

Metodologias Experimentais em Informática

Regressão Linear

Artur Coutinho - 2014230432 — Diogo Amores - 2015231975
Maria Inês Roseiro - 2015233281

20 de Dezembro de 2019

1 Introdução

Este relatório foi desenvolvido no âmbito da cadeira de **Metodologias Experimentais em Informática** tendo como objectivo prever um modelo de relação entre uma variável independente (y) e uma ou mais variáveis dependentes (x). Todo este processo tem como objetivo a obtenção de uma resposta para a pergunta proposta:

De que forma algoritmos de ordenamento são afetados por memory faults?

Os scripts para a análise de dados, geração dos gráficos e regressões lineares foram realizados com recurso à linguagem R.

2 Regressão Linear

Este método tem como objectivo avaliar a relação entre duas variáveis contínuas através de uma equação matemática de resposta a uma variável contínua dependente (Y), em função de uma ou mais variáveis independentes (X).

Definition 1. Um modelo de regressão linear standardised tem como equação $y = a + bx$, onde x é a resposta prevista do output, a e b são os parâmetros de regressão estimados.

2.1 Correlação e modelos utilizados

De maneira a avaliar a qualidade dos modelos aplicados, é utilizado o coeficiente de determinação (r^2), que devolve a proporção da variância da variável dependente relativamente à variável independente.

Note-se que os limites do coeficiente de determinação (r^2) se encontram entre 0 e 1. Se a relação entre o *input* e o *output* for total, então a variância total é explicada pelo modelo, com $r^2 = 1$. Devido a isto, o objectivo do nosso modelo é obter o valor mais alto possível para r^2 .

Avaliamos diferentes tipos de modelos:

- Equação Linear *standard* : $y = a + bx$;
- Modelo exponencial : $y = e^{a+bx}$;
- Modelo quadrático: $y = (a + bx)^2$
- Modelo recíproco: $y = 1/(a + bx)$
- Modelo logarítmico: $y = a + b * \ln(x)$
- Modelo de potência: $y = a * x^b$

De maneira a avaliar os modelos são propostas as seguintes assunções:

- Relação linear entre a variável dependente e a independente, avaliado através do coeficiente de determinação (r^2);
- Independência dos resíduos;
- Distribuição normal dos resíduos;
- Variância igual dos resíduos.

3 Resultados da análise de Modelos

De maneira a avaliar os diferentes modelos referidos, foram avaliados os diferentes coeficientes de determinação (r^2), procurando sempre que esse valor fosse o máximo possível, dado que um maior (r^2) implica uma relação mais forte entre as variáveis em estudo.

3.1 Resultados para número de elementos do array (n)

Model	Algorithm	Multiple R-squared	Adjusted R-squared
Std Linear	bubble	0.7441	0.7428
Exponential	bubble	0.7499	0.7487
Quadratic	bubble	0.7907	0.7897
Reciprocal	bubble	0.4143	0.4114
Logarithmic	bubble	0.6448	0.643
Power Model	bubble	0.8632	0.8626
Std Linear	quick	0.6778	0.6762
Exponential	quick	0.7205	0.7191
Quadratic	quick	0.6783	0.6767
Reciprocal	quick	0.4306	0.4277
Logarithmic	quick	0.6664	0.6647
Power Model	quick	0.8038	0.8028
Std Linear	merge	0.6891	0.6876
Exponential	merge	0.7612	0.76
Quadratic	merge	0.674	0.6723
Reciprocal	merge	0.4706	0.4679
Logarithmic	merge	0.6467	0.6449
Power Model	merge	0.8227	0.8218
Std Linear	insertion	0.7337	0.7324
Exponential	insertion	0.7513	0.75
Quadratic	insertion	0.7411	0.7398
Reciprocal	insertion	0.6261	0.6242
Logarithmic	insertion	0.7354	0.7341
Power Model	insertion	0.8129	0.8119

3.2 Resultados para o limite de valores da sequência inicial (max_r)

Model	Algorithm	Multiple R-squared	Adjusted R-squared
Std Linear	bubble	0.03509	0.03022
Exponential	bubble	0.02784	0.0229
Quadratic	bubble	0.03175	0.02686
Reciprocal	bubble	0.0192	0.01425
Logarithmic	bubble	0.04279	0.03796
Power Model	bubble	0.03527	0.03039
Std Linear	quick	1.706e-06	-0.005049

Model	Algorithm	Multiple R-squared	Adjusted R-squared
Exponential	quick	0.0003259	-0.004723
Quadratic	quick	8.027e-05	-0.00497
Reciprocal	quick	0.001047	-0.003998
Logarithmic	quick	0.0006178	-0.00443
Power Model	quick	5.014e-05	-0.005
Std Linear	merge	0.0239	0.02449
Exponential	merge	0.02635	0.02143
Quadratic	merge	0.02779	0.02288
Reciprocal	merge	0.02385	0.01892
Logarithmic	merge	0.02659	0.02167
Power Model	merge	0.02282	0.01789
Std Linear	insertion	0.001229	-0.03815
Exponential	insertion	0.001773	-0.003269
Quadratic	insertion	0.001514	-0.003529
Reciprocal	insertion	0.001223	-0.002908
Logarithmic	insertion	0.002132	0.003821
Power Model	insertion	0.001793	-0.03249

3.3 Resultados para a probabilidade de ocorrência de erro (*eps*)

Model	Algorithm	Multiple R-squared	Adjusted R-squared
Std Linear	bubble	0.4425	0.4396
Exponential	bubble	0.4646	0.4619
Quadratic	bubble	0.4642	0.4615
Reciprocal	bubble	0.3996	0.3966
Logarithmic	bubble	0.4674	0.4647
Power Model	bubble	0.3931	0.3901
Std Linear	quick	0.5307	0.5283
Exponential	quick	0.7215	0.7201
Quadratic	quick	0.6461	0.6443
Reciprocal	quick	0.7446	0.7433
Logarithmic	quick	0.7327	0.7314
Power Model	quick	0.7051	0.7036
Std Linear	merge	0.515	0.5126
Exponential	merge	0.7595	0.7583
Quadratic	merge	0.6614	0.6597
Reciprocal	merge	0.7714	0.7703
Logarithmic	merge	0.8166	0.8157
Power Model	merge	0.7496	0.7484
Std Linear	insert	0.5415	0.5392
Exponential	insert	0.6711	0.6694
Quadratic	insert	0.6186	0.6166
Reciprocal	insert	0.6944	0.6929
Logarithmic	insert	0.7622	0.761
Power Model	insert	0.714	0.7126

3.4 Modelos aplicados

Procurámos então obter o valor mais alto dos coeficientes de determinação (r^2) de entre os diferentes modelos analisados. No entanto, e de forma a tentar estabelecer uma relação mais homogénea, os modelos escolhidos foram similares para cada variável independente, ou seja, os valores de n seguem o mesmo modelo para os quatro algoritmos de ordenamento, bem como os valores de max_r e eps .

Assim sendo, apesar de os valores mais altos se encontrarem pontualmente em outros modelos, como, por exemplo, relativamente à variável independente eps , o valor (r^2) de *Recíproco* para o *Quick Sort* (0.7446) é superior ao modelo considerado, o *Logarítmico* (0.7327).

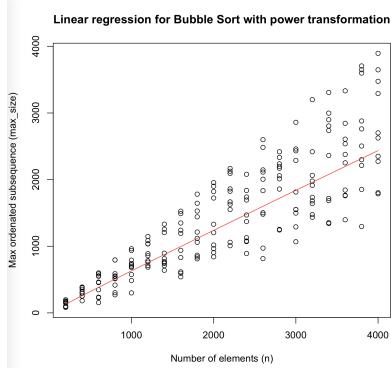
Foram escolhidos os seguintes modelos para cada variável independente:

- n - modelo de potência;
- max_r - modelo logarítmico;
- eps - modelo logarítmico.

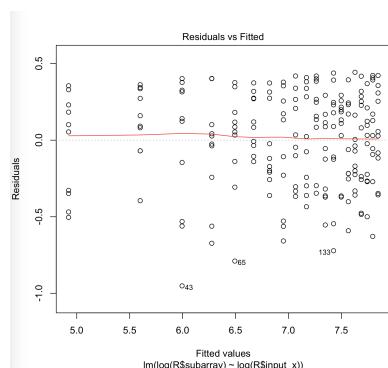
4 Resultados dos modelos escolhidos

4.1 Número de elementos do array (n)

4.1.1 Bubble Sort

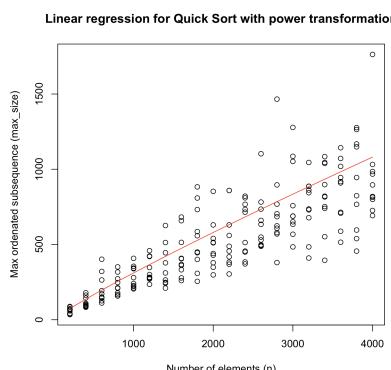


(a) Regressão linear com modelo logarítmico max_r

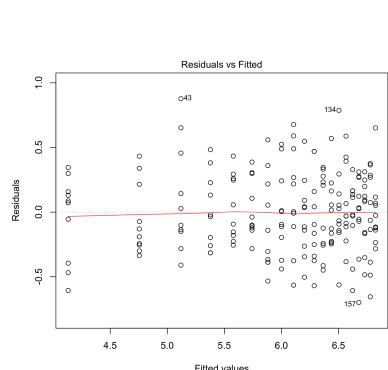


(b) Resíduos VS valores fitted

4.1.2 Quick Sort

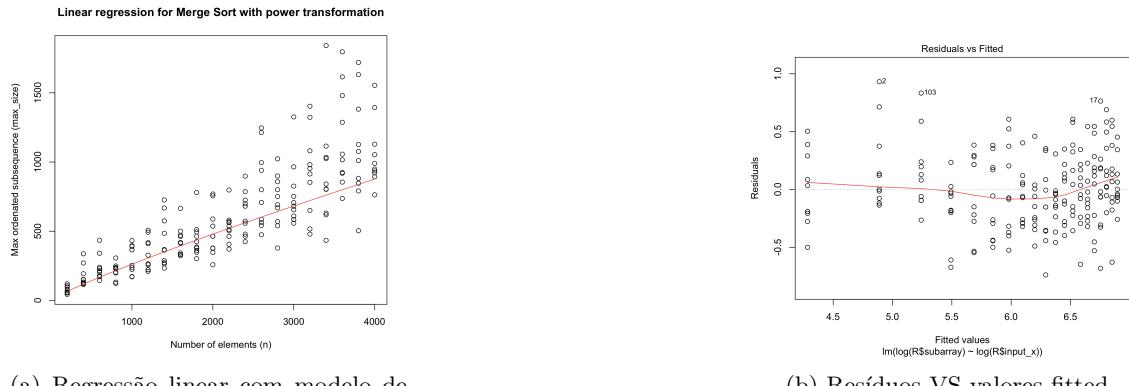


(a) Regressão linear com modelo de potência max_r

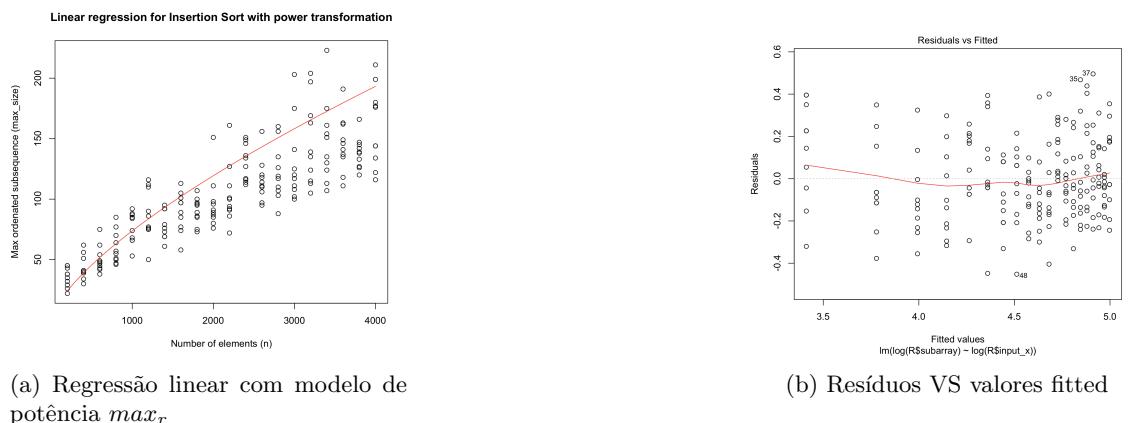


(b) Resíduos VS valores fitted

4.1.3 Merge Sort

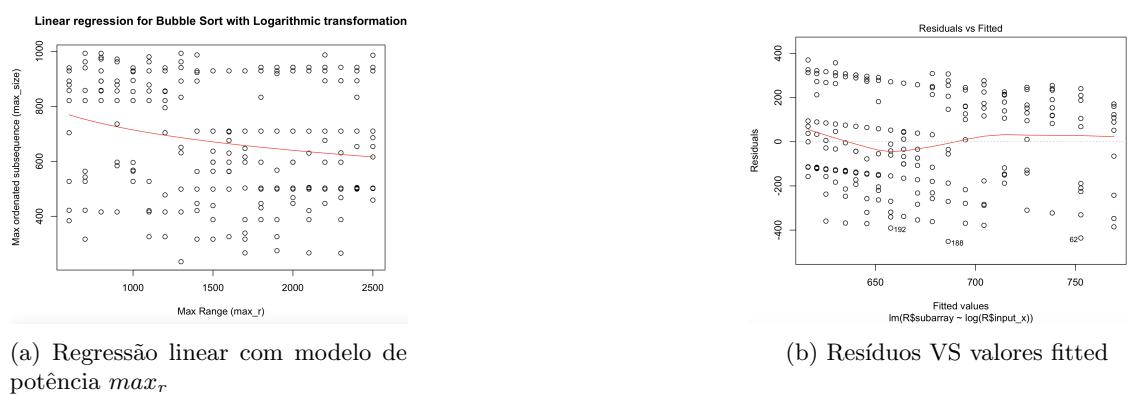


4.1.4 Insertion Sort

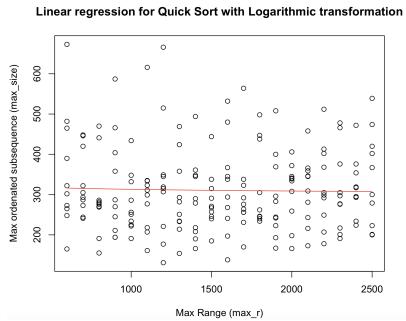


4.2 Variação dos valores da sequência (\max_r)

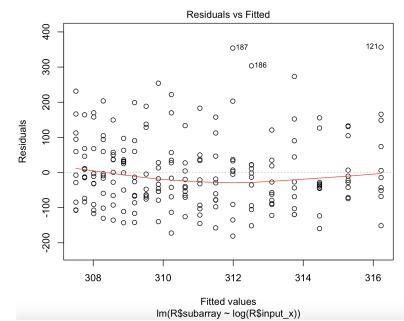
4.2.1 Bubble Sort



4.2.2 Quick Sort

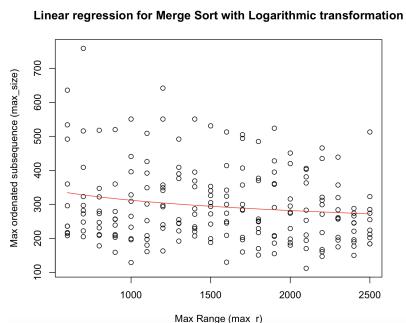


(a) Regressão linear com modelo logarítmico \max_r

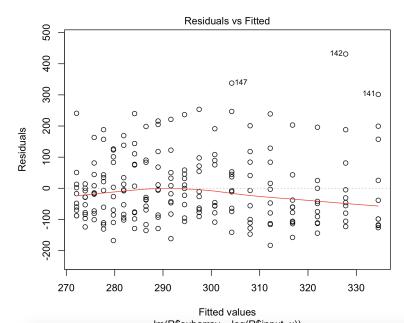


(b) Resíduos VS valores fitted

4.2.3 Merge Sort

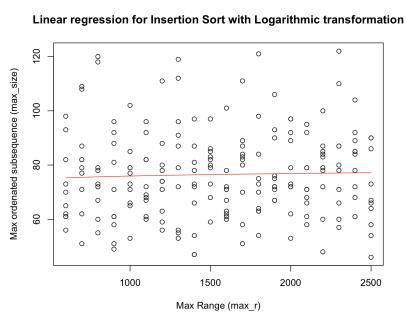


(a) Regressão linear com modelo logarítmico \max_r

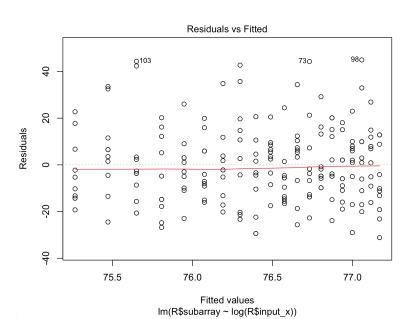


(b) Resíduos VS valores fitted

4.2.4 Insertion Sort



(a) Regressão linear com modelo logarítmico \max_r



(b) Resíduos VS valores fitted

4.3 Probabilidade de Falha (eps)

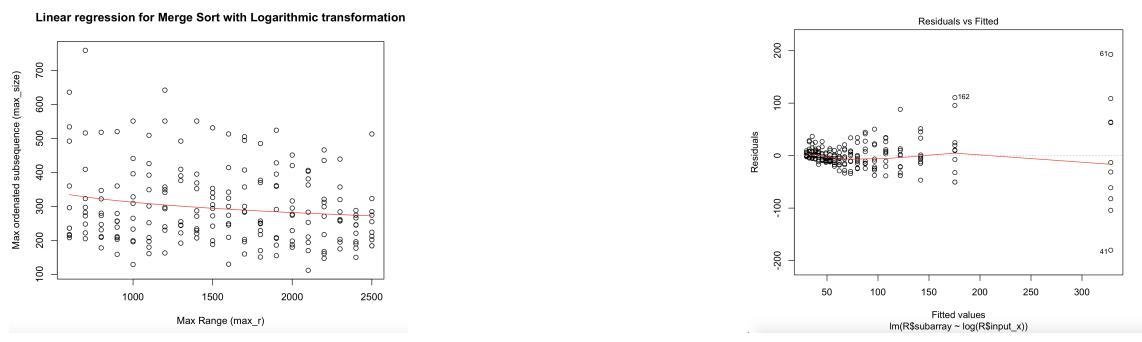
4.3.1 Quick Sort



(a) Regressão linear com modelo logarítmico eps

(b) Resíduos VS valores fitted

4.3.2 Merge Sort



(a) Regressão linear com modelo logarítmico eps

(b) Resíduos VS valores fitted

4.3.3 Insertion Sort



(a) Regressão linear com modelo logarítmico eps

(b) Resíduos VS Valores fitted

4.4 Tabela de coeficientes de curvas

Algorithm	Ind. Variable	A	B
bubble	n	0.74578	0.97556
quick	n	0.60574	0.6951
merge	n	0.63495	0.87176
insertion	n	0.62761	0.89838
bubble	max_r	1455.37	-107.26
quick	max_r	355.296	-6.108
merge	max_r	614.43	43.75
insertion	max_r	66.735	1.334
bubble	eps	134.594	-90.636
quick	eps	10.562	-32.865
merge	eps	-18.526	-50.194
insertion	eps	10.583	-10.0390

5 Conclusões

Após a análise dos resultados da análise de modelos, é visível uma diferença significativa entre os valores de relação entre variáveis. Podemos então afirmar, de acordo com os nossos resultados, que existe uma relação consistente entre a variável dependente max_{size} e a variável independente n , sendo que o modelo estudado verifica um coeficiente médio de determinação (r^2) de 0.825, sugerindo uma relação forte.

Também os valores de eps sugerem uma relação consistente com max_{size} , ainda que ligeiramente mais fraca do que a anteriormente mencionada, com um coeficiente médio de determinação (r^2) de 0.6932.

A nível de max_r , e de acordo os valores obtidos, podemos concluir que a relação entre esta variável independente e max_{size} é bastante fraca, com um coeficiente médio de determinação (r^2) 0.08.

Relativamente à relação linear de cada um dos modelos, já com as transformações aplicadas, verifica-se que de uma maneira geral, a linha vermelha sugere linearidade nos dados apresentados, ainda que o modelo onde essa propriedade melhor se verifica seja para o caso da variável independente n . Nos valores de eps é onde se verifica uma maior variância dos valores, onde

Por último, todos os modelos analisados seguem uma distribuição aproximadamente normal, e a nível da igualdade na variância de resíduos, a variável independente n foi a variável que nos apresentou gráficos com uma linha o mais homóloga possível, bem como uma distribuição equalitária dos pontos.

Concluindo, é possível afirmar a existência de uma relação forte entre os valores de n e max_{size} , assim como entre eps e max_{size} . Contrariamente ao pensado e obtido no relatório anterior, não consideramos a existência de uma relação entre max_r e max_{size} .

6 Referências

1. <http://www.learnbymarketing.com/tutorials/linear-regression-in-r/>
2. https://courses.lumenlearning.com/suny-natural-resources-biometrics/chapter/chapter-7-correlation-and-simple-linear-regression/?fbclid=IwAR0t1_c0ONC23hp6ot_KfWp783q7L07Yil5lxAMgmP7Jch-XZWEyzxyvBcI
3. <https://stattrek.com/regression/linear-transformation.aspx>
4. Slides de apoio à cadeira