

# 1 Introdução

## 1.1 O que é Análise de Regressão?

A **Análise de Regressão** é uma técnica estatística para modelar e investigar a relação entre duas ou mais variáveis, com o objectivo de explicar determinado fenómeno em estudo e nomeadamente prever a evolução desse fenómeno.

### Exemplos:

- Estudar a relação entre a resistência de um tipo de plástico e o tempo que decorre a partir do fim do processo de moldagem até à medição da resistência;
- Estudar a relação entre o consumo de tabaco e várias variáveis, tais como, sexo, idade, nível de instrução, salário, etc.
- Estudar a relação entre o peso e a idade e a altura de um indivíduo.
- Estudar a relação entre o peso de uma criança e o peso dos seus pais.

Esta relação é representada por um **modelo matemático** ou **modelo de regressão**, ou seja, por uma equação que associa a **variável dependente ou resposta**,  $Y$ , com as  $p$  **variáveis independentes ou explicativas**,  $X_1, X_2, \dots, X_p$ .

A variável resposta  $Y$  é a variável que se deseja modelar, prever a partir da variável independente  $X$ . Enquanto que a variável  $X$  admite-se conhecida, e com base na qual se pretende tirar conclusões sobre a variável resposta.

Em alguns casos, a relação entre as variáveis é exacta (determinística), do tipo:

$$Y = f(x_1, x_2, \dots, x_p)$$

onde  $f$  é uma dada função.

### Exemplo:

Seja  $X$  a temperatura ambiente em graus Celsius e  $Y$  a temperatura ambiente em graus Fahrenheit. O modelo

$$Y = 32 + \frac{9}{5}X$$

é um modelo linear determinístico.

No entanto, na maior parte dos casos o modelo determinístico não pode ser usado porque a relação entre as variáveis não é exacta.

### Exemplo:

A relação entre a altura  $X$  (em cm) e o peso  $Y$  (em Kg) de indivíduos (dois indivíduos medindo o mesmo comprimento podem não ter o mesmo peso). Neste caso, existe uma relação do tipo:

$$Y = f(x_1, x_2, \dots, x_p) + \epsilon$$

onde  $\epsilon$  é um erro aleatório de média zero e variância pequena.

**Exemplos:**

- Modelo de Regressão Linear Simples:  $Y = \beta_0 + \beta_1 x + \epsilon$
- Modelo de Regressão Linear Múltipla:  $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_p x_p + \epsilon$
- Modelo de Regressão Não Linear:  $Y = \beta_0 + \exp \beta_1 x_1 + \epsilon$

O termo **linear** refere-se à linearidade dos parâmetros  $(\beta_0, \beta_1, \dots)$ .

Assim os modelos da mesma forma que os modelos de regressão linear descritos mas onde em vez de  $x$  e  $y$  figuram funções destas variáveis também são considerados modelos de regressão linear:

$$\log Y = \beta_0 + \beta_1 x^2 + \epsilon$$

O modelo linear:

- engloba um grande número de modelos específicos: regressão linear simples e múltipla, análise da variância, análise de covariância;
- serve de base para numerosas generalizações: modelos lineares generalizados; regressão não linear.

**Exercício:** Os seguintes modelos são (ou podem ser transformados) em modelos de regressão linear?

- a)  $Y = \beta_0 + \beta_1 \ln(x) + \epsilon$
- b)  $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$
- c)  $Y = \beta_0 + \beta_1 e^x + \epsilon$
- d)  $Y = \beta_0 e^{\beta_1 x + \epsilon}$
- e)  $Y = \frac{1}{\beta_0 + \beta_1 x + \epsilon}$
- f)  $Y = \beta_0 x^{\beta_1} \epsilon$
- g)  $Y = \frac{x}{\beta_0 + \beta_1 x}$

## 1.2 Aplicações

A regressão é hoje uma das técnicas estatísticas mais usadas em todas as áreas da Ciência, por exemplo:

- Agricultura: prever a produção de leite;
- História: estimar a idade de objectos históricos;
- Medicina: estudar a relação entre a sensibilidade à insulina e o índice de massa corporal das mulheres;
- etc.

## 1.3 Origem da Análise da Regressão

A origem da análise de regressão remonta ao século XIX (1877) e a sua primeira aplicação foi apresentada pelo cientista Galton no artigo "Typical Laws of Heredity" em Inglaterra. Num estudo de biologia mostrou a existência de uma relação linear entre o diâmetro dos grãos de ervilhas "pais" e o diâmetro médio dos grãos descendentes. Ele concluiu que:

- os grãos de ervilhas "filhos" de "pais" de diâmetro maior que a média têm tendência a ter diâmetro maior do que a média mas mais pequeno do que os "pais";
- os grãos de ervilhas "filhos" de "pais" de diâmetro menor que a média têm tendência a ter diâmetro menor do que a média mas maiores do que os "pais".

Desta forma Galton pensou que sempre que uma característica se desenvolvia extraordinariamente num indivíduo, os seus descendentes tendiam a regredir no que respeita a essa característica.

## 1.4 Etapas da Análise de Regressão

- Formulação do problema
- Selecção das variáveis de interesse
- Recolha dos dados
- Formulação do modelo
- Escolha do método de estimação dos parâmetros
- Estimação dos parâmetros
- Validação do modelo

## 1.5 Dados

Os dados são normalmente apresentados na forma:

Número observação	Variável Resposta	$X_1$	$X_2$	$\dots$	$X_p$
1	$y_1$	$x_{11}$	$x_{12}$	$\dots$	$x_{1p}$
2	$y_2$	$x_{21}$	$x_{22}$	$\dots$	$x_{2p}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
n	$y_n$	$x_{n1}$	$x_{n2}$	$\dots$	$x_{np}$

onde  $x_{ij}$  refere-se à observação  $i$  da  $j$ -ésima variável.

As variáveis poder ser classificadas em **quantitativas** e **qualitativas**.

Em geral, a variável resposta é uma variável quantitativa. Quando a variável resposta é binária, aplica-se a **Regressão Logística**

## 1.6 Representação Gráfica dos Dados

Seja  $X$  e  $Y$  duas variáveis aleatórias contínuas e pretende-se estudar a relação entre elas. Para isso medem-se os valores de  $X$  e  $Y$  numa amostra de  $n$  indivíduos, obtendo-se os pares de observações  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ . O método mais simples para observar a relação entre as variáveis é representar estes pontos num gráfico de duas dimensões, no eixo horizontal representam-se os valores da variável  $X$  e no eixo vertical os da variável  $Y$ . Este gráfico designa-se por **diagrama de dispersão**, em inglês "scatter plot", e sugere-nos qual a função mais adequada para ajustar os dados.

**Exemplo:** Interessa estudar a relação entre a resistência de um determinado tipo de plástico ( $Y$ ) e o tempo que decorre a partir da conclusão do processo de moldagem até ao momento de medição da resistência ( $x$ [horas]). As observações que se seguem foram efectuadas em 12 peças construídas com este plástico, escolhidas aleatoriamente.

$x_i$	32	48	72	64	48	16	40	48	48	24	80	56
$y_i$	230	262	323	298	255	199	248	279	267	214	359	305

- Construir o diagrama de dispersão. é:
- O modelo de regressão linear parece ser adequado?

Algumas questões podem ser colocadas:

- Qual a "melhor" equação de recta  $Y = \beta_0 + \beta_1 x$  para as 12 observações?
- E se as 12 observações são apenas uma amostra aleatória duma população mais vasta, o que se pode afirmar sobre a recta populacional?

## 1.7 Método de Estimação

O problema da análise de regressão consiste em estimar os **parâmetros ou coeficientes do modelo**,  $\beta_0, \beta_1, \dots, \beta_p$  a partir de uma amostra de dados, de modo que os erros sejam os mais pequenos possíveis.

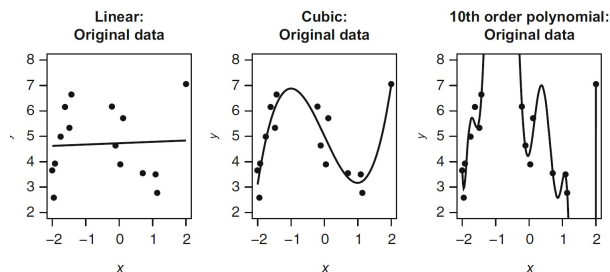
A forma mais habitual de estimar os coeficientes de regressão baseia-se no **Método dos Mínimos Quadrados**. No entanto outros métodos podem ser usados: **Método da Máxima Verosimilhança**, **Método Ridge**, etc.

As estimativas dos parâmetros  $\beta_0, \beta_1, \dots, \beta_p$  são denotadas por,  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$  e a equação do modelo estimada é dada por:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p$$

## 1.8 Ideias prévias da modelação

- Todos os modelos são apenas aproximações da realidade. Uns são melhores que outros.
- O princípio da parcimónia na modelação: de entre os modelos considerados adequados, é preferível o mais simples.
- Não há (necessariamente) relação de causa e efeito entre variável resposta e preditores. A Estatística só pode mostrar que há associação. Uma eventual existência de relação causa e efeito é exterior à Estatística.
- Os modelos estatísticos não são determinísticos: descrevem uma relação de fundo, sabendo-se que há variação das observações em torno dessa relação de fundo. Essa variabilidade tem de ser incorporada no modelo.
- Um modelo muito simples ou muito complexo não modela bem os dados. Modelos complexos podem se ajustar bem aos dados fornecidos, mas geralmente não generalizam bem para outros conjuntos de dados (isso é chamado de *overfitting*)



- No estudo de modelos estatísticos há aspectos diferentes:
  - faceta descritiva: ajustar modelo a dados observados, qualquer que seja a sua origem.

- faceta inferencial: se os dados são uma amostra aleatória duma população, procurar tirar conclusões sobre a população.
- Advertência sobre transformações linearizantes: a regressão linear simples não modela directamente relações não lineares entre  $x$  e  $y$ . Pode modelar uma relação linear entre variáveis transformadas. Linearizar e obter os parâmetros do modelo e depois desfazer a transformação linearizante não produz os mesmos valores de parâmetros que resultariam de minimizar a soma de quadrados dos resíduos directamente na relação não linear, através duma Regressão não linear.
- Para modelos de regressão, existem duas motivações principais:
  - Predição: Produzir previsões precisas a partir de dados novos ou futuros. Por exemplo, pretende-se determinar se existe uma relação (potencialmente causal) entre fumar e o volume expiratório forçado (FEV). Pretende-se saber se fumar tem um efeito sobre o FEV, e em que direcção. Para esse fim, a significância dos coeficientes no modelo são de interesse.
  - Interpretação: Compreender como as variáveis se relacionam entre si. Por exemplo, pretende-se estabelecer a faixa de peso normal para crianças de uma determinada idade. Aqui, o objetivo é poder avaliar se uma criança específica está fora da faixa normal. Neste caso, uma curva de predição relacionando peso à idade é desejada.

**Exemplo:** Considere as vendas (em milhares de unidades) de um produto específico em função dos orçamentos de publicidade (em milhares de dólares) na TV, rádio e jornal.

Suponha que, lhe era solicitado, com base nestes dados, um plano de marketing para o próximo ano que resultará em altas vendas do produto. Que informações seriam úteis para resolver a solicitação?

- Há uma relação entre o orçamento de publicidade e as vendas?
- Quão forte é a relação entre o orçamento de publicidade e as vendas?
- Quais os orçamentos de publicidade que estão associadas às vendas?
- A relação é linear?

**Exercício:** Indica dois exemplos, numa área do seu interesse, onde a análise de regressão pode ser aplicada, respondendo às seguintes questões:

- a) Qual o objectivo do estudo?
- b) Identifica a variável resposta e as variáveis explicativas. Classifica-as.
- c) Indica um possível modelo para o estudo.