Lista 2

Modelos Lineares Generalizados - 2/2023

César Augusto Galvão - 19/0011572

Laiza Mendes - 20/0067028

Table of contents

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 durantes as análises).

Uma amostra dos dados é exibida na tabela a seguir:

y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
19.2	0	3.85	66	9.35	5.65	2.40	3.25	0.33	19	0.06
18.3	0	3.73	79	11.15	6.95	3.15	3.80	0.36	21	0.08
17.1	0	3.88	73	9.40	5.75	2.10	3.65	0.40	18	0.07
17.3	0	3.86	99	12.85	7.70	3.90	3.80	0.35	22	0.08
16.8	0	3.98	75	8.55	5.05	2.05	3.00	0.49	12	0.06
16.5	0	3.85	61	10.30	6.20	2.50	3.70	0.38	20	0.07

Cada variável do banco de dados apresenta as seguintes características:

- y: classificação de qualidade (20 no maximo);
- x1: variedade de vinho (0 — Cabernet Sauvignon, 1 — Shiraz);
- x2: nível de pH;
- x3: SO2 total (ppm);
- x4: densidade de cor;
- x5: cor de vinho;
- x6: cor de pigmento polimérico;
- x7: cor de antocianina;
- x8: antocianinas totais (g/L);
- x9: grau de ionização das antocianinas (porcentagem);
- x10: antocianinas ionizadas (porcentagem).

Diante dos dados acima, para a aplicação de um modelo de regressão é necessário verificar os níveis de correlação entre as variáveis. Com isso, podemos ter uma noção de quais variáveis podem apresentar multicolinearidade num modelo de regressão linear múltiplo.

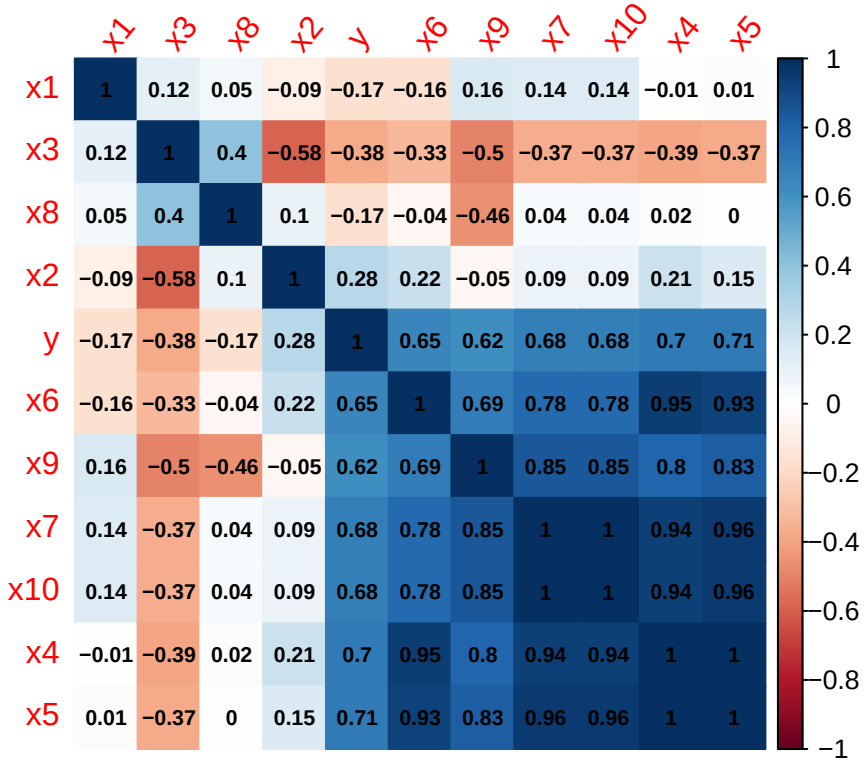


Figure 1: Correlograma das variáveis disponíveis

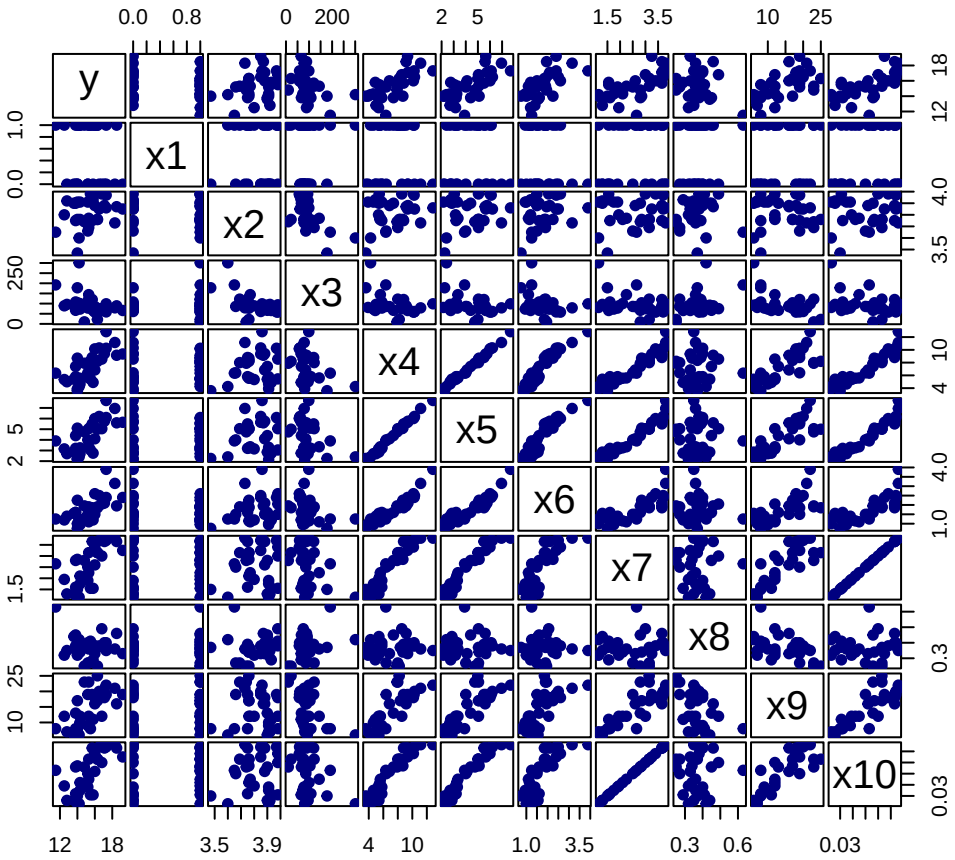


Figure 2: Correlações dois a dois das variáveis disponíveis.

Nota-se a partir dos gráficos apresentados que:

- A variável x1 tem apenas correlações fracas;
- A variável x2 tem apenas uma correlação negativa moderada com x3;
- A variável x3 tem correlação moderada com todas as variáveis, exceto com x1;
- A variável x4 tem correlação forte positiva com y, x5, x6, x7, x9, x10, além de uma correlação moderada negativa com x3;
- A variável x5 tem correlação forte positiva com y, x4, x6, x7, x9, x10, além de uma correlação moderada negativa com x3;
- A variável x6 tem correlação forte positiva com y, x4, x5, x7, x10, além de uma correlação moderada negativa e positiva com, respectivamente, x3 e x9;
- A variável x7 tem correlação forte positiva com x4, x5, x6, x9, correlação moderada negativa e positiva com, respectivamente, x3 e y, além de uma correlação perfeita com x10;
- A variável x8 tem apenas correlação moderada positiva e negativa, respectivamente, com x3 e x9;
- A variável x9 tem correlação forte positiva com x4, x5, x7, x10, além de uma correlação moderada positiva com y e x6 e negativa com x3 e x8;
- A variável x10 tem correlação forte positiva com x4, x5, x6, x9, correlação moderada positiva com y e negativa com x3, além de uma correlação perfeita com x7.

Independentemente das correlações, vamos aplicar um modelo de regressão linear múltiplo completo, com todas as variáveis, para analisar e modelar a relação entre a variável dependente ou resposta y e as variáveis independentes ou explicativas $x_i \ \forall \ i = \{1, 2, ..., 10\}$.

Table 1: Modelo de regressão linear múltipla completo, aplicado para todas as variáveis.

Coeficiente	Estimativa	EP	Estatística t	p-valor
(Intercept)	-12.208	14.612	-0.836	0.412
x1	-0.846	0.586	-1.443	0.162
x2	7.418	3.512	2.112	0.046
x3	0.010	0.009	1.220	0.235
x4	-1.947	2.221	-0.877	0.390
x5	4.895	3.218	1.521	0.142
x6	-1.434	1.813	-0.791	0.437
x8	-11.425	7.881	-1.450	0.161
x9	-0.108	0.220	-0.490	0.629

Nota-se que duas variáveis, x7 e x10, apresentam estimadores no modelo como `NA`. Isso contece porque há um elevado grau de correlação entre elas, o que atrapalha na estimação dos parâmetros de ambas com relação a variável resposta y.

Ignorando por enquanto esse fato, podemos fazer uma análise prévia dos resíduos desse modelo e averiguar se há alguma inconsistência. Para tal, faz-se uma análise gráfica dos e, posteriormente, um teste de normalidade dos resíduos e de autocorrelação entre os resíduos, além de um teste de heterocedasticidade do modelo.

Analisando os gráficos a seguir, nota-se do primeiro gráfico que a linha está aproximadamente horizontal. Logo, tem-se linearidade no modelo. No segundo gráfico vê-se que os resíduos aparentam ter distribuição normal. Já no terceiro gráfico, aparenta-se haver homocedasticidade, pois os resíduos estão dispersos como retângulo. Por fim, no quarto gráfico, para observar outliers e pontos influentes, não aparenta possuir resíduos outliers pois os resíduos estão entre -3 e +3.

Considerando os testes aplicados, nota-se que não há outliers nos resíduos, não rejeita-se normalidade dos resíduos, não há evidências de que há heterocedasticidade e não rejeita-se a independência dos resíduos, logo, eles não apresentam autocorrelação.

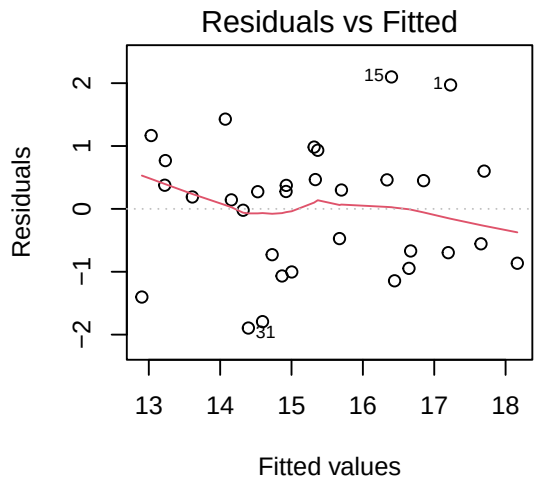
Table 2: Medidas estatísticas dos resíduos do modelo completo.

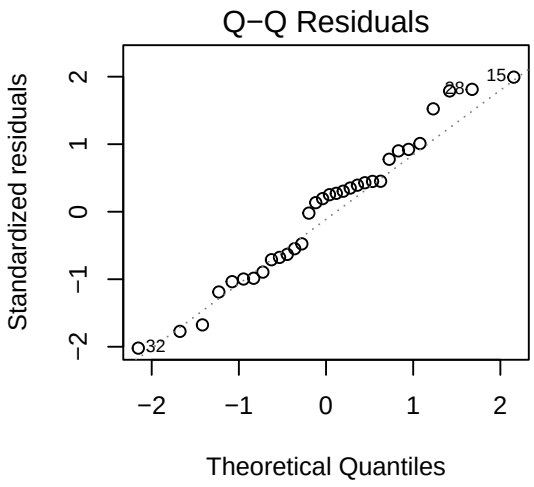
minimum	q1	median	mean	q3	maximum
-2.02	-0.757	0.223	0.01	0.533	1.993

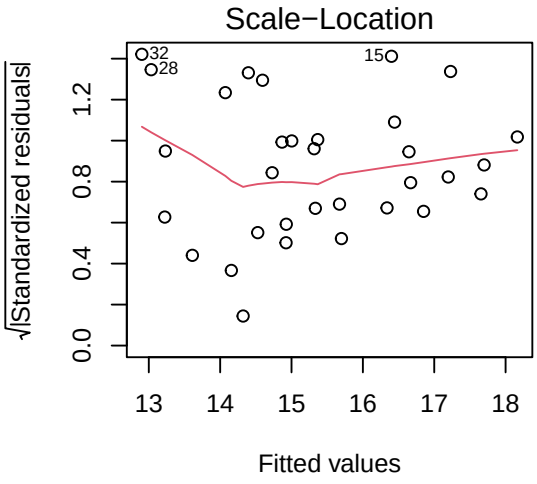
Table 3: Testes de hipótese para normalidade, heteroscedasticidade e independência.

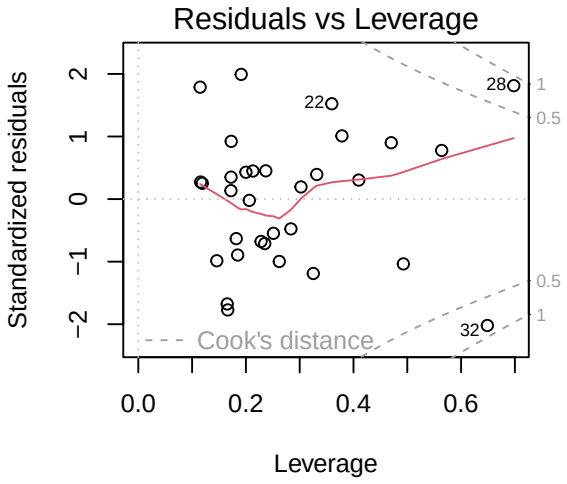
Teste	Estatística de teste	p-valor
Shapiro-Wilk normality test	0.974	0.615
studentized Breusch-Pagan test	3.036	0.932
Durbin-Watson Test	1.973	0.792

Já que os resíduos aparentam seguir as características esperadas, vamos avaliar se há multicolinearidade no modelo, uma vez que haviam fortes correlações entre as variáveis.


Residuals vs Fitted
Residuals
Fitted values
15
1
31


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
32
28
15


Scale−Location
√|Standardized residuals|
Fitted values
32
28
15


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
22
28
32

a) Proponha algum método para resolver o problema da multicolinearidade no conjunto de dados

Nota-se do tópico anterior que x7 e x10 são fortemente correlacionadas com outras variáveis, o que impossibilita de se estimar os parâmetros dessas variáveis para o modelo. Logo, podemos retirá-las e analisar novamente o modelo, avaliando se ainda há multicolinearidade nele por meio da medida VIF, calculada para cada variável.

Table 4: Modelo de regressão linear sem x7 e x10

Coeficiente	Estimativa	EP	Estatística t	p-valor
(Intercept)	-12.208	14.612	-0.836	0.412
x1	-0.846	0.586	-1.443	0.162
x2	7.418	3.512	2.112	0.046
x3	0.010	0.009	1.220	0.235
x4	-1.947	2.221	-0.877	0.390
x5	4.895	3.218	1.521	0.142
x6	-1.434	1.813	-0.791	0.437
x8	-11.425	7.881	-1.450	0.161
x9	-0.108	0.220	-0.490	0.629

Table 5: VIF para modelo sem x7 e x10

Variável	VIF
x1	1.971
x2	4.092
x3	4.513
x4	603.519
x5	511.870
x6	33.320
x8	7.931
x9	36.171

Analisando os VIFs, nota-se que apesar de se retirar x7 e x10 ainda há multicolinearidade no modelo. Poderíamos agora testar três métodos para se retirar a multicolinearidade, sem precisar retirar as variáveis: centrar, escalonar ou padronizar empiricamente as variáveis preditoras. No entanto, após testar os três métodos, percebeu-se que não mudou os índices de multicolinearidade. Logo, esses métodos não foram eficazes. Portanto, testaremos a remoção da variável de maior VIF, x4.

Table 6: Modelo de regressão linear sem x4, x7 e x10

Coeficiente	Estimativa	EP	Estatística t	p-valor
(Intercept)	-8.931	14.057	-0.635	0.531
x1	-1.008	0.553	-1.821	0.081
x2	6.267	3.242	1.933	0.065
x3	0.013	0.008	1.550	0.134
x5	2.250	1.115	2.017	0.055
x6	-2.517	1.320	-1.907	0.069
x8	-11.994	7.816	-1.534	0.138
x9	-0.079	0.217	-0.365	0.718

Table 7: VIF para modelo sem x4, x7 e x10

Variável	VIF
x1	1.775
x2	3.520
x3	4.131
x5	62.068
x6	17.840
x8	7.877
x9	35.369

Ainda notando-se alta multicolinearidade, podemos eliminar a variável x5 de maior VIF, ou seja, maior grau de correlação com as demais, e analisar o que acontece.

Table 8: Modelo de regressão linear sem x4, x5, x7 e x10

Coeficiente	Estimativa	EP	Estatística t	p-valor
(Intercept)	-21.075	13.460	-1.566	0.130
x1	-1.235	0.574	-2.150	0.041
x2	8.114	3.295	2.462	0.021
x3	0.015	0.009	1.695	0.102
x6	-0.211	0.700	-0.302	0.765
x8	1.827	3.987	0.458	0.651
x9	0.310	0.105	2.953	0.007

Table 9: VIF para modelo sem x4, x5, x7 e x10

Variável	VIF
x1	1.702
x2	3.239
x3	4.076
x6	4.463
x8	1.826
x9	7.379

Agora sim foi possível eliminar a multicolinearidade e seguir para uma seleção do melhor modelo final.

b) Usando algum método de seleção de variáveis, obtenha o modelo final para o conjunto de dados

Considerando os métodos de seleção de variáveis Forward, Bacward e Stepwise, pode-se selecionar o melhor modelo com base no critério de AIC.

Pelo método de seleção de variáveis *stepwise backward*, obtém-se o melhor modelo, com menor AIC (Critério de Informação de Akaike), no formato $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1\, x_{1i} + \widehat{\beta}_2\, x_{2i} + \widehat{\beta}_3\, x_{3i} + \widehat{\beta}_9\, x_{9i}$, excluindo-se as vaiáveis x4, x5, x6, x7, x8, x10 do modelo.

Table 10: Modelo de regressão linear com x1, x2, x3 e x9

Coeficiente	Estimativa	EP	Estatística t	p-valor
(Intercept)	-19.615	10.347	-1.896	0.069
x1	-1.122	0.446	-2.516	0.018
x2	7.911	2.494	3.172	0.004
x3	0.014	0.007	2.118	0.043
x9	0.280	0.051	5.488	0.000

Table 11: VIF para modelo linear com x1, x2, x3 e x9

Variável	VIF
x1	1.098
x2	1.986
x3	2.754
x9	1.862

Com isso, seleciona-se o seguinte modelo final, com todas as variáveis significativas e sem multicolinearidade:

$$y_i = \beta_0 + \beta_1 \, x_{i1} + \beta_2 \, x_{i2} + \beta_3 \, x_{i3} \; + \beta_4 \, x_{i9} + \varepsilon_i, \quad i = 1, 2, \ldots, 32. \tag{1}$$

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.

Para testar as hipóteses

$$H_0 : \beta_j = 0, \quad H_a : \beta_j \neq 0 \tag{2}$$

monta-se a seguinte tabela de análise de variância:

Table 12: Tabela ANOVA para o modelo linear com x1, x2, x2, x3 e x9

Fonte de Var.	g.l.	SQ	QM	F	p-valor
x1	1	2.805	2.805	1.968	0.172
x2	1	6.745	6.745	4.732	0.039
x3	1	6.223	6.223	4.365	0.046
x9	1	42.940	42.940	30.123	0.000
Residuals	27	38.488	1.425	NA	NA

Pode-se concluir que para um nível de significância de 0.05, a variável x1 não tem seu coeficiente significativamente diferente de zero para o modelo, o que significa que pode-se considerar a sua remoção. Isto é, essa variável não contribui de forma estatisticamente significativa para a explicação da variável y. Dessa forma, analisa-se novamente a ANOVA.

Nota-se ainda que, ao retirar x1, x3 também passa a não contribuir de forma estatisticamente significativa para a explicação da variável y, apresentando coeficiente significativamente igual a 0 para um nível de significância de 0,05. Logo, retira-se essa variável também.

Table 13: Modelo de regressão linear com x2, x3 e x9

Coeficiente	Estimativa	EP	Estatística t	p-valor
(Intercept)	-16.652	11.215	-1.485	0.149
x2	7.206	2.704	2.665	0.013
x3	0.010	0.007	1.428	0.164

Table 13: Modelo de regressão linear com x2, x3 e x9 *(continued)*

Coeficiente	Estimativa	EP	Estatística t	p-valor
x9	0.244	0.053	4.571	0.000

Table 14: Tabela ANOVA para o modelo linear com x2, x3 e x9

Fonte de Var.	g.l.	SQ	QM	F	p-valor
x2	1	7.483	7.483	4.410	0.045
x3	1	6.757	6.757	3.982	0.056
x9	1	35.450	35.450	20.892	0.000
Residuals	28	47.510	1.697	NA	NA

Por fim, para $\alpha = 0.05$, temos o modelo final para esses dados, que é

$$y_i = \beta_0 + \beta_1 \, x_{i2} + \beta_2 \, x_{i9} + \varepsilon_i, \quad i = 1, 2, \ldots, 32, \tag{3}$$

obtido considerando os resultados das tabelas a seguir.

Table 15: Modelo de regressão linear com x2 e x9

Coeficiente	Estimativa	EP	Estatística t	p-valor
(Intercept)	-4.660	7.566	-0.616	0.543
x2	4.507	1.968	2.291	0.029
x9	0.195	0.042	4.695	0.000

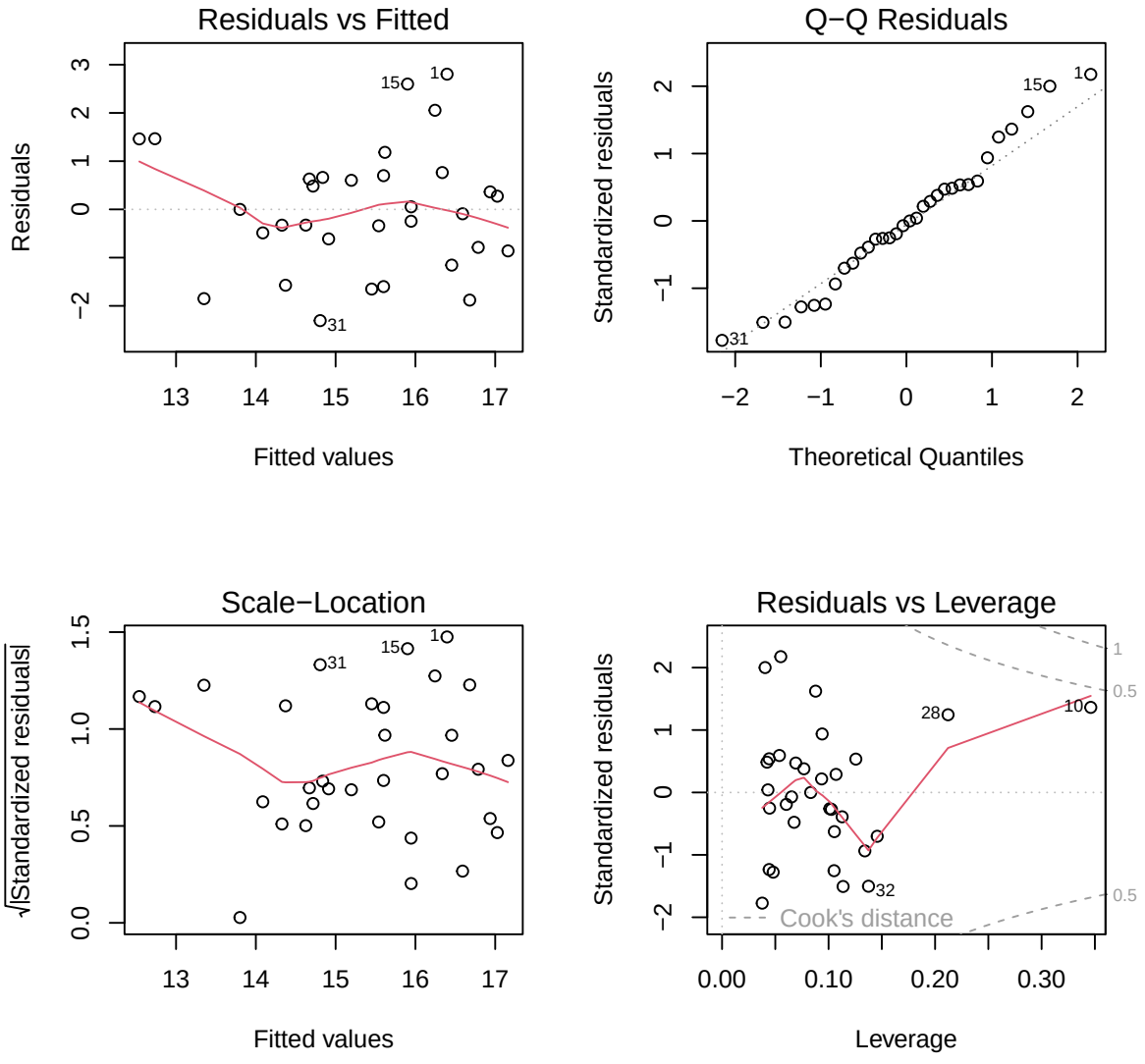
Table 16: Tabela ANOVA para o modelo linear com x2 e x9

Fonte de Var.	g.l.	SQ	QM	F	p-valor
x2	1	7.483	7.483	4.258	0.048
x9	1	38.747	38.747	22.045	0.000
Residuals	29	50.970	1.758	NA	NA

d) Com base no modelo obtido no item anterior, faça uma análise de resíduos e conclua.

Para uma análise de resíduos, fazemos uma análise gráfica e, posteriormente, teste de normalidade, de autocorrelação e de heterocedasticidade do dos resíduos.

Analisando os gráficos, nota-se que estão aproximadamente distribuídos em torno de zero, que os resíduos apresentam ter distribuição normal, que aparenta-se haver homocedasticidade e que não aparenta possuir resíduos outliers.



Confirmando as informações gráficas através dos testes, para um nível de significância de 0,05, nota-se que não há outliers nos resíduos, não rejeita-se normalidade dos resíduos, não há evidências de que há heterocedasticidade e não rejeita-se a independência dos resíduos, logo, eles não apresentam autocorrelação.

Table 17: Testes de hipótese para normalidade, heteroscedasticidade e independência.

Teste	Estatística de teste	p-valor
Shapiro-Wilk normality test	0.977	0.697
studentized Breusch-Pagan test	1.067	0.587
Durbin-Watson Test	1.421	0.090

Table 18: Medidas descritivas dos resíduos padronizados.

minimum	q1	median	mean	q3	maximum
-1.773	-0.646	-0.036	0.005	0.536	2.176

Portanto, como os resíduos estão de acordo com as suposições esperadas, pode-se concluir que o modelo escolhido no item anterior como modelo final, com a relação entre y, x2 e x9, está adequado. Fazendo uma análise desse modelo de regressão linear múltipla, poderia-se afirmar que para cada 1 unidade a mais de PH do vinho, a qualidade do vinho aumentaria, em média, 4.5 (t= 2.291; p-valor = 0.0294). Já para cada aumento de um grau de ionização das antocianinas (em porcentagem), a qualidade do vinho aumentaria, em média, 0.19 (t= 4.695; p-valor = 5.91e-05).

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 estudara 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 durantes 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

Um histograma da variável resposta é exibido a seguir, sugerindo assimetria na sua distribuição.

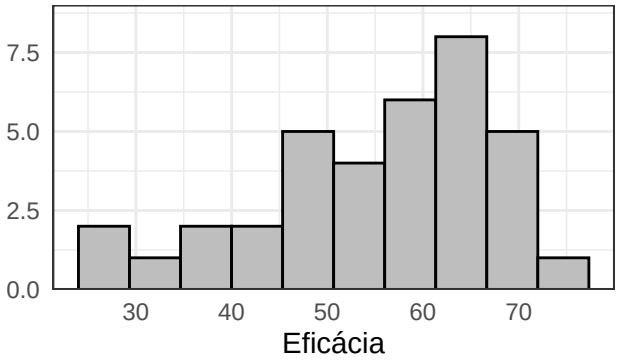


Figure 3: Histograma da variável resposta

No entanto, se montamos um histograma da variável resposta para cada valor de tratamento, vemos que há uma discrepência na sua distribuição. O tratamento A está bem concentrando em eficácias mais altas, enquanto o B está mais concentrado ao centro da métrica de eficácia e o C está amplamente distribuído.

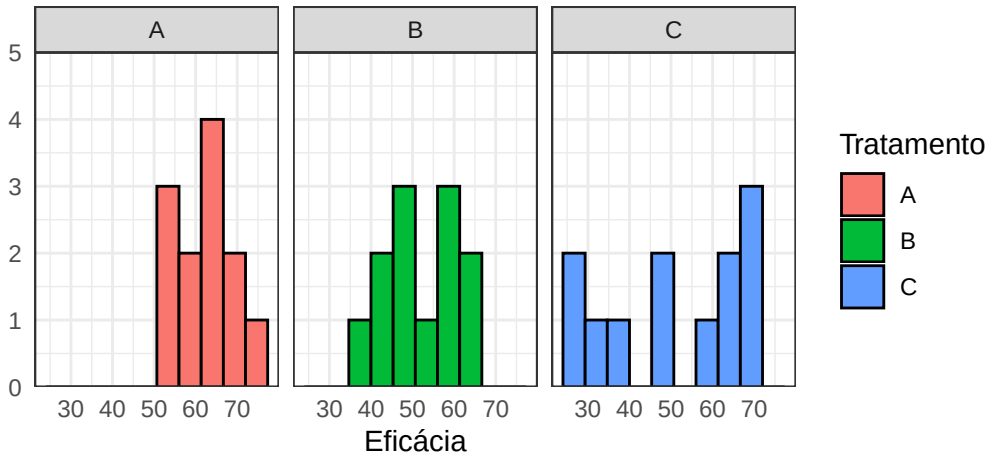


Figure 4: Histograma da variável resposta segregado por tratamento

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

Inicialmente, consideremos apenas um gráfico de dispersão entre a variável resposta e a única variável numérica, Idade. É possível notar uma relação que pode ou não ser linear, mas também há indícios de heteroscedasticidade. As demais variáveis são dicotômicas, então não há necessidade de se montar dispersões para elas.

Além disso, se segregamos a dispersão por grupos de tratamento, notamos que pode ser preferível um modelo que considere comportamentos de cada grupo separadamente.

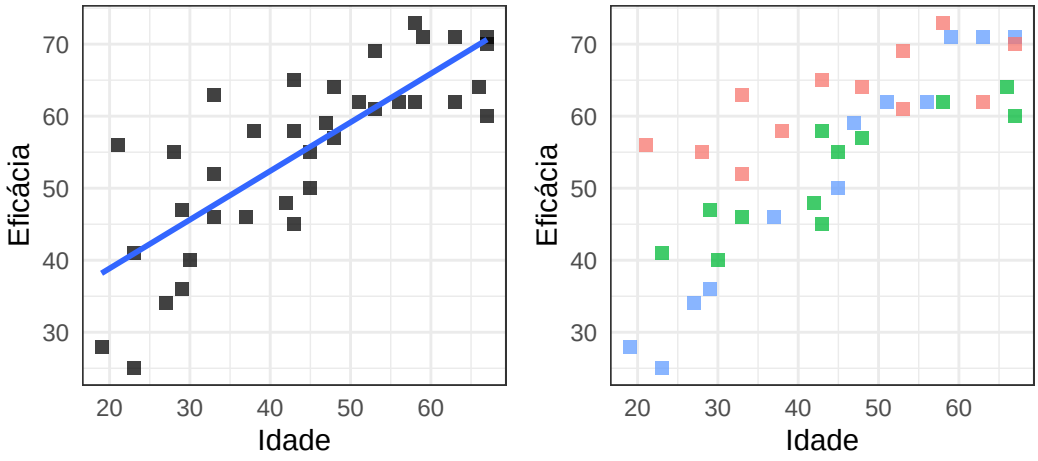


Figure 5: Gráficos de dispersão geral e segregado por tratamento

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

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{4}$$

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

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

Se simplesmente ajustamos um modelo de regressão linear – sem os termos de interação – utilizando (5) como referência na função `lm()`, obtemos os seguintes resultados:

Table 19: 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 (5) 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 20: 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 interessantemente 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 varáveis independentes

i) Lista de todos os submodelos possíveis

A partir do modelo (5), construimos 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 (5) e, conforme suspeitas, verificamos que não apenas agora o coeficiente β_3, correspondente ao tratamento B puro, é significativamente diferente de zero, como a interação dos tratamentos também o é.

Table 21: 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

Se avaliarmos os estimadores por intervalos de confianca a 95% de significância, apenas o intercepto compreende zero. De fato, dada a magnitude do intervalo, isto é consonante com o p-valor obtido para este estimador, de 0.07. Este valor está um pouco acima da significância sugerida, mas decide-se por mantê-lo por interpretabilidade do modelo.

Table 22: Intervalos de confiança para estimadores dos coeficientes de regressão

Estimador	LI(2.5%)	LS(97.5%)
(Intercept)	-0.630	13.052
idade	0.886	1.181
A	30.920	51.688
B	12.310	33.104
idade:A	-0.925	-0.480
idade:B	-0.735	-0.284

Há várias mudanças na interpretação dos coeficientes estimados em relação ao primeiro modelo ajustado. Primeiramente, vemos que o efeito mínimo dos tratamentos está bem diferente, com interceptos $C < B < A$ e 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.

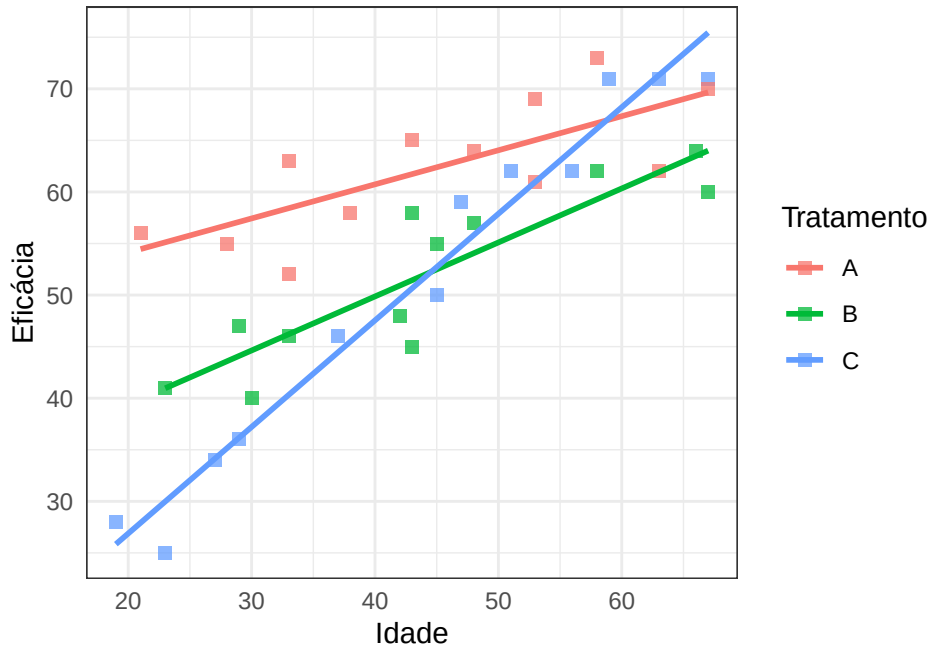


Figure 6: Curvas de regressão para modelo com interações.

Cabe recapitularmos que o grupo de referência é o tratamento C, o que força a interpretação de que o intercepto β_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.

Podemos expressar então o modelo ajustado da seguinte forma:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1\, x_{1i} + \hat{\beta}_2\, x_{Ai} + \hat{\beta}_3\, x_{Bi} + \hat{\beta}_4\, x_{1i}\, x_{Ai} + \hat{\beta}_5\, x_{1i}\, x_{Bi}$$

$$= \begin{cases} \hat{\beta}_0 + \hat{\beta}_1\, x_{1i}, & \text{se tratamento C} \\ (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_4) x_{1i}, & \text{se tratamento A} \\ (\hat{\beta}_0 + \hat{\beta}_3) + (\hat{\beta}_1 + \hat{\beta}_5) x_{1i}, & \text{se tratamento B} \end{cases}$$

$$= \begin{cases} 6.21 + 1.03\, x_{1i}, & \text{se tratamento C} \\ 47.51 + 0.33\, x_{1i}, & \text{se tratamento A} \\ 28.91 + 0.53\, x_{1i}, & \text{se tratamento B} \end{cases} \tag{6}$$

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

Finalmente, montamos a tabela de ANOVA do modelo. Em contraposição ao modelo anterior, sem interações, vemos que uma pequena parcela da soma de quadrados é atribuída aos resíduos. De fato, o coeficiente associado ao efeito puro do tratamento B ainda explica muito pouco da variação do modelo. No entanto, para que possamos considerar a interação Idade × Tratamento B, que testa significativamente para $H_a : \beta_j \neq 0$, mantemos o efeito puro.

Table 23: 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

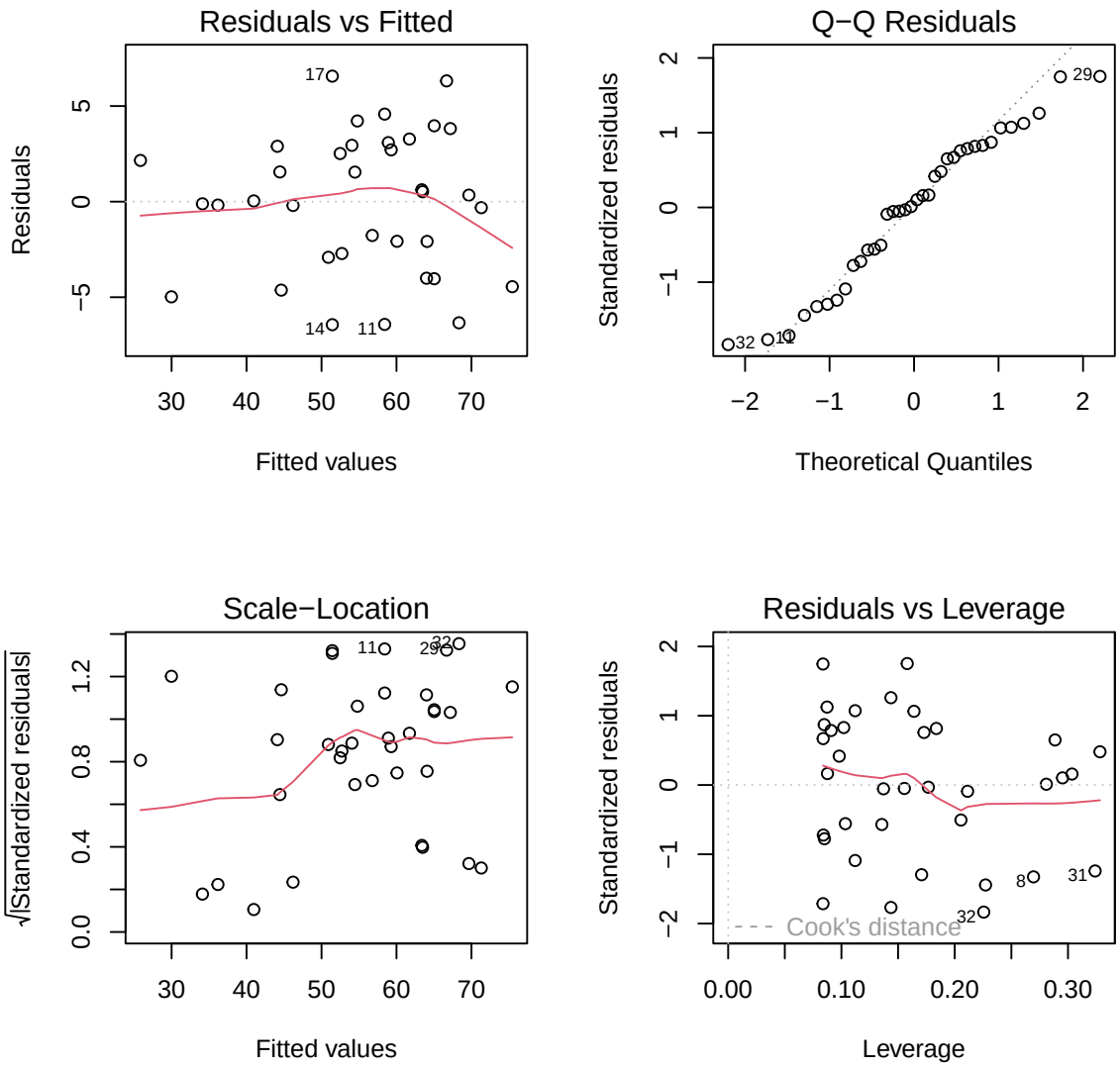
Iniciamos a análise dos resíduos do modelo, supondo-se que $\varepsilon_i \sim N(0, \sigma^2)$, com testes de hipótese para normalidade, heteroscedasticidade e independência dos dados. Avaliando apenas os p-valores dos testes, não rejeitamos as hipóteses nulas e podemos considerar a normalidade dos resíduos, constância da variância e independência dos dados.

Table 24: Testes de hipótese para normalidade, heteroscedasticidade e independência.

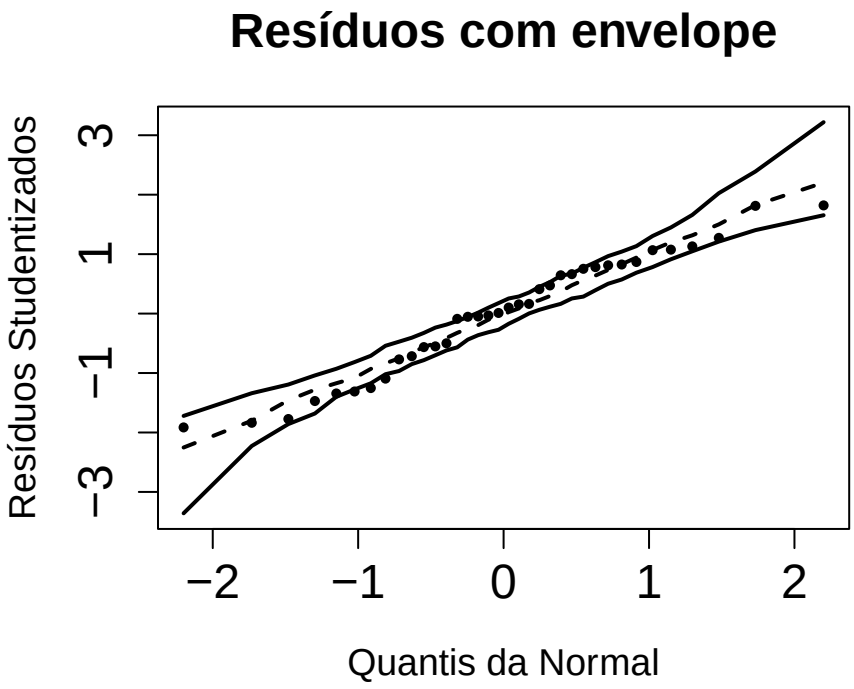
Teste	Estatística de teste	p-valor
Shapiro-Wilk normality test	0.963	0.263
studentized Breusch-Pagan test	2.996	0.701
Durbin-Watson Test	1.633	0.214

Se analisarmos graficamente, notamos que os resíduos parecem uniformemente distribuídos em torno de zero e com caudas mais pesadas no QQplot. Isto favorece a hipótese de não normalidade, mas contradiz os resultados do teste realizado. Verificaremos isso mais adiante.

Além disso, se avaliamos o gráfico de resíduos por sua alavancagem, vemos que há alguns pontos com distância de Cook muito próxima a dois. De fato, se utilizamos a função `stats::influence.measures()` encontramos 4 pontos de alavancagem por algum dos critérios apresentados. No entanto, nenhum deles está discrepante em ambos os eixos, de modo que podemos considerá-los bons pontos de alavancagem.


Residuals vs Fitted
Residuals
Fitted values
17
14
11
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
29
32
10
Scale−Location
√|Standardized residuals|
Fitted values
11
29
32
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
8
31
32

Por fim, investigamos um pouco melhor os resíduos studentizados, gerando um evenlopamento por bootstrap paramétrico[1]. De fato há pontos na borda da região designada como intervalo de confiança. No entanto, considerando a distância deles ao intervalo e os resultados dos demais testes, mantemos a afirmação de normalidade dos resíduos e não são propostos mais ajustes ao modelo.



Apêndice

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

```
if(!("pacman" %in% installed.packages())){install.packages("pacman")}

pacman::p_load(tidyverse, tidymodels, kableExtra, corrplot, plotrix, lmtest, psych, car, p


dados <- read.table("dados lista 2/Q02_data.txt", header=T)
```

[1] https://github.com/cesar-galvao/modelos_lineares/blob/main/envelope_function.R

```r
dados %>%
  ggplot(aes(eficacia))+
  geom_histogram(color = "black", fill = "gray", bins = 10)+
  scale_y_continuous(limits = c(0, 9),
                     expand = expansion(mult = 0, add = 0))+
  labs(x = "Eficácia", y = "")+
  theme_bw()+
  theme(axis.ticks = element_blank())

dados %>%
  ggplot(aes(eficacia, fill = tratamento))+
  geom_histogram(bins = 10, color = "black")+
  scale_y_continuous(limits = c(0, 5),
                     expand = expansion(mult = 0, add = 0))+
  labs(x = "Eficácia", y = "", fill = "Tratamento")+
  theme_bw()+
  theme(axis.ticks = element_blank())+
  facet_wrap(~tratamento)

geral <- ggplot(dados, aes(x = idade, y = eficacia))+
  geom_point(shape = 15, size = 2, alpha = .75)+
  geom_smooth(method = "lm", se = FALSE)+
  labs(y = "Eficácia", x = "Idade")+
  theme_bw()+
  theme(axis.ticks = element_blank())

trat <- ggplot(dados, aes(x = idade, y = eficacia, color = tratamento))+
  geom_point(shape = 15, size = 2, alpha = .75)+
  labs(y = "Eficácia", x = "Idade")+
  theme_bw()+
  theme(axis.ticks = element_blank(),
        legend.position = "none")

plot_grid(geral, trat)

#da encoding à variavel tratamento
dados_dummy <- dados %>%
  mutate(A = if_else(tratamento == "A", 1, 0),
         B = if_else(tratamento == "B", 1, 0)) %>%
  dplyr::select(-tratamento)
```

```r
#monta fit aditivo do modelo
fit_depressao <- lm(eficacia ~ (.), data = dados_dummy)

#tabela do modelo
fit_depressao %>%
  summary() %>%
  tidy()

anova(fit_depressao) %>%
  tidy()

fit_depressao_intera <- lm(eficacia ~ (.) + idade*A + idade*B, data = dados_dummy)

fit_depressao_intera %>%
  summary() %>%
  tidy()

nomes <- row.names(confint(fit_depressao_intera, level=0.95))

tabela <- confint(fit_depressao_intera, level=0.95) %>%
  as_tibble() %>%
  mutate(Estimador = nomes) %>%
  dplyr::select(Estimador, everything())

ggplot(dados, aes(x = idade, y = eficacia, color = tratamento))+
  geom_point(shape = 15, size = 2, alpha = .75)+
  geom_smooth(method = "lm", se = FALSE)+
  labs(y = "Eficácia", x = "Idade", color = "Tratamento")+
  theme_bw()+
  theme(axis.ticks = element_blank())

anova(fit_depressao_intera) %>%
  tidy()

bind_rows(
  ## Perform the normal Shapiro-Wilk test for the residuals
  shapiro.test(residuals(fit_depressao_intera)) %>% tidy(),

  ## Perform breush-pagan test for hetereocedascity
  (bptest(fit_depressao_intera) %>% tidy())[, c(1, 2, 4)],
```

```r
  ## Perform Durbin-Watson test for Independence
  (durbinWatsonTest(fit_depressao_intera) %>% tidy())[,c(1, 2, 4)]
) %>%
  dplyr::select(method, everything())

par(mfrow=c(2,2))
plot(fit_depressao_intera)

source("dados lista 2/envelope_function.R")
envelope_LR(fit_depressao_intera, OLS = T, main.title = "Resíduos com envelope")
```