MAE-350 - MODELOS DE REGRESSÃO I

5ª Lista de Exercícios

Profa. Silvia N. Elian

1) Seja $E(y|x) = \beta_0 + \beta_1 x$ o modelo de regressão linear simples. Deduza um intervalo de confiança para $\beta_0 + \beta_1$.

2) No exemplo, Y = tempo de reação, X = idade, Z = acuidade Visual, obter através do cálculo matricial:

a) $\hat{\beta}_0, \hat{\beta}_1$ e $\hat{\beta}_2$ b) ANOVA

c) $\hat{Var}(\hat{\beta}_0), \hat{Var}(\hat{\beta}_1), \hat{Var}(\hat{\beta}_2)$ e as estimativas de todas as possíveis covariâncias entre os $\hat{\beta}$.

d) Um intervalo de confiança para o tempo médio de reação de indivíduos com 28 anos e acuidade visual 80.

3) Sejam

X_1	-1	1	-1	1	0	0	0
X_2	-1	-1	1	1	0	1	2
Y	1	4	8	9	3	8	9

a) Ajuste para esses dados o modelo $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_{11} x_1^2$

b) Teste ao nível $\alpha = 0{,}05$ a hipótese $H_0: \beta_{11} = 0, \beta_1 = \beta_2$

c) Baseado no resultado deste teste, qual seria o modelo adotado?

4) Seja $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots \hat{\beta}_k x_k$ o modelo de regressão linear com k variáveis independentes, onde $(X'X)^{-1} = (a_{ij})_{i,j} = 0,1 \ldots k$

a) Deduza um intervalo de confiança para $\beta_1 - \beta_2$.

b) Como seria o teste $H_0: \beta_1 = \beta_2$ contra $H_a: \beta_1 > \beta_2$?

5) Com o objetivo de ajustar o modelo $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \in$, foi obtida uma amostra de 8 observações. Os valores das variáveis independentes estão na tabela

X_1	X_2	X_3	X_4
-1	-1	-1	-3
-1	-1	1	-1
-1	1	-1	-1
-1	1	1	1
1	-1	-1	-1
1	-1	1	1
1	1	-1	1
1	1	1	3

Verifique que, embora o valor dos coeficientes de correlação entre pares de variáveis independentes seja sempre inferior a 0, 58, existe multicolinearidade perfeita.

6) Deseja-se estudar a quantidade de nitrogênio (Y) que não é absorvida por um processo químico, como função das variáveis:

X_1 = fluxo de ar X_2 = temperatura da água usada na refrigeração

X_3 = concentração ácida

Decidiu-se usar regressão linear como técnica exploratória para escolher o melhor modelo de dependência. Assim, as 7 possíveis equações de regressão produziram as somas de quadrados da regressão apresentadas abaixo. Escolha o melhor modelo e justifique.

Modelo	$f(X_1)$	$f(X_2)$	$f(X_3)$	$f(X_1,X_2)$	$f(X_1,X_3)$	$f(X_2,X_3)$	$f(X_1,X_2,X_3)$
SSregressão	1750	1580	330	1880	1760	1590	1890

$n = 21 \quad SST = 2070$

7) Mostre que uma fórmula alternativa para calcular a soma de quadrados da regressão é

$$SSR = \sum_{i=1}^{n} \hat{y}_i^2 - n\bar{y}^2.$$

8) Sejam X_1, X_2 e X_3 variáveis independentes e Y variável dependente e os modelos:

i) $y = \alpha X_1^{\beta} X_2^{\gamma} X_3^{\delta} \in$

ii) $y = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \in}$

iii) $y = \dfrac{1}{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \in}$

iv) $y = \dfrac{1}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \in}}$

a) Transforme cada um dos modelos acima num modelo de regressão linear.

b) Para cada modelo transformado, especifique quais são as variáveis independentes e dependentes, o resíduo e os parâmetros.

9) Sejam três regressões onde o número de observações e os valores das variáveis independentes são os mesmos. Na primeira, a variável dependente é Y_1, na segunda Y_2 e na terceira $Y_3 = Y_1 + Y_2$.

Sendo $\hat{\underset{\sim}{\beta}}_1, \hat{\underset{\sim}{\beta}}_2, \hat{\underset{\sim}{\beta}}_3$ os estimadores de mínimos quadrados dos parâmetros respectivamente da 1a., 2a, e 3a. equação de regressão, prove que . $\hat{\underset{\sim}{\beta}}_3 = \hat{\underset{\sim}{\beta}}_1 + \hat{\underset{\sim}{\beta}}_2$

10) Um ensaio de adubação forneceu os seguints resultados:

	X= dose de adubo por hectare	Y = produção por hectare	
	0	6	8
	1	16	18
	2	18	20
	3	12	14
onde	$\bar{y} = 14$	$\Sigma(yi - \bar{y})^2 = 176$	

Admitindo um modelo da forma $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \in$

a) Obtenha os estimadores de mínimos quadrados de β_0, β_1 e β_2.
b) Teste a hipótese que a produção máxima é obtida aplicando 2 doses de adubo por hectare. (Obs.: Se $y = ax^2 + bx + c$, o ponto de máximo é $x = -\frac{b}{2a}$).
c) Teste a hipótese $H_0 : \beta_2 = 0$ contra $H_a : \beta_2 \neq 0$.

11) O administrador de um hospital deseja estudar a relação entre o nível de satisfação do paciente (Y), a idade do paciente (X_1, em anos), severidade da doença (X_2, um índice) e nível de ansiedade (X_3, um índice). Para isto, selecionou 23 pacientes e coletou os dados abaixo, onde grandes valores de Y, X_2 e X_3 estão respectivamente associados com mais satisfação, aumento da severidade da doença e mais ansiedade.

i	1	2	3	4	5	6	7	8	9
X_{1i}	50	36	40	41	28	49	42	45	52
X_{2i}	51	46	48	44	43	54	50	48	62
X_{3i}	2,3	2,3	2,2	1,8	1,8	2,9	2,2	2,4	2,9
Y_i	48	57	66	70	89	36	46	54	26

i	10	11	12	13	14	15	16	17	18
X_{1i}	29	29	43	38	34	53	36	33	29
X_{2i}	50	48	53	55	51	54	49	56	46
X_{3i}	2,1	2,4	2,4	2,2	2,3	2,2	2,0	2,5	1,9
Y_i	77	89	67	47	51	57	66	79	88

i	19	20	21	22	23
X_{1i}	33	55	29	44	43
X_{2i}	49	51	52	58	50
X_{3i}	2,1	2,4	2,3	2,9	2,3
Y_i	60	49	77	52	60

a) Ajuste a estes dados o modelo de regressão $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \in_i$ e descreva a equação de regressão obtida. Qual o significado prático de β_2 neste caso ?
b) Faça um box plot dos resíduos. Este gráfico sugere a presença de outliers ?
c) Construa o gráfico dos resíduos contra $\hat{y}$, e contra cada variável independente envolvida. Construa o normal probability plot e com base nestas análises e na do item b), tire conclusões.
d) Teste, ao nível de significância 0,10 a existência de relação entre a variável resposta Y e as três variáveis explicativas consideradas. Calcule aproximadamente o p-value.
e) Calcule o coeficiente de correlação múltipla deste modelo.

12) No exercício anterior, admitindo válidas todas as suposições do modelo de regressão,

a) Obtenha um intervalo de confiança para o nível médio de satisfação de pacientes com 35 anos, grau de severidade da moléstia 45 e nível de ansiedade 2,2. Use coeficiente de confiança de 0,90.
b) Obtenha um intervalo de previsão para o nível de satisfação de um novo paciente com 50 anos, severidade da moléstia 45 e nível de ansiedade 2,0.

13) No exercício 11, verifique se a variável X_3 pode ser descartada do modelo, dado que X_1 e X_2 estão mantidas. Use $\alpha = 0,025$.

14) Repita o problema para testar se X_2 e X_3 podem ser descartadas dado que X_1 é mantida. Use $\alpha = 0,025$.

Resolver utilizando o R.

15) Sejam as seguintes características relativas ao preço de 20 terrenos

Terreno	área (em milhares de pés quadr.) = X_1	grau de declive = X_2	Vista Panorâmica = X_3 escala: 1 (pobre) a 9 (excelente)	Preço = Y (mil dólares)
1	14.7	1.5	2	4.1
2	14.2	1.8	2	3.9
3	12.7	2.9	1	3.2
4	13.8	1.0	1	2.9
5	14.4	0.5	2	3.9
6	17.4	1.0	2	4.1
7	21.8	5.7	4	5.8
8	14.0	5.4	6	5.1
9	17.5	17.5	9	6.8
10	23.0	14.5	9	6.8
11	18.3	14.4	9	6.5
12	19.4	12.2	9	7.0
13	15.2	5.0	8	5.8
14	18.3	13.1	6	5.1
15	21.7	15.2	8	5.3
16	16.7	10.1	8	4.9
17	13.6	7.4	7	6.0
18	14.5	5.8	7	5.3
19	12.1	5.1	7	4.8
20	17.4	17.3	1	4.3

Usando as variáveis independentes X_1, X_2 e X_3, qual seria a equação de regressão mais indicada para o problema?
Obs.: Possíves equações (2^3 possíveis equações)

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$ $\quad$ $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ $\quad$ $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2$ $\quad$ $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_3 x_3$ $\quad$ $\hat{y} = \bar{y}$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_3 x_3$ $\quad$ $\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$

Para selecionar o modelo utilize o teste F parcial como critério básico.
A manutenção do modelo selecionado estará condicionada à análise posterior. Nesta análise, não esquecer de:

1) Cacular os coeficientes de Explicação: R^2 e $\overline{R}^2$.
2) Fazer o teste de falta de ajuste (se possível).
3) Fazer uma análise de resíduos rigorosa, incluindo diagnóstico.
4) Interpretar todas as conclusões obtidas, estudar o significado prático das estimativas dos parâmetros, etc...
5) Com base nos itens estudados, apresentar uma conclusão geral, salientando as qualidades do modelo selecionado.