Universidade Federal da Paraíba
Coordenação do Curso de Ciência de Dados e Inteligência Artificial

Regressão Linear

Prof. Gilberto Farias

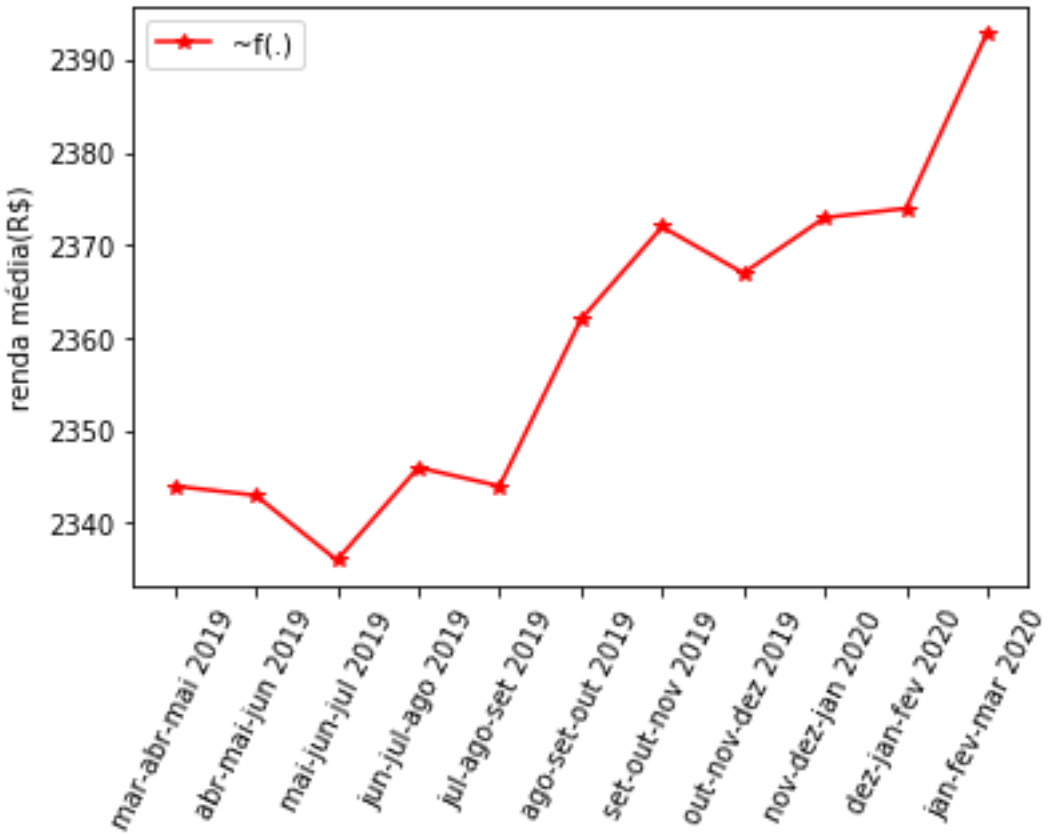
Roteiro

- Regressão linear
  - Prática 1 : renda média brasileira
- Introd. Séries Temporais
  - Prática 2 : tendência e sazonalidade
  - Exercício : Covid 19

Como prever a renda média brasileira nos meses futuros??

Trimestre	Renda Média (R$)
mar-abr-mai 2019	2344
abr-mai-jun 2019	2343
mai-jun-jul 2019	2336
jun-jul-ago 2019	2346
jul-ago-set 2019	2344
ago-set-out 2019	2362
set-out-nov 2019	2372
out-nov-dez 2019	2367
nov-dez-jan 2020	2373
dez-jan-fev 2020	2374
jan-fev-mar 2020	2393

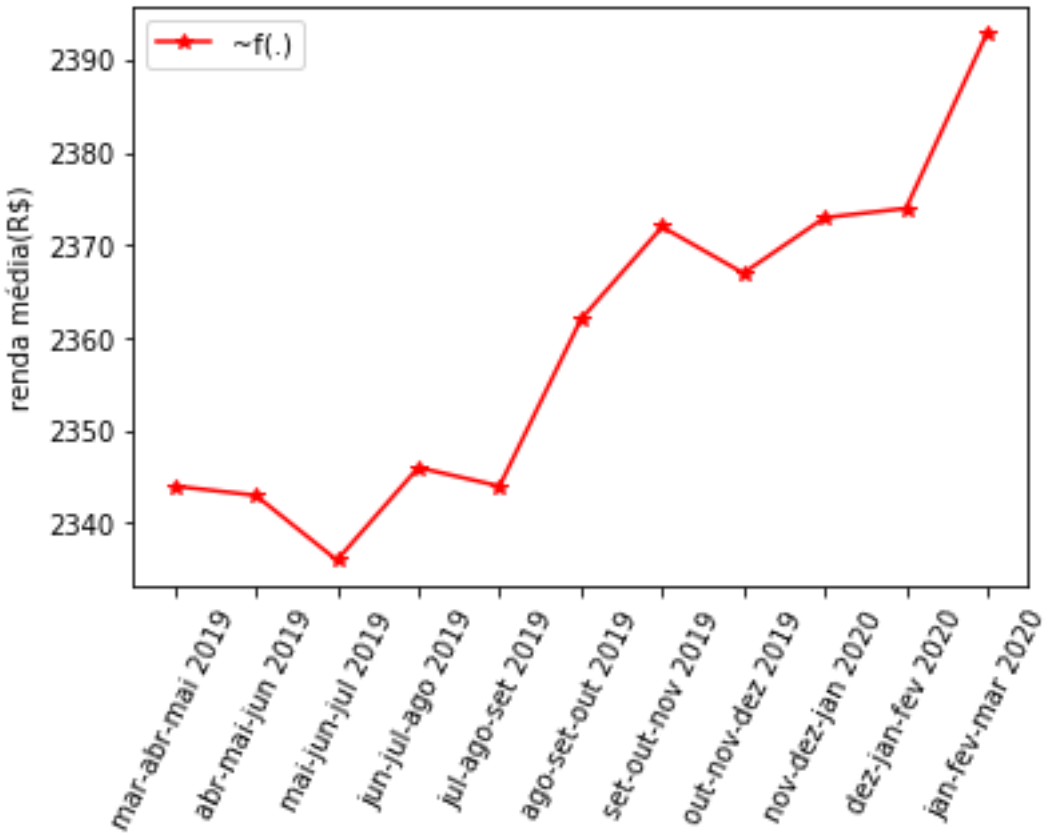
Fonte: IBGE



$f(x)???$

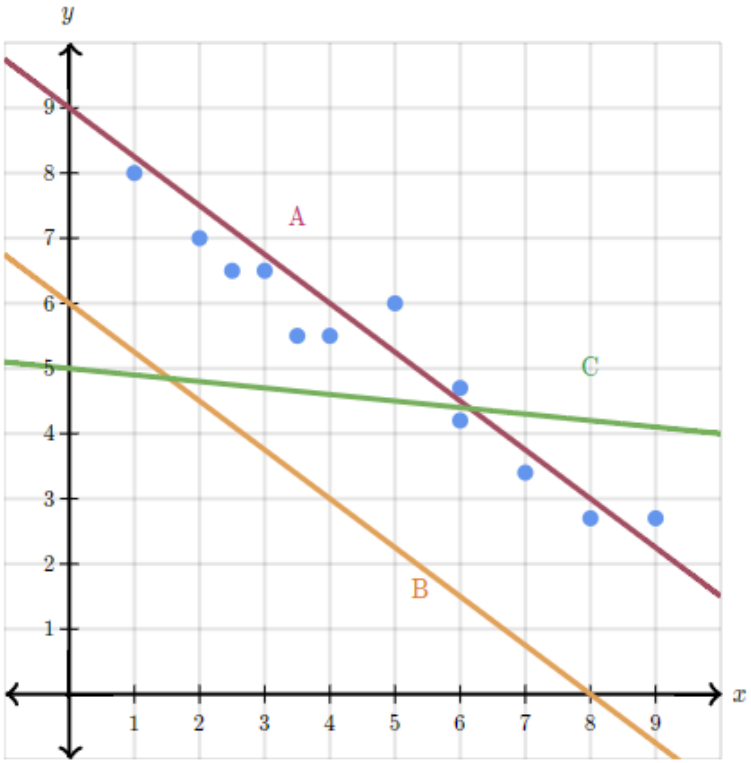
Representação dos dados de entrada

x	$f(x)$
1	2344
2	2343
3	2336
4	2346
5	2344
6	2362
7	2372
8	2367
9	2373
10	2374
11	2393



$f(x)???$

Qual a função linear $h(x)$ descreve melhor os pontos?
A, B ou C?



Fonte: Khan Academy

$$h(x) = ax + b$$

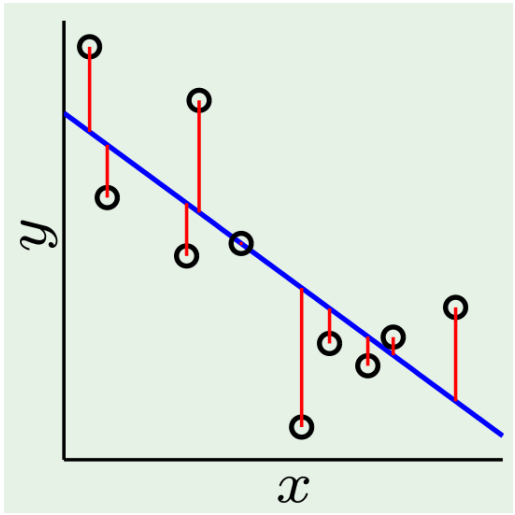
$$h(x) = w_1 x_1 + w_0 x_0, \quad \text{onde } x_0 = 1$$

$$h(x) = \sum_{i=0}^{d} w_i x_i = \mathrm{w}^{\mathrm{T}}\mathrm{x}$$

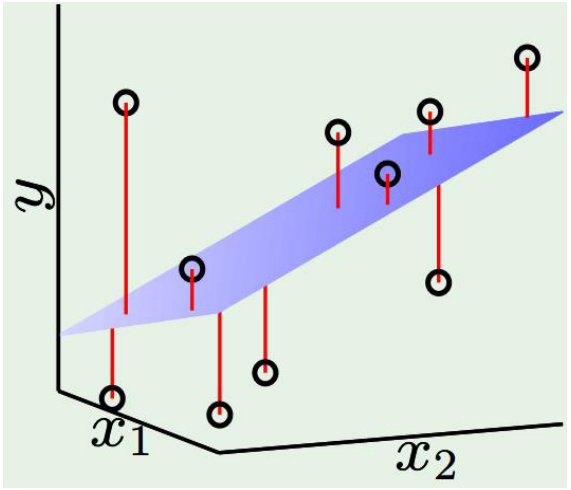
Como aproximar $h(x) = w^T x$ de $f(x)$??

- Na regressão linear é utilizado o erro quadrático $(h(x) - f(x))^2$

$$E(h) = \frac{1}{N} \sum_{n=1}^{N} (h(\mathbf{x}_n) - y_n)^2$$



Fonte: notas de aula Yaser Abu Mostafa



Fonte: notas de Yaser Abu Mostafa

Forma vetorial de $\boldsymbol{E(h)}$

$$E(h) = \frac{1}{N}\sum_{n=1}^{N}(h(\mathrm{x}_n) - y_n\ )^2$$

$$E(\mathrm{w}) = \frac{1}{N}\sum_{n=1}^{N}(\mathrm{w}^{\mathrm{T}}\mathrm{x}_n - y_n\ )^2$$

$$E(\mathrm{w}) = \frac{1}{N}\|(\mathrm{X}\mathrm{w} - y)\|^2$$

onde

$$\mathrm{X} = \begin{bmatrix} -\ \mathrm{x}_1^{\mathrm{T}}\ - \\ -\ \mathrm{x}_2^{\mathrm{T}}\ - \\ \vdots \\ -\ \mathrm{x}_N^{\mathrm{T}}\ - \end{bmatrix} = \begin{bmatrix} 1{,}1 \\ 1{,}2 \\ \vdots \\ 1{,}11 \end{bmatrix}$$

trimestre

x_0, x_1

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 2344 \\ 2343 \\ \vdots \\ 2393 \end{bmatrix}$$

renda média

Minimizando a função erro $\boldsymbol{E(w)}$ e computando **w**

$$E(\mathrm{w}) = \frac{1}{N}\|(\mathrm{X}\mathrm{w} - \mathrm{y})\|^2$$

Ponto mínimo de $E(w)$ $\longrightarrow$ $\nabla E(\mathrm{w}) = \frac{2}{N}\mathrm{X}^T(\mathrm{X}\mathrm{w} - \mathrm{y}) = 0$

$$\mathrm{X}^T\mathrm{X}\mathrm{w} = \mathrm{X}^T\mathrm{y} \longrightarrow \mathrm{w} = (\mathrm{X}^T\mathrm{X})^{-1}\,\mathrm{X}^T\mathrm{y}$$

$\mathbf{w} = \mathbf{X}^\dagger\mathbf{y}$ onde $\mathbf{X}^\dagger = (\mathrm{X}^\mathrm{T}\mathrm{X})^{-1}\,\mathrm{X}^\mathrm{T}$

pseudo-inversa de X

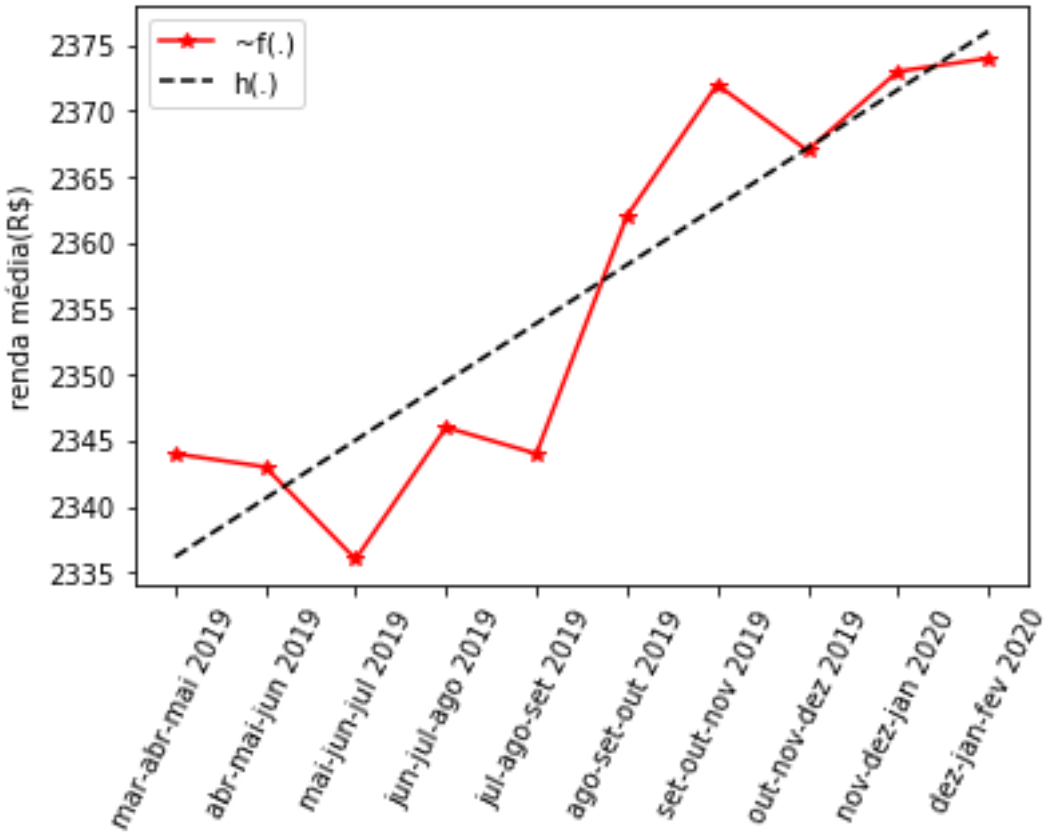
Algoritmo de Regressão Linear

1. Compute a matriz X e o vetor y do conjunto de dados (x_1, y_1), ..., (x_N, y_N);

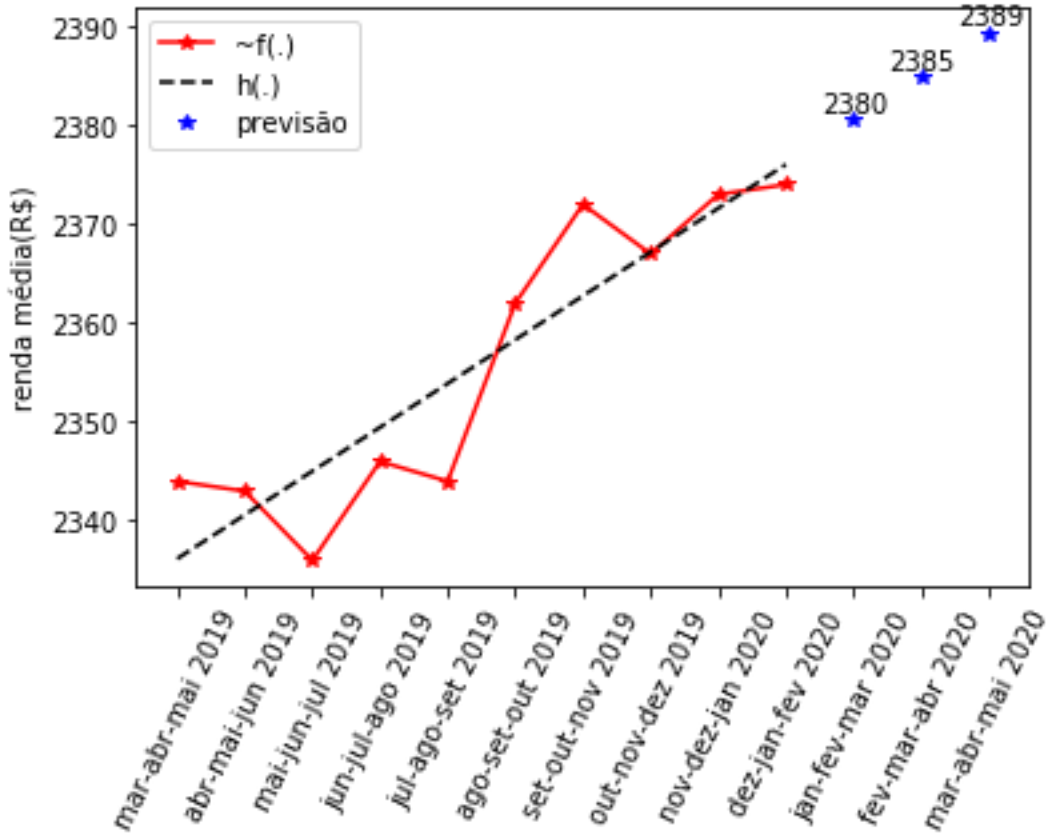
2. Compute a pseudo-inversa de $\mathbf{X}^{\dagger} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\,\mathrm{X}^{\mathrm{T}}$;

3. Retorne $\boldsymbol{w} = \boldsymbol{X}^{\dagger}\mathbf{y}$.

Regressão linear aplicada ao rendimento médio

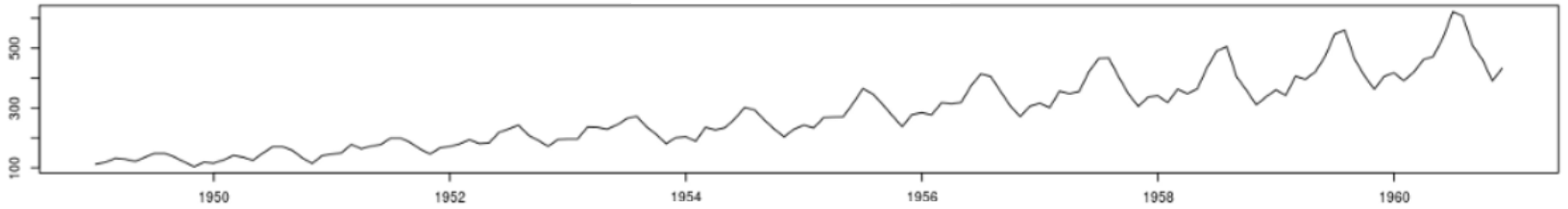


Extrapolando os valores para meses futuros

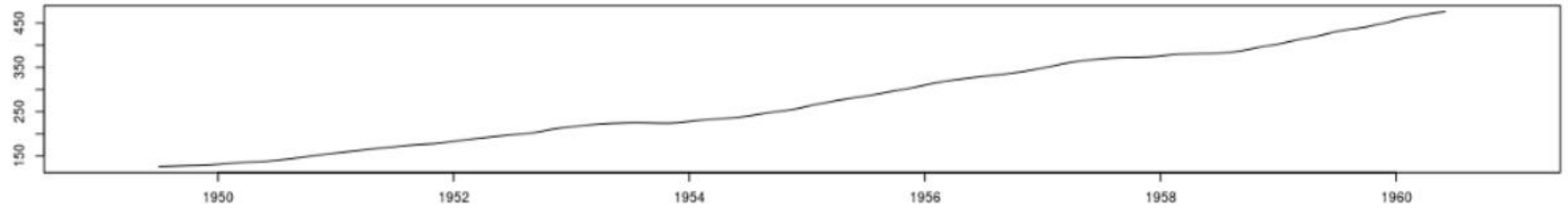


Series Temporais: $y_t = T_t + S_t + \varepsilon_t$

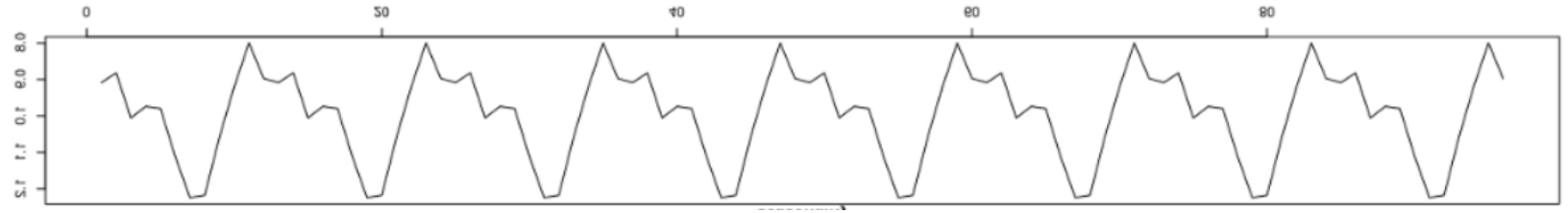
Série Temporal



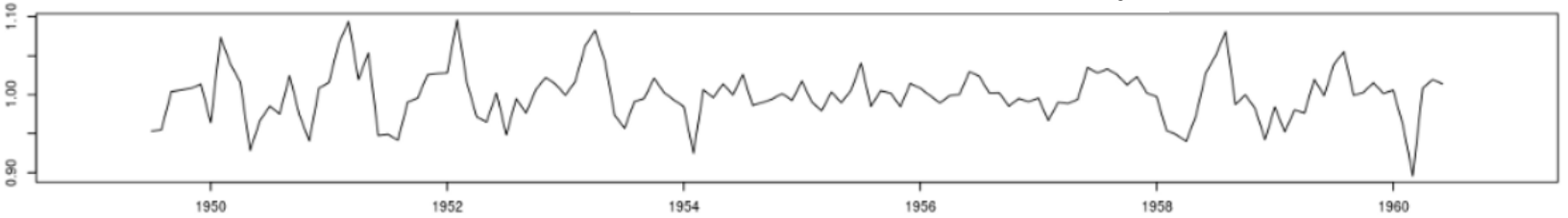
Componente de Tendência (T_t)



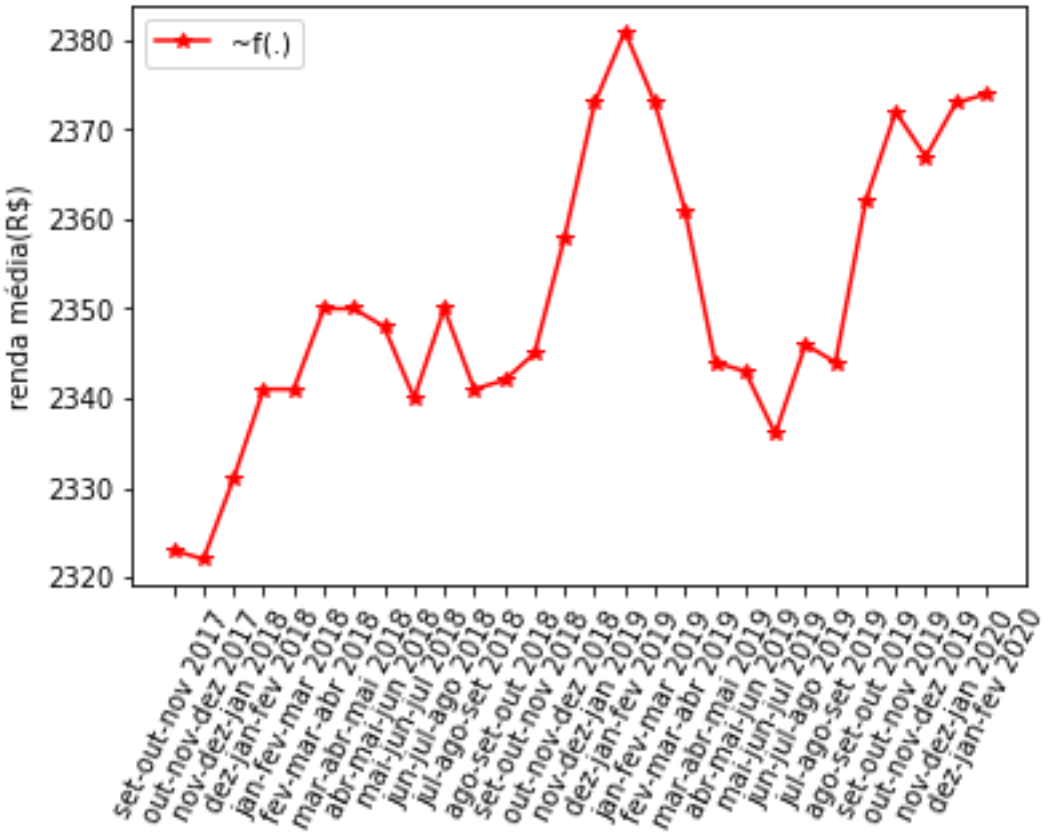
Componente de Sazonalidade (S_t)



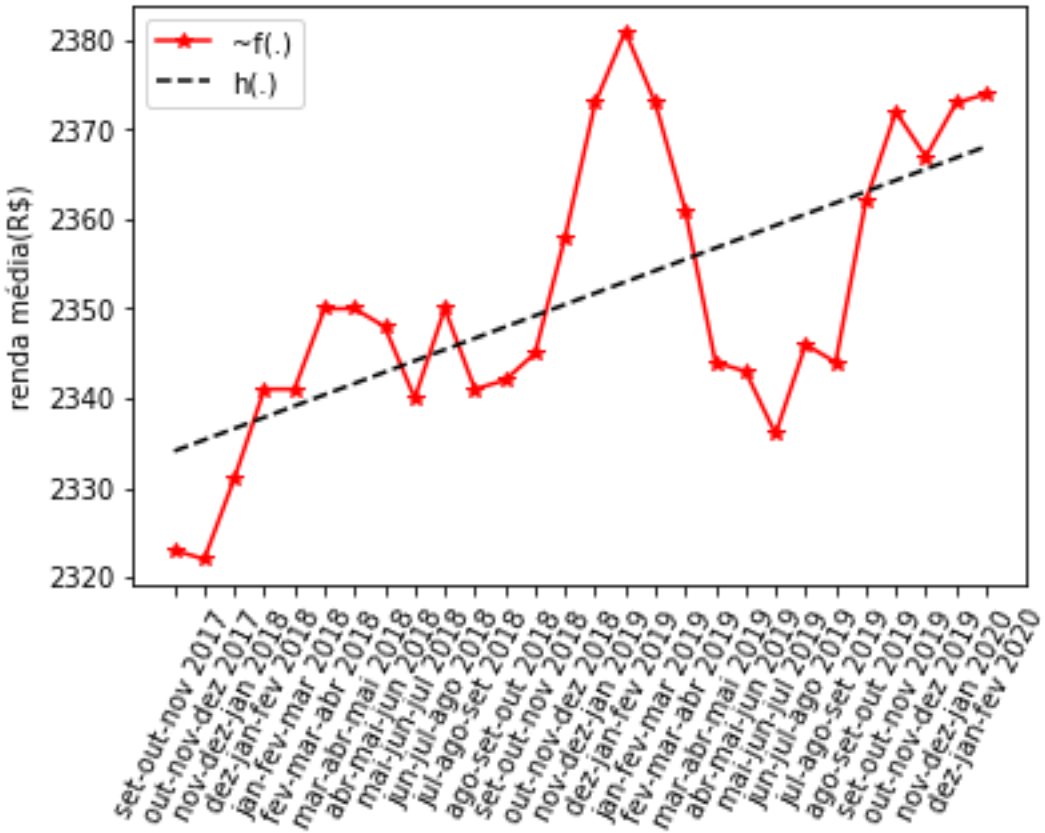
Componente aleatória (ε_t)



Qual a **tendência** da renda média brasileira? 2017 - 2020

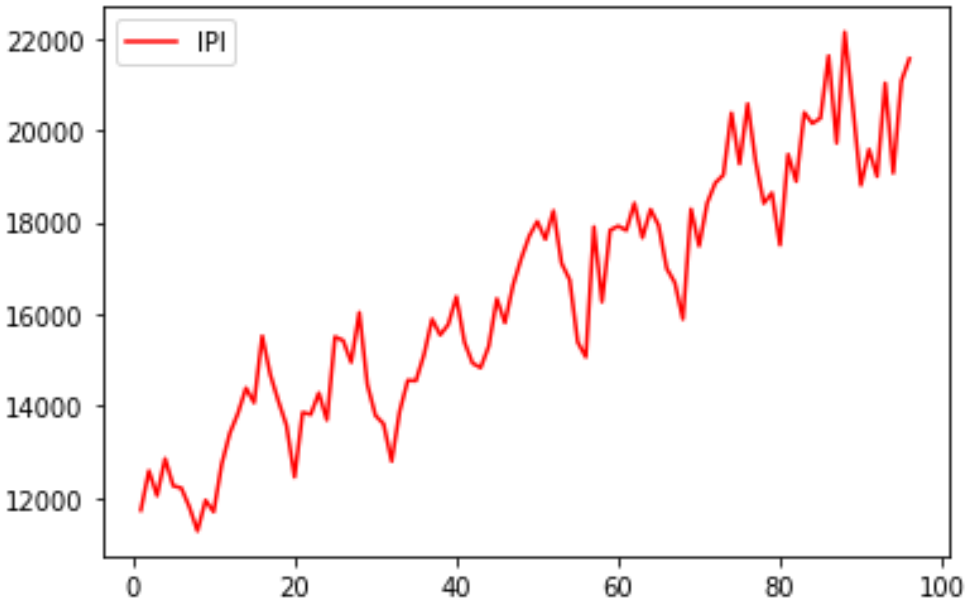


Aplica regressão linear



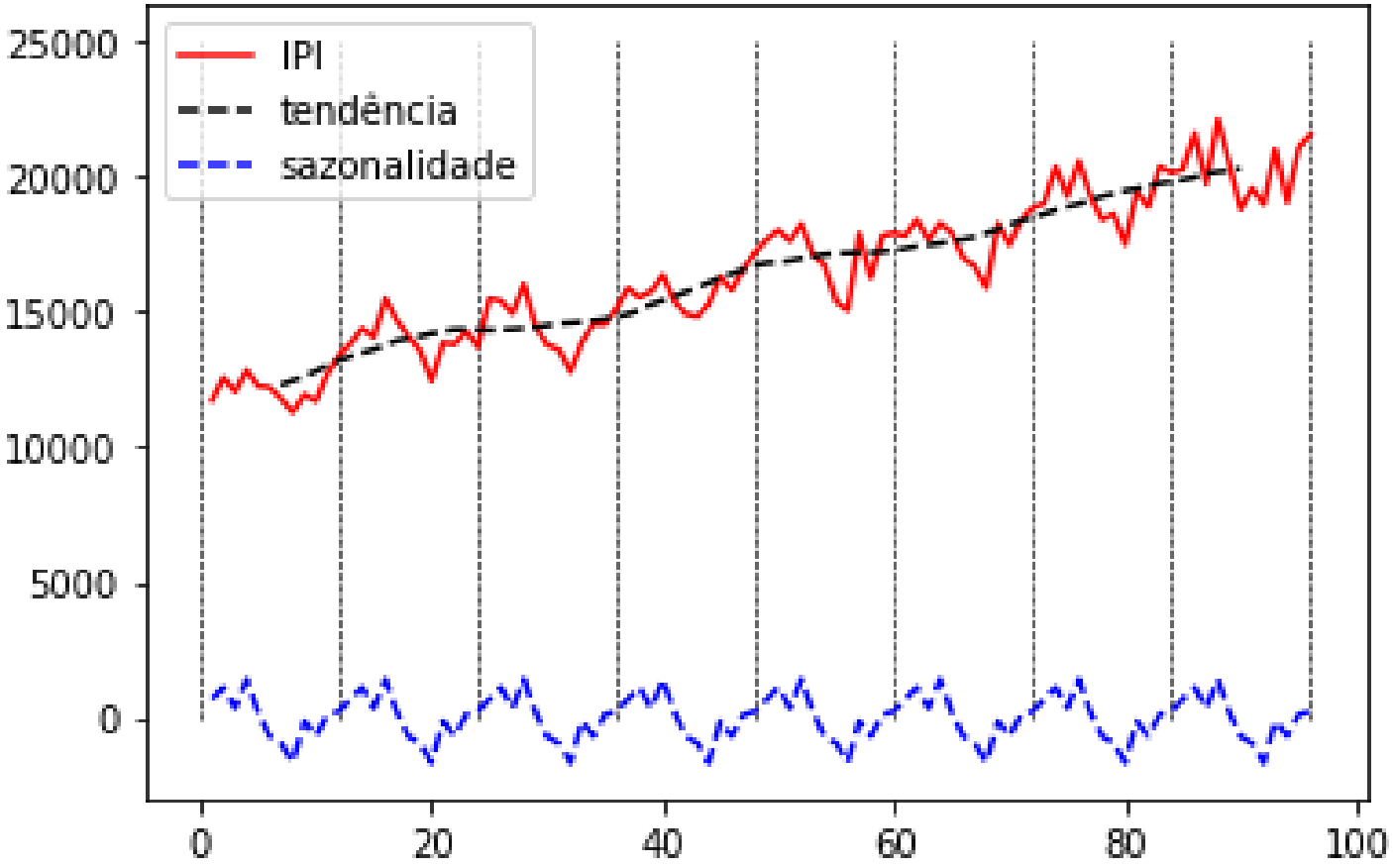
Qual a sazonalidade da série de Índice de Produto Industrial do Brasil (IPI)

Ano	Jan.	Fev.	Mar.	Abr.	Mai.	Jun.	Jul.	Ago.	Set.	Out.	Nov.	Dez.
1969	7.780	7.351	8.317	8.036	8.424	8.300	8.985	8.589	8.564	8.614	8.102	8.044
1970	8.209	7.738	8.828	9.150	8.960	9.282	9.934	9.546	9.572	10.272	9.991	9.537
1971	8.761	8.501	9.642	9.058	9.256	9.799	10.828	11.063	10.652	11.278	10.661	10.500
1972	9.759	9.876	10.664	10.110	11.055	11.615	11.730	12.587	12.046	12.852	12.259	12.214
1973	11.798	11.278	11.945	11.695	12.734	13.405	13.836	14.388	14.069	15.519	14.680	14.104
1974	13.577	12.451	13.856	13.812	14.280	13.692	15.502	15.423	14.947	16.031	14.462	13.791
1975	13.608	12.794	13.889	14.555	14.545	15.114	15.886	15.541	15.770	16.375	15.386	14.927
1976	14.829	15.297	16.330	15.807	16.623	17.196	17.691	18.012	17.625	18.244	17.102	16.744
1977	15.385	15.062	17.896	16.262	17.820	17.911	17.818	18.410	17.658	18.273	17.922	16.987
1978	16.681	15.886	18.281	17.478	18.412	18.849	19.023	20.372	19.262	20.570	19.304	18.407
1979	18.633	17.497	19.470	18.884	20.308	20.146	20.258	21.614	19.717	22.133	20.503	18.800
1980	19.577	18.992	21.022	19.064	21.067	21.553	22.513	–	–	–	–	–



Decompondo as componentes da série temporal

$$y_t = T_t + S_t + \varepsilon_t$$



Sazonalidade com Regressão Linear