Aprendizagem Supervisionada

Supervised Machine Learning

Do livro indicado, Schutt & O'Neil (2014), ler:
capítulo 2 (pp. 24-37)
capítulo 3 (pp. 51-71)

Modelação de dados e Machine Learning

Modelação: como se constroem modelos a partir dos dados que se recolhem.

O que é um modelo?

Um modelo é a nossa tentativa de representar a natureza por uma determinada perspectiva: modelos arquitetónicos, físicos, biológicos, matemáticos…

Modelação de dados

Modelos estatísticos

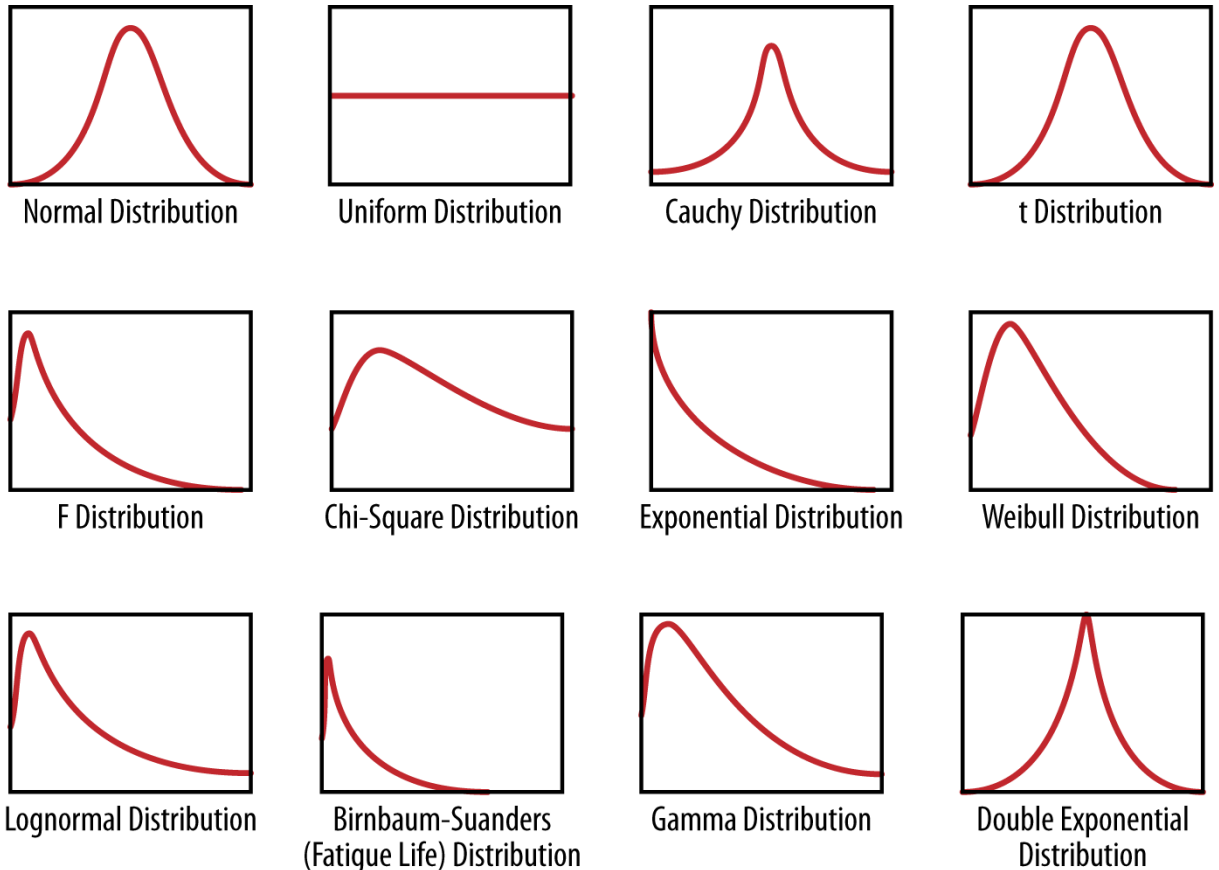
Simplicidade: simplicidade vs precisão

Os modelos simples são fáceis de interpretar e de compreender

Um modelo simples pode ser rápido de obter e validar e pode representar 90% do processo; um modelo mais complexo pode levar meses a conseguir e representar 92%...

Distribuição de probabilidades

As distribuições de probabilidade são as fundações dos modelos estatísticos.



Distribuição de probabilidades

- Os **processos naturais** tendem a gerar registos/observações cuja forma empírica pode ser aproximada por **funções matemáticas** com alguns **parâmetros** que podem ser **estimados** a partir desses dados.

- Nem todos os processos geram dados que se assemelhem a **distribuições conhecidas**, mas uma grande maioria gera.

- Podem utilizar-se estas funções para construir **blocos** nos nossos modelos.

Algoritmos básicos

- Regressão linear
- k-Nearest Neighbors (k-NN)
- k-means

Regressão Linear

A **regressão linear** é um dos métodos estatísticos mais comuns, que permite exprimir uma relação matemática entre duas ou mais variáveis ou atributos.

Ao utilizá-lo, admite-se que existe uma **relação linear** entre a variável dependente (saída do modelo) e a(as) variável(-eis) independente(s) (preditores).

Regressões Lineares Simples e Múltipla

Regressão Linear

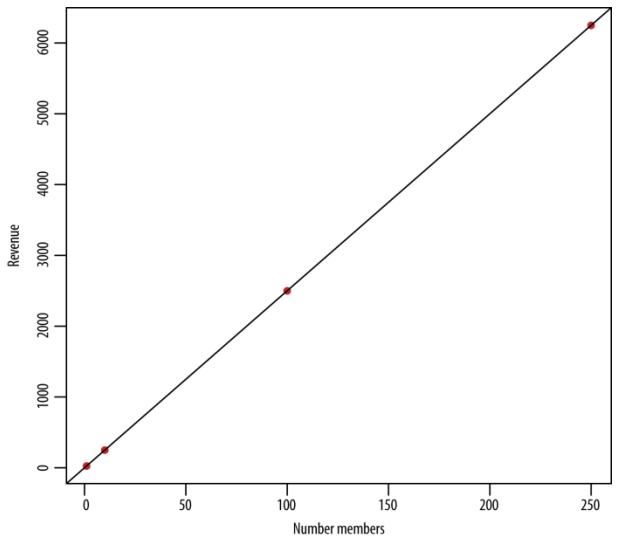
Determinismo:

$$y = f(x) = \beta_0 + \beta_1 * x.$$

a equação da recta é descrita por uma constante e um declive, mas é sempre determinista.

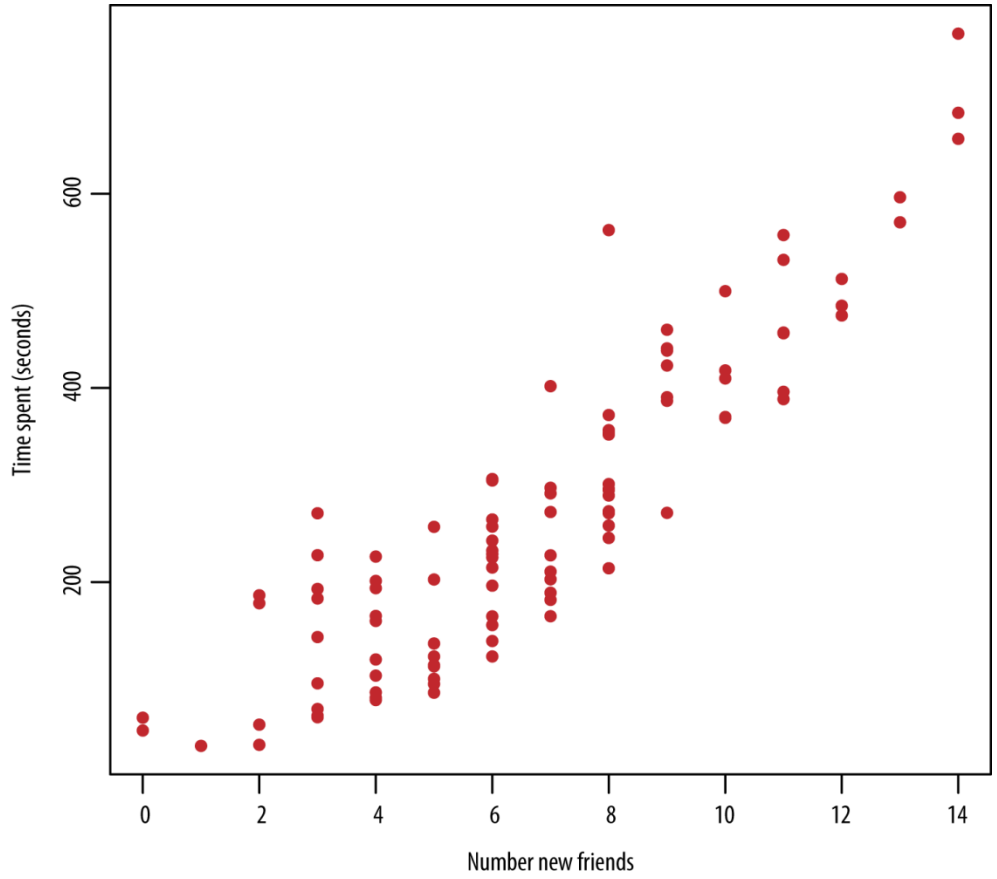
$$S = \{(x, y) = (1, 25), (10, 250), (100, 2500), (200, 5000)\}$$

$$y = 25x$$



Regressão Linear

- Relação ‘parece’ linear
- Não é determinista

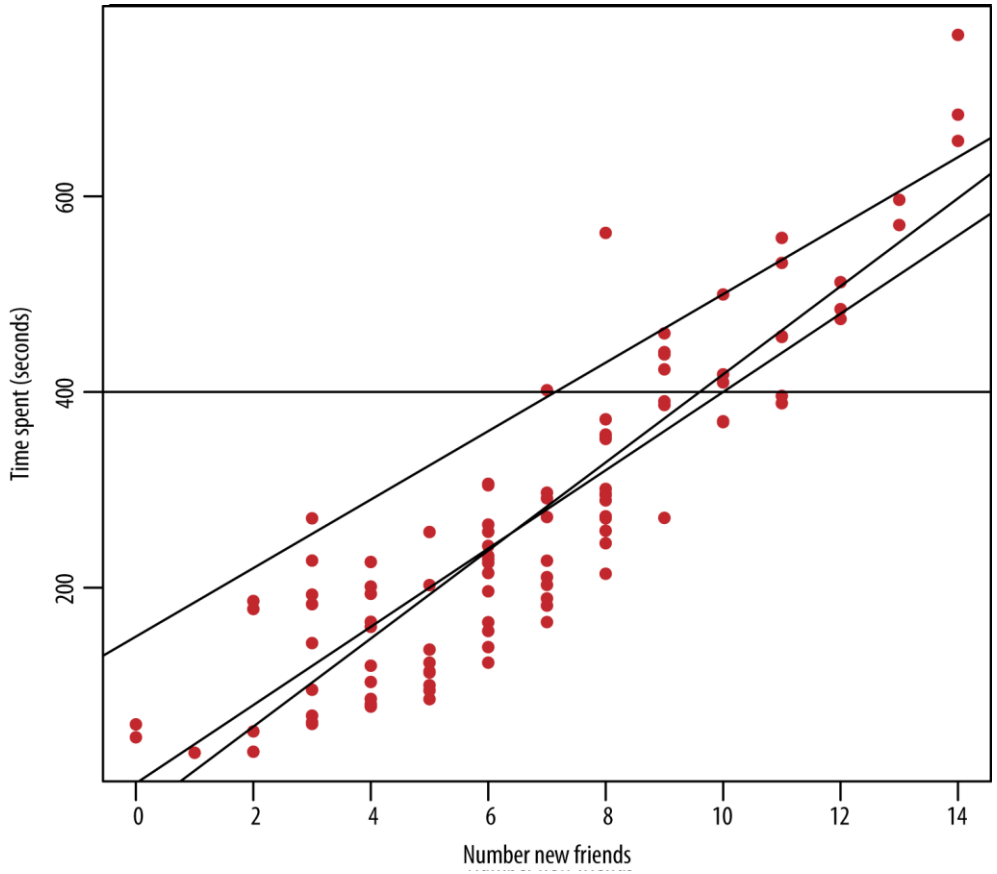


Regressão Linear

- Relação ‘parece’ linear
- Não é determinista

Com o modelo queremos encontrar:

- tendência
- variabilidade



Regressão Linear

Como a relação é linear, assumimos a forma:

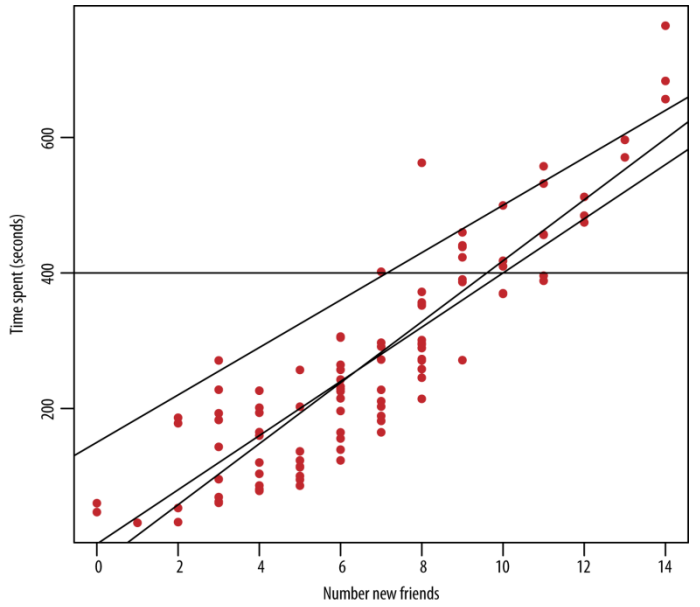
$$y = \beta_0 + \beta_1 x$$

Resta-nos agora encontrar os parâmetros β0 e β1, utilizando os dados observados

$$(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$$

Já temos o modelo!

Agora temos de o ajustar…



Método dos mínimos quadrados

Objetivo: encontrar a recta que permita minimizar a distância de todos os pontos à recta de ajuste

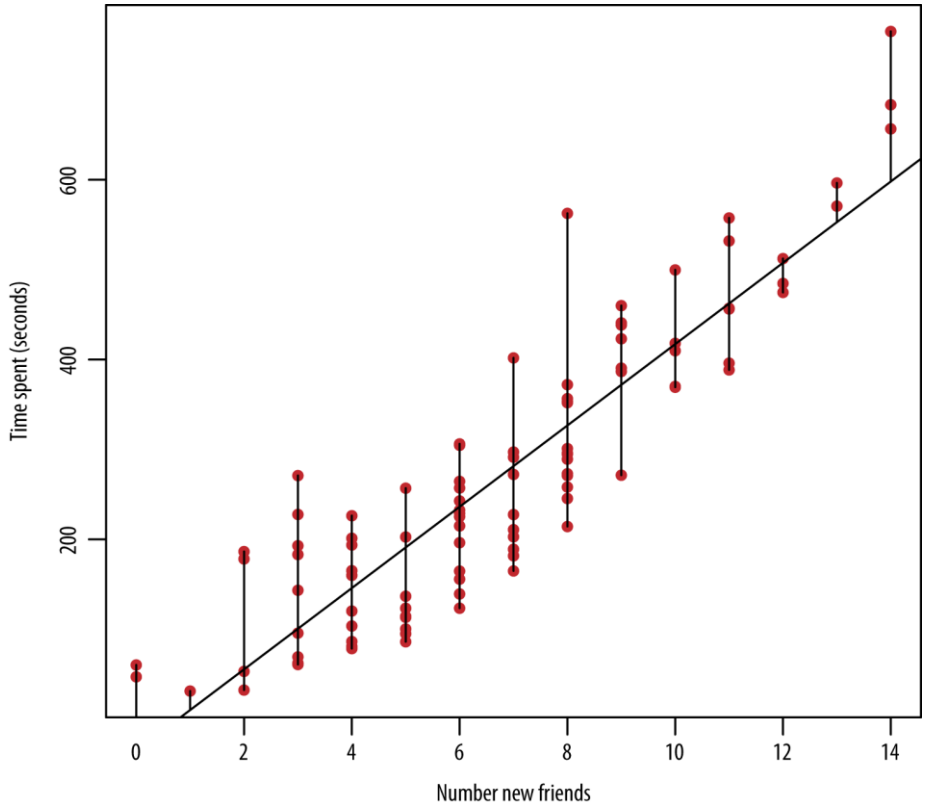
Estimação pelo método

dos mínimos quadrados

$$RSS(\beta) = \sum_i (y_i - \beta x_i)^2$$

$$RSS(\beta) = (y - \beta x)^t (y - \beta x)$$

$$\hat{\beta} = (x^t x)^{-1} x^t y$$

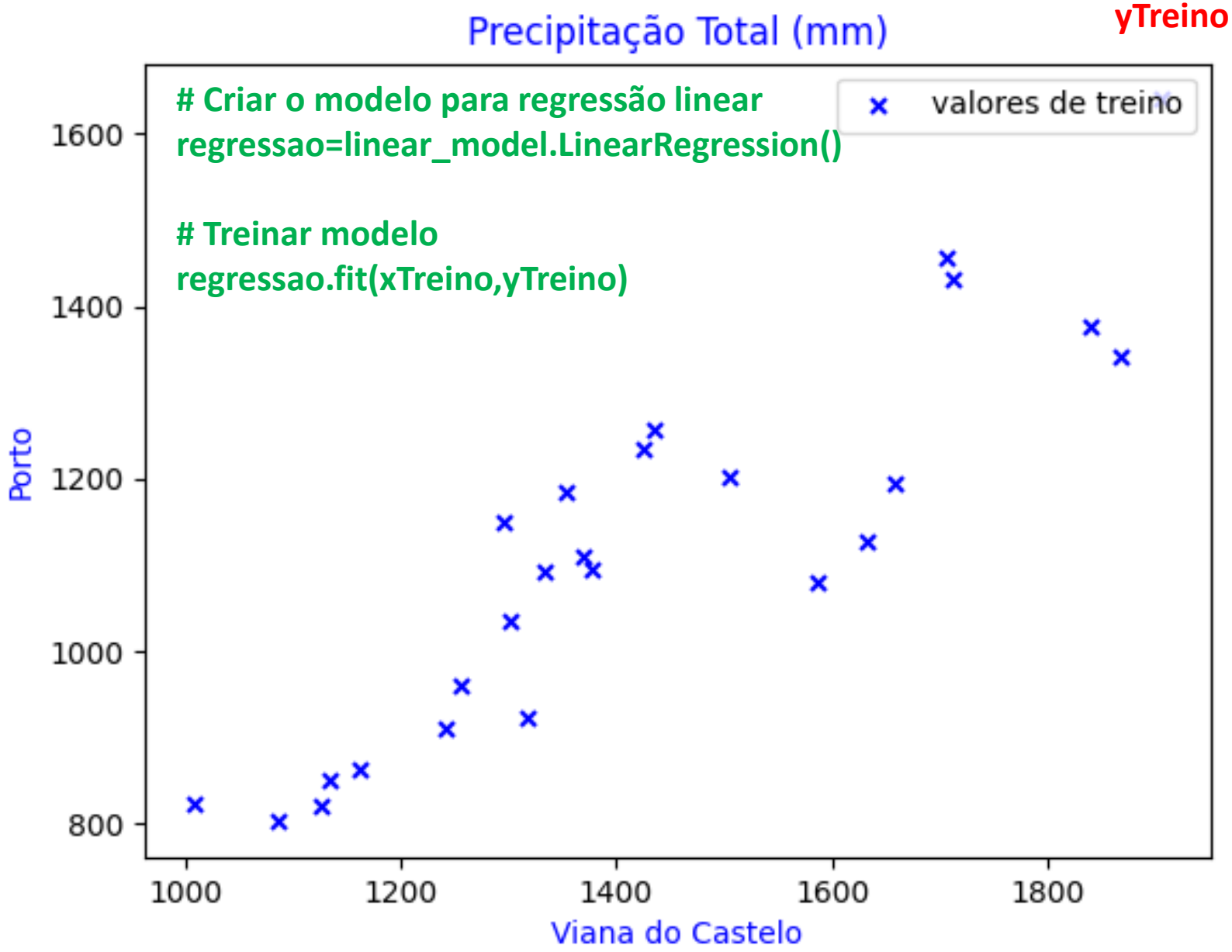


Regressão Linear

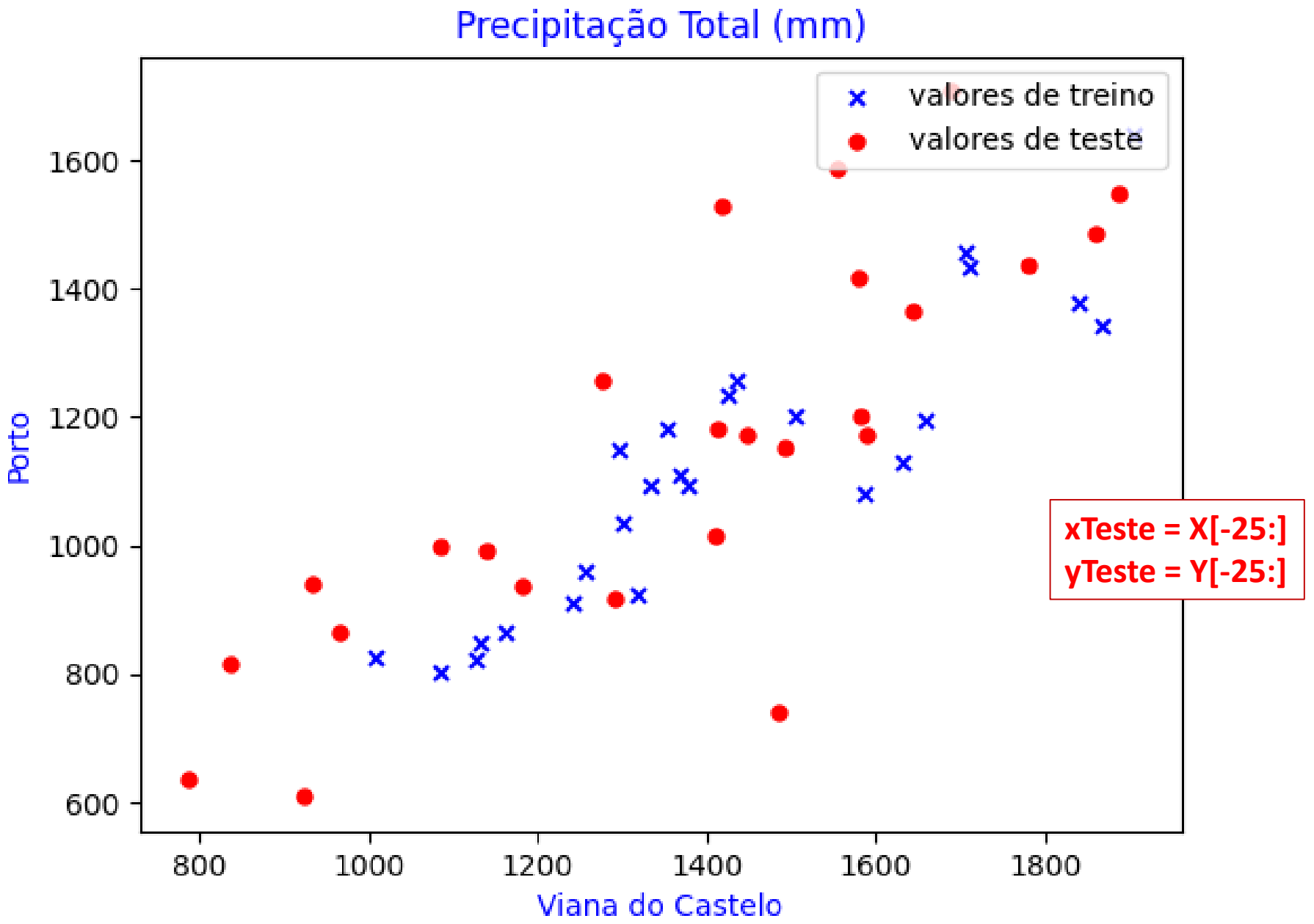
- Preparação dos Dados
- Escolha do método (Fitting Method)
- Escolha do modelo ou gama de modelos
- Ajuste do modelo aos dados
- Análise da qualidade e ajuste do modelo obtido
- Prever ou simular resposta com novos dados
- Partilha do modelo obtido

Ajuste do modelo aos dados: treino

xTreino = X[:-25]
yTreino = Y[:-25]



Análise da qualidade e ajuste do modelo obtido: teste

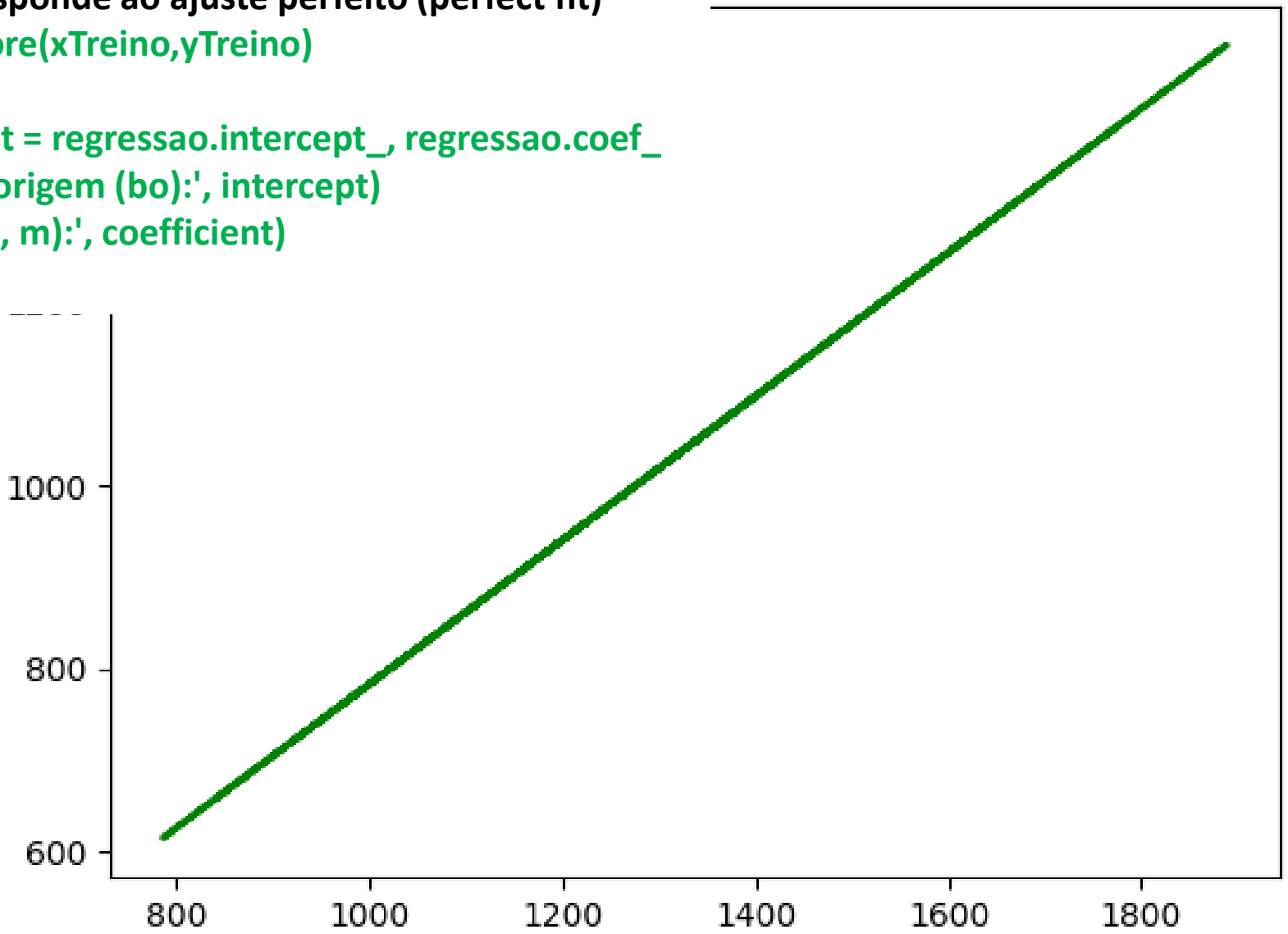


Representação do modelo obtido: equação y = bo + mx

O valor R^2 = 1 corresponde ao ajuste perfeito (perfect fit)

```
r_sq = regressao.score(xTreino,yTreino)

intercept, coefficient = regressao.intercept_, regressao.coef_
print('ordenada na origem (bo):', intercept)
print('declive (slope, m):', coefficient)
```



Prever ou simular resposta com novos dados

Qual será a precipitação no Porto se a precipitação anual for de 1300 mm em Viana do Castelo?