

Entrega 1 - Exploratory Data Analysis and Linear Regression

Metodologias Experimentais em Informática | 2022/2023 | 1.º Semestre

Eva Teixeira	n.º 2019215185	uc2019215185@student.uc.pt	PL1
João Dionísio	n.º 2019217030	uc2019217030@student.uc.pt	PL2
João Oliveira	n.º 2022184283	uc2022184283@student.uc.pt	PL4

1. Introdução

No âmbito da unidade curricular de Metodologias Experimentais em Informática foi proposto aos alunos a elaboração de um trabalho prático para consolidar e a pôr em prática os conceitos abordados nas aulas teóricas e práticas.

Posto isto, foi-nos sugerido que efetuássemos uma avaliação do desempenho de três algoritmos diferentes para o problema do *Maximum Flow*. O *Maximum Flow problem* consiste em encontrar, para um determinado grafo, o vazamento máximo que pode ser enviado de um vértice para o outro. Um grafo é caracterizado pela capacidade de cada arco e o máximo de vazamento, sendo que o número de arestas e vértices é variável.

2. Declaração do problema

É necessário, para grafos equiparados (com o mesmo número de arestas e de vértices), testar o desempenho de três algoritmos diferentes - Dinic, EK, MPM. Para tal, é necessário obter métricas que refletem a velocidade, tais como o tempo de execução, na obtenção da capacidade máxima de vazamento.

3. Variáveis

Para este problema, existe dois tipos principais de variáveis:

- Variáveis dependentes
 - Tempo de execução
- Variáveis independentes
 - Número de vértices de um grafo
 - Probabilidade do arco
 - Capacidade máxima

4. Cenário e dados recolhidos

Para o procedimento experimental, foi necessário:

- 1) Gerar grafos aleatórios com parâmetros diferentes:
 - a) número de vértices: variando entre 50 e 950 em incrementos de 50;
 - b) probabilidade do arco: variando entre 10 a 100 em incrementos de 10;
 - c) para cada incremento de probabilidade do arco foram gerados 30 grafos diferentes (com base na *seed*), para cada par (número de vértice, probabilidade de arco).

- 2) Execução dos testes: para cada um dos três algoritmos, foram obtidos os valores de tempo de execução que cada algoritmo demora para os diferentes grafos;
 - a) para um determinado número de vértices e probabilidade de arco de um grafo, foi feita a média do tempo de execução dos 30 grafos equiparados visto a medição em causa ser não determinística.
 - b) Análise exploratória dos dados.

Os testes foram realizados numa máquina com processador i7-9750H (6 cores, 12 threads), 16 GB de RAM. A velocidade base de cada núcleo do processador é 2.6GHz, com Turbo Boost até 4.5GHz.

5. Análise e discussão

A geração dos grafos teve como parâmetros: valores de vértice de 50 a 950 em incrementos de 50; probabilidade de arco de 10 a 100 em incrementos de 10 e seeds aleatórias.

Para estes testes decidimos variar a seed, probabilidade de arco e o número de vértices. Para cada valor de vértice e probabilidade de arco diferente, foram utilizados 30 grafos com seeds diferentes. A média do tempo de execução de cada conjunto de 30 grafos representa o *runtime* do par valor de vértice e probabilidade de arco, para obter valores mais concretos.

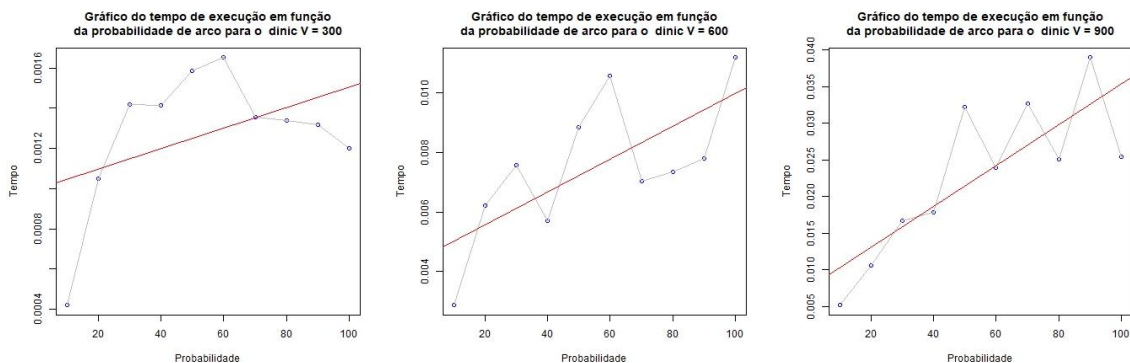
Para melhor interpretar os resultados obtidos, foram geradas regressões lineares e gráficos para valores de vértices de 300, 600 e 900 e para valores de probabilidade de arco de 20, 50, 70 e 100. Para calcular a complexidade dos algoritmos em função da probabilidade do arco, foi utilizada a seguinte fórmula: $A = P \times V \times \frac{V-1}{2}$

Dinic

Expectativa: Sabe-se que o algoritmo *Dinic* tem uma complexidade temporal de $O(V^2 \times A)$. Substituindo A com a fórmula anterior, obtém-se $O(V^2 \times P \times V \times \frac{V-1}{2}) = \dots = O(V^4 \times P)$. Com isto, espera-se que o número de vértices tenha um comportamento biquadrático em função do tempo, e que a variação da probabilidade do arco tenha um comportamento linear.

Probabilidade de arco - transformação linear

Os gráficos abaixo representam a variação do tempo em função da probabilidade do arco dos grafos, para os valores de vértices de 300, 600 e 900.



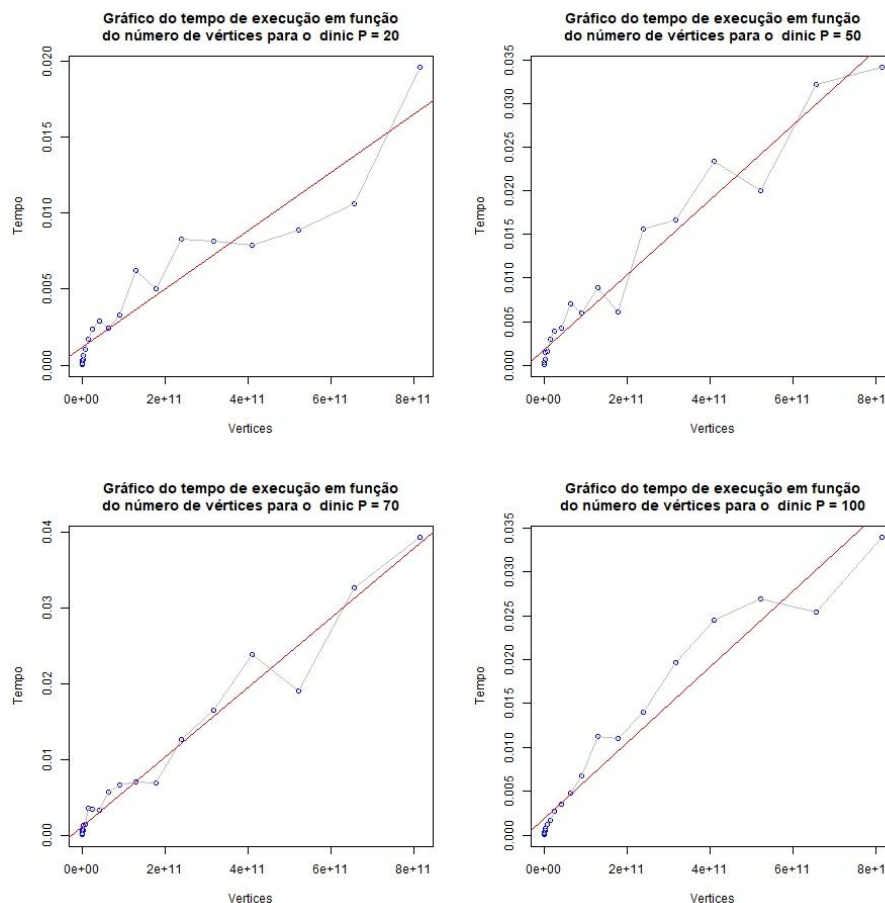
(Figura 1: Gráficos do tempo de execução em função da probabilidade de arco para $V = 300, 600$ e 900)

Residual standard error: 0.001825 on 8 degrees of freedom
Multiple R-squared: 0.484, Adjusted R-squared: 0.4196
F-statistic: 7.505 on 1 and 8 DF, p-value: 0.02547

(Figura 2: Regressão Linear para V = 600)

Estes valores comprovam que a transformação linear não é a transformação mais adequada para este caso.

Número de vértices - Transformação biquadrática



(Figura 3: Gráficos do tempo de execução em função do número de vértices para valores de P = 20, 50, 70 e 100)

Residual standard error: 0.002311 on 17 degrees of freedom
Multiple R-squared: 0.9571, Adjusted R-squared: 0.9545
F-statistic: 378.9 on 1 and 17 DF, p-value: 4.658e-13

(Figura 4: Regressão linear da transformação biquadrática para P = 50)

Aplicando a transformação biquadrática à regressão linear, observa-se que os valores estão conforme o esperado e que o aumento do número de vértices traduz-se num aumento de tempo elevado à quarta.

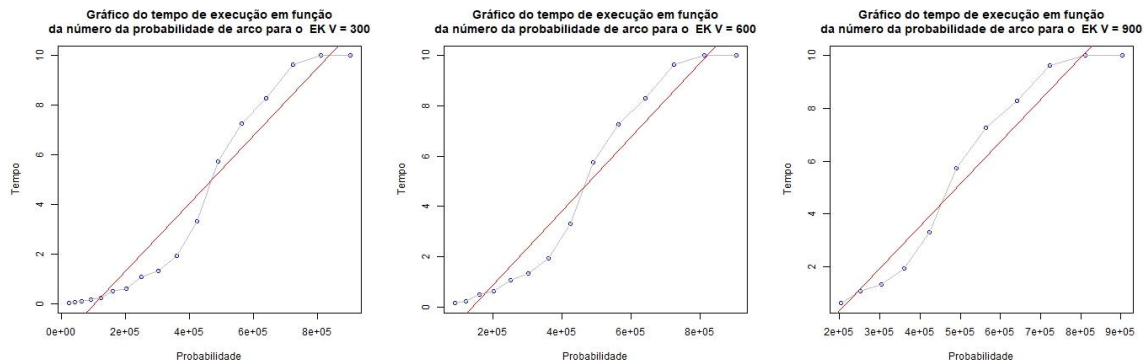
O erro padrão residual apresenta um valor muito baixo, o que nos leva a concluir que a regressão atual se ajusta aos dados do gráfico. Além disso, o R^2 apresenta valores acima de 0.95, revelando um elevado ajuste dos pontos do gráfico à linha de regressão utilizada.

Edmonds-Karp (EK)

Expectativa: Sabendo que o EK tem uma complexidade de $O(A^2 \times V)$, o que corresponde a $O((P \times (V \times (V - 1)/2))^2 \times V) = O(V^5 \times P^2)$, espera-se que o número de vértices tenham um comportamento quártico em função do tempo e que a probabilidade de arco tenha um comportamento quadrático.

Resultados

Probabilidade de arco - Transformação quadrática



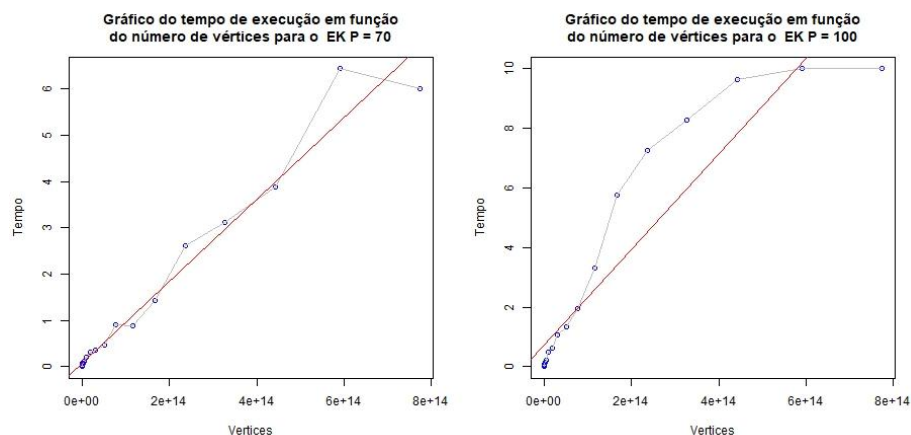
(Figura 5: Gráficos do tempo de execução em função da probabilidade de arco para V = 300, 600 e 900)

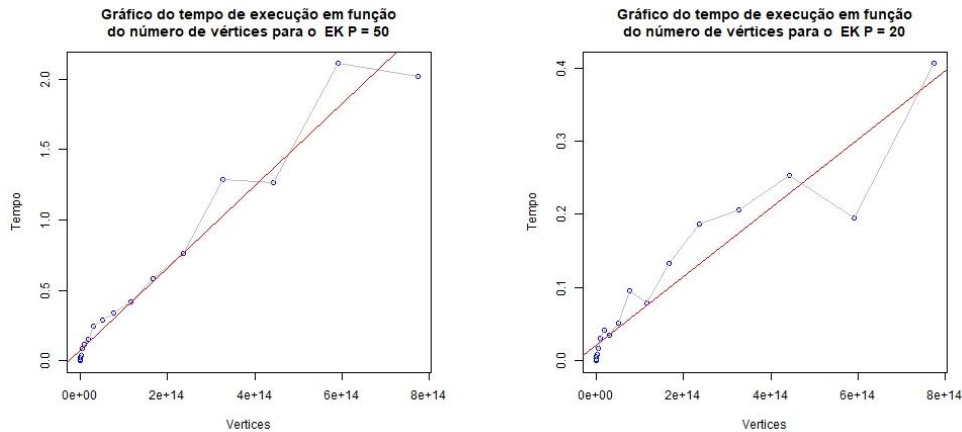
Residual standard error: 0.909 on 12 degrees of freedom
Multiple R-squared: 0.9523, Adjusted R-squared: 0.9484
F-statistic: 239.7 on 1 and 12 DF, p-value: 2.705e-09

(Figura 6: Regressão linear da transformação quadrática para V = 600)

Esta regressão linear apresenta um coeficiente de determinação de 0.948, um valor elevado que comprova que a reta se ajusta aos pontos do gráfico. O Residual *standard error* não apresenta um valor infinitamente baixo, mas, ainda assim, também não é muito elevado.

Número de vértices- Transformação quártica





(Figura 7: Gráficos do tempo de execução em função do número de vértices para valores de P = 20, 50, 70 e 100)

Residual standard error: 0.1351 on 17 degrees of freedom
Multiple R-squared: 0.962, Adjusted R-squared: 0.9598
F-statistic: 430.8 on 1 and 17 DF, p-value: 1.63e-13

(Figura 8: Regressão linear da transformação quártica para P=50)

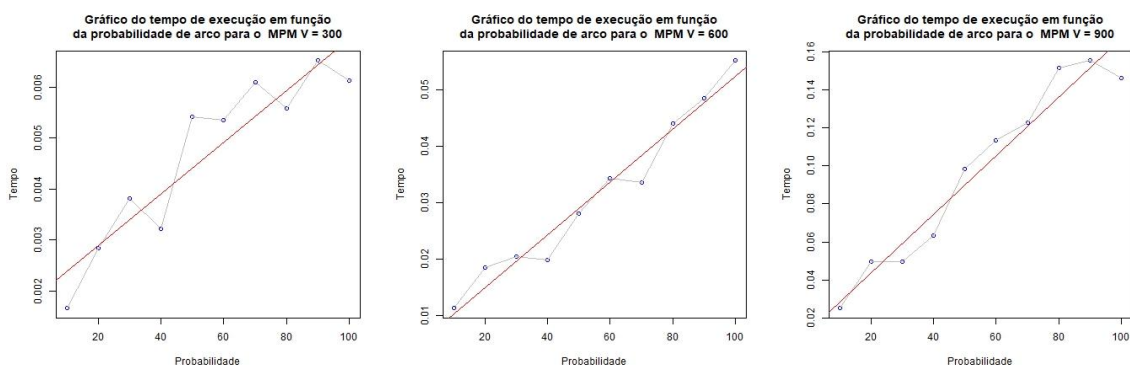
Através dos gráficos observados e dos valores registados após a transformação quártica, verifica-se que esta regressão se adapta aos valores. O valor de R^2 é, relativamente à regressão linear, o que revela que esta reta de regressão se ajusta aos valores do gráfico.

Malhotra, Pramod-Kumar and Maheshwari (MPM)

Expectativas: Sabendo que o MPM tem uma complexidade de $O(V^3)$, espera-se que o número de vértices tenha um comportamento cúbico em função do tempo e que a probabilidade de arco tenha, na teoria, um comportamento constante.

Resultados

Probabilidade de arco - Transformação linear



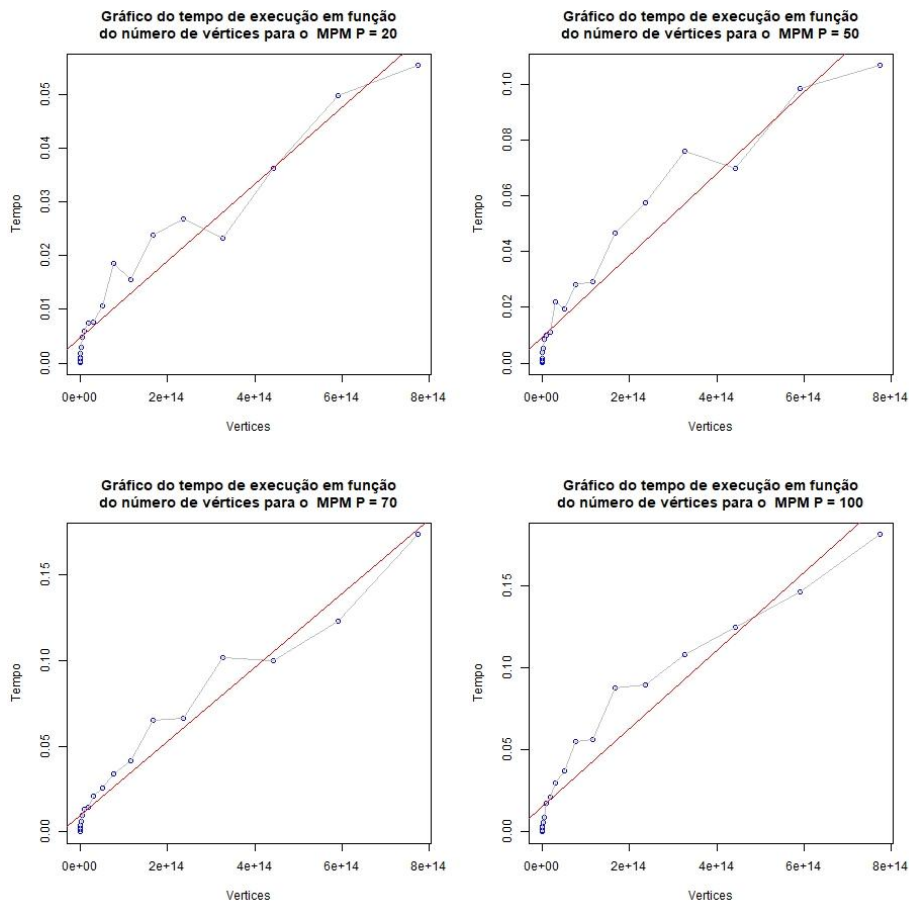
(Figura 9: Gráficos do tempo de execução em função da probabilidade de arco para V = 300, 600 e 900)

Residual standard error: 0.002903 on 8 degrees of freedom
Multiple R-squared: 0.9638, Adjusted R-squared: 0.9593
F-statistic: 213 on 1 and 8 DF, p-value: 4.764e-07

(Figura 10: Regressão linear da transformação cúbica para $V = 600$)

Ao contrário do esperado, a regressão linear do tempo em função da probabilidade de arco é adequada, sendo o valor do R^2 muito próximo de 1, o que pode ser justificado se a complexidade for, por exemplo, $O(V^3 + P)$. Isto faria com que P , apesar de ter alguma influência linear, fosse desprezado aquando do cálculo da complexidade.

Número de vértices - Transformação cúbica



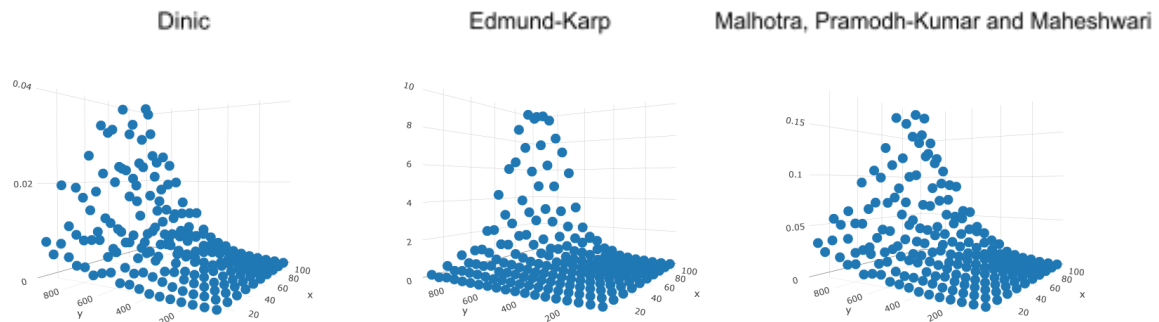
(Figura 11: Gráficos do tempo de execução em função do número de vértices para valores de $P = 20, 50, 70$ e 100)

Residual standard error: 0.009229 on 17 degrees of freedom
Multiple R-squared: 0.9321, Adjusted R-squared: 0.9281
F-statistic: 233.4 on 1 and 17 DF, p-value: 2.314e-11

(Figura 12: Regressão linear da transformação cúbica para $P=50$)

Aplicando-se uma transformação cúbica na regressão do MPM, tal como esperado, ajusta-se perfeitamente aos valores do gráfico, como pode ser comprovado pelo valor do coeficiente de determinação (0,9281). Esta transformação possui um baixo valor de erro padrão residual comparativamente ao mesmo da regressão linear.

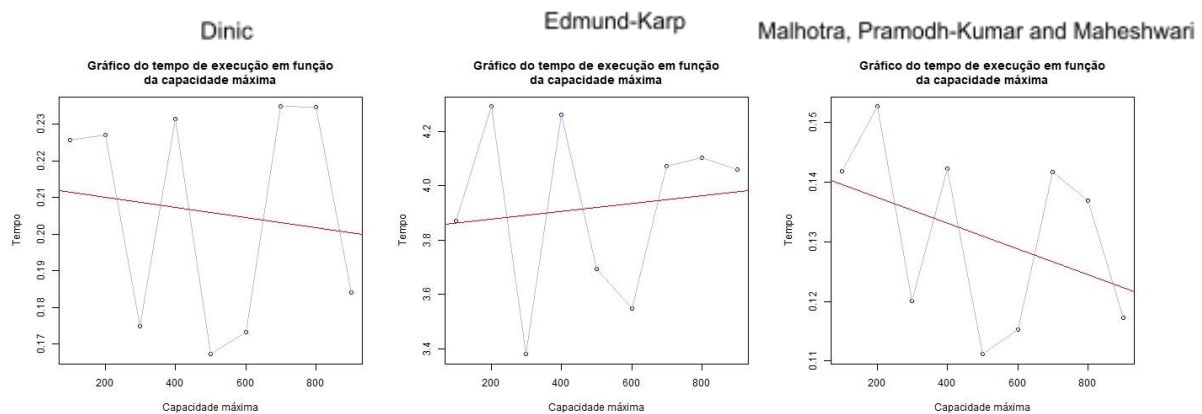
6. Gráfico Tridimensional



(Figura 13: Gráficos do tempo de execução em função do número de vértices e da probabilidade de arco)

Através dos gráficos tridimensionais ilustrados acima, podemos comprovar algumas das conclusões obtidas nos plots observados e analisados anteriormente. O algoritmo Dinic, apesar de não ser tão linear em função da probabilidade de arco, é o mais rápido, seguindo-se o MPM e só por fim o EK. Apesar do EK ter um tempo de execução, em média, superior aos restantes, é muito mais previsível. É também possível observar que, apesar do tempo de execução do algoritmo MPM apenas ser influenciado, teoricamente, pelo número de vértices, na realidade verificou-se que o aumento da probabilidade de arco também influencia o tempo.

7. Capacidade máxima



(Figura 11: Gráficos do tempo de execução em função da capacidade máxima para valores de $V = 1000$, $P = 50\%$)

Dinic

Residual standard error: 0.0317 on 7 degrees of freedom
Multiple R-squared: 0.01553, Adjusted R-squared: -0.1251
F-statistic: 0.1104 on 1 and 7 DF, p-value: 0.7494

Edmund-Karp

Residual standard error: 0.3375 on 7 degrees of freedom
Multiple R-squared: 0.01516, Adjusted R-squared: -0.1255
F-statistic: 0.1077 on 1 and 7 DF, p-value: 0.7523

Malhotra, Pramodh-Kumar and Maheshwari

Residual standard error: 0.01479 on 7 degrees of freedom
Multiple R-squared: 0.1541, Adjusted R-squared: 0.0333
F-statistic: 1.276 on 1 and 7 DF, p-value: 0.2959

Observando estes gráficos e as respectivas regressões lineares, podemos concluir que muito provavelmente não existe qualquer correlação entre o tempo e a capacidade máxima.

8. Conclusão

No final da realização de todos os testes e respetiva análise dos dados e gráficos gerados, concluímos que, entre os 3 algoritmos analisados, o Dinic é o mais rápido. Além disso, conseguimos comprovar as expectativas que tínhamos antes de gerar os valores e gráficos, relativamente ao desempenho (tempo de execução) e ao comportamento dos algoritmos e das variáveis independentes de cada um.