Modelagem via regressão linear aplicado em dados nutricionais de cereais matinais

João Inácio Scrimini[a], Joelmir Junior[a], Renata Stone[a]

[a] *Departamento de Estatística, Universidade Federal de Santa Maria*

Resumo

Os índices nutricionais são pouco explorados, o que levam a uma insuficiência das informações, no entanto, atualmente, o interesse por hábitos saudáveis é cada vez maior, exigindo o aprimoramento da modelagem de dados nutricionais. O presente estudo tem por objetivo definir um modelo de regressão linear que preveja qual a classificação nutricional de um cereal matinal a partir dos seus componentes nutricionais, ou seja, conseguir avaliar o quanto os componentes nutricionais do cereal podem influenciar na sua classificação. Para a seleção do modelo foram realizadas análises de diagnóstico e influência nos ajustes proposto. Na sequência foi realizada a validação do modelo e uma posterior testagem em cerais recentes e diferentes dos considerados no banco de dados.

1. Introdução

Na busca pela resolução de problemas das ciências exatas, testes que exploram a relação entre duas ou mais variáveis são amplamente utilizados. Pode-se analisar, por exemplo, como a variação dos salários interfere na taxa de desemprego ou como a altura dos pais pode determinar a altura dos filhos, dentre outros. Esta associação pode ser considerada determinística ou não determinística dependendo da forma em que o modelo prevê a resposta para o que buscamos (MONTGOMERY e RUNGER, 2003).

Na disciplina de Análise de Regressão, busca-se conhecer e dominar técnicas para analisar medidas observadas simultaneamente, procurando entender o relacionamento entre as mesmas, além de desenvolver um modelo de regressão linear simples e múltiplo satisfatório que explique uma variável através de outras que estão relacionadas entre si de maneira não determinística (MONTGOMERY e RUNGER, 2003).

No ajuste destes modelos, deve-se ponderar sempre a veracidade de certas suposições que garantem a validade do mesmo. Além disso, a escolha das variáveis é muito importante, pois a quantidade mínima garante uma economia na pesquisa, no entanto isto não pode prejudicar o estudo (Rodrigues, 2012). Para selecioná-las, utiliza-se diversas técnicas de análise que filtram as variáveis regressoras que melhor explicam o modelo, o qual também deve ser preferido com base em índices que indicam qual mais favorável.

Para aplicação destes conhecimentos, foi proposto o presente trabalho, no qual decidimos fazer a modelagem via modelo de regressão linear para dados nutricionais de cereais matinais. A rotulagem nutricional garante a segurança alimentar e zela pela saúde da população, facilitando a ciência do consumidor sobre as propriedades nutricionais dos alimentos, para assim favorecer escolhas conscientes e equilibradas. No entanto, os índices nutricionais são pouco

explorados, o que levam a uma insuficiência das informações e logo problemas de saúde generalizados. Recentes mudanças nos hábitos alimentares da população, sendo esses saudáveis, ocasionam diversos estudos utilizando dados nutricionais, aumentando a necessidade deles serem desenvolvidos e aprimorados com o passar do tempo (Barbosa, 2014).

Os cereais são produtos com alto teor de proteína, fibras, enriquecidos com vitaminas e sais minerais garantindo seu valor nutritivo. Podem, também, conter outros cereais, mel, açúcar, chocolate, ou frutas. Apesar de apresentarem uma composição variada, esses alimentos são constituídos principalmente por trigo, milho e arroz (Costa, 2012). Tais componentes podem influenciar nas escolhas alimentares das pessoas, os rótulos alimentícios vêm sendo estudados como fonte de informação nutricional aos consumidores.

Sendo assim, no banco de dados proposto para estudo, foram consideradas 77 observações de marcas de cereais, assim como seus índices nutricionais. O objetivo deste estudo é avaliar o quanto os componentes nutricionais do cereal podem influenciar na sua classificação através da modelagem via modelo de regressão linear. A partir do ajuste do modelo, buscamos entender a relação destas variáveis para classificar cada cereal com uma nota de 0 a 100, definindo assim as marcas mais saudáveis com melhor nota.

O trabalho está organizado conforme segue. A Seção 2 apresenta o estudo, análise descritiva das variáveis, modelagem, assim como a análise de diagnóstico e influência. A Seção 3 apresenta a validação do modelo. Por fim na Seção 4 as conclusões.

2. Seleção do modelo

2.1. Variáveis em estudo

O estudo foi direcionado para o banco de dados nutricional de aproximadamente 80 opções de cereais matinais disponíveis no supermercado local Wegmans no ano de 1990, coletado por alunos da Cornell University. Esse banco de dados em questão é uma adaptação disponível em Chris Crawford (2017). Nesse estudo ao todo serão 77 observações e 9 variáveis que estão descritas na Tabela 1.

Tabela 1: Variáveis do estudo sobre os cereais.

Variável	**Descrição**
Nota	Classificação de 0 a 100
Calorias	Calorias por porção (50kcal a 160kcal)
Proteína	Gramas de proteína (1g a 6g)
Gordura	Gramas de gordura (0g a 5g)
Sódio	Miligramas de sódio (0mg a 320mg)
Fibra	Gramas de fibra dietética (0g a 14g)
Açúcares	Gramas de açúcares (0g a 15g)
Vitaminas	Porcentagem típica de vitaminas e minerais (0, 25 ou 100)
Peso	Peso de uma porção (0,5 a 1,5), unidade (1) como xícara(200g)

Agora na Tabela 2 é apresentado um resumo das variáveis do estudo contendo suas medidas descritivas de posição, com valor mínimo, 1º quantil, mediana, valor médio, 3º quantil e valor máximo. Percebe-se que a maioria das variáveis por serem grandezas de medidas tem seu valor minimo igual a zero. É importante ressaltar, também, que a variável de desfecho nota,

para os 77 cereais observados nos dados, apresentou nota mínima de 18.4 e nota máxima de 93.70.

Tabela 2: Análise descritiva dos dados em estudo.

Variável	Mínimo	1º Quantil	Mediana	Média	3º Quantil	Máximo
Nota	18.04	33.17	40.40	42.67	50.83	93.70
Calorias	50.0	100.0	110.0	106.9	110.0	160.0
Proteínas	1.000	2.000	3.000	2.545	3.000	6.000
Gordura	0.000	0.000	1.000	1.013	2.000	5.000
Sódio	0.0	130.0	180.0	159.7	210.0	320.0
Fibra	0.000	1.000	2.000	2.152	3.000	14.000
Açúcares	-1.000	3.000	7.000	6.922	11.000	15.000
Vitaminas	0.00	25.00	25.00	28.25	25.00	100.00
Peso	0.50	1.00	1.00	1.03	1.00	1.50

Uma vez conhecidas as variáveis do estudo e suas análises descritivas, pode-se encontrar o modelo inicial contendo todas as variáveis,

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9, \tag{1}$$

em que β_1 é o intercepto do modelo de regressão linear, y a variável de desfecho nota do cereal e o vetor de covariáveis $(x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9)^\top$ = (Calorias, Proteína, Gordura, Sódio, Fibra, Açúcares, Vitaminas, Peso)$^\top$.

No modelo inicial (1) somente x_2 e x_9 não apresentaram significância a um nível de 5%, como pode ser visto na Tabela 3. O modelo apresentou coeficiente de determinação (R^2) de 0.9731 e R^2 ajustado ($\overline{R^2}$) de 0.97.

Pelos gráficos de dispersão, na Figura 1, não conseguimos identificar nenhuma possível regressão em alguma outra função que não seja linear. Portanto, aplica-se o método de Stepwise para analisar através do critério de informação de Akaike (AIC) o melhor modelo a ser proposto, retirando possíveis covariáveis não explicativas. Com isso, é proposto a eliminação das

Tabela 3: Coeficientes para o modelo inicial.

Variável	Estimativa ($\beta's$)	Erro padrão	Estatística t	p-valor ($>\|t\|$)	
Intercepto	59.750777	2.115304	28.247	< 2e-16	***
Calorias	-0.003967	0.036619	-0.108	0.91406	
Proteína	1.743235	0.355717	4.901	6.22e-06	***
Gordura	-4.167492	0.376992	-11.055	< 2e-16	***
Sódio	-0.047095	0.003759	-12.527	< 2e-16	***
Fibra	2.532346	0.202207	12.524	< 2e-16	***
Açúcares	-1.654571	0.089735	-18.439	< 2e-16	***
Vitaminas	-0.038255	0.013913	-2.750	0.00764	**
Peso	-2.207353	3.949774	-0.559	0.57810	

Código de significância: 0^{***} 0.001^{**} 0.01^{*} $0.05^{\cdot}$ 0.1 1

Figura 1: Gráfico das dispersões de todas as variáveis

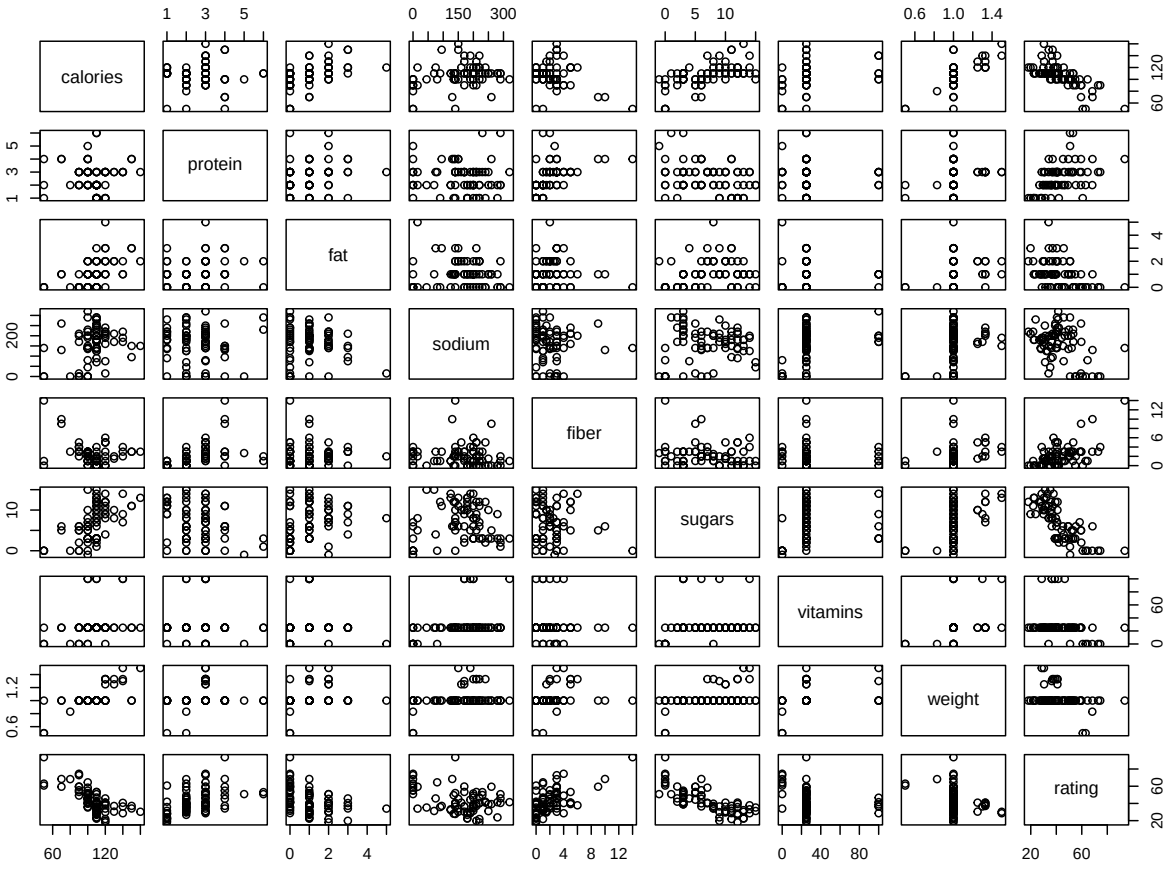


Tabela 4: Coeficientes para o primeiro ajuste no modelo.

Variável	Estimativa ($\beta's$)	Erro padrão	Estatística t	p-valor ($>\|t\|$)	
Intercepto	57.992712	1.110893	52.204	$<$ 2e-16	***
Proteína	1.629558	0.328131	4.966	4.64e-06	***
Gordura	-4.195320	0.306549	-13.686	$<$ 2e-16	***
Sódio	-0.048150	0.003559	-13.528	$<$ 2e-16	***
Fibra	2.517498	0.135743	18.546	$<$ 2e-16	***
Açúcares	-1.702690	0.072224	-23.575	$<$ 2e-16	***
Vitaminas	-0.041355	0.013456	-3.073	0.00302	**

Código de significância: 0*** 0.001** 0.01* 0.05· 0.1 1

covariáveis x_2 e x_9, calorias e peso, resultando no novo modelo ajustado:

$$y = \beta_1 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8. \tag{2}$$

O ajuste desse modelo proposto através do método de Stepwise apresentou na Tabela 4 significância de 5% em todas as covariáveis. Seu R^2 sofreu uma pequena redução para 0.9727 e $\overline{R^2}$ um pequeno aumento para 0.9704, ambos são valores altos.

2.2. Análise de diagnóstico e influência

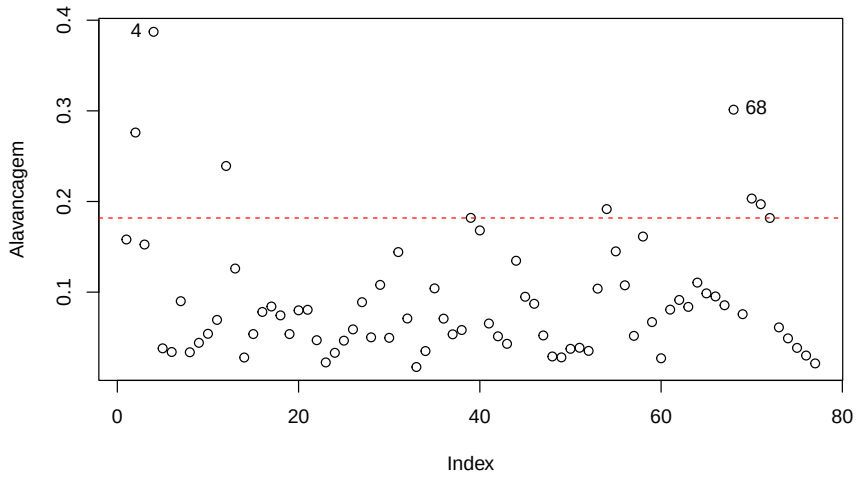
Sendo assim, vamos para a análise de influência para testar se existe alguma observação influente nos dados, causando alguma interferência significativa no ajuste do modelo. Um dos testes considerados na análise de influência é a alavancagem, que consiste na identificação de pontos de alavanca, observações com valores atípicos no espaço das variáveis explicativas, eles são potencialmente, entretanto não necessariamente, pontos influentes, que podem alterar de maneira significativa as estimativas e correspondentes erros padrões. Na Figura 2 pelo gráfico nota-se dois pontos mais sobressalientes aos demais, que são os pontos 4 e 68, mas não se observa uma diferença muito significativa neles.

O teste DFFIT na Figura 3 considera o quanto cada observação influencia sobre o y estimado ($\hat{y}$), medindo a alteração no valor ajustado. Percebe-se que com exceção da observação 58 todas as observações localizaram-se dentro do limite, indicando influências semelhantes das observações com $\hat{y}$. A observação 58 possui alta influência sobre o y estimado.

No teste DFBETAS considera-se os setes betas separadamente e a observação 58 está sendo influente em todos eles, em alguns de uma maneira mais intensa e em outros de forma mais leve, como pode ser visto na Figura 4. Alguns outros pontos, também, tiveram um pouco de influência em alguns betas, como a observação 12 nos betas 1 e 3, também a observação 71 apenas no beta 8.

Como foi visto na alavancagem a disposição dos pontos no espaço das variáveis é importante. Algumas observações podem exercer influência sobre as estimativas dos parâmetros, os valores previstos e as estatísticas utilizadas, por isso a aplicação de mais um teste de diagnóstico, a distância de Cook. Na Figura 5 percebe-se que somente a observação 58 está muito distante, indicando ser influente no modelo.

Figura 2: Gráfico para a alavancagem



Na Figura 6 está disponível o gráfico dos resíduos, percebe-se que há um outlier, ou seja uma observação que não é bem ajustada pelo modelo e se distancia das demais. Nesse caso o outlier é a observação 58 que nas análises já vem mostrando-se influente no modelo. Sendo a única observação fora do limite e muito distante.

Já avaliando a Figura 7 com o gráfico do envelope simulado baseado nos resíduos studentizados nota-se que os resíduos não estão tão bem ajustados nas bandas de confiança como desejado, a medida que nos extremos se tem mais pontos fora. Os resíduos studentizados são vantajosos por incorporar as variâncias individuais dos resíduos no escalonamento. Além de facilitarem a visualização de outliers, ou seja uma observação que não é bem ajustada pelo modelo e se distância das demais. No gráfico no canto inferior esquerdo percebe-se, também, a presença de um outlier.

A partir dos testes e análises realizados foi identificado a necessidade de retirar do modelo a observação 58, pois ela se destacou na análise de influência, este cereal possui uma observação faltando (-1) na variável Açúcares talvez um erro na coleta ou de digitação, mas nesse caso ele é um grande influente no ajuste do melhor modelo. Portanto, o modelo foi reajustado sem essa observação e todos os testes foram refeitos. Agora são considerados 76 cereais.

A observação 58 foi, inicialmente, retirada do modelo inicial (1), pois ela poderia estar influenciando também na exclusão ou aceitação de alguma covariável. Esse modelo havia apresentado R^2 de 0.9731 e $\overline{R^2}$ de 0.97. Nota-se que após o novo ajuste as covariáveis x_2 e x_9 seguem não sendo significativas a 5%, entretanto houve um significativo aumento do R^2 e $\overline{R^2}$ agora sendo 0.9918 e 0.9908, respectivamente.

Portanto continuaremos os testes, já que a observação 58 era, sim, influente no ajuste do modelo. Assim, testando se o modelo agora foi ajustado corretamente ou teremos que fazer alguma alteração ainda.

Figura 3: Gráfico para o DFFIT

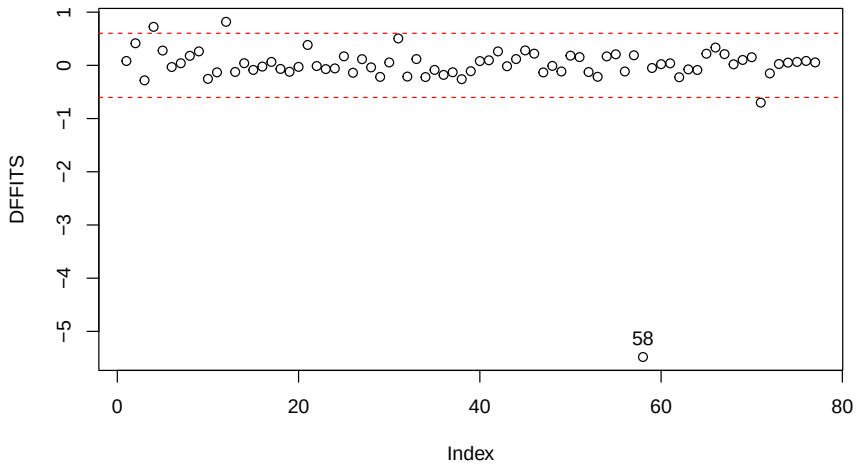


Tabela 5: Coeficientes para o segundo ajuste no modelo.

Variável	Estimativa ($\beta's$)	Erro padrão	Estatística t	p-valor ($>\|t\|$)	
Intercepto	58.412861	0.620282	94.17	$<$ 2e-16	***
Proteína	2.106591	0.186887	11.27	$<$ 2e-16	***
Gordura	-3.988571	0.171713	-23.23	$<$ 2e-16	***
Sódio	-0.052559	0.002016	-26.08	$<$ 2e-16	***
Fibra	2.399082	0.076273	31.45	$<$ 2e-16	***
Açúcares	-1.770267	0.040629	-43.57	$<$ 2e-16	***
Vitaminas	-0.047672	0.007519	-6.34	2.04e-08	***

Código de significância: 0*** 0.001** 0.01* 0.05· 0.1 1

Primeiramente, vamos refazer o teste de AIC para ver se essa observação retirada influenciou para alguma redefinição das covariáveis explicativas do modelo. Pelo teste seguimos com o modelo (2), as covariáveis x_2 e x_9, mas novos coeficientes como pode ser visto na Tabela 5.

Através desse ajuste o modelo proposto apresentou significância de 5% em todas as covariáveis. Seu R^2 passou para 0.9916 e $\overline{R^2}$ para 0.9909, ambos são valores altos. Como podemos perceber, pelo teste de AIC temos o mesmo modelo ajustado de antes, apenas agora sem a observação 58. Portanto, essa observação não teve influência para retirar ou colocar outra variável que fosse explicativa no modelo. Vamos agora para os testes de influência, testando se existe alguma outra possível observação que seja influente significativamente no modelo. Na Figura 8 temos o gráfico de Alavancagem e vemos os pontos 2, 4, 12 e 67 um pouco mais distante do resto, mas nada muito significativo.

No novo teste DFFIT, Figura 9, percebe-se que com exceção da observação 4, novamente, um pouco mais distante, mas também nada muito significativo, todas as observações localizaram-se ou dentro ou bem próximas do limite, indicando influências semelhantes das observações com $\hat{y}$.

Nesses setes gráficos para os DFBETAS presente na Figura 10 vemos que não temos nenhuma observação que se destaque, as observações 4, 31 e 67, apareceram mais vezes como possíveis influentes, mas nada que possamos concluir ainda de forma significativa.

Pelo gráfico da distância de Cook, Figura 11, nota-se que o ponto 4 é um dos que está sendo mais influente até agora. Percebe-se, também, as observações 31, 67 e 70, mas nada muito significativo.

Novamente ao ser analisado os resíduos, Figura 12, identifica-se que todos os pontos que apresentaram um pouco de variação praticamente não foram mencionados antes, a não ser a observação 31. Mas não vemos nenhuma variação muito significativa para uma possível influência de alguma outra observação. Agora gerando o gráfico para o envelope simulado baseado nos resíduos studentizados, Figura 13, percebe-se que a diferença é muito significativa. Agora todos os resíduos estão praticamente dentro das bandas de confiança.

Outro fator importante a ser considerado é a mudança na distribuição dos resíduos perceptível na Figura 14, que contem os histogramas do primeiro ajuste em comparação com o do segundo ajuste. Antes percebe-se que não tínhamos uma distribuição normal, entretanto agora já vemos que os resíduos estão seguindo uma distribuição normal.

Pelos testes feitos, pode-se notar que o modelo provavelmente já está bem ajustado. Portanto, vamos considerar algumas observações antes de prosseguimos para os testes das suposições do modelo. Pelos testes de influência, algumas das observações podiam estar influenciando de alguma forma no ajuste do modelo. Foi testado todas as observações mais influentes mencionadas, como a 4, 31 e 67. Mas, nenhuma apresentou alguma diferença significativa no modelo. Essas etapas não serão mostradas, pois ficaria muito repetitivo e irrelevante, também no propósito do trabalho.

Como o modelo está aparentemente ajustado, vamos para os testes das suposições. Testando se o modelo realmente está descrevendo corretamente a variável nota, classificação dos cereais, com bases nas covariáveis definidas.

3. Validação do modelo

A especificação de um modelo de regressão linear baseia-se em várias suposições. A validação do modelo ajustado depende da verificação dessas suposições, afinal não é prudente contar com o modelo até que a validade dessas suposições seja verificada:

S0 : O modelo está corretamente especificado;

S1 : A média dos erros é zero;

S2 : Homoscedasticidade dos erros;

S3 : Não haver autocorrelação;

S4 : Ausência de Multicolinearidade;

S5 : Normalidade dos erros.

Para testar [S0] é usado o teste RESET de especifição sob hipótese nula ($H0$) de o modelo está corretamente especificado. O p-valor encontrado $0.06261 > 0.05$ (valor do alpha) indica que não se rejeita $H0$, logo o modelo está corretamente especificado.

Para testar [S1] é usado o teste t para média dos erros sob $H0$ de a média dos erros ser igual a zero. O p-valor encontrado $0.9966 > 0.05$ (valor do alpha) indica que não se rejeita $H0$, logo a média dos erros é igual a zero.

Para testar [S2] é usado o teste de Bressch-Pagan (Koenker) de Heteroscedasticidade sob $H0$ de que os erros são homoscedásticos. O p-valor encontrado $0.962 > 0.05$ (valor do alpha) indica que não se rejeita $H0$, logo os erros são homoscedásticos.

Para testar [S3] é usado o teste de Durbin-Watson de autocorrelação sob $H0$ de não haver autocorrelação. O p-valor encontrado $0.404 > 0.05$ (valor do alpha) indica que não se rejeita $H0$, logo não há autocorrelação.

Para testar [S4] é usado Fatores de Inflação de Variância (VIF) para detectar multicolinearidade, em que o ideal é resultarem o valor 1. Os valores encontrados para x_3, x_4, x_5, x_6, x_7 e x_8 são aproximadamente 1, como pode ser visto na Tabela 6, logo não há multicolinearidade.

Tabela 6: Fatores de Inflação de Variância para as variáveis do modelo ajustado.

Variável	Proteína	Gordura	Sódio	Fibra	Açúcares	Vitaminas
VIF	1.634900	1.234248	1.137387	1.382335	1.307380	1.156292

Para testar [S5] é usado o teste Jarque-Bera de Normalidade sob $H0$ de os erros possuírem distribuição normal. O p-valor encontrado $0.8215 > 0.05$ (valor do alpha) indica que não se rejeita $H0$, logo os erros possuem distribuição normal.

Portanto, todas as suposições foram satisfeitas e validadas, informando que o modelo ajustado final,

$$y = \beta_1 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8,$$

em que β_1 é o intercepto do modelo de regressão linear, y a variável de desfecho nota do cereal e o vetor de covariáveis $(x_3,x_4,x_5,x_6,x_7,x_8)^\top$ = (Proteína, Gordura, Sódio, Fibra, Açúcares, Vitaminas)$^\top$, está adequado para modelagem do dados em estudo.

Em comparação com o modelo inicial (1) o ajuste resultou que somente as covariáveis x_2 e x_9 e a observação 58 foram excluídas do modelo final. Esse modelo apresentou $\overline{R^2}$ de 0.9909 e R^2 de 0.9916, logo aproximadamente 99% da variação da classificação do cereal é explicada pelas covariáveis propostas.

Substituindo os valores dos betas estimados presentes na Tabela 5, temos:

$$y = 58{,}413 + 2{,}107x_3 - 3.989x_4 - 0{,}053x_5 + 2{,}399x_6 - 1{,}77x_7 - 0{,}048x_8.$$

Nota-se que as covariáveis gordura, sódio, açúcares e vitaminas influenciam negativamente na média de y, enquanto proteína e fibra positivamente. A covariável fibra é a que mais influencia positivamente e gordura a que mais influencia negativamente.

3.1. Testagem do modelo

Para aplicarmos testes no modelo proposto, selecionamos os seguintes cereais, atualmente disponíveis em mercados, vale lembrar que o banco de dados até então usado é do ano de 1990:

- Cereal 1: Cereal Matinal NESCAU Tradicional 120g;
- Cereal 2: Cereal Matinal KELLOGG'S SUCRILHOS 730g;
- Cereal 3: Cereal Matinal sem glúten tradicional integral VITALIN 200g;
- Cereal 4: Cereal Matinal Integral com Flocos de Trigo, Aveia, Arroz e Milho sem Adição de Açúcar Nestlé Nesfit Caixa 220g.

Ambos são externos aos dados usados na definição do modelo de regressão linear e deles foram retirados os valores nutricionais referentes a uma porção(30g) e suas respectivas notas estimadas pelo modelo, mostrados na Tabela 7.

Tabela 7: Aplicação do modelo em outros cereais.

Variável	Cereal 1	Cereal 2	Cereal 3	Cereal 4
Proteínas (g)	1.1	1.9	1.7	3.2
Gordura (g)	0	1.4	0.5	0
Sódio (mg)	75	110	180	81
Fibra (g)	0.5	1.7	2	3.8
Açúcares (g)	12	9	0	0
Vitaminas (%)	25	25	0	25
Nota	35.55	38	55.34	68.82

Dos quatro cereais selecionados, dois são tradicionais (Cereal 1 e 2), já os outros dois (Cereal 3 e 4) são integrais, sendo que o Cereal 3, também, não contem glúten. A ideia foi identificar qual dos quatro cereais tem uma maior nota nutricional e também poder comparar os tradicionais e integrais entre eles. Como podemos ver pela Tabela 7, os dois cereais tradicionais tiveram notas bem próximas e consideravelmente baixas, sendo 35.55 para o Cereal 1 e 38 para o Cereal 2. Já os cerais integrais, mesmo não contendo notas muito elevadas, sendo 55.31 para o Cereal 3 e 68.82 para o Cereal 4, já contém notas bem mais agradáveis comparadas aos tradicionais. Diferente dos dois tradicionais que foram bem próximos entre eles, temos uma diferença de 13.51 pontos entre os integrais. Logo, o Cereal 4 nessa situação destacou-se como o melhor cereal, com nota 68.82, e o Cereal 1 como a menor avaliação, com nota 35.55, em comparação com os outros cereais matinais considerados.

4. Conclusões

O estudo, portanto, propôs um modelo de regressão linear para dados nutricionais de cereais matinais, contendo como covariáveis proteína, gordura, sódio, fibra, açúcares, vitaminas. Do modelo inicial foram retidas calorias e peso, assim como a observação 58, que apresentou forte influencia durante as análises. Aproximadamente 99% da variação da classificação do cereal é explicada pelas covariáveis propostas, o que indica um ótimo ajuste. Também pode-se concluir que a covariável fibra é a que mais influencia positivamente e gordura a que mais influencia negativamente na média da variável de desfecho. Para testar o modelo ajustado foi considerados outros cereais matinais, atuais, além do banco de dados, nessa aplicação destacaram-se com maiores notas os integrais, comprovando a eficiência do modelo em avaliar o

quanto os componentes nutricionais do cereal podem influenciar na sua classificação e definição de mais saudável.

Referências

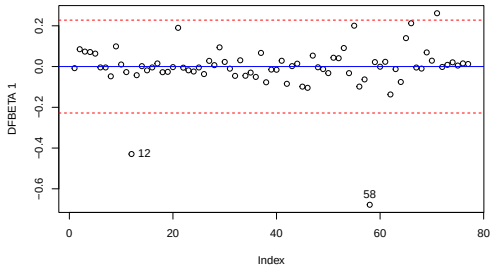
Barbosa, A. L. C., 2014. Alternativas para análise de dados nutricionais.

Chris Crawford, 2017. 80 cereals. Access date: 26 ago. 2021.
URL `https://www.kaggle.com/crawford/80-cereals`

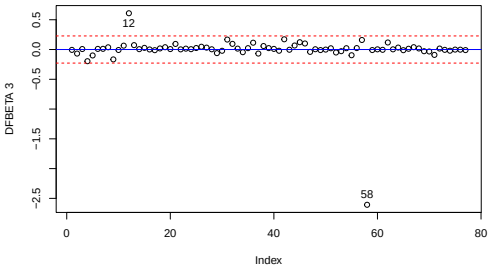
Costa, M. A. C. A., 2012. Avaliação exploratória da rotulagem nutricional de cereais de pequeno-almoço.

MONTGOMERY, D. C., RUNGER, G. C., 2003. Estatística aplicada e probabilidade para engenheiros, 2ª. Ed. Rio de Janeiro: Editora LTC.

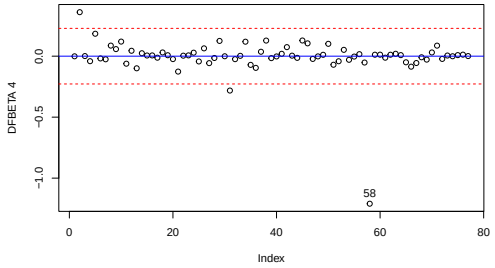
Rodrigues, S. C. A., 2012. Modelo de regressão linear e suas aplicações. Tese de Doutorado, Universidade da Beira Interior.



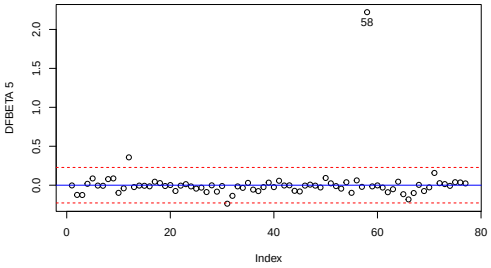
(a) DFBETA1



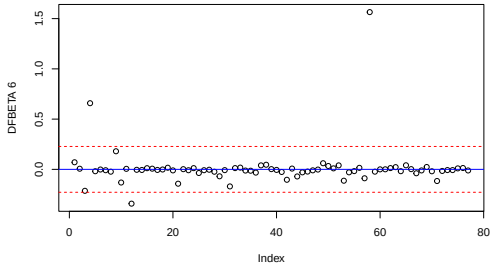
(b) DFBETA3



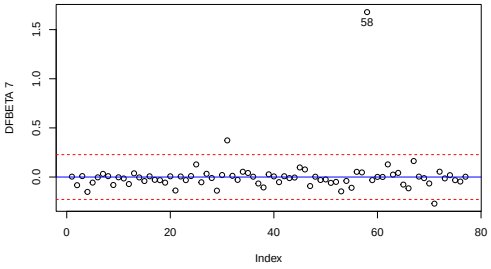
(c) DFBETA4



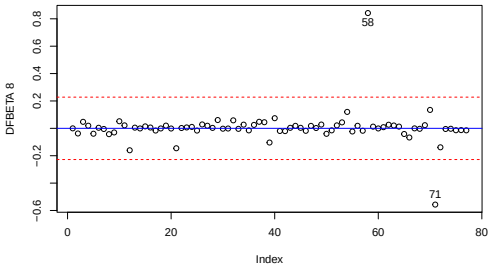
(d) DFBETA5



(e) DFBETA6



(f) DFBETA7



(g) DFBETA8

Figura 4: Gráficos para os DFBETAS.

Figura 5: Gráfico para a distância de Cook

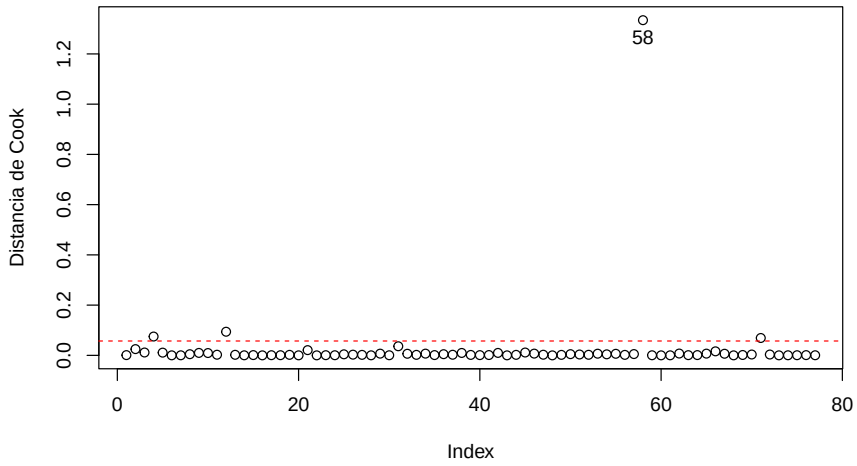


Figura 6: Gráfico para os resíduos

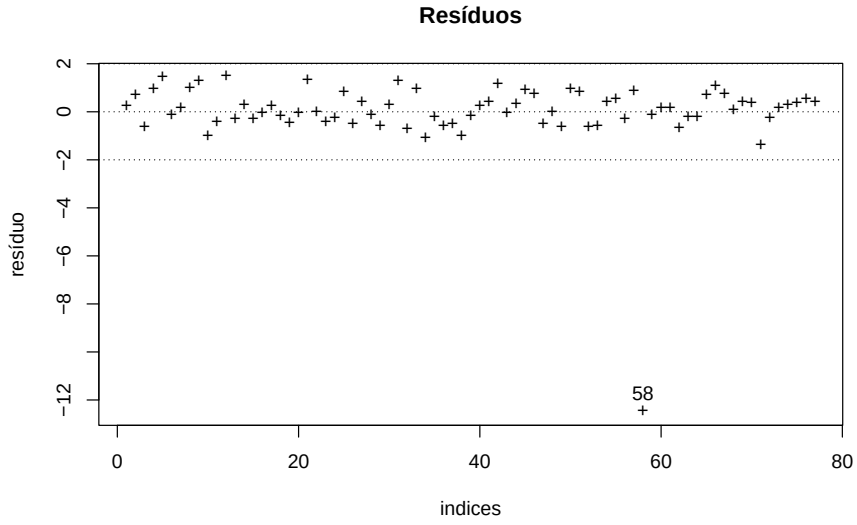


Figura 7: Gráfico para o envelope simulado baseado nos resíduos studentizados

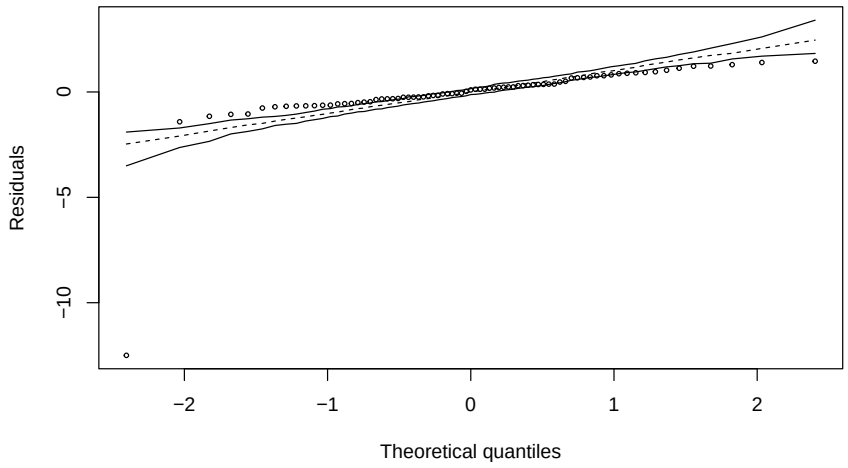


Figura 8: Gráfico para a alavancagem

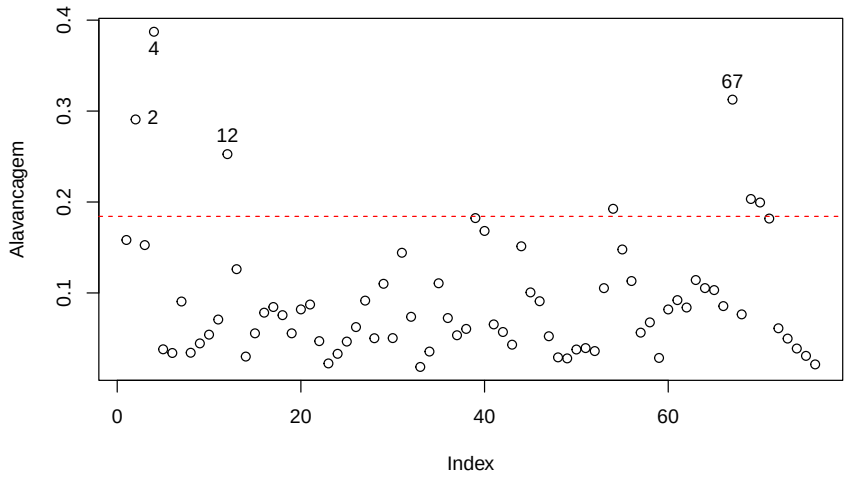
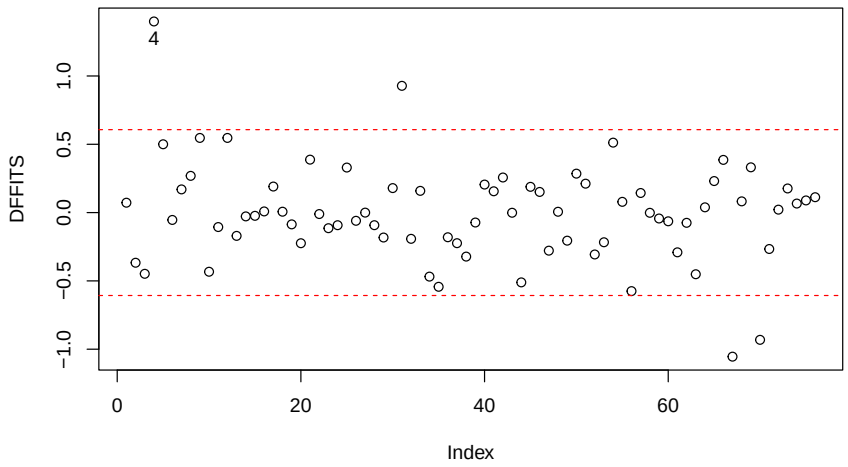
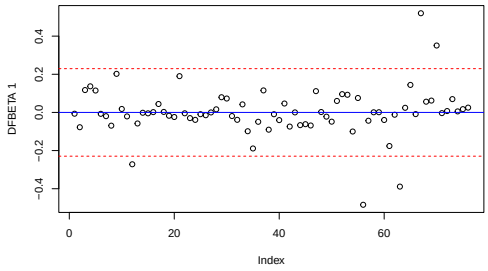
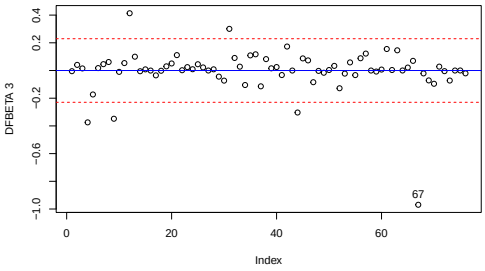


Figura 9: Gráfico para o DFFIT

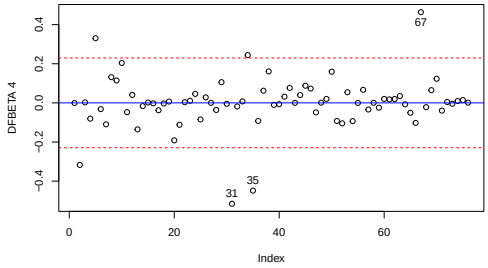




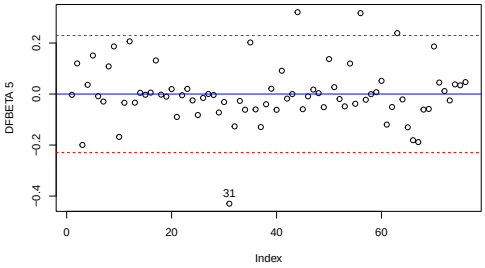
(a) DFBETA1



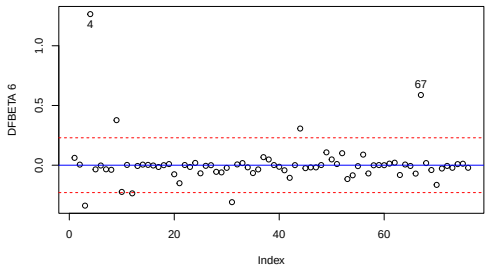
(b) DFBETA3



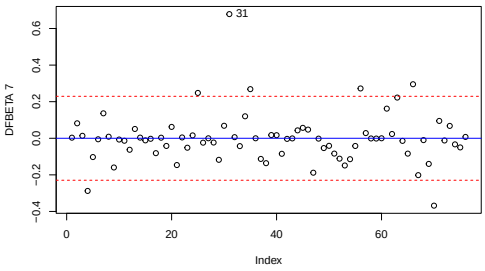
(c) DFBETA4



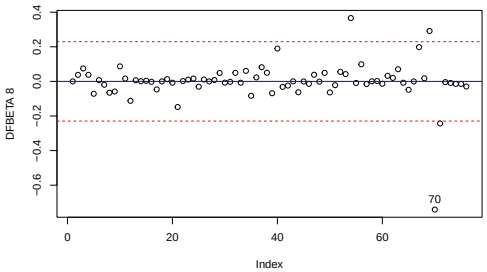
(d) DFBETA5



(e) DFBETA6



(f) DFBETA7



(g) DFBETA8

Figura 10: Gráficos para os DFBETAS.

Figura 11: Gráfico para a distância de Cook

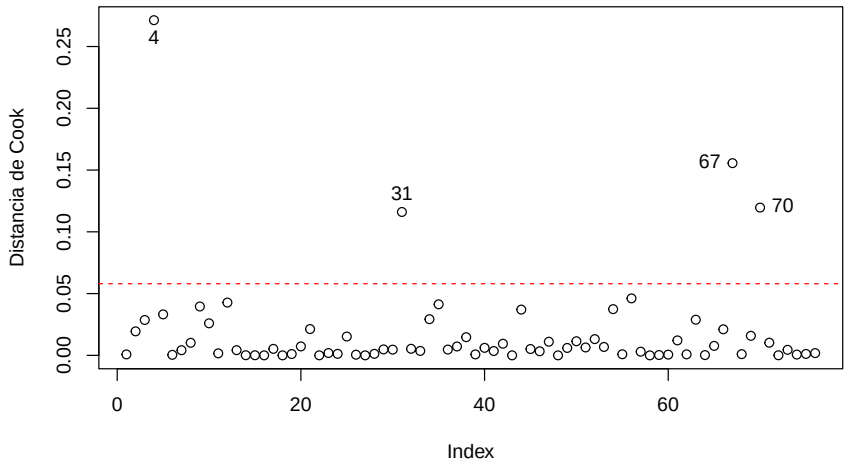


Figura 12: Gráfico para os resíduos

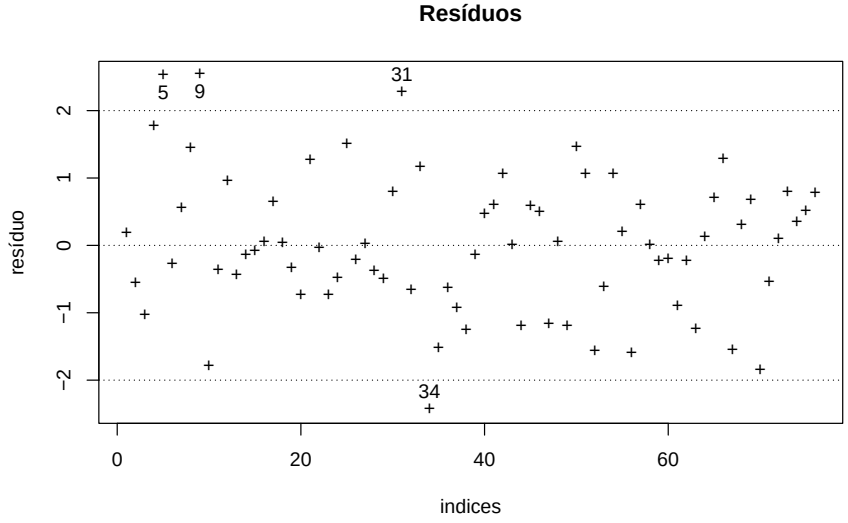
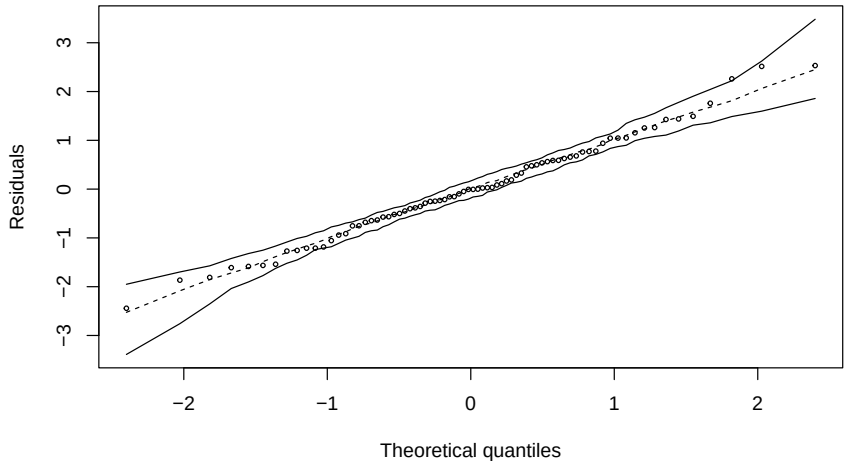
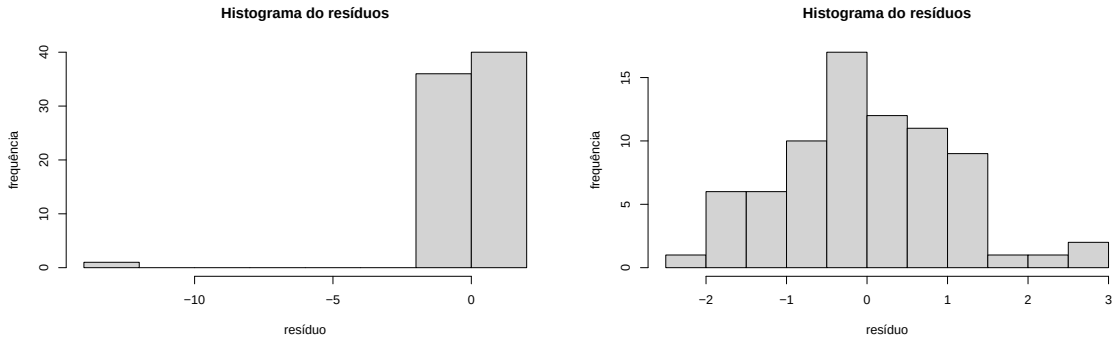


Figura 13: Gráfico para o envelope simulado baseado nos resíduos studentizados





(a) Primeiro ajuste (b) Segundo ajuste

Figura 14: Histograma dos resíduos para o primeiro e segundo ajuste do modelo.