Regressão Linear Simples

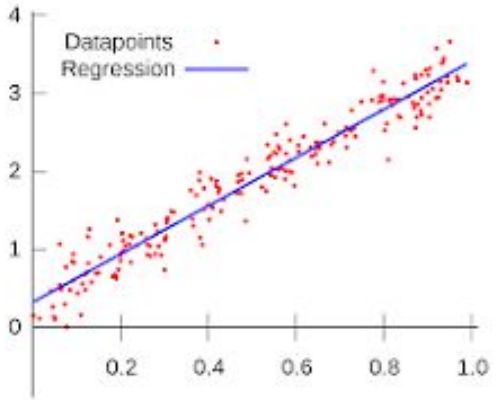
Aprendizado de máquina - Supervisionado

- Dados para treinamento são rotulados
- Modelos tentam prever a saída (nesse caso valores numéricos continuos) e se adaptam, visando minimizar o erro
- Busca criar um modelo que consiga aprender a relação entre as entradas e a saída
- Ao final o modelo deve ser capaz de receber dados não vistos e prever a saída correta

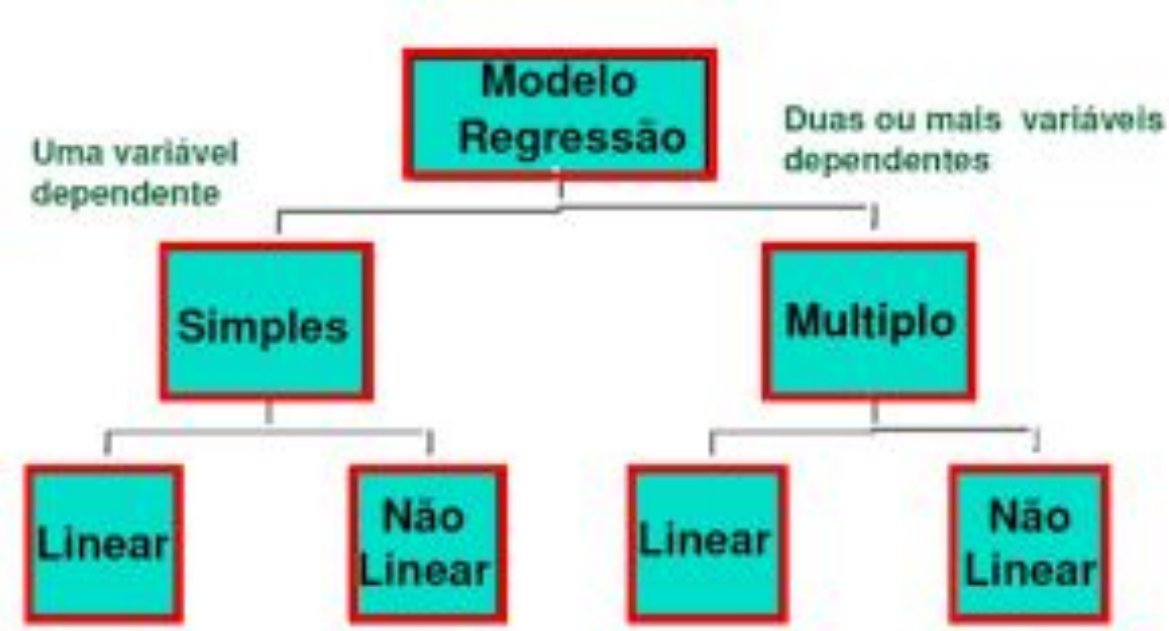
Regressão Linear

- A regressão é um modelo matemático que tenta estudar a relação de uma variável resposta com base em um ou mais características, chamadas de variáveis independentes ou atributos preditivos.

- Existem diversos tipos de regressão (Logística, Polinomial, Lasso, Ridge,...)
- Faz parte do aprendizado supervisionado
- Podemos realizar duas tarefas: **Predição** e **Inferência**



Regressão


Modelo Regressão
Uma variável dependente
Duas ou mais variáveis dependentes
Simples
Multiplo
Linear
Não Linear
Linear
Não Linear

Inferência Estatística

- Tirar conclusões a respeito de uma variável resposta, a partir de variáveis independentes
- Com isso, podemos fazer estimativas e testar hipóteses sobre uma população
- Não podemos inferir causalidade com a regressão
- Ex: Tentar estudar a relação entre a venda no número de pizzas com a população de estudantes em uma cidade. (Ex retirado do livro Estatística Aplicada à ADM/ECONO)

Regressão Linear Simples

- Temos uma variável dependente e uma variável independente
- Tentamos explicar linearmente a variação da VD com a nossa VI
- Foco em inferência

Estimação dos parâmetros

- É necessário estimar os coeficientes, já que não são conhecidos

- Método dos mínimos quadrados: Tentamos minimizar a soma dos erros quadráticos do modelo e com isso chegamos a estimativas dos nosso coeficientes.

- Assim, após estimarmos nossos coeficientes precisamos testar a validade deles.

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2}$$

ESTIMATIVAS

$$b_0 = \bar{y} - b_1 \bar{x}$$

x_i → Valor da variável independente para a i-ésima observação (X[i])
y_i → Valor da variável dependente para a i-ésima observação (y[i])
$\bar{x}$ → Valor médio da variável independente
$\bar{y}$ → Valor médio da variável independente

Métricas de Validação

- Visam verificar o quão satisfatoriamente a equação de regressão estimada ajusta os dados.
- R2 -> Coeficiente de determinação
- Interpretação: Porcentagem da soma total dos quadrados que pode ser explicada usando a reta de regressão. Ou seja, se R2 = 0.9, então 90% da variabilidade da VD pode ser explicada por meio da relação linear com a VI.
- Correlação -> sqrt(R2). Intensidade da associação linear entre duas variáveis. (Leva em conta o sinal da estimação de b1)

Sum of Squares Total →

$$SST = \sum (y - \bar{y})^2$$

Sum of Squares Regression →

$$SSR = \sum (y' - \bar{y'})^2$$

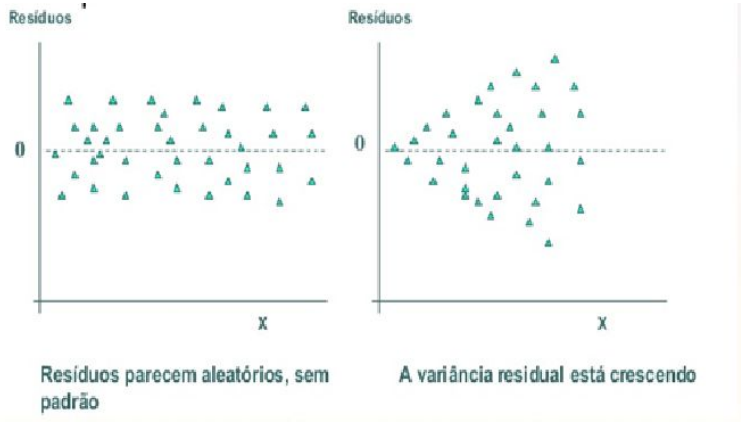
Sum of Squares Error →

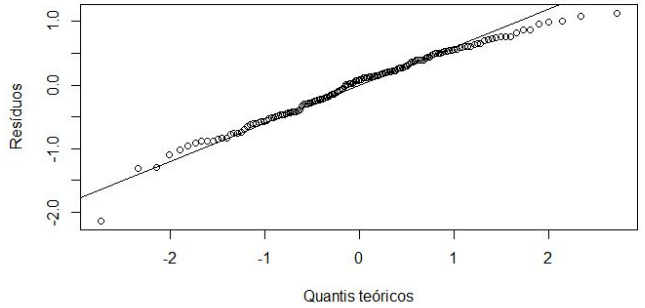
$$SSE = \sum (y - y')^2$$

$$R^2 = \frac{SSR}{SST} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

Suposições do modelo

- Para realizarmos a análise da regressão é necessário fazer algumas suposições sobre o nosso modelo.
  1. O erro é uma variável aleatória com valor esperado igual à 0
  2. Variância do erro é a mesma para todos os valores da variável independente. (Podemos verificar em um grafo que contenha Var(Erro) x VI)
  3. Erro é independente
  4. O erro é normalmente distribuído (Podemos verificar com QQplot)





Teste de Significância

- Utilizado para testar se uma relação de regressão é significativa
- No teste, iremos testar se os coeficientes são iguais a zero ou diferentes de zero
- Como os coeficientes foram estimados anteriormente, precisamos verificar se eles podem ser utilizados
- O teste tem o seguinte formato: (H0: Hip. Nula // H1: Hip. Alternativa)

$$IGC_i = \beta_0 + \beta_1 fastfood_i + u_i$$

$$\begin{cases} H_0: \beta_1 = 0 \longrightarrow \text{o consumo de } \textit{fastfood} \text{ não é relevante para explicar o IGC de um indivíduo} \\ H_A: \beta_1 \neq 0 \longrightarrow \text{o consumo de } \textit{fastfood} \text{ é relevante para explicar o IGC de um indivíduo} \end{cases}$$

Nível de Significância

- Probabilidadade de cometermos o erro do Tipo 1 quando a hipótese nula é verdadeira.
- Valores comuns: 0.01 , 0.05, 0.1
- Interpretação:
  - Caso a nossa hipótese nula seja recusada à um nível de significância de 0.01, isso significa, que a chance de H0 ser falsa é de 99%, ou seja, temos 1% de chance de cometer o erro do tipo 1
  - Caso a nossa hipótese nula seja aceita à um nível de significância de 0.01, não podemos inferir nada a respeito da probabilidade de cometer o erro do tipo 2

	Decisão	
Verdade	Aceitar H_0	Rejeitar H_0
H_0 verdadeiro	— $(1-\alpha)$	Erro Tipo I (α)
H_0 falso	Erro Tipo II β	— $(1-\beta)$

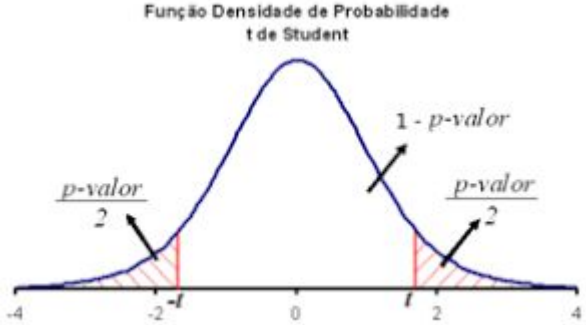
OBS: Como não podemos inferir a respeito do Erro do Tipo 2 o recomendado, caso H0 seja aceito, é falar “não rejeitar H0”

Intervalo de confiança

- Garante que há X%(nível de significância) de chance do coeficiente B1 estar entre dois valores
- Estes valores são: b1 - (margem de erro) e b1 + (margem de erro)
- A margem de erro é o erro de estimação, que envolve a estatística (T ou P) multiplicado pelo desvio padrão estimado de b1
- Caso a o valor hipotético (H0) esteja dentro do intervalo de confiança, não rejeitamos H0, caso contrário, rejeitamos H0

Teste-T

- Teste estatístico para validar uma hipótese
- Desvio padrão da população é desconhecido
- Geralmente utilizado para amostras pequenas (N < 30)
- A estatística de teste segue uma distribuição t-Student com n-2 graus de liberadade
- 2 * valor-p < nivel_significância



Exemplo

VALOR-P

	α						
v	0.40	0.30	0.20	0.15	0.10	0.05	0.025
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.538	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960

	α						
v	0.02	0.015	0.01	0.0075	0.005	0.0025	0.0005
1	15.894	21.205	31.821	42.433	63.656	127.321	636.578
2	4.849	5.643	6.965	8.073	9.925	14.089	31.600
3	3.482	3.896	4.541	5.047	5.841	7.453	12.924
4	2.999	3.298	3.747	4.088	4.604	5.598	8.610
5	2.757	3.003	3.365	3.634	4.032	4.773	6.869
6	2.612	2.829	3.143	3.372	3.707	4.317	5.959
7	2.517	2.715	2.998	3.203	3.499	4.029	5.408
8	2.449	2.634	2.896	3.085	3.355	3.833	5.041
9	2.398	2.574	2.821	2.998	3.250	3.690	4.781
10	2.359	2.527	2.764	2.932	3.169	3.581	4.587
11	2.328	2.491	2.718	2.879	3.106	3.497	4.437
12	2.303	2.461	2.681	2.836	3.055	3.428	4.318
13	2.282	2.436	2.650	2.801	3.012	3.372	4.221
14	2.264	2.415	2.624	2.771	2.977	3.326	4.140
15	2.249	2.397	2.602	2.746	2.947	3.286	4.073
16	2.235	2.382	2.583	2.724	2.921	3.252	4.015
17	2.224	2.368	2.567	2.706	2.898	3.222	3.965
18	2.214	2.356	2.552	2.689	2.878	3.197	3.922
19	2.205	2.346	2.539	2.674	2.861	3.174	3.883
20	2.197	2.336	2.528	2.661	2.845	3.153	3.850
21	2.189	2.328	2.518	2.649	2.831	3.135	3.819
22	2.183	2.320	2.508	2.639	2.819	3.119	3.792
23	2.177	2.313	2.500	2.629	2.807	3.104	3.768
24	2.172	2.307	2.492	2.620	2.797	3.091	3.745
25	2.167	2.301	2.485	2.612	2.787	3.078	3.725
26	2.162	2.296	2.479	2.605	2.779	3.067	3.707
27	2.158	2.291	2.473	2.598	2.771	3.057	3.689
28	2.154	2.286	2.467	2.592	2.763	3.047	3.674
29	2.150	2.282	2.462	2.586	2.756	3.038	3.660
30	2.147	2.278	2.457	2.581	2.750	3.030	3.646
40	2.123	2.250	2.423	2.542	2.704	2.971	3.551
60	2.099	2.223	2.390	2.504	2.660	2.915	3.460
120	2.076	2.196	2.358	2.468	2.617	2.860	3.373
∞	2.054	2.170	2.326	2.432	2.576	2.807	3.290

Exemplo prático

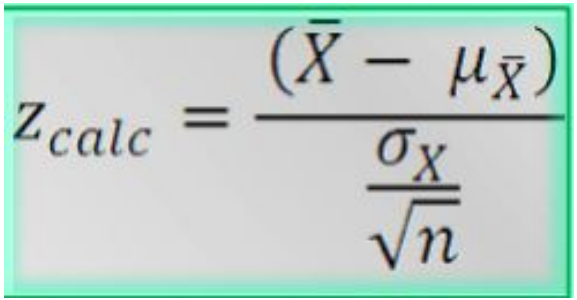
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	38.6517	9.456	4.087	0.000	19.826	57.478
Region[T.E]	-15.4278	9.727	-1.586	0.117	-34.793	3.938
Region[T.N]	-10.0170	9.260	-1.082	0.283	-28.453	8.419
Region[T.S]	-4.5483	7.279	-0.625	0.534	-19.039	9.943
Region[T.W]	-10.0913	7.196	-1.402	0.165	-24.418	4.235
Literacy	-0.1858	0.210	-0.886	0.378	-0.603	0.232
Wealth	0.4515	0.103	4.390	0.000	0.247	0.656

Omnibus:	3.049	Durbin-Watson:	1.785
Prob(Omnibus):	0.218	Jarque-Bera (JB):	2.694
Skew:	-0.340	Prob(JB):	0.260
Kurtosis:	2.454	Cond. No.	371.

Onto our coefficients!

Teste-Z

- Teste estatístico para validar uma hipótese
- Desvio padrão da população é conhecido
- Geralmente utilizado para amostras grandes (N > 30)
- A distribuição do teste estatístico sob a hipótese nula pode ser aproximada por uma distribuição normal
- 2(valor-z) < nivel_significância



/X = Média amostral
ux = Valor hipotético

sigma_x = Desvio padrão
n = Tamanho amostral

Exemplo

z	0,09	0,08	0,07	0,06	0,05	0,04	0,03	0,02	0,01	0

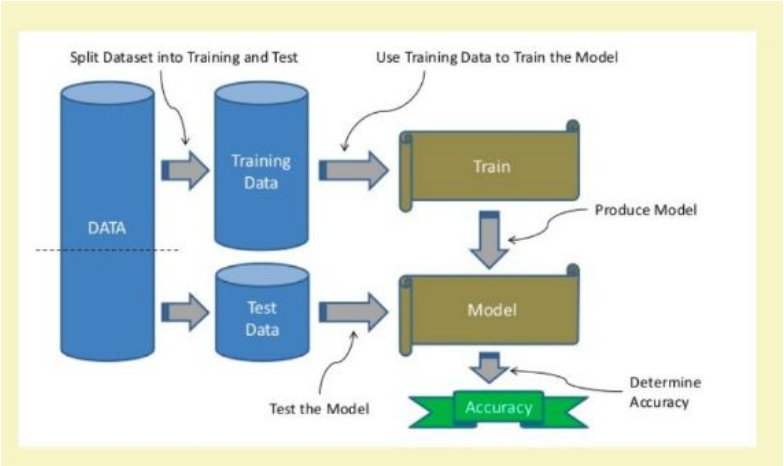
-3,7										
-3,6	0,00011	0,00012	0,00012	0,00013	0,00013	0,00014	0,00014	0,00015	0,00015	0,00016
-3,5	0,00017	0,00017	0,00018	0,00019	0,00019	0,0002	0,00021	0,00022	0,00022	0,00023
-3,4	0,00024	0,00025	0,00026	0,00027	0,00028	0,00029	0,0003	0,00031	0,00032	0,00034
-3,3	0,00035	0,00036	0,00038	0,00039	0,0004	0,00042	0,00043	0,00045	0,00047	0,00048
-3,2	0,0005	0,00052	0,00054	0,00056	0,00058	0,0006	0,00062	0,00064	0,00066	0,00069
-3,1	0,00071	0,00074	0,00076	0,00079	0,00082	0,00084	0,00087	0,0009	0,00094	0,00097
-3	0,001	0,00104	0,00107	0,00111	0,00114	0,00118	0,00122	0,00126	0,00131	0,00135
-2,9	0,00139	0,00144	0,00149	0,00154	0,00159	0,00164	0,00169	0,00175	0,00181	0,00187
-2,8	0,00193	0,00199	0,00205	0,00212	0,00219	0,00226	0,00233	0,0024	0,00248	0,00256
-2,7	0,00264	0,00272	0,0028	0,00289	0,00298	0,00307	0,00317	0,00326	0,00336	0,00347
-2,6	0,00357	0,00368	0,00379	0,00391	0,00402	0,00415	0,00427	0,0044	0,00453	0,00466
-2,5	0,0048	0,00494	0,00508	0,00523	0,00539	0,00554	0,0057	0,00587	0,00604	0,00621
-2,4	0,00639	0,00657	0,00676	0,00695	0,00714	0,00734	0,00755	0,00776	0,00798	0,0082
-2,3	0,00842	0,00866	0,00889	0,00914	0,00939	0,00964	0,0099	0,01017	0,01044	0,01072
-2,2	0,01101	0,0113	0,0116	0,01191	0,01222	0,01255	0,01287	0,01321	0,01355	0,0139
-2,1	0,01426	0,01463	0,015	0,01539	0,01578	0,01618	0,01659	0,017	0,01743	0,01786
-2	0,01831	0,01876	0,01923	0,0197	0,02018	0,02068	0,02118	0,02169	0,02222	0,02275
z	0,09	0,08	0,07	0,06	0,05	0,04	0,03	0,02	0,01	0
-1,9	0,0233	0,02385	0,02442	0,025	0,02559	0,02619	0,0268	0,02743	0,02807	0,02872
-1,8	0,02938	0,03005	0,03074	0,03144	0,03216	0,03288	0,03362	0,03438	0,03515	0,03593
-1,7	0,03673	0,03754	0,03836	0,0392	0,04006	0,04093	0,04182	0,04272	0,04363	0,04457
-1,6	0,04551	0,04648	0,04746	0,04846	0,04947	0,0505	0,05155	0,05262	0,0537	0,0548
-1,5	0,05592	0,05705	0,05821	0,05938	0,06057	0,06178	0,06301	0,06426	0,06552	0,06681
-1,4	0,06811	0,06944	0,07078	0,07215	0,07353	0,07493	0,07636	0,0778	0,07927	0,08076
-1,3	0,08226	0,08379	0,08534	0,08692	0,08851	0,09012	0,09176	0,09342	0,0951	0,0968
-1,2	0,09853	0,10027	0,10204	0,10383	0,10565	0,10749	0,10935	0,11123	0,11314	0,11507
-1,1	0,11702	0,119	0,121	0,12302	0,12507	0,12714	0,12924	0,13136	0,1335	0,13567
-1	0,13786	0,14007	0,14231	0,14457	0,14686	0,14917	0,15151	0,15386	0,15625	0,15866
-0,9	0,16109	0,16354	0,16602	0,16853	0,17106	0,17361	0,17619	0,17879	0,18141	0,18406
-0,8	0,18673	0,18943	0,19215	0,19489	0,19766	0,20045	0,20327	0,20611	0,20897	0,21186
-0,7	0,21476	0,2177	0,22065	0,22363	0,22663	0,22965	0,2327	0,23576	0,23885	0,24196
-0,6	0,2451	0,24825	0,25143	0,25463	0,25785	0,26109	0,26435	0,26763	0,27093	0,27425
-0,5	0,2776	0,28096	0,28434	0,28774	0,29116	0,2946	0,29806	0,30153	0,30503	0,30854
-0,4	0,31207	0,31561	0,31918	0,32276	0,32636	0,32997	0,3336	0,33724	0,3409	0,34458
-0,3	0,34827	0,35197	0,35569	0,35942	0,36317	0,36693	0,3707	0,37448	0,37828	0,38209
-0,2	0,38591	0,38974	0,39358	0,39743	0,40129	0,40517	0,40905	0,41294	0,41683	0,42074
-0,1	0,42465	0,42858	0,43251	0,43644	0,44038	0,44433	0,44828	0,45224	0,4562	0,46017
0	0,46414	0,46812	0,4721	0,47608	0,48006	0,48405	0,48803	0,49202	0,49601	0,5
z	0,09	0,08	0,07	0,06	0,05	0,04	0,03	0,02	0,01	0

Cuidados na Regressão Linear Simples

- Rejeitar H0, e concluir que a relação entre x e y é significativa **não nos permite concluir uma relação de causa efeito**
- Não podemos concluir que alterações em x provocam alterações em y simplesmente por encontrar relações significativas
- **Não podemos concluir que a relação entre x e y seja linear**
- Podemos afirmar que x e y estão relacionados e que temos uma relação linear que explica a parte significativa da variabilidade em y ao longo dos valores observados de x

Tópico Adicional

- Data Leakage: Vazar informações do teste para o treino ou vice-versa
- Enviesar o modelo
- Vaza informações do treino para o teste, que na teoria não deveria ter contato com o treino



Casos

- Preencher valores missing com a média da coluna
- Qualquer encoder que leve em consideração a classe do objeto
- Métodos de normalização (Min Max, Standart,...)

Aplicação

```python
# Escalonamento dos dados usando StandardScaler
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Divisão dos dados em treinamento e teste
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)
```

```python
# Divisão dos dados em treinamento e teste
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Escalonamento dos dados usando StandardScaler após a divisão
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```