

DETECTOR DE RISCO DE DOENÇAS CARDIOVASCULARES

ID	Idade (>50?)	Diabetes	Hipertensão	Risco
1	Sim	Sim	Sim	Doença
2	Sim	Sim	Não	Doença
3	Sim	Não	Sim	Doença
4	Sim	Não	Não	Não
5	Não	Sim	Sim	Doença
6	Não	Sim	Não	Não
7	Não	Não	Sim	Não
8	Não	Não	Não	Não

DETECTOR DE RISCO DE DOENÇAS CARDIOVASCULARES

ID	Idade (>50?)	Diabetes	Hipertensão	Risco
1	Sim	Sim	Sim	Doença
2	Sim	Sim	Não	Doença
3	Sim	Não	Sim	Doença
4	Sim	Não	Não	Não
5	Não	Sim	Sim	Doença
6	Não	Sim	Não	Não
7	Não	Não	Sim	Não
8	Não	Não	Não	Não

1º Analisamos os riscos:

- Total: 8 pacientes
- 4 com doença
- 4 sem doença

DETECTOR DE RISCO DE DOENÇAS CARDIOVASCULARES

ID	Idade (>50?)	Diabetes	Hipertensão	Risco
1	Sim	Sim	Sim	Doença
2	Sim	Sim	Não	Doença
3	Sim	Não	Sim	Doença
4	Sim	Não	Não	Não
5	Não	Sim	Sim	Doença
6	Não	Sim	Não	Não
7	Não	Não	Sim	Não
8	Não	Não	Não	Não

1º Analisamos os riscos:

- Total: 8 pacientes
- 4 com doença
- 4 sem doença

2º Estimando as probabilidades a priori:

$$P(\text{Doença}) = 4/8 = 0,5$$

$$P(\text{Não Doença}) = 4/8 = 0,5$$

DETECTOR DE RISCO DE DOENÇAS CARDIOVASCULARES

ID	Idade (>50?)	Diabetes	Hipertensão	Risco
1	Sim	Sim	Sim	Doença
2	Sim	Sim	Não	Doença
3	Sim	Não	Sim	Doença
4	Sim	Não	Não	Não
5	Não	Sim	Sim	Doença
6	Não	Sim	Não	Não
7	Não	Não	Sim	Não
8	Não	Não	Não	Não

3º Estimando as probabilidades condicionais

Idade > 50:

$$P(\text{Idade}=Sim \mid \text{Doença}) = 3/4 = 0,75$$

$$P(\text{Idade}=N\ão \mid \text{Doença}) = 1/4 = 0,25$$

$$P(\text{Idade}=Sim \mid \text{N}\ão \text{ Doença}) = 1/4 = 0,25$$

$$P(\text{Idade}=N\ão \mid \text{N}\ão \text{ Doença}) = 3/4 = 0,75$$

DETECTOR DE RISCO DE DOENÇAS CARDIOVASCULARES

4º Classificando um novo paciente

Suponha um paciente:
• Idade >50? Sim
• Diabetes? Não
• Hipertensão? Sim

Probabilidade de Doença

$$\begin{aligned} &= P(\text{Doença}) \times P(\text{Idade}=Sim|\text{Doença}) \times P(\text{Diab}=Não|\text{Doença}) \\ &\quad \times P(\text{Hip}=Sim|\text{Doença}) \\ &= 0,5 \times 0,75 \times 0,5 \times 0,5 = 0,09375 \end{aligned}$$

Probabilidade de Não Doença

$$\begin{aligned} &= P(\text{Não}) \times P(\text{Idade}=Sim|Não) \times P(\text{Diab}=Não|Não) \\ &\quad \times P(\text{Hip}=Sim|Não) \\ &= 0,5 \times 0,25 \times 0,5 \times 0,25 = 0,015625 \end{aligned}$$

DETECTOR DE RISCO DE DOENÇAS CARDIOVASCULARES

4º Classificando um novo paciente

Suponha um paciente:
• Idade >50? Sim
• Diabetes? Não
• Hipertensão? Sim

Probabilidade de Doença = $0,09375 = 9,375\%$

Probabilidade de Não Doença = $0,015625 = 1,5625\%$

Comparando: $0,0937 > 0,0156$
Classificamos como Doença (risco alto)

VARIACÕES DE NAIVE BAYES

Bernoulli Naive Bayes



Dados Binários
(ex: presença/ausência de palavra)

Multinomial Naive Bayes



Contagens
(ex: número de vezes que a palavra aparece)

Gaussiana Naive Bayes



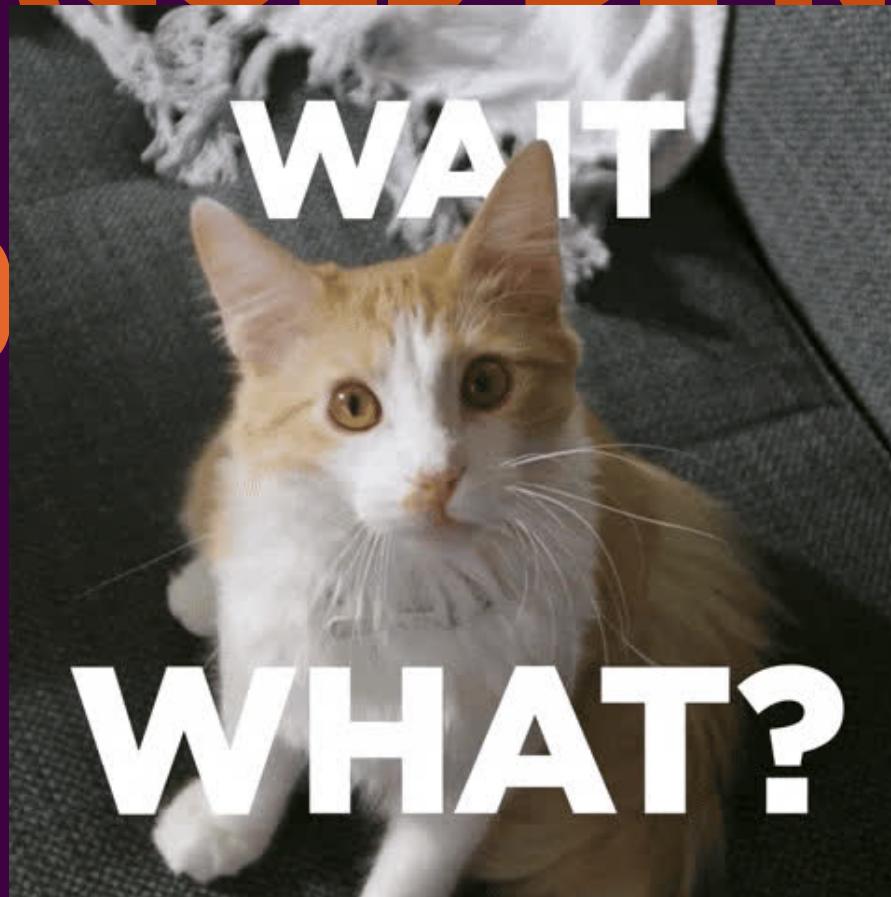
Atributos Contínuos
(assumem distribuição normal)

VARIACÕES DE NAIVE

Bernoulli Naive Bayes

!

Dados Binários
(ex: presença/ausência
de palavra)



Gaussian Naive Bayes

!

Atributos Contínuos
(presumem distribuição
normal)

RESUMINDO

- É um classificador probabilístico baseado no Teorema de Bayes, que assume que as variáveis de entrada são independentes entre si, dado a classe — por isso o nome "naive" (ingênuo).
 - Suposição de independência nem sempre é realista
 - Pode ter desempenho inferior se as features forem fortemente correlacionadas
- Em contrapartida:
 - Necessita de menos espaço na memória
 - Requer menos da CPU
 - Não necessita de um conjunto de dados extenso

DOCUMENTAÇÃO



[Documentação Naives Bayes](#)

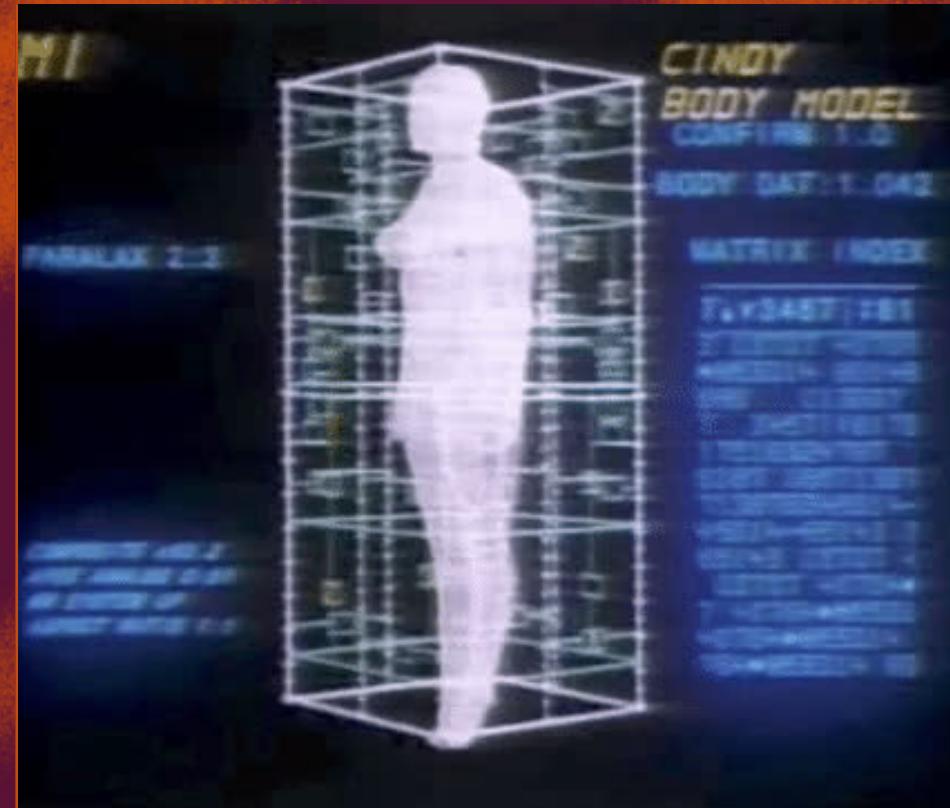
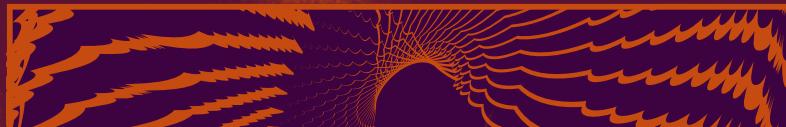


EDG



Regressão Linear

- Como podemos modelar relações lineares entre variáveis?

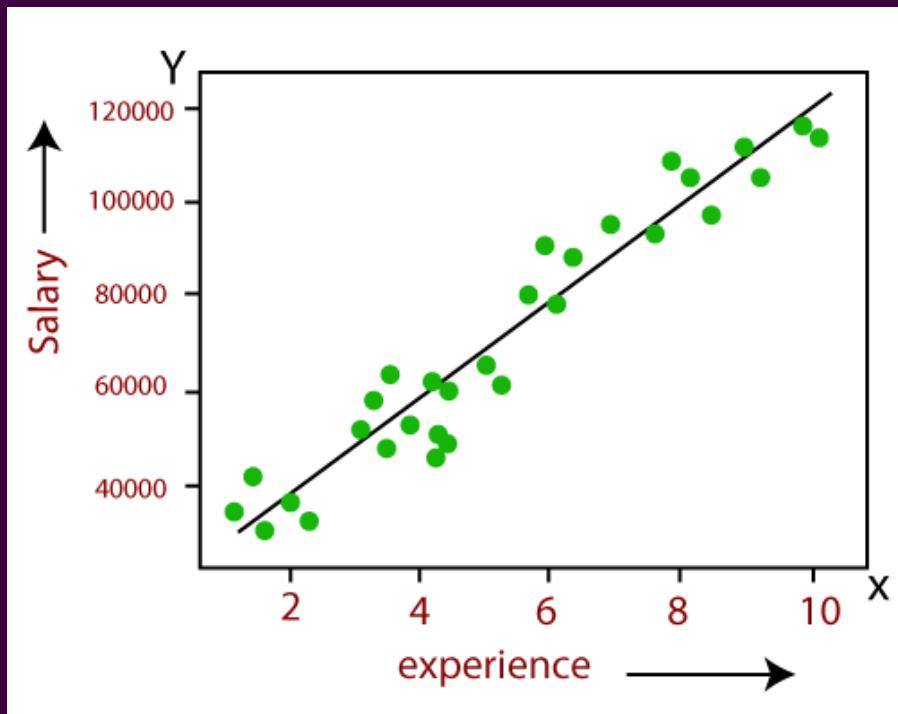


REGRESSÃO LINEAR

Imagine que você tem um gráfico com vários pontos representando Salários dos funcionários de uma empresa:

- No eixo X, a quantidade de experiência.
- No eixo Y, o valor do salário.

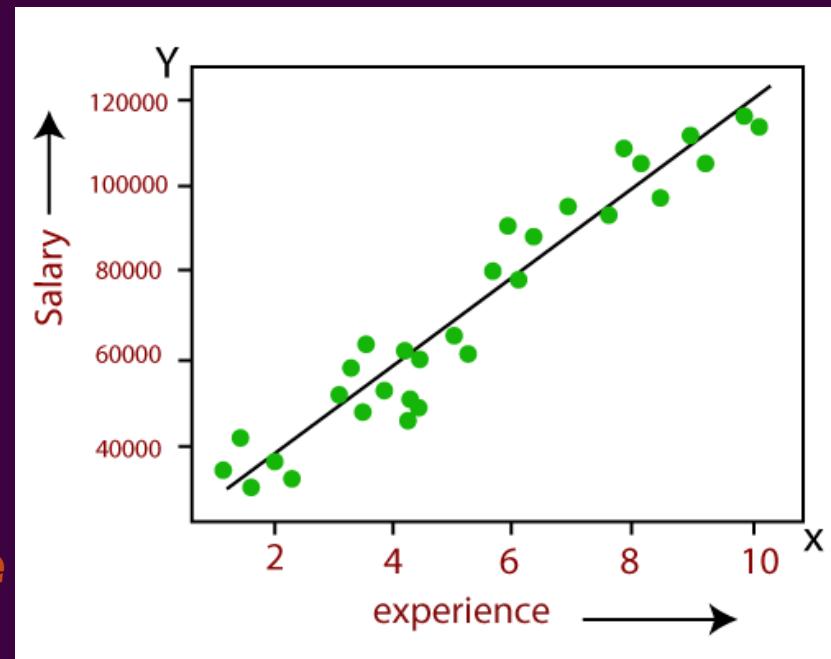
A Regressão Linear tenta traçar uma linha reta que melhor se ajusta a esses pontos. Essa linha representa uma relação matemática simples entre as variáveis.



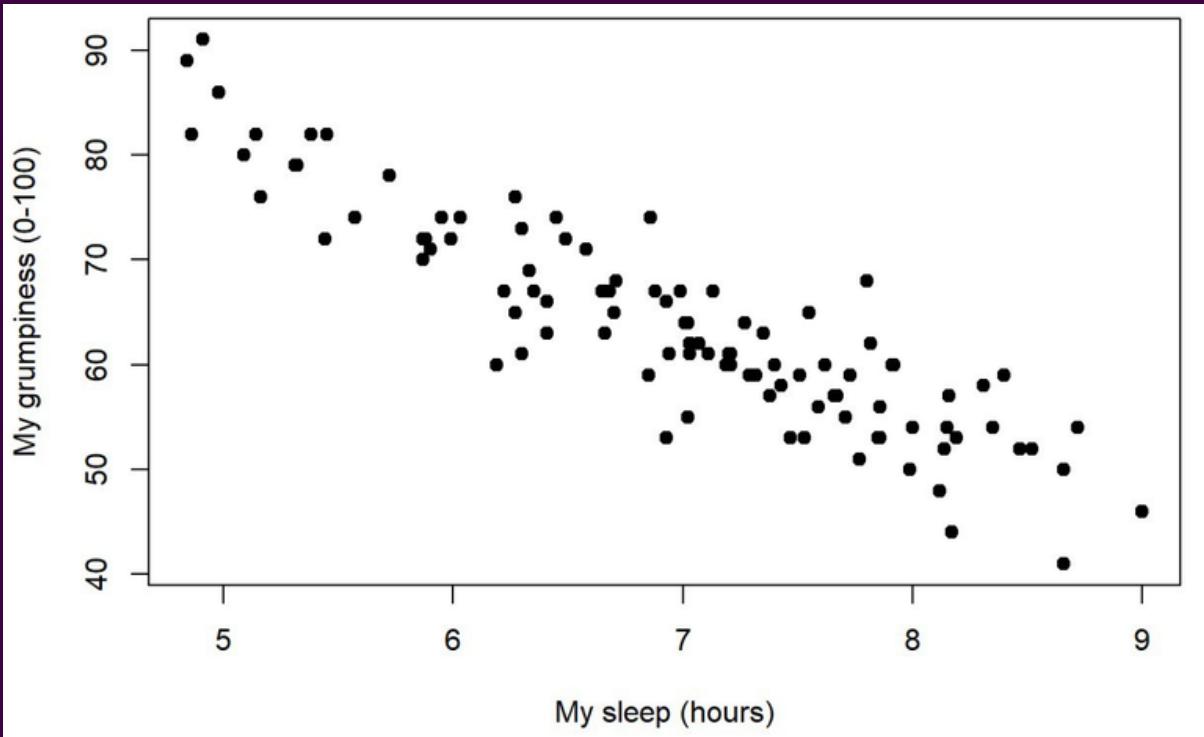
REGRESSÃO LINEAR

$$y = a \cdot x + b$$

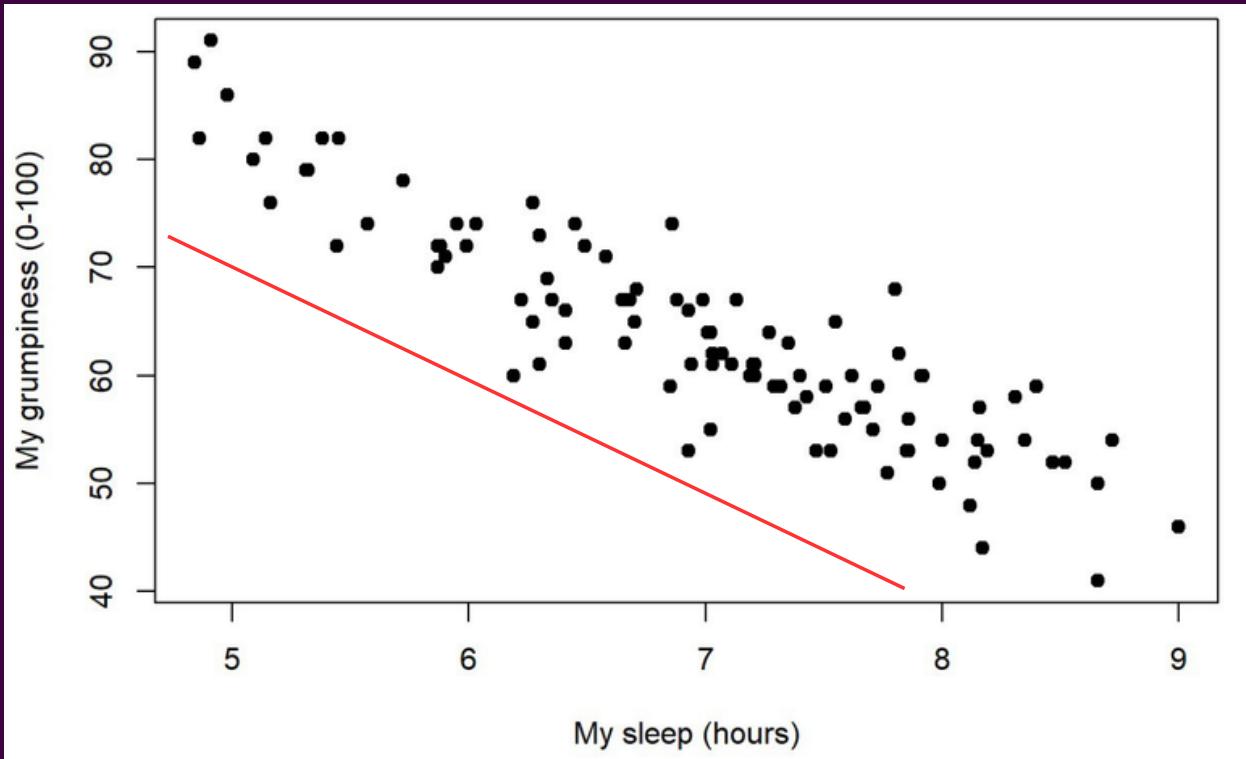
- **y: valor que queremos prever (ex: valor do salário)**
- **x: valor que temos (ex: quantidade de experiência)**
- **a: inclinação da linha (o quanto o preço sobe para cada nível de experiência a mais)**
- **b: onde a linha cruza o eixo y (valor inicial, mesmo com tamanho zero)**



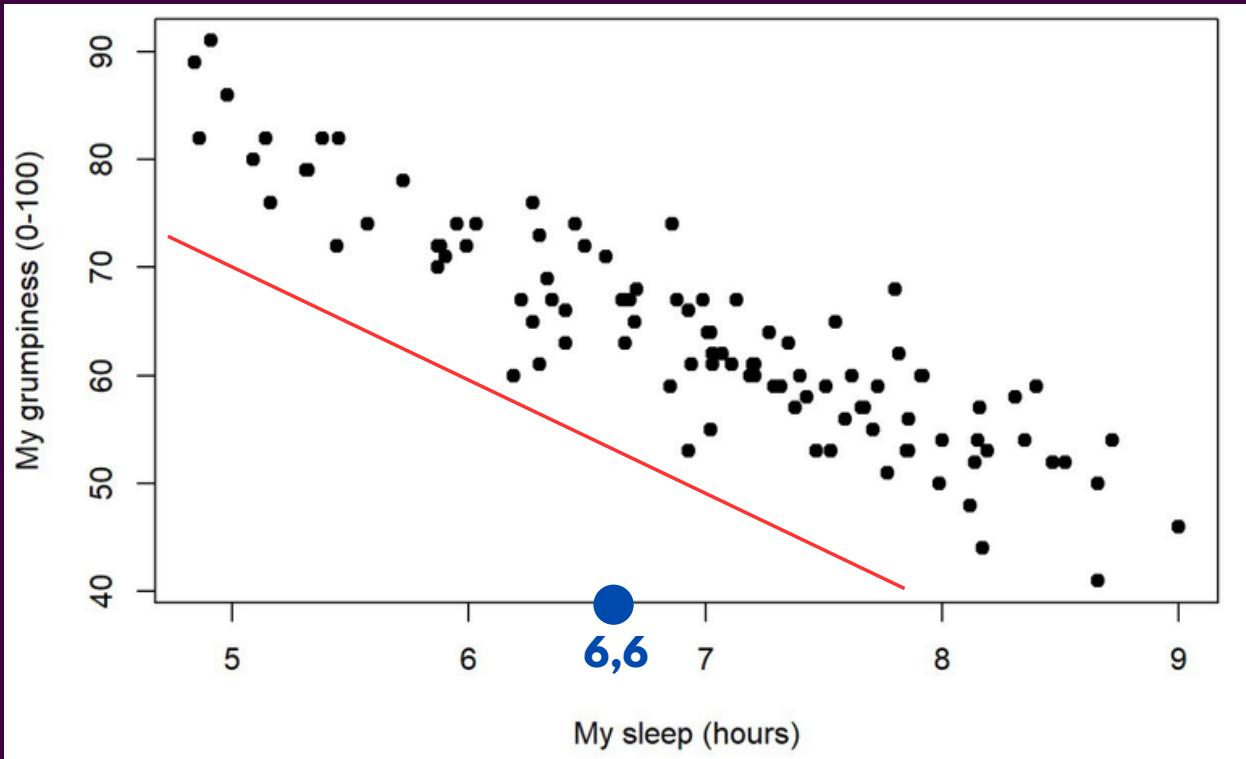
REGRESSÃO LINEAR



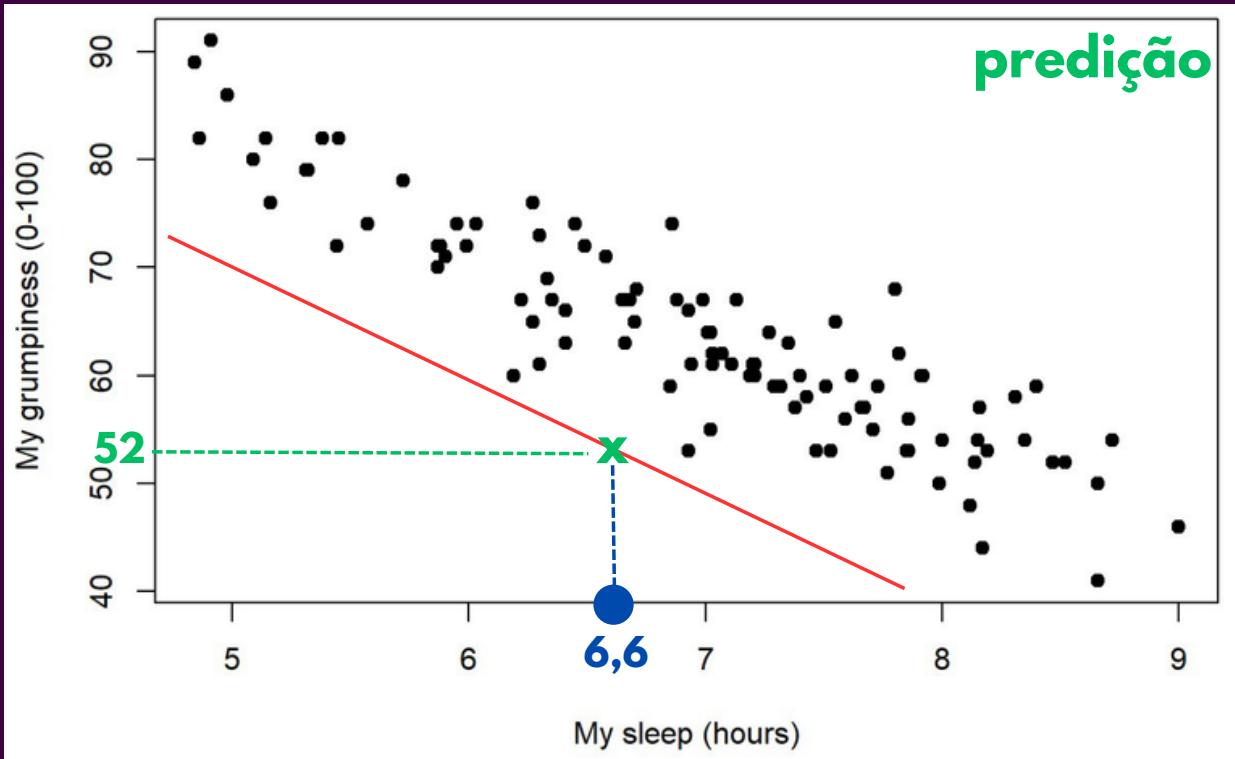
REGRESSÃO LINEAR



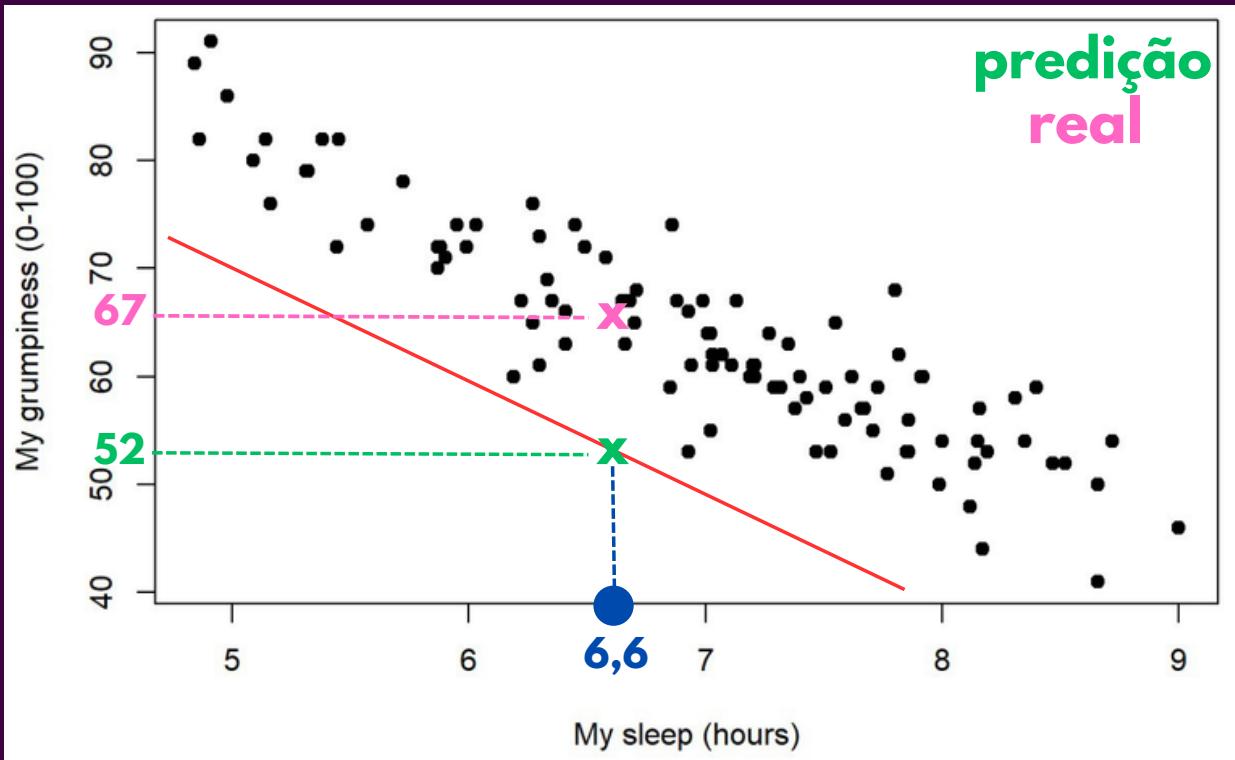
REGRESSÃO LINEAR



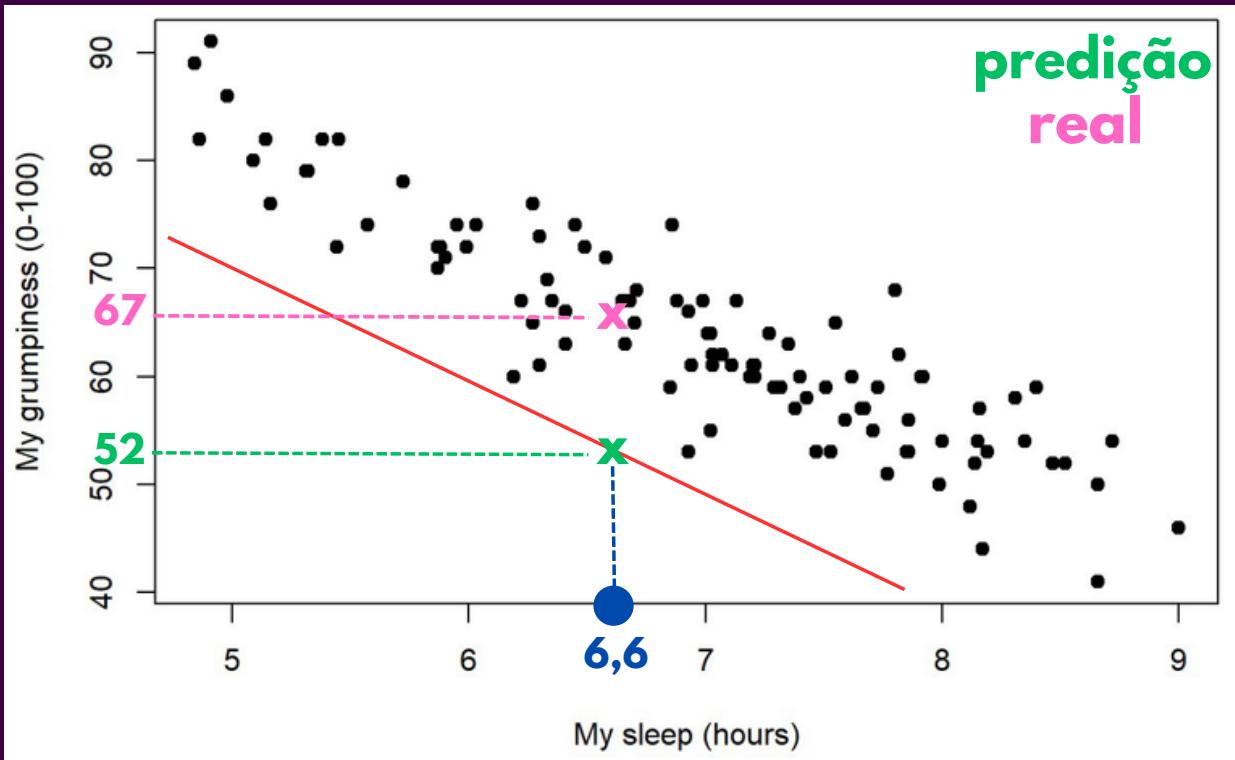
REGRESSÃO LINEAR



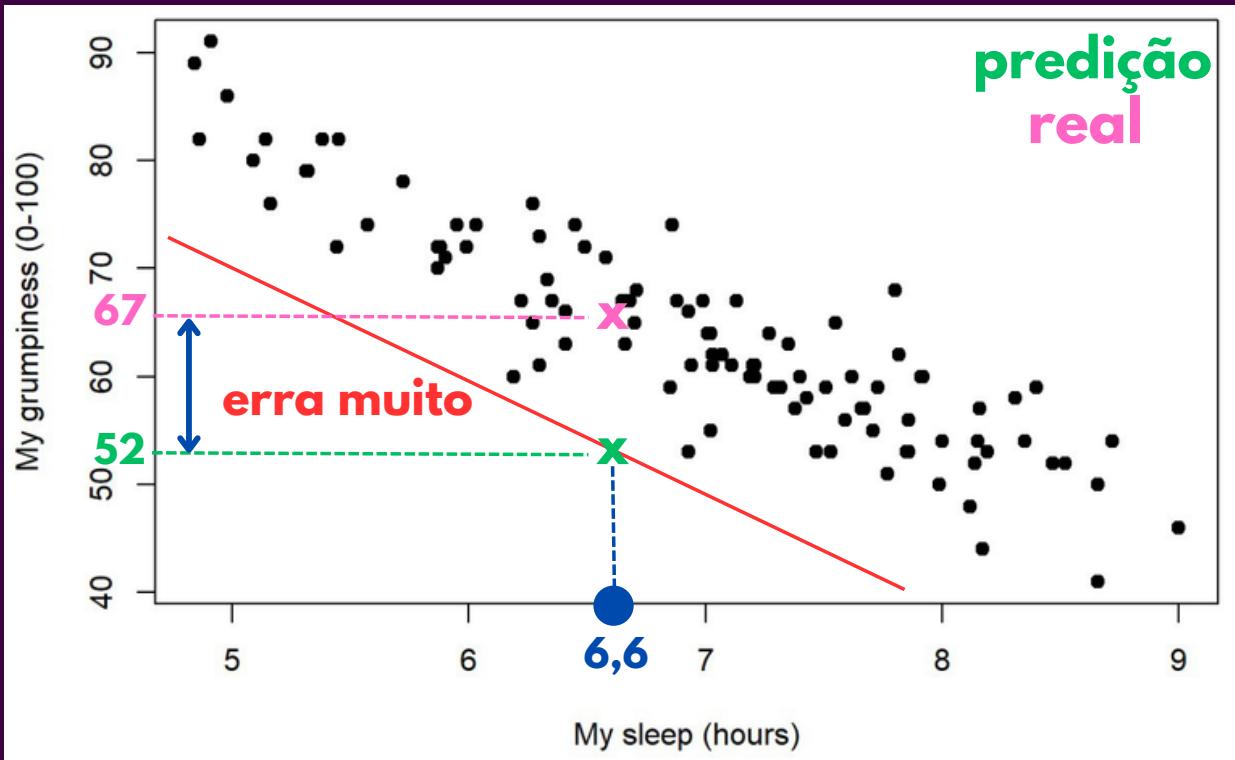
REGRESSÃO LINEAR



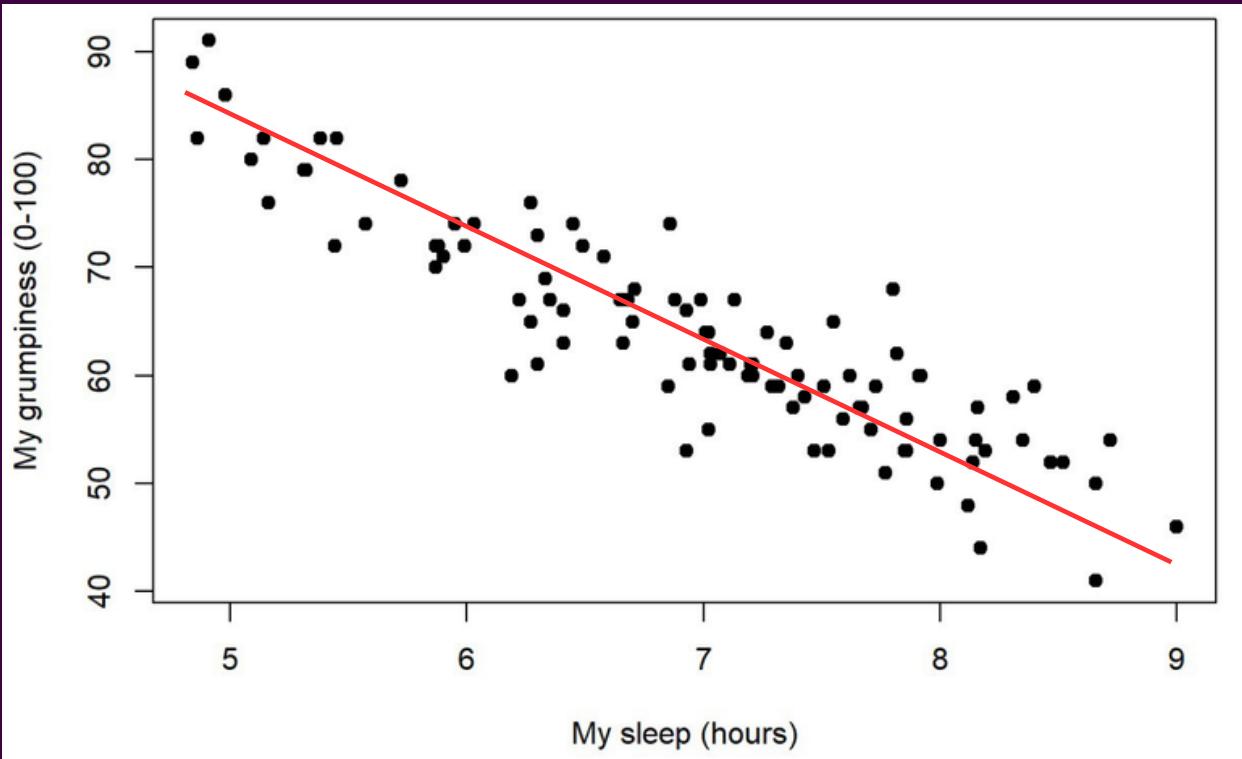
REGRESSÃO LINEAR



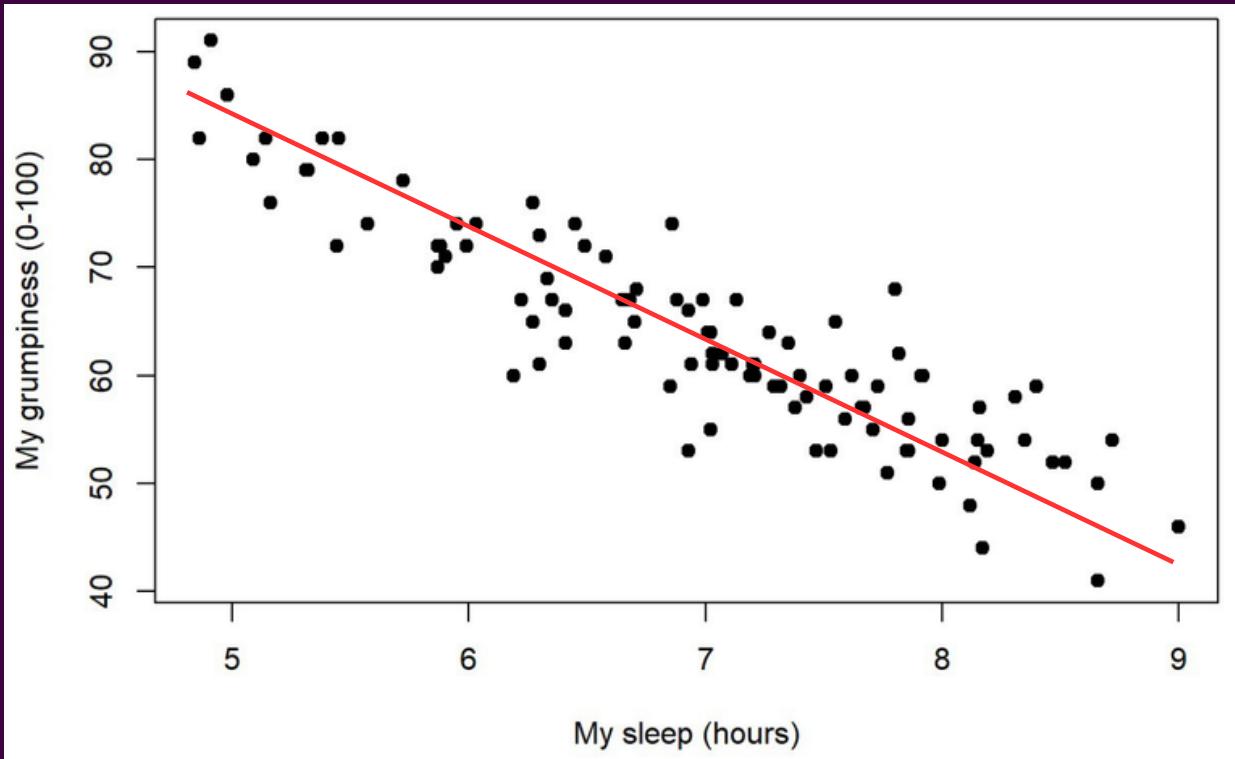
REGRESSÃO LINEAR



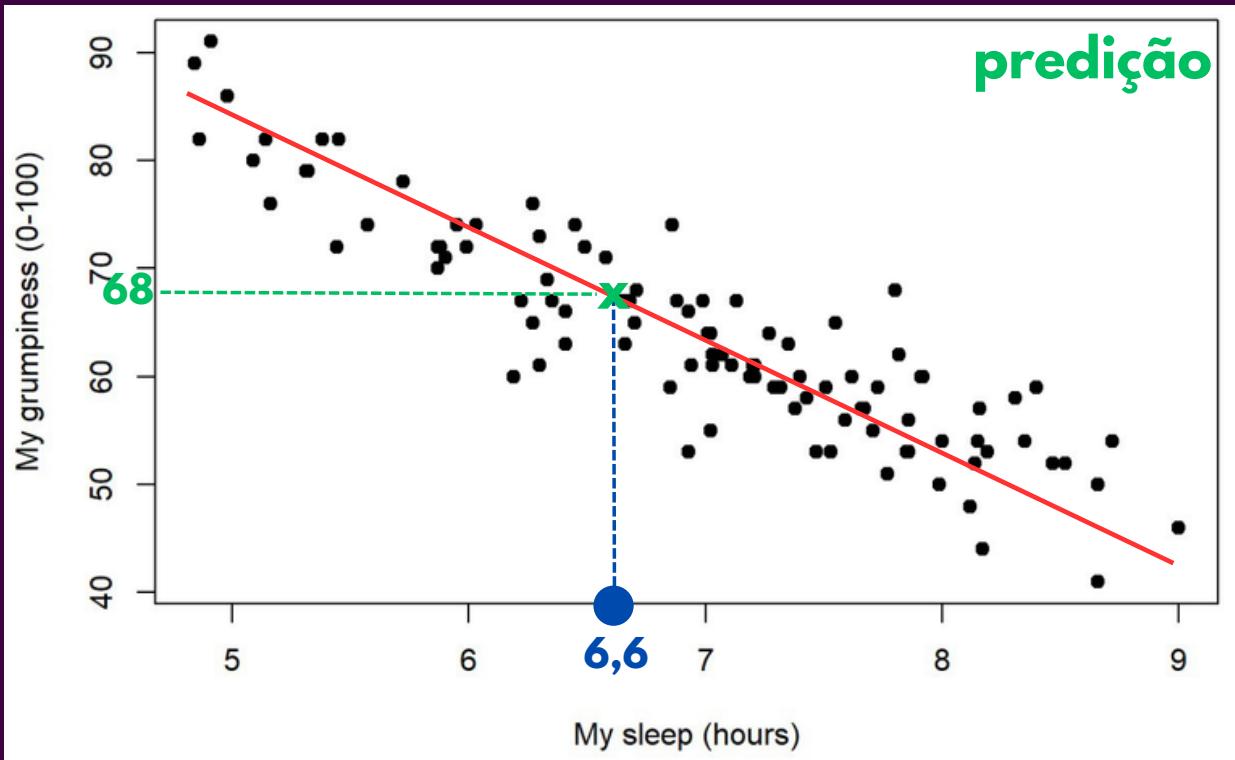
REGRESSÃO LINEAR



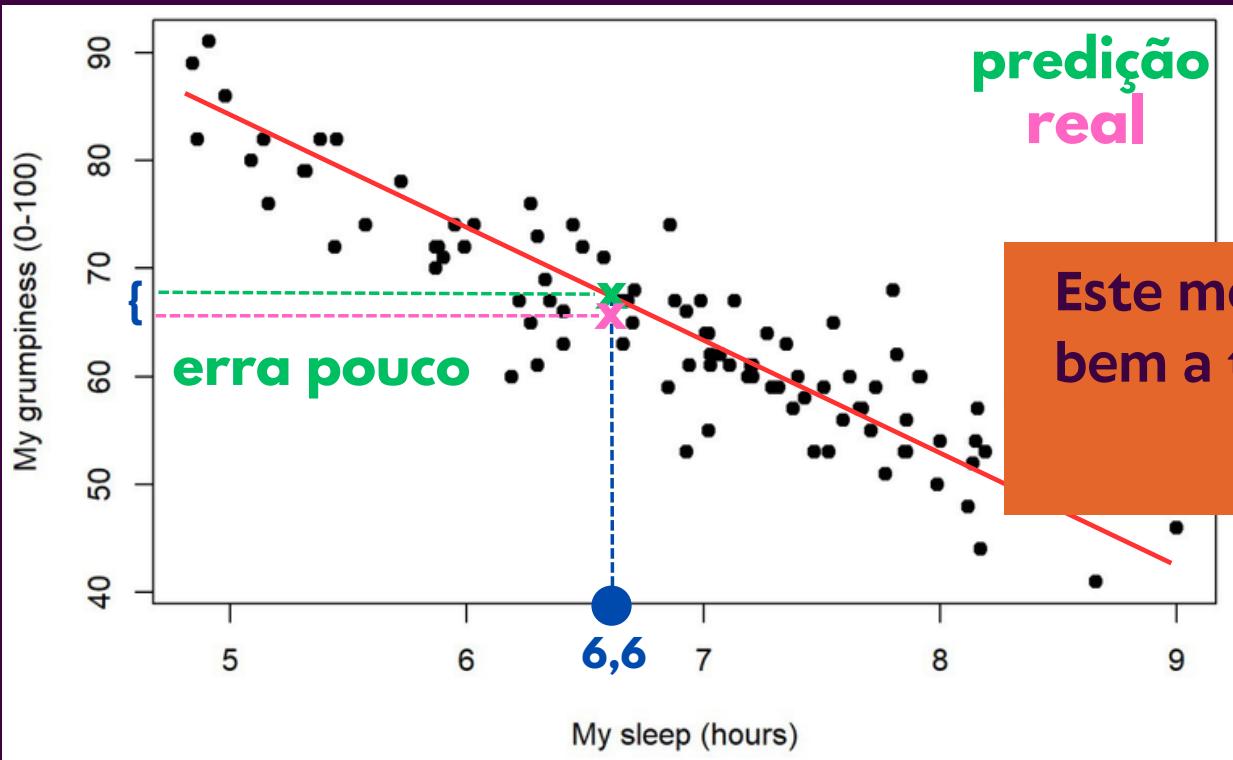
REGRESSÃO LINEAR



REGRESSÃO LINEAR

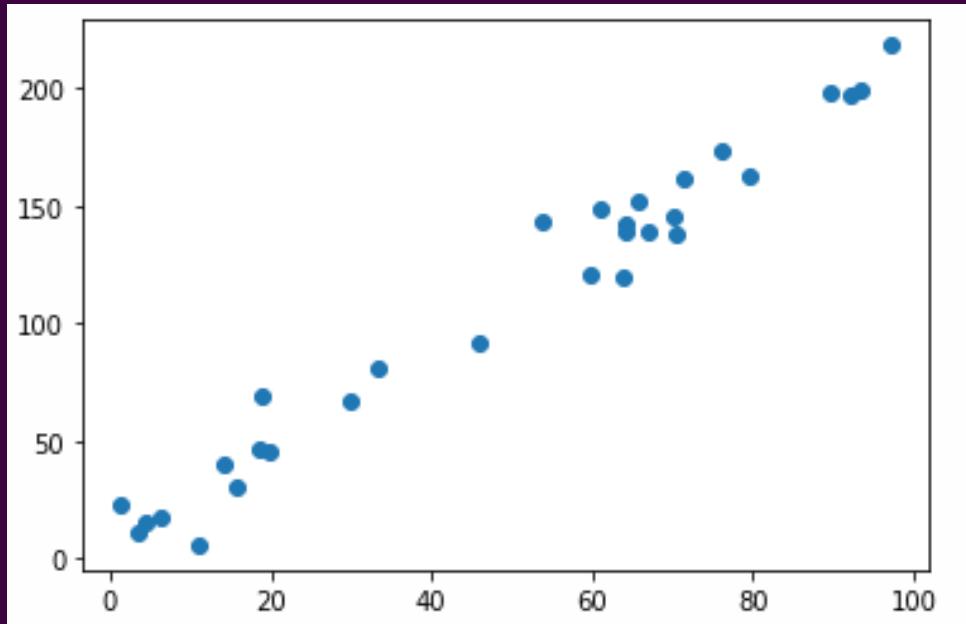


REGRESSÃO LINEAR



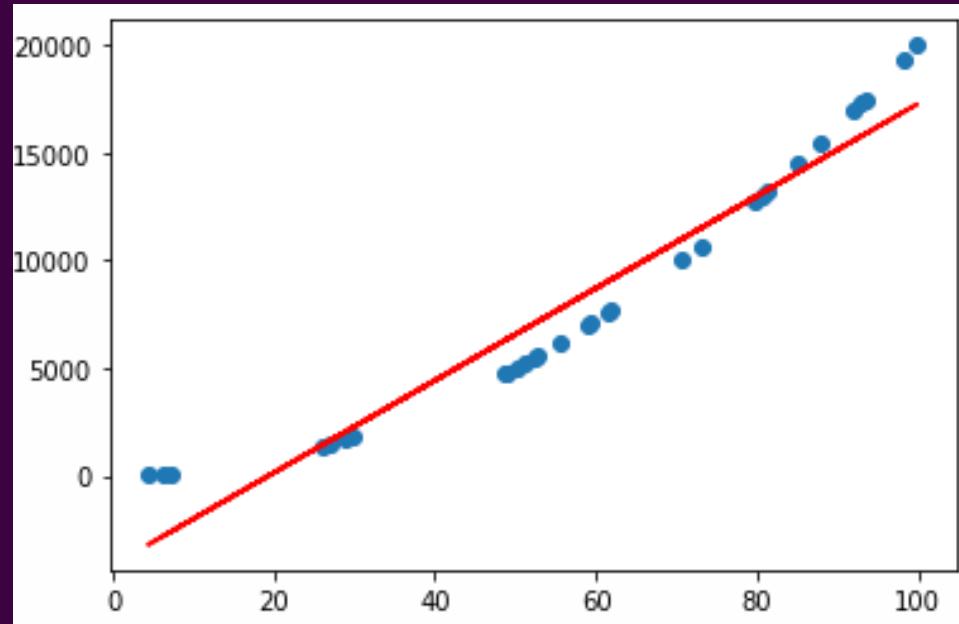
RUÍDO

Como os dados se comportam na **realidade**.
Tendências lineares não são incomuns



NÃO LINEARIDADE

Apesar de minimizar um erro, segue-se uma suposição do comportamento dos dados



MINIMIZAÇÃO DO ERRO

Como o modelo aprende?

- Ele ajusta os valores de a e b para que a linha passe o mais perto possível de todos os pontos do gráfico. Isso é feito minimizando o erro entre os valores reais e os valores previstos. Esse erro é medido geralmente com:

Distância vertical entre cada ponto e a linha

MINIMIZAÇÃO DO ERRO

Como o modelo aprende?

- Ele ajusta os valores de a e b para que a linha passe o mais perto possível de todos os pontos do gráfico. Isso é feito minimizando o erro entre os valores reais e os valores previstos. Esse erro é medido geralmente com:

Erro quadrático médio (MSE)

ERRO QUADRÁTICO MÉDIO

Erro quadrático médio (MSE)

- É a média do quadrado das distâncias verticais entre os pontos reais e os pontos previstos pela linha.
- Em vez de só calcular a diferença, o MSE eleva ao quadrado cada erro (para evitar que erros negativos cancelem os positivos) e depois faz a média de todos

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2$$

Mean Error Squared

ERRO QUADRÁTICO MÉDIO

Erro quadrático médio (MSE)

- É a média do quadrado das distâncias verticais entre os pontos reais e os pontos previstos pela linha.
- Em vez de só calcular a diferença, o MSE eleva ao quadrado cada erro (para evitar que erros negativos cancelem os positivos) e depois faz a média de todos

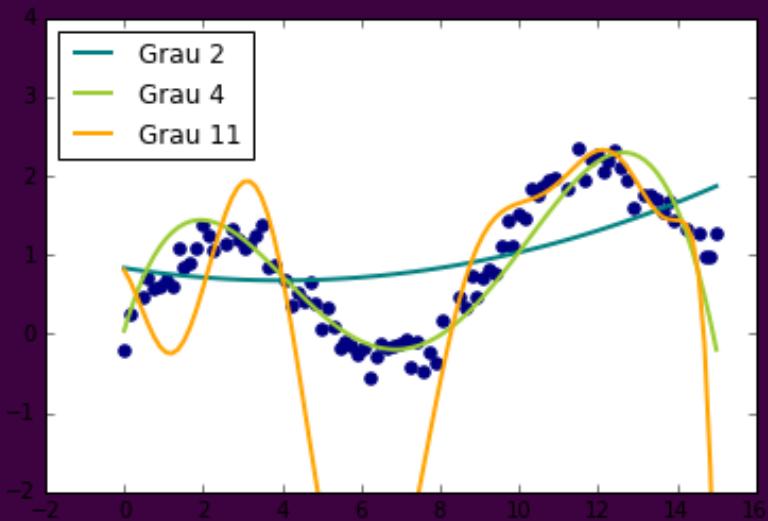
A distância vertical é o erro individual. O MSE é o resumo geral desses erros — ele diz quão boa (ou ruim) a linha está como um todo.



MAS E SE ...

A DISTRIBUIÇÃO NÃO FOR LINEAR?

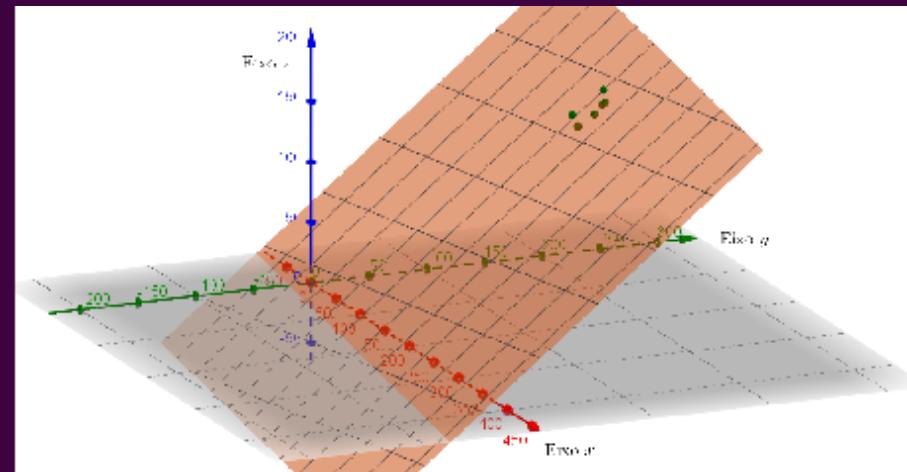
REGRESSÃO POLINOMIAL



$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon.$$

CONSIDERARMOS +1 CARACTERÍSTICA?

REGRESSÃO LINEAR MÚLTIPLA



$$y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n + \epsilon$$

DOCUMENTAÇÃO



[Documentação Regressão Linear](#)

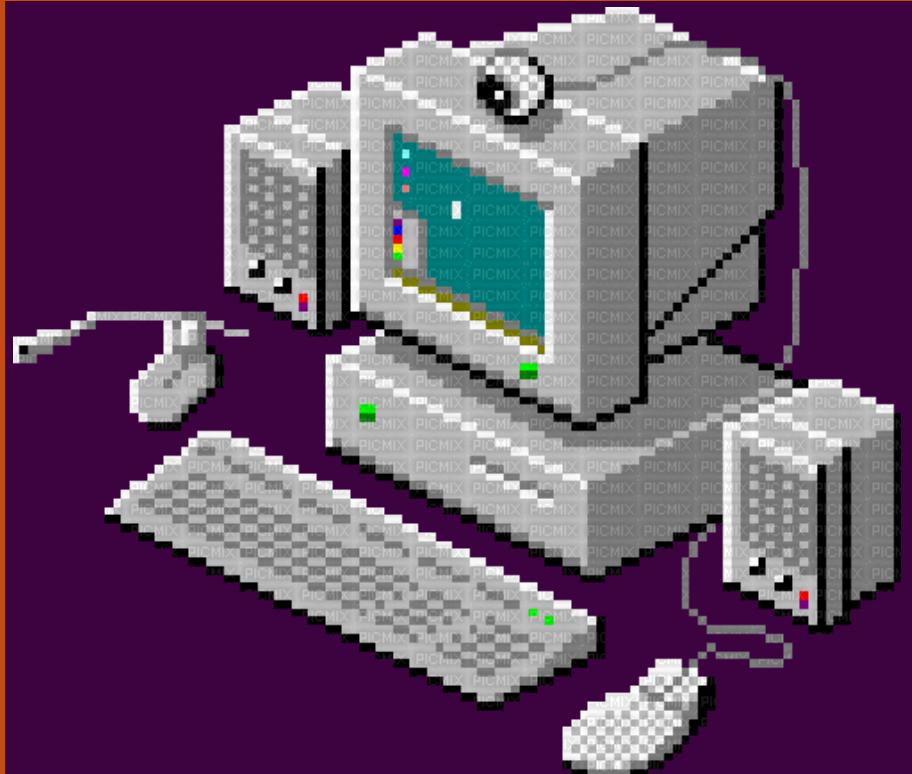


EDG



Regressão Logística

- Como adaptar a regressão para problemas de classificação?



REGRESSÃO LOGÍSTICA

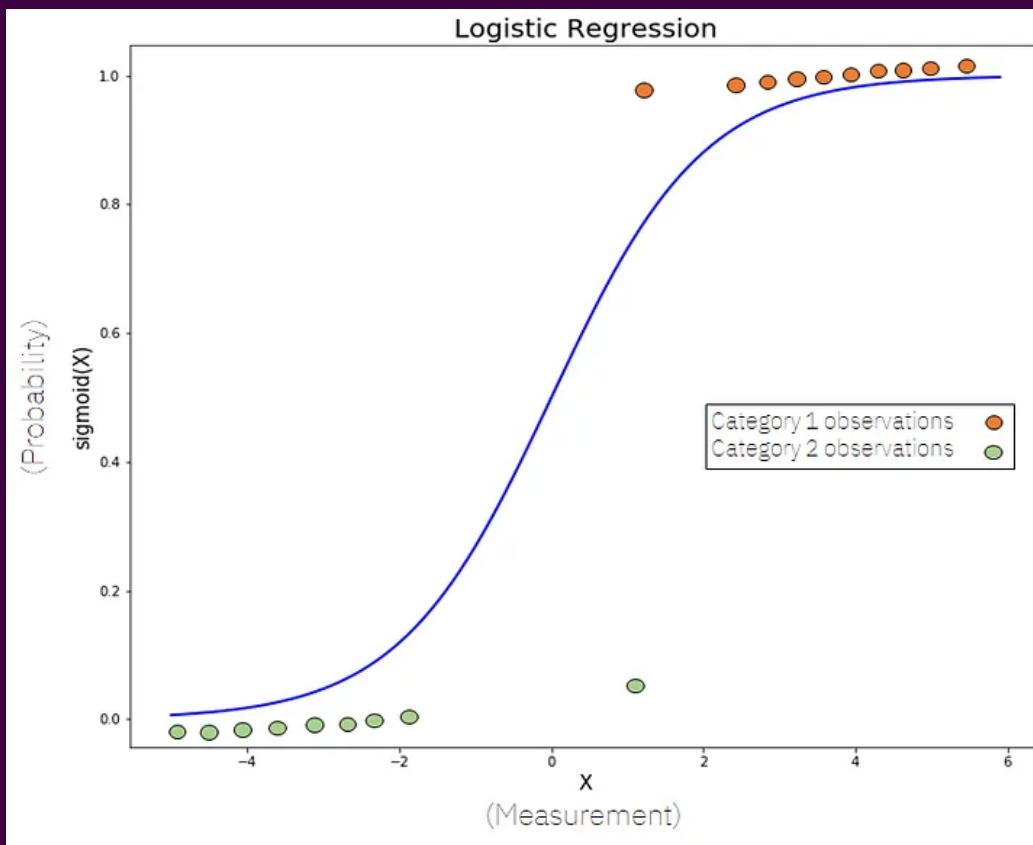


Imagen: [amalaj7.medium](#)

REGRESSÃO LOGÍSTICA

Em geral, a regressão logística como toda outra regressão, prevê probabilidades, mas com o objetivo principal de, na verdade, classificar observações de forma binária, como verdadeiro ou falso, sim ou não, 0 ou 1

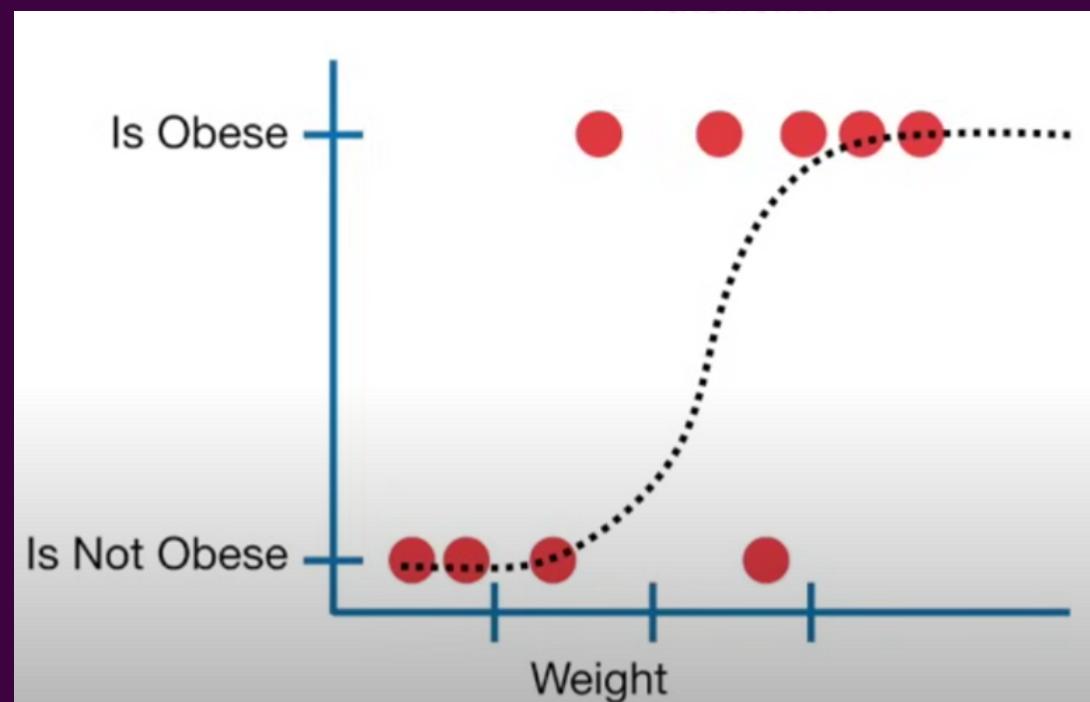


Imagen: StatQuest with Josh Tarmer

LIKELIHOOD

Mas como que ela faz esse cálculo?

O cálculo desse modelo em geral é feito pelo likelihood, onde cada dado é calculado por base na probabilidade dos dados, onde temos a sensibilidade e a rotulação dos dados para que a curva se move com esses parâmetros.

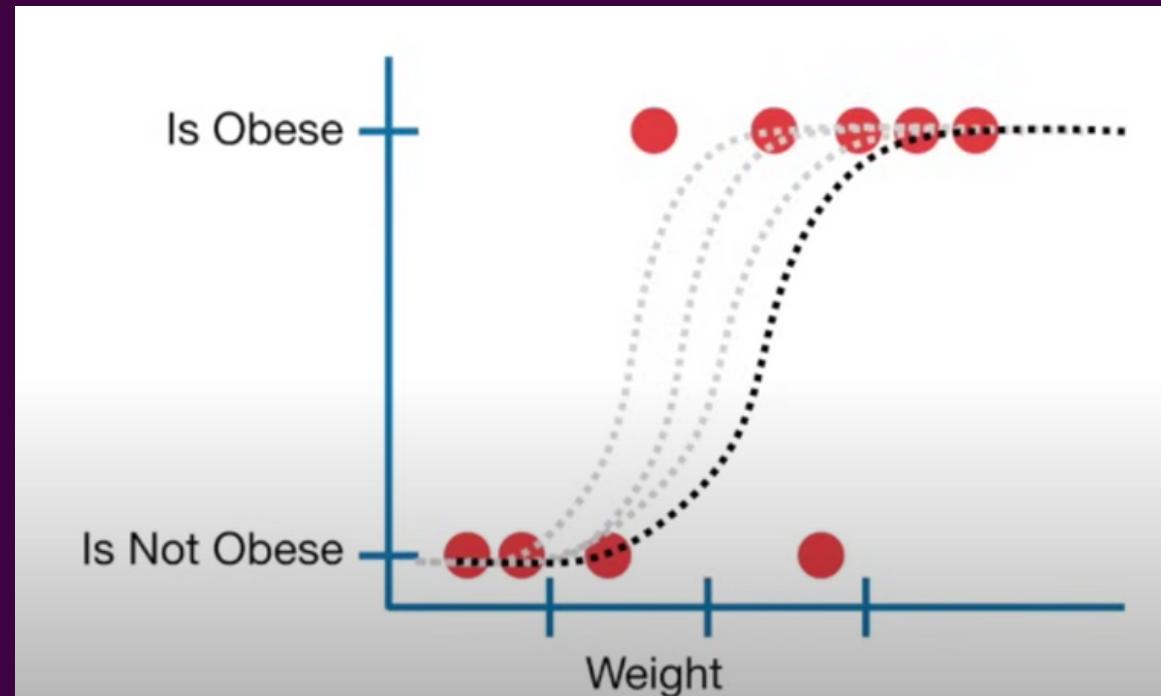
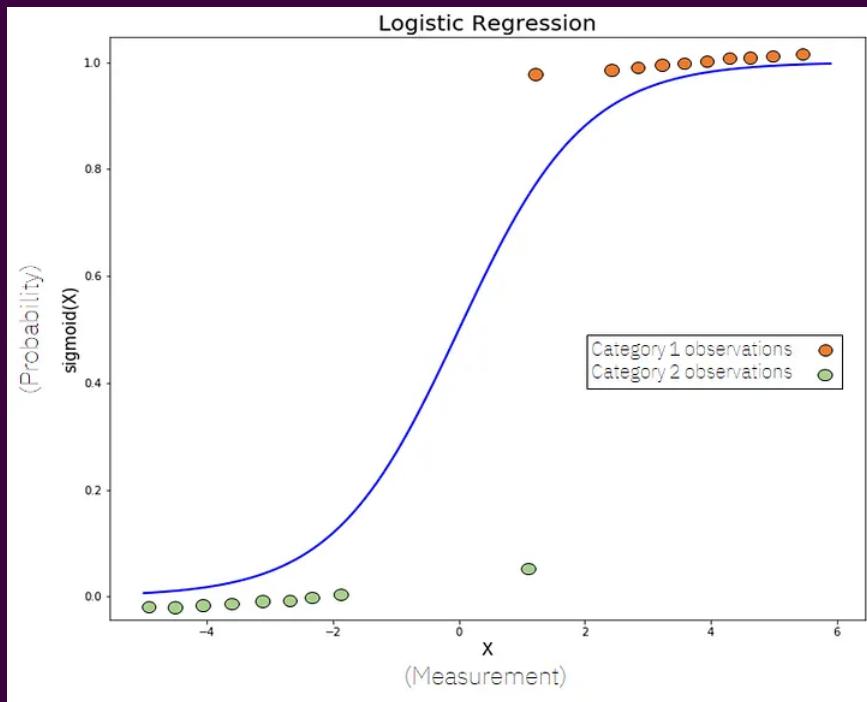


Imagen: StatQuest with Josh Tarmer

REGRESSÃO LOGÍSTICA



$$f(x) = \frac{1}{1 + e^{-x}}$$

Imagen: [amalaj7.medium](#)



ESG



Regularização

- Como evitar que nosso modelo se ajuste demais aos dados de treino?
- Regularização L1 e L2



REGULARIZAÇÃO EM REGRESSÕES

L2 Ridge

$$LossFunction = \frac{1}{N} \sum_{i=1}^N (\hat{Y} - Y)^2 + \lambda \sum_{i=1}^N \theta_i^2$$

Alguns coeficientes se
aproximam de zero

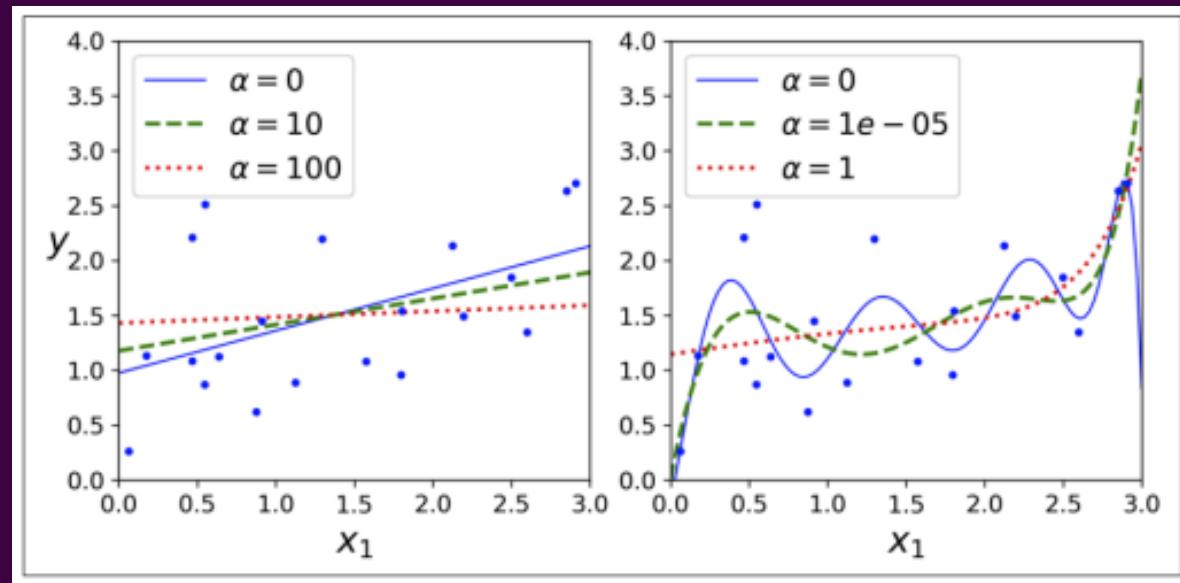


Imagen: [amalaj7.medium](#)

REGULARIZAÇÃO EM REGRESSÕES

L1 Lasso

$$LossFunction = \frac{1}{N} \sum_{i=1}^N (\hat{Y} - Y)^2 + \lambda \sum_{i=1}^N |\theta_i|$$

Alguns coeficientes se
aproximam de zero

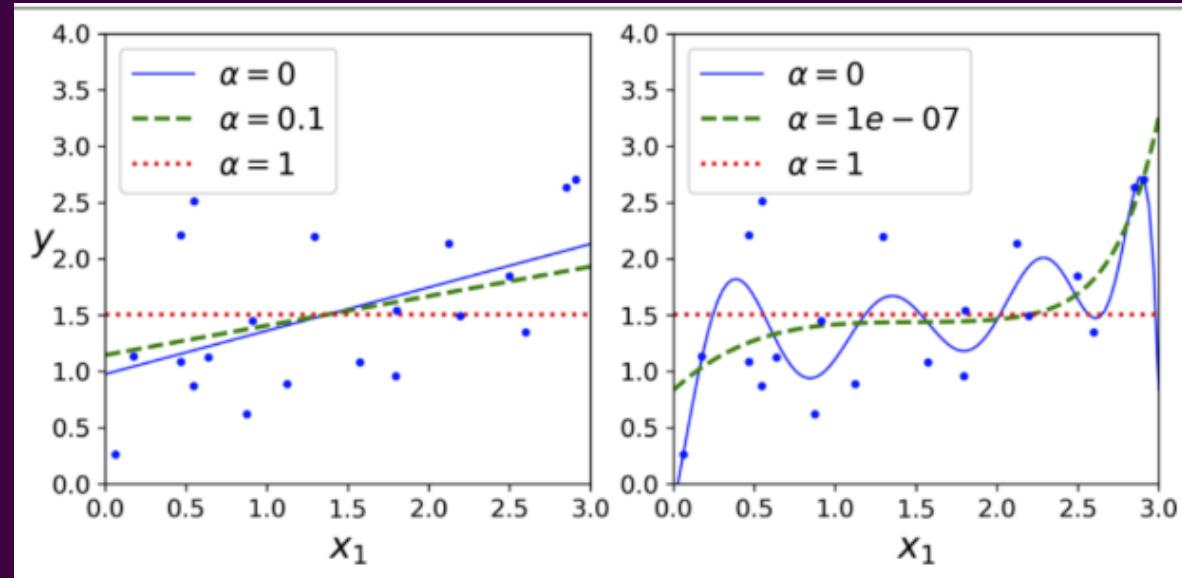


Imagen: [amalaj7.medium](#)