Lista de Exercícios 3

Exercício 1

Você está interessado em desenvolver um modelo para estimar o tempo de execução de um algoritmo. Você tem razões para acreditar que dois parâmetros de configuração do algoritmo, A e B, sejam bons previsores para o tempo de execução do mesmo. Além disto, você acredita que cada um destes parâmetros impacte o tempo de execução de forma linear. Em outras palavras, você acredita que o tempo de execução do algoritmo possa ser previsto pela seguinte equação:

Tempo = $b_0 + b_A x_A + b_B x_B$

Utilizando os dados de **28** experimentos com seu algoritmo (cada um com valores diferentes de A e B), você realiza uma regressão linear múltipla e encontra os seguintes resultados:

$b_0 = 13.5$, $b_A = 0.47$ e $b_B = 1.72$

$sb_0 = 2.2$, $sb_A = 0.018$ e $sb_B = 0.18$

SST = 850.6 e SSR 670.3

Exercício 1

a) Qual porcentagem da variação dos dados é explicada pelos supostos previsores?

$$R^2 = \frac{SSR}{SST} = \frac{670.3}{850.6} = 78.80\%$$

Exercício 1

b) Os previsores explicam uma fração significativa da variação? Justifique

Isto equivale a perguntar se a regressao é significativa. Precisa fazer o teste-F.

SSR = 670.3

SSE = 850.6 – 670.3 = 180.3

MSR = SSR/k = 670.3/2 = 335.15

MSE = SSE/(n-k-1) = 180.3/(28 - 2 - 1) = 7.212

F-calculado = MSR/MSE = 46.47

$F_{[90;\ 2,25]} = 2.53$ (com 90%) $F_{[95;\ 2,25]} = 3.39$ (com 95%)

Assim o teste F passa com 95% de confiança

Exercício 1

c) Quais previsores são significativos? Justifique.

$t_{0.975,25} = 2.06$

IC de 95% de b_A :

$b_A \pm \ t_{0.975,25} \ sb_A$

$0.47 \ \pm \ 2.06 * 0.018$: 0.47 ± 0.037 -> SIGNIFICATIVO

IC de 95% de $b_B = 1.72$

$b_B \pm \ t_{0.975,25} \ sb_B$

$1.72 \pm \ 2.06 * 0.18$: 1.72 ± 0.37 -> SIGNIFICATIVO

Exercício 1

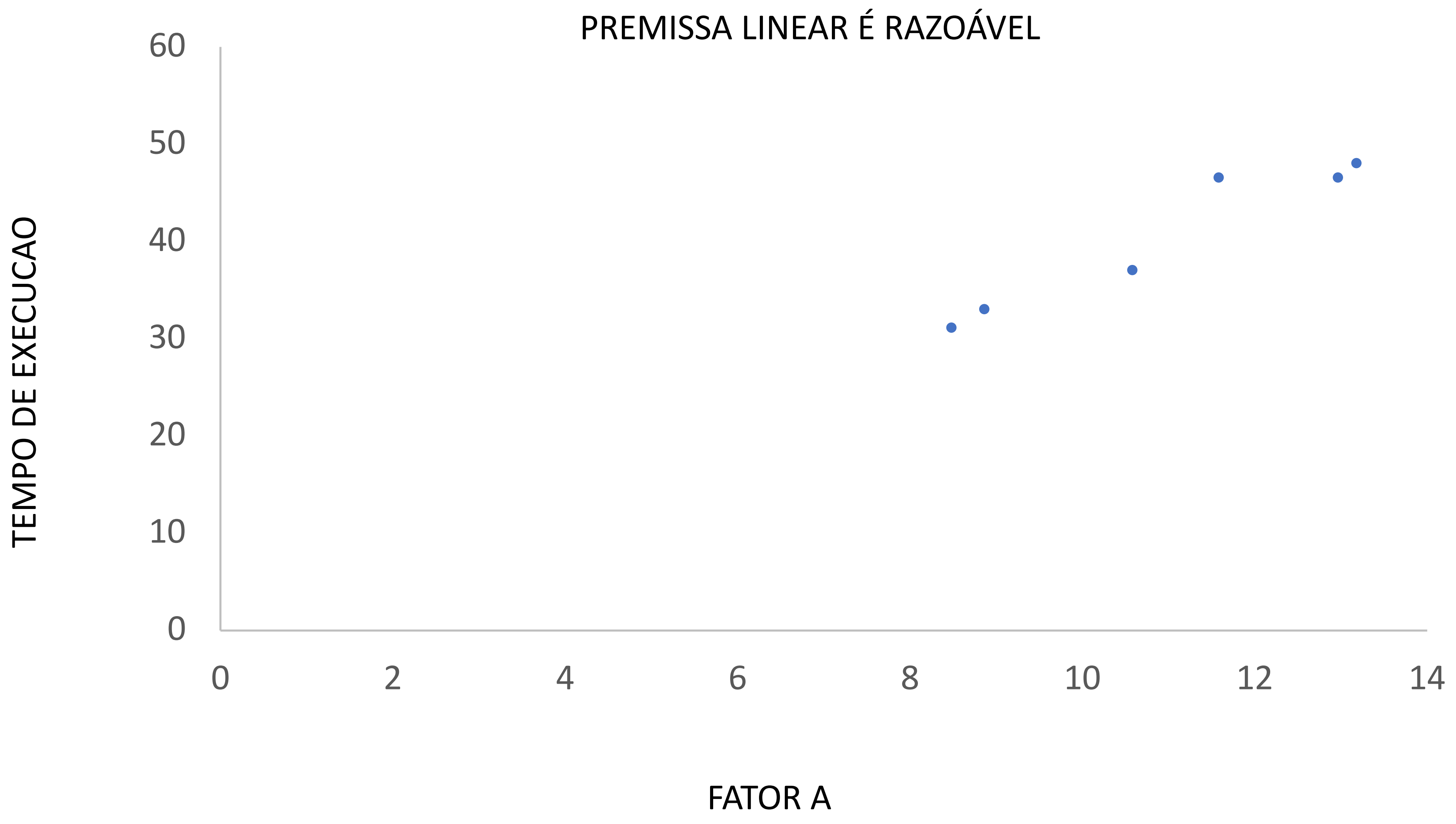
d) Há algum indício de multicolinearidade entre A e B? Por quê? Em caso afirmativo, descreva o que deverá ser feito para tratar este problema.

Não, não há já que todos os cálculos (R2, teste-F e IC de preditores) são consistentes e apontam para um bom modelo.

Exercício 2

Considere o exemplo da questão anterior. Suponha que você deseja verificar a relação custo-benefício de modelos alternativos. Em particular, você deseja verificar o benefício de um modelo baseado em um único previsor, o fator A. Porém, você perdeu os resultados referentes a grande parte dos experimentos realizados, restando apenas os dados abaixo. Supondo que a realização de novos experimentos seja inviável, você decide comparar o modelo anterior (analisado na questão anterior) com um modelo novo, derivado a partir dos dados abaixo. Qual a sua conclusão? Em outras palavras: o modelo com apenas um previsor (fator A) compensa?? Desenvolva a resposta.

A	8.5	8.9	10.6	11.6	13	13.2
Tempo de execução	30.9	32.7	36.7	46.3	46.2	47.8


PREMISSA LINEAR É RAZOÁVEL
TEMPO DE EXECUCAO
FATOR A
0
10
20
30
40
50
60
0
2
4
6
8
10
12
14

Exercício 2

$$\bar{x} = \frac{1}{n}\sum x_i$$ = 10.9667

$$\sum xy = \sum x_i y_i$$ = 2711.34

$$\bar{y} = \frac{1}{n}\sum y_i$$ = 40.1

$$\sum x^2 = \sum x_i^2$$ = 741.62

$$b_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$ = 72.76/20.01333 = 3.6356

$$b_0 = \bar{y} - b_1\bar{x}$$ = 0.2298

Tempo = $b_0 + b_1 x_A$

A	8.5	8.9	10.6	11.6	13	13.2
Tempo de execução	30.9	32.7	36.7	46.3	46.2	47.8
Valor Predito	31.13	32.59	38.77	42.40	47.49	48.22
Erro	0.23	-0.11	2.07	-3.89	1.29	0.42

Exercício 2

A	8.5	8.9	10.6	11.6	13	13.2
Tempo de execução	30.9	32.7	36.7	46.3	46.2	47.8
Valor Predito	31.13	32.59	38.77	42.40	47.49	48.22
Erro	0.23	-0.11	2.07	-3.89	1.29	0.42
Erro^2	0.054	0.013	4.27	15.19	1.67	0.18

SSE = 21.3755

SST = SSY – n$\bar{y}^2$ = 9933.96 – 6*(40.1)^2 = 285.9

R2 = $\frac{SST - SSE}{SST}$ = 92.52% -> SIM

Exercício 2

A	8.5	8.9	10.6	11.6	13	13.2
Tempo de execução	30.9	32.7	36.7	46.3	46.2	47.8
Valor Predito	31.13	32.59	38.77	42.40	47.49	48.22
Erro	0.23	-0.11	2.07	-3.89	1.29	0.42
Erro^2	0.054	0.013	4.27	15.19	1.67	0.18

SSE = 21.3755

SST = SSY – n$\bar{y}^2$ = 9933.96 – 6*(40.1)^2 = 285.9

R2 = $\frac{SST - SSE}{SST}$ = 92.52% -> SIM, o modelo com 1 previsor é melhor (R2 maior)

Exercício 3

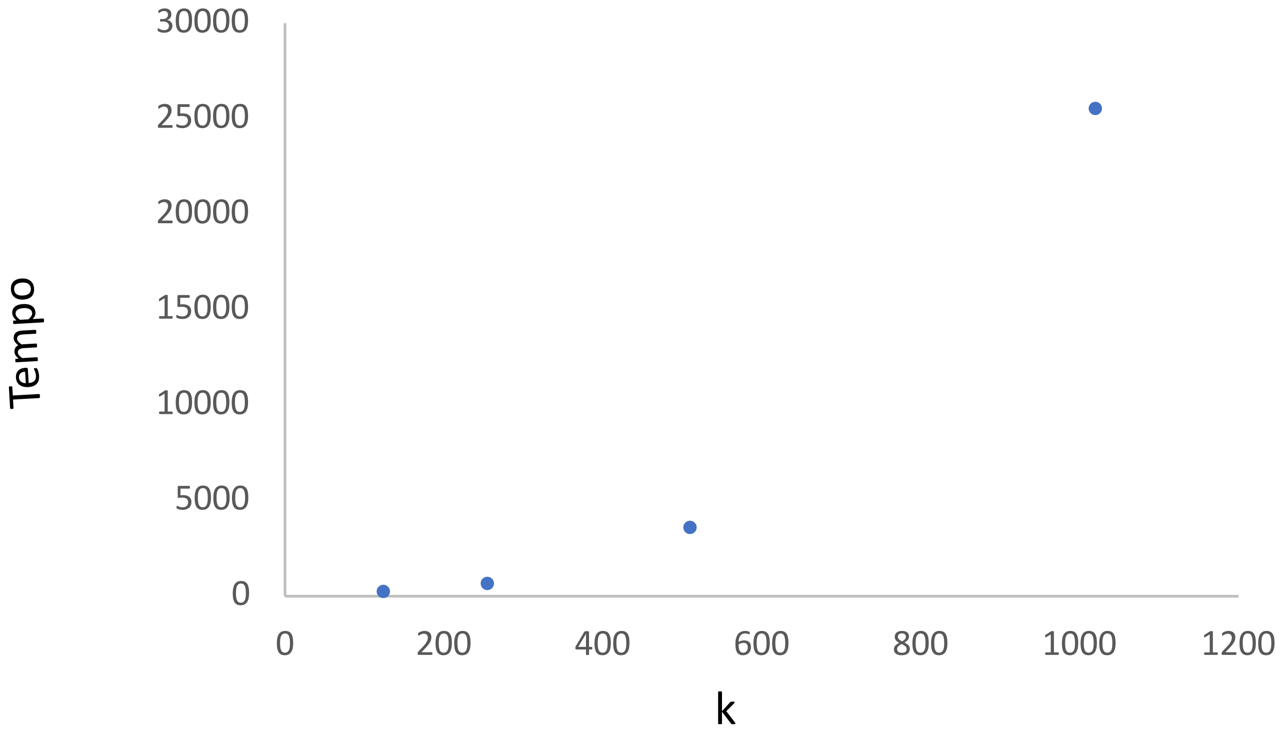
Os tempos para criptografar registros de k bits em um computador A são mostrados na tabela abaixo. Responda:

k	Tempo (ms)
128	93
256	478
512	3408
1024	25410

Exercício 3

Os tempos para criptografar registros de k bits em um computador A são mostrados na tabela abaixo. Responda:

k	Tempo (ms)
128	93
256	478
512	3408
1024	25410



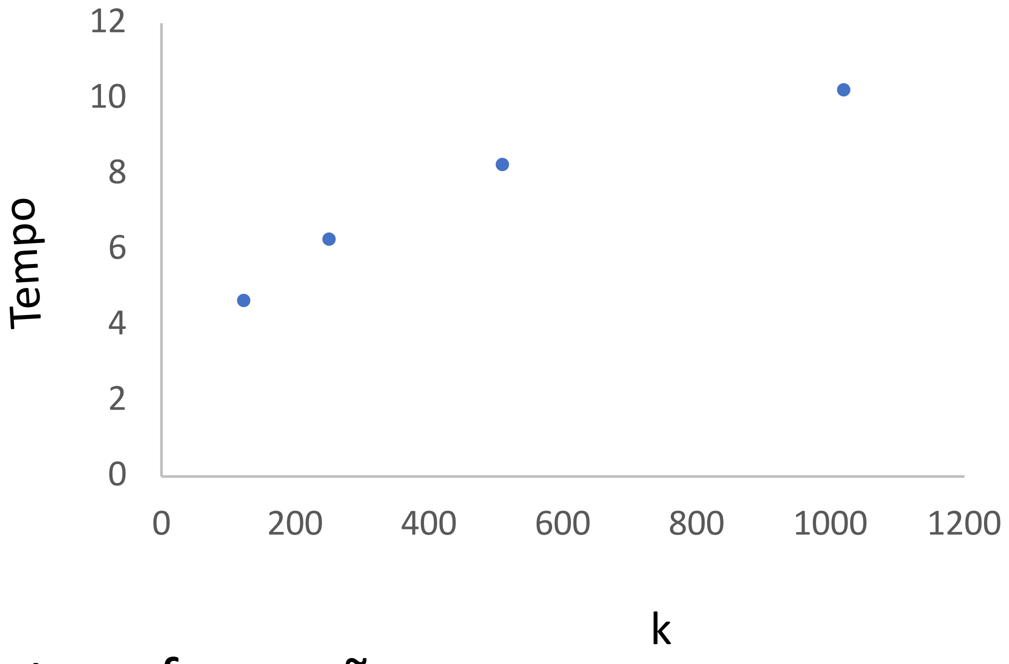
Relação claramente não linear, deve tentar transformação

$y = ae^{bx}$ -> $y' = \ln(y) = b_0 + b_1 x$

Exercício 3

Os tempos para criptografar registros de k bits em um computador A são mostrados na tabela abaixo. Responda:

k	Tempo (ms)	ln (tempo)
128	93	4.5326
256	478	6.1696
512	3408	8.1339
1024	25410	10.1429



Relação claramente não linear, deve tentar transformação

$y = ae^{bx}$ -> $y' = \ln(y) = b_0 + b_1x$

Exercício 3

a) Uma regressão linear pode ser aplicada diretamente aos dados? Justifique sua resposta. Em caso negativo, discuta o que você faria para obter uma relação entre tempo de criptografia e tamanho de registro. Apresente um modelo para esta relação.

Não, não pode pois a relação é claramente não linear. Deve-se fazer uma transformação. Optou-se por $y' = \ln(y)$

$$y' = \ln(y) = b_0 + b_1 x$$

Exercício 3

b) Determine os valores dos parâmetros do seu modelo.

$$\bar{x} = \frac{1}{n}\sum x_i \quad = 480$$

$$\sum xy = \sum x_i y_i \quad = 16710.4677$$

$$\bar{y} = \frac{1}{n}\sum y_i \quad = 7.2447$$

$$\sum x^2 = \sum x_i^2 \quad = 1392640$$

$$b_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} \quad = 0.005945$$

$$b_0 = \bar{y} - b_1\bar{x} \quad = 4.3909$$

Exercício 3

c) Qual a porcentagem da variação é explicada pela regressão? Você está satisfeito com seu modelo? Se não, qual seria o seu próximo passo?

SSE = 1.0509

SST = SSY – SS0 = 227.6469 – $4*7.2447^2$ = 17.7015

R2 = $\frac{SST - SSE}{SST}$ = 94.06%

Modelo muito bom. Estou satisfeito. Poderia-se tentar outra transformação (e.g., $y = ax^b$ -> ln(y) = ln(a) + bln(x)-> ln(y) = b0 + b1ln(x)) e verificar se ele é ainda melhor

Exercício 3

d) Quais parâmetros são significativos, com uma confiança de 90%?

$$s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{1.0509}{4-2}} = 0.7249$$

$$s_{b0} = se\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum x^2 - n\bar{x}^2}} = 0.7249 * \sqrt{\frac{1}{4} + \frac{480^2}{1392640 - 4 * 480^2}} = 0.6232$$

$$s_{b1} = \frac{se}{\sqrt{\sum x^2 - n\bar{x}^2}} = \frac{0.7249}{\sqrt{1392640 - 4 * 480^2}} = 0.001056$$

$t_{0.95,2} = 2.92$

IC de 90% para b0: 4.3909 ± 2.92* 0.6232 (SIGNIFICATIVO)

b1: 0.005945 ± 2.92* 0.001056 : 0.005945 ± 0.003084

(SIGNIFICATIVO)

Exercício 3

e) Qual o tempo esperado para criptografar um registro de 384 bits? Quais limites você colocaria para esta estimativa se você aceita um erro máximo de 10% para uma única medida futura?

$$y' = \ln(y) = 4.3909 + 0.005945*384 = 6.6739 \;\; -> \;\; y = e^{6.6739} = 791.5416$$

IC de 90% na escala logaritma

$$s_y = s_e\sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum x^2 - n\bar{x}^2}} = 0.7249\sqrt{1 + \frac{1}{4} + \frac{(384-480)^2}{1392640 - 4*480^2}} = 0.8168$$

$t*\ 0.8168/\ 6.6739 \leq 0.1$ $\quad t_{(1-\alpha/2),2} \leq 0.8171$ $\quad 1-\boldsymbol{\alpha}/2 \leq 0.7$

Confiança máxima (1- $\boldsymbol{\alpha}$) de 40%

Exercício 4

O tempo de execução de um algoritmo foi medido em função de 2 parâmetros. Uma regressão linear múltipla, utilizando 7 observações, levou aos seguintes resultados:

j	**0**	**1**	**2**
$\mathbf{b_j}$	-0.1614	0.1182	0.0165
$\mathbf{c_{jj}}$	0.6297	0.0280	0.0012

Coeficiente de correlação múltipla: 0.99

$s_e = 1.2$

Exercício 4

a) A regressão é significativa com 90% de confiança?

R = 0.99 -> R^2 = 98.01%

$s_e = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{\frac{SSE}{7-2-1}} = 1.2$ -> SSE = $1.2^2 * 4$ = 5.76

-> MSE = $s_e^2 = 1.2^2 = 1.44$

$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR+SSE} = \frac{SSR}{SSR+5.76} = 0.9801$

SSR= 0.9801*SSR + 5.6454 -> SSR = 283.6872

MSR = SSR/k = 283.6872/2 = 141.8436

F_{comp} = MSR/MSE = 98.5025

F(0.9, 2, 4) = 4.32 -> F_{comp} > F(0.9, 2, 4) -> SIM a regressão é significativa

Exercício 4

b) Quais variáveis previsoras são significativas com 90% de confiança?

IC de 90% para b_1 e b_2 -> $t_{0.95,4} = 2.132$

desvios padrões

$s_{b1} = s_e\sqrt{c_{11}} = 1.2\sqrt{0.0280}. = 0.2008$ $s_{b2} = s_e\sqrt{c_{22}} = 1.2\sqrt{0.0012}. = 0.04157$

IC:

b_1: 0.1182 ± 2.132* 0.2008 = 0.1182 ± 0.4281 -> Não é significativo

b_2: 0.0165 ± 2.132* 0.04157 = 0.0165 ± 0.0886 -> Não é significativo

Exercício 4

c) Qual variável previsora você consegue estimar com maior precisão, para uma confiança de 90%? Justifique apresentando a precisão que você pode atribuir a cada variável com a dada confiança.

Erro máximo associado à estimativa de b_1: 0.4281/0.1182 = 3.6218

Erro máximo associado à estimativa de b_2: 0.0886/ 0.0165 = 5.3712

Logo, aparentemente, a maior precisão é da variável b1. Entretanto, esta precisão está muito baixa (erro máximo por um fator de quase 4). Além disto, há problemas de inconsistências na regressão o que levanta questionamentos sobre validade do modelo.

Exercício 4

d) Você está satisfeito com seu modelo? Justifique. Se não, qual seria seu próximo passo?

Não; há sinais de multicolineariedade. Seria interessante avaliar as regressões simples com cada previsor separadamente e ver qual leva ao melhor resultado.