Definição & Exemplos

A estatística permite melhorar o entendimento de eventos, observando a correlação entre diferentes variáveis. O coeficiente de correlação de Pearson — ou correlação linear —, por exemplo, mensura a relação entre diferentes variáveis quantitativas. Além de quantificar o grau de associação entre variáveis, é comum querer prever o valor esperado dado as variáveis explicativas, desafio esse solucionado pela regressão linear.

Diferentemente da classificação, que visa a obtenção de um rótulo, a regressão visa a obtenção de um determinado valor. Em linhas gerais, a regressão linear trata da modelagem da relação entre variáveis numéricas, sendo as mesmas identificadas como:

- variável dependente y - valor previsto;
- variável(is) independente(s) X - atributo(s) previsor(es).

Figura 1: Exemplos de gráficos de regressão (esquerda) e de classificação (direita).

Como exemplos de regressão linear, podemos elucidar os seguintes:

Idade (x) → Valor do plano de saúde (y)[1]
Temperatura, umidade e pressão do ar (x) → Velocidade do vento (y)[2]
Pressão aplicada no tubo (x) → Espessura da garrafa (y)[1]
Número de propagandas(x) → Número de vendas (y)[1]
Teor de açúcar e gordura na dieta (x) → Incidência de obesidade (y)[2]

---

[1] Regressão linear simples, pois há apenas uma variável explicativa.
[2] Regressão linear múltipla, pois há mais de uma variável explicativa.

Graficamente, podemos observar um exemplo contendo a relação entre o índice de massa corpórea (IMC) - atributo previsor - e a circunferência da cintura - valor previsto.

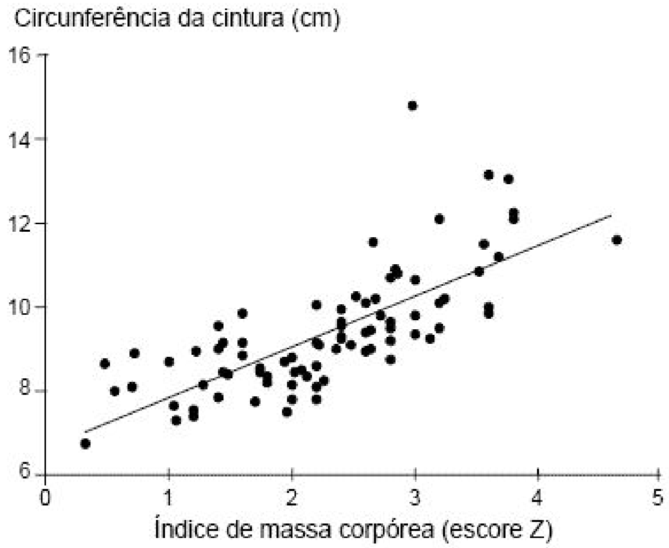

Figura 2: Relação entre o IMC e a circunferência da cintura.

Do ponto de vista da regressão linear simples, temos a seguinte equação:

$$y = b_0 + b_1 * x_1 \rightarrow y = a * x + b$$

onde y é o atributo que queremos prever, b_0 é uma constante, b_1 é um coeficiente angular e x_1 é a variável explicativa. No exemplo visto na Figura 2, podemos considerar o y como sendo o valor da circunferência da cintura e x_1 como o valor do IMC. Os principais objetivos são determinar os melhores valores para as variáveis b_0 e b_1; ambas as variáveis são responsáveis pela localização da reta no gráfico.

Para verificar se a reta se ajustou bem aos dados, ou seja, se determinamos os melhores coeficientes b_0 e b_1, é utilizada a função MSE (*Mean Squared Error*)[3].

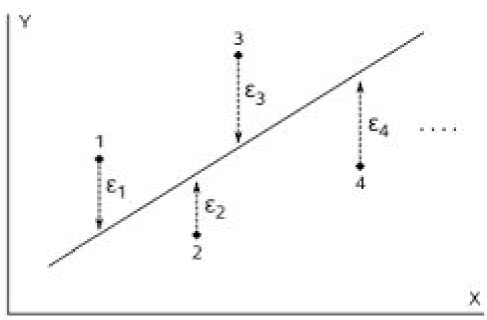

Figura 3: Erros (resíduos) da reta ajustada.

[3] O MSE penalisa os erros maiores, ao longo dos registros.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

O principal objetivo dessa equação é calcular o erro de todas as nossas amostras e, consequentemente, mostrar-nos se é uma boa reta ou não.

Para cada amostra, pegamos o y e $\tilde{y}$, representando o valor real e o valor previsto pela reta, respectivamente. O erro é calculado da seguinte forma:

1. Subtraímos o valor $\tilde{y}_i$ de y_i, ($y_i - \tilde{y}_i = \varepsilon_n$).
2. Calculamos o quadrado do resultado (ε_n).
3. Fazemos um somatório e dividimos pela quantidade de exemplos, isto é,

$$\sum \frac{(\varepsilon_n)^2}{n}$$

Dessa forma, a reta que melhor representa os dados é aquela que tiver o menor MSE possível. Para visualizarmos o processo de cálculo do erro, vamos ao seguinte exemplo:

Valor real	Valor calculado	Erro
1500	1525	$(1500 - 1525)^2 = 625$
1753	1740	$(1753 - 1740)^2 = 169$
1897	1890	$(1897 - 1890)^2 = 49$
2066	2088	$(2066 - 2088)^2 = 484$

Logo o MSE para esse exemplo será:

$$\text{MSE} = \frac{625 + 169 + 49 + 484}{4} = 331,75$$

1. Design Matrix

- Utilizamos esse artifício para bases de dados com *poucos atributos*;
- Trabalhamos com o conceito de inversão de matrizes.

2. Gradient Descent[4] (mais utilizado)

- Possui um desempenho melhor em meio a *muitos atributos*.

Explicando um pouco melhor o *Gradient Descent*, temos o seguinte gráfico:

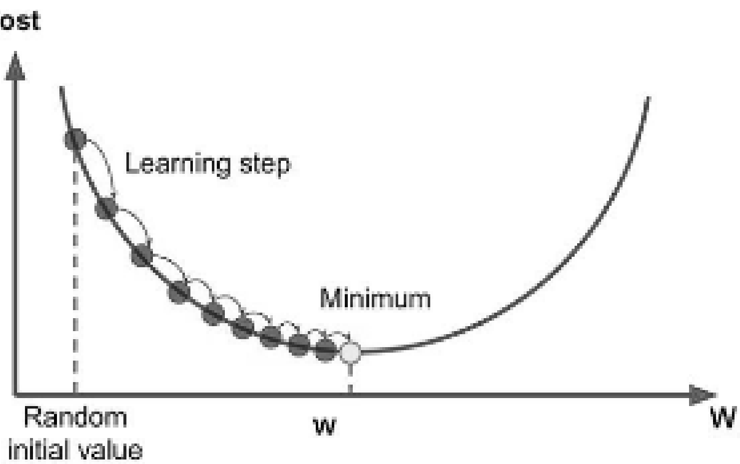

Figura 4: Interpretação gráfica do Gradient Descent.

O principal objetivo em utilizar o *Gradient Descent* está em atingir o mínimo global de uma determinada curva cuja relação seja descrita entre os erros e os pesos. Nesse caso, “atingir o mínimo global” significa obter o menor erro possível.

[4] Também utilizado em algoritmos de Regressão Logística.

Diferentemente da Regressão Linear Simples, a Múltipla conta com mais de uma variável explanatória no seu desenvolvimento. A equação é dada por:

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + ... + b_n * x_n$$

onde $x_1, x_2, ..., x_n$ são os variáveis explanatórias do conjunto de dados em questão e $b_0, b_1, ..., b_n$ são os coeficientes associados às variáveis. Resumidamente, a Regressão Linear Múltipla é utilizada quando se quer analisar a os efeitos, sobre y, de 2 ou mais atributos previsores.