

# Data Science Bootcamp



Host Oficial  
**sãojudas**  
universidade



**MÓDULO #3**

# **Modelos Regressivos**

**Talita Correa**  
**Gisely Alves**



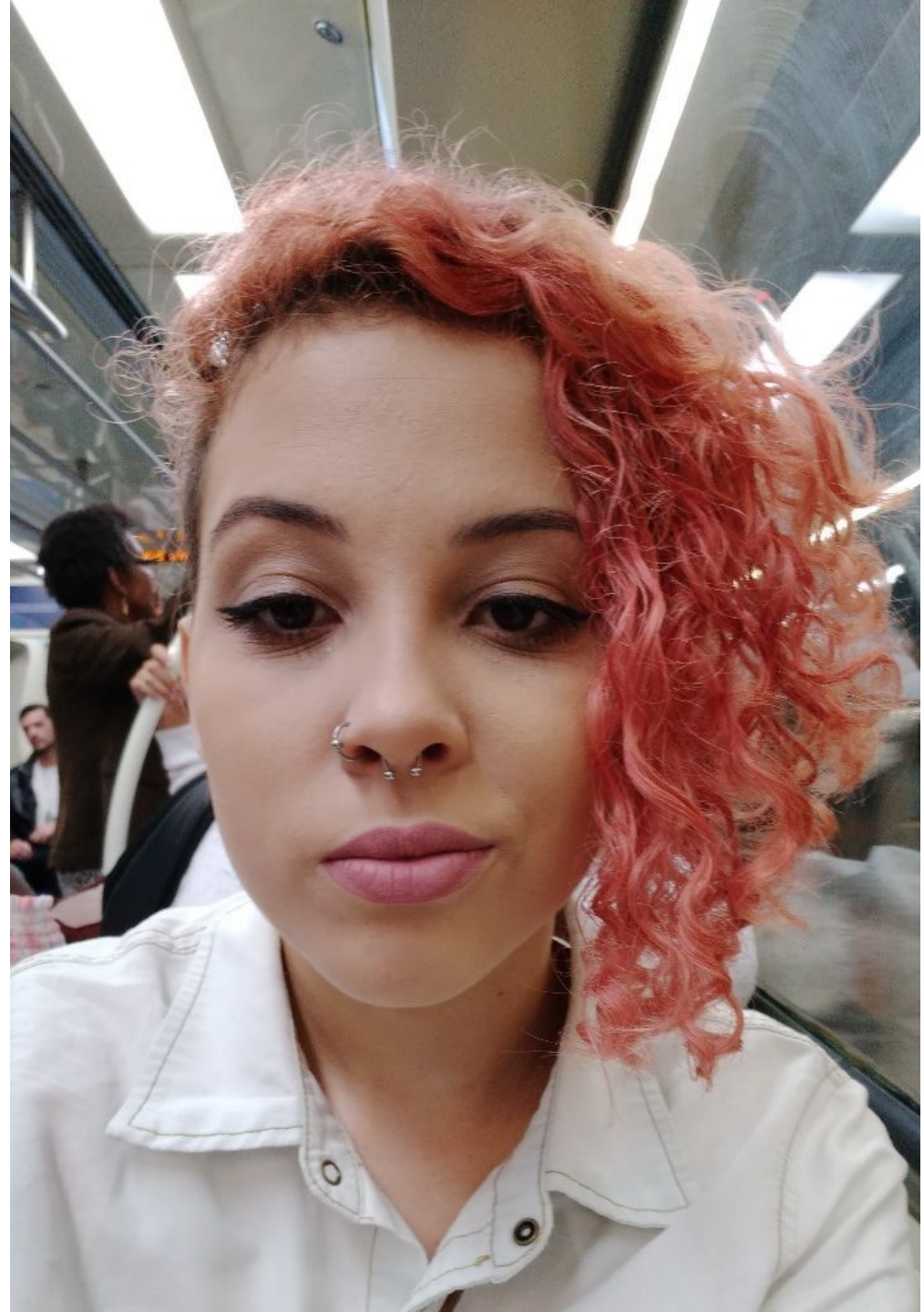
# Gisely Alves

Desenvolvedora

 @giselyalves13

**M** gisely.alves    **DEV** giselyalves13

 linkedin.com/in/giselybrandao



# Talita Correa

Gerente de Data & Analytics



@talitacb



talitacb



[linkedin.com/in/talitacorreabarcelos/](https://linkedin.com/in/talitacorreabarcelos/)



# Preço de uma Casa



R\$ 70.000,00



?



R\$ 160.000,00



# Preço de uma Casa



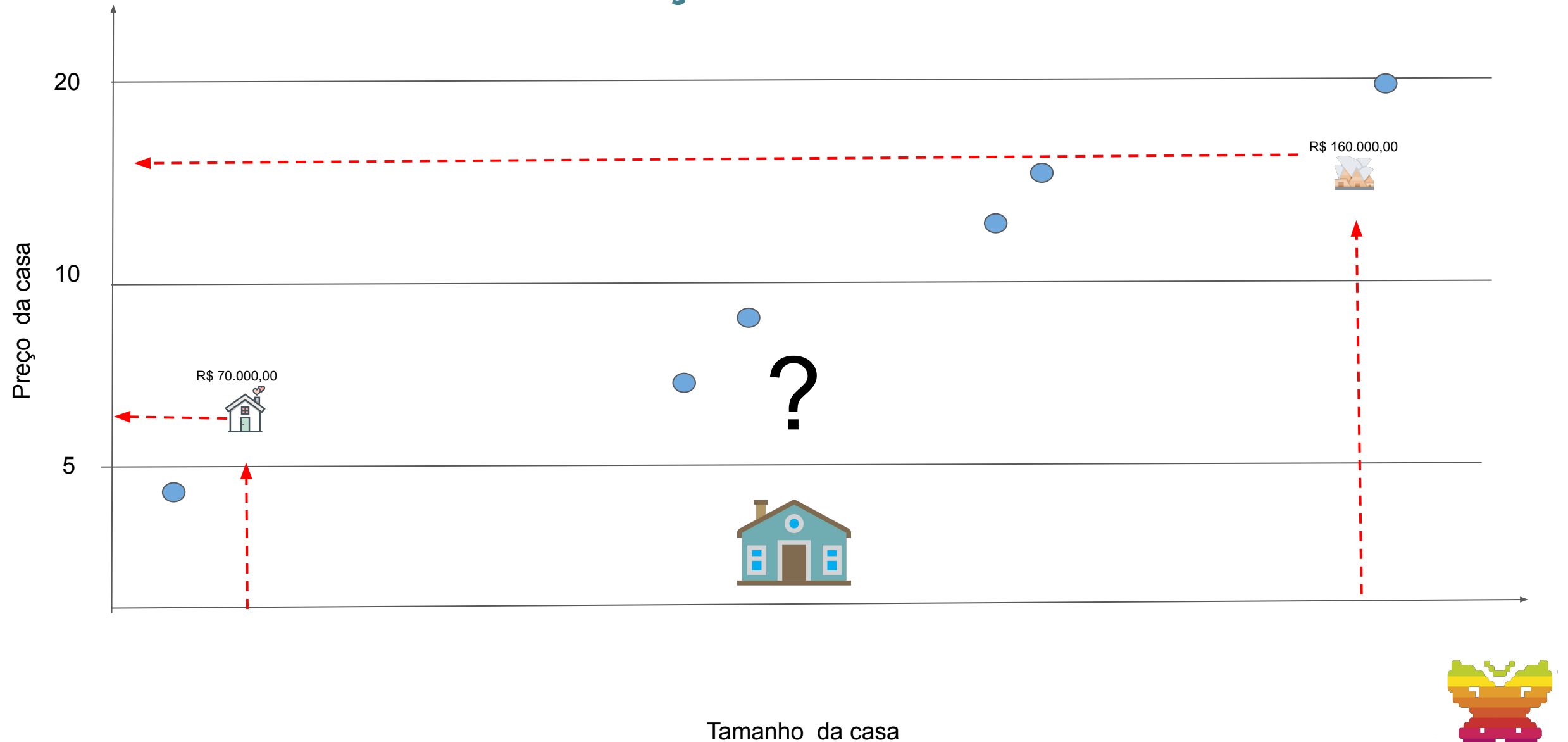


?

- R\$ 80.000,00?
- R\$120.000,00?
- R\$190.000,00?

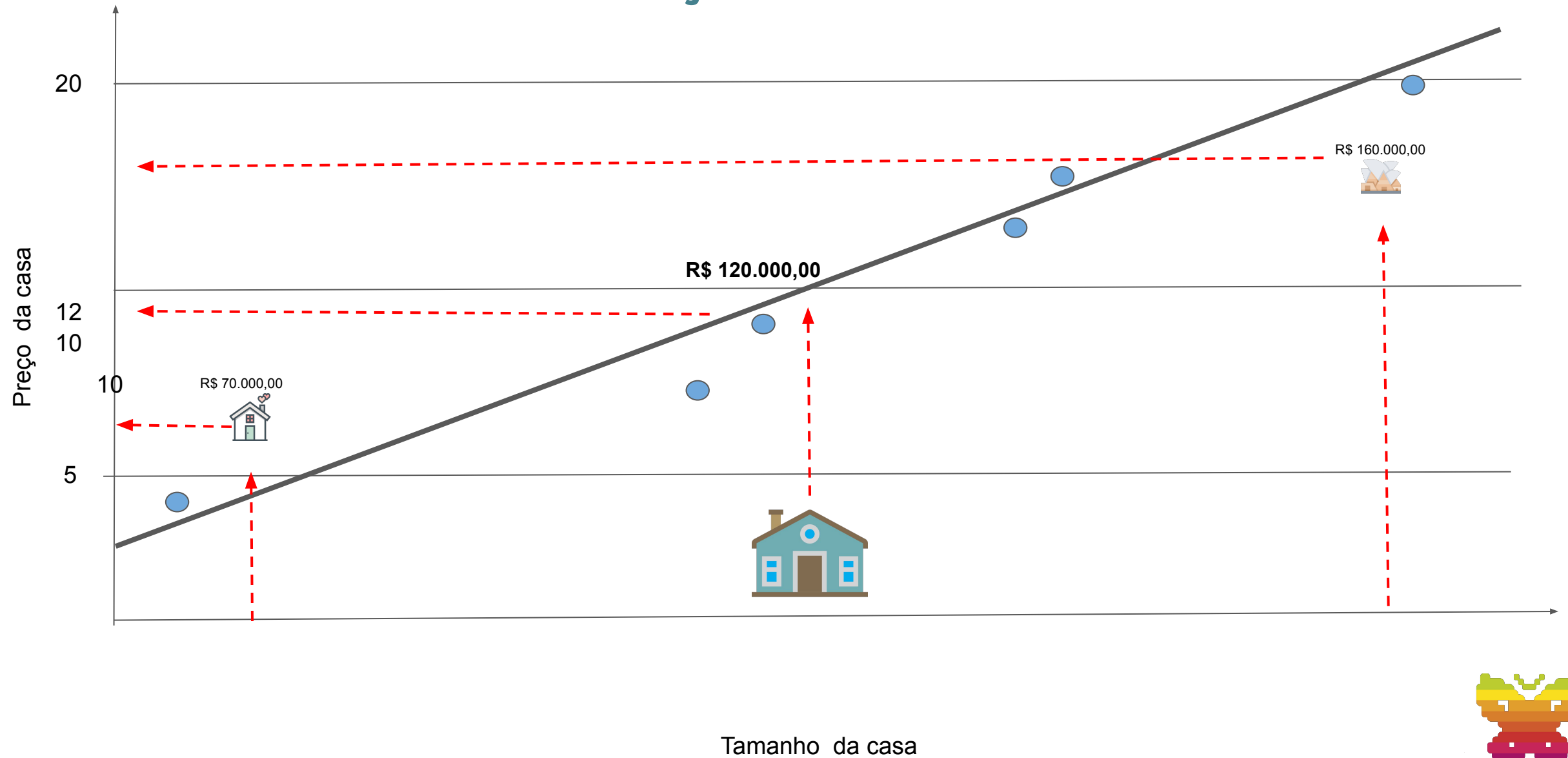


# Preço de uma Casa



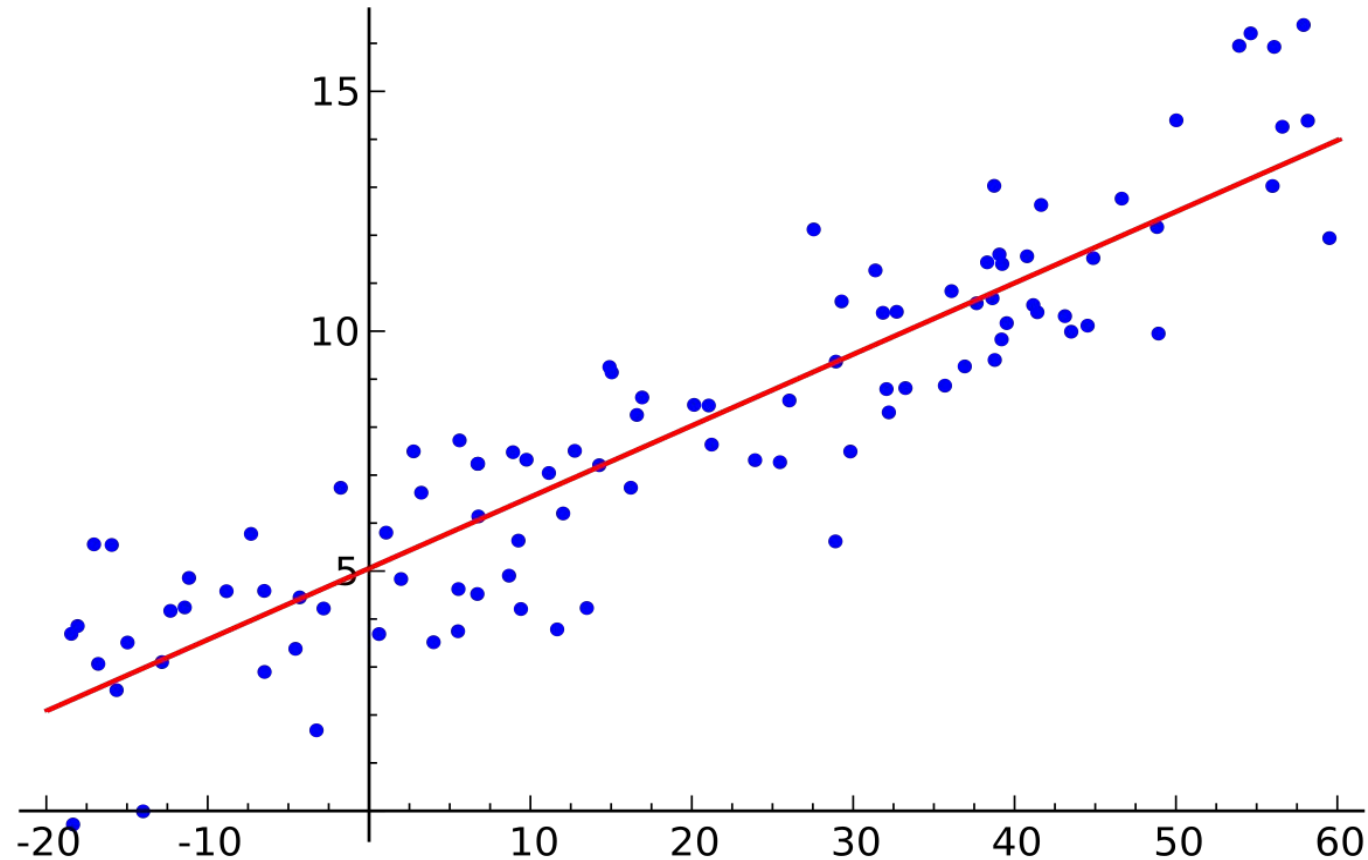


# Preço de uma Casa



# O que é uma regressão?

Uma linha que descreve a relação entre os dados



# Regressões

Variáveis dependentes → Resposta

Variáveis independentes → Podem estar relacionadas com a resposta

Respostas quantitativas

Regressões Lineares e Não Lineares

**Objetivo: Encontrar a “reta” que melhor descreve a relação entre as variáveis**



# O que uma regressão pode nos dizer?

- ▷ Existe relação entre a minha resposta e as outras variáveis?
- ▷ Quão forte é a relação dessas variáveis com a variável resposta?
- ▷ Qual delas contribui mais para variável resposta?
- ▷ O quanto cada variável afeta a resposta?
- ▷ Essa relação é realmente linear?
- ▷ Quão acurado conseguimos predizer as respostas?
- ▷ Explicável



# Regressão Linear Simples

$Y$  = variável resposta

$$Y \approx \alpha + \beta X.$$

$\alpha$  e  $\beta$  = coeficientes

$X$  = variável independente

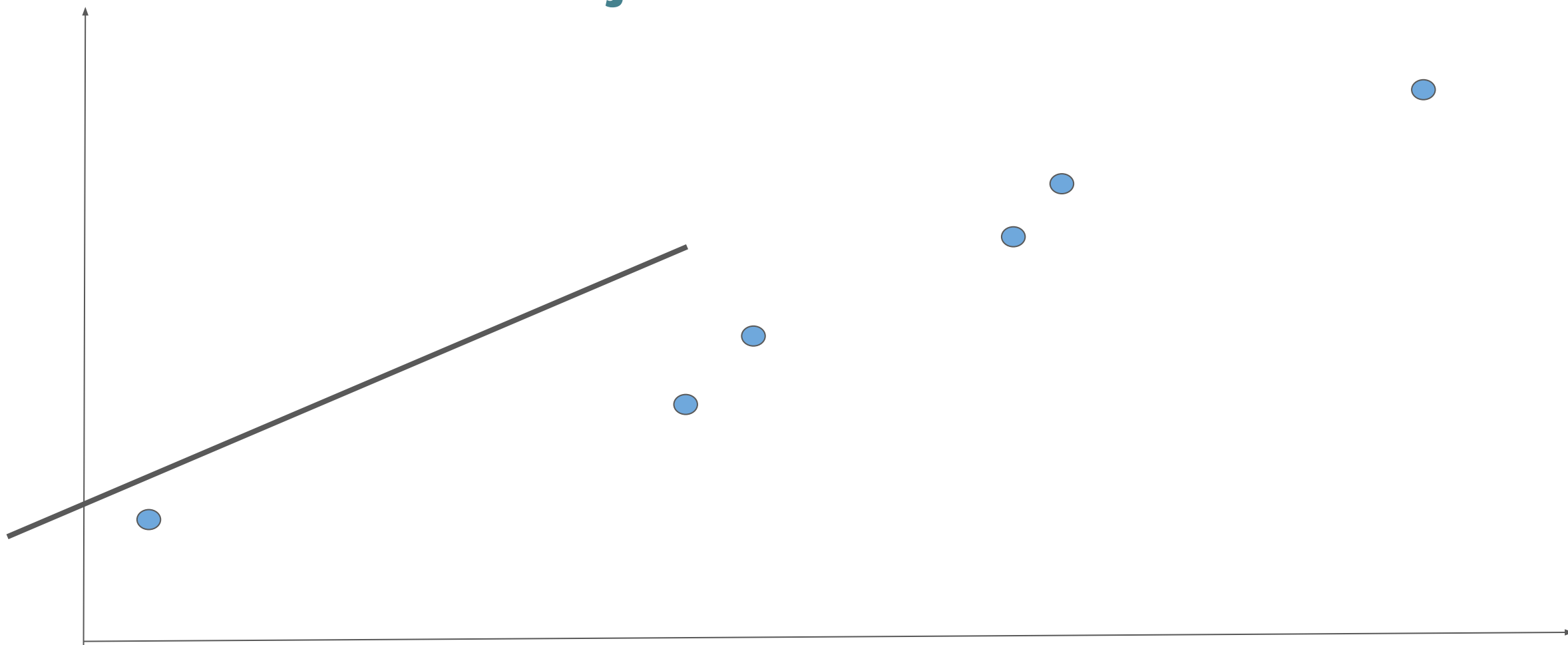
Queremos encontrar os melhores  $\alpha$  e  $\beta$  para a relação.



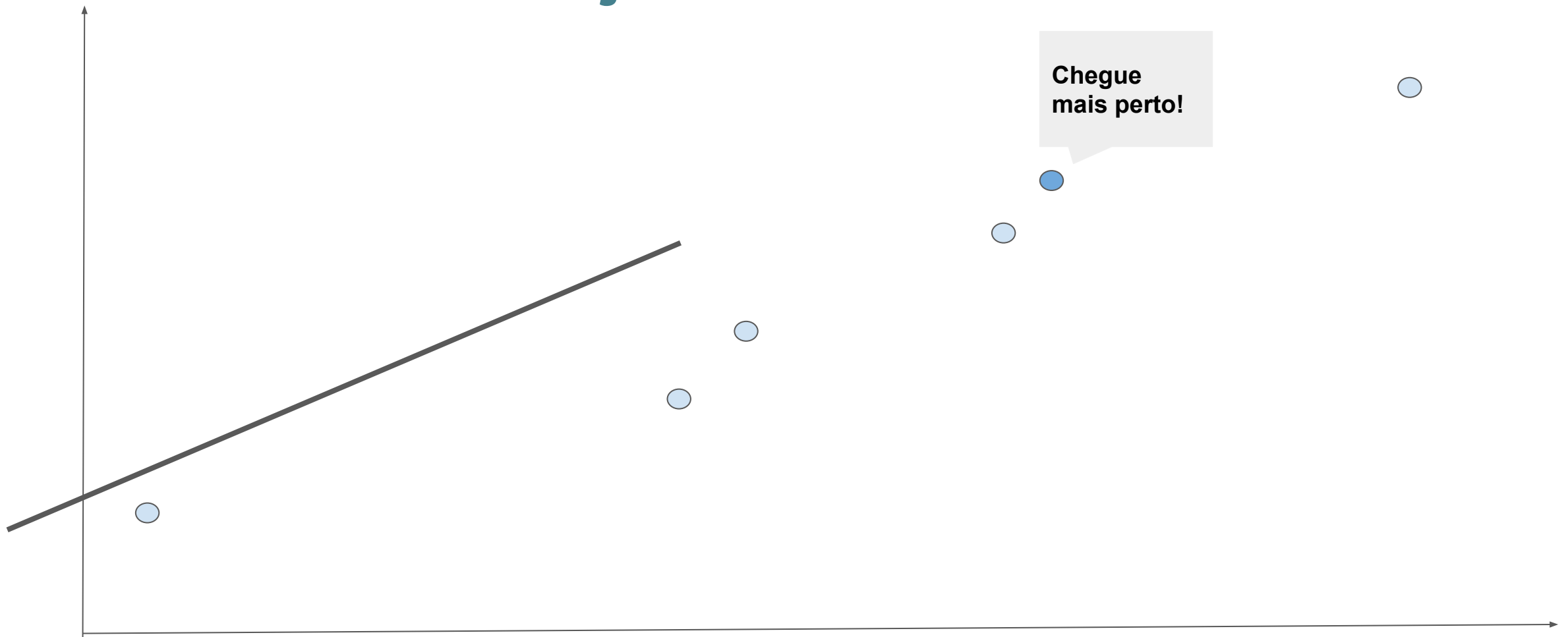
# Como encontrar melhores os coeficientes?



# Ajustando a Linha

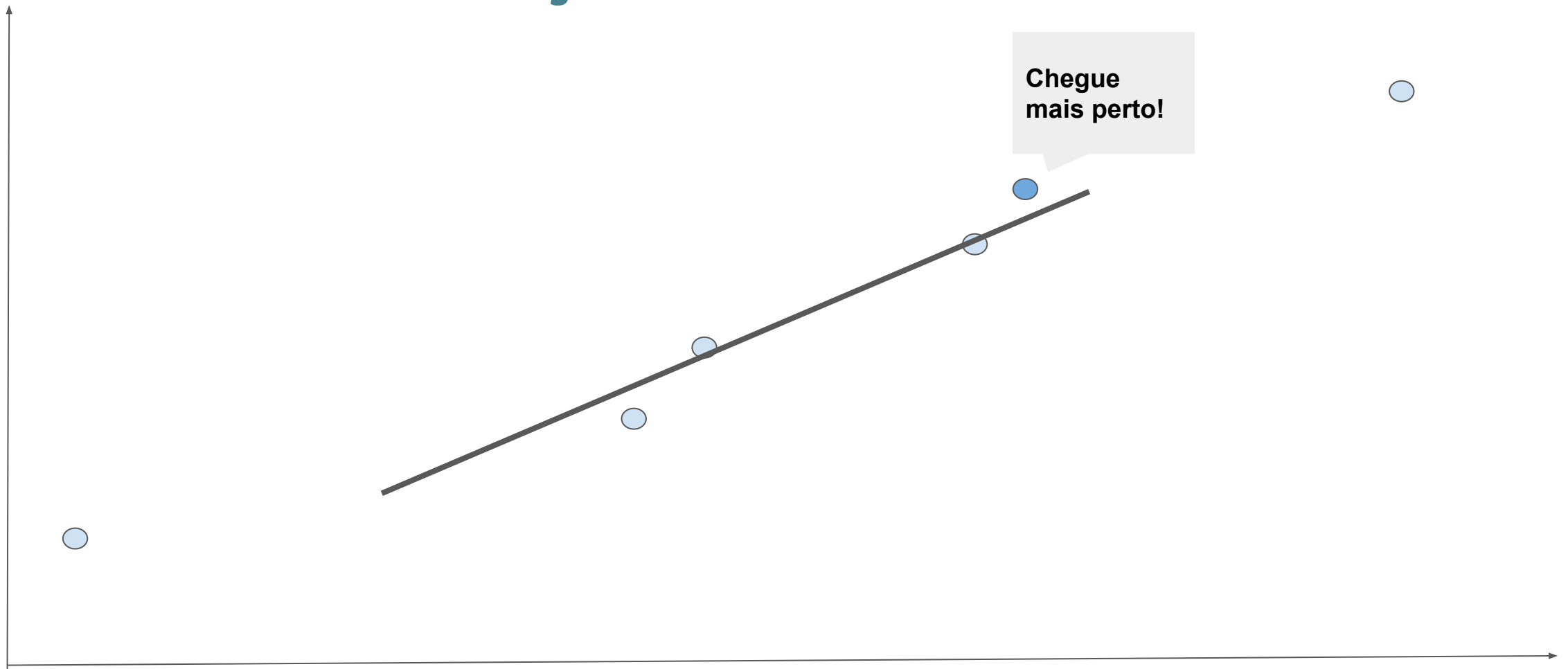


# Ajustando a Linha

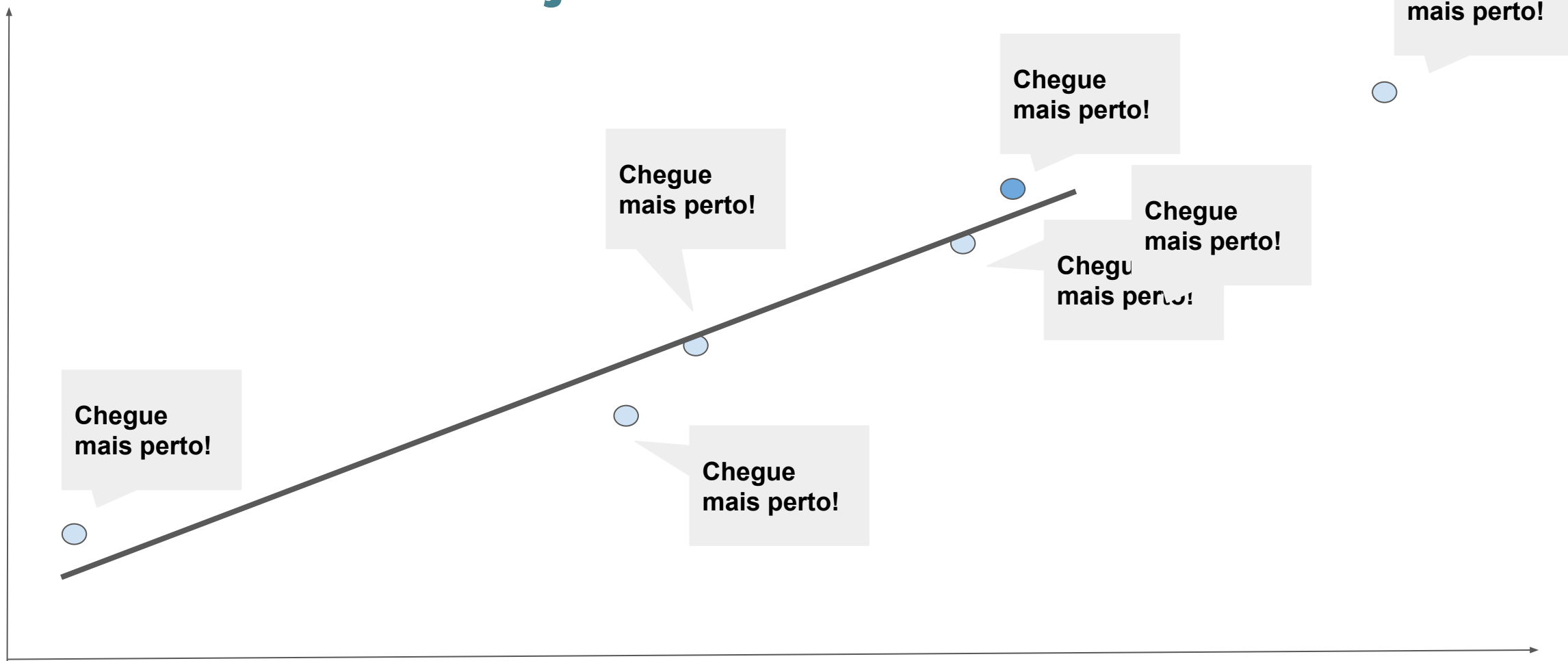




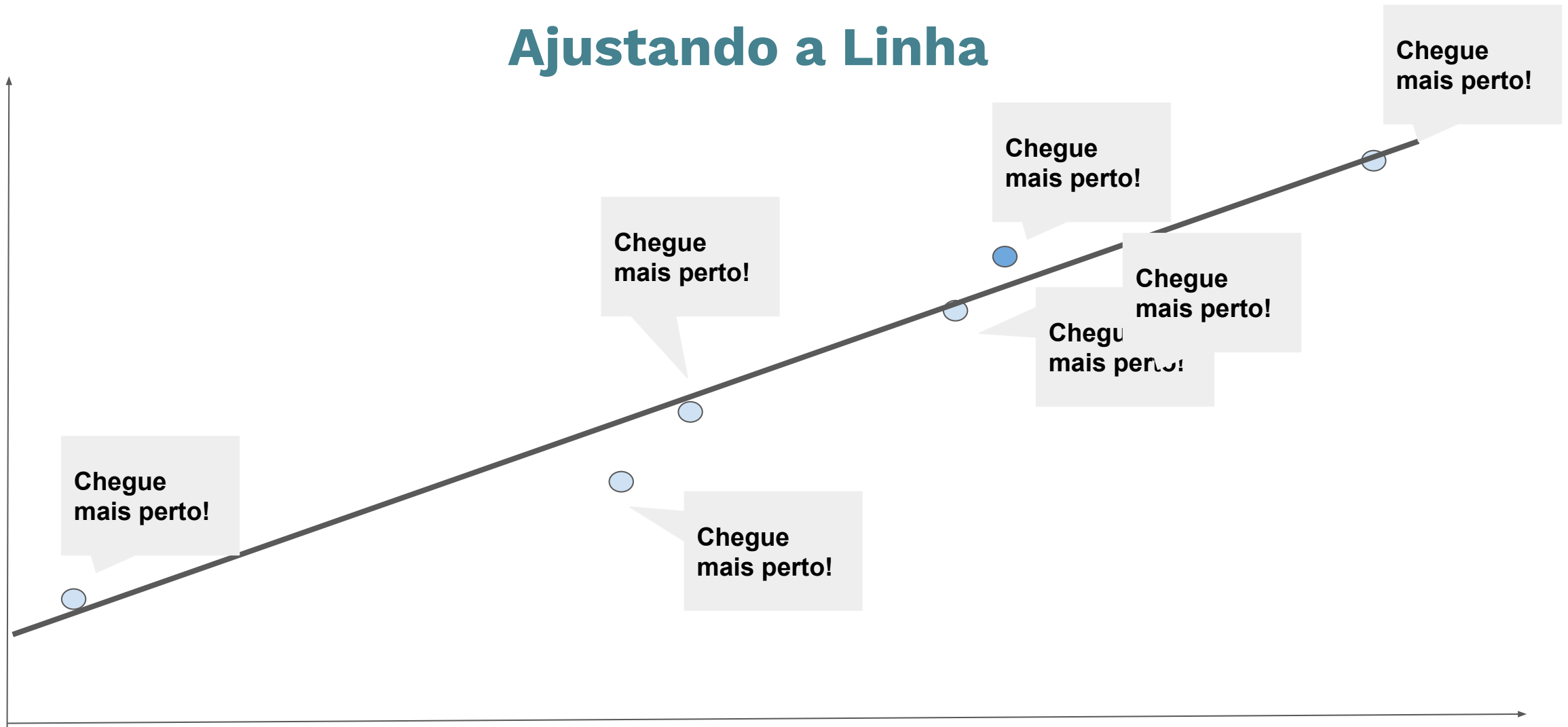
# Ajustando a Linha



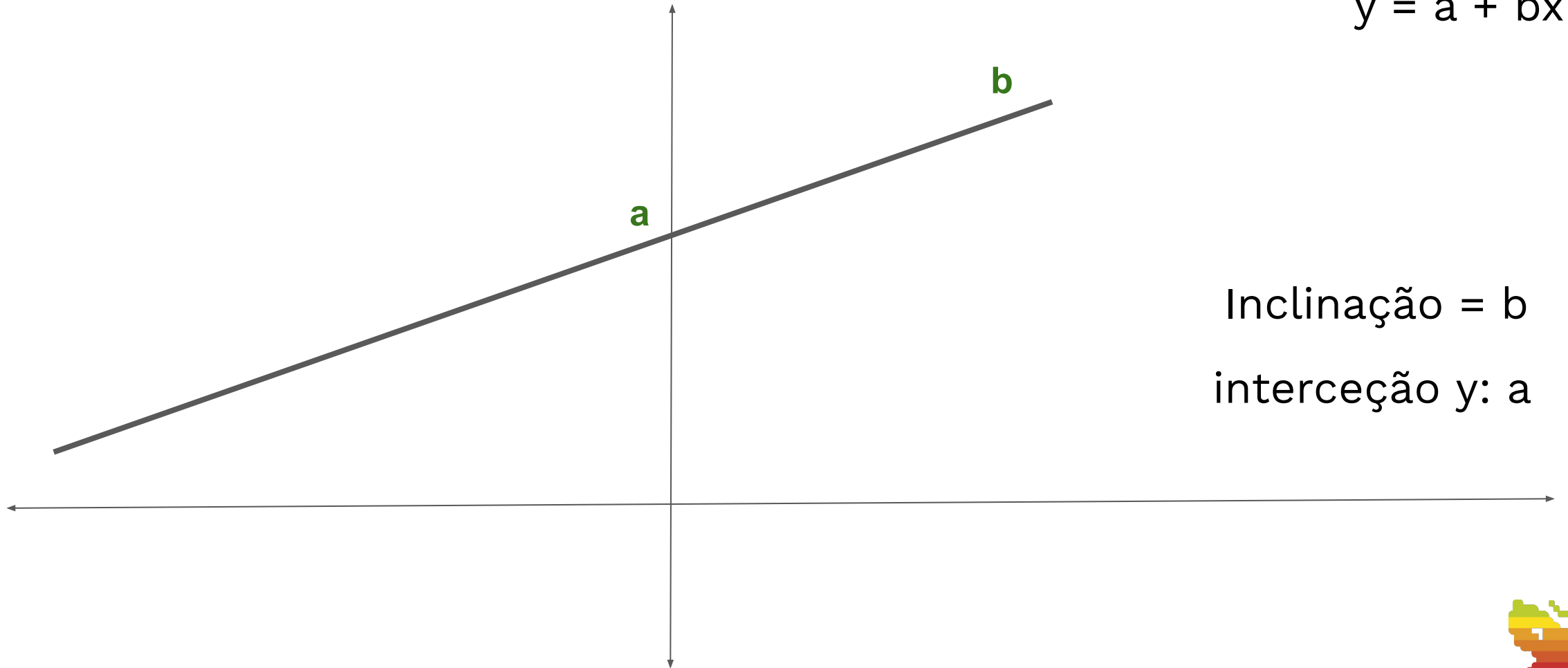
# Ajustando a Linha



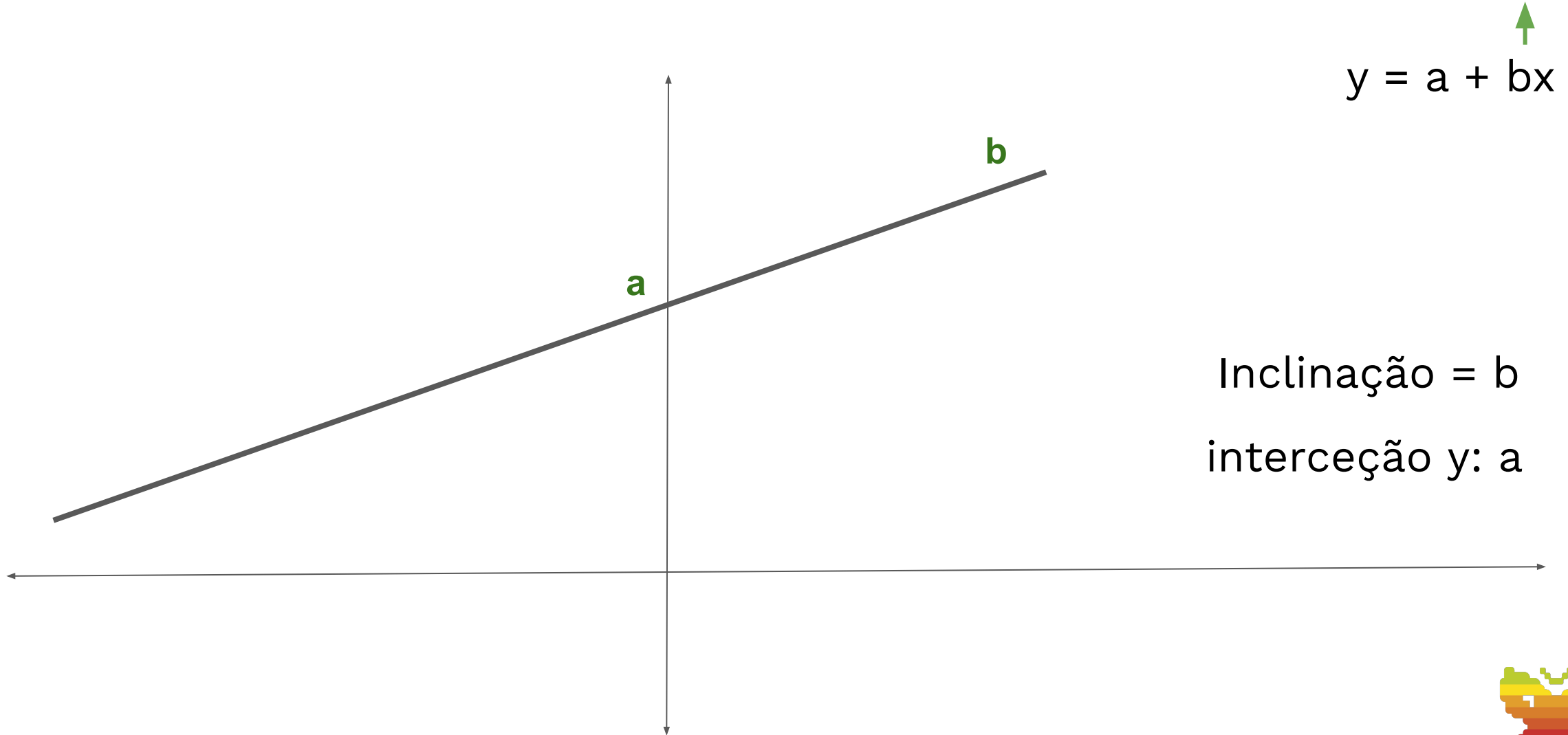
# Ajustando a Linha



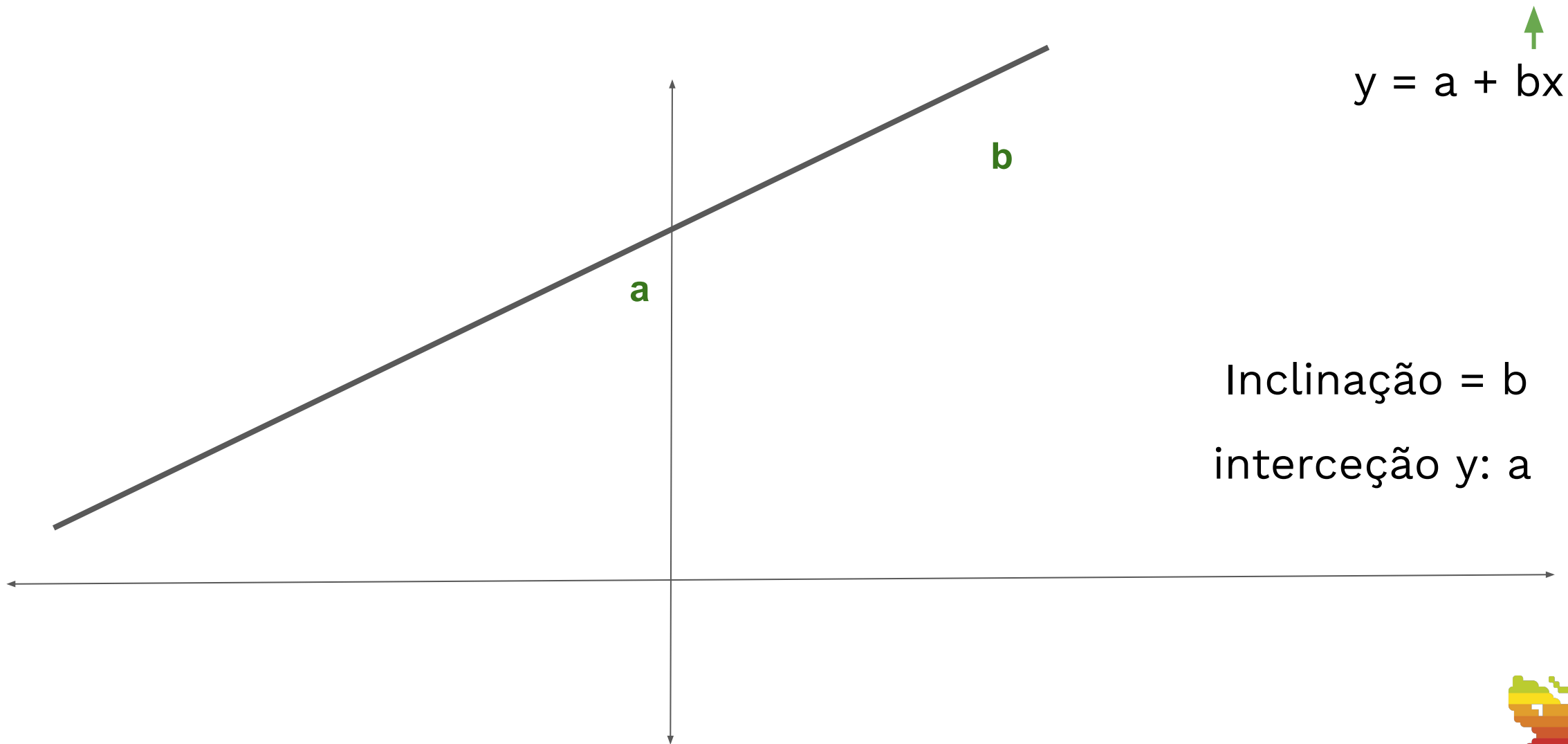
# Movendo as linhas



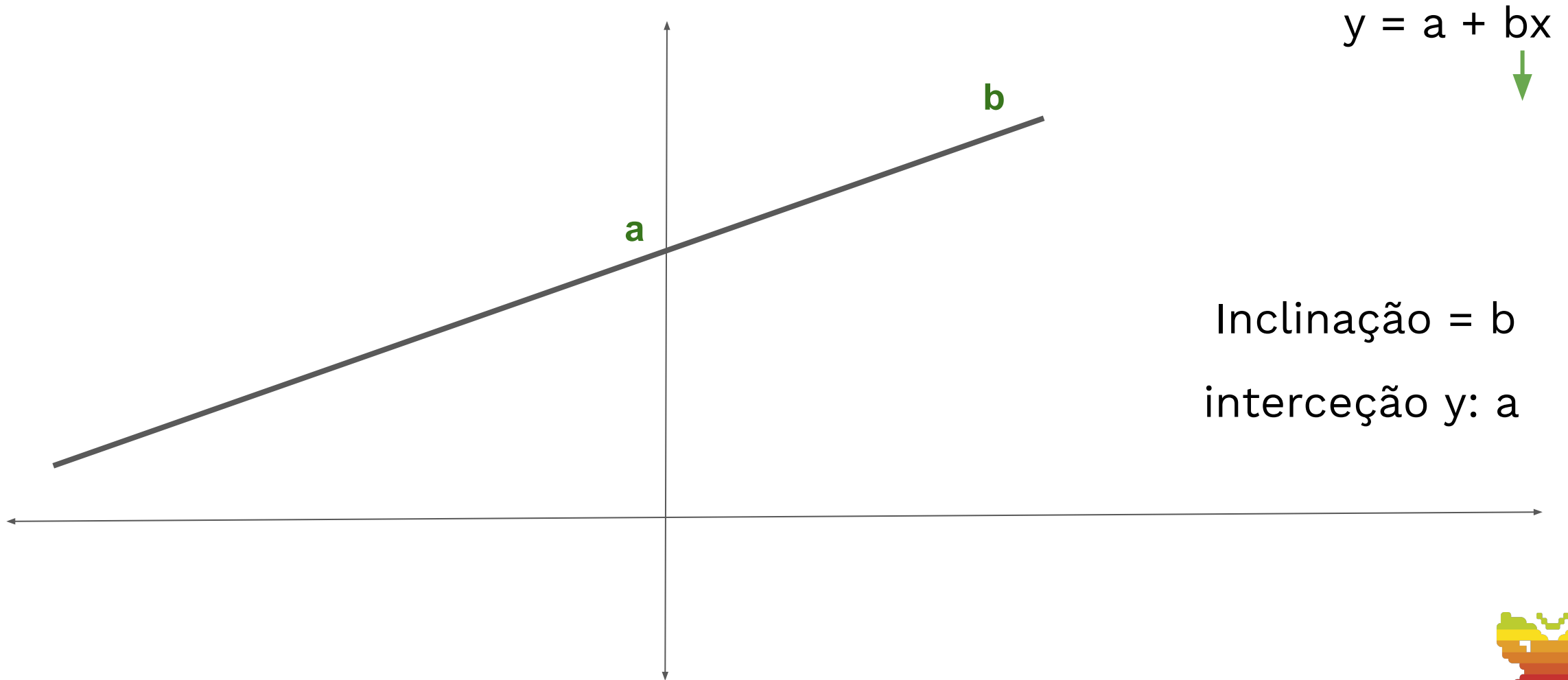
# O que acontece se aumentarmos o $b$ ?



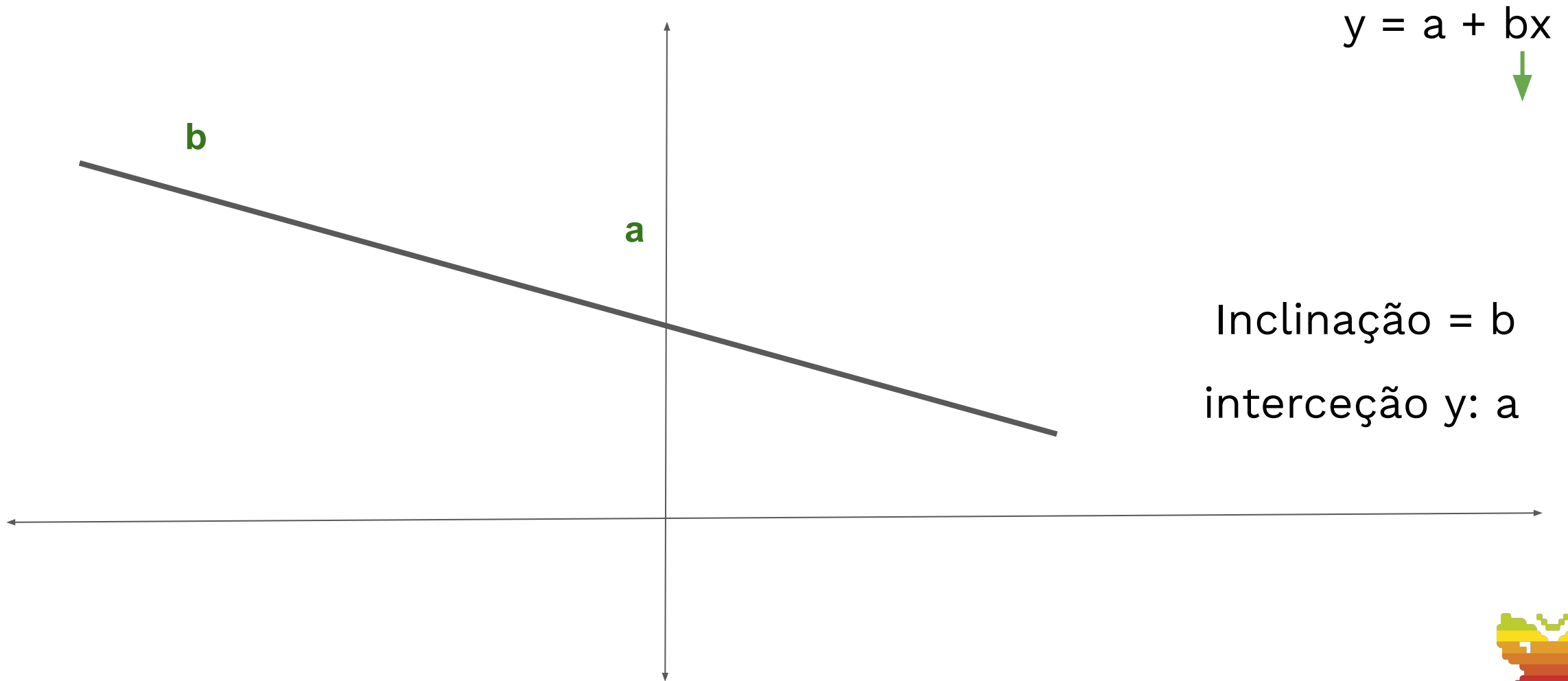
# A linha se move



# O que acontece se diminuirmos o $b$ ?



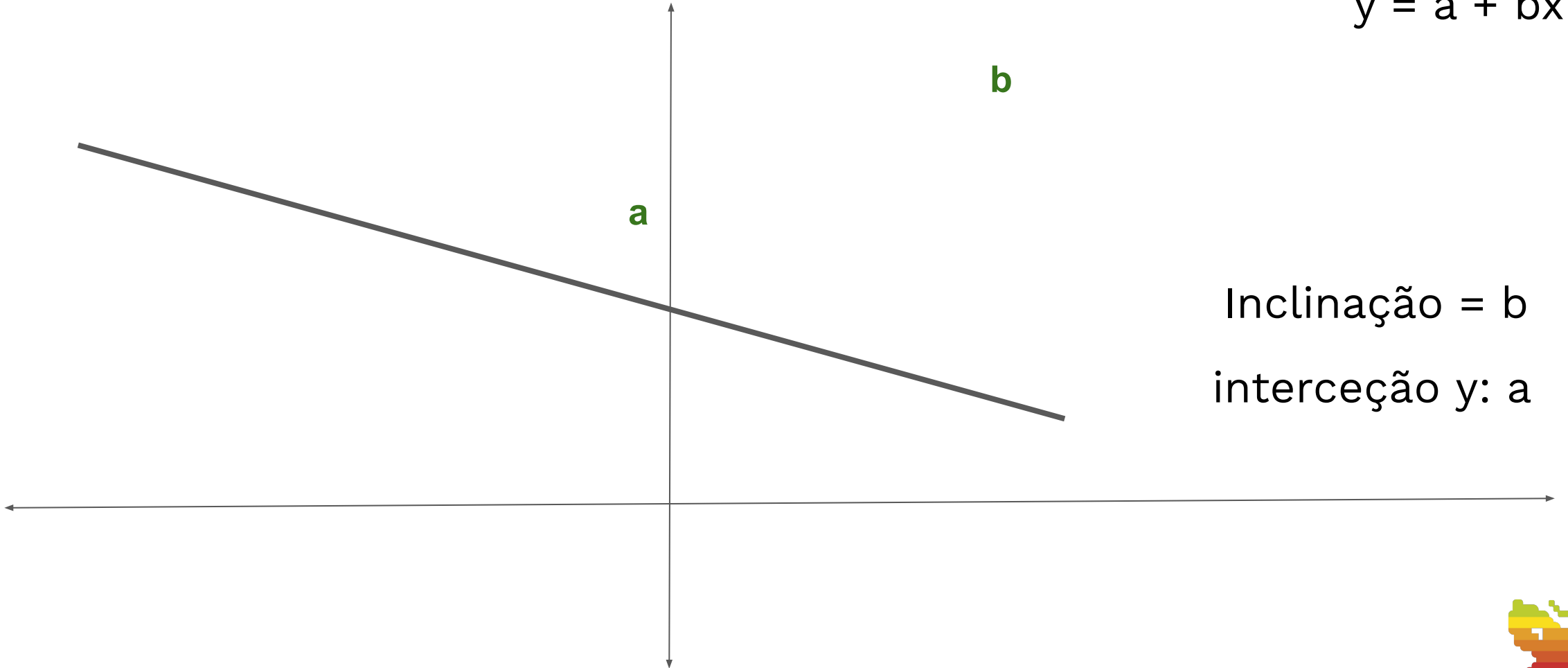
# A linha rotacional como abaixo





# O que acontece se aumentarmos o $a$ ?

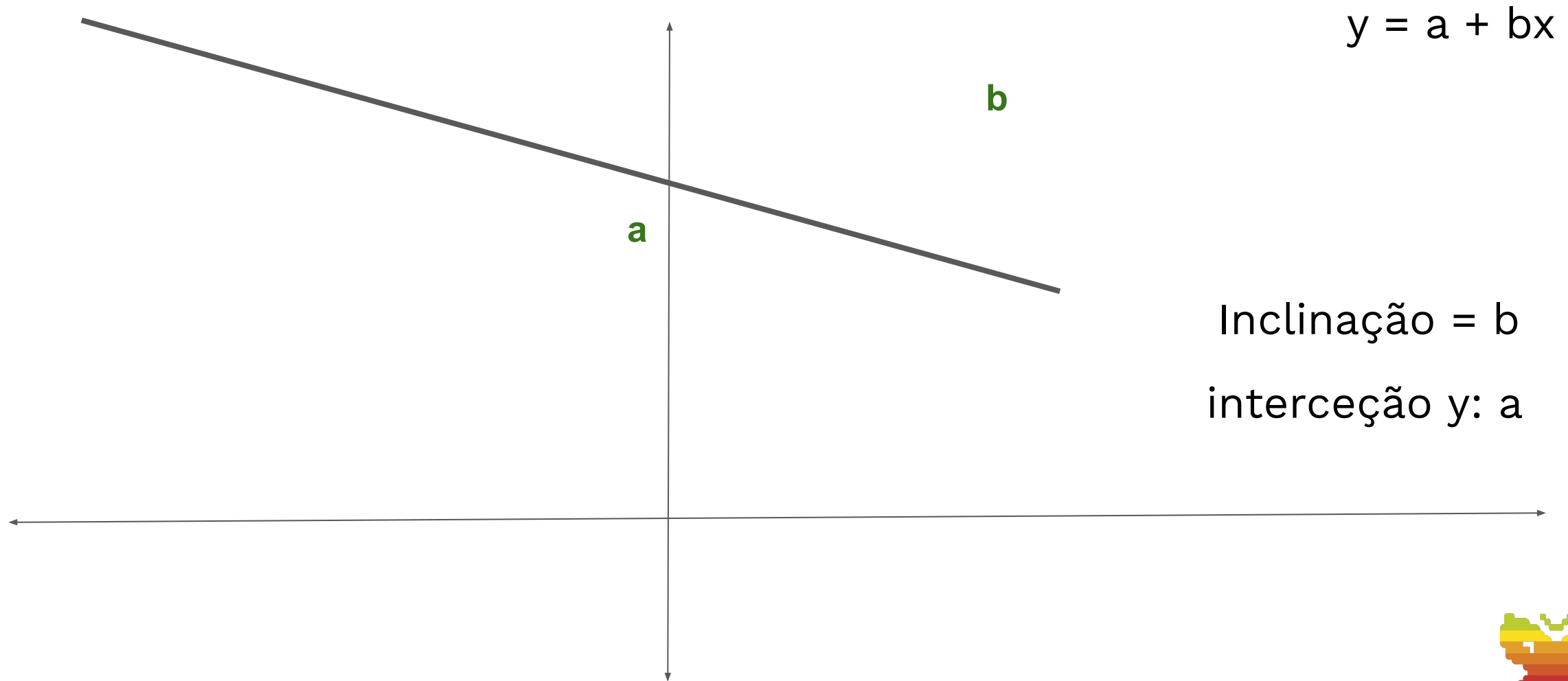
$$y = a + bx$$



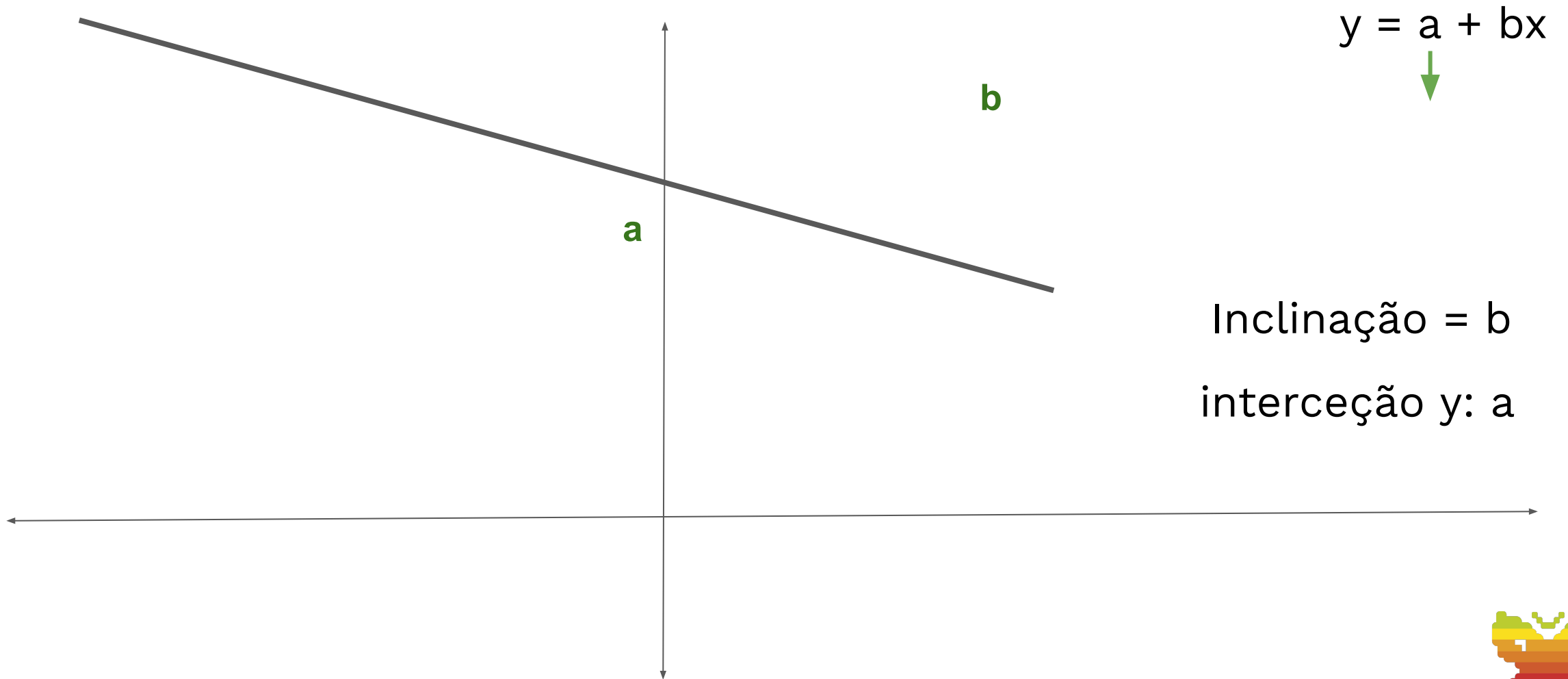
Inclinação =  $b$   
interceção  $y$ :  $a$



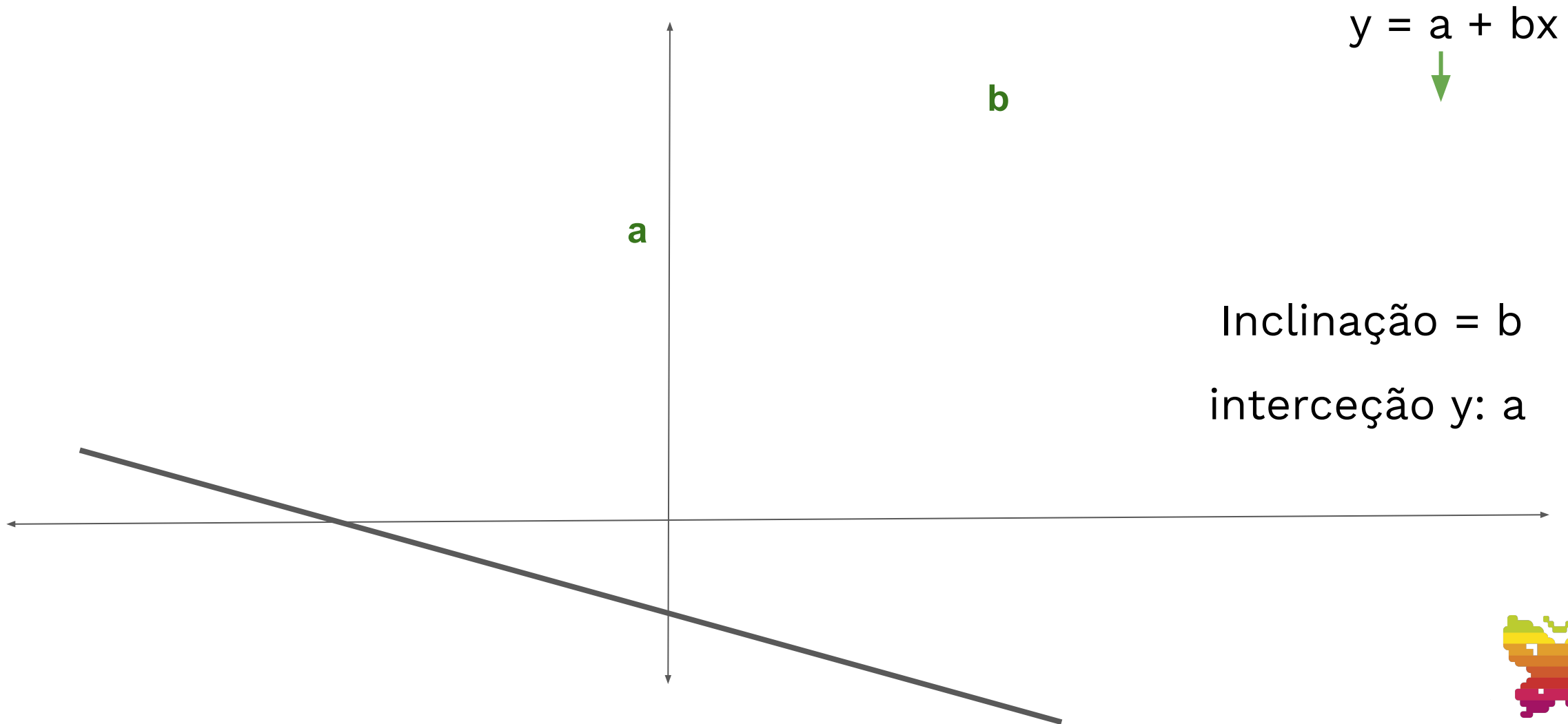
# A linha se move de forma paralela



# O que acontece se diminuirmos o $a$ ?



# A linha desce

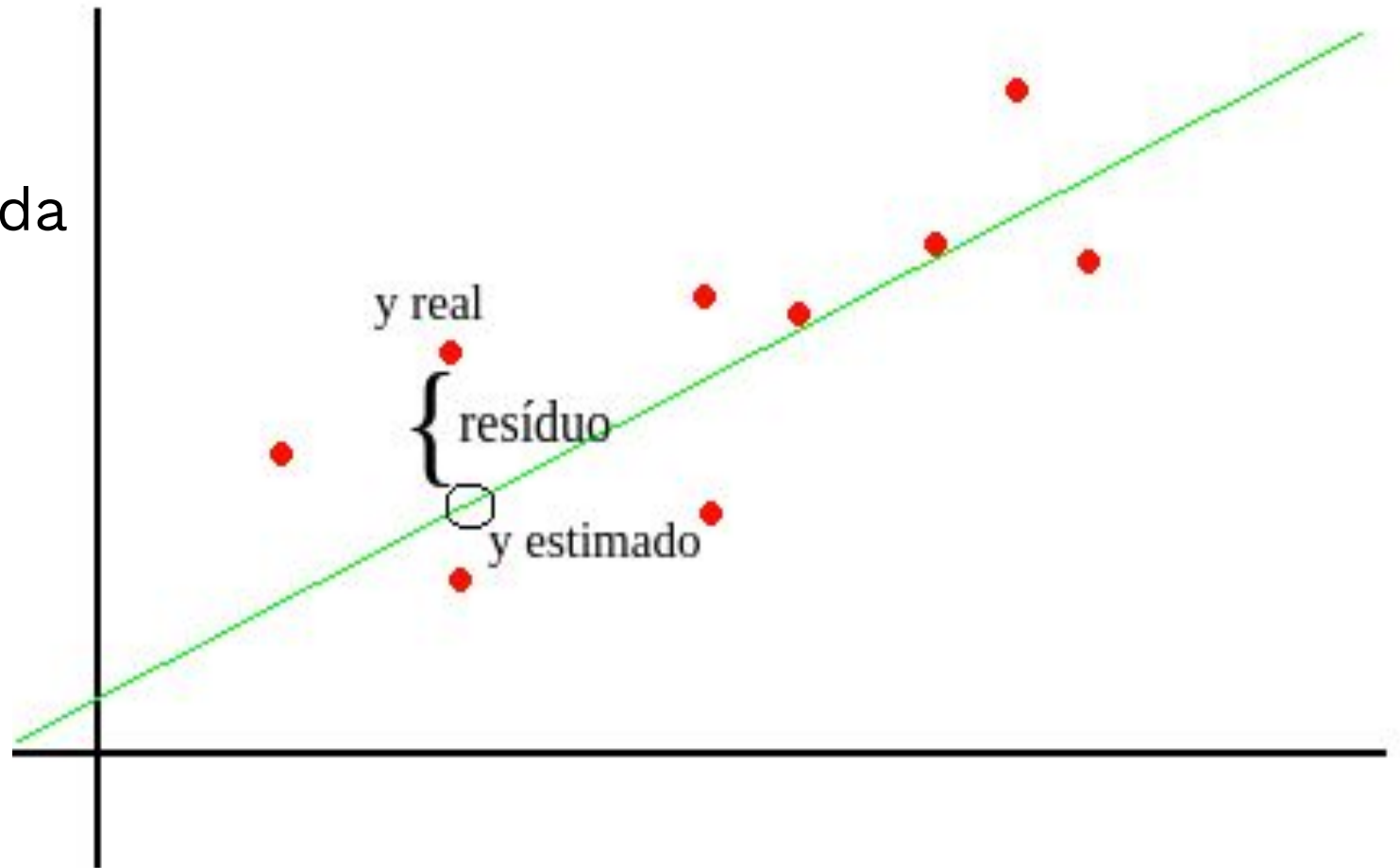


# Como encontrar melhores os coeficientes?

$y = a + bx$  estimados de cada ponto

$Y \text{ (real)} - y \text{ estimado} \rightarrow$   
Resíduos

Objetivo: Diminuir os  
resíduos do modelos



# Soma dos resíduos ao quadrado

Minimizar essa equação:

y(real) - y (estimado)

$$(y_1 - a - bx_1)^2 + (y_2 - a - bx_2)^2 + (y_3 - a - bx_3)^2 + \dots$$

$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$



Com isso encontramos os melhores  $\alpha$  e  $\beta$  para a relação. 👍



# Como ler os coeficientes?

lucro de um produto x horas trabalhadas

$$\text{lucro} = 20 + 300 \cdot \text{horas}$$



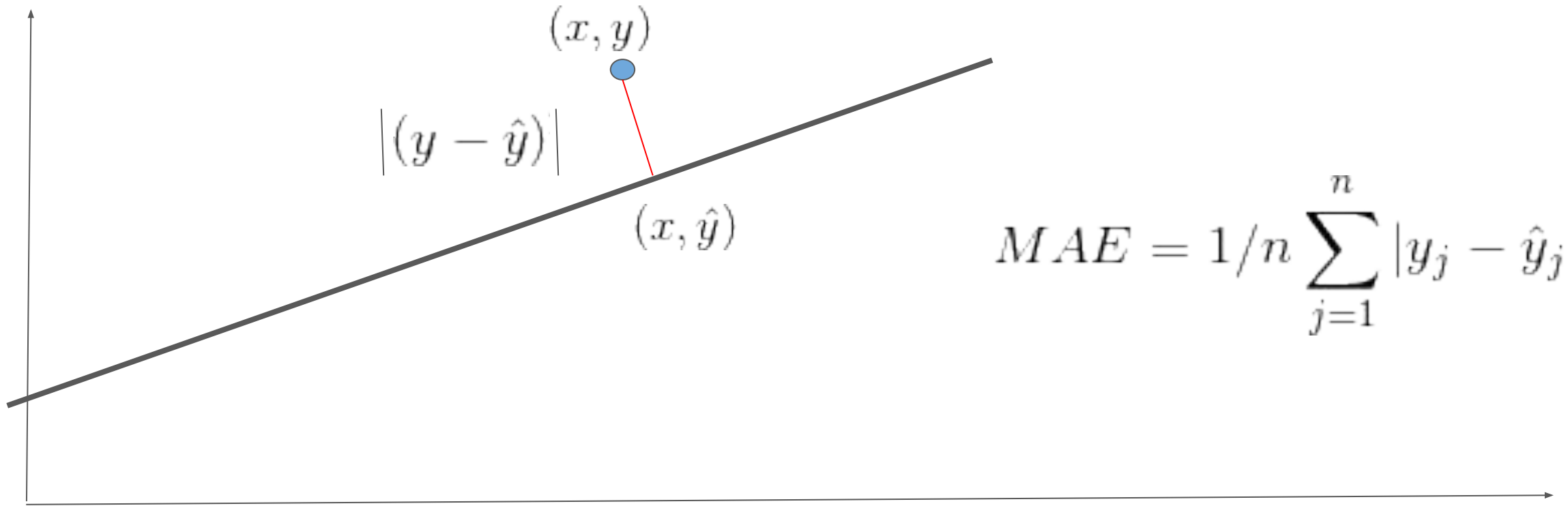
# Funções de Erro

- ▷ Mean Absolut Error
- ▷ Mean Squared Error

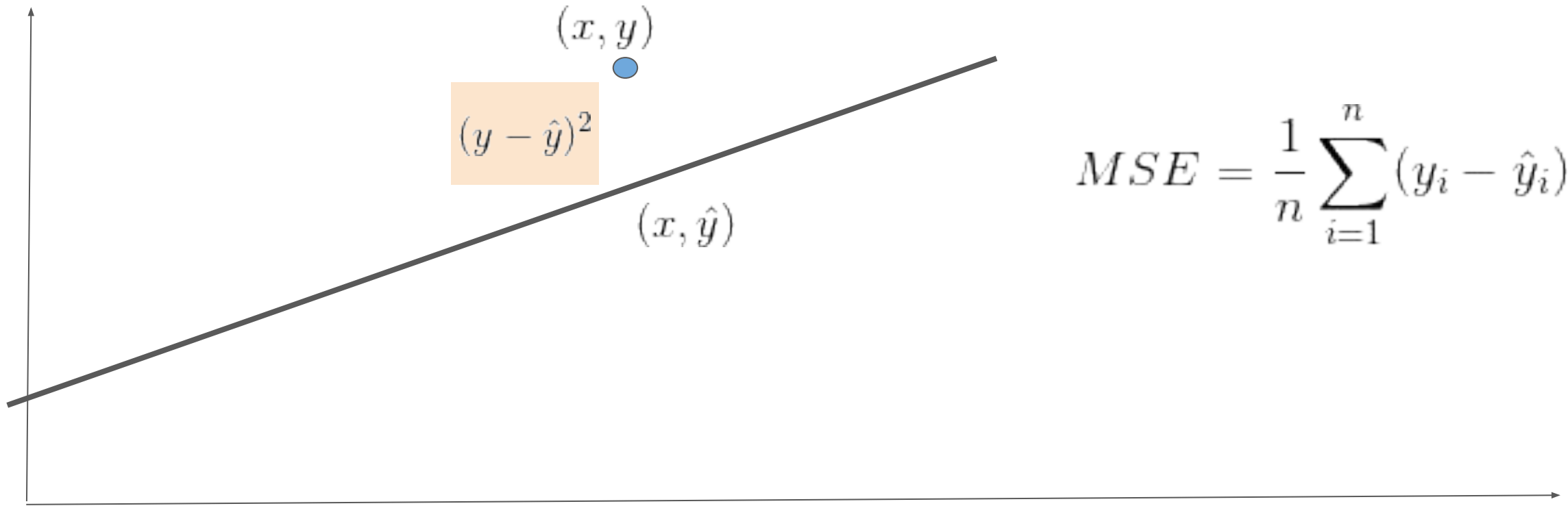




# Mean Absolute Error



# Mean Squared Error



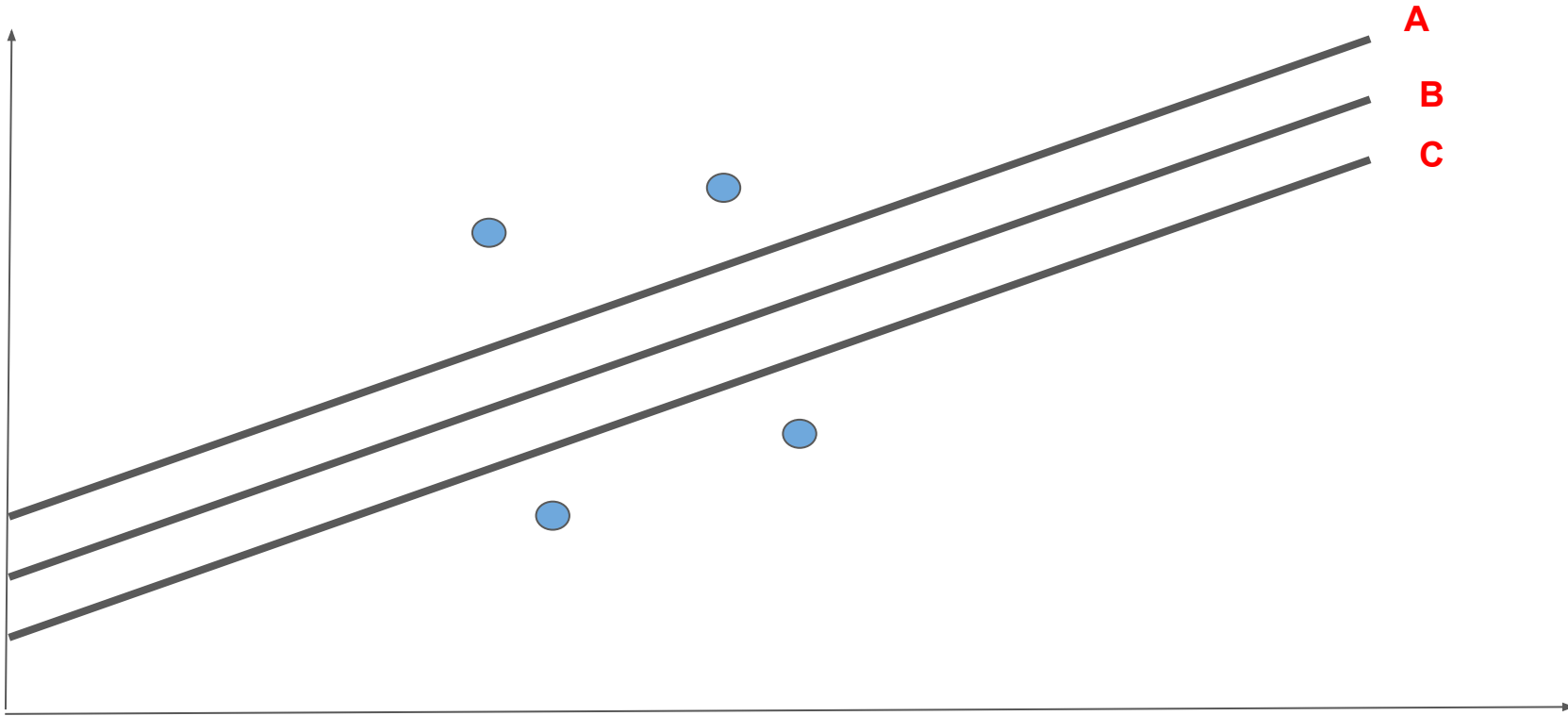
$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$



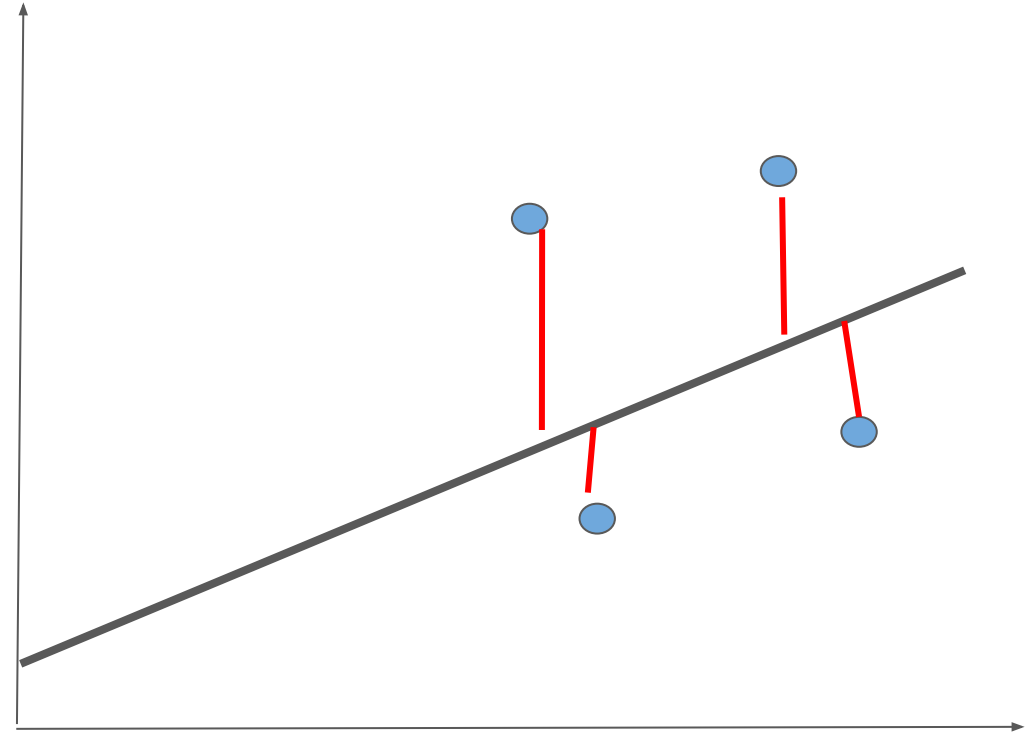
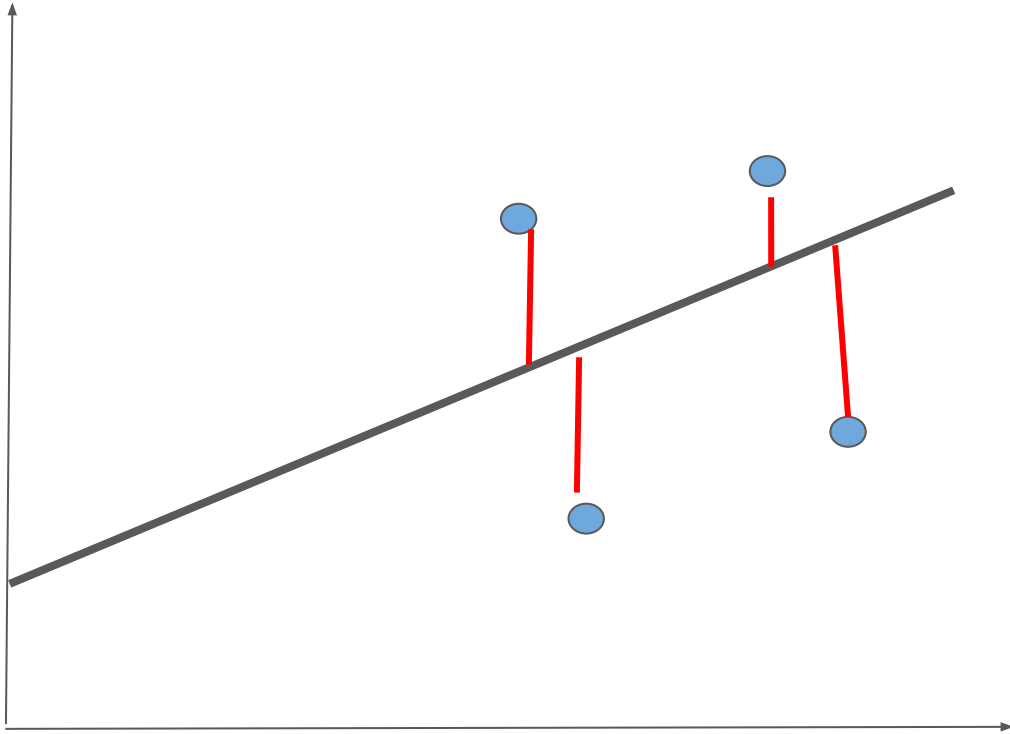
**MSE ou MAE?**



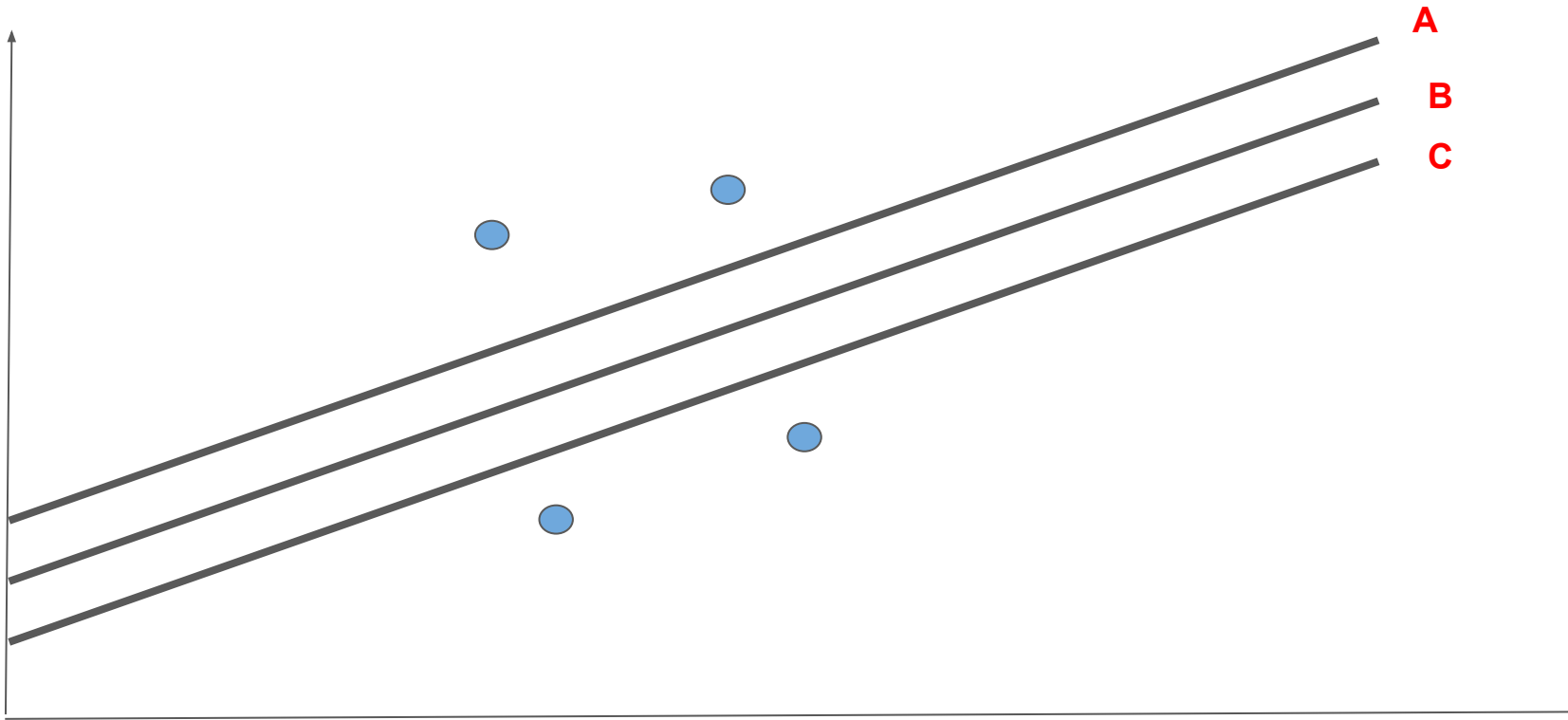
# Qual dessas linhas tem menor MAE?



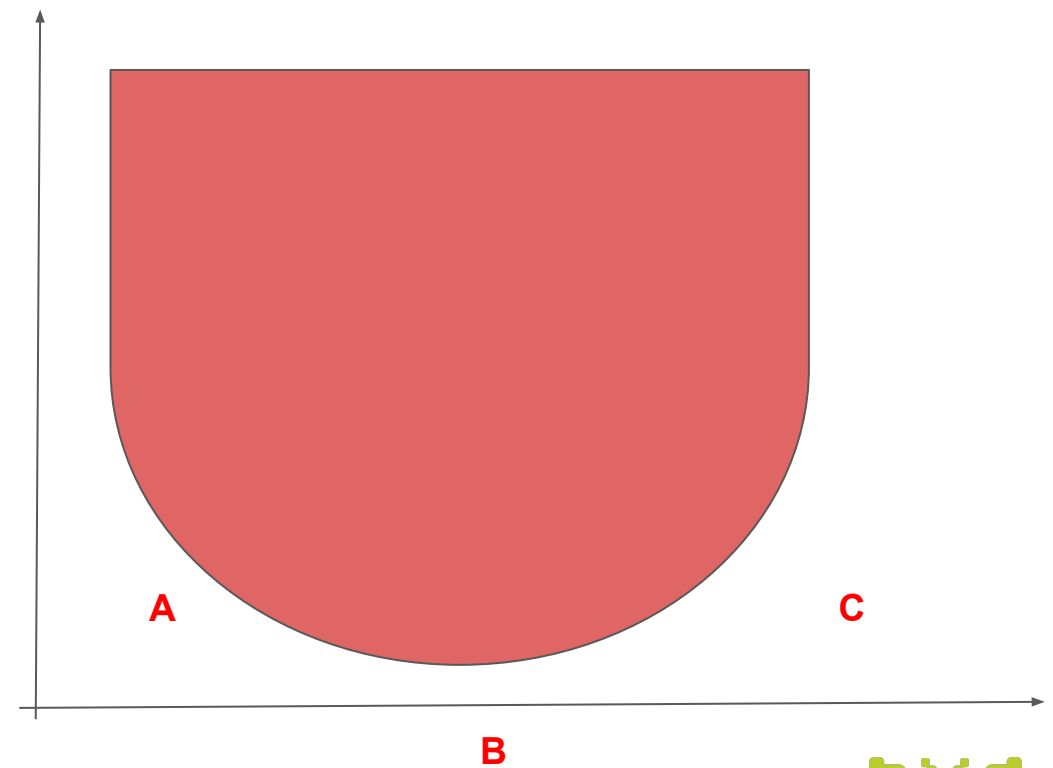
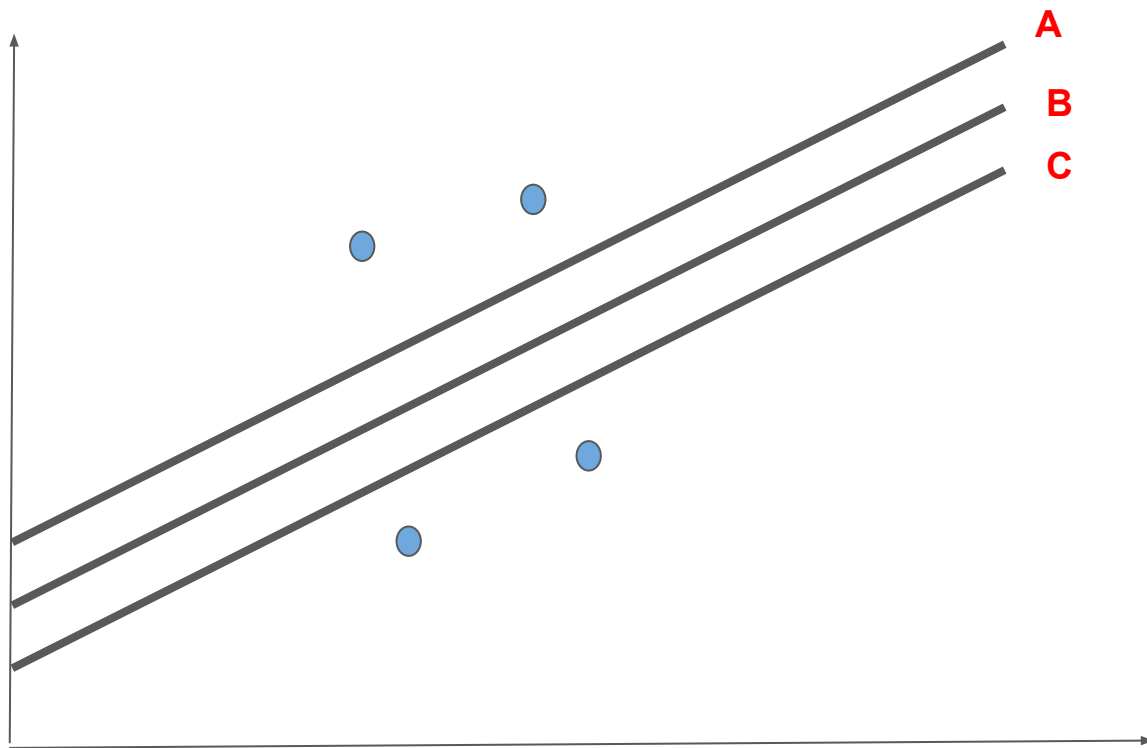
# O mesmo MAE!



# Qual dessas linhas tem menor MSE?



# A resposta é B!



# R Squared of Regression

- ▷ "Quanto que eu mudo da entrada (y) é explicado pela mudança na minha entrada (x)"

Se for esse o resultado, o modelo não está bom

$$0.0 < r^2 < 1.0$$

Quanto mais perto desse resultado, melhor seu modelo





A woman with long dark hair is sitting at a desk in an office, working on a laptop. She is looking at the screen with her hand on the trackpad. In the background, another person is visible, also working at a desk. The scene is dimly lit, with light coming from the windows. The text "Vamos à pratica!!!" is overlaid on the image in white.

Vamos à pratica!!!

# Biblioteca Python para ML



[https://scikit-learn.org/stable/supervised\\_learning.html](https://scikit-learn.org/stable/supervised_learning.html)

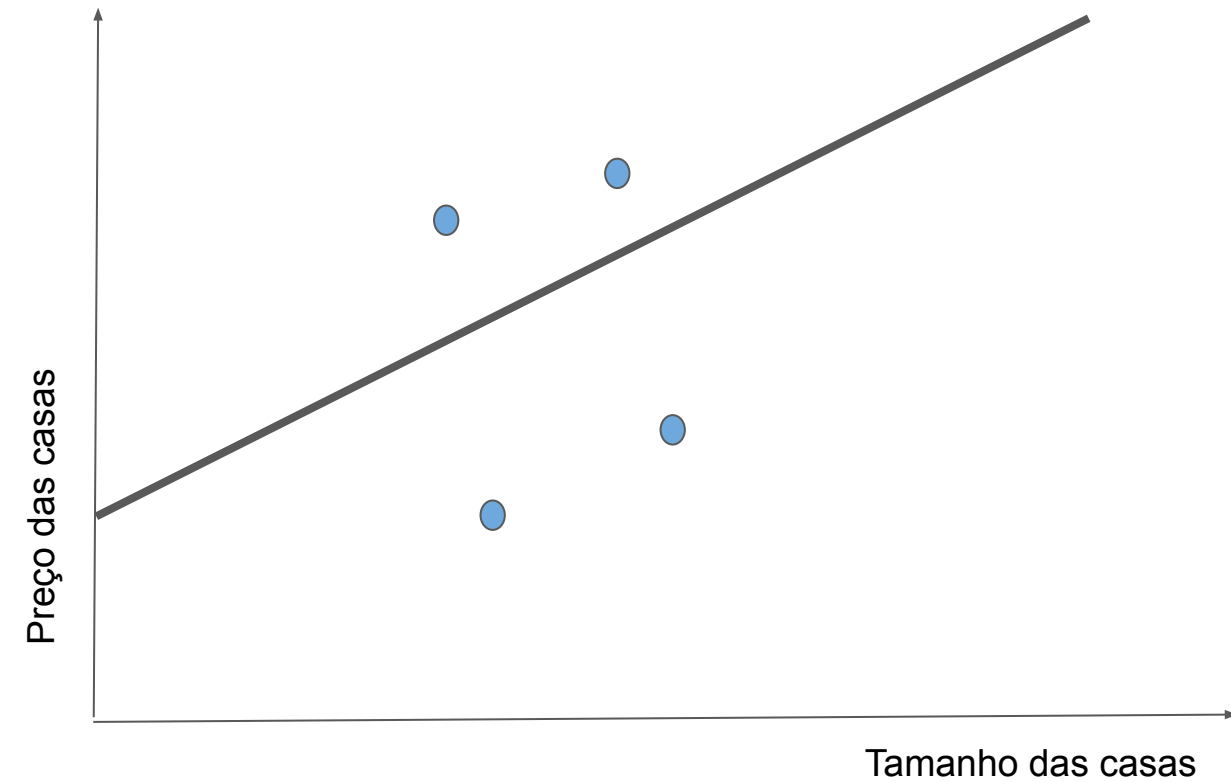
[https://scikit-learn.org/stable/modules/generated/sklearn.linear\\_model.LinearRegression.html](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html)



A woman with long dark hair is sitting at a desk in an office, working on a laptop. She is looking at the screen with her hand on the trackpad. In the background, another person is visible, also working at a desk. The office environment is dimly lit, with computer monitors and various office supplies visible on the desks.

Vamos à pratica!!!

# Regressão Linear Múltipla

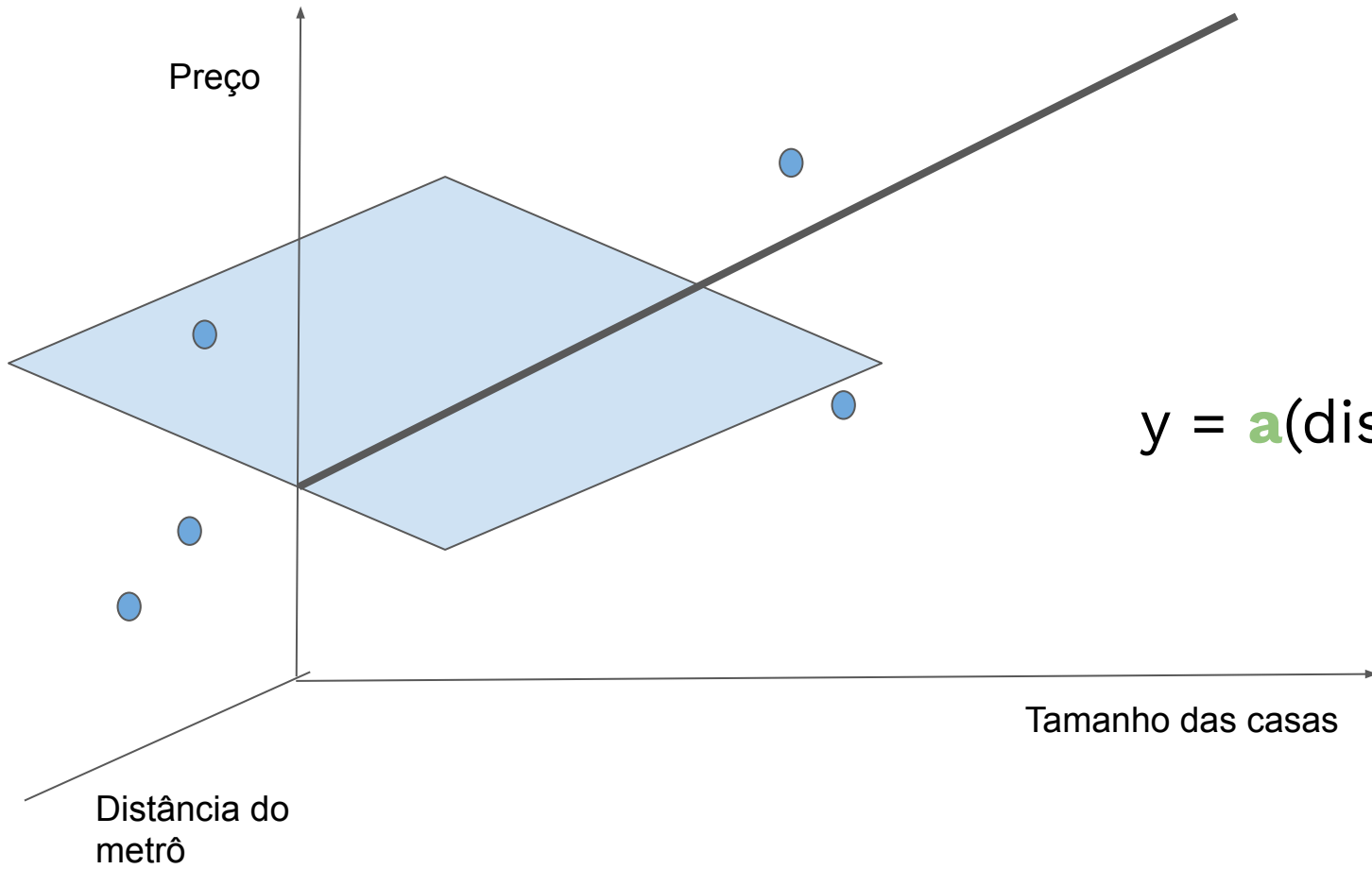


Predição: Linha

$$y = \mathbf{a} + \mathbf{b}(\text{tamanho})$$



# Regressão Linear Múltipla



Predição: Plano

$$y = \mathbf{a}(\text{distância metrô}) + \mathbf{b}(\text{tamanho}) + z$$



# Regressão Linear Múltipla

**n** dimensões

$x_1, x_2, \dots, x_{n-1}$

|        | Tamanho | Distância Metro | ... | N Quartos | Preço    |
|--------|---------|-----------------|-----|-----------|----------|
| Casa 1 | 900     | 600             | ... | 2         | R\$ 100k |
| Casa 2 | 560     | 400             | ... | 1         | R\$50k   |
| ...    | ...     | ...             | ... | ...       |          |
| Casa n | 2000    | 700             | ... | 4         | \$250k   |

$$y = a + b_1x_1 + b_2x_2 + \dots + b_nx_n$$

← **n** columnas →



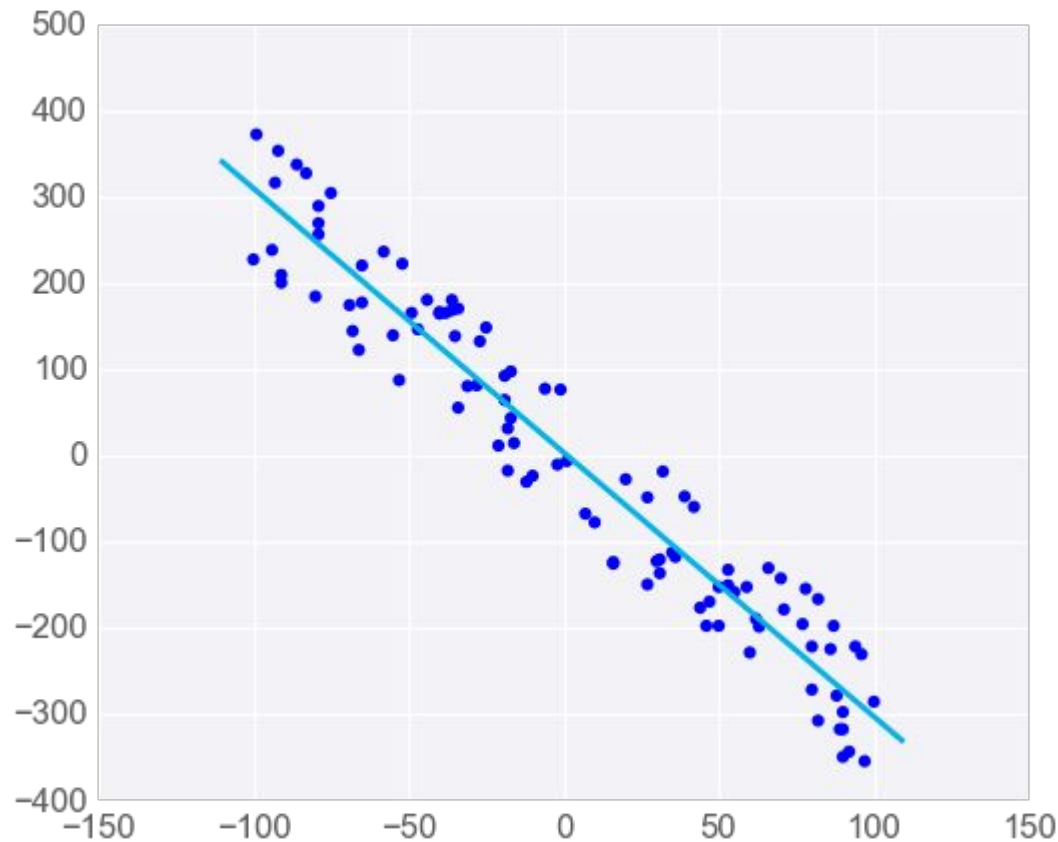


A woman with long dark hair is sitting at a desk in an office, working on a laptop. She is looking at the screen with her hand on the trackpad. In the background, another person is visible, also working at a desk. The office environment is dimly lit, with computer monitors and various office supplies visible on the desks.

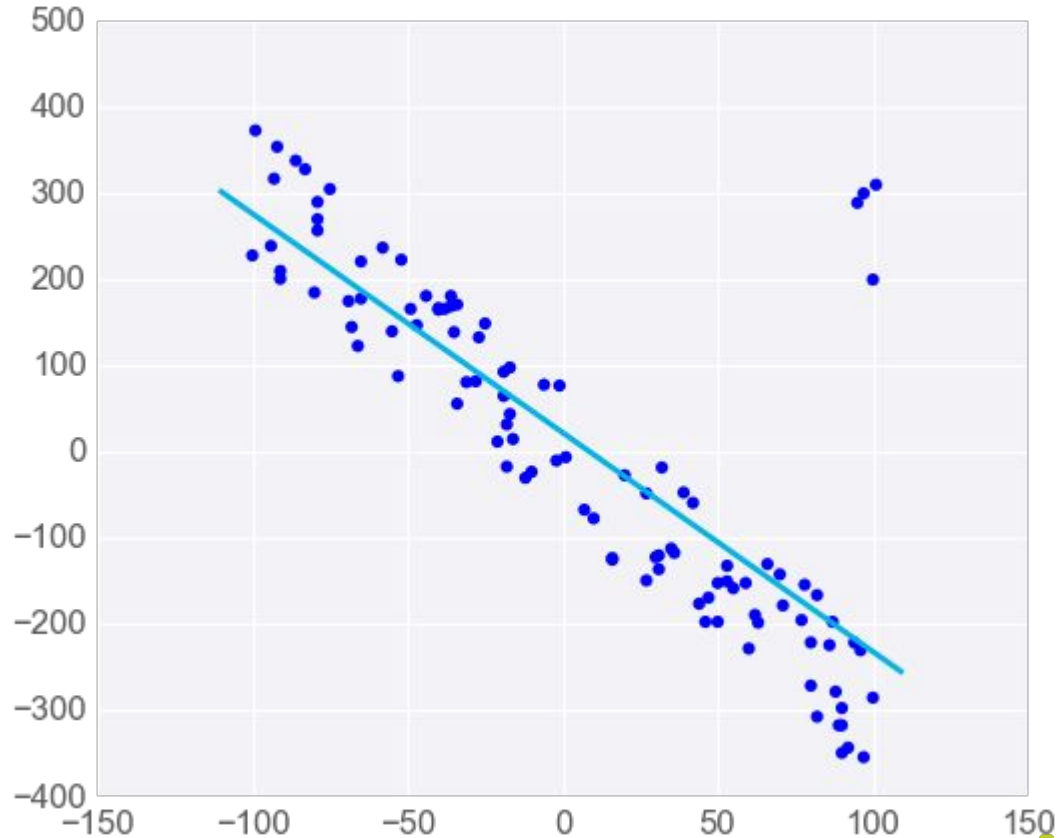
Vamos à pratica!!!

# Quando usar Regressão Linear?

Regressão linear é sensível a Outliers



Sem outliers



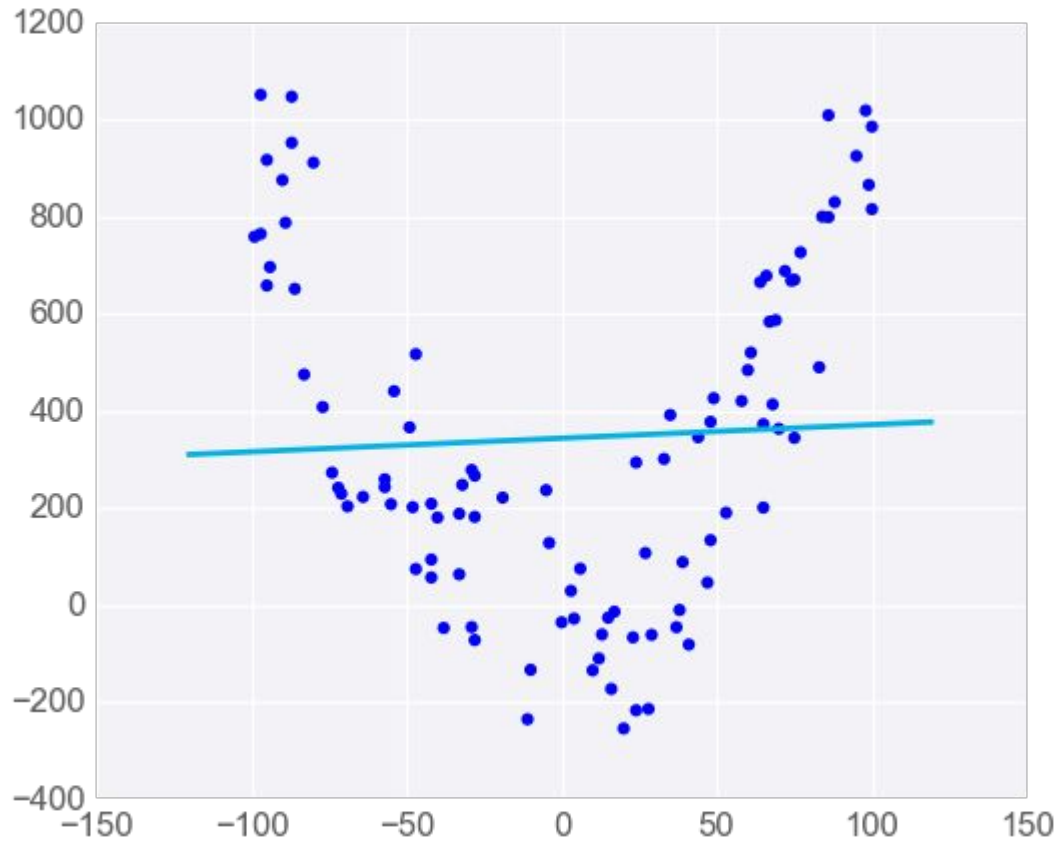
Com outliers



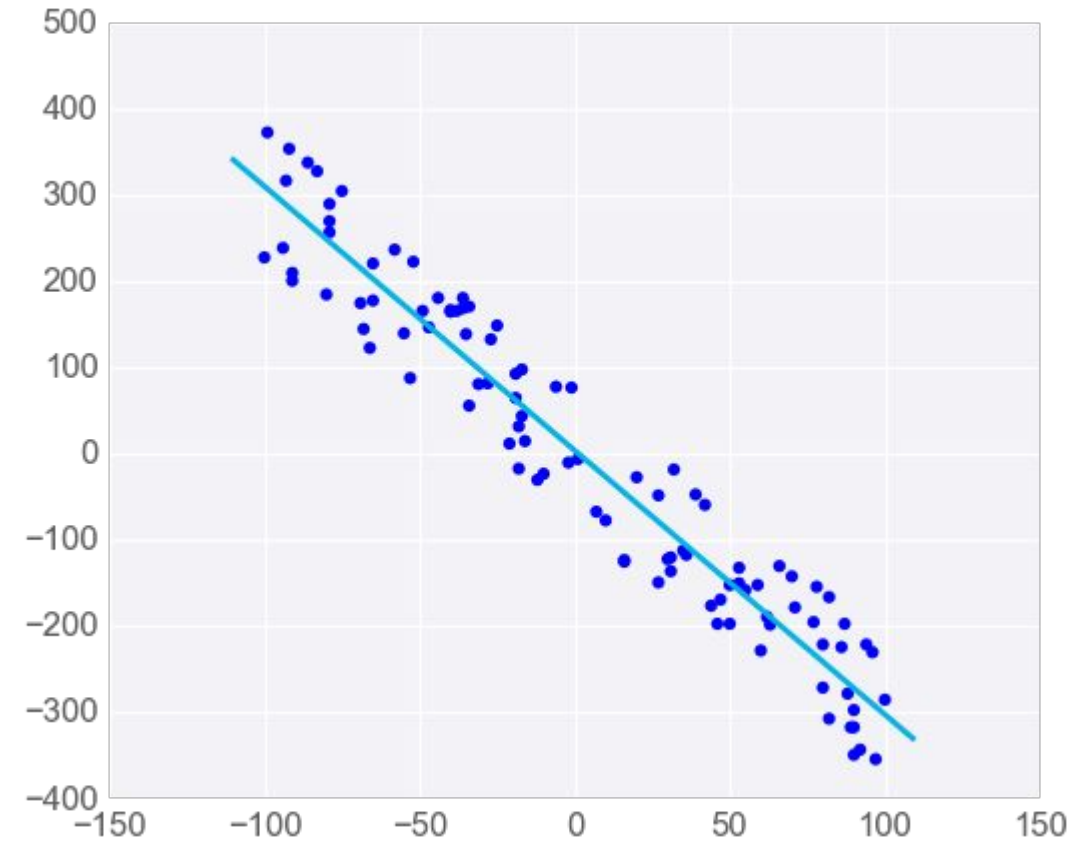


# Quando usar Regressão Linear?

Regressão linear funciona melhor quando os dados são lineares



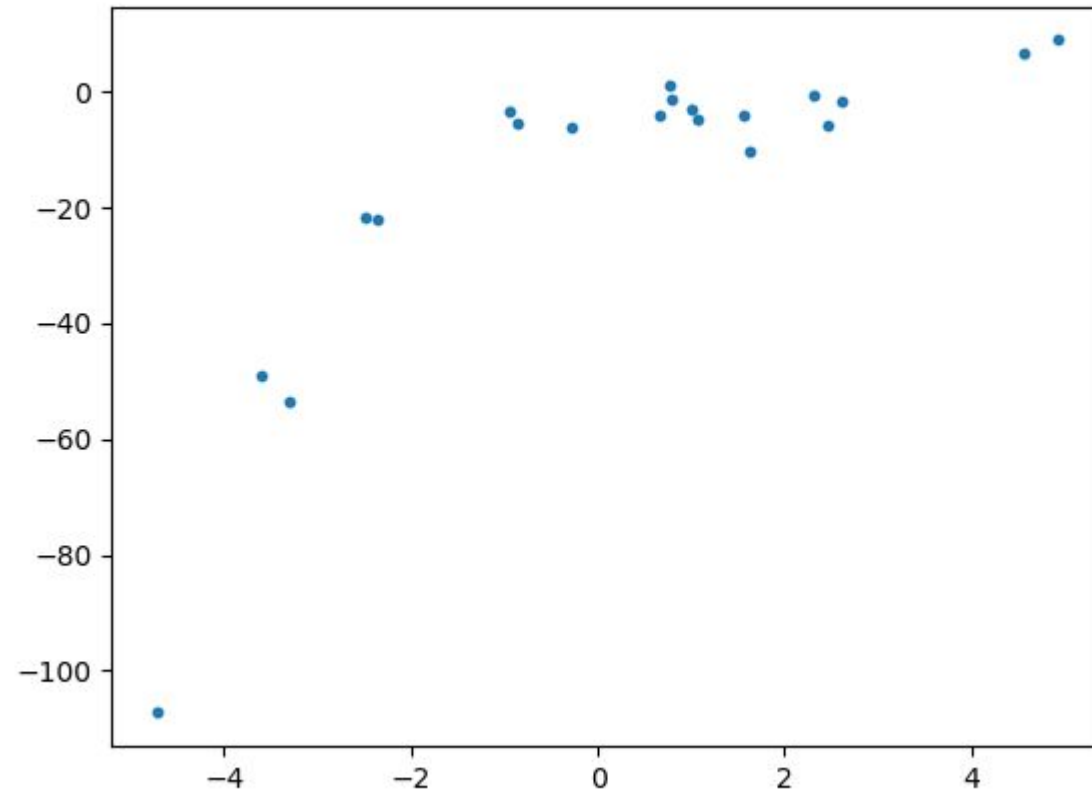
Exemplo dados não lineares



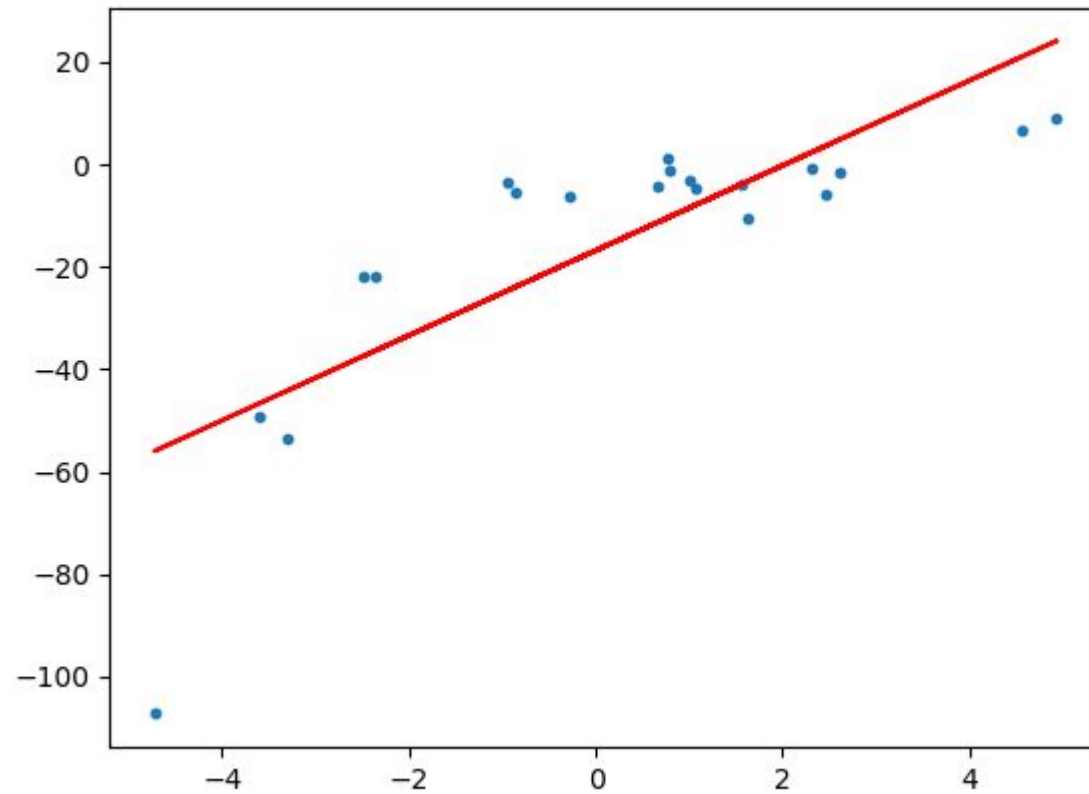
Exemplo dados lineares



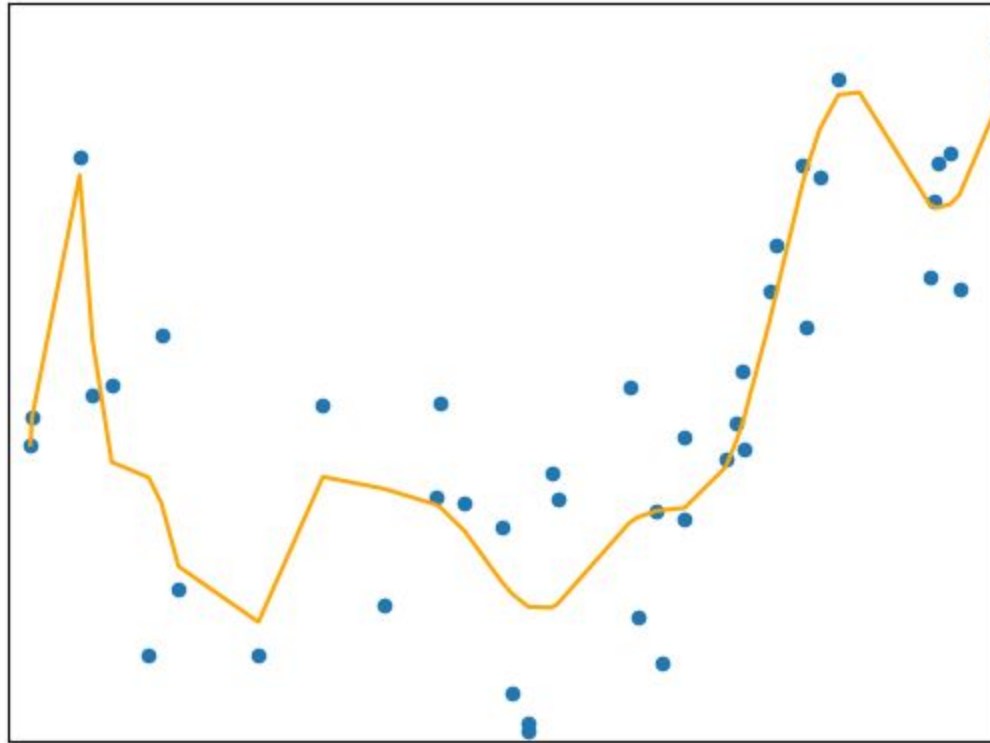
# E quando os dados se parecem com isso?



# Ajustando linha....



# Regressão Polinomial



$$\hat{y} = 2x^3 - 8x^2 - 5x + 4$$



A woman with long dark hair is sitting at a desk in an office, working on a silver laptop. She is looking at the screen, which is partially visible on the left. Her right hand is on the trackpad. In the background, another person is blurred, also working at a desk. The office environment includes various cables and equipment on the desk. The overall lighting is dim, with a focus on the woman and her work area.

Vamos à pratica!!!

# Feature Scaling (Transformações)

Feature Scaling é uma maneira de transformar seus dados em um intervalo comum de valores. Existem duas escalas comuns mais usadas:

- 1 - Normalização
- 2 - Estandardização



# Estandartização

A padronização é realizada utilizando cada valor de sua coluna, subtraindo a média da coluna e depois dividindo pelo desvio padrão da coluna. No python seria algo como:

```
df["altura"] = (df["altura"] - df["altura"].mean()) / df["altura"].std()
```



# Normalização

Com a normalização, os dados são redimensionados entre 0 e 1. Usando o mesmo exemplo anterior, podemos executar a normalização no Python da seguinte maneira:



```
df["altura_normal"] = (df["altura"] - df["altura"].min()) /  
(df["altura"].max() - df["altura"].min())
```





# Parametrização

Premissas que os modelos assumem para facilitar o ajuste.

- Mais rápidos
- Menos dados 
- Limitados 

Regressão linear:

- Variáveis independentes são não correlacionadas entre si.
- Distribuição normal



A woman with long dark hair is sitting at a desk in an office, working on a laptop. She is looking at the screen with her hand on the trackpad. In the background, another person is visible, also working at a desk. The office environment is dimly lit, with computer monitors and various office supplies visible on the desks.

Vamos à pratica!!!

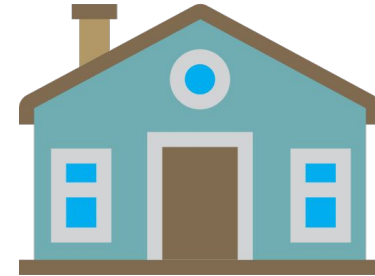
# Previsões Séries Temporais

Jan/2019



R\$ 120.00,00

Jan/2025



?



# SÉRIES TEMPORAIS

OBSERVAÇÕES PASSADAS



FUTURO

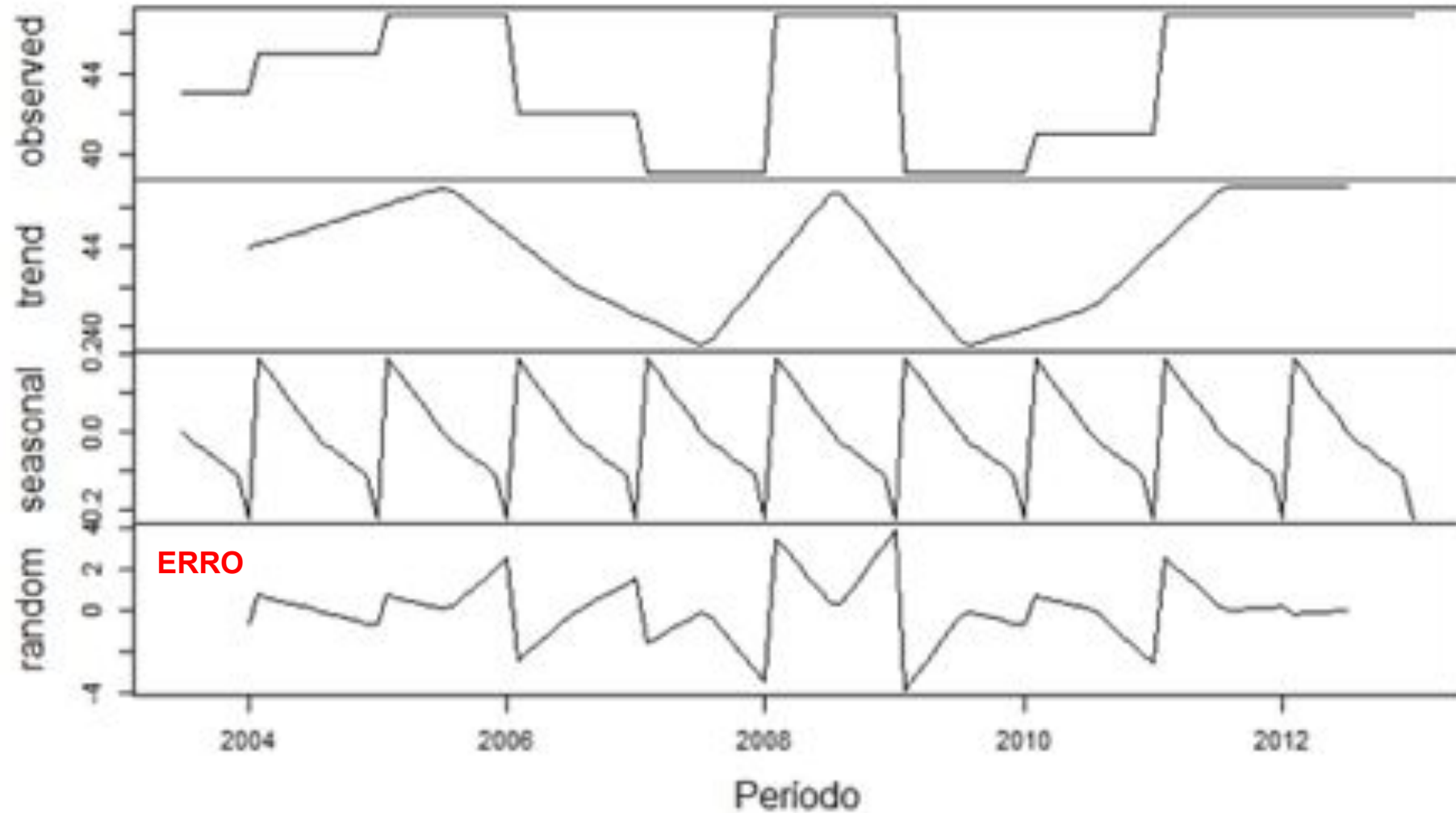


# Previsões Séries Temporais

- É sobre um intervalo de tempo contínuo
- Ordem Importa
- Existem medições sequenciais nesse intervalo
- Há espaçamento igual entre cada duas medições consecutivas
- Cada unidade de tempo dentro do intervalo de tempo tem no máximo um ponto de dados

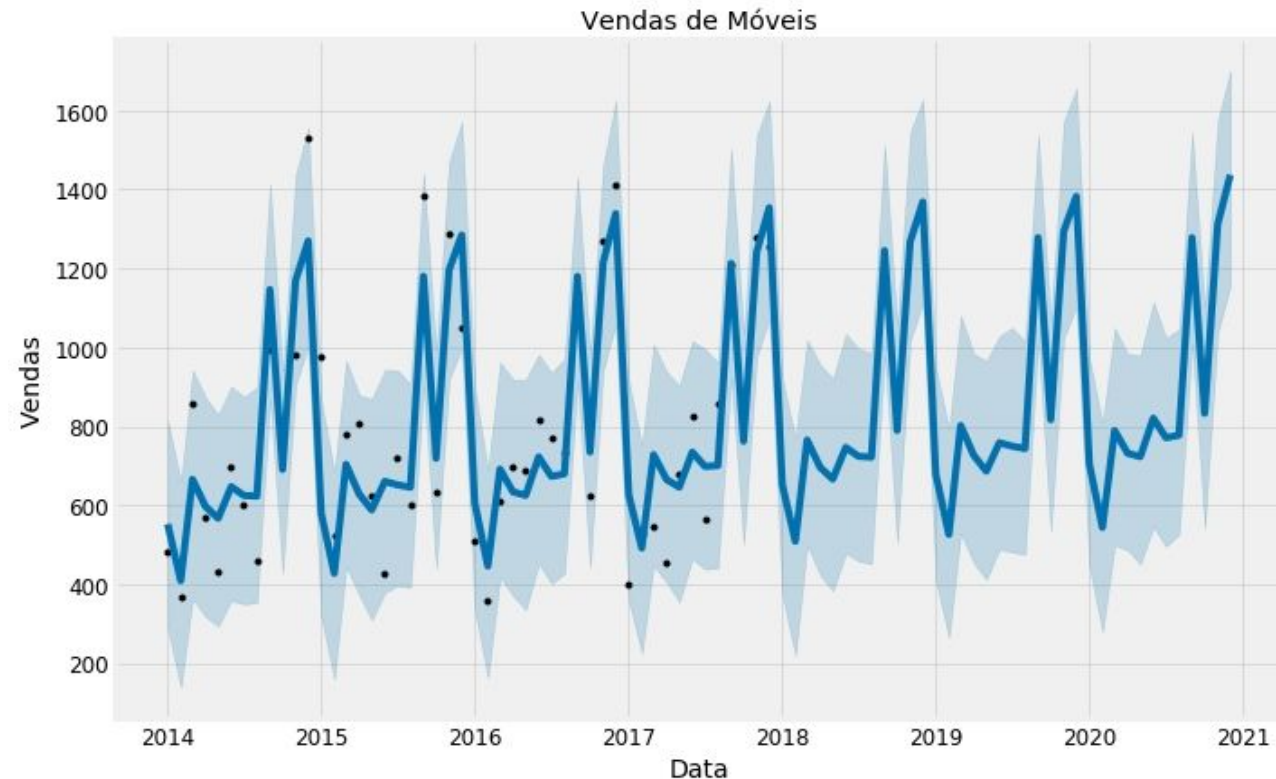


# Gráfico de Decomposição



# Como trabalhar com Séries Temporais

Identificando padrões que se repetem ao longo do tempo para prever o que vai acontecer no futuro



# Tendência





# Sazonalidade



# Padrões Cíclicos

Ibovespa  
INDEXBVMF: IBOV

103.767,42 +1.724,31 (1,69%) ↑

4 de jul 15:29 BRT · Exoneração de responsabilidade

Um dia 5 dias 1 mês Um ano 5 anos Máx



# Mercado vê crise se estender para 2017

**Ainda vai piorar**  
Projeções para a economia mostram que quadro recessivo será mais longo que o esperado

Veja as estimativas de especialistas

|                                    | 2016  | 2017  |
|------------------------------------|-------|-------|
| PB (variação %)                    | -2,59 | 0,86  |
| Produção Industrial                | -3,45 | 1,98  |
| Inflação (em %)                    | 6,93  | 5,20  |
| Dólar (R\$/US\$)                   | 4,25  | 4,23  |
| Taxa básica de juros (em % ao ano) | 15,25 | 12,75 |

**Bolsa em queda livre**

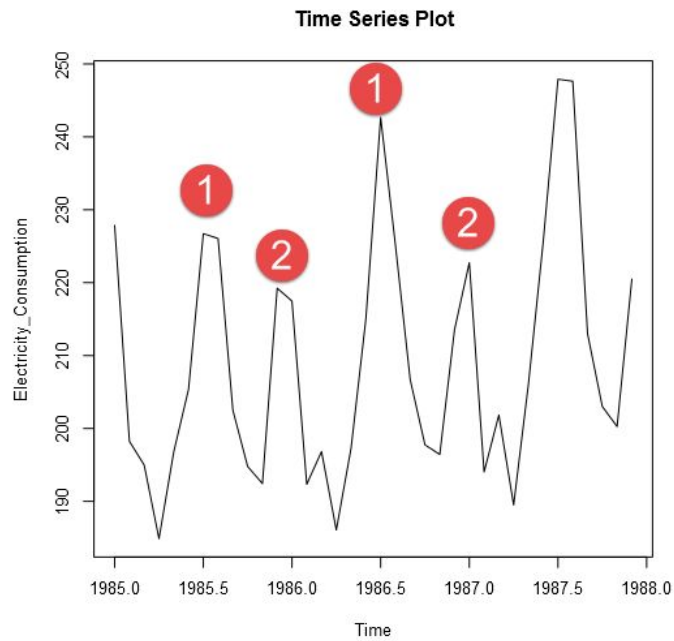
Novo deterioração das expectativas da economia, a sombra da desconfiança da China e a sinalização de novo aumento de juros nos Estados Unidos fizeram o Ibovespa, principal indicador da Bolsa de Valores de São Paulo (BM&FBOVESPA), romper o piso dos 90 mil pontos, logo após a abertura de 17 de março de 2016, no auge da crise mundial. No pregão de ontem, o índice recuou 1,63%, aos

# Bolsa de Nova York quebra e o mundo inteiro treme

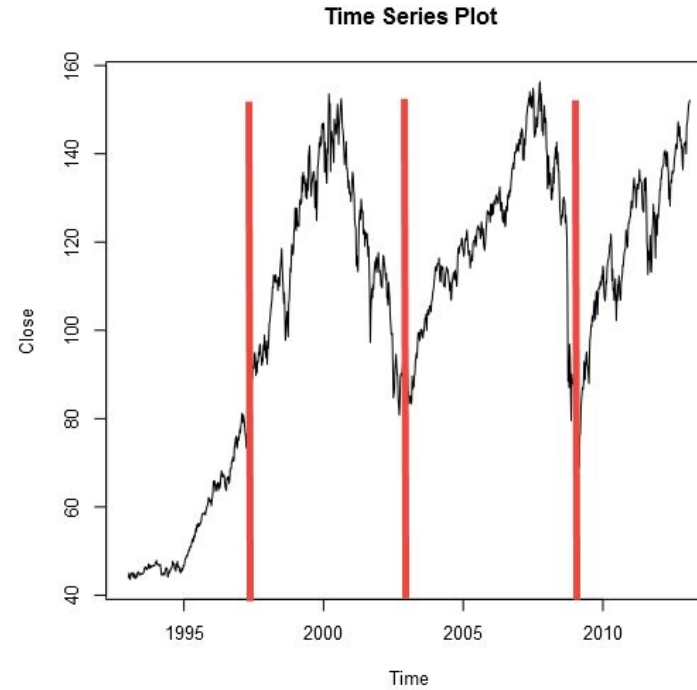
**REALIDADE CAPITALISTA**

2016 - A Bolsa de Valores de Nova York sofreu a maior queda de todos os tempos, encerrando o ano com quedas recordes. A crise de confiança internacional que se refletiu em Wall Street, afetando também a bolsa de valores brasileira, levou a uma queda de 10% no Ibovespa em 2016. A queda de 10% no Ibovespa em 2016, no auge da crise mundial, no pregão de ontem, o índice recuou 1,63%, aos

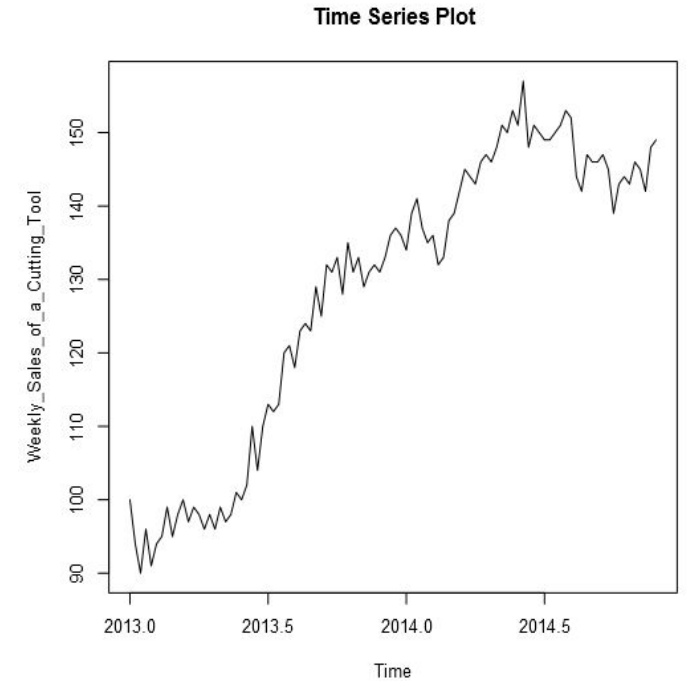
# Padrões Cíclicos vs Sazonalidade



TEMPORAL



CICLÍCO



SEM PADRÃO



# Modelos Autoregressivos

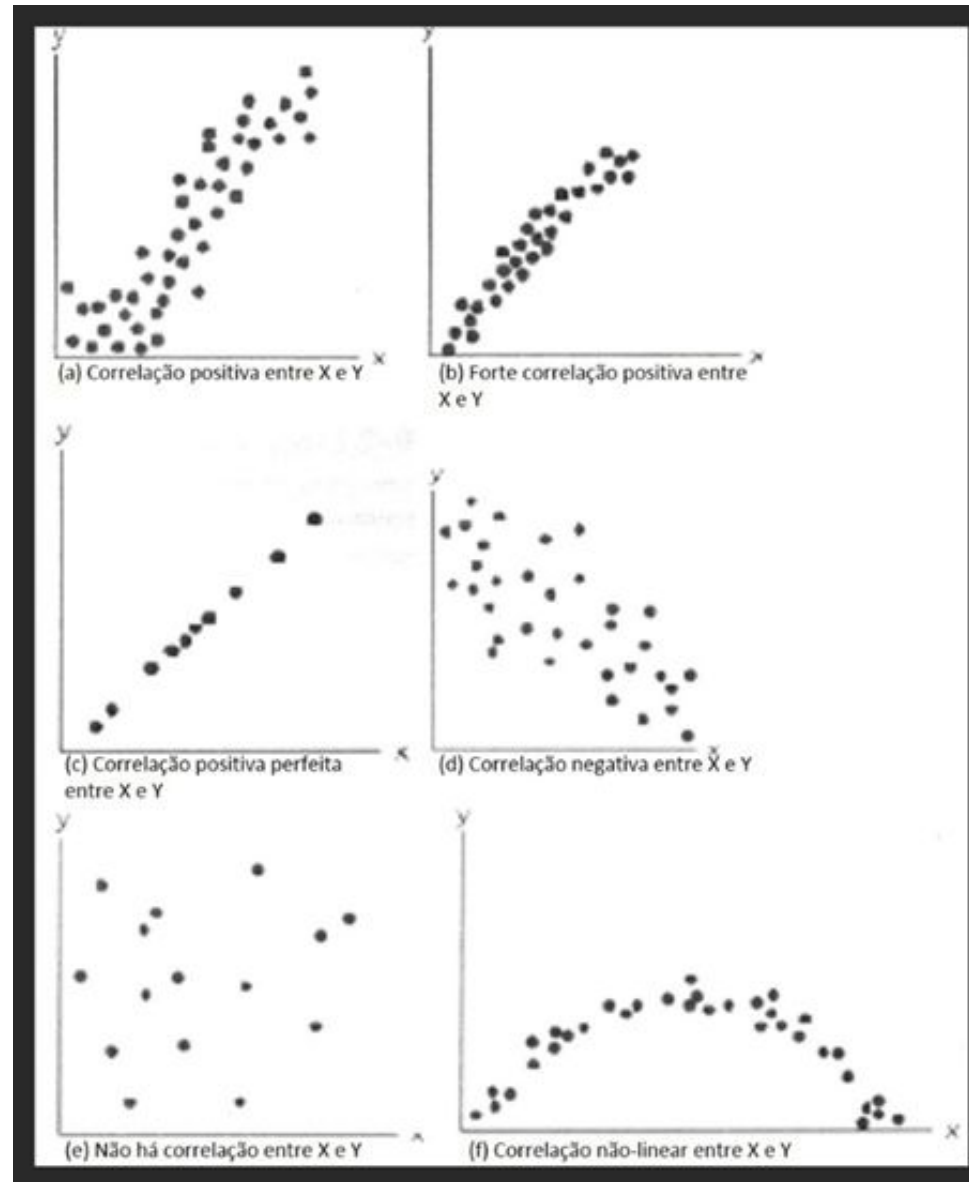
São uma regressão linear

$$y = a + bx$$

$$X(t+1) = a + b_1 * X(t-1) + b_2 * X(t-2)$$



# Correlação positiva vs negativa



# AUTOCORRELAÇÃO

Demonstra o quanto uma série está correlacionada com seus valores passados.



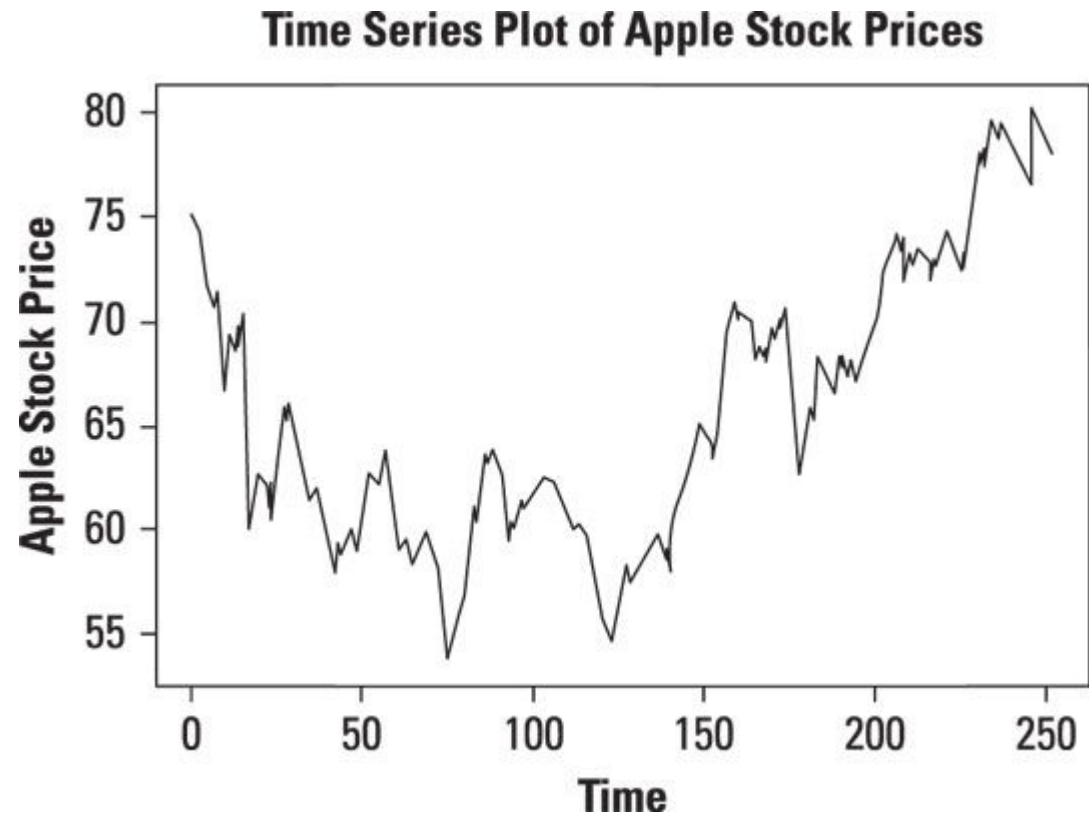
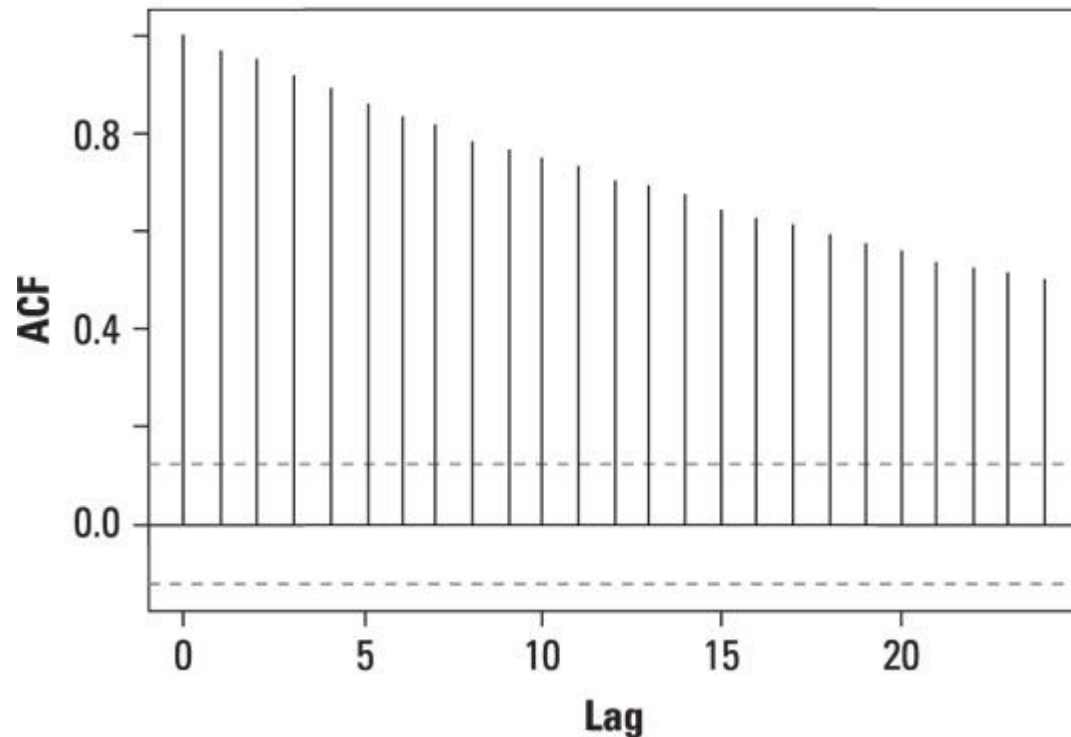
Usamos esse gráfico para ver a correlação entre os pontos, incluindo o lag





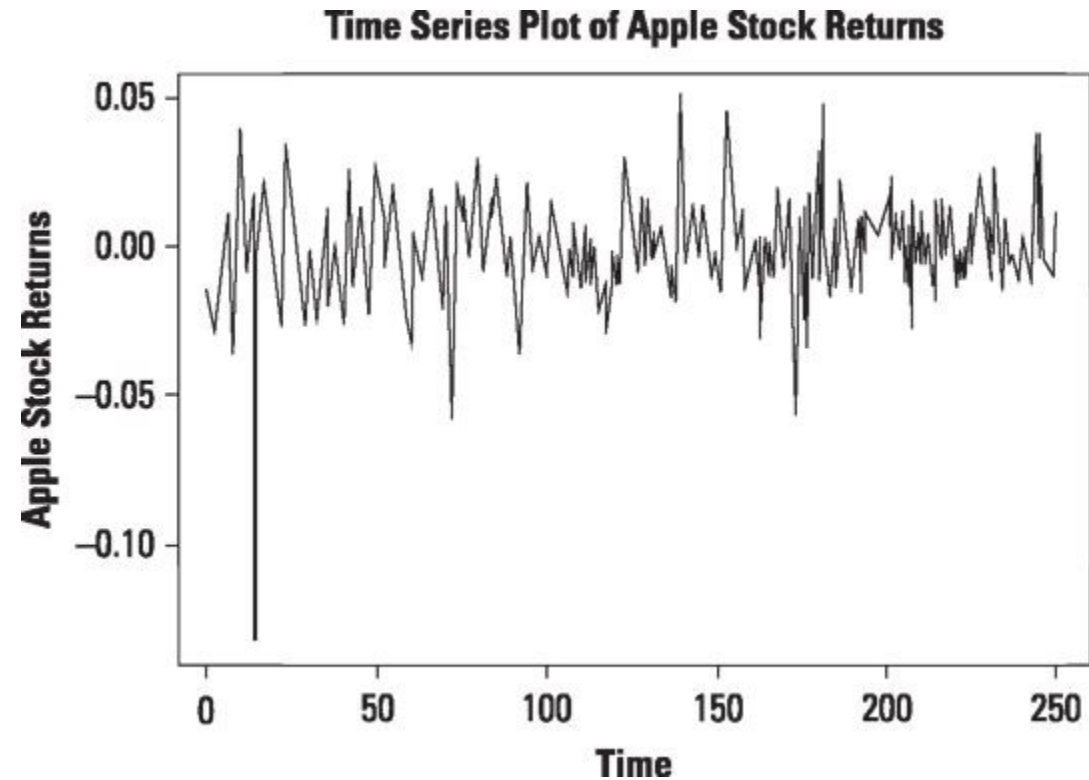
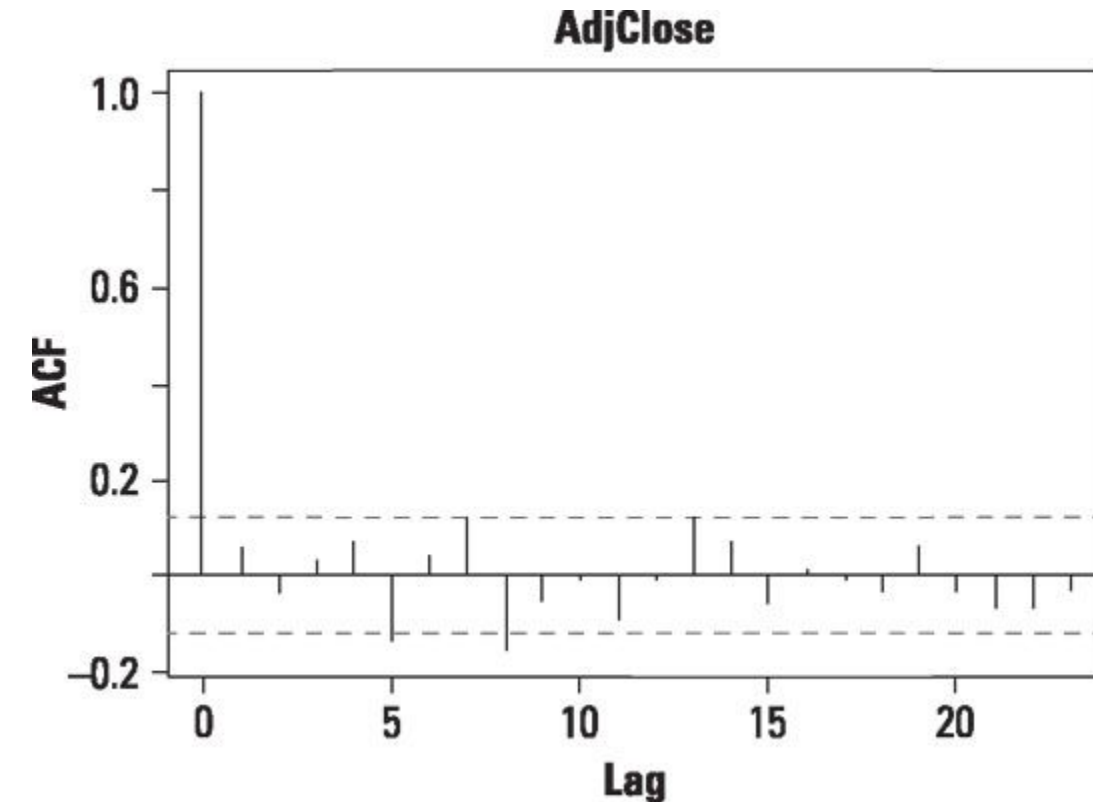
# EXEMPLO - AÇÕES APPLE

O gráfico abaixo demonstra ações da Apple de 1 de janeiro 2013 a 31 de dezembro de 2013.



# EXEMPLO - AÇÕES APPLE

O gráfico abaixo demonstra ações da Apple diariamente de 1 de janeiro 2013 a 31 de dezembro de 2013.





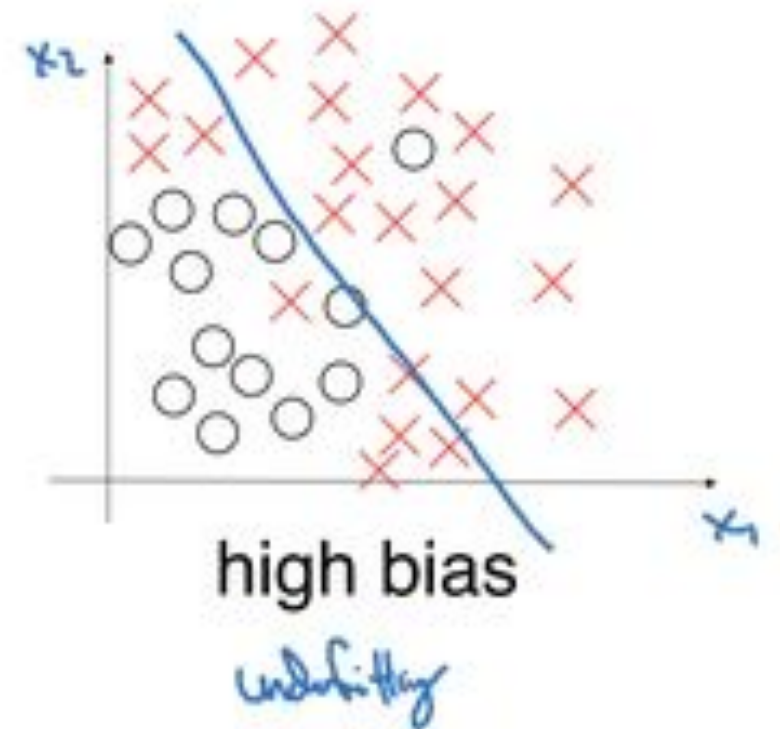
A dimly lit office scene with a woman in the foreground, seen from the side, working on a laptop. Her hand is on the trackpad. In the background, another person is blurred, also working at a desk with multiple monitors. The overall atmosphere is professional and focused.

Vamos à pratica!!!

# Viés e Variância

A incapacidade de um método de capturar a verdadeira relação entre variáveis e o objeto a ser predito é o BIAS/VIÉS.

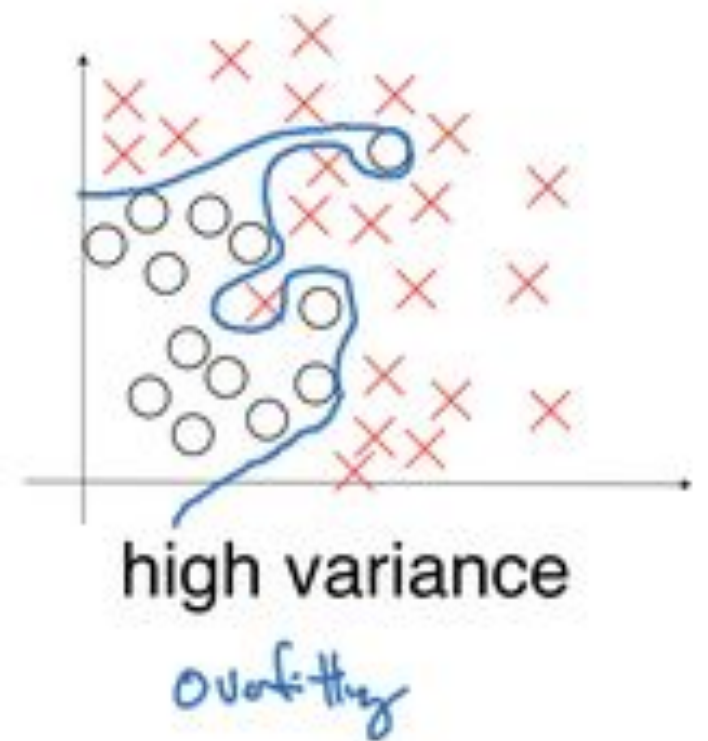
Viés alto = modelo não está aprendendo nada.



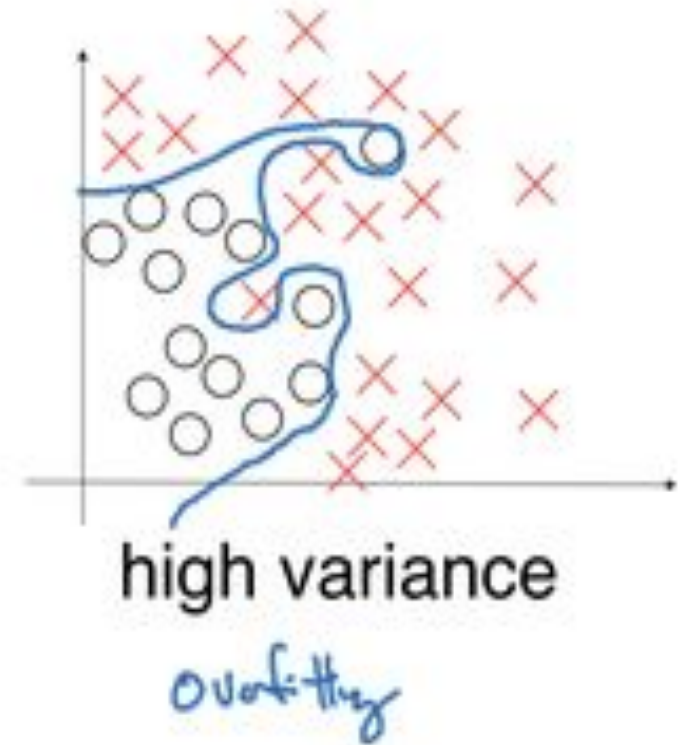
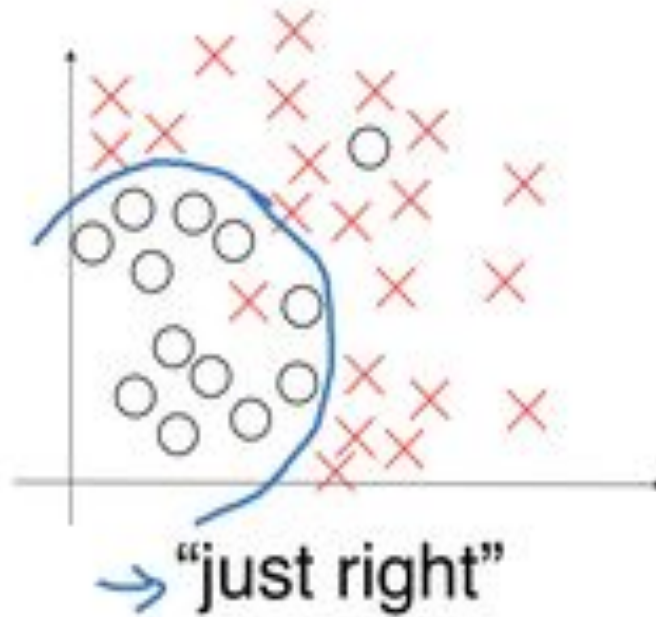
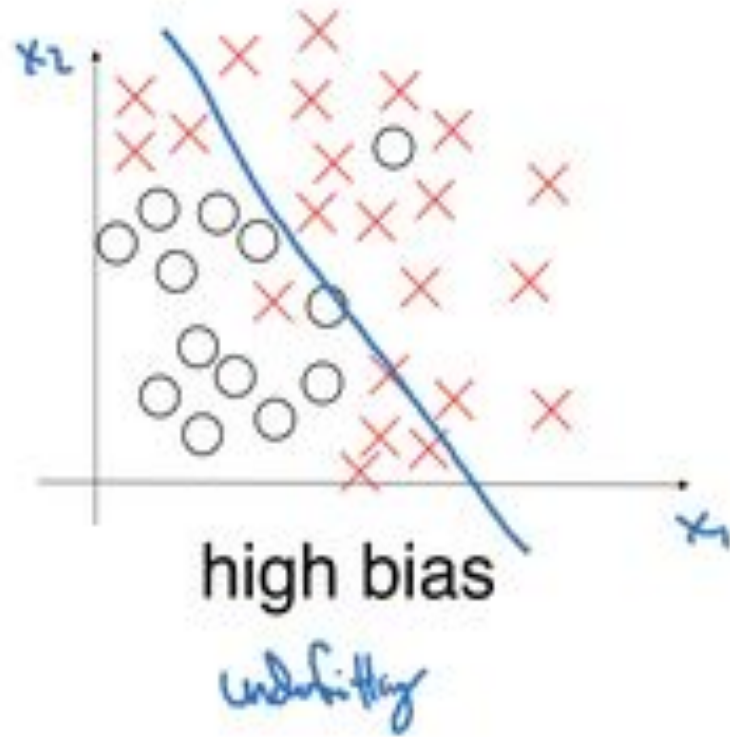
# Viés e Variância

A variância é a sensibilidade de um modelo ao ser usado com outros datasets diferentes do treinamento.

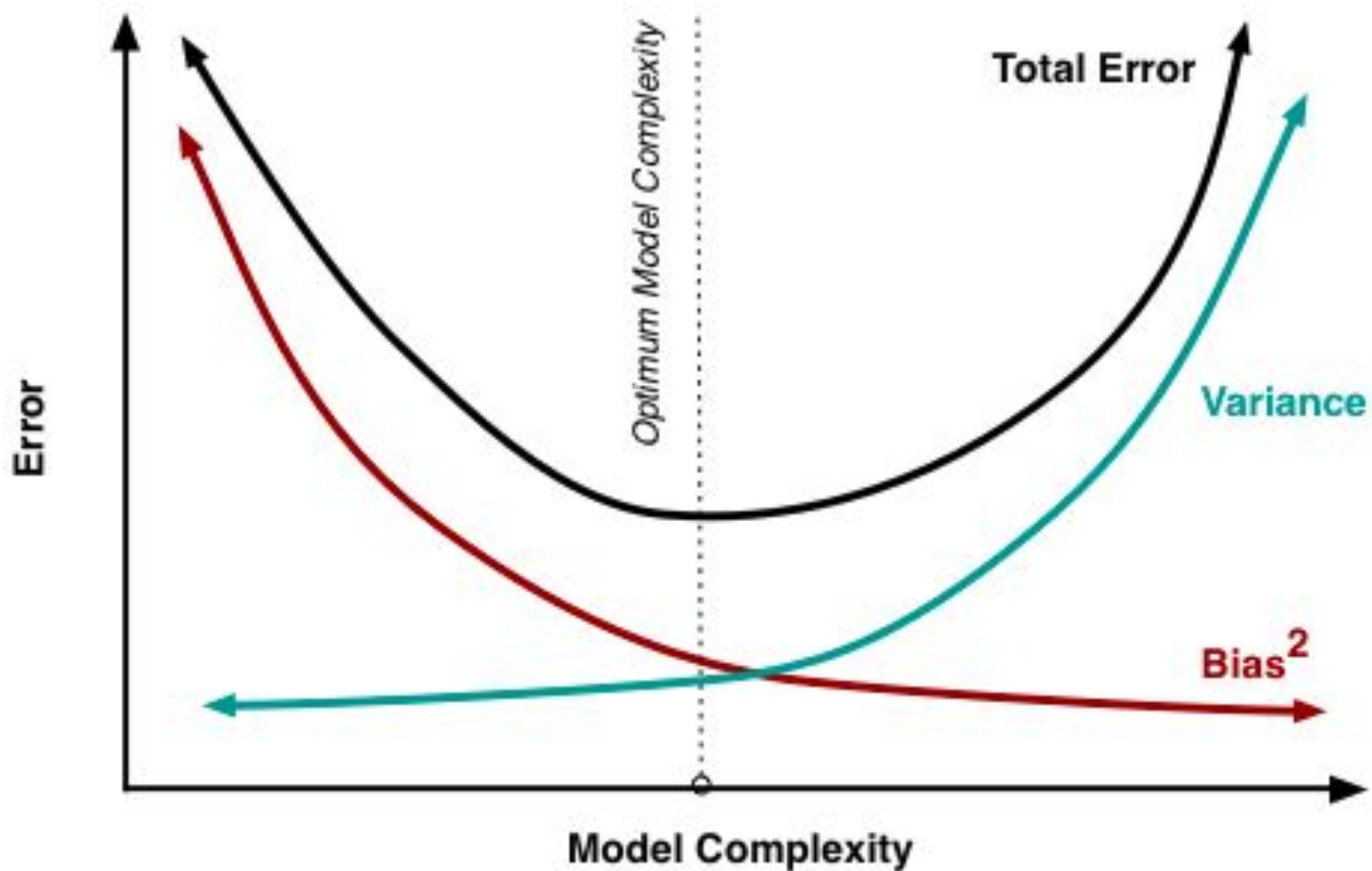
Se o modelo é muito sensível aos dados de treinamento, quando colocado em teste irá errar justamente a variação entre os datasets.



# Viés e Variância



# Viés e Variância - Trade off



# Regularização

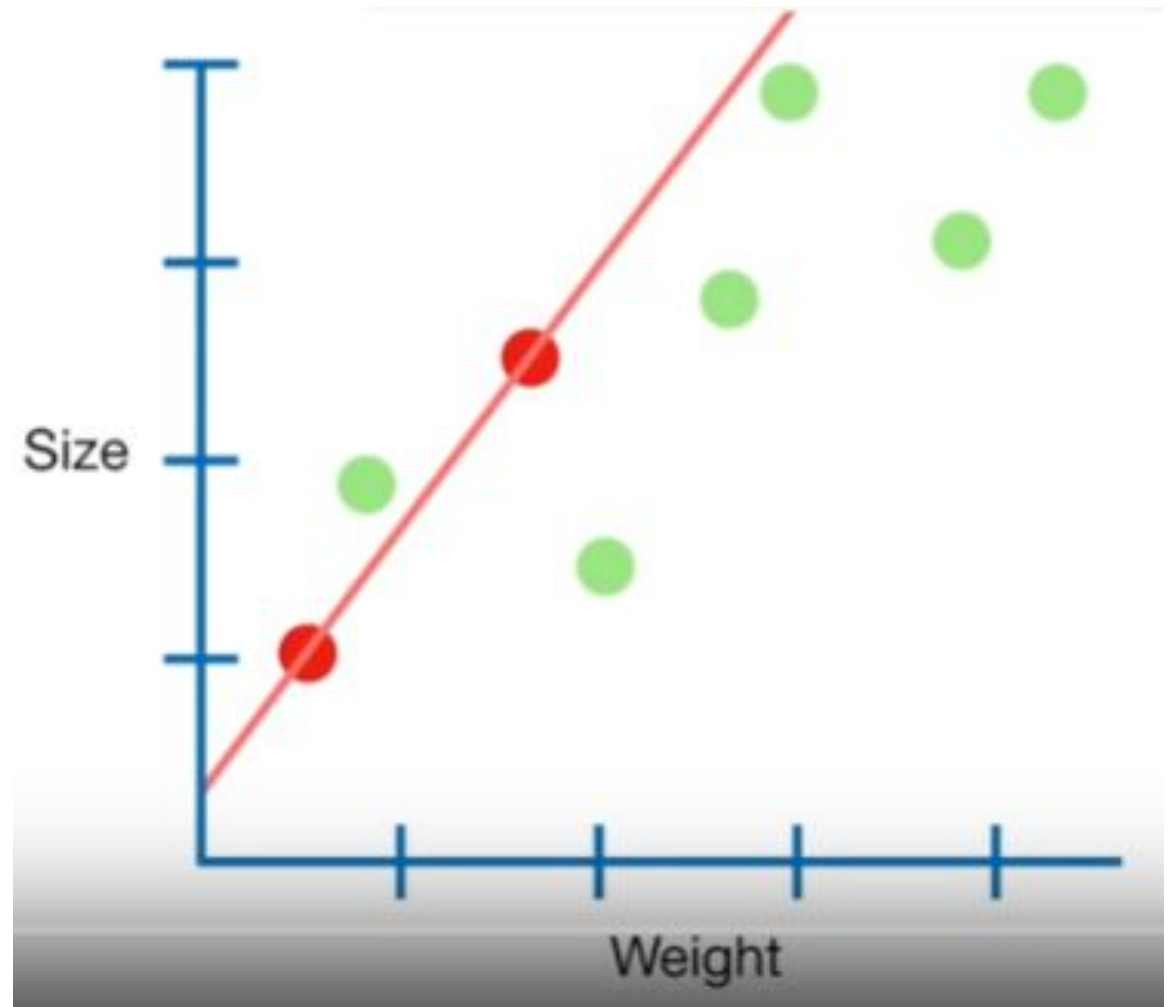
A regularização introduz um RUÍDO no modelo para diminuir o viés e a variância.





# Regularização

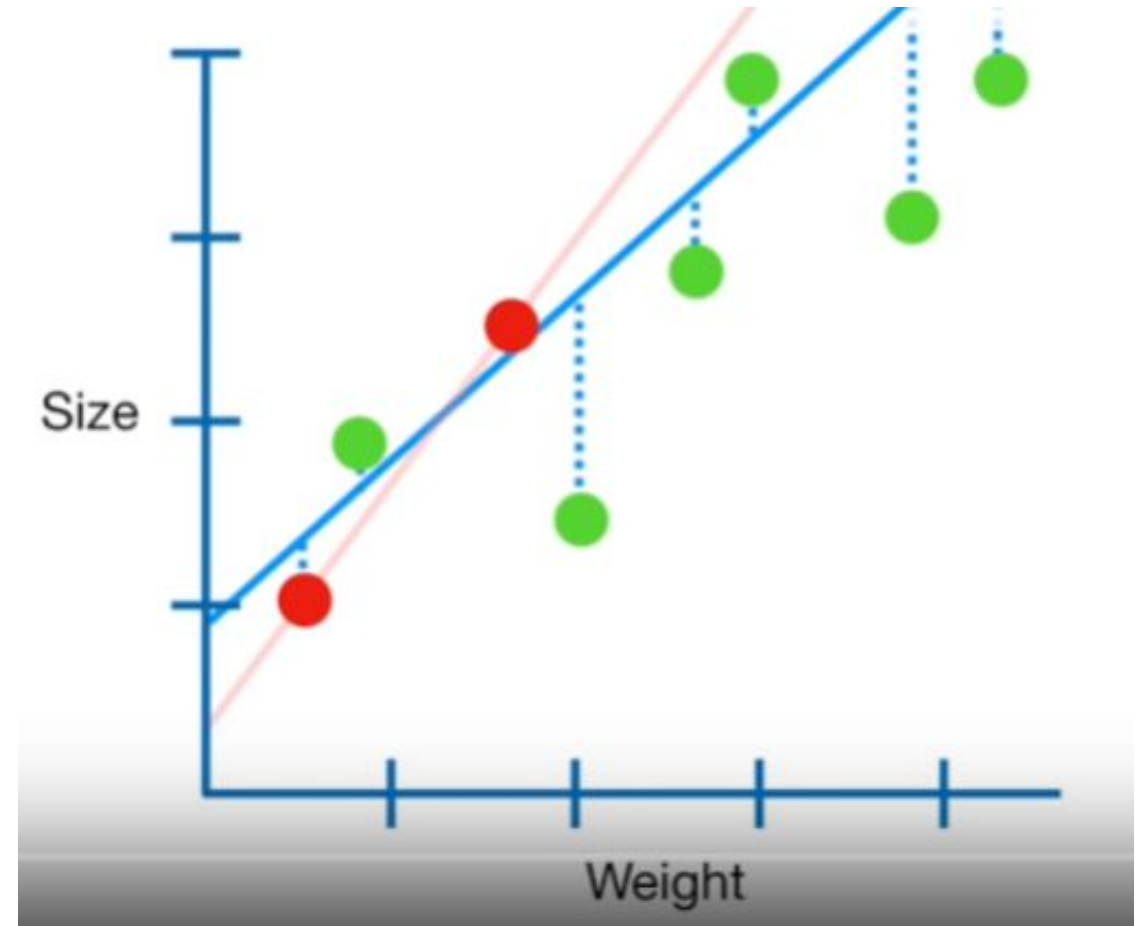
Essa é uma regressão comum que minimiza a soma dos resíduos ao quadrado



# Ridge - L1

Minimiza a soma dos resíduos ao quadrados + uma penalização em cima de todos os parâmetros, exceto a intersecção com y.

SRQ +  
 $\lambda * (\text{parâmetro}_1^2 + \text{parâmetro}_2^2 + \dots)$



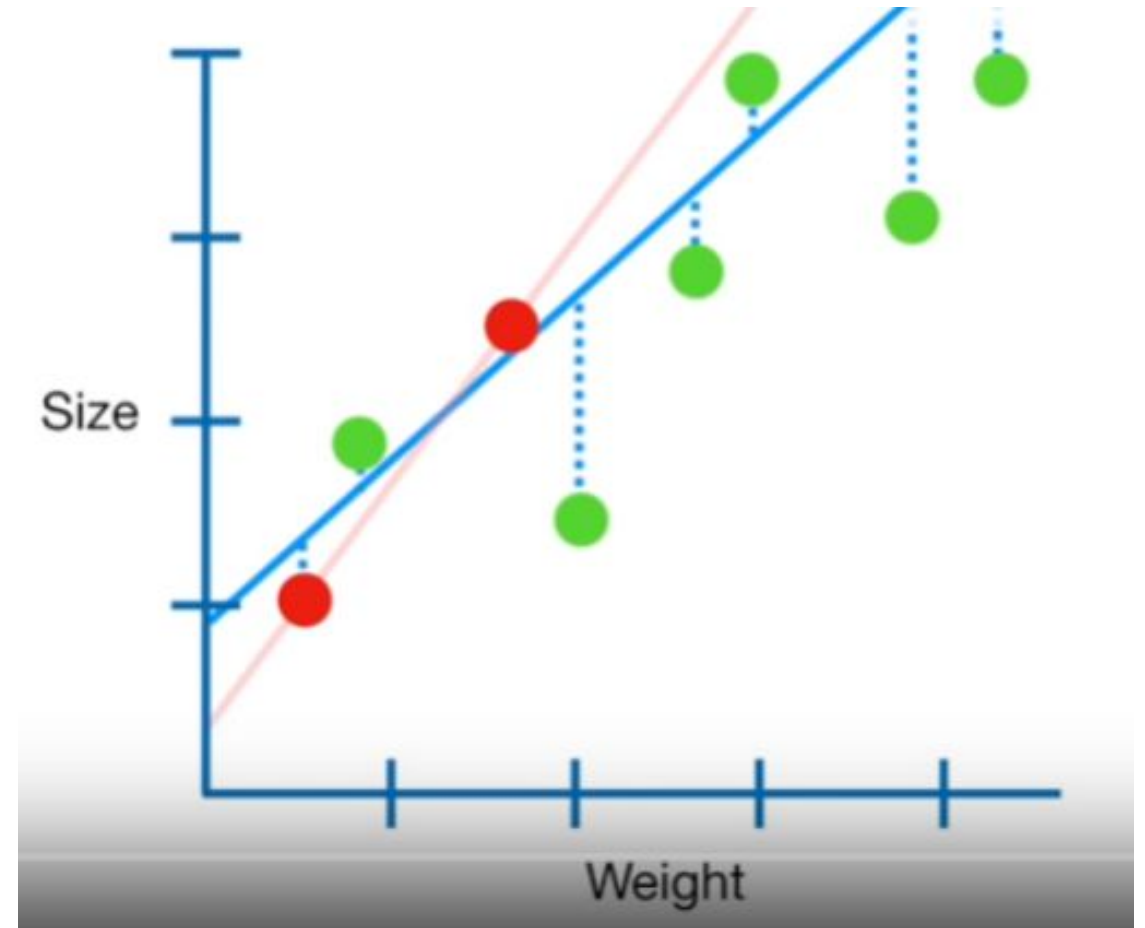


# Lasso - L2

Também penaliza os parâmetros.  
Usa os parâmetros em módulo.  
Minimiza a seguinte função:

$$\text{SRQ} + \lambda * (|\text{parâmetro}_1| + |\text{parâmetro}_2| + \dots)$$

Beleza, mas e daí?



# Lasso - L2

A grande diferença entre Lasso e Ridge é que a Lasso pode levar os parâmetros a (exatamente) 0 enquanto a Ridge apenas próximo à 0.

Isso significa que:

Parâmetros ruins usados na modelagem podem ser EXCLUÍDOS da equação.



# ElasticNet

$$\text{SRQ} + \lambda_1 * (|\text{parâmetro}_1| + |\text{parâmetro}_2| + \dots) + \lambda_2 * (\text{parâmetro}_1^2 + \text{parâmetro}_2^2 + \dots)$$

Lasso + Ridge

3 funções a serem minimizadas = parâmetros não muito bons acabam ficando com um peso MUITO baixo mas não zeram.

