Associação entre 2 variáveis

Apontamentos sobre a regressão linear

Page

Como analisar a associação entre 2 variáveis numéricas?

- Se a relação entre 2 variáveis for linear, e ao confrontarmos 2 amostras num diagrama de dispersão, devemos observar um conjunto de pontos que se dispõem aproximadamente sobre uma reta. Por vezes, os desvios em relação à reta são mínimos, mas noutras, os pontos apresentam bastante dispersão, tornando difícil a identificação da dita relação linear
    - Calcular medidas de associação (coeficientes de correlação)
    - Realizar um teste de hipóteses para averiguar se os valores das medidas de associação observados nos dados são significativos

O que significa a existência de correlação?

- As variáveis podem estar correlacionadas porque uma delas depende da outra (há uma relação de causalidade)
- As variáveis podem estar correlacionadas porque são interdependentes (ex: idade do marido, idade da esposa)
- As 2 variáveis podem estar correlacionadas porque ambas são influenciadas por uma 3ª variável e é o facto de ambas "responderem" às variações nessa variável que explica a correlação (ex: nº de insolações e produção de trigo)

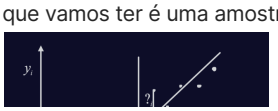
Regressão Linear

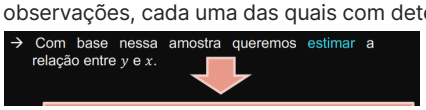
- Conjunto vasto de técnicas estatísticas usadas para modelar relações entre 2 ou mais variáveis e predizer, com base na relação observada o valor de uma variável dependente a partir de um conjunto de variáveis independentes.
- Existem 2 tipos de regressão linear:
    - **Regressão linear simples -** considera apenas uma variável dependente e uma variável independente
    - **Regressão linear múltipla -** a relação entre as 2 variáveis é linear

Regressão linear simples

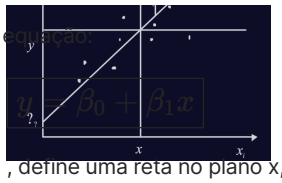
- Quando medimos correlação linear, estamos a medir o grau de associação linear entre as variáveis. Tanto faz de correlação entre x e y, como correlação entre y e x
- Quando fazermos regressão linear, também queremos estudar relação entre variáveis, mas queremos estudar se uma das variáveis depende da outra
- Na regressão linear simples há uma variável explicativa (ou independente) e uma variável explicada (ou dependente). O que queremos saber é se a variável explicativa (ou não) a explicar o comportamento da variável explicada

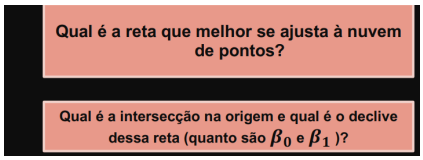
- Se a relação entre y e x fosse exata todas as observações estariam na reta
- Mas a relação não é exata, há outros fatores aleatórios que influenciam y, para além de x
- Há pontos acima da reta (desvios positivos) e pontos abaixo da reta (desvios negativos)
- O que vamos ter é uma amostra de observações, cada uma das quais com determinados valores de x e y





- A equação:





, define uma reta no plano x,y

β0 → Representa a ordenada na origem

β1 → Representa o declive

- Os parâmetros β0 e β1 são chamados de coeficientes de regressão

Parâmetros da regressão

- Como as observações estão afetadas de erros, não é possível saber o valor exato dos coeficientes β0 e β1
- **Método dos mínimos quadrados -** método para estimar os parâmetros β0 e β1, com base na informação de uma amostra
- Para uma dada reta, podemos calcular os desvios em relação à reta (desvios positivos e negativos compensam-se)

- Chamamos valores preditos a:

$$\beta_1 = n\frac{\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$\beta_0 = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

Coeficiente de determinação

- O coeficiente de determinação informa que a fração da variabilidade de uma característica é explicada estatisticamente pela outra variável, sendo determinado pelo quadrado do coeficiente de correlação

$$R^2 = \left(\frac{cov(x,y)}{SxSy}\right)^2$$

- Pode ser interpretado como a redução relativa na variação de yi associada à utilização da variável independente, xi
- Quanto maior o coeficiente de correlação, maior será a proporção na redução da variação total de yi quando conhecemos, xi
- É muitas vezes interpretado com proporção da variação total de yi "explicada" por xi

Exemplo 1:

Um serviço de enfermagem que presta cuidados a idosos realizou um rastreio dos fatores de risco das doenças cardiovasculares nos homens. Para tal foi recolhida uma amostra de 10 idosos com os quais se pretendeu analisar se (X) "a sobrecarga ponderal em relação ao ideal", em %, está relacionada com (Y) "o valor dos triglicerídeos no plasma", em mg/dl :

x_i	y_i	x_i^2	y_i^2	$x_i.y_i$
10	140	100	19600	1400
5	135	25	18225	675
0	130	0	16900	0
20	165	400	27225	3300
15	170	225	28900	2550
13	150	169	22500	1950
20	205	400	42025	4100
0	120	0	14400	0
15	160	225	25600	2400
10	145	100	21025	1450
108	1520	1644	236400	17825