

# Lista 2

## Modelos Lineares Generalizados - 2/2023

César Augusto Galvão - 19/0011572

Laiza Mendes - 20/0067028

### Table of contents

<b>Questão 1</b>	<b>2</b>
a) Proponha algum método para resolver o problema da multicolinearidade no conjunto de dados . . . . .	2
b) Usando algum método de seleção de variáveis, obtenha o modelo final para o conjunto de dados . . . . .	2
c) Apresente a tabela de Análise de Variância para testar a significância global dos coeficientes do modelo final. Apresente as hipóteses de teste e conclua. . . . .	2
d) Com base no modelo obtido no item anterior, faça uma análise de resíduos e conclua.	2
<b>Questão 2</b>	<b>2</b>
a) Ajuste um modelo de regressão linear e interprete os resultados obtidos . . . . .	3
b) Obtenha a tabela ANOVA para o modelo obtido no item (a) e interprete os resultados	4
c) Considere a possibilidade de incluir a interação entre as variáveis independentes .	4
<b>Apêndice</b>	<b>8</b>

## Questão 1

Considere os dados sobre a qualidade do vinho tinto, apresentados no ficheiro `Q01-data.txt`. Ajuste o modelo de regressão linear múltipla, e faça uma análise completa desses dados. Que conclusões você tira dessa análise? (use 5% de significância durante as análises).

- a) **Proponha algum método para resolver o problema da multicolinearidade no conjunto de dados**
- b) **Usando algum método de seleção de variáveis, obtenha o modelo final para o conjunto de dados**
- c) **Apresente a tabela de Análise de Variância para testar a significância global dos coeficientes do modelo final. Apresente as hipóteses de teste e conclua.**
- d) **Com base no modelo obtido no item anterior, faça uma análise de resíduos e conclua.**

## Questão 2

Uma equipe de pesquisadores de saúde mental deseja comparar três métodos de tratamento da depressão grave (A, B e C=referência). Eles também gostariam de estudar a relação entre idade e eficácia do tratamento, bem como a interação (se houver) entre idade e tratamento. Cada elemento da amostra aleatória simples de 36 pacientes, foi selecionado aleatoriamente para receber o tratamento A, B ou C. Os dados obtidos podem ser encontrados no ficheiro `Q02-data.txt`. A variável dependente  $y$  é a eficácia do tratamento; as variáveis independentes são: a idade do paciente no aniversário mais próximo e o tipo de tratamento administrado (use 1% de significância durante as análises).

Uma amostra dos dados é exibida na tabela a seguir:

eficacia	idade	tratamento
56	21	A
41	23	B
40	30	B
28	19	C
55	28	A
25	23	C

### a) Ajuste um modelo de regressão linear e interprete os resultados obtidos

Temos um potencial modelo de regressão linear que pode ou não conter interações entre as variáveis, o qual pode ser expresso em sua forma saturada, em que  $X_1$  é a variável idade e  $X_2$  a variável tratamento

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (1)$$

ou, de forma análoga, desmembrando  $X_2$  em variáveis *dummy*  $X_A$  e  $X_B$ , indicadores da presença do tratamento  $A$  e  $B$ , ambas assumindo valor 0 quando se trata do tratamento  $C$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{Ai} + \beta_3 x_{Bi} + \beta_4 x_{1i} x_{Ai} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i. \quad (2)$$

Se simplesmente ajustamos um modelo de regressão linear – sem os termos de interação – utilizando (2) como referência na função `lm()`, obtemos os seguintes resultados:

Table 1: Modelo de regressão linear para tratamentos sem interação com idade sobre eficácia

Coeficiente	Estimativa	EP	Estatística t	p-valor
(Intercept)	22.291	3.505	6.359	0.000
idade	0.664	0.070	9.522	0.000
A	10.253	2.465	4.159	0.000
B	0.445	2.464	0.181	0.858

Ou seja, se considerarmos independentemente idade, tratamento A e tratamento B, podemos considerar que:

- Há uma linha de base na eficácia de aproximadamente 22.3, i.e. sob o tratamento C;
- A eficácia base para o tratamento A é de 32.3;
- A eficácia base para o tratamento B é de 22.75 – mas poderíamos desconsiderar este coeficiente, se nos guiarmos pelo p-valor;
- Cada ano a mais de vida incrementa a eficácia em 0.644.

É possível considerar que um tamanho de amostra pequeno tenha grande influência sobre a significância de  $H_0 : \beta_3 = 0$  do modelo. No entanto, trata-se de um fenômeno para o qual o tratamento pode estar estreitamente associado à idade, caso em que teríamos que considerar o modelo (2) por completo.

**b) Obtenha a tabela ANOVA para o modelo obtido no item (a) e interprete os resultados**

Se montarmos uma tabela de Análise de Variância para o modelo de regressão linear ajustado, obtemos os resultados a seguir:

Table 2: Tabela ANOVA para o modelo linear sem interações

Fonte de Var.	g.l.	SQ	QM	F	p-valor
idade	1	3424.432	3424.432	94.015	0.000
A	1	803.804	803.804	22.068	0.000
B	1	1.189	1.189	0.033	0.858
Residuals	32	1165.575	36.424	NA	NA

Nota-se que a maioria da variância explicada pelo modelo está associada à variável idade, enquanto a soma de quadrados das variáveis de tratamento juntas não superam a soma de quadrados dos resíduos.

Se conjugarmos os resultados deste item com os do item a) vemos que isoladamente apenas idade, e interessante apenas o tratamento A, parecem ser variáveis que realmente contribuem para a explicação do fenômeno.

**c) Considere a possibilidade de incluir a interação entre as variáveis independentes**

Supõe-se que  $\varepsilon_i \sim N(0, \sigma^2)$ .

**i) Lista de todos os submodelos possíveis**

A partir do modelo (2), construímos todos os possíveis submodelos. Considerando que temos três covariáveis e dois termos de interação, temos  $\sum_{n=1}^5 \binom{6}{n} = 62$  modelos

1.  $y_i = \beta_0 + \varepsilon_i$
2.  $y_i = \beta_1 x_{1i} + \varepsilon_i$
3.  $y_i = \beta_2 x_{Ai} + \varepsilon_i$
4.  $y_i = \beta_3 x_{Bi} + \varepsilon_i$
5.  $y_i = \beta_4 x_{1i} x_{Ai} + \varepsilon_i$
6.  $y_i = \beta_5 x_{1i} x_{Bi} + \varepsilon_i$

7.  $y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$
8.  $y_i = \beta_0 + \beta_2 x_{Ai} + \varepsilon_i$
9.  $y_i = \beta_0 + \beta_3 x_{Bi} + \varepsilon_i$
10.  $y_i = \beta_0 + \beta_4 x_{1i} x_{Ai} + \varepsilon_i$
11.  $y_i = \beta_0 + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
12.  $y_i = \beta_1 x_{1i} + \beta_2 x_{Ai} + \varepsilon_i$
13.  $y_i = \beta_1 x_{1i} + \beta_3 x_{Bi} + \varepsilon_i$
14.  $y_i = \beta_1 x_{1i} + \beta_4 x_{1i} x_{Ai} + \varepsilon_i$
15.  $y_i = \beta_1 x_{1i} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
16.  $y_i = \beta_2 x_{Ai} + \beta_3 x_{Bi} + \varepsilon_i$
17.  $y_i = \beta_2 x_{Ai} + \beta_4 x_{1i} x_{Ai} + \varepsilon_i$
18.  $y_i = \beta_2 x_{Ai} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
19.  $y_i = \beta_3 x_{Bi} + \beta_4 x_{1i} x_{Ai} + \varepsilon_i$
20.  $y_i = \beta_3 x_{Bi} \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
21.  $y_i = \beta_4 x_{1i} x_{Ai} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
22.  $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{Ai} + \varepsilon_i$
23.  $y_i = \beta_0 + \beta_1 x_{1i} + \beta_3 x_{Bi} + \varepsilon_i$
24.  $y_i = \beta_0 + \beta_1 x_{1i} + \beta_4 x_{1i} x_{Ai} + \varepsilon_i$
25.  $y_i = \beta_0 + \beta_1 x_{1i} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
26.  $y_i = \beta_0 + \beta_2 x_{Ai} + \beta_3 x_{Bi} + \varepsilon_i$
27.  $y_i = \beta_0 + \beta_2 x_{Ai} + \beta_4 x_{1i} x_{Ai} + \varepsilon_i$
28.  $y_i = \beta_0 + \beta_2 x_{Ai} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
29.  $y_i = \beta_0 + \beta_3 x_{Bi} + \beta_4 x_{1i} x_{Ai} + \varepsilon_i$
30.  $y_i = \beta_0 + \beta_3 x_{Bi} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
31.  $y_i = \beta_0 + \beta_4 x_{1i} x_{Ai} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
32.  $y_i = \beta_1 x_{1i} + \beta_2 x_{Ai} + \beta_3 x_{Bi} + \varepsilon_i$
33.  $y_i = \beta_1 x_{1i} + \beta_2 x_{Ai} + \beta_4 x_{1i} x_{Ai} + \varepsilon_i$
34.  $y_i = \beta_1 x_{1i} + \beta_2 x_{Ai} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
35.  $y_i = \beta_1 x_{1i} + \beta_3 x_{Bi} + \beta_4 x_{1i} x_{Ai} + \varepsilon_i$

36.  $y_i = \beta_1 x_{1i} + \beta_3 x_{Bi} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
37.  $y_i = \beta_1 x_{1i} + \beta_4 x_{1i} x_{Ai} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
38.  $y_i = \beta_2 x_{Ai} + \beta_3 x_{Bi} + \beta_4 x_{1i} x_{Ai} + \varepsilon_i$
39.  $y_i = \beta_2 x_{Ai} + \beta_3 x_{Bi} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
40.  $y_i = \beta_2 x_{Ai} + \beta_4 x_{1i} x_{Ai} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
41.  $y_i = \beta_3 x_{Bi} + \beta_4 x_{1i} x_{Ai} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
42.  $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{Ai} + \beta_3 x_{Bi} + \varepsilon_i$
43.  $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{Ai} + \beta_4 x_{1i} x_{Ai} + \varepsilon_i$
44.  $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
45.  $y_i = \beta_0 + \beta_1 x_{1i} + \beta_3 x_{Bi} + \beta_4 x_{1i} x_{Ai} + \varepsilon_i$
46.  $y_i = \beta_0 + \beta_1 x_{1i} + \beta_3 x_{Bi} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
47.  $y_i = \beta_0 + \beta_1 x_{1i} + \beta_4 x_{1i} x_{Ai} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
48.  $y_i = \beta_0 + \beta_2 x_{Ai} + \beta_3 x_{Bi} + \beta_4 x_{1i} x_{Ai} + \varepsilon_i$
49.  $y_i = \beta_0 + \beta_2 x_{Ai} + \beta_3 x_{Bi} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
50.  $y_i = \beta_0 + \beta_2 x_{Ai} + \beta_4 x_{1i} x_{Ai} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
51.  $y_i = \beta_0 + \beta_3 x_{Bi} + \beta_4 x_{1i} x_{Ai} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
52.  $y_i = \beta_1 x_{1i} + \beta_2 x_{Ai} + \beta_3 x_{Bi} + \beta_4 x_{1i} x_{Ai} + \varepsilon_i$
53.  $y_i = \beta_1 x_{1i} + \beta_2 x_{Ai} + \beta_3 x_{Bi} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
54.  $y_i = \beta_1 x_{1i} + \beta_2 x_{Ai} + \beta_4 x_{1i} x_{Ai} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
55.  $y_i = \beta_1 x_{1i} + \beta_3 x_{Bi} + \beta_4 x_{1i} x_{Ai} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
56.  $y_i = \beta_2 x_{Ai} + \beta_3 x_{Bi} + \beta_4 x_{1i} x_{Ai} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
57.  $y_i = \beta_1 x_{1i} + \beta_2 x_{Ai} + \beta_3 x_{Bi} + \beta_4 x_{1i} x_{Ai} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
58.  $y_i = \beta_0 + \beta_2 x_{Ai} + \beta_3 x_{Bi} + \beta_4 x_{1i} x_{Ai} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
59.  $y_i = \beta_0 + \beta_1 x_{1i} + \beta_3 x_{Bi} + \beta_4 x_{1i} x_{Ai} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
60.  $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{Ai} + \beta_4 x_{1i} x_{Ai} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
61.  $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{Ai} + \beta_3 x_{Bi} + \beta_5 x_{1i} x_{Bi} + \varepsilon_i$
62.  $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{Ai} + \beta_3 x_{Bi} + \beta_4 x_{1i} x_{Ai} + \varepsilon_i$

## ii) Interpretação de coeficientes de regressão de fatores de interação

Agora experimentamos ajustar exatamente o modelo (2) e, conforme suspeitas, verificamos que não apenas agora o coeficiente  $\beta_3$ , correspondente ao tratamento B, é significativamente diferente de zero, como a interação dos tratamentos também o é.

Table 3: Modelo de regressão linear para tratamentos com interação com idade sobre eficácia

Coeficiente	Estimativa	EP	Estatística t	p-valor
(Intercept)	6.211	3.350	1.854	0.074
idade	1.033	0.072	14.288	0.000
A	41.304	5.085	8.124	0.000
B	22.707	5.091	4.460	0.000
idade:A	-0.703	0.109	-6.451	0.000
idade:B	-0.510	0.110	-4.617	0.000

Há várias mudanças na interpretação dos coeficientes estimados em relação ao primeiro modelo ajustado. Primeiramente, vemos que a linha de base dos tratamentos está bem diferente com efeitos  $C < B < A$ , com uma grande diferença entre o primeiro e último tratamento.

O efeito da interação entre idade e tratamentos pode ser melhor explicada se analisarmos graficamente primeiro. A figura a seguir ilustra as curvas de regressão para cada tratamento.

Cabe recapitularmos que o grupo de referência é o tratamento C, o que força a interpretação de que o intercepto  $\beta_0$  do modelo é o seu efeito de tratamento isolado e  $\beta_1 x_{1i}$  se refere à interação entre o tratamento C e a variável idade.

Enquanto este coeficiente, do grupo de referência, é positivo e muito próximo a 1, os demais são negativos. No entanto, isto não significa que suas retas têm coeficiente angular negativo. O sentido e magnitude desses estimadores indica quanto a inclinação das demais retas está deslocada no sentido horário em relação à referência. Pela tabela, notamos que  $\beta_1 > \beta_4 > \beta_5$ . Analogamente, se  $\theta_j, j = A, B, C$  for o ângulo da curva de regressão em relação às abscissas, notamos que  $\theta_C > \theta_B > \theta_A$ .

Em termos reais, o modelo sugere que há uma grande influência da idade sobre a eficácia do tratamento C, enquanto essa influência é menor para o tratamento B e menor ainda para o tratamento A.

## iii) tabela ANOVA

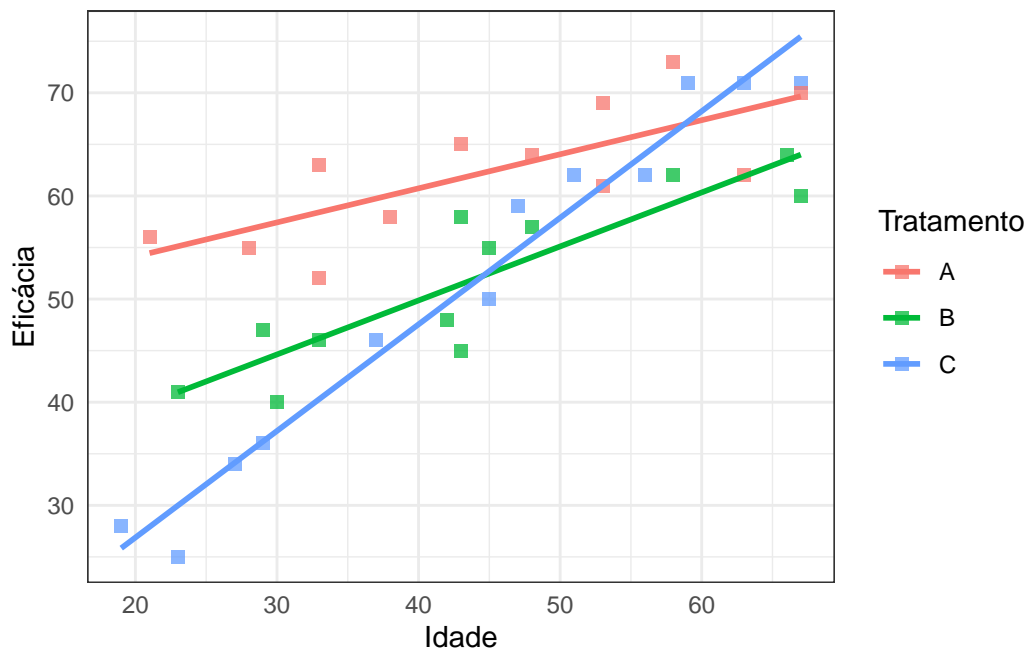


Figure 1: Curvas de regressão para modelo com interações.

Table 4: Tabela ANOVA para o modelo linear com interações

Fonte de Var.	g.l.	SQ	QM	F	p-valor
idade	1	3424.432	3424.432	222.295	0.000
A	1	803.804	803.804	52.178	0.000
B	1	1.189	1.189	0.077	0.783
idade:A	1	375.002	375.002	24.343	0.000
idade:B	1	328.424	328.424	21.319	0.000
Residuals	30	462.148	15.405	NA	NA

#### iv) Análise completa dos resíduos do modelo

### Apêndice

Todo o projeto de composição deste documento pode ser encontrado aqui: <https://github.com/cesar-galvao/mlg>