Aprendizado de Máquins

Regressão

Porque fazer regressão

- Suponha que você seja um gerente de vendas tentando prever os números do próximo mês.
- Você sabe que dezenas, talvez até centenas de fatores, desde o clima até a promoção de um concorrente e o boato de um modelo novo e aprimorado podem afetar o número.
- Talvez as pessoas da sua organização tenham até uma teoria sobre o que terá maior efeito nas vendas.

"Confie em mim. Quanto mais chuva temos, mais vendemos. ”

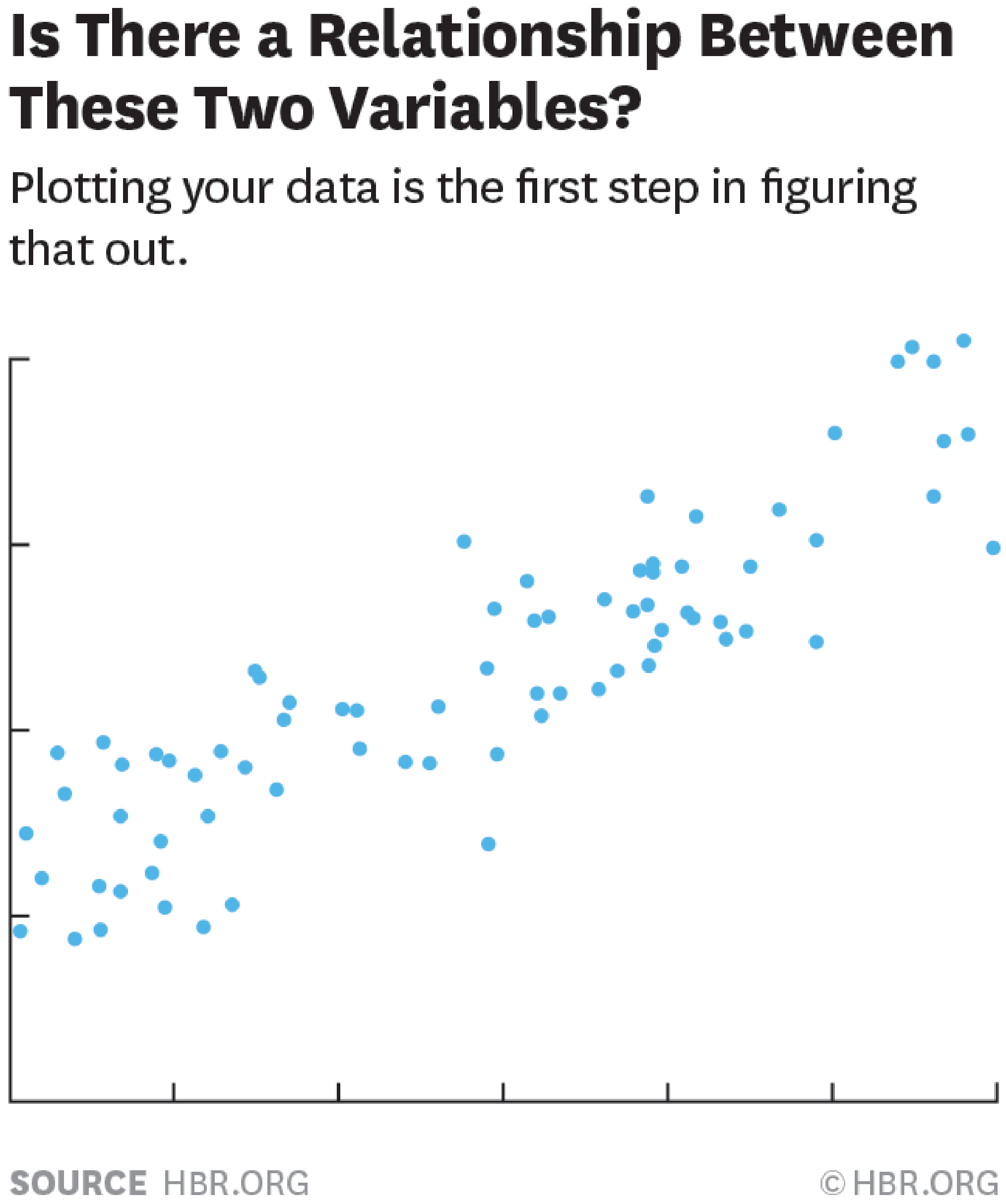
"Seis semanas após a promoção do concorrente, as vendas aumentam."

Porque fazer regressão

- A análise de regressão é uma maneira de classificar matematicamente quais dessas variáveis realmente têm impacto.
- Responde às perguntas: Quais fatores são mais importantes?
- O que podemos ignorar? Como esses fatores interagem entre si?
- E, talvez o mais importante, até que ponto estamos certos sobre todos esses fatores?
- Na análise de regressão, esses fatores são chamados de variáveis. Você tem sua variável dependente - o principal fator que você está tentando entender ou prever.

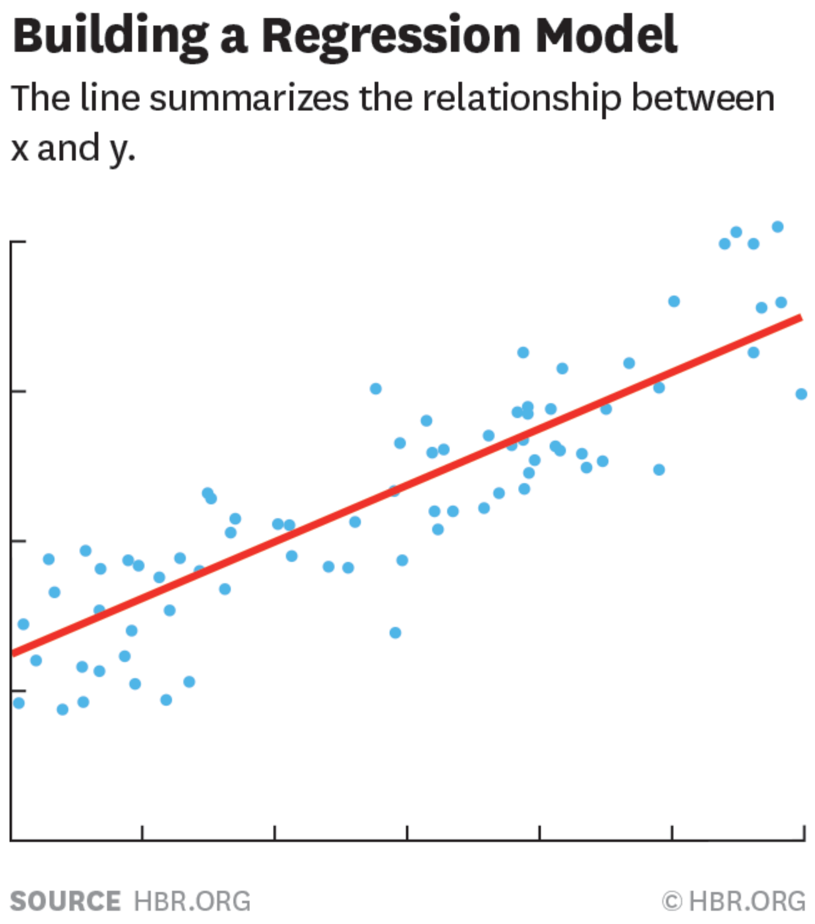
Como funciona?

- Para realizar uma análise de regressão, você reúne os dados nas variáveis em questão.
- Você usa todos os seus números de vendas mensais nos últimos três anos, por exemplo, nos últimos três anos e em qualquer dado sobre as variáveis independentes nas quais você está interessado. P
- ortanto, neste caso, digamos que você descubra também a precipitação média mensal nos últimos três anos.
- Em seguida, você plota todas essas informações em um gráfico semelhante a este:



Como funciona?

- O eixo y é a quantidade de vendas (a variável dependente, a coisa em que você está interessado, está sempre no eixo y) e o eixo x é a precipitação total.
- Cada ponto azul representa os dados de um mês - quanto choveu naquele mês e quantas vendas você fez no mesmo mês.
- Olhando para esses dados, você provavelmente percebe que as vendas são maiores nos dias em que chove muito. É interessante saber, mas em quanto?
- Se chover 3 polegadas, você sabe quanto vai vender? E se chover 10 cm?
- Agora imagine desenhar uma linha no gráfico acima, que percorre aproximadamente o meio de todos os pontos de dados.
- Essa linha o ajudará a responder, com certeza grau de certeza, quanto você normalmente vende quando chove uma certa quantidade.



A Tarefa ade Regressão

- Na resolução da tarefa de classificação de dados, o objetivo é predizer o rótulo para um exemplar qualquer que não pertence ao conjunto de dados de treinamento.
- Portanto, o uso de um modelo preditivo f promove a atribuição de um rótulo y a um exemplar $\vec{x}$ qualquer, ou seja, $y = f(\vec{x})$, sendo y uma variável do tipo categórico.
- Por outro lado, quando y é do tipo numérico (contínuo ou discreto), diz-se ter um problema de regressão ou predição numérica

O que são variáveis categóricas, discretas e contínuas?

- Variáveis quantitativas podem ser classificadas como discretas ou contínuas.
- **Variável categórica**
  - As variáveis categóricas contêm um número finito de categorias ou grupos distintos. Os dados categóricos podem não ter uma ordem lógica. Por exemplo, os preditores categóricos incluem gênero, tipo de material e método de pagamento.
- **Variável discreta**
  - Variáveis discretas são variáveis numéricas que têm um número contável de valores entre quaisquer dois valores. Uma variável discreta é sempre numérica. Por exemplo, o número de reclamações de clientes ou o número de falhas ou defeitos.
- **Variável contínua**
  - Variáveis contínuas são variáveis numéricas que têm um número infinito de valores entre dois valores quaisquer. Uma variável contínua pode ser numérica ou de data/hora. Por exemplo, o comprimento de uma peça ou a data e hora em que um pagamento é recebido.

Tipos de Regressão

- Basicamente, os modelos de regressão podem ser divididos em:
  - linear simples ou multivariado, ou
  - não linear simples ou multivariado.
- A diferença entre regressão do tipo linear e regressão do tipo não linear está na função f a ser utilizada por exemplo:
  - Uma função que representa a equação da reta ou do plano se aplica para regressão linear,
  - Uma função que representa uma equação exponencial se aplica para regressão não linear.
- O tipo simples ou multivariado é definido a partir da quantidade de atributos descritivos utilizados para estimar o valor de y:
  - um único atributo é utilizado no tipo simples, e
  - mais de um atributo é usado no tipo multivariado.

Aplicações da Regressão

- Em se tratando de aplicação, a regressão é usada para estimar valores a partir de um conjunto de dados históricos.
- Isto é o que acontece, por exemplo, em problemas de indicadores econômicos ou de mercado futuro, nos quais se tenta prever o próximo valor analisando os dados de algumas variáveis (atributos descritivos) historicamente armazenadas em um conjunto de dados.
- Para o caso do restaurante, que vimos na ultima aula, exemplos poderiam ser a estimação da quantidade de bebidas que deve ser estocada para determinado período ou o número de clientes que provavelmente comparecerá ao restaurante em um dia especial.

Que tipo de regressão utilizar

- Para decidir entre usar uma regressão linear ou não linear, geralmente se faz uma análise inicial dos dados, de forma a verificar o tipo de distribuição que os atributos assumem.
- Usar recursos de visualização de dados, a exemplo de um gráfico de dispersão, pode ser de grande auxílio.
- No caso da regressão não linear, ainda é preciso verificar qual seria a melhor função de ajuste a ser usada, como polinomial, potência, logarítmica etc.
- A solução para a tarefa de regressão pode ser obtida a partir de métodos estatísticos baseados em premissas e condições relacionadas com o tipo de distribuição dos dados, ou de técnicas de aprendizado indutivo, que não necessitam de informação prévia sobre o tipo de distribuição dos dados.

Regressão Linear

A regressão linear consiste em uma análise estatística que envolve duas variáveis:

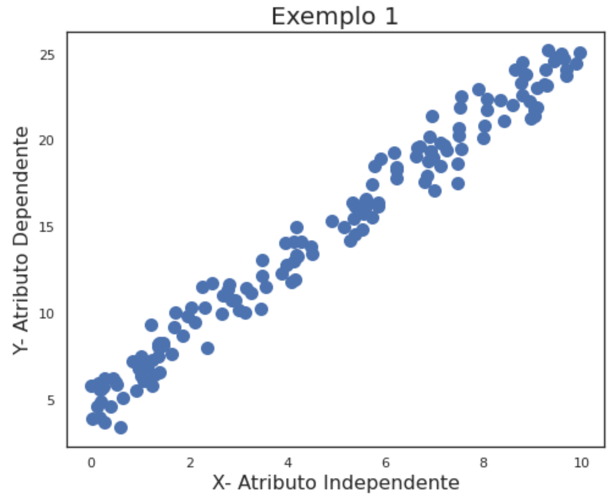
- a de resposta, explicada, dependente ou, como definido aqui, rótulo de um exemplar (y);
- e a preditora, explicativa, independente ou, como aqui definido, conjunto de atributos descritivos ($\vec{x}$).

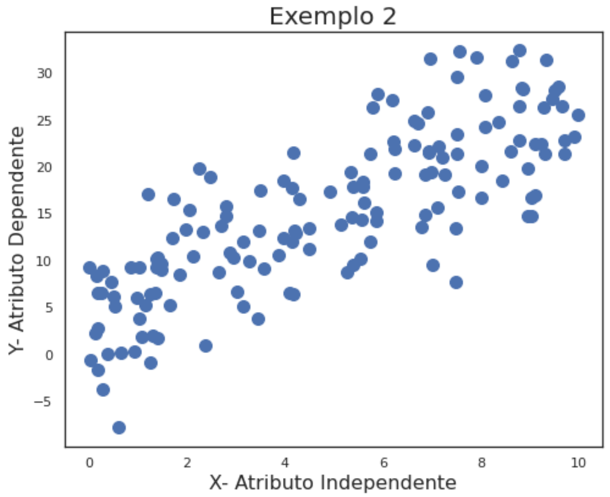
Um modelo de regressão linear considera que o valor da variável de resposta (ou dependente) y pode ser estimado por uma combinação linear das variáveis explicativas (ou independentes) $\vec{x}$.

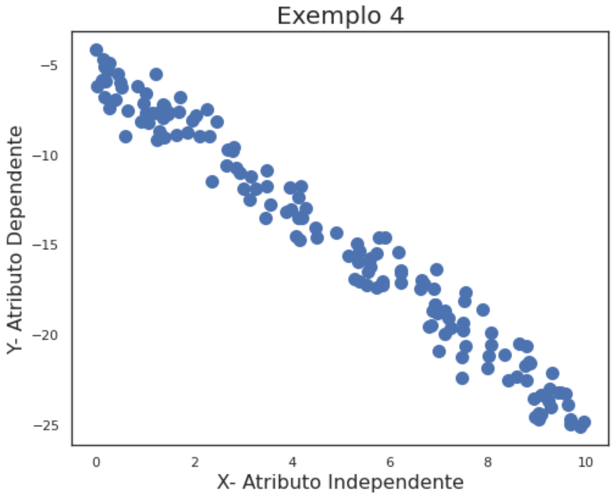
Modelo de regressão linear simples

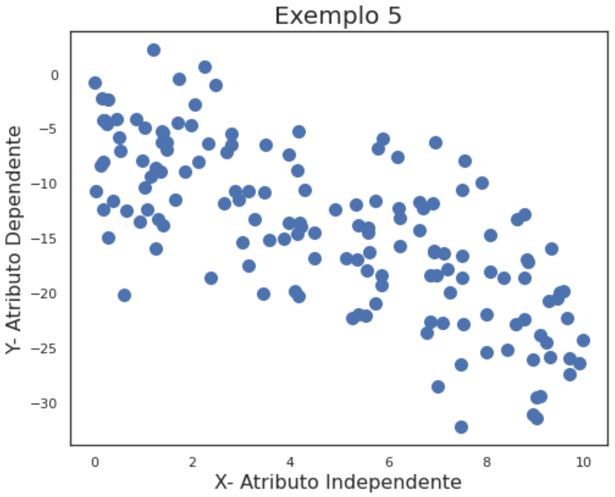
- o modelo de regressão para uma única variável preditora x, ou regressão linear simples, pode ser definido pela seguinte equação da reta:
  - $y = a + bx$
- em que a e b são coeficientes de regressão e especificam o intercepto do eixo y e a inclinação da reta, respectivamente.
- Os coeficientes de regressão podem também ser entendidos como pesos.

Situações onde podemos usar a regressão linear simples



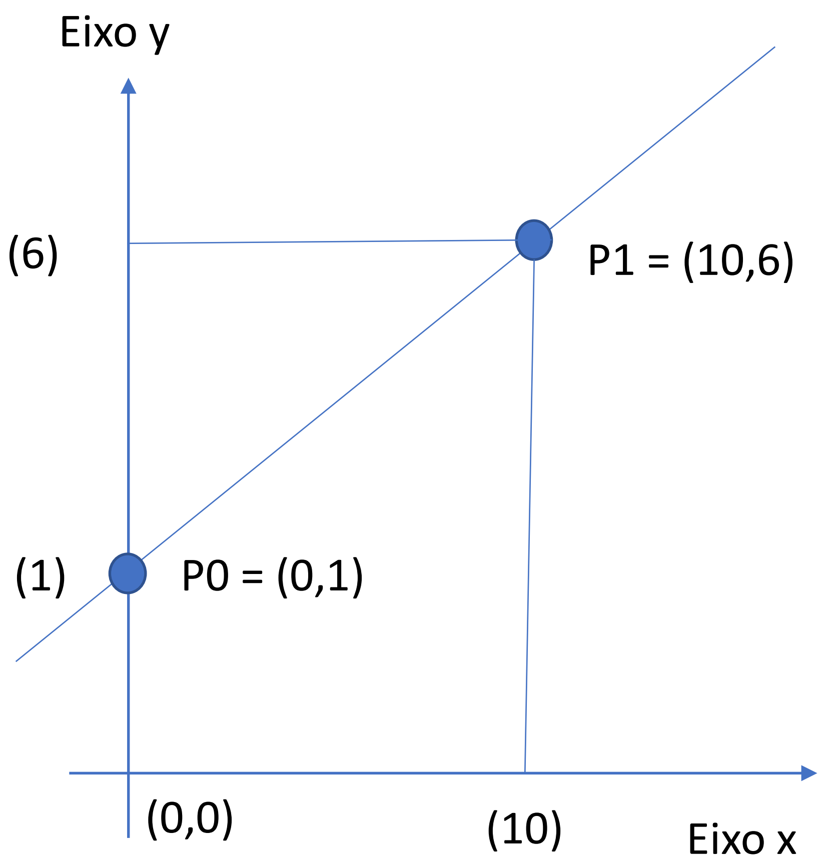






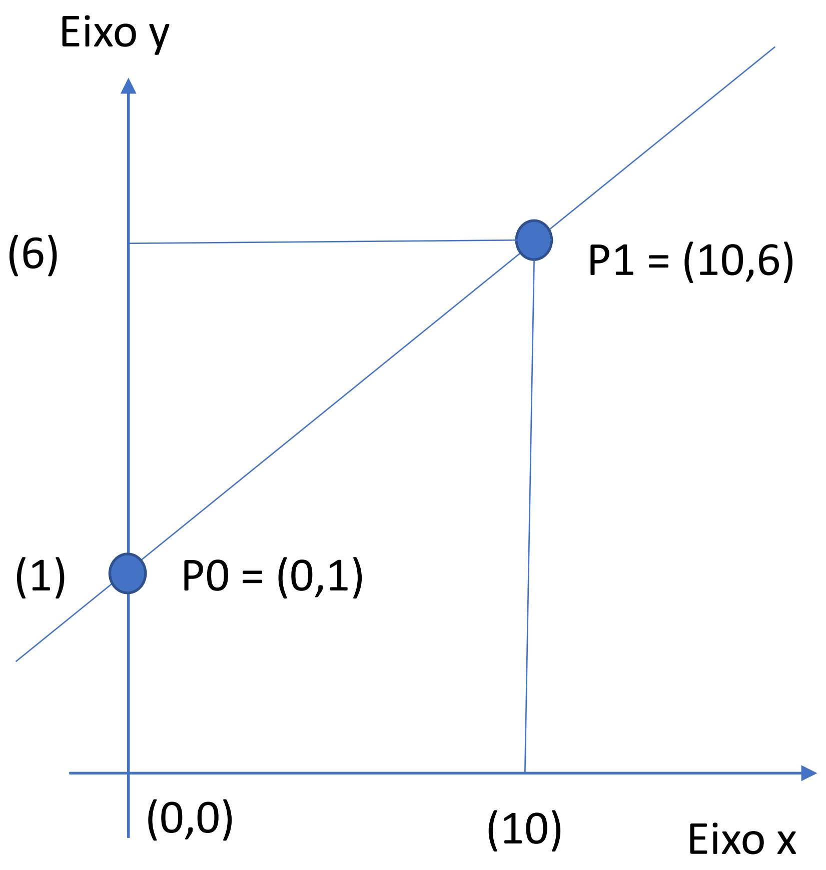
Equação reduzida da reta

- Toda equação na forma:
- $y = ax + b$
- é chamada equação reduzida da reta,
- Em que a é o coeficiente angular e b a ordenada do ponto n qual a reta cruza o eixo y.
- b também é chamdo de intercepto
- A equação reduzida pode ser obtida diretamente da equação geral $ax + by + c = 0$



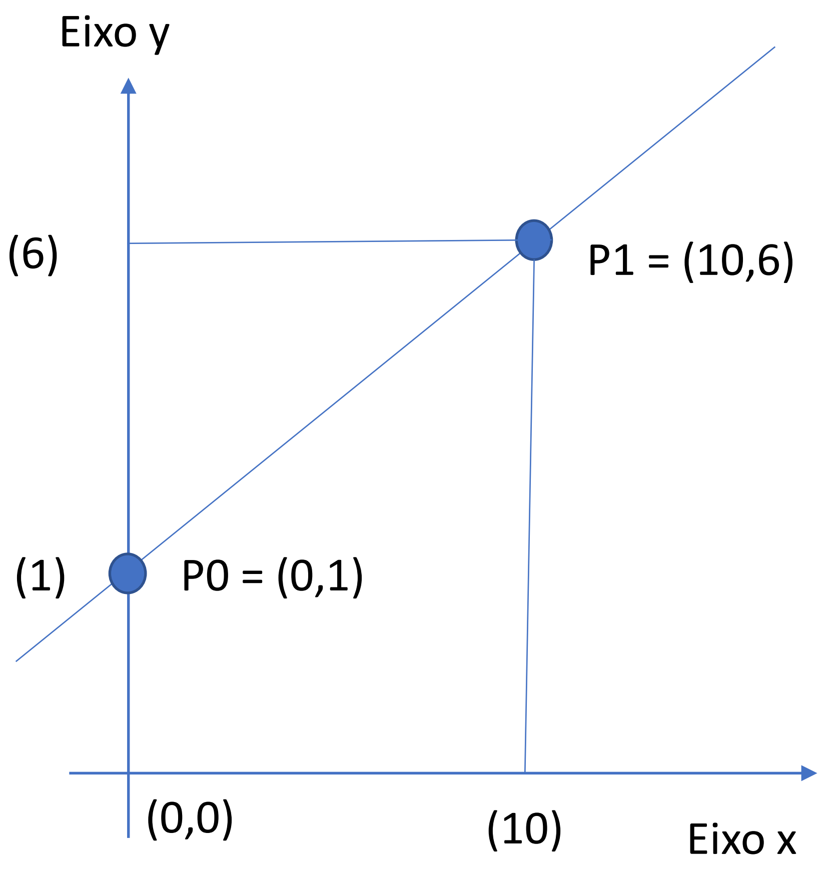
Equação reduzida da reta – coeficiente angular

- Dada equação na forma:
- $y = ax + b$
- O coefinciente angular pode ser calculado por:
- $a = (yp1-yp0)/(xp1-xp0)$
- $yp1 = 6$
- $yp0 = 1$
- $xp1 = 10$
- $xp0 = 0$
- Portanto:
- $a = (6-1)/(10-0)$
- $a = 5/10 = ½ = 0.5$



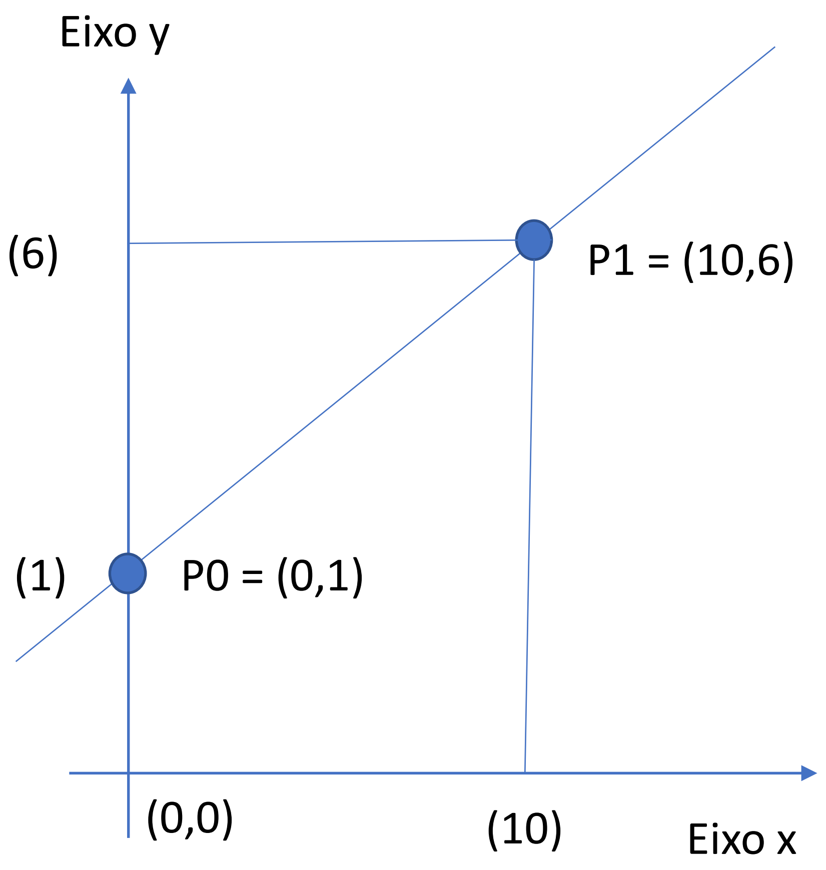
Equação reduzida da reta – intercepto

- Dada equação na forma:
- $y = ax + b$
- O intercepto é a ordenada do ponto n qual a reta cruza o eixo y
- A reta cruza o eixo y no ponto P0 onde a ordenada é y
- Portanto b= 1



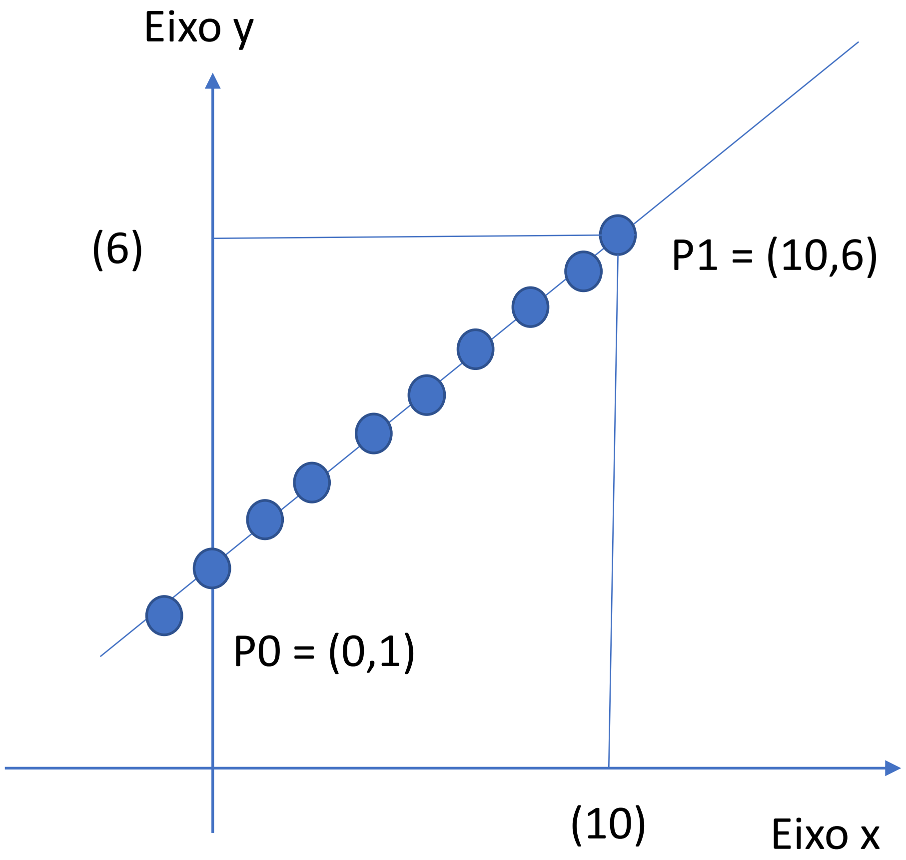
Equação reduzida da reta – Calculo dos pontos pertencente a equação a reta

- Dada equação na forma:
- y = ax + b
- O coeficiente angular é 0.5
- O intercepto é 1
- A equação da reta é
- y = 0.5*x + 1
- Por exemplo para x =0 y =1
- Para x = 10 y = 6
- y = 0.5*10 +1 = 5+1 = 6

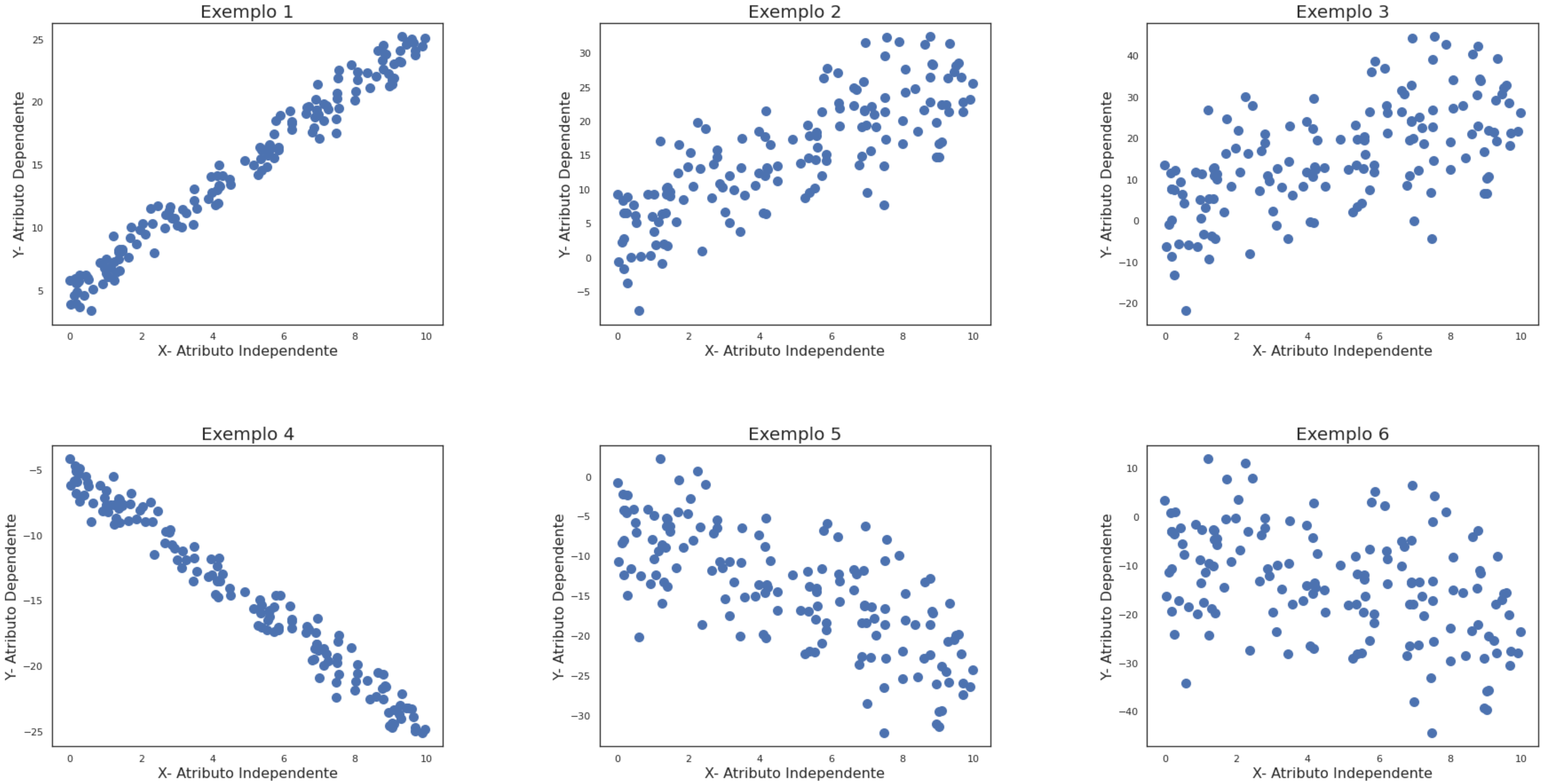


Assim com a equação pode-se calcular os pontos da reta

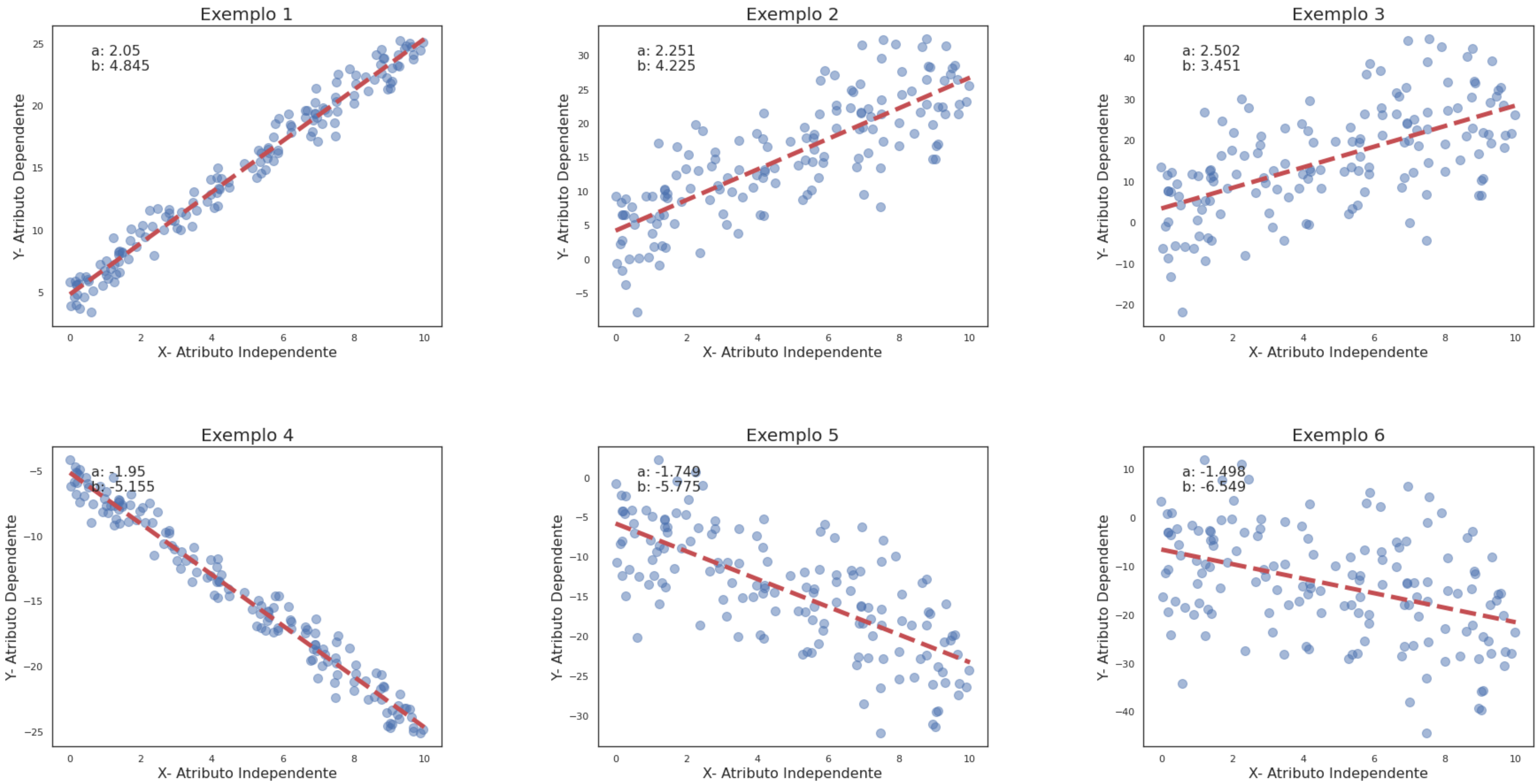
X	Y= 0.5*X+1	
-1	Y = 0.5*(-1)+1	Y=0.5
0	Y = 0.5*(0)+1	Y=1.0
1	Y = 0.5*(1)+1	Y=1.5
2	Y = 0.5*(2)+1	Y=2.0
3	Y = 0.5*(3)+1	Y=2.5
4	Y = 0.5*(4)+1	Y=3.0
5	Y = 0.5*(5)+1	Y=3.5
6	Y = 0.5*(6)+1	Y=4.0
7	Y = 0.5*(7)+1	Y=4.5
8	Y = 0.5*(8)+1	Y=5.0
9	Y = 0.5*(9)+1	Y=5.5
10	Y = 0.5*(10)+1	Y=6.0



Situações onde podemos usar a regressão linear simples



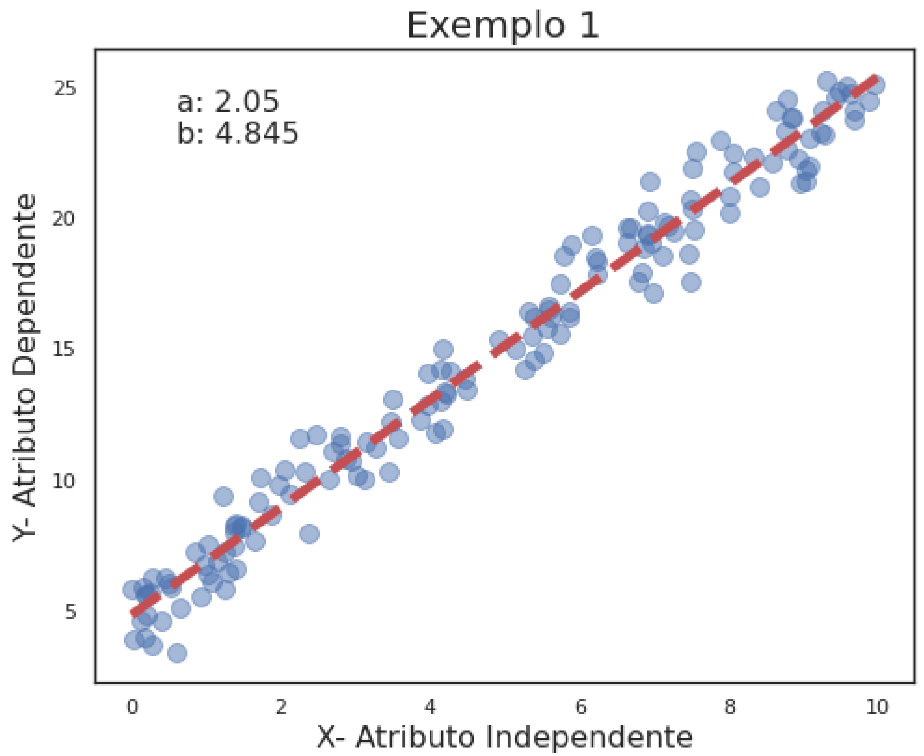
Após a regressão obtemos equações de retas para cada caso



Exemplo 1

- Equação:
- Y = 2.05*x + 4.845
- Por exemplo se x

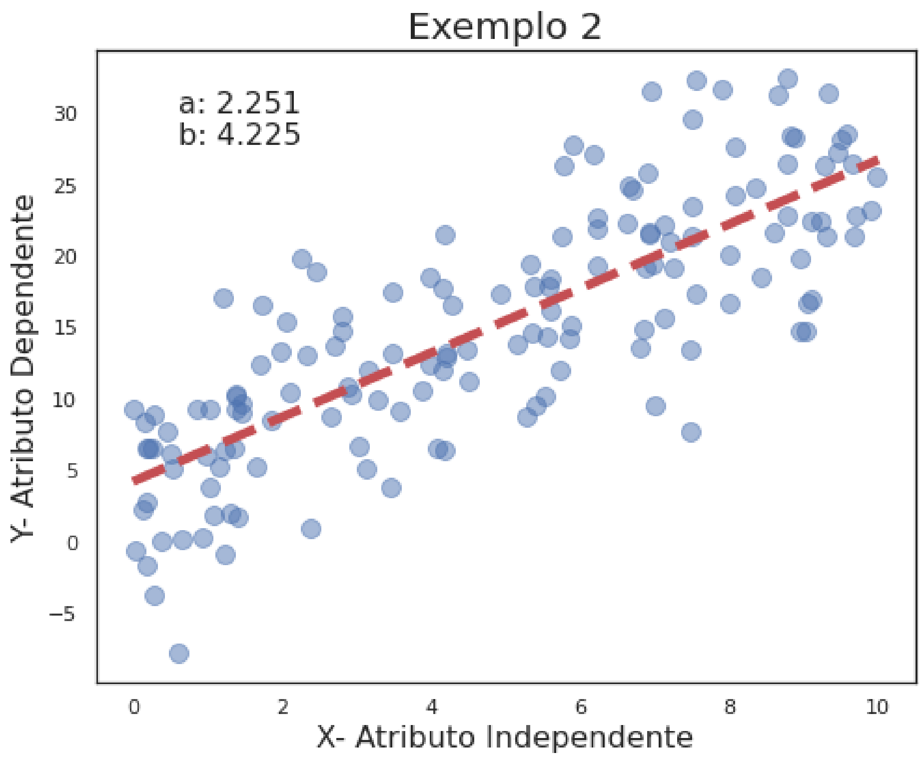
X	Y
2	9
5	15
10	25



Exemplo 2

- Equação:
- Y = 2.25*x + 4.225
- Por exemplo se x

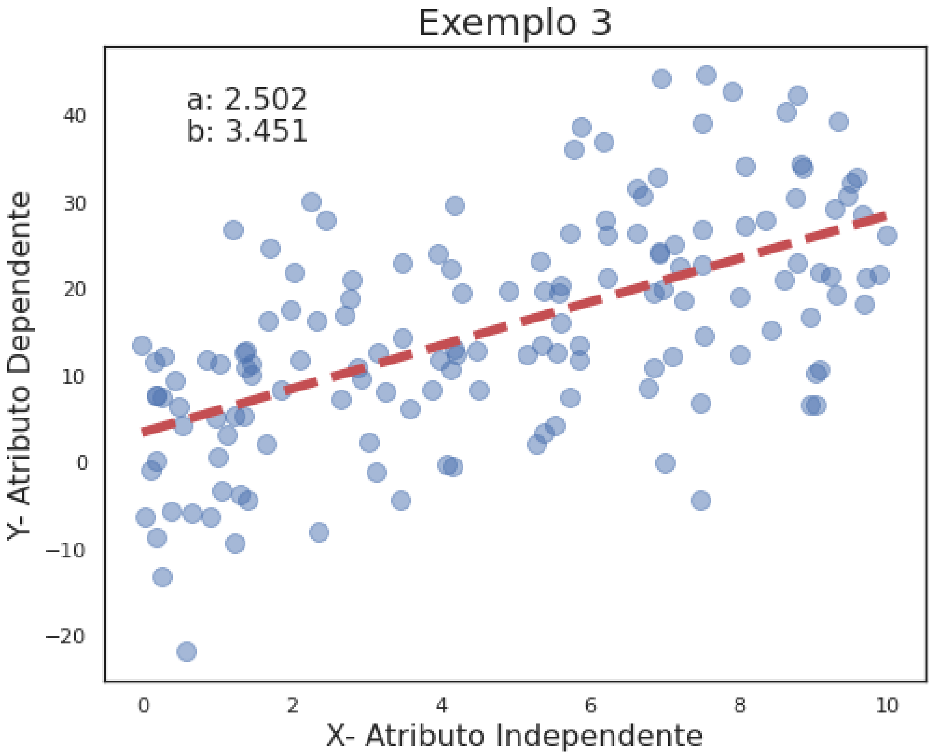
X	Y
2	8.725
5	15.475
10	26.725



Exemplo 3

- Equação:
- Y = 2.502*x + 3,451
- Por exemplo se x

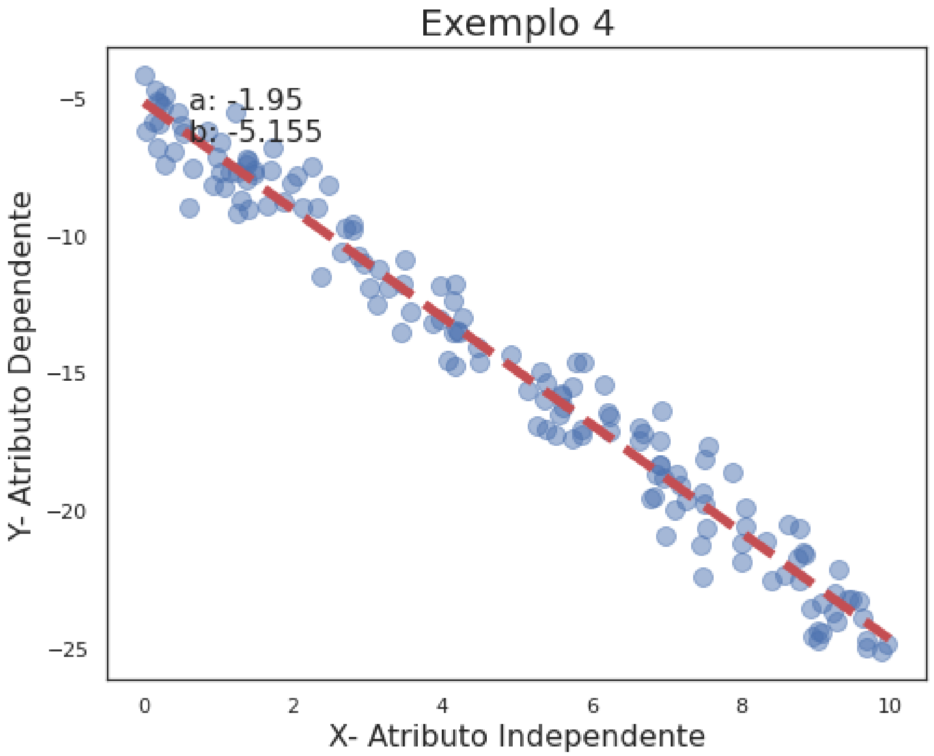
X	Y
2	8.425
5	15.931
10	28.441



Exemplo 4

- Equação:
- Y = -1.95*x - 5.155
- Por exemplo se x

X	Y
2	-9.055
5	-14.905
10	-24.655

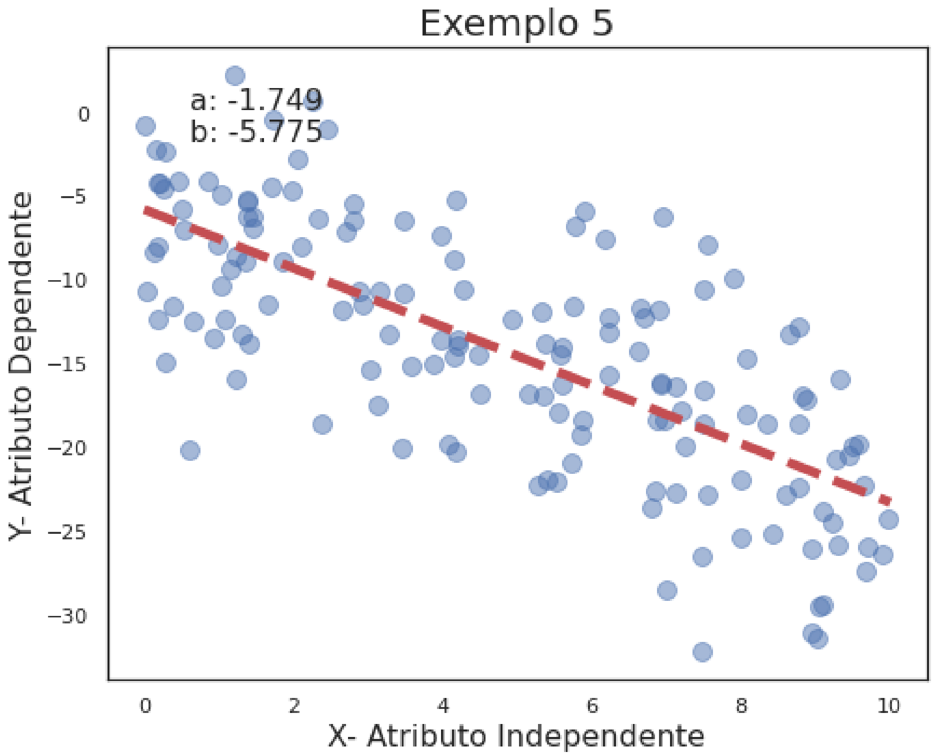


Exemplo 6

- Equação:
- Y = -1.75*x + -5.775
- Por exemplo se x

X
2
5
10

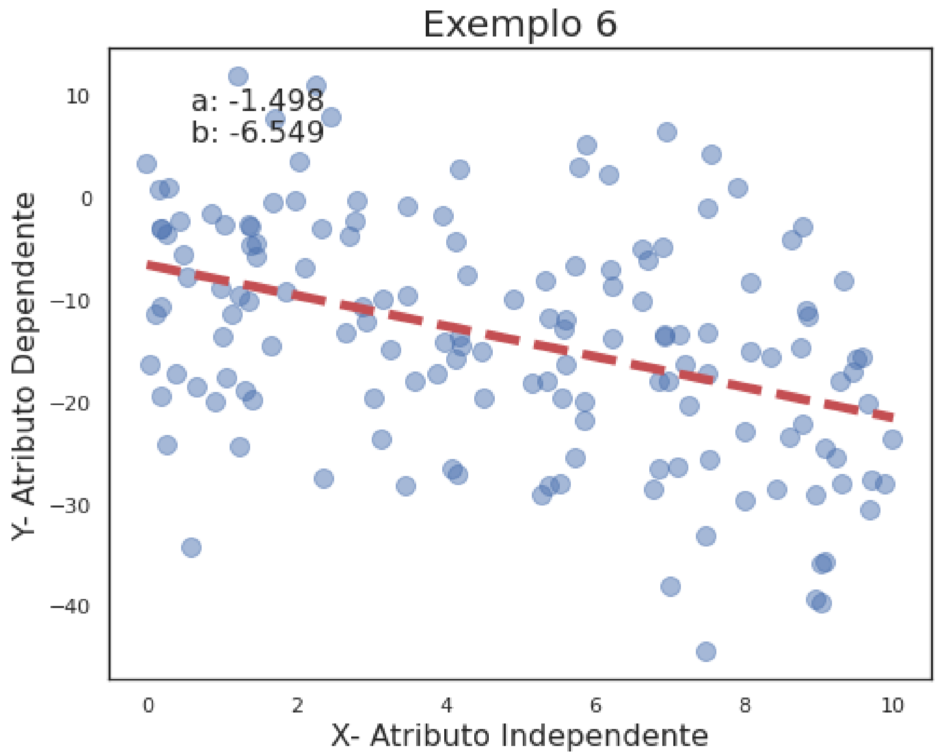
Y
-9.275
-14.525
-23.275



Exemplo 6

- Equação:
- Y = -1.5*x -6.549
- Por exemplo se x

X	Y
2	-9.529
5	-13.999
10	-21.449



Comparando os resultados

X	Y	Y	Y
2	9	8.725	8.425
5	15	15.475	15.931
10	25	26.725	28.441

X	Y	Y	Y
2	-9.055	-9.275	-9.529
5	-14.905	-14.525	-13.999
10	-24.655	-23.275	-21.449

Qual o Melhor Modelo ?

Métricas de avaliação R^2

- O coeficiente de determinação, geralmente indicado como R^2.
- Representa a proporção de variância (de y) que foi **explicada** pelas variáveis independentes no modelo.
- Ele fornece uma indicação da qualidade do ajuste e, portanto, uma medida de **quão bem as amostras não vistas provavelmente serão previstas pelo modelo**, por meio da proporção da variação explicada.
- Como essa variação depende do conjunto de dados, R^2 pode não ser significativamente comparável entre diferentes conjuntos de dados.
- A melhor pontuação possível é 1,0 e pode ser negativa (porque o modelo pode ser arbitrariamente pior). R^2

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

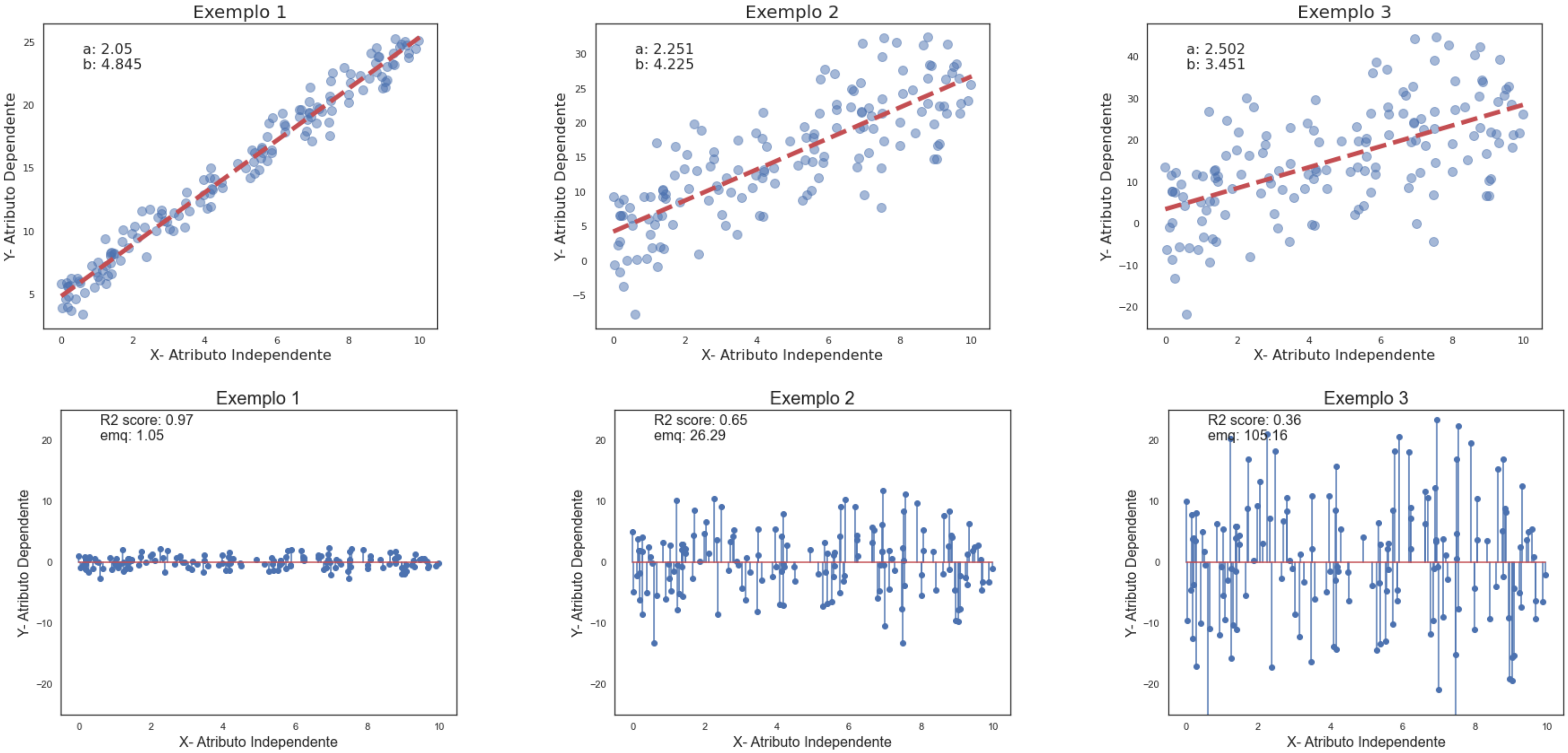
where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \epsilon_i^2$.

Métricas de avaliação Erro Médio Quadrático

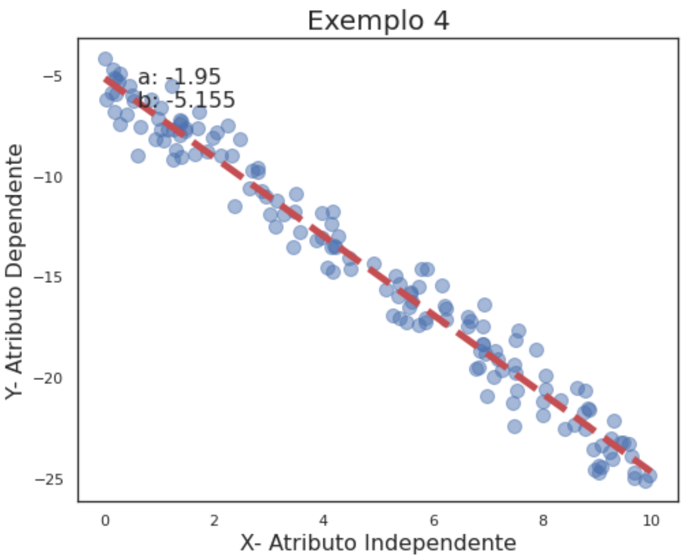
- O erro médio quadrático, uma métrica correspondente ao valor esperado do erro ou perda quadrática (quadrática).

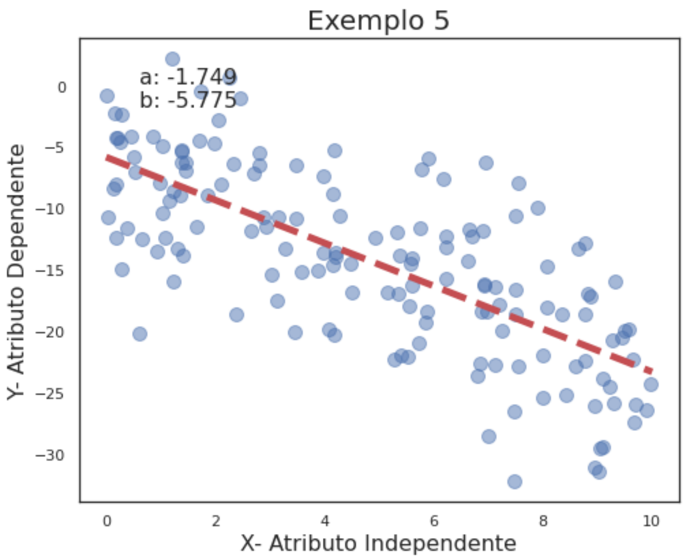
$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2.$$

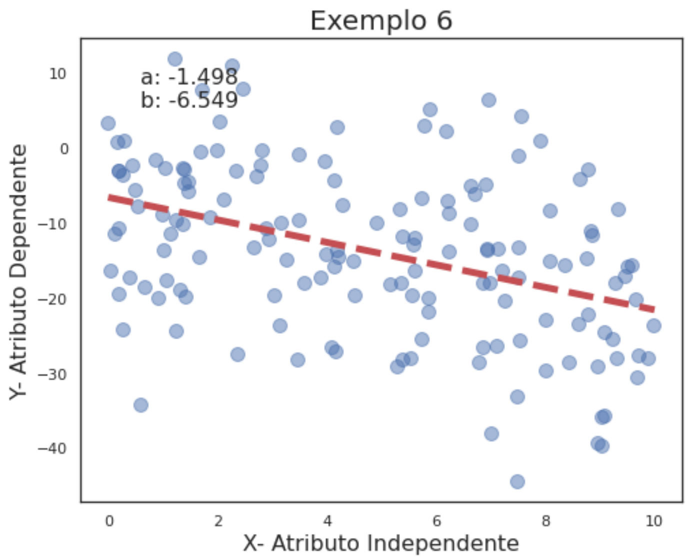
As Métricas nos Exemplos

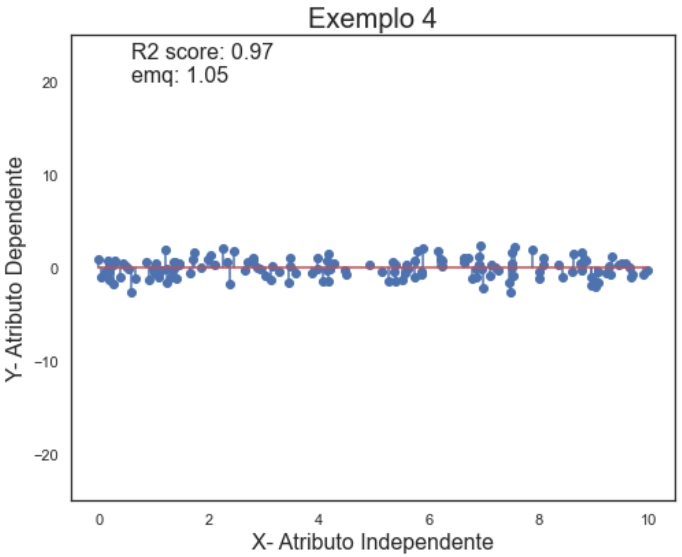


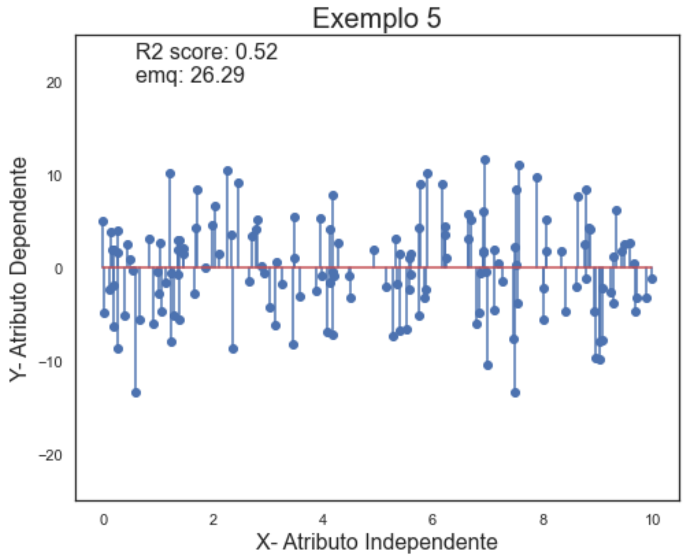
As Métricas nos Exemplos

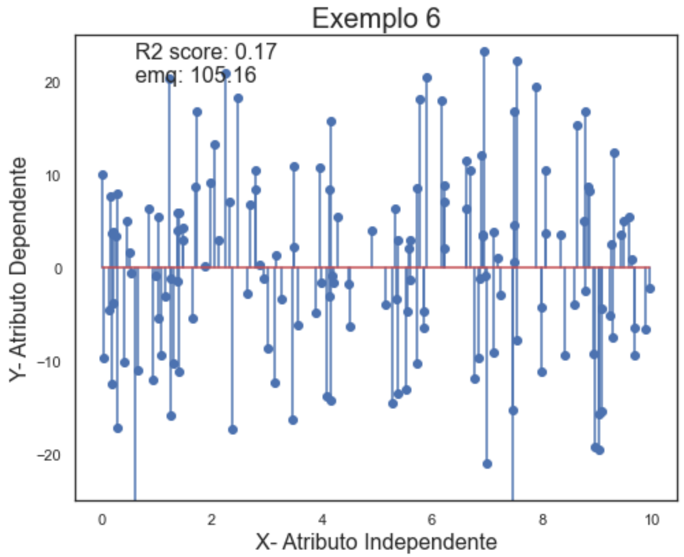










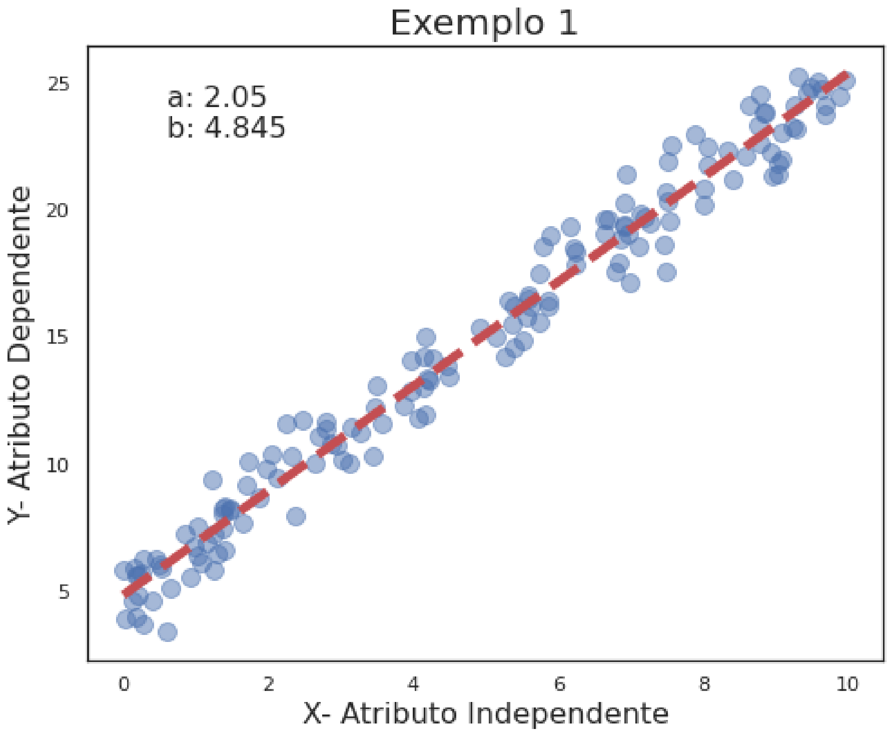


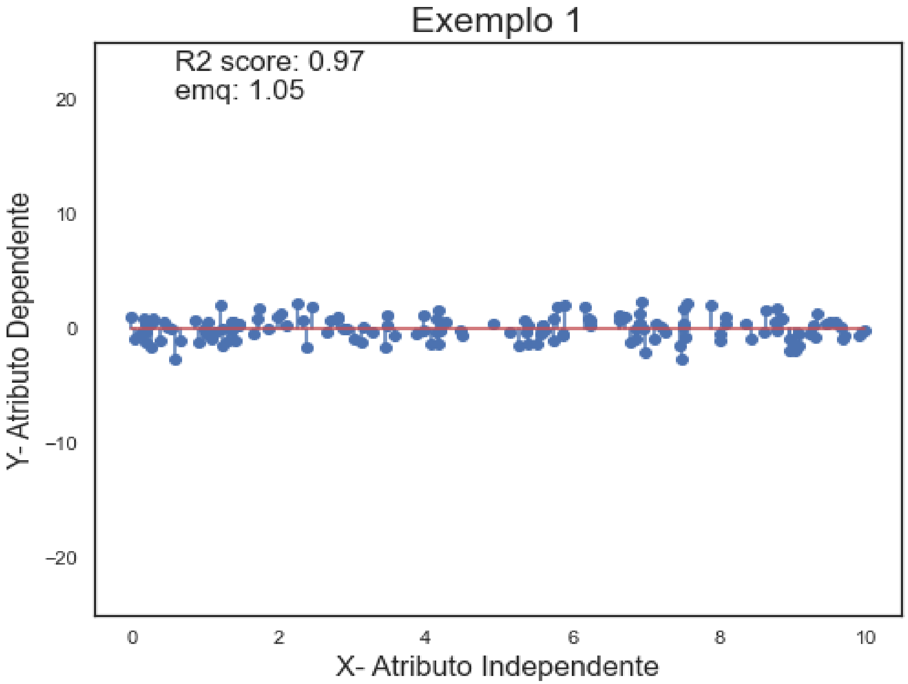
Exemplo 1

- Equação:
- Y = 2.05*x + 4.845
- Por exemplo se x

X	Y
2	9
5	15
10	25

$R^2 = 0.97$
EMQ = 1.05



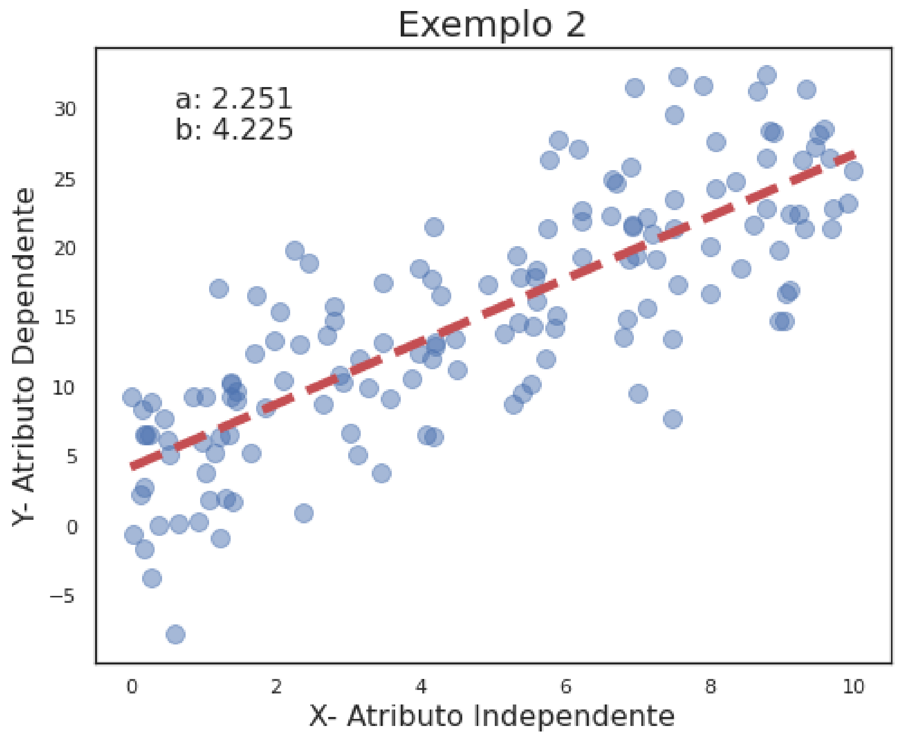


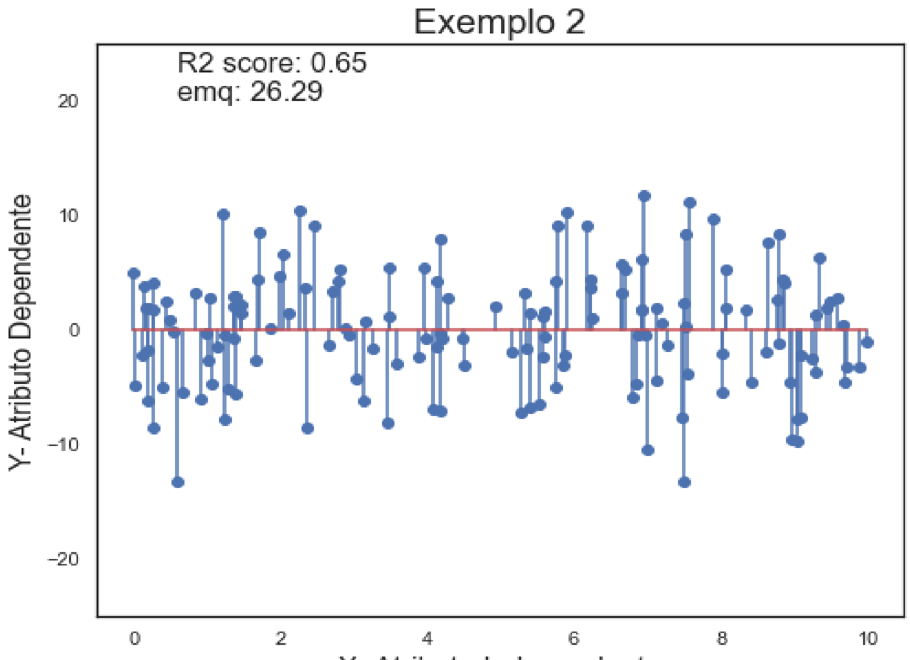
Exemplo 2

- Equação:
- Y = 2.25*x + 4.225
- Por exemplo se x

X	Y
2	8.725
5	15.475
10	26.725

$R^2 = 0.65$
EMQ = 26.29



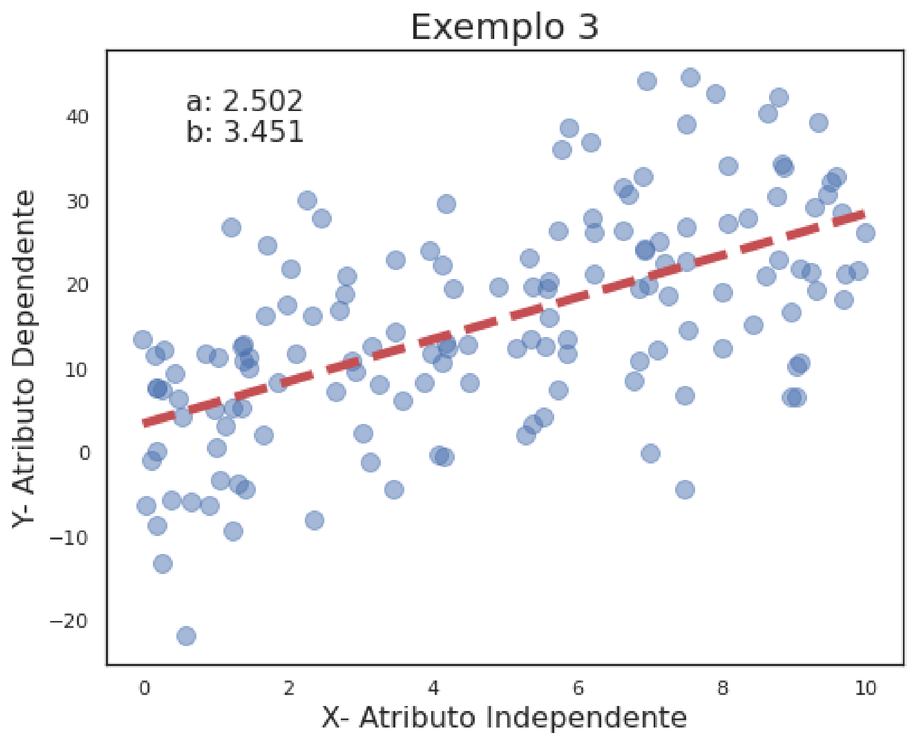


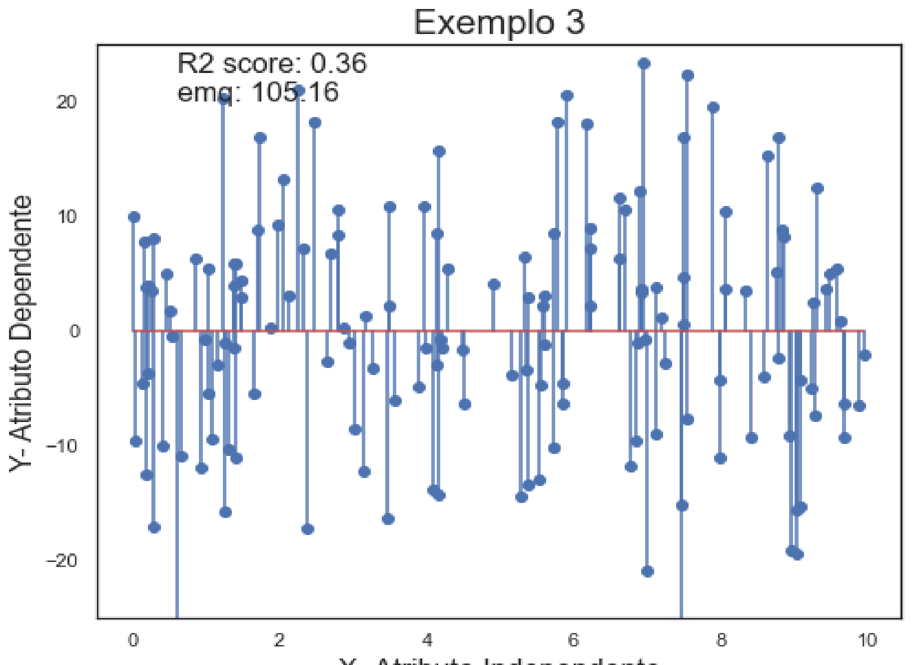
Exemplo 3

- Equação:
- Y = 2.502*x + 3,451
- Por exemplo se x

X	Y
2	8.425
5	15.931
10	28.441

$R^2 = 0.36$
EMQ = 105



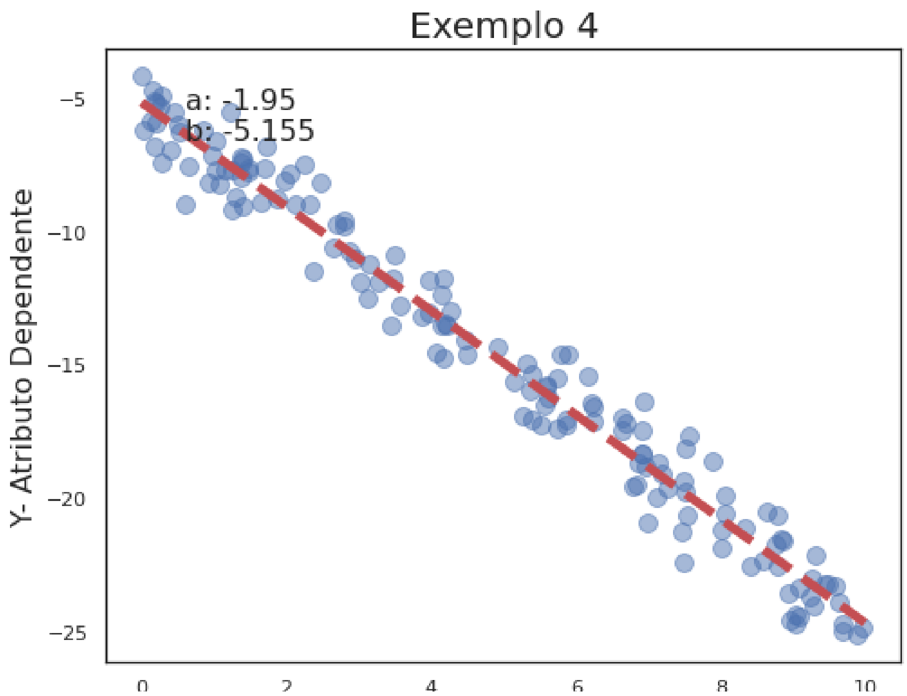


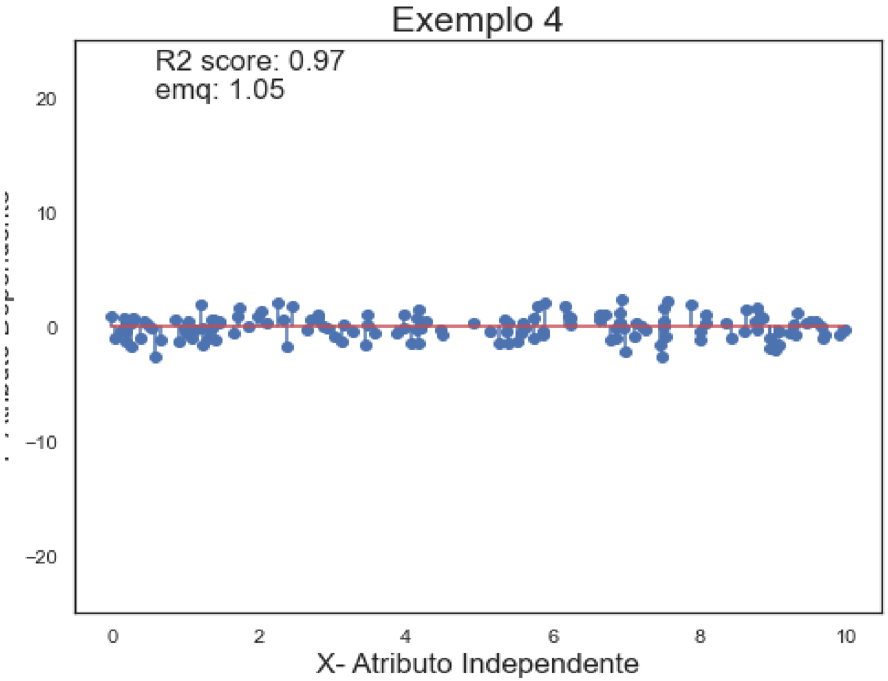
Exemplo 4

- Equação:
- Y = -1.95*x - 5.155
- Por exemplo se x

X	Y
2	-9.055
5	-14.905
10	-24.655

R^2 = 0.97
EMQ = 1.05



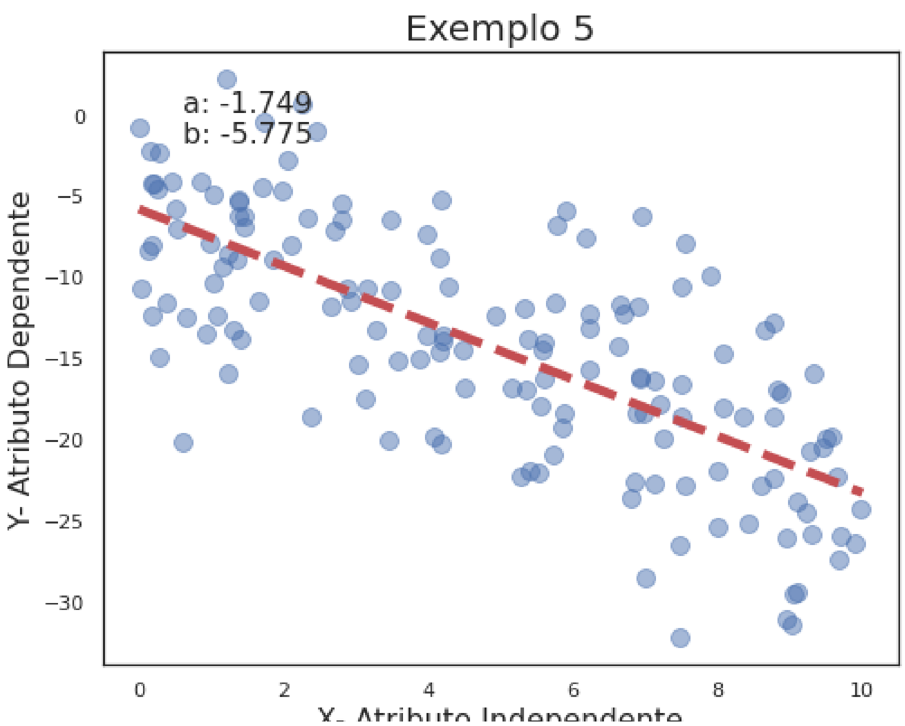


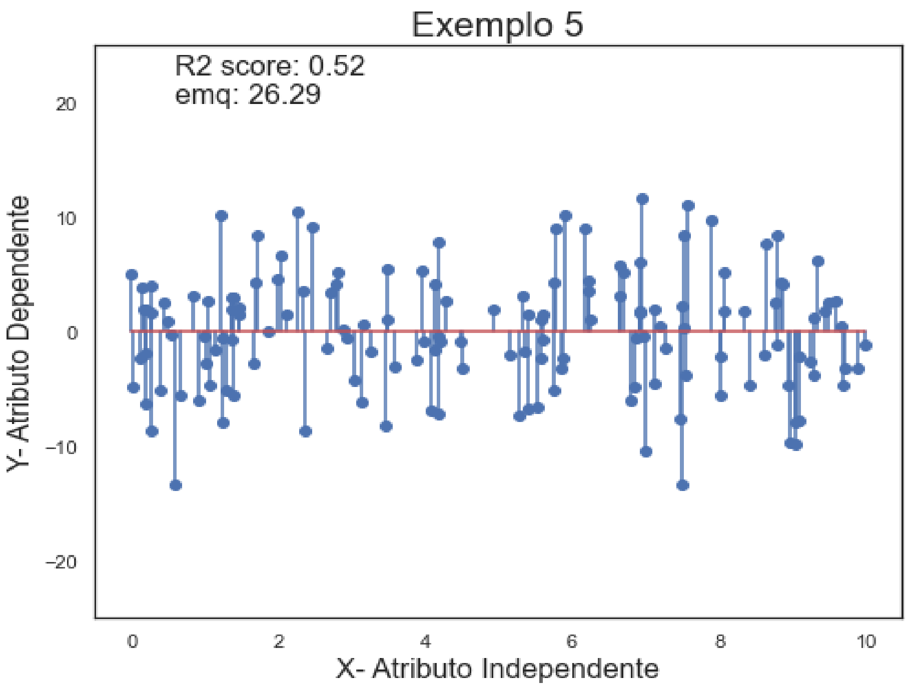
Exemplo 5

- Equação:
- Y = -1.75*x + -5.775
- Por exemplo se x

X	Y
2	-9.275
5	-14.525
10	-23.275

$R^2 = 0.52$
EMQ = 26.29



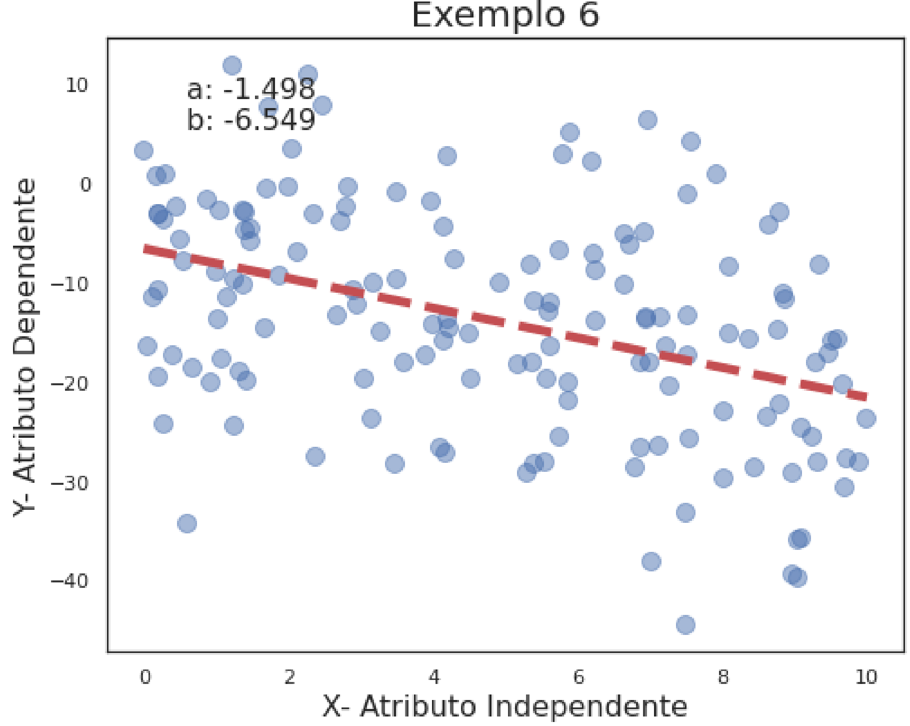


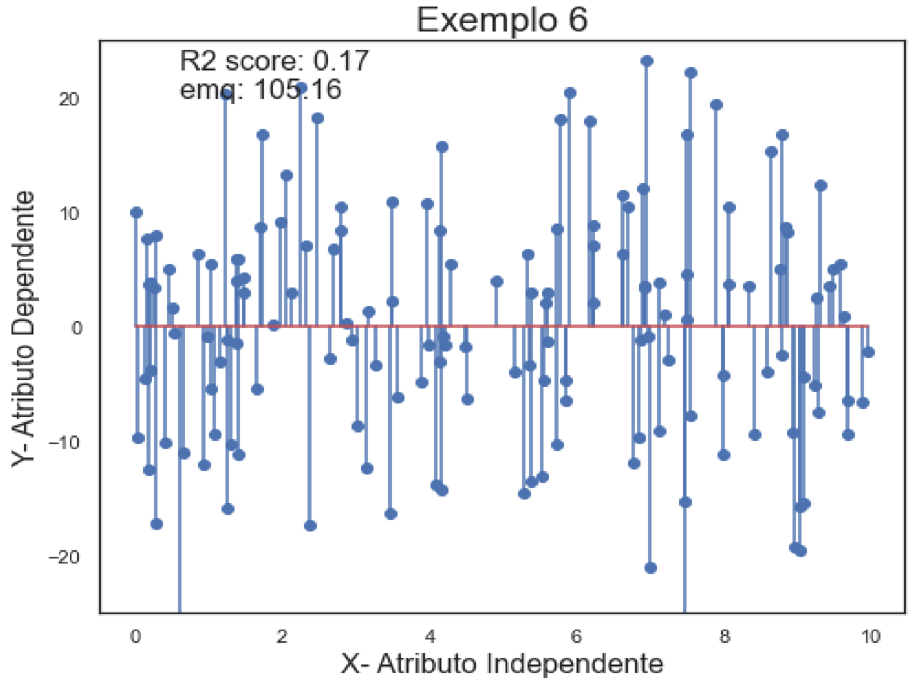
Exemplo 6

- Equação:
- Y = -1.5*x -6.549
- Por exemplo se x

X	Y
2	-9.529
5	-13.999
10	-21.449

$R^2 = 0.17$
EMQ = 105





Comparando os resultados

X	Y	Y	Y
2	9	8.725	8.425
5	15	15.475	15.931
10	25	26.725	28.441

	$R^2 = 0.97$	$R^2 = 0.65$	$R^2 = 0.36$
	EMQ = 1.05	EMQ = 26.29	EMQ = 105

X	Y	Y	Y
2	-9.055	-9.275	-9.529
5	-14.905	-14.525	-13.999
10	-24.655	-23.275	-21.449

	$R^2 = 0.97$	$R^2 = 0.52$	$R^2 = 0.17$
	EMQ = 1.05	EMQ = 26.29	EMQ = 105

Qual o Melhor Modelo ?

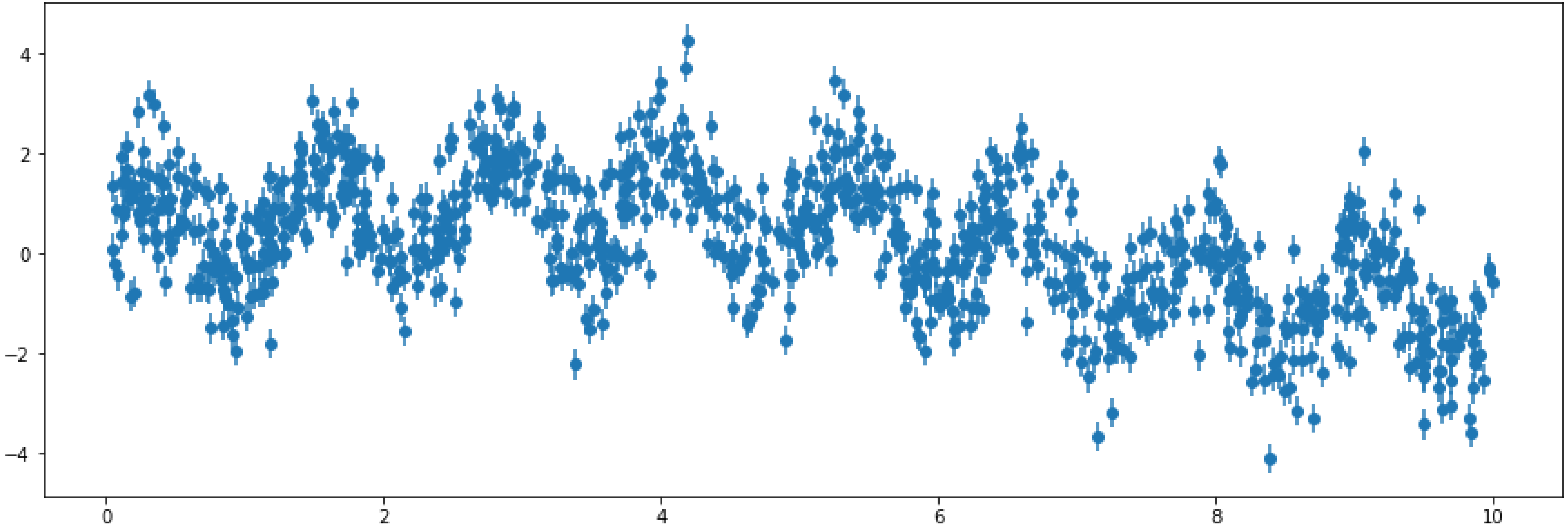
Afinal qual o melhor modelo ?

- Todo modelo possui um erro
- Quanto maior o R2 e menor o emq, melhor ajustado
- Perceba que quanto maior a dispersõa dos dados menor o R2 e maior o erro
- Portanto avaliara o modelo com base em métricas é uma forma de dizer como ele se ajusta aos nosso dados
- Mas se os dados forem muito dispersos o modelo possui um maior erro e portanto sua predição ser menos acurada.

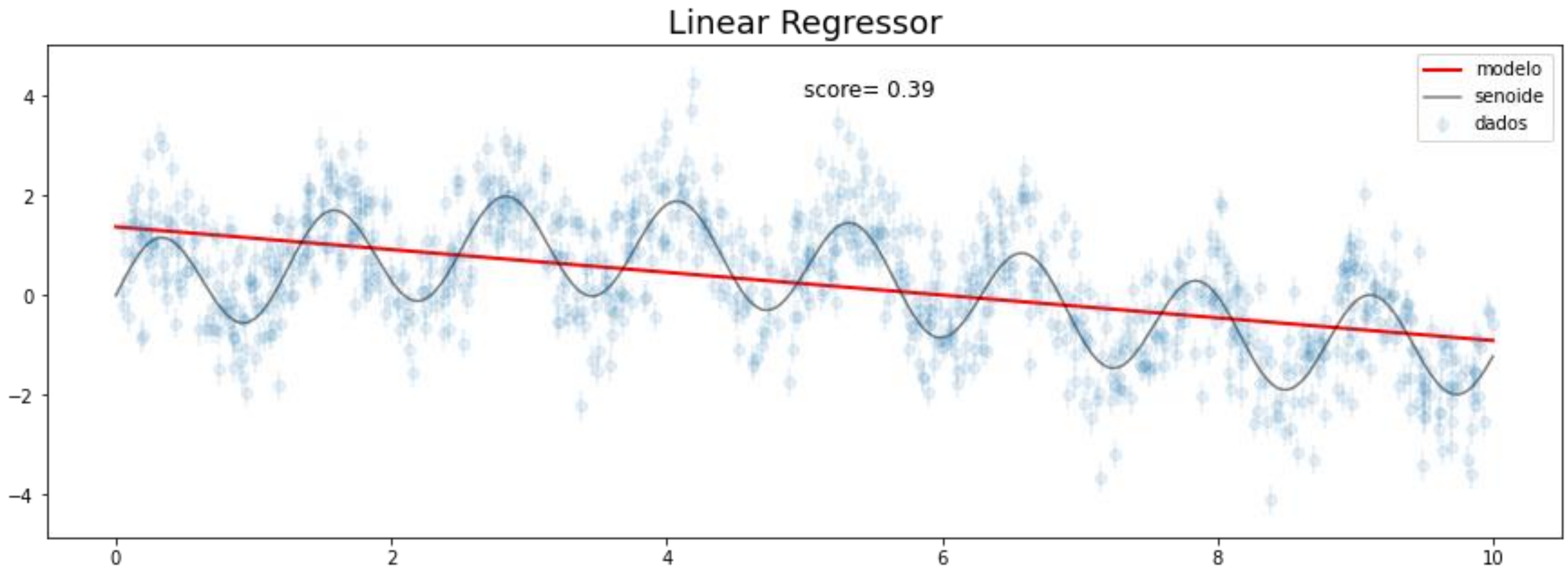
Regressão não linear

- Muitas vezes os dados não estão alinhado em trono de uma linha
- E nesses casos a regressão linear pode não ser uma boa escolha
- A maioria dos modelos de classificação que estudamos na aula passada podem ser usados para a regressão
- Vamos ver um caso de regressão não linear e os resultados de vários modelos na sua modelagem

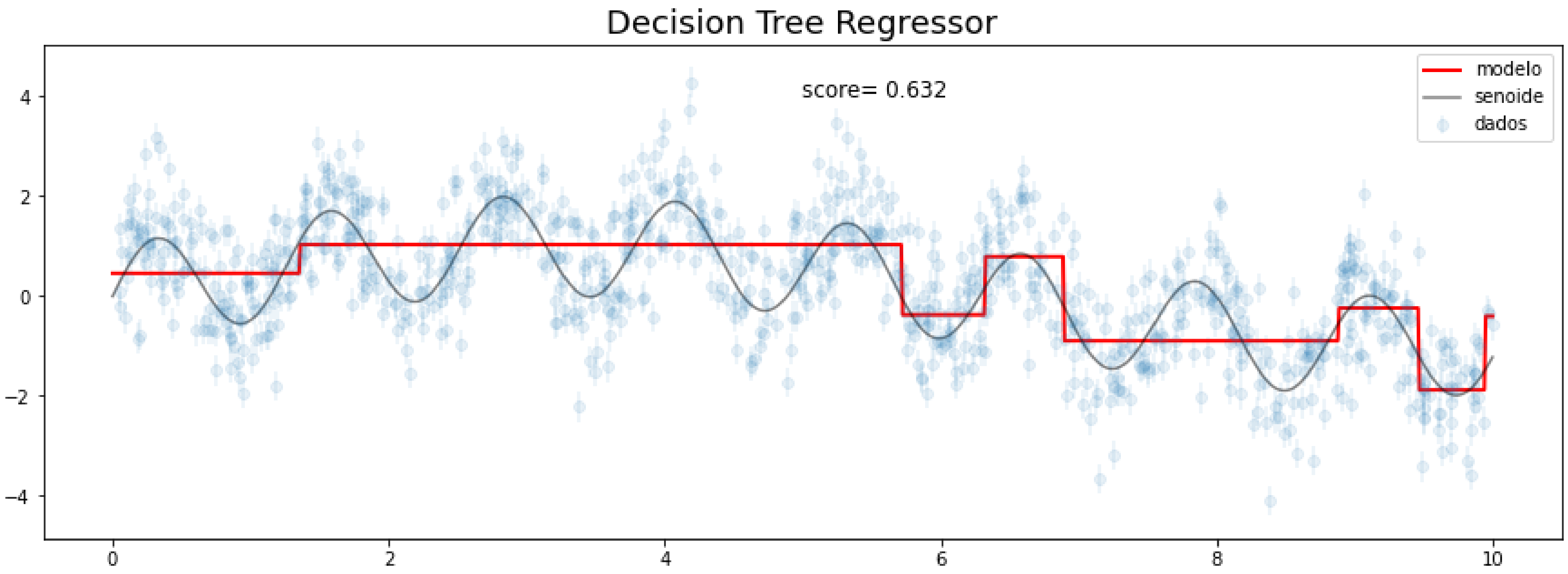
Nosso dados



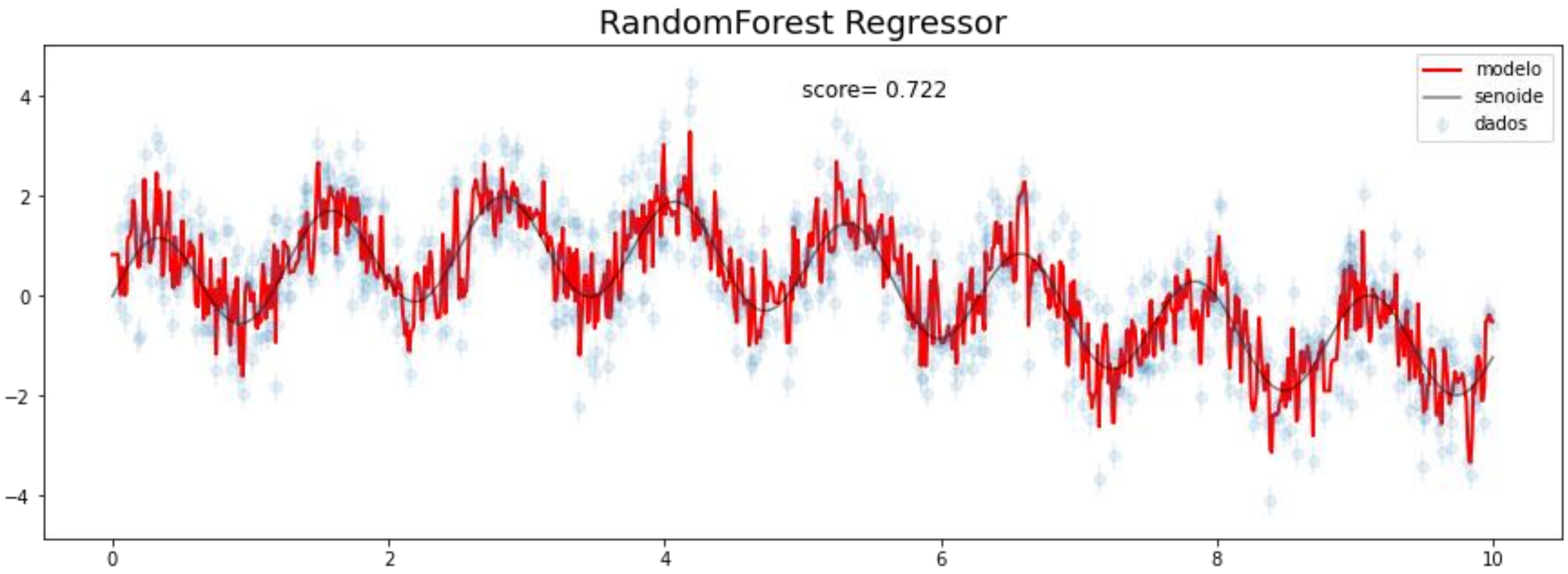
Usando a regressão linear



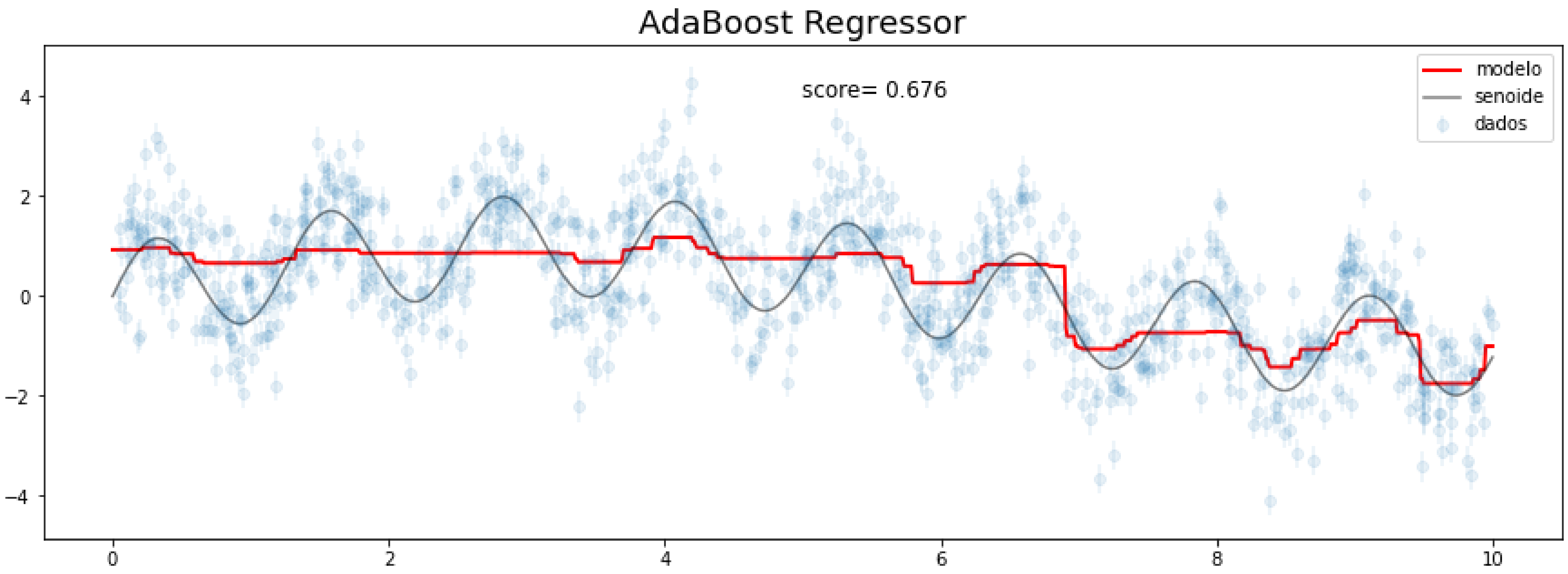
Usando uma Árvore de Decisão



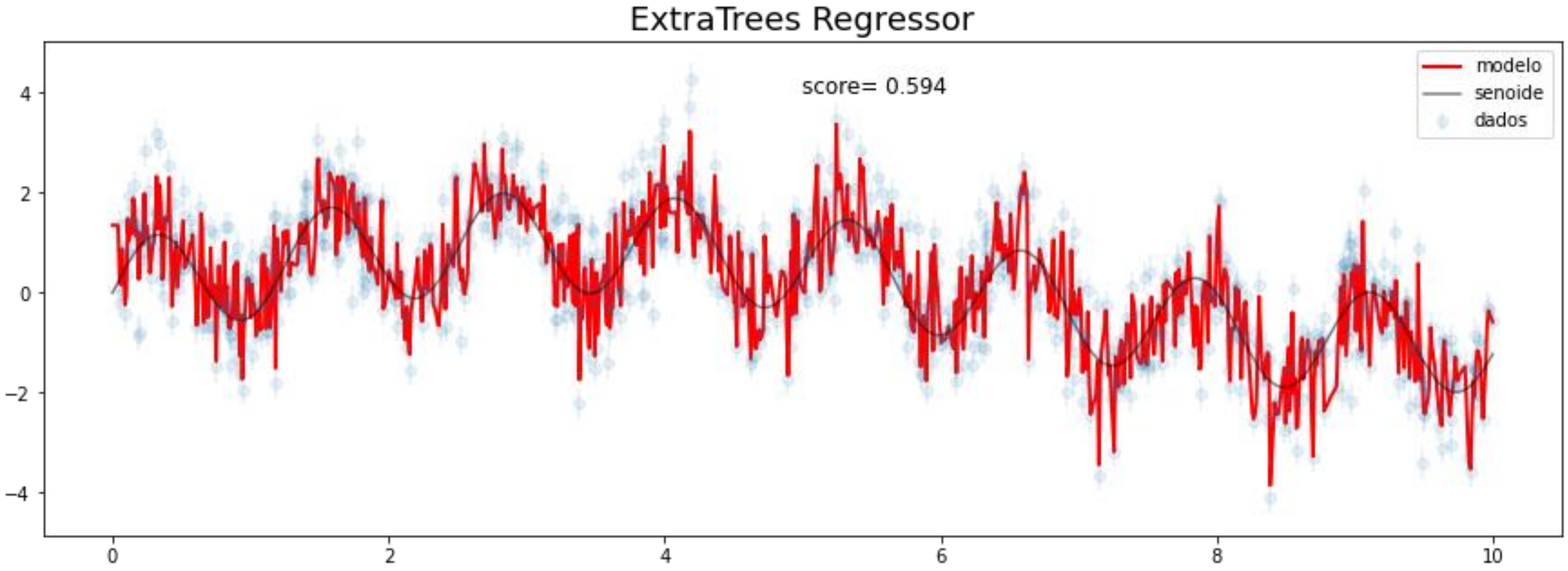
Usandoa Floresta Randômica



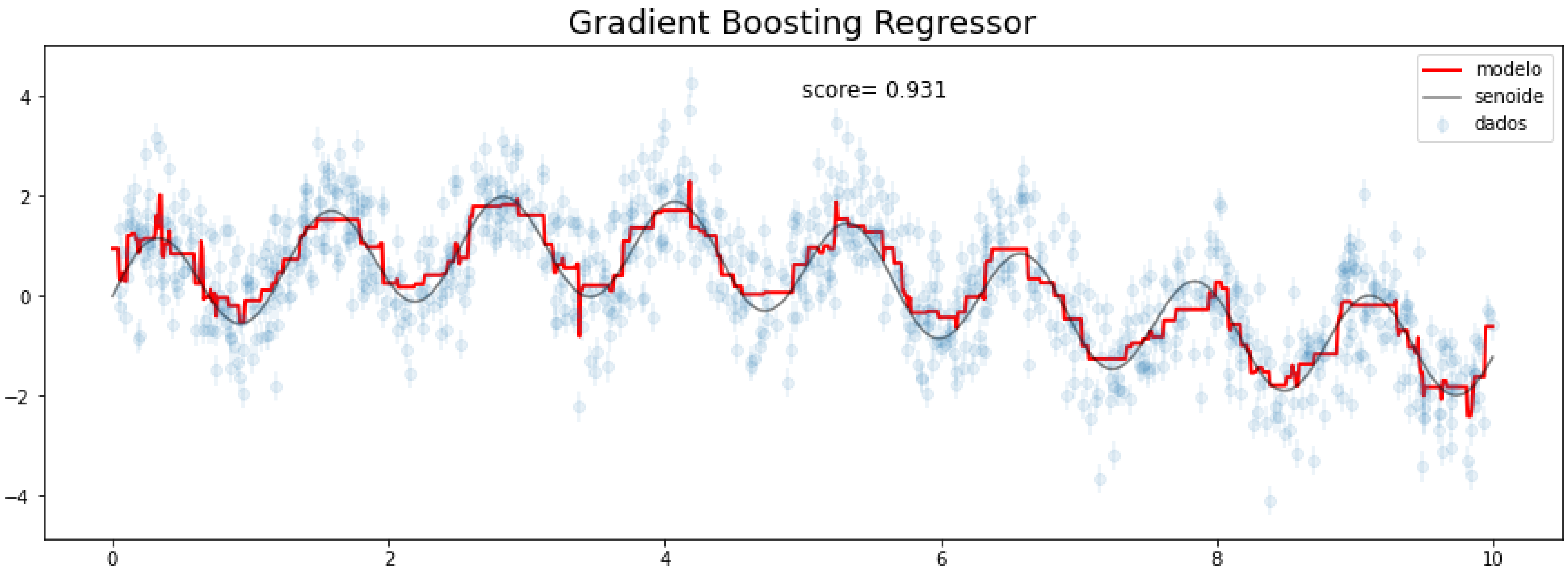
Usando o AdaBoost



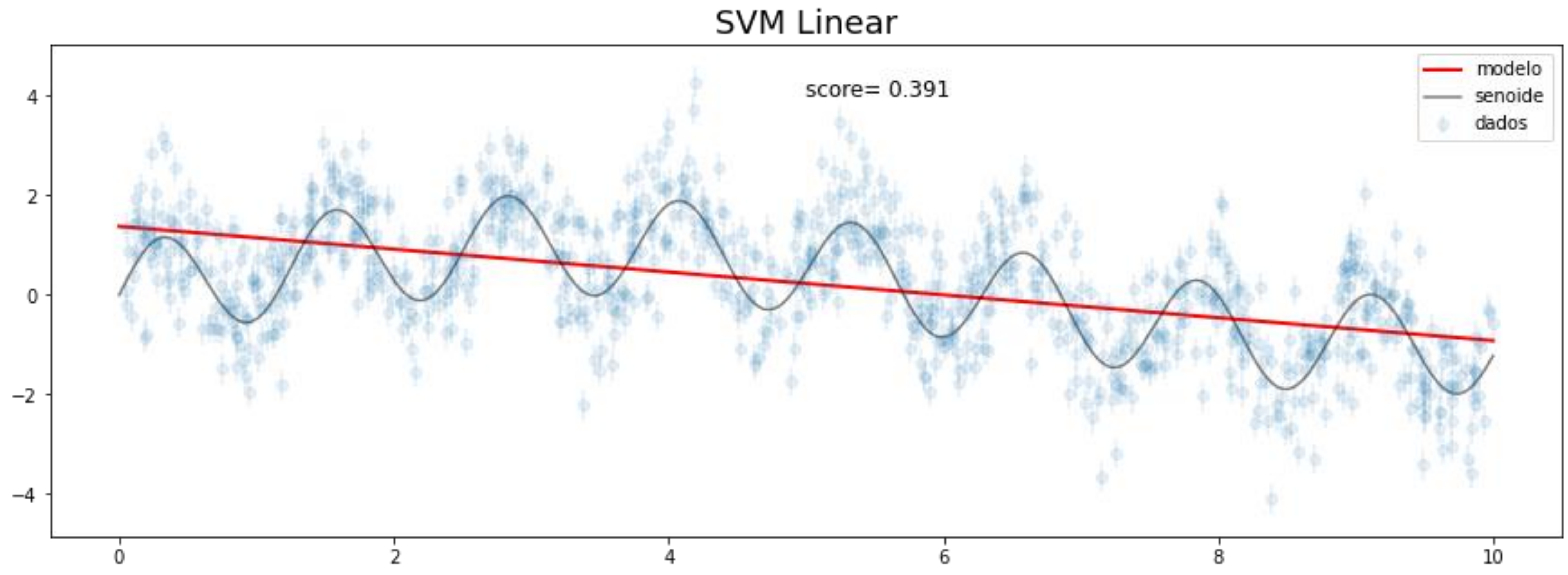
Usando o ExtraTrees



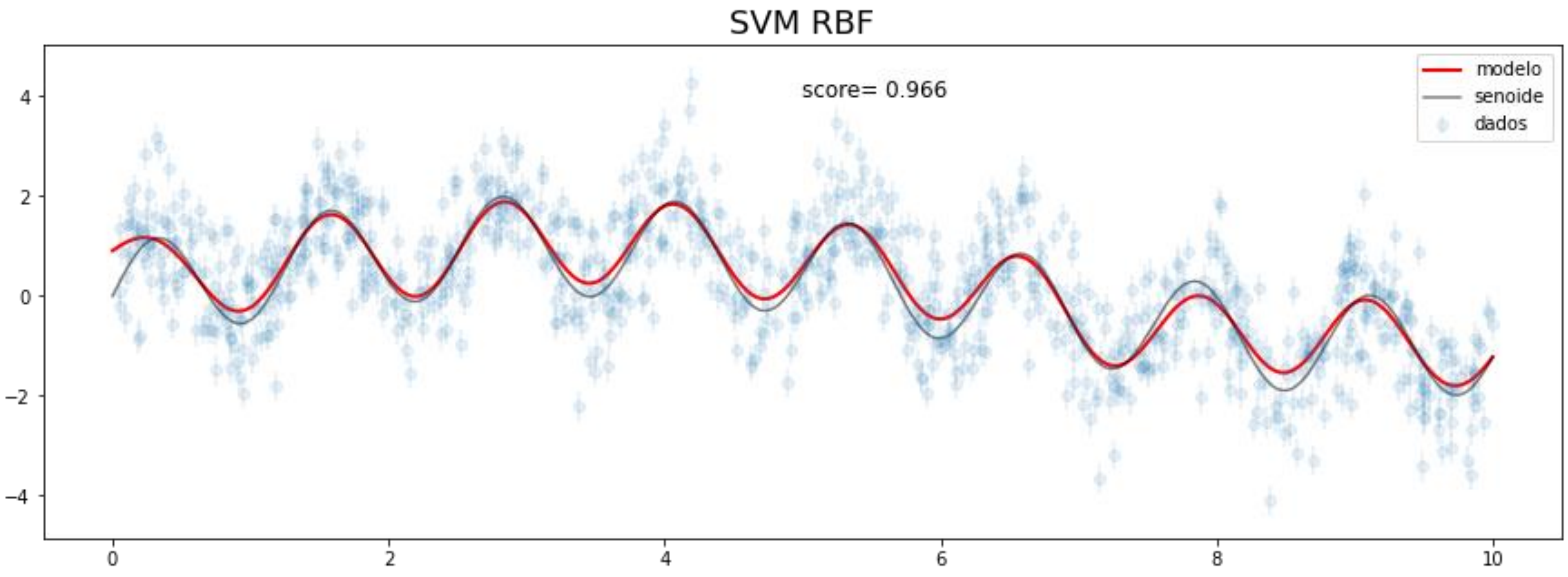
Usando o Gradiente Boosting



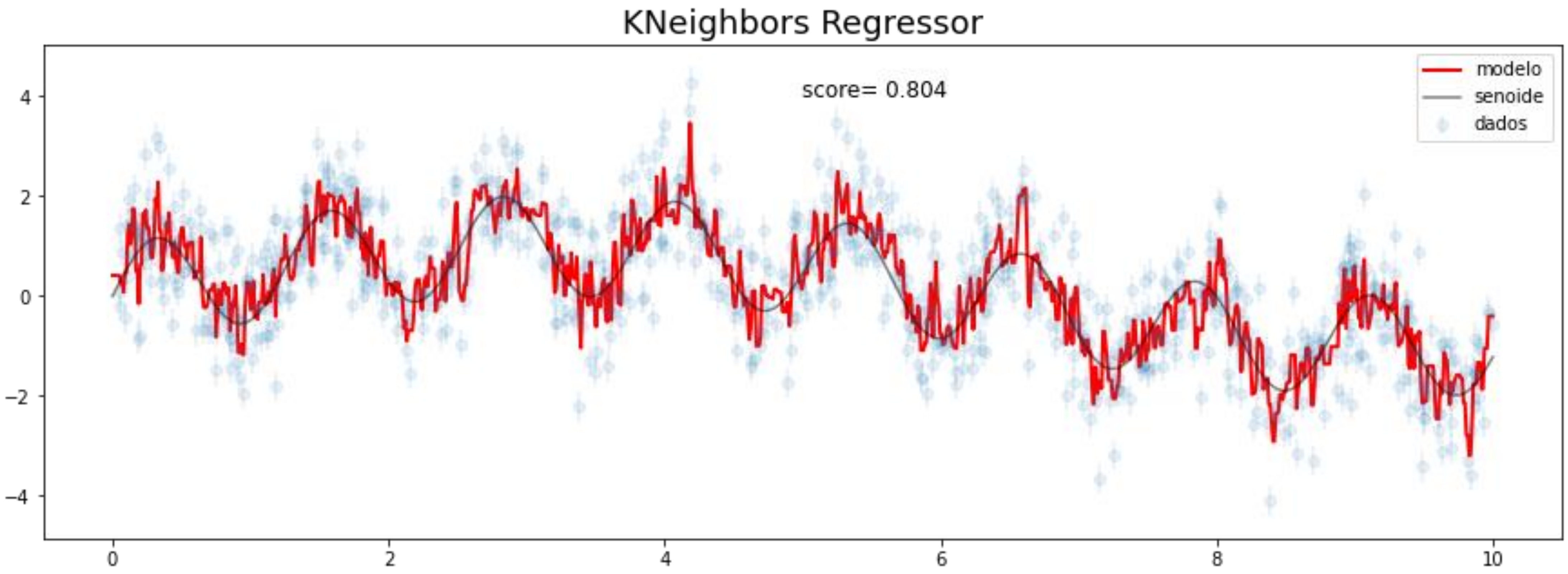
Usando o SVM Linear



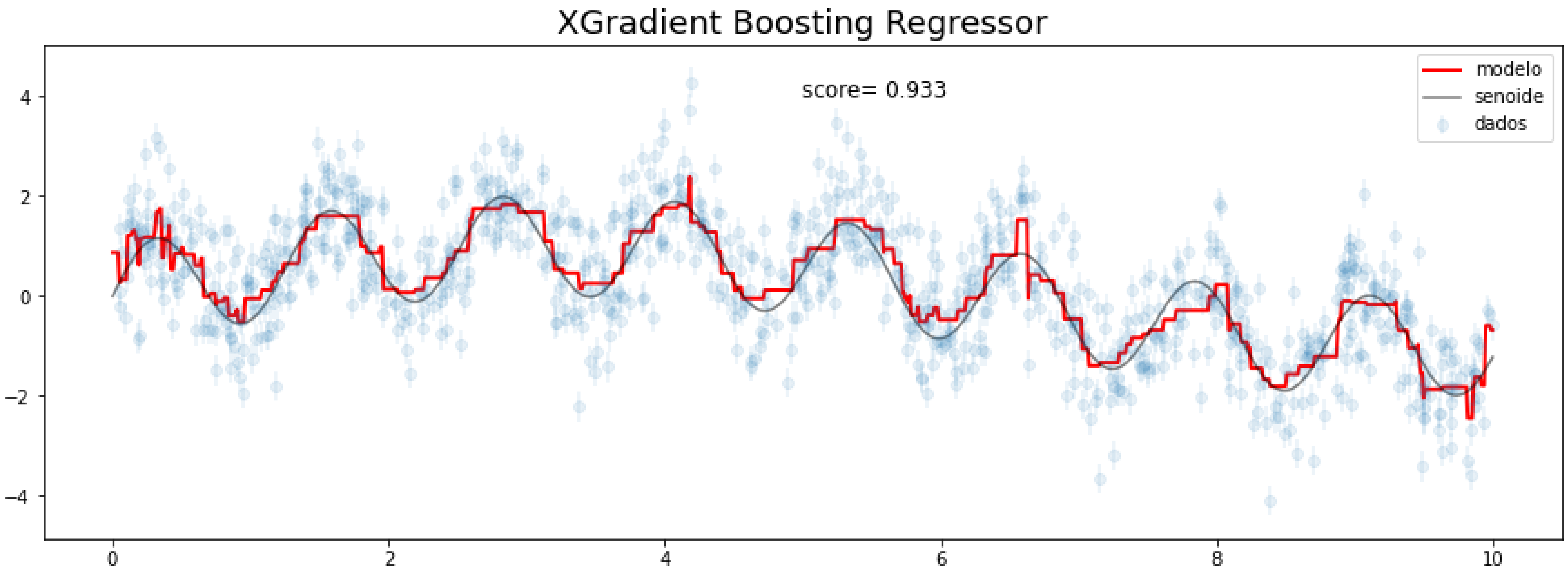
Usando o SVM RBF



Usando o k Vizinhos Mais Próximos



Usando o XGBoost



Usando CatBoost