizados dados amostrais para determinar uma equação e deste modo representar a relação observada entre as variáveis em estudo.

Introdução

Dado n pares de valores de duas variáveis aleatórias x_i e y_i , com $i = 1, 2, 3, \dots, n$, admitindo que Y é função de X pode-se estabelecer uma **regressão linear simples**, cujo modelo estatístico é dado por:

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

Onde β_0 e β_1 são parâmetros do modelo e e_i os erros aleatórios.

Pressuposições

- 1 A relação entre X e Y é linear;
- 2 Os valores de X são fixos;
- 3 A média dos erros é nula, ou seja, a $E(e_i)$;
- 4 Para um valor de X , a variância do erro é sempre σ^2 , isto é, $V(e_i) = \sigma^2$. Diz-se então que o erro é homocedástico ou que há homocedasticidade (do erro ou da variável independente), de forma que:

$$E(e_i^2) = \sigma^2 \quad \text{ou} \quad E[Y_i - E(Y_i|X_i)]^2 = \sigma^2$$

- 5 O erro de uma observação é independente do erro de outra observação, isto é, $E(e_i e_j) = 0 \quad \forall i \neq j$;
- 6 Os erros tem distribuição normal.

Estimativa dos parâmetro do modelo

O primeiro passo na análise de regressão é obter as estimativas de β_0 e β_1 . O método usual de é o do **Mínimos Quadrados (MMQ)**, esse método consiste em adotar como estimativa dos parâmetros os valore que minimizam a soma de quadrados dos erros.

Estimativa dos parâmetro do modelo

Sistema de Equações

$$\begin{cases} \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \\ \sum_i X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \end{cases} \Rightarrow \begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum_i X_i = \sum_i Y_i \\ \hat{\beta}_0 \sum_i X_i + \hat{\beta}_1 \sum_i X_i^2 = \sum_i X_i Y_i \end{cases}$$

Estimativas dos parâmetros

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$
$$\hat{\beta}_1 = \frac{\sum_i x_i y_i - \frac{\left(\sum_i x_i\right)\left(\sum_i y_i\right)}{n}}{\sum_i x_i^2 - \frac{\left(\sum_i x_i\right)^2}{n}} = \frac{SPD_{xy}}{SQD_x}$$

Coeficiente de determinação simples R^2

O coeficiente de determinação simples, denotado por R^2 ou r^2 e expresso em porcentagem, é dado por:

$$R^2 = \frac{\hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (Y_i - \bar{Y})^2},$$

Forma simplificada

$$R^2 = (r_{xy})^2 \times 100$$

O R^2 indica a proporção de variação de Y em função de X e quanto é a variação total que está sendo explicada pela regressão (modelo).

Exemplo

Considere que o superior imediato de uma fábrica de pregos, está interessado em elevar a produção da fábrica sem alterar a mão-de-obra, o maquinário e a quantidade de matéria prima utilizados na produção de pregos. Para tanto, ele elege o engenheiro de produção como responsável para alcançar níveis elevados na produção de pregos. Como sugestão o superior imediato menciona ao engenheiro que estudos passados indicaram que a temperatura da máquina está relacionada com o desempenho na produção de pregos. O engenheiro coletou um conjunto de dados (Tabela) da temperatura da máquina e a quantidade produzida de pregos.

Exemplo

Temperatura (°C)	Produção de pregos (t)
40	10
70	16
100	20
120	24
170	30

Exemplo

Pede-se:

- a)* Definir as variável dependente (Y) e independente (X).
- b)* Obter o grau de relação entre as variáveis X e Y .
- c)* Obter a equação ajustada.
- d)* Obtenha o coeficiente de determinação simples.
- e)* Determine a estimativa de Y para uma temperatura de 130 °C.

Importante

