Questão 7

Introdução

Um estudo realizado na Faculdade de Odontologia da Universidade de São Paulo tinha como objetivo comprar duas escovas de dente (convencional e monobloco) com respeito à redução de um índice de placa bacteriana. De um total de 32 crianças foram coletados os índices de placa bacteriana (IPB) antes e depois da escovação. Cada criança foi submetido a 4 tratamentos (escova convencional com dentifrício, escova convencional sem dentifrício,escova monobloco com dentifrício, escova monobloco sem dentifrício), com espaçamento de 1 semana entre eles para eliminar possíveis efeito residuais.

Nosso objetivo é comparar os quatro grupos formados pelas combinações dos fatores escova e dentifrício, assumindo que as observações não são correlacionadas e são homocedásticas, e determinar o grupo com melhor desemprenho. Para isso serão proposto modelos com duas variáveis respostas diferentes uma é $\frac{IPBdepois}{IPBantes}$ (será denotada no resto do relatório como Razão A) e a outra é $\frac{IPBdepois-IPBantes}{IPBantes}$ (será denotada no resto do relatório como Razão B). Para ambas quanto menor o seu valor melhor o desempenho da escovação. Ambas as variáveis respostas serão modeladas pelas variáveis explicativas tipo de escova e uso do dentifrício. A variável dentifrício será codificada da forma: 1 - uso de dentífrico; 0 - sem uso de dentifrício.

Para tal análises será usada a metodologia dos modelos lineares homocedásticos, metodologias de verificação da qualidade do ajuste e comparação de modelos apropriado, veja Azevedo (2019). Todas as análises serão feitas com auxílio computacional do R.

Análise Descritiva

Primeiramente devido a natureza da variável resposta duas observações tiveram que ser removidas do grupo, a do indivíduo 26 no dia 2 e a do indivíduo 24 também no dia 2, pois seus valores valores de IPB antes são iguais a zero e consequentemente a variável resposta não é um número. O ideal seria contatar o pesquisador para conversar sobre esses casos e decidir o que fazer, mas como não é possível tivemos que tomar essa ação.

Em ambas as variáveis respostas, Razão A e Razão B, notamos através dos boxplots das figuras 1 e 2 e das tabelas 1 e 2 que a escova convencional parece performar melhor que a monobloco. Também observamos que a escovação com dentifrício apresenta um melhor resultado do que sem. Os boxplots das figuras 1 e 2 também sugerem uma assimetria na distribuição dos dados, que não é ideal em um modelo linear normal.

Análise Inferencial

Iremos primeiramente ajustar dois modelos lineares homocedásticos, o para a Razão A será chamado de Modelo A e o para a Razão B será chamado de Modelo B. Em ambos consideraremos as duas variáveis respostas e a sua interação. Os modelos serão ajustados segundo o método dos mínimos quadrados ordinários e a significância de casa parâmetro será testada a partir de vários testes, veja Azevedo(2019).

Para ambos os modelos teremos a mesma equação do modelo. Seja Y_{ijk} a razão da k-ésima criança para o tipo de escova i = (convencional, monobloco) e para o uso de dentífrico j = (0,1). Temos o seguinte modelo:

$$Y_{ijk} = \mu + \beta_{0i} + \beta_{1j} + \beta_{0i}\beta_{1j} + \varepsilon_{ijk}$$

onde μ é o valor esperado de Y_{ijk} para quando a criança usar a escova tradicional e não usar dentífrico, $\beta_{0convencional} = 0$ e $\beta_{10} = 0$ pois estamos usando casela de referência, $\beta_{0tradicional}$ é o efeito da criança usar escova tradicional na média, β_{11} é o efeito do uso do dentífrico e $\beta_{0i}\beta_{1j}$ é o efeito de interação das duas variáveis. A parte aleatório do modelo é dada por $\varepsilon_{ijk} \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$.

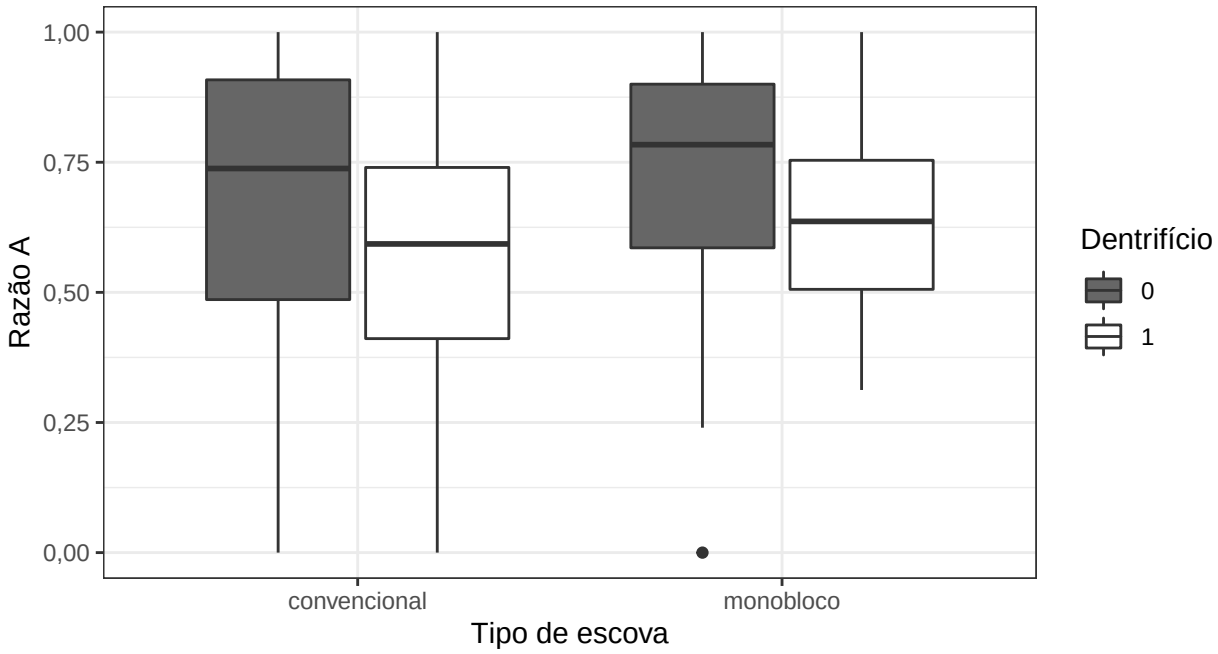

Figura 1: Boxplot para a Razão A pelo tipo de escova e pelo uso de dentrifício.

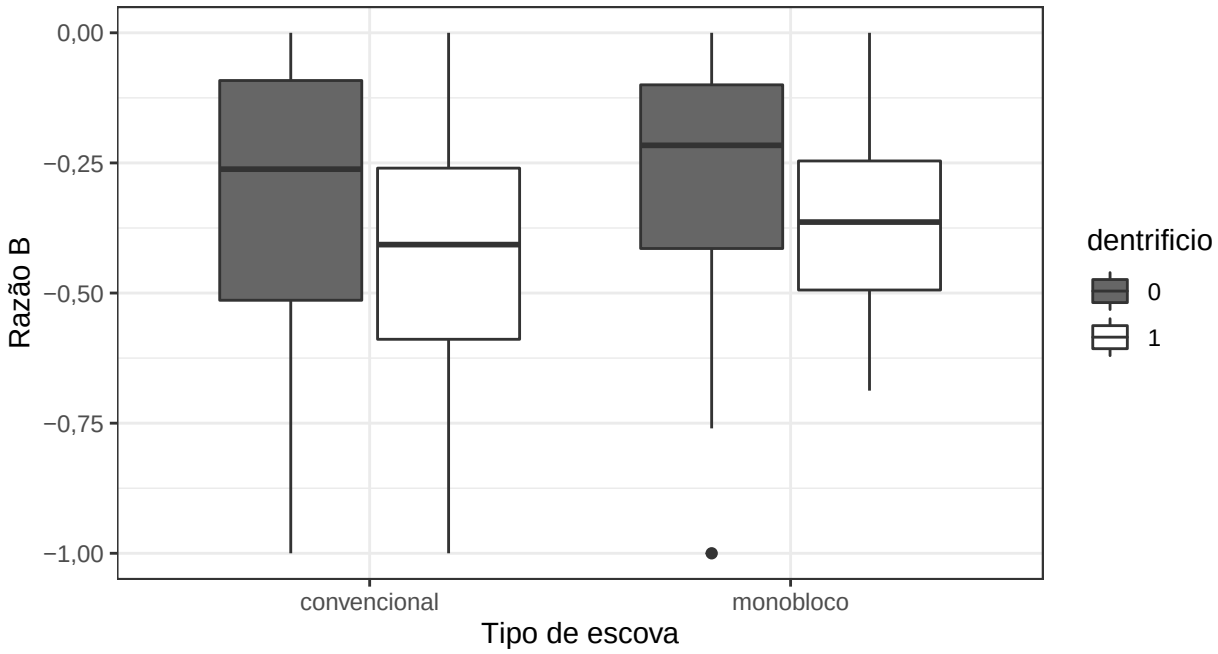

Figura 2: Boxplot para a Razão B pelo tipo de escova e pelo uso de dentrifício.

Tabela 1: Medidas descritivas para a Razão A.

Escova	Dentrifício	Média	Desvio Padrão	Coeficiente de Variação
convencional	0	0,6947	0,2521	36,29%
convencional	1	0,5827	0,2509	43,07%
monobloco	0	0,7272	0,2528	34,77%
monobloco	1	0,6412	0,1882	29,35%

Tabela 2: Medidas descritivas para a Razão B.

Escova	Dentrifício	Média	Desvio Padrão	Coeficiente de Variação
convencional	0	-0,3053	0,2521	82,58%
convencional	1	-0,4173	0,2509	60,13%
monobloco	0	-0,2728	0,2528	92,67%
monobloco	1	-0,3588	0,1882	52,45%

Os modelos são válido sobre as seguintes suposições: (i) $\varepsilon_{ijk} \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$; (ii) as observações são independentes; (iii) e a variância é constante.

Os modelos foram ajustado segundo o método dos mínimos quadrados ordinários e a significância de seus parâmetros foi determinada a partir de um teste de hipótese utilizando as estatística t. O resultado dos ajuste se encontra na tabela 3 e 3, podemos notar que ambas praticamente só o μ é significante para o model. O β_{11} pode ser considerado significativo dependendo do nível de confiança e pelas análises descritivas isso parece ser plausível. Para entender mais de quais fatores são significantes fizemos uma análise de covariância. Observando o resultado desses testes nas tabelas 5 e 6 chegamos a conclusão que o uso de dentifrício afeta o modelo.

Para ambos os modelos as análises de resíduos, figuras 3 e 4 indicaram quebra das suposições. O gráficos dos resíduos pelos valores ajustados indicam a presença de Heteroscedasticidade, os histogramas e os gráficos quantis-quantis indicam a ausencia de normalidade devido a forte assimetria. Apenas a suposição de independência parece se manter.

Devido ao mal ajuste dos modelos propostos, proporemos dois novos modelos que irão considerar apenas o fator do uso do dentífrico. Os modelos reduzidos serão dados por:

$$Y_{jk} = \mu + \beta_j + \varepsilon_{jk}$$

onde μ é o valor esperado na variável resposta quando a criança escova sem dentífrico, $\beta_0 = 0$ por adotarmos casela de referência e β_1 é o efeito esperado para quando a criança escova com dentífrico. Ainda temos as mesmas suposições: (i) $\varepsilon_{ij} \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$; (ii) as observações são independentes; (iii) e a variância é constante.

As novas estimativas podem ser vistas nas tabelas 7 e 8, dessa vez todos os seu parâmetros são significantes. No diagnóstico do modelo, nas figuras 5 e 6, podemos novamente rejeitar a suposição de normalidade, embora o problema de heterocedasticidade tenha sido resolvido ou pelo menos amenizado. Em relação aos pontos influentes ou alavanca, para ambos os modelos nenhum ponto se confirmou. Na medida H nenhum dos pontos passam do limiar e os pontos que se destacam na distância de cook na influenciam muito nas estatísticas do modelo e não mudam a conclusão.

Na comparação dos modelos os modelos A e B, e A e B reduzidos se mostraram praticamente idênticos, como visto na tabela 9. Porém ao comparar os modelos reduzidos aos completo os modelos reduzidos parecem melhor.

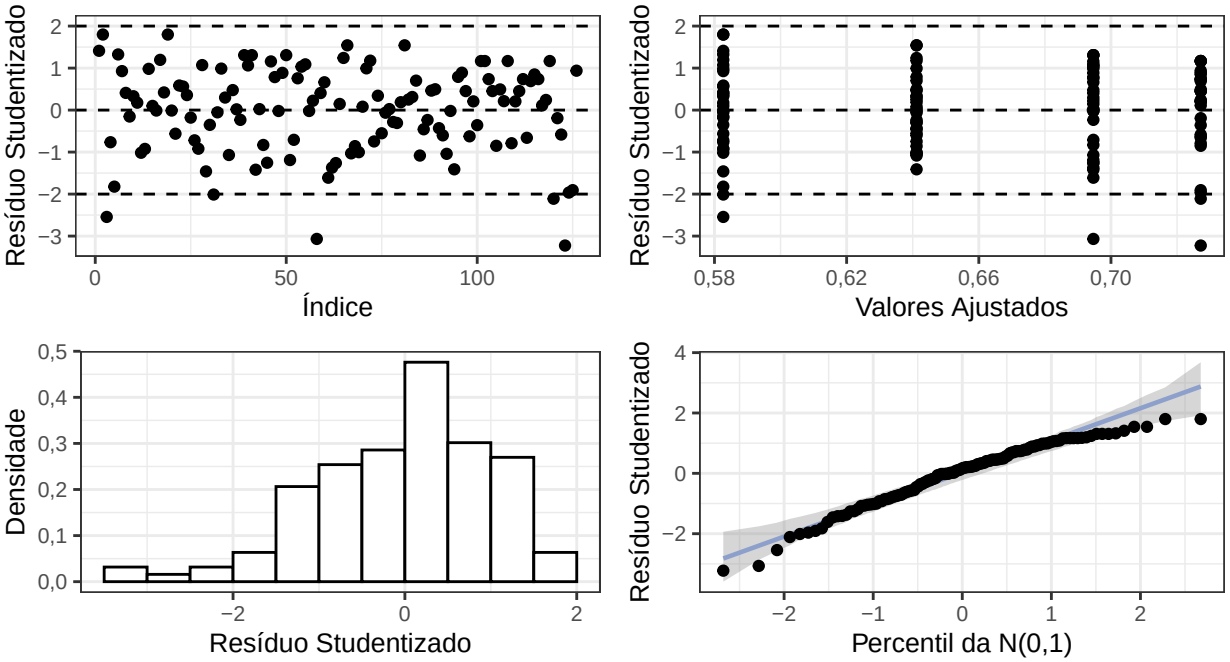

Figura 3: Diagnóstico do modelo A

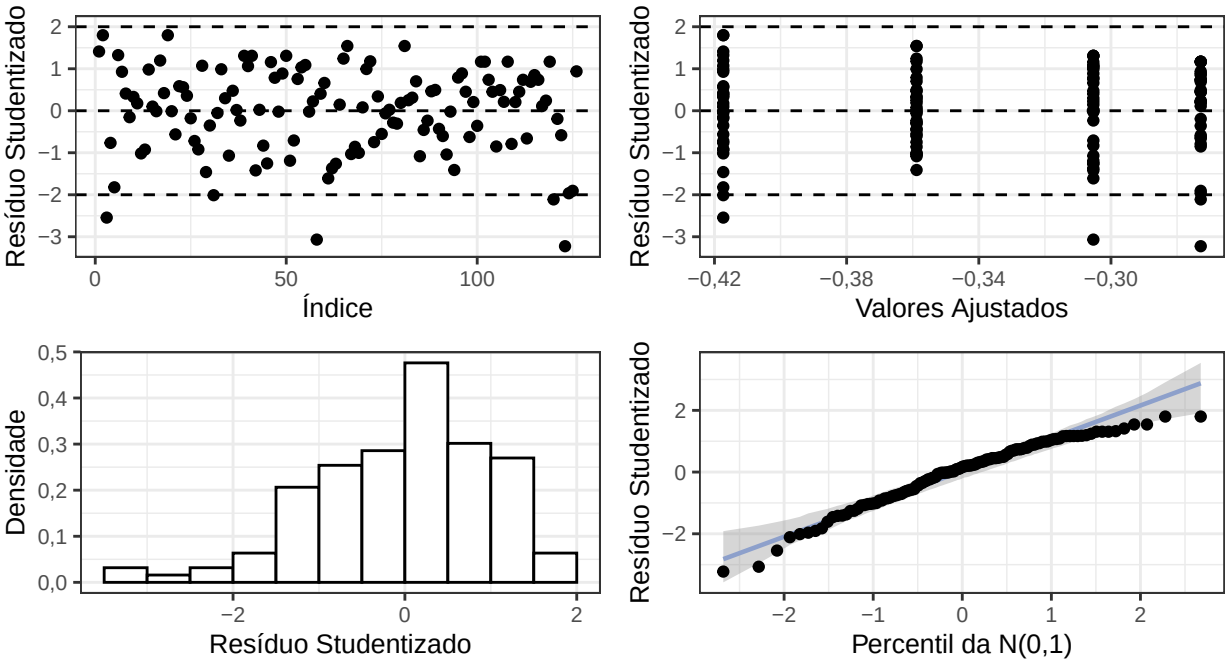

Figura 4: Diagnóstico do modelo B

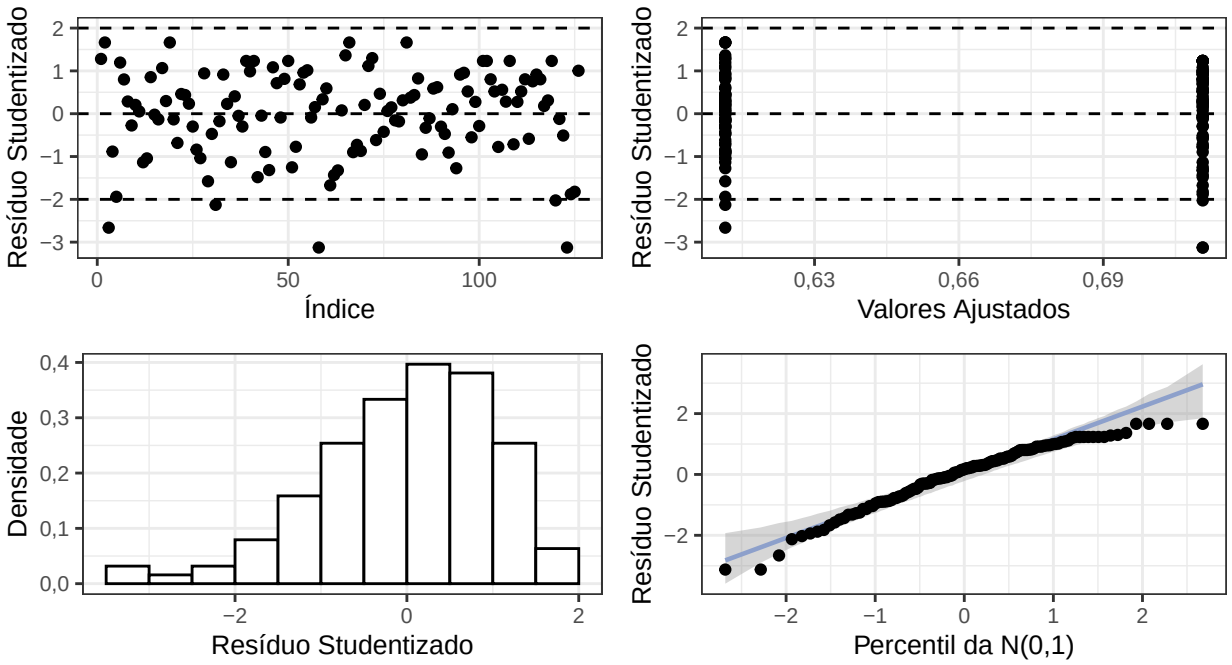

Figura 5: Diagnóstico do modelo A reduzido

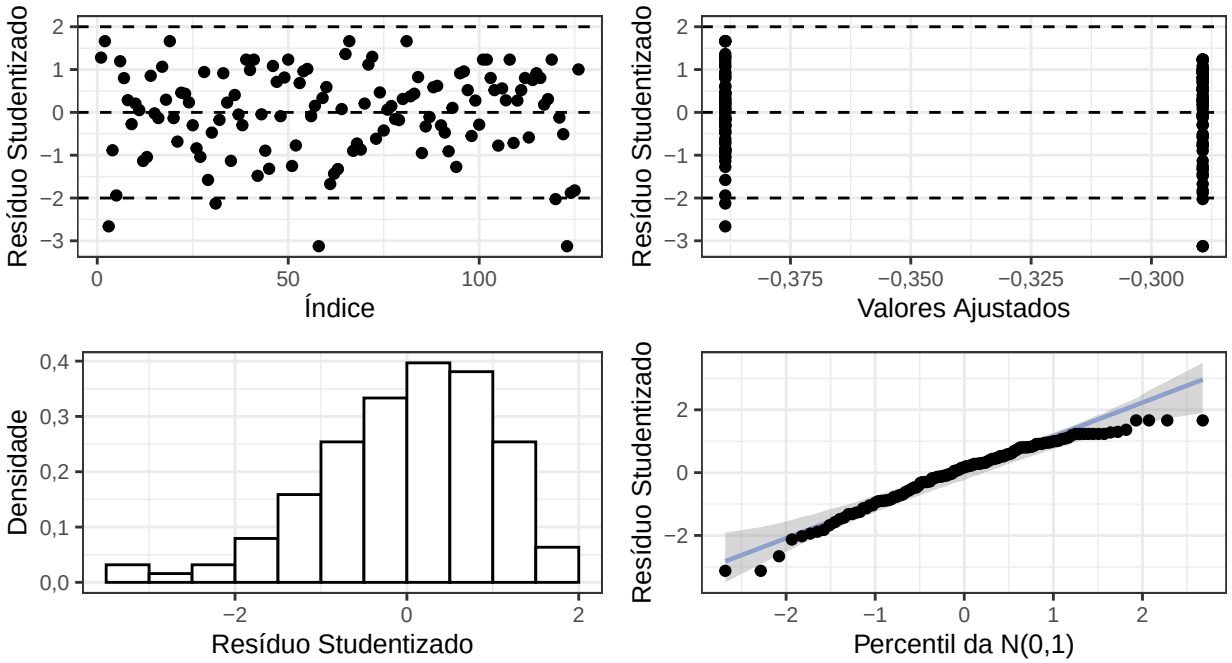

Figura 6: Diagnóstico do modelo B reduzido

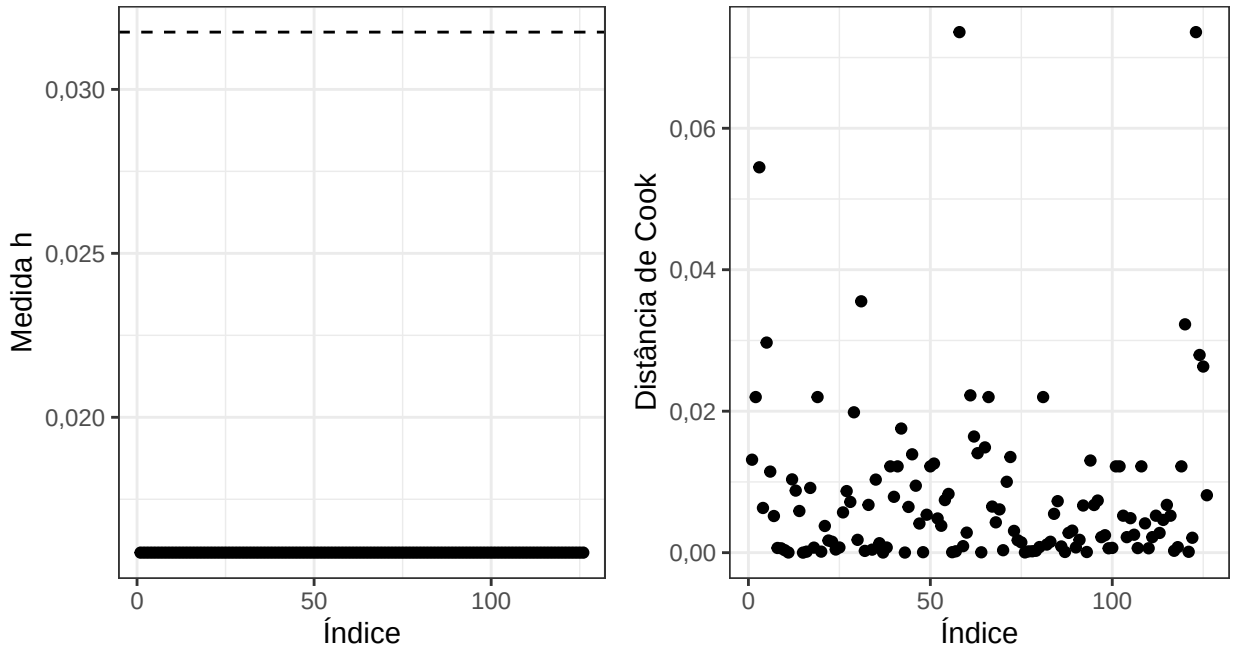

Figura 7: Análise de pontos influentes e alavancas para o modelo A reduzido

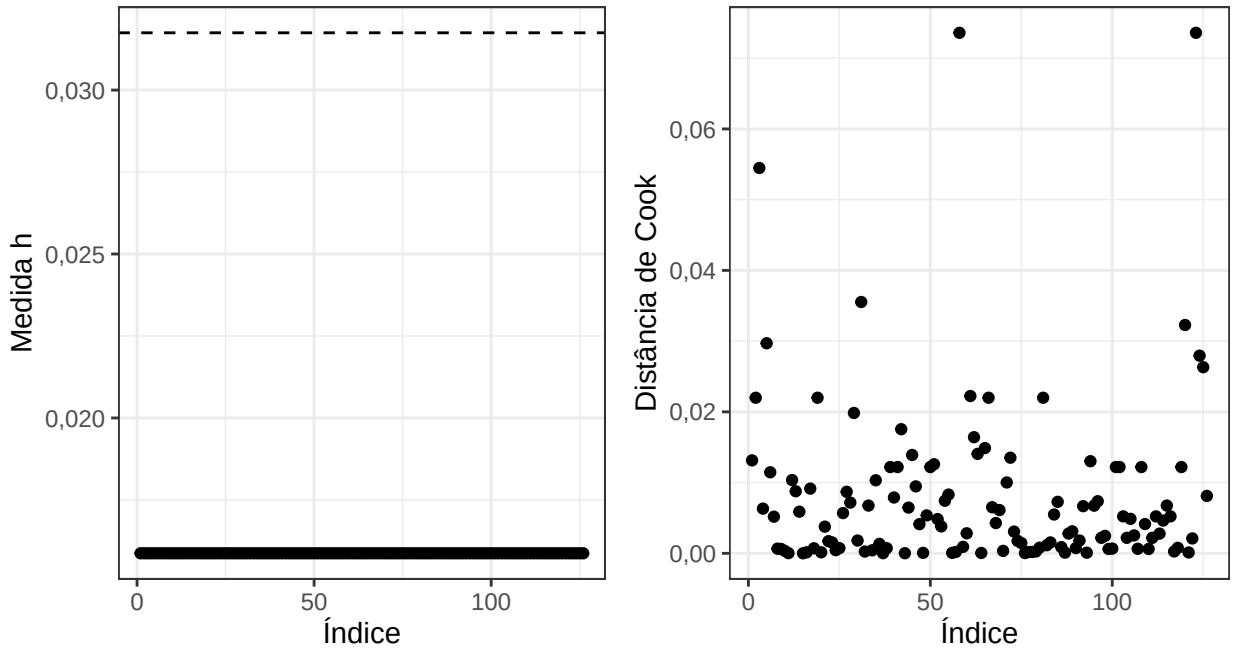

Figura 8: Análise de pontos influentes e alavancas para o modelo B reduzido

Tabela 3: Estimativa para os parâmetros do modelo A

Termo	Estimativa	Erro Padrão	IC (95%)	Estatística t	p-valor
μ	0,695	0,042	[0,612 ; 0,777]	16,520	< 0.001
$\beta_{0monobloco}$	0,032	0,060	[-0,085 ; 0,15]	0,542	0,589
β_{11}	-0,112	0,059	[-0,229 ; 0,005]	-1,883	0,062
$\beta_{0monobloco}\beta_{11}$	0,026	0,085	[-0,14 ; 0,192]	0,307	0,76

Tabela 4: Estimativa para os parâmetros do modelo B

Termo	Estimativa	Erro Padrão	IC (95%)	Estatística t	p-valor
μ	-0,305	0,042	[-0,388 ; -0,223]	-7,261	< 0.001
$\beta_{0monobloco}$	0,032	0,060	[-0,085 ; 0,15]	0,542	0,589
β_{11}	-0,112	0,059	[-0,229 ; 0,005]	-1,883	0,062
$\beta_{0monobloco}\beta_{11}$	0,026	0,085	[-0,14 ; 0,192]	0,307	0,76

Conclusão

Todos os modelos ajustados não se mostraram apropriados, ou seja quebraram algumas das suposições necessárias. Devido a isso as estatísticas e portanto as conclusões sobre os dados podem estar erradas. Em nossa análise o tipo de escova se mostrou não significativo no modelo, porem a análise descritiva sugere um efeito do mesmo, talvez se usássemos de um modelo mais apropriado mais apropriado esse fator se provesse significativo.

Trabalhando com os modelos que temos em mão os modelos reduzidos parecem mais apropriados que os modelos completos e aparentemente quebram menos suposições. Considerando as análises de diagnóstico e as medidas de comparação os modelos reduzidos parecem uma melhor escolha. Os modelos não se mostram diferentes em relação a razão A e razão B, portanto escolhemo o modelo de mais fácil interpretabilidade, que ao nosso ver é a razão A. Pois nessa razão quanto mais perto de 0 mais efetivo foi a escovação e para razão B quanto mais perto de -1.

Referências

Azevedo, C. L. N (2019). Notas de aula sobre Análise de regressão, http://www.ime.unicamp.br/~cnaber/Material_ME613_1S_2019.htm

Paula, G. A. (2013). Modelos de regressão com apoio computacional, versão pré-eliminar, https://www.ime.usp.br/~giapaula/texto_2013.pdf

Tabela 5: Análise de covariância para o modelo A

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrados Médios	Estatística	p-valor
β_{0i}	1	0,0652	0,0652	1,1520	0,2853
β_{1j}	1	0,3100	0,3100	5,4786	0,0209
$\beta_{0i}\beta_{1j}$	1	0,0053	0,0053	0,0941	0,7596
Erro	122	6,9029	0,0566	NA	NA

Tabela 6: Análise de covariância para o modelo B

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrados Médios	Estatística	p-valor
β_{0i}	1	0,0652	0,0652	1,1520	0,2853
β_{1j}	1	0,3100	0,3100	5,4786	0,0209
$\beta_{0i}\beta_{1j}$	1	0,0053	0,0053	0,0941	0,7596
Erro	122	6,9029	0,0566	NA	NA

Tabela 7: Estimativa para os parâmetros do modelo A reduzido

Termo	Estimativa	Erro Padrão	IC (95%)	Estatística t	p-valor
μ	0,711	0,030	[0,652 ; 0,769]	23,786	< 0.001
β_1	-0,099	0,042	[-0,182 ; -0,016]	-2,348	0,02

Tabela 8: Estimativa para os parâmetros do modelo B reduzido

Termo	Estimativa	Erro Padrão	IC (95%)	Estatística t	p-valor
μ	-0,289	0,030	[-0,348 ; -0,231]	-9,684	< 0.001
β_1	-0,099	0,042	[-0,182 ; -0,016]	-2,348	0,02

Tabela 9: Medidas de comparação do modelo.

	Modelo A	Modelo B	Modelo A Reduzido	Modelo B Reduzido
AIC	1,626195	1,626195	-1,0934252	-1,0934252
BIC	15,807604	15,807604	7,4154206	7,4154206
AICc	1,956773	1,956773	-0,9958642	-0,9958642
SABIC	-1,677900	-1,677900	-3,7454723	-3,7454723
HQCIC	4,235365	4,235365	-0,7888403	-0,7888403
-2log.lik	-8,373805	-8,373805	-7,0934252	-7,0934252