UNIVERSIDADE
FEDERAL DO CARIRI

UNIVERSIDADE FEDERAL DO CARIRI
CENTRO DE CIÊNCIAS E TECNOLOGIA
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Probabilidade e Estatística - Unidade IV - Projeto Rosilda Benício de Souza

Wanderson Faustino Patricio
João Isaac Alves Farias
Karla Mikaelly Paz de Almeida

Juazeiro do Norte, 15 de Julho de 2023

1 Apresentação dos dados

Durante várias etapas da programação é necessário executar buscas em bancos de dados. Existem vários algoritmos de busca, cada um deles tendo um desempenho específico e um tempo de execução característico, que pode variar dependendo do sistema operacional, da memória do computador utilizado para executar o algoritmo, da ordem dos elementos, etc.

Dentre os parâmetros que podem alterar o tempo de execução de uma busca está o tamanho da entrada fornecida, ou seja, a quantidade de elementos que deverão ser visitados até ser encontrado o elemento desejado.

Para esse estudo analisaremos o tempo de execução de um algoritmo de busca linear em uma lista não ordenada. Como o funcionamento do algoritmo baseia-se em visitar todos os elementos da lista um a um, em sequência, até encontrar o elemento chave, caso a lista não possua o número tido como parâmetro da procura, o programa terá que visitar todos os elementos da lista até terminar sua execução. Desta forma, considerando que a visita a um elemento seja realizada em um tempo constante, o tempo de execução total do algoritmo dependerá linearmente do tamanho da lista.

Escrevendo em notação "big O"temos:

$$T(n) \in O(n) \Rightarrow T(n) \approx A + B \cdot n$$

Onde $T(n)$ é o tempo que o algoritmo leva para varrer uma lista de tamanho n.

Para minimizar possíveis erros devido a outros fatores além do tamanho da entrada faremos o valor de n "varrer valores grandes"($n \geq 10^6$).

Executando o algoritmo e registrando os resultados exibidos pelo compilador podemos montar a seguinte tabela:

n (em milhões)	Tempo (s)
1	0.8850
2	0.9041
3	0.9262
4	0.9489
5	0.9710
6	0.9768
7	0.9940
8	1.0210
9	1.0310
10	1.0530
11	1.0870
12	1.0820
13	1.1100
14	1.1210
15	1.1510
16	1.1750
17	1.1860
18	1.2230
19	1.2360
20	1.2480

n (em milhões)	Tempo (s)
21	1.2680
22	1.2840
23	1.3130
24	1.3200
25	1.3410
26	1.3570
27	1.3760
28	1.4020
29	1.4020
30	1.4430
31	1.4620
32	1.4440
33	1.4970
34	1.5020
35	1.5110
36	1.5360
37	1.5470
38	1.5640
39	1.6760
40	1.6080

Tabela T x n (dados coletados pela equipe)

2 Gráfico de Dispersão

Plotando o gráfico de dispersão dos pontos teremos

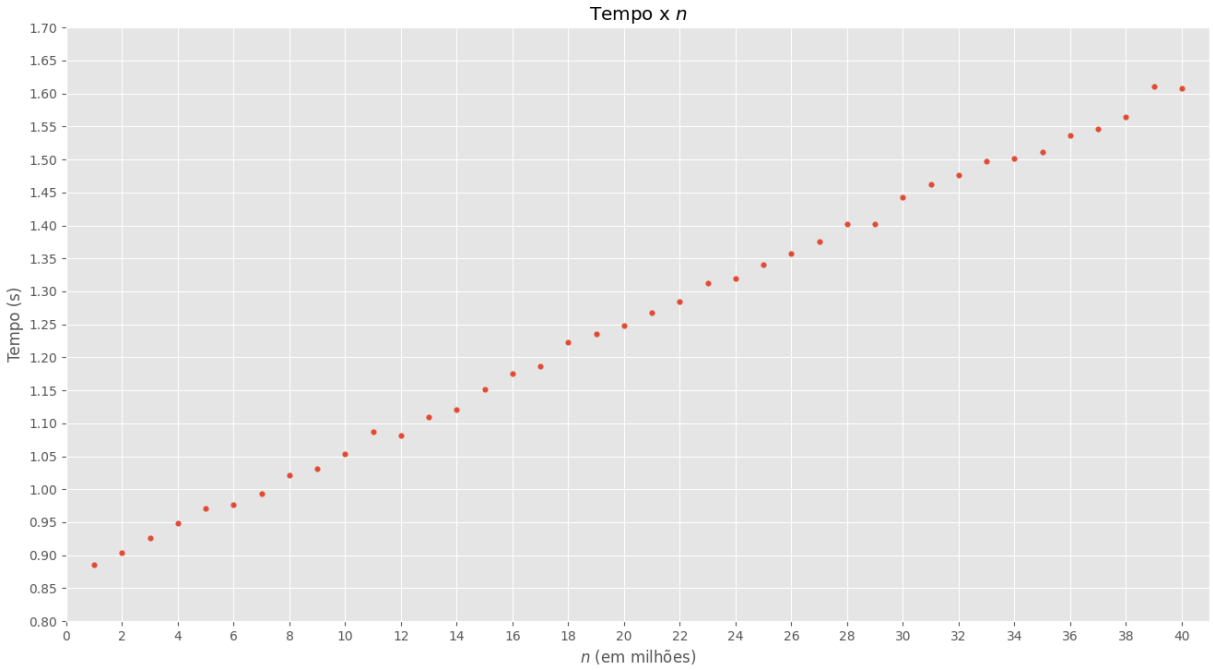

Figura 1: Gráfico de dispersão dos pontos T x n

Através da análise gráfica vemos indícios que as duas variáveis estão correlacionadas entre si.

Ademais, o gráfico mostra uma tendência linear positiva com baixa dispersão entre os pontos, ou seja, os pontos encontrados através do experimento estão de maneira bem ajustada a uma reta. Portanto, podemos inferir que o tempo de execução do algoritmo possui dependência linear com o tamanho da entrada.

considerando o tamanho n como a variável x, o tempo de execução como a variável y e N a quantidade de pontos, podemos calcular o coeficiente de correlação linear de Pearson para a distribuição:

$$r = \frac{N \cdot \sum(xy) - \left(\sum x\right) \cdot \left(\sum y\right)}{\sqrt{N \cdot \left(\sum x^2\right) - \left(\sum x\right)^2} \cdot \sqrt{N \cdot \left(\sum y^2\right) - \left(\sum y\right)^2}} = 0,99921$$

Como o coeficiente de correlação tende a 1 percebemos uma correlação muito forte.

3 Teste de Hipótese para a correlação

Consideremos como hipótese nula a proposição de que não existe correlação entre as variáveis, e hipótese alternativa a hipótese que a correlação entre elas é positiva.

$$\begin{cases} H_o: \ \rho = 0 \\ H_1: \ \rho > 0 \end{cases}$$

Calculando o t através da fórmula

$$t_{ob} = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

encontramos

$$t_{ob} = 155,44$$

A um nível de significância de 1% para 38 graus de liberdade e um teste bilateral a direita temos um t de student tabelado de

$$t_c = 2,423$$

Como $t_{ob} >> t_c$ rejeitamos a hipótese nula, e concluímos que as variáveis possuem correlação positiva.

OBS: Para o t_c encontramos um coeficiente de correlação mínimode

$$r_c = 0,36582$$

que está bem abaixo de r, o que colabora para a rejeição da hipótese nula.

4 Equação de regressão

A equação de regressão é uma aproximação paraos pontos a uma reta, de tal forma que os erros em relação aos pontos experimentais seja minimizado. Tal equação é dada pela reta

$$\hat{T} = a + b \cdot n$$

Com

$$b = \frac{\sum x(y-\bar{y})}{\sum x(y-\bar{y})} = 0,000000187$$

e

$$a = \bar{y} - b\bar{x} = 0,8709$$

Teremos, portanto, uma aproximação para o tempo de execução em função da entrada dada por

$$\hat{T}(n) = 0,8709 + 0,000000187 \cdot n$$

Colocando a reta de regressão e os pntos experimentais em um gráfico temos:

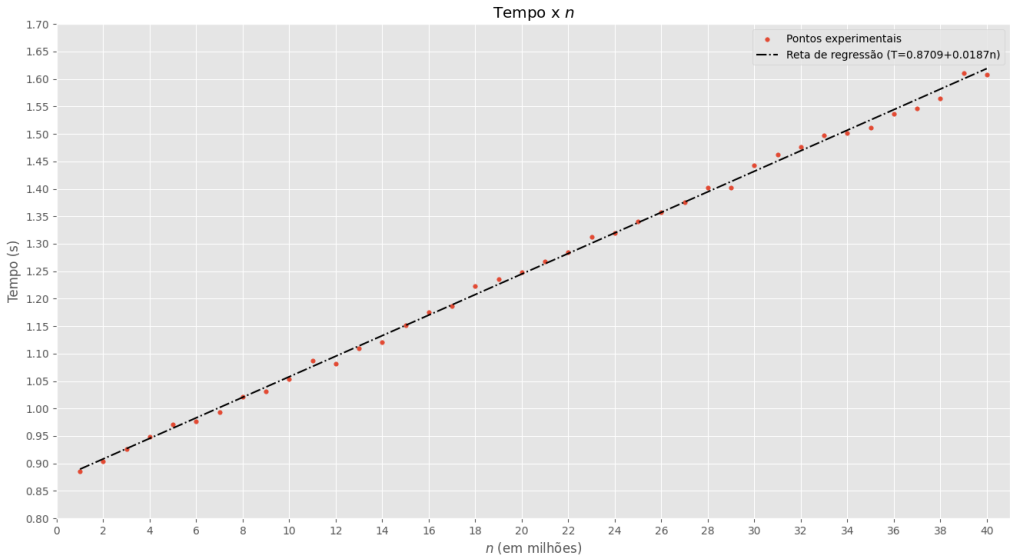

Figura 2: Reta ajustada

Fazendo uma tabela com o erro relativo para cada ponto em relação à reta:

n (em milhões)	$e_i = y_i - (a + bx_i)$ (s)
1	-0.0046
2	-0.0042
3	-0.0008
4	0.0032
5	0.0066
6	-0.0063
7	-0.0078
8	0.0005
9	-0.0082
10	-0.0049
11	0.0104
12	-0.0133
13	-0.0040
14	-0.0117
15	-0.0004
16	0.0049
17	-0.0028
18	0.0155
19	0.0098
20	0.0031

n (em milhões)	$e_i = y_i - (a + bx_i)$ (s)
21	0.0044
22	0.0017
23	0.0120
24	0.0003
25	0.0026
26	-0.0001
27	0.0002
28	0.0075
29	-0.0112
30	0.0111
31	0.0114
32	0.0077
33	0.0090
34	-0.0047
35	-0.0144
36	-0.0081
37	-0.0158
38	-0.0175
39	0.0108
40	-0.0109

Através da tabela percebemos que o erro de cada medida é muito pequeno em relação aos tempos medidos.

5 Análise da variável independente

Calculando o coeficiente de determinação encontramos

$$R^2 = \frac{\sum(a + bx - \bar{y})^2}{\sum(y - \bar{y})^2} = 0,99843 = 99,843\%$$

Concluímos que a variância no tempo de execução do programa depende $99,843\%$ do tamanho da entrada e $0,157\%$ devido a outros fatores intervenientes.

5.1 Teste ANOVA para o b

Consideremos como hipótese nula a proposição de que não existe dependência do tempo de execução com o tamanho da entrada, e hipótese alternativa a hipótese que há dependência.

$$\begin{cases} H_o: \ b = 0 \\ H_1: \ b \neq 0 \end{cases}$$

Analisando as somas dos quadrados e a razão f encontramos

Fonte de Variação	SQ	gl	QM	f
Regressão	SQR = 1.8592017	1	QMR = 1.8592017	f=710720.03
Erro	SQE = 0.0000994	38	QME = 0.0000026	

O f tabelado para 1% de significência, com 1 grau de liberdade no numerador e 38 no denominador é

$$f_c = 7,31$$

Como $f >> f_c$ rejeitamos H_o e concluímos que o tempo tem dependência com o tamanho da entrada.

6 Considerações finais

A partir das análises gráficas e dos testes de hipóteses podemos concluir ,com confiança de 99%, que o tempo de execução de um algoritmo de busca linear sobre um vetor não ordenado depende linearmente do tamanho da entrada fornecida. Portanto, consideramos que a variável independente escolhida (tamanho da amostra) é extremamente significante para o nosso modelo.