



Modelos Lineares



ANA CAROLINA

Graduada e Mestre em Estatística

- Fundadora do Descomplica Estatística
- Cientista de Dados em TakeBlip



AGENDA

- **Introdução à modelagem**
Correlação e causalidade
- **Modelo Linear simples**



Introdução à modelagem

Correlação e causalidade

T Modelagem Estatística

Modelos estatísticos são técnicas (matemáticas) que ajudam:

- A descrever, prever ou classificar características e comportamentos.
- A comprovar ou rejeitar as hipóteses levantadas através do uso de probabilidades.

O nosso objetivo é responder perguntas, tais como:

- ❖ Será que, se eu investir X em marketing, o retorno nas vendas será Y?
- ❖ Qual a probabilidade de um cliente de um banco se tornar inadimplente ao pegar um empréstimo?

T Tipos de variáveis

Quantitativas

(Número)

Discreta

Geralmente se referem a contagens, e assumem apenas valores inteiros.

Exemplo:

- número de filhos (0, 1, 2, 3, ...)
- Número de compras

Contínua

Geralmente se referem a medições e podem assumir valores fracionados.

Exemplo:

- peso (kg)
- altura (m)
- IMC (Kg/m²)

Qualitativas

(Característica)

Nominal

Não há uma ordem definida entre as categorias.

Exemplo:

- cor dos olhos (preto, azul, verde)

Ordinal

Há uma ordem definida entre as categorias.

Exemplo:

- escolaridade (primário, médio, superior)
- grau de obesidade (leve, moderado, grave, mórbida)

A vertical bar with a gradient from green at the top to blue at the bottom.

**Vocês já ouviram falar em
correlação e causalidade?**

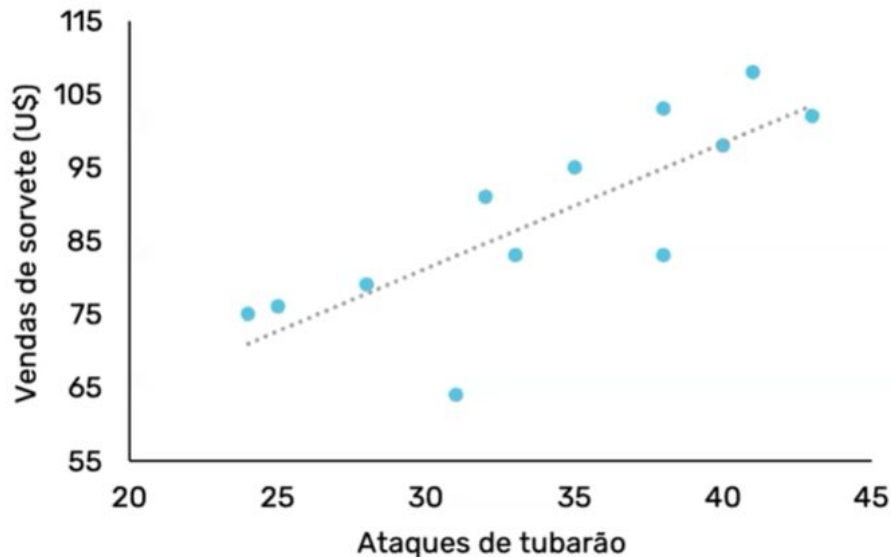
T Correlação

- **A correlação descreve uma associação entre variáveis:** quando uma variável muda, a outra também muda.

A correlação é um indicador estatístico da relação entre as variáveis. Ou seja, se essas variáveis mudam juntas (covariam).

Causalidade

- **Causalidade significa que mudanças em uma variável provocam mudanças na outra; existe uma relação de causa e efeito entre as variáveis.**



Será que eu posso falar que quanto mais ataques de tubarão, mais sorvete eu consigo vender?



- Quando uma variável muda, a outra também muda. Elas estão covariando. (Correlacionadas).

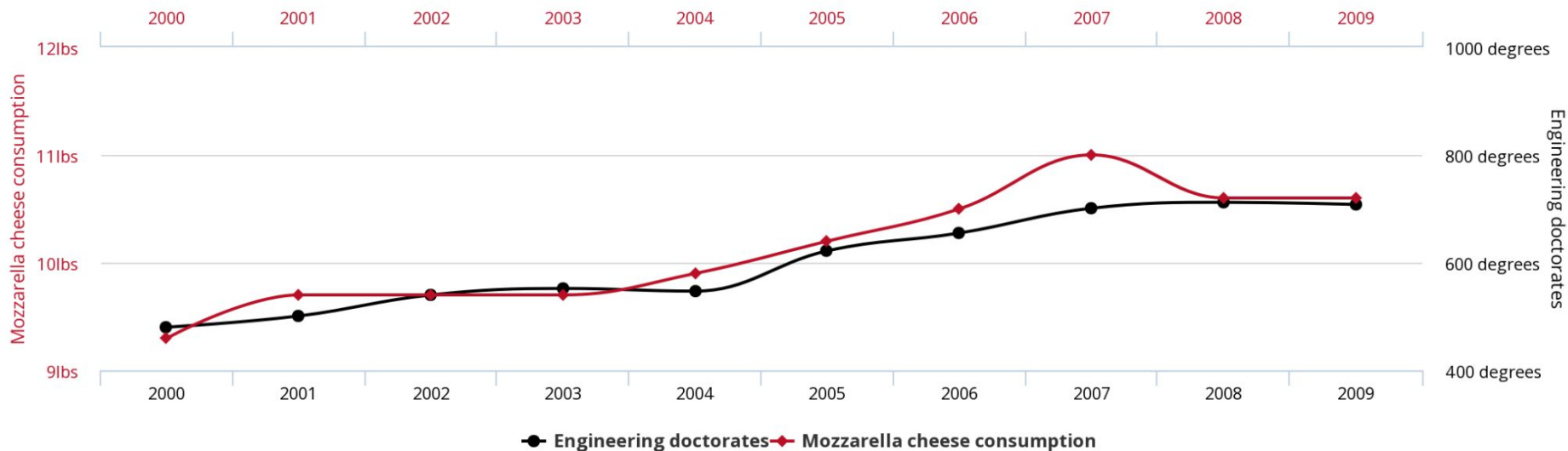
Correlação não implica em causalidade.

Consumo de queijo Muçarela - per capita

correlacionado com

Doutorados em engenharia civil concedidos

Correlation: 95.86% ($r=0.958648$)

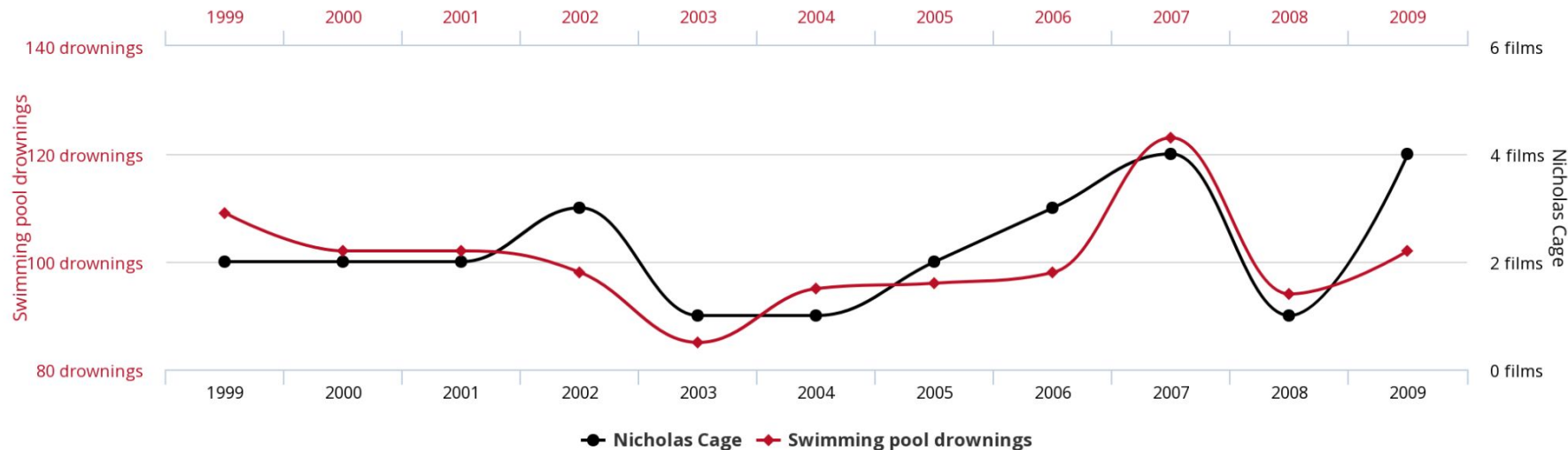


Número de pessoas que se afogaram ao cair em uma piscina

correlacionado com

Filmes em que Nicolas Cage aparece

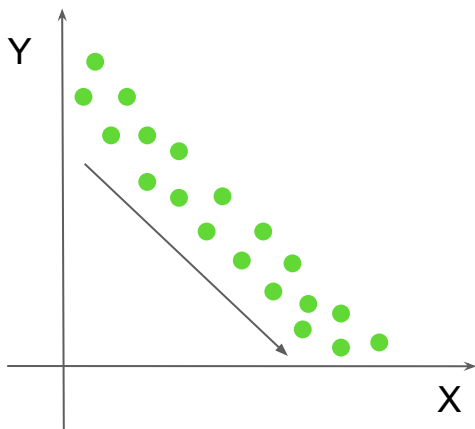
Correlation: 66.6% ($r=0.666004$)



T Correlação de Pearson $-1 \leq \rho \leq 1$

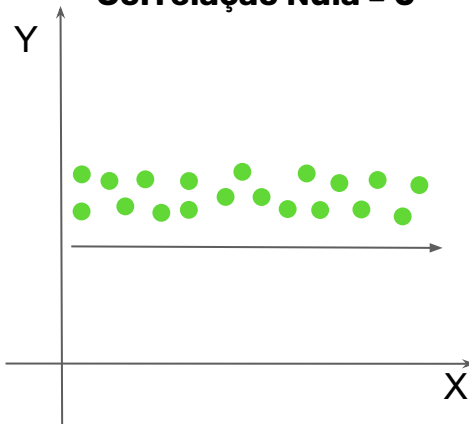
É uma medida com o objetivo de medir o grau de correlação linear entre duas variáveis.

Correlação Negativa = -1



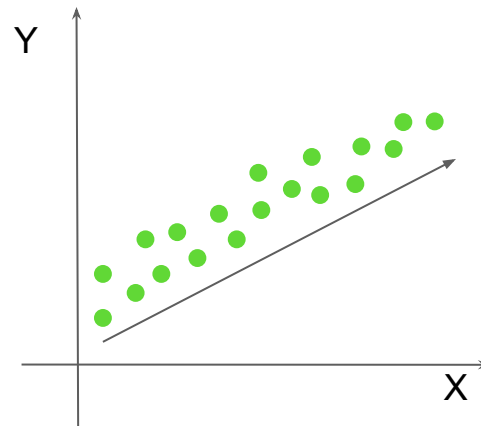
A variável **Y** tende a **diminuir** na mesma proporção em que **X aumenta**.

Correlação Nula = 0




Com o **aumento** na variável **X**, **não se observa variação em Y**.

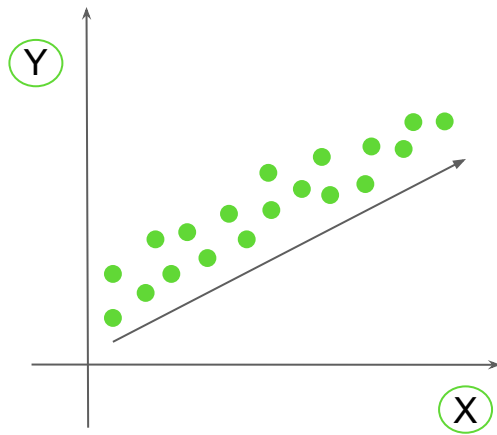
Correlação Positiva = 1



A variável **Y** tende a **aumentar** na mesma proporção em que **X aumenta**.

T Correlação de Pearson

$$\rho = \frac{Cov(X, Y)}{S_x S_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$




X = Temperatura
Y = Vendas de sorvete

A correlação pode ser calculada.

Mas como identificar se existe causalidade?

Precisamos ajustar um modelo estatístico, como o modelo de Regressão Linear.

A vertical bar with a gradient from green at the top to blue at the bottom.

Modelo Linear Simples

T Fatura do cartão de crédito

Vamos analisar um conjunto de dados simulado composto por informações de 400 portadores de cartão de crédito **e tentar validar se existe relação de causa e efeito entre o Limite e o valor da fatura do Cartão de Crédito.** (Exemplo retirado do livro - "[An Introduction to Statistical Learning with Applications in R 2ª edição](#)")

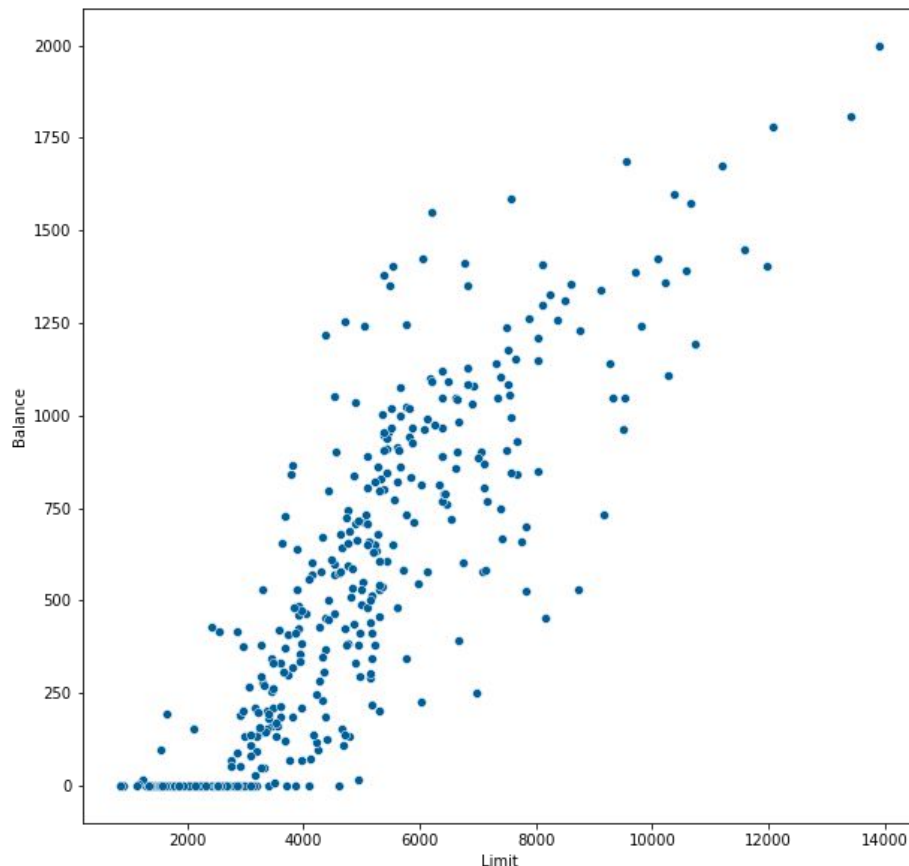
	Fatura	Limite
0	333	3606
1	903	6645
2	580	7075
3	964	9504
4	331	4897
...
394	734	5758
395	560	4100
396	480	3838
397	138	4171
399	966	5524

310 rows × 2 columns

Correlação:
Fatura x Limite = 0.796

- Será que existe relação de causa e efeito entre essas duas variáveis?
- Quanto maior o limite fornecido, maior o valor da fatura ?

Correlação:
Fatura x Limite = 0.796

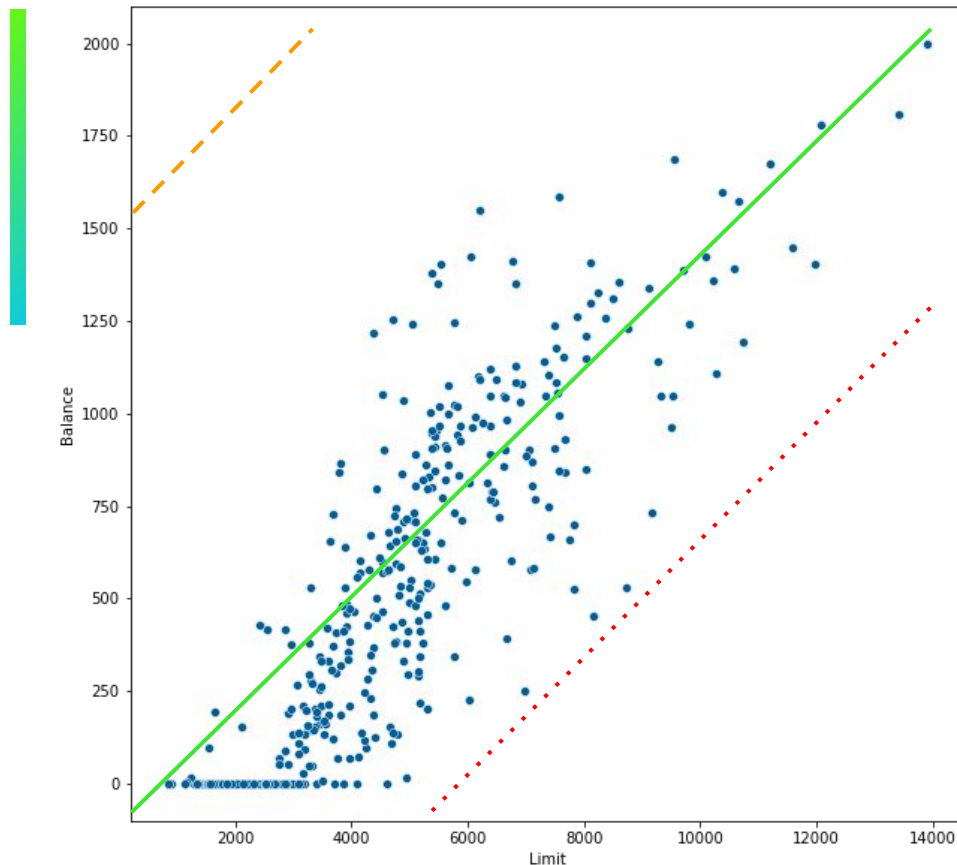


- Será que conseguimos saber qual vai ser o valor da fatura de um cartão de crédito utilizando os dados do limite?

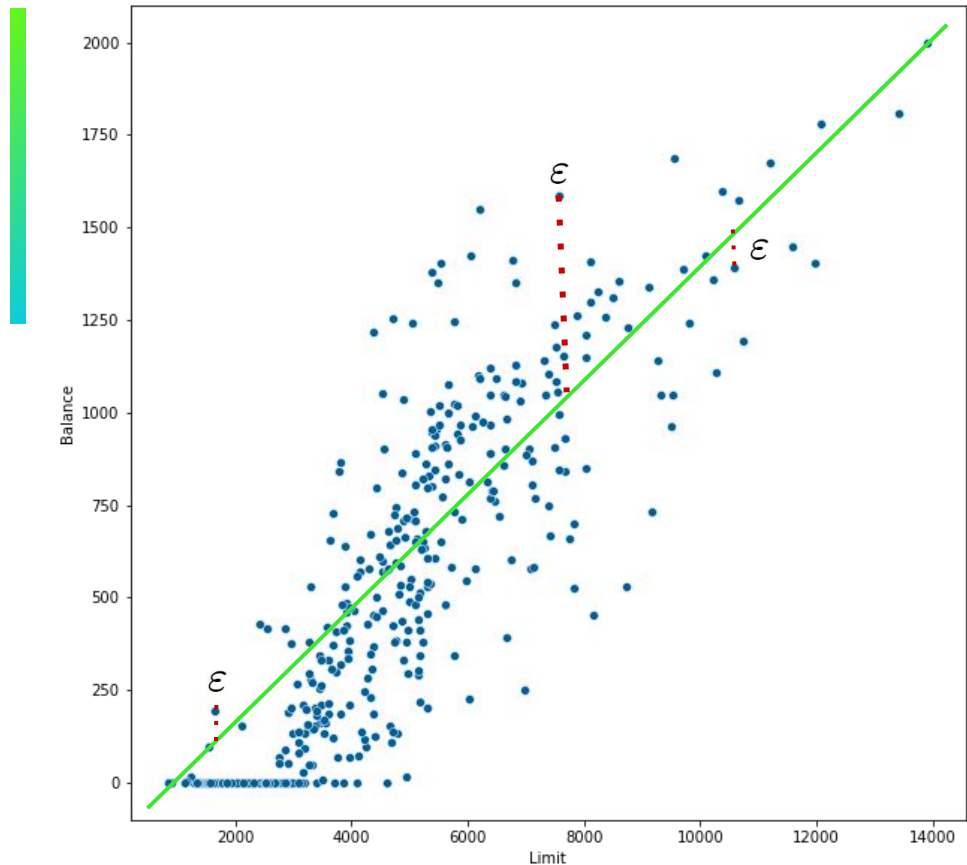
Para responder essas perguntas, vamos pensar em uma modelagem.

T

Correlação:
Fatura x Limite = 0.796



**Traçamos uma reta próximo
aos pontos, porque queremos
errar o menos possível.**



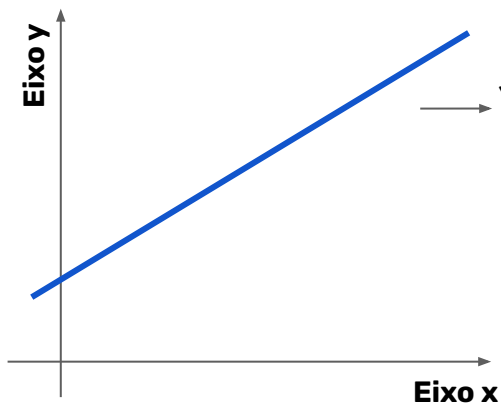
- Estamos buscando uma modelagem, para explicar ou prever os pontos, mas estamos fazendo algo que certamente vai ter um erro.

Queremos minimizar o erro que vamos cometer.



Vamos relembrar os tempos de escola

$$a + bx \longrightarrow \text{Equação da reta}$$



$$Y = a + bx$$

a = ponto onde a reta toca o eixo y (quando $x = 0$)

b = inclinação da reta

Equação da reta

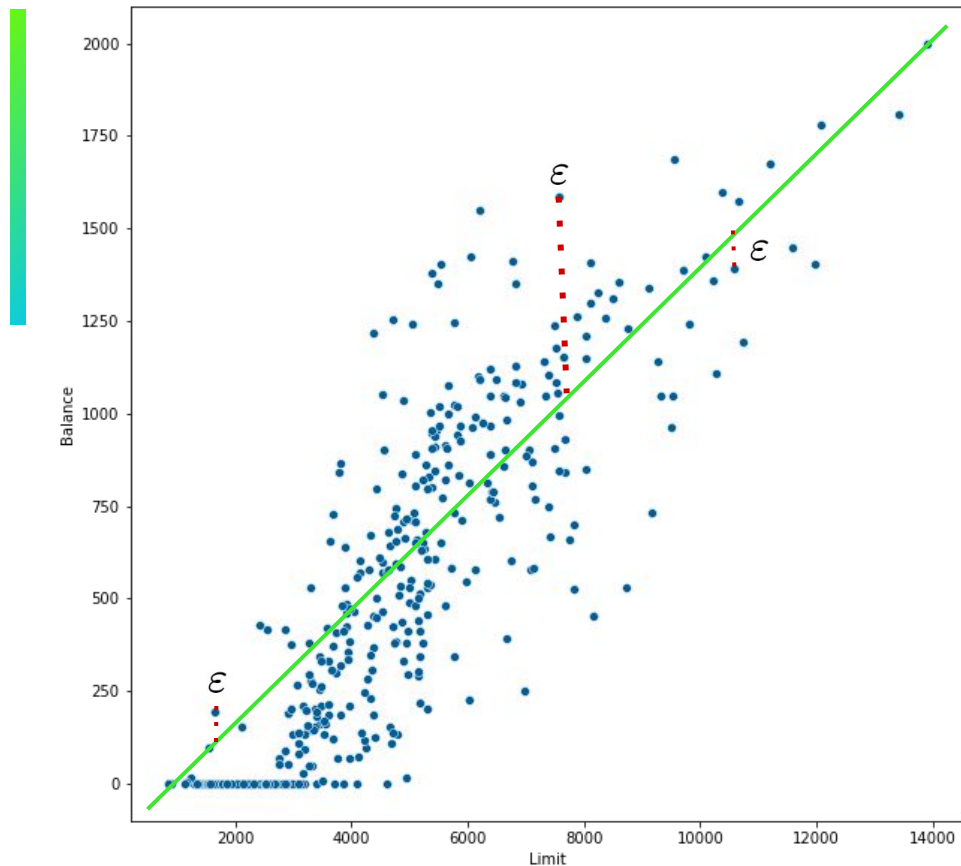
$$y = a + bx$$

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

β_0 = ponto onde a reta toca o eixo y (quando $x = 0$)
 β_1 = inclinação da reta

y : Fatura média do cartão (variável que eu quero explicar ou prever).

x : Limite de crédito (variável que vai auxiliar na explicação ou previsão).



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



Precisamos
encontrar uma reta
que minimize esse
erro.

T Estruturando a nossa equação:

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

$$\varepsilon = y - \underbrace{\beta_0 + \beta_1 x}_{\substack{\text{Função matemática que} \\ \text{vamos usar para fazer a} \\ \text{estimação}}} = \hat{y}$$

Valor verdadeiro que queremos estimar

$$\varepsilon = y - \hat{y}$$

O erro é o valor verdadeiro menos o valor estimado pela nossa equação matemática.

Queremos que o erro seja o menor possível.

T Estruturando a nossa equação:

	Fatura	Limite
0	333	3606
1	903	6645
2	580	7075
3	964	9504
4	331	4897
...
394	734	5758
395	560	4100
396	480	3838
397	138	4171
399	966	5524

$$\varepsilon = y - \beta_0 - \beta_1 x$$

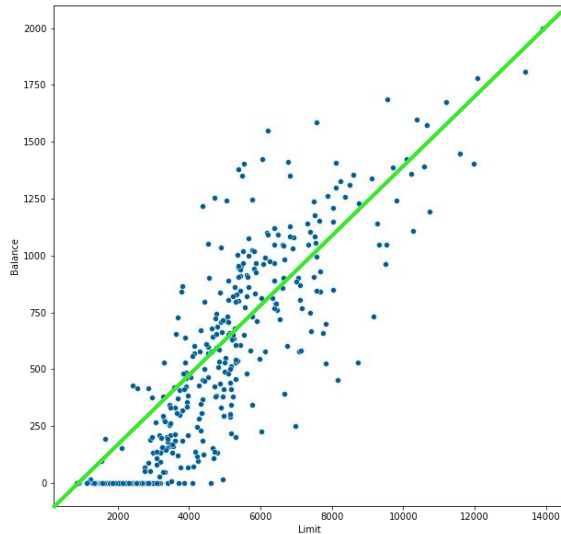
$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

↓
indivíduo ou
observação na
amostra

$$\sum_{i=1}^n \varepsilon_i = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)$$

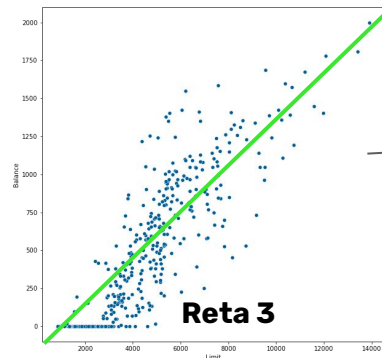
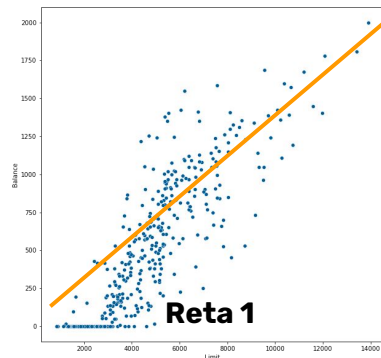
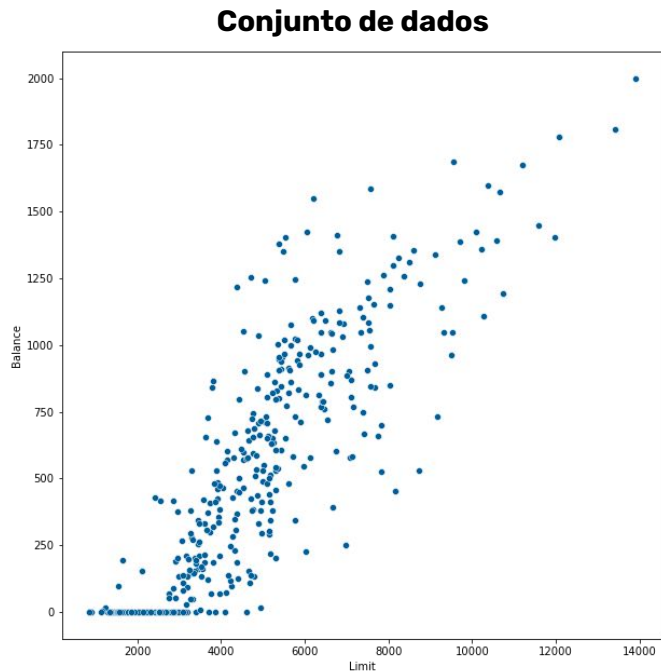
n = Tamanho da base de dados

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

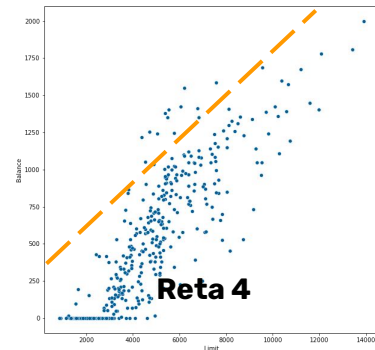
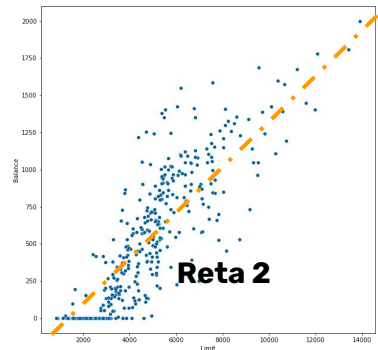




Estimadores por Mínimos Quadrados



Reta que minimiza o erro que cometemos.



$$\sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$



Estima-se os parâmetros desconhecidos através dos dados da amostra.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

O modelo estima os parâmetros desconhecidos através dos dados da amostra.

$$Fatura = \hat{\beta}_0 + \hat{\beta}_1 * Limite$$

$$Fatura = 120 + 0.5 * Limite$$

Fatura média estimada se considerarmos o valor do limite igual a 0.

Interpretação $\hat{\beta}_0$

A cada incremento de 1 dólar no limite, aumenta-se em média o valor da fatura em 0.5 dólares.

Interpretação $\hat{\beta}_1$

T Previsão

$$Fatura = 120 + 0.5 * Limite$$

Suponha que fizemos essa análise, e agora queremos saber **qual será em média o valor da fatura de uma pessoa se o limite concedido for de 5000 dólares.**

$$Fatura = 120 + 0.5 * 5000$$

$$Fatura = \$2620$$

A previsão média da fatura para uma pessoa que possui limite de \$5000 é de \$2620.

Regressão linear simples

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Diagram illustrating the components of the simple linear regression equation:

- y_i : **variável dependente/ target**
- β_0 : Valor de y quando $x = 0$ (intercepto)
- β_1 : Inclinação da reta: acréscimo ou decréscimo em y para cada acréscimo de unidade em x
- x_i : **variável independente/ explicativa/ feature**
- ε_i : Erro da estimação

A variável dependente, y , é uma variável contínua.

T

A vertical bar with a gradient from light green at the top to light blue at the bottom.

RESUMO E TAKEAWAYS

TAKEAWAY #1

Quando duas variáveis mudam juntas (covariam), elas têm uma correlação.

Ter correlação não implica que mudanças em uma variável provoca mudanças na outra.
(Correlação não implica em causalidade).

Correlação é apenas um indicativo de causalidade.

TAKEAWAY #2

Um modelo linear simples identifica e explica a **relação linear entre a variável **target (Y)** e a variável **independente (X)**.**

Utilizamos uma fórmula matemática para construir essa relação e esperamos que os valores estimados por ela sejam próximos dos valores verdadeiros.

TAKEAWAY #3

A vertical bar with a gradient from green at the top to blue at the bottom.

A variável de interesse (y) é contínua.

O Método de Mínimos Quadrados (MMQ), ou Mínimos Quadrados Ordinários (MQO) ou OLS (do inglês Ordinary Least Squares) é uma técnica de otimização matemática que procura encontrar o melhor ajuste para um conjunto de dados tentando minimizar a soma dos erros ao quadrado.

