Universidade de Brasília

DEPARTAMENTO DE ESTATÍSTICA

16 de janeiro de 2025

Lista 1

Prof. Drª. Terezinha Kessia de Assis Ribeiro

Modelos Lineares Generalizados

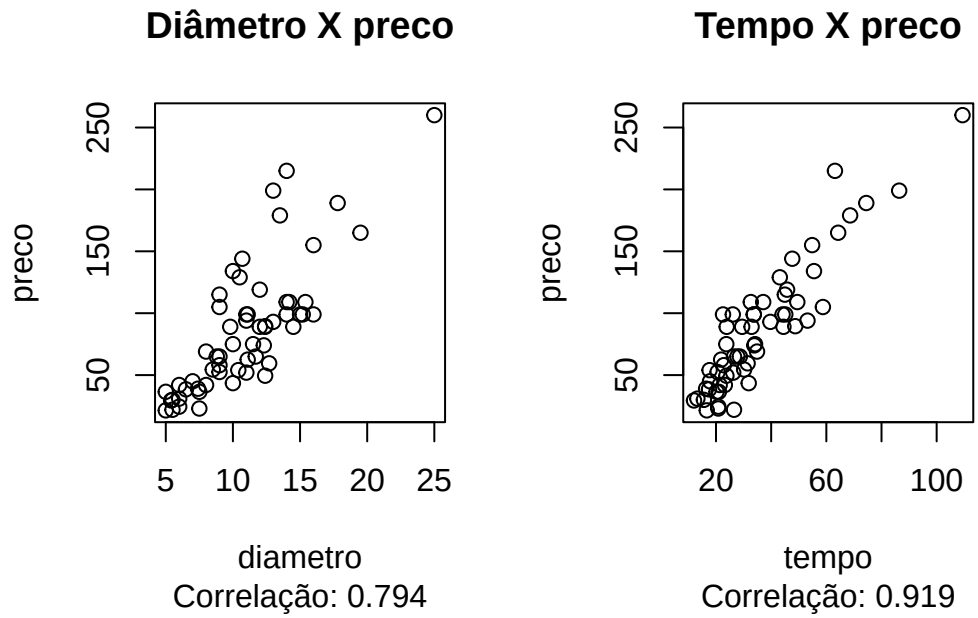
Aluno: Bruno Gondim Toledo | Matrícula: 15/0167636

1 Questão 1

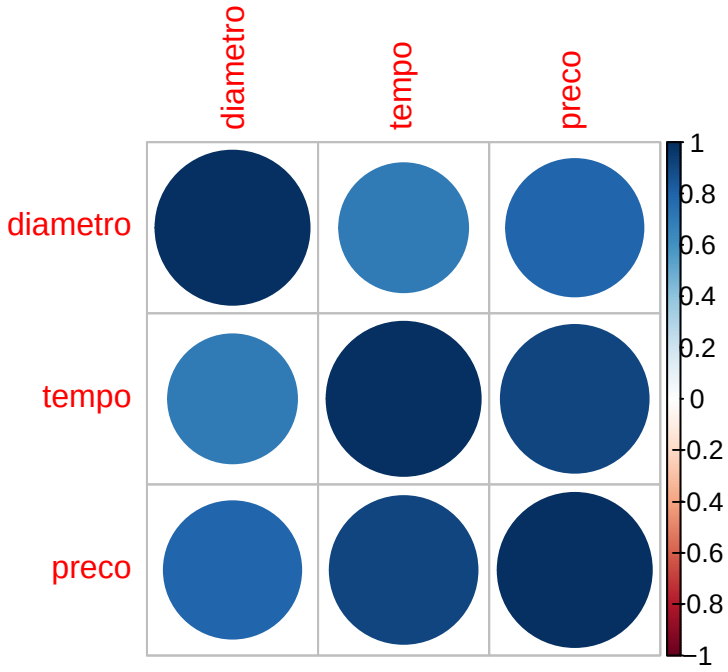
Para os dados da Nambe Mills, temos três variáveis quantitativas (diâmetro do produto em polegadas, tempo total de lixamento e polimento em minutos, preço em dólares do item) e uma variável categórica (tipo do item: cumbuca, cassarola, louça, prato e travessa). Realizar-se-á uma análise descritiva dessas variáveis, com foco na variável preço em dólares do item, que é o que se pretende responder a partir das demais covariáveis. Em seguida, será realizada uma modelagem afim de explicar o preço a partir das covariáveis pertinentes.

1.1 Correlação linear

Podemos verificar a correlação linear da variável preço ante as covariáveis diâmetro e tempo, para tentar inferir sobre a linearidade destas.



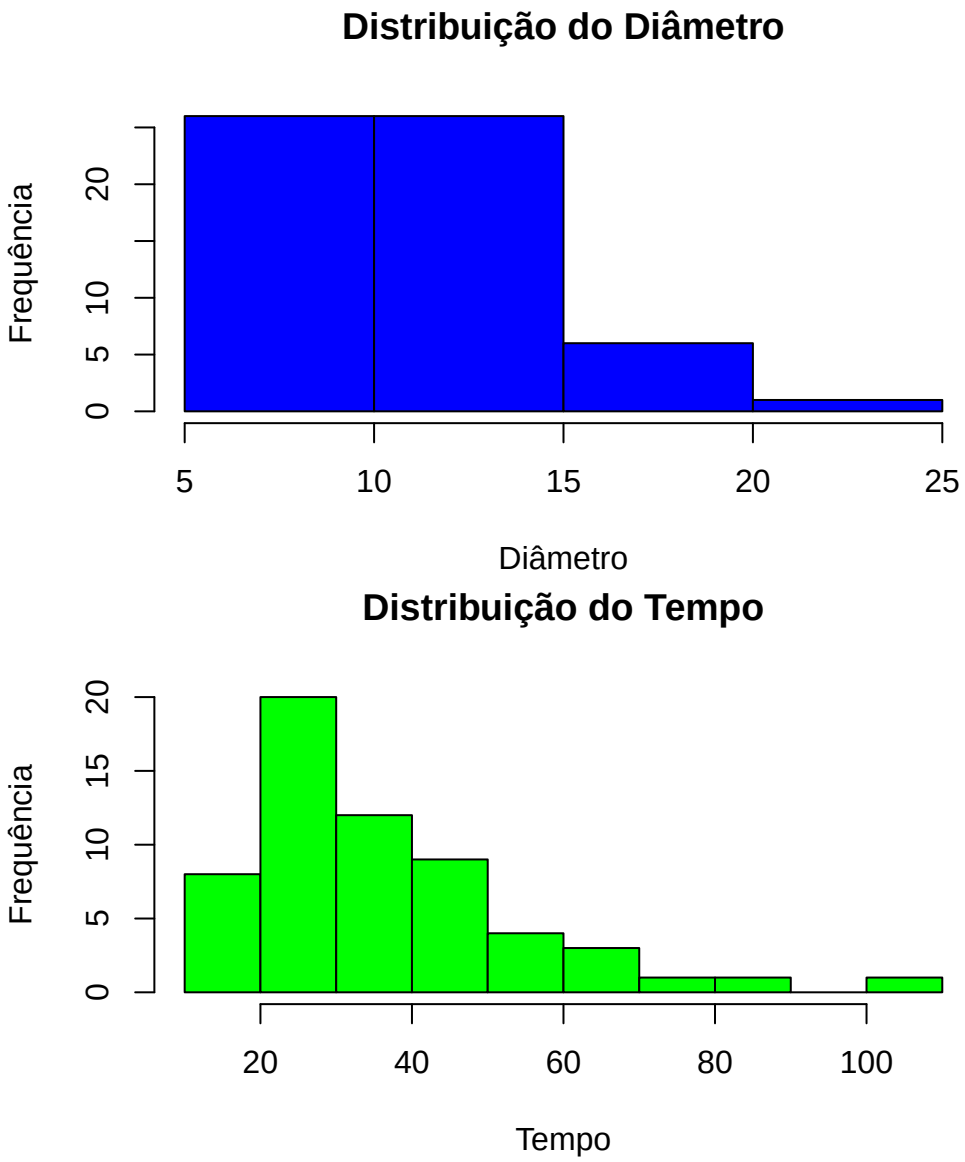
Pelas Figuras, observa-se correlação linear forte com ambas as covariáveis, principalmente com a covariável tempo.

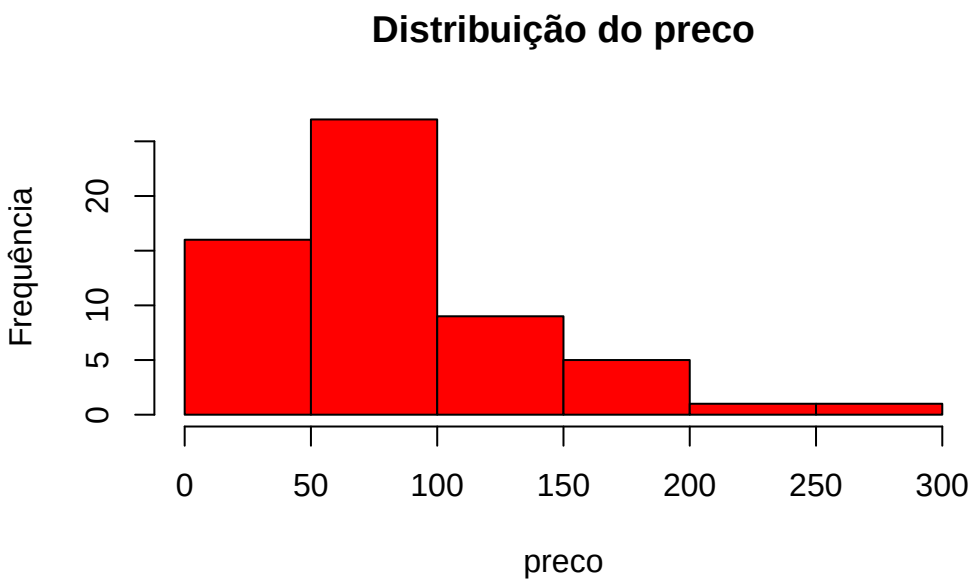


Verificando também correlações entre todas as variáveis e covariáveis, observamos que para todos os casos a correlação é forte. Isso pode levar a um problema de multicolinearidade quando da modelagem dessa variável.

1.2 Forma das distribuições empíricas

Podemos investigar a forma das distribuições empíricas das variáveis resposta e explicativas, já investigando o formato aproximado da distribuição dessas variáveis afim de, por exemplo, observar se a distribuição Gaussiana é adequada para descrever estas distribuições.

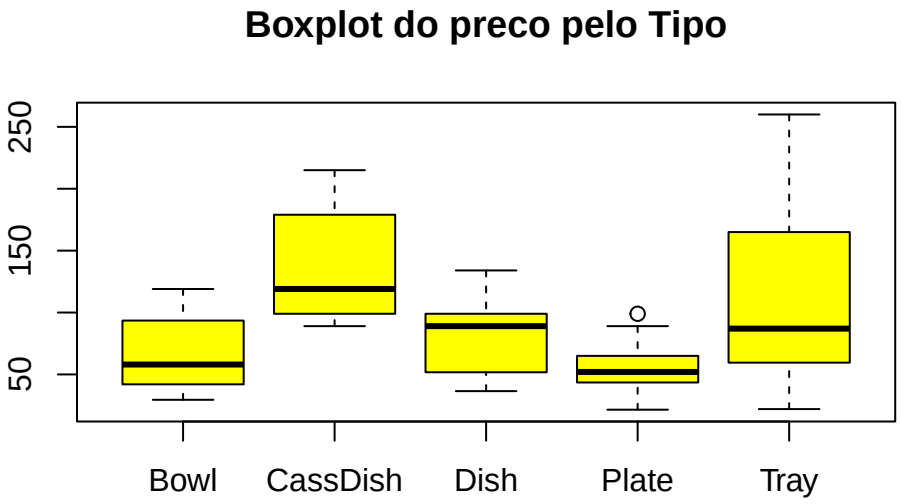




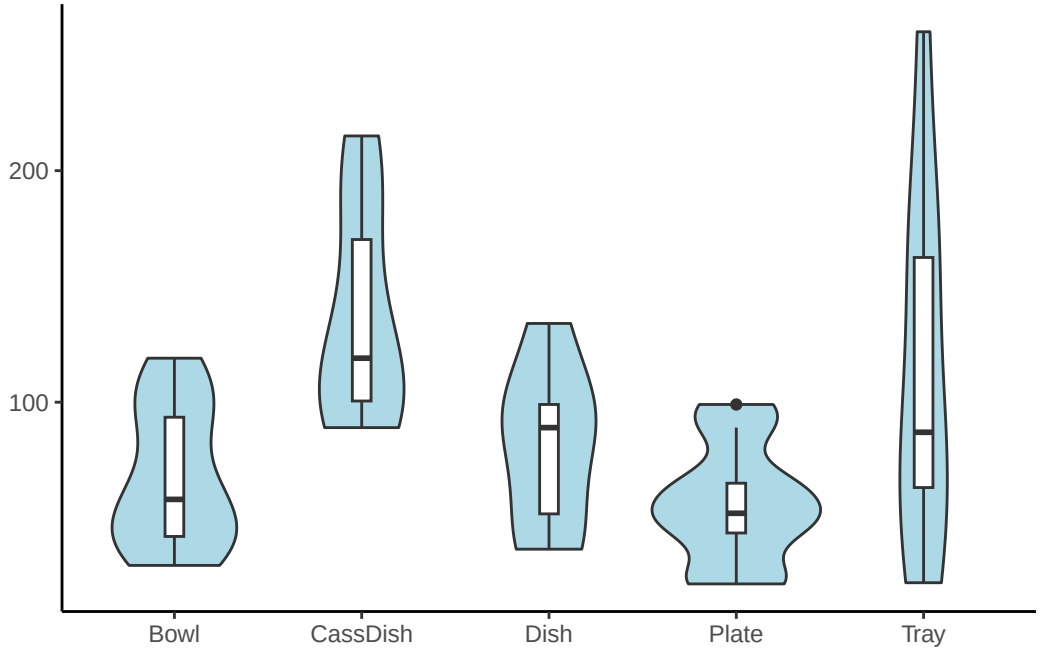
Vemos que para todas as variáveis, em especial a variável preço, o histograma das distribuições sugerem que estas são positivas e alongadas à direita. Este é um resultado que faria sentido sem nem precisar dos gráficos, visto que preço, tempo e diâmetro são grandezas que não se espera que assumam valores negativos, por exemplo. Estes histogramas sugerem que uma modelagem Gaussiana não seja a mais adequada, provavelmente. Podemos também calcular medidas-resumo, com destaque para nossa variável resposta — preço:

Mín.	1º Quartil	Mediana	Média	3º Quartil	Máx.	Assimetria	Curtose
21.5	47.25	75	86.38	107	260	1.18	4.34

Podemos observar um deslocamento a direita da média em relação a mediana, o que indica o alongamento a direita. Existe também uma grande amplitude entre o menor e o maior preço observado, e o coeficiente de assimetria confirma a análise visual do histograma sobre a assimetria a direita da variável. O coeficiente de curtose indica excesso de curtose, indicando também um possível problema de aderência à distribuição Gaussiana.



O boxplot do preço discriminado pelo tipo de produto indica que possivelmente teremos influência desta variável categórica na modelagem. Alguns produtos como prato e travessa apresentam medianas similares, o que indica que nem todos os fatores possivelmente serão considerados relevantes, porém produtos como cassarola e cumbuca apresentam mediana de preço significativamente diferente pelo boxplot, o que possivelmente será detectado pela modelagem.



O gráfico de violino pode ser utilizado também para visualização da variável preço, afim de observar a continuidade dos valores listados, auxiliando hipóteses formuladas sobre os histogramas e boxplot acerca do comportamento da variável, neste caso diferenciada pelos tipos de produtos.

1.3 Modelagem

Uma alta correlação entre as covariáveis numéricas com a variável resposta (e entre si) indicam que pode ser útil realizar a modelagem da variável preco em função das covariáveis dos dados. A praxis manda iniciar a tentativa com um modelo simples e parcimonioso: a regressão linear gaussiana; ainda que pelo diagnóstico preliminar a forma da distribuição empírica da variável preço indica que possivelmente teremos problemas com esta modelagem.

1.4 Modelo 1: Gaussiana com link identidade

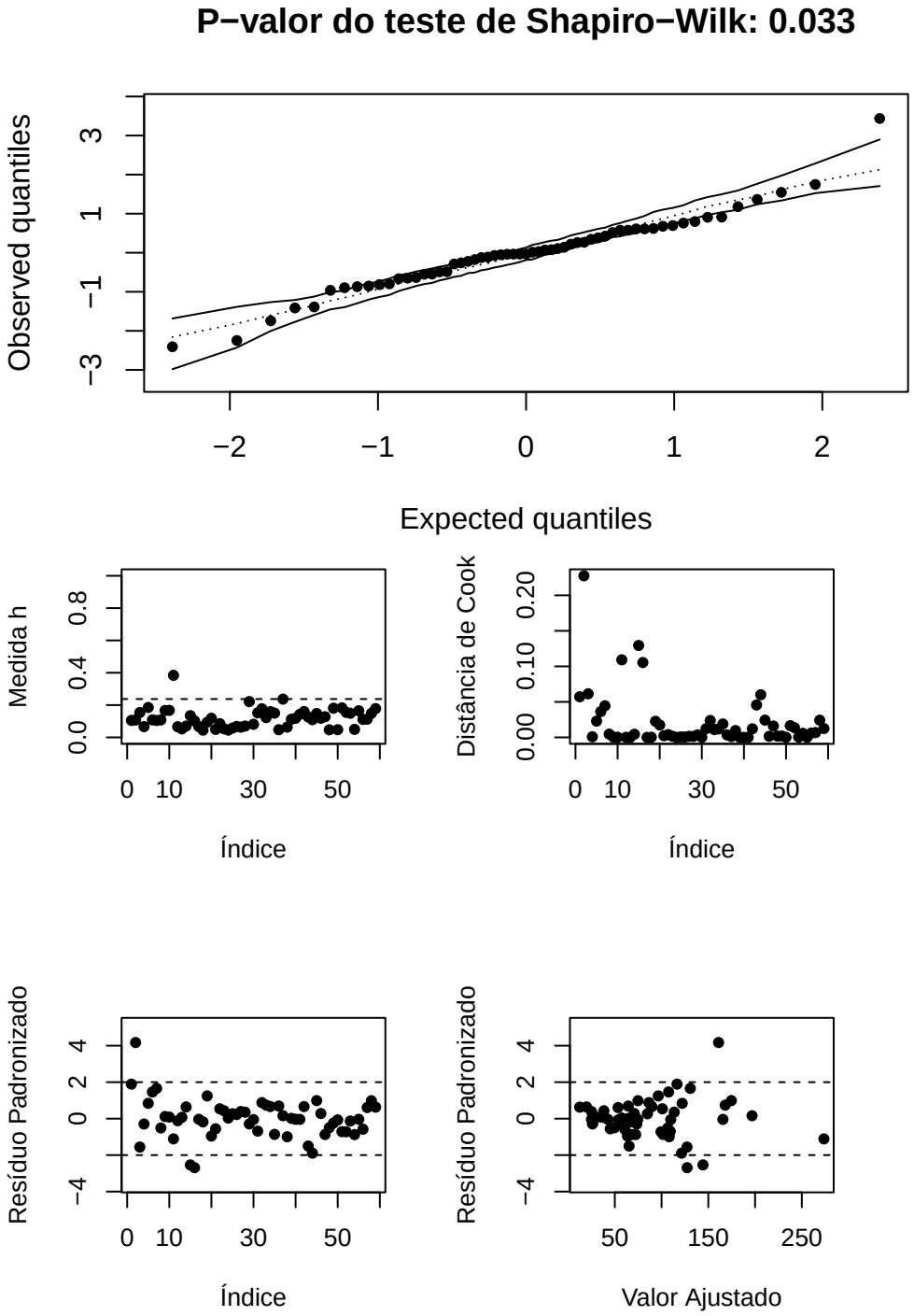
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-29.6037	6.7030	-4.4165	0.0001
diametro	5.2186	0.8282	6.3009	0.0000
tempo	1.7600	0.1834	9.5954	0.0000
tipoCassDish	6.4426	6.9775	0.9233	0.3601
tipoDish	1.7567	7.0127	0.2505	0.8032
tipoPlate	-13.3523	6.3764	-2.0940	0.0412
tipoTray	-19.7189	6.8135	-2.8941	0.0055

Pelas estimativas do modelo, podemos observar que tanto a variável diametro quanto a variável tempo foram significativas para explicar a variável preço. A variavel tipo também apresentou significância para alguns produtos: fixado o produto “Bowl” (cumbuca), observamos que para $\alpha = 0,05$ os produtos “Plate” (prato) e “Tray” (travessa) são siginificativos para explicar o preço. Para o modelo normal, é interessante observar os coeficientes R² e R² ajustado, que apresentaram valores 0.9167935 e 0.90719 respectivamente, indicando que o modelo explica bem os dados.

Podemos realizar o procedimento automático *stepwise* para seleção de covariáveis deste modelo, buscando assim talvez obter um modelo ainda melhor e/ou parcimonioso que o modelo saturado apresentado anteriormente.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-29.6037	6.7030	-4.4165	0.0001
diametro	5.2186	0.8282	6.3009	0.0000
tempo	1.7600	0.1834	9.5954	0.0000
tipoCassDish	6.4426	6.9775	0.9233	0.3601
tipoDish	1.7567	7.0127	0.2505	0.8032
tipoPlate	-13.3523	6.3764	-2.0940	0.0412
tipoTray	-19.7189	6.8135	-2.8941	0.0055

Pelos resultados do procedimento, nota-se que o modelo saturado é justamente o melhor modelo segundo esta metodologia, indicando que podemos proceder para os diagnósticos dos pressupostos deste modelo.



O modelo de regressão linear abarca uma série de pressupostos afim de tornar seus resultados verossímeis, sendo eles independência, normalidade e homocedasticidade dos resíduos. Pelo primeiro gráfico desta seção, do envelope simulado dos resíduos com o

teste de Shapiro-Wilk para normalidade, já verificamos uma fuga para um nível $\alpha = 95\%$ do pressuposto de normalidade. Demais pressupostos podem ser analisados, mas visto que já para o primeiro temos um problema, tentarei outro modelo afim de obter ajuste que cumpra pressupostos.

Conforme discutido em aula, os ajustes MLG perdem em explicabilidade e exatidão, mas fornecem modelos mais flexíveis ante a estes pressupostos, reduzindo o diagnóstico a análises mais flexíveis dos erros.

1.5 Modelo 2: Gaussiana com link log

Como aparentava termos obtido um bom ajuste com a regressão linear múltipla, é possível que introduzindo uma função de ligação log já seja possível contornar alguns dos problemas do diagnóstico do modelo.

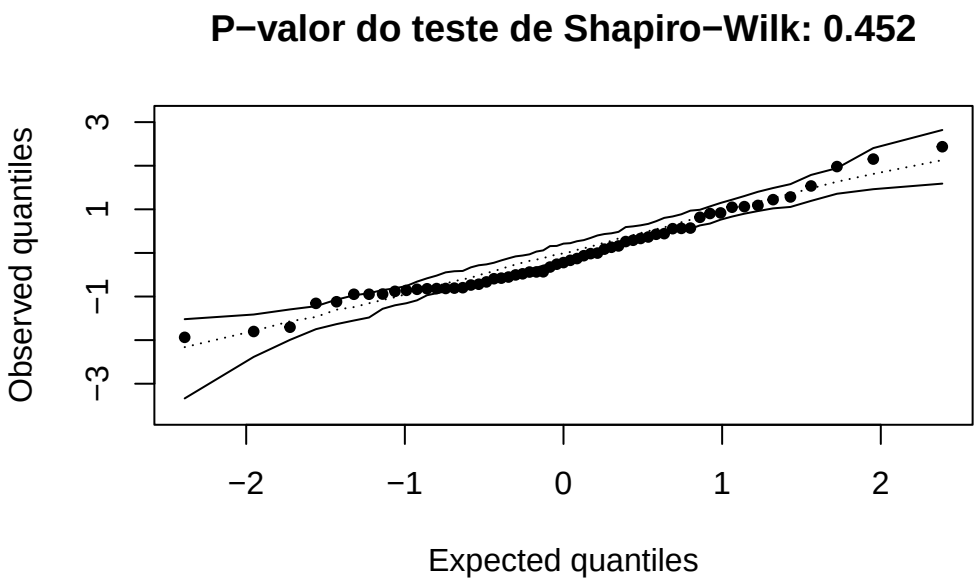
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.4597	0.1175	29.4375	0.0000
diametro	0.0498	0.0129	3.8668	0.0003
tempo	0.0108	0.0022	5.0283	0.0000
tipoCassDish	0.2419	0.0924	2.6177	0.0116
tipoDish	0.1139	0.1210	0.9410	0.3510
tipoPlate	-0.2108	0.1430	-1.4738	0.1466
tipoTray	-0.1741	0.1242	-1.4021	0.1668

Notamos que para o modelo GLM Gaussiano com ligação log, obtemos significância dos parâmetros bastante comparáveis com o modelo anterior, com a diferença da significância dos tipos de produtos que são pertinentes sob um valor $\alpha = 0,05$; neste caso apenas a cassarola sendo significativa ante ao produto de referência: a cumbuca. Dos coeficientes R^2 e R^2 ajustado, com valores 0.8384917 e 0.81532 respectivamente, pode-se dizer que este modelo aparenta performar bem sobre os dados

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.4597	0.1175	29.4375	0.0000
diametro	0.0498	0.0129	3.8668	0.0003
tempo	0.0108	0.0022	5.0283	0.0000
tipoCassDish	0.2419	0.0924	2.6177	0.0116
tipoDish	0.1139	0.1210	0.9410	0.3510
tipoPlate	-0.2108	0.1430	-1.4738	0.1466
tipoTray	-0.1741	0.1242	-1.4021	0.1668

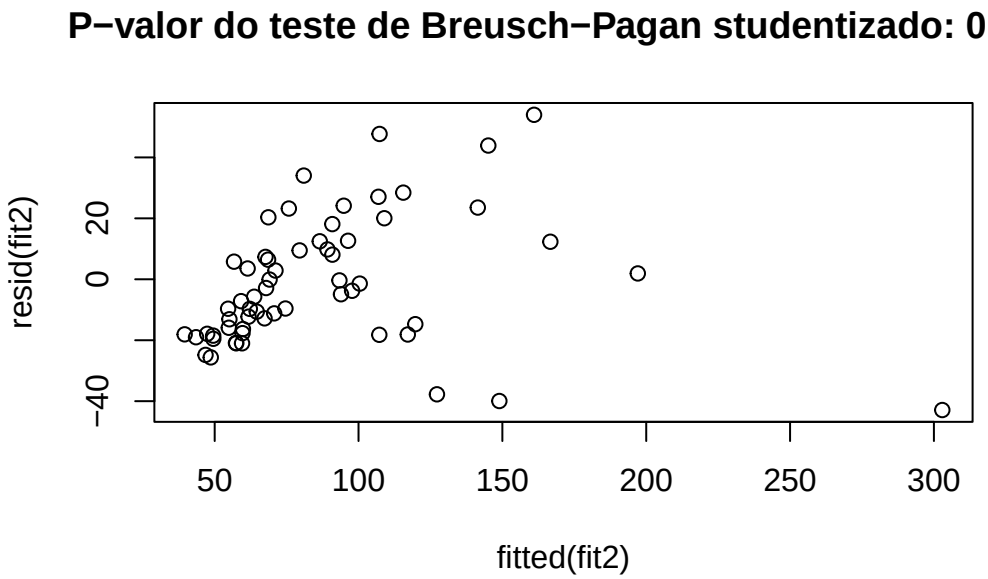
O Procedimento *Stepwise* para este modelo também chega a uma conclusão análoga à anterior, indicando que o modelo saturado é o melhor modelo, permitindo assim partir para a análise de diagnósticos

1.5.1 Normalidade



O pressuposto de normalidade aparenta ser atendido para este modelo.

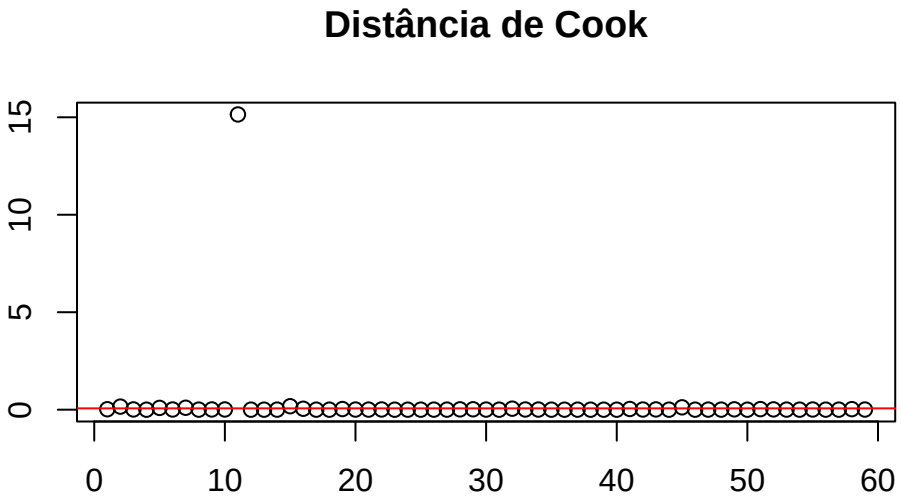
1.5.2 Homocedasticidade



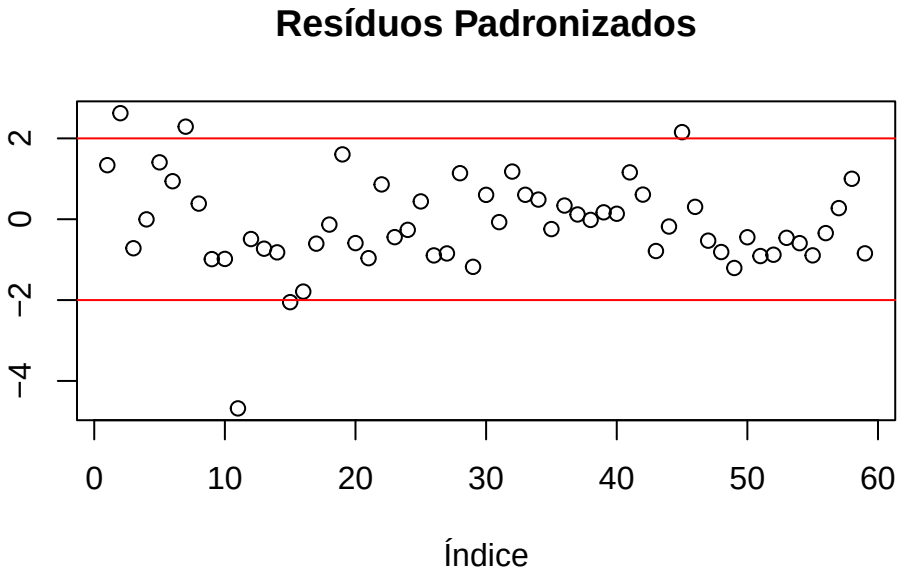
O gráfico dos erros contra os valores ajustados, além do teste de Breusch-Pagan, indicam que este é um modelo heterocedástico. Para os MLGs, em geral isto não é um problema, visto que este tipo de modelagem permite a heterocedasticidade dos dados. Porém, para a família Gaussiana, isso deve ser observado com cautela.

1.5.3 Influência

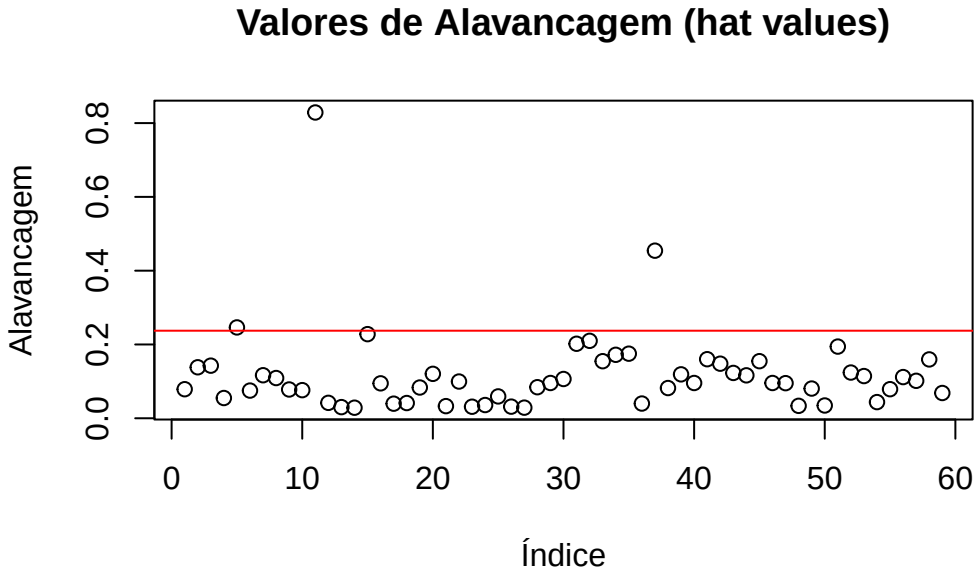
Observaremos alguns dos gráficos habitualmente utilizados para análise de pontos influentes



Pelo gráfico da distância de cook, notamos um valor extremamente, que é justamente o ponto 11: Uma travessa com preço 260: o maior valor observado da variável! Também é o produto de maior diâmetro, e maior tempo de polimento, portanto possivelmente trata-se de uma peça diferenciada das demais mesmo.



Pelo gráfico dos resíduos padronizados, notamos que a maior parte das observações se encontram a até 2 desvios do zero, com um outlier marcante: o ponto 11 novamente. Este é um ponto que possivelmente será um outlier em qualquer modelagem que for executada. Os pontos 2, 5, 7, 15 e 45 são os outros que também estão fora do valor típico de referência |2|.



Para análise de alavancagem, observamos novamente o ponto 2 com alguma alavancagem, o ponto 37 com alavancagem significativa, e principalmente o ponto 11 novamente, com a maior alavancagem individual.

1.5.4 Conclusão

A modelagem normal com ligação identidade já se mostrou util para descrever estes dados, ainda que com problemas em pressupostos. A utilização da ligação log aparenta ter melhorado este problemas de pressupostos, entretanto ao custo de perda de interpretabilidade. A análise de influência mostra que alguns pontos (em especial o ponto 11) possivelmente melhoraria ainda mais a eficiência desta modelagem. Como a questão solicita a utilização de outros modelos, não faz sentido uma análise tão aprofundada dos resultados obtidos até aqui. Mas a termos práticos, a família Gaussiana, ainda que não seja a mais adequada por conta da forma da distribuição empírica da variável resposta, já seria suficiente para obter estimativas úteis acerca do preço baseado nas covariáveis tipo, diâmetro e tempo.

Irei agora propor modelagem GLM com família Gama, utilizando funções de ligação pertinentes para esta família afim de obter um bom ajuste para estes dados.

1.6 Modelo 3: Gama com link identidade

A distribuição gama é muito mais flexível que a distribuição normal, e a forma da distribuição empírica da variável preço se assemelha visualmente a uma distribuição gama com parâmetros $k = \theta = 2$.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-22.1457	3.5253	-6.2819	0.0000

	Estimate	Std. Error	t value	Pr(>|t|)
diametro	4.9811	0.5416	9.1975	0.0000
tempo	1.6042	0.1820	8.8128	0.0000
tipoCassDish	8.0129	7.7931	1.0282	0.3086
tipoDish	0.6000	4.1518	0.1445	0.8857
tipoPlate	-11.8241	2.7140	-4.3566	0.0001
tipoTray	-24.2059	3.2266	-7.5020	0.0000

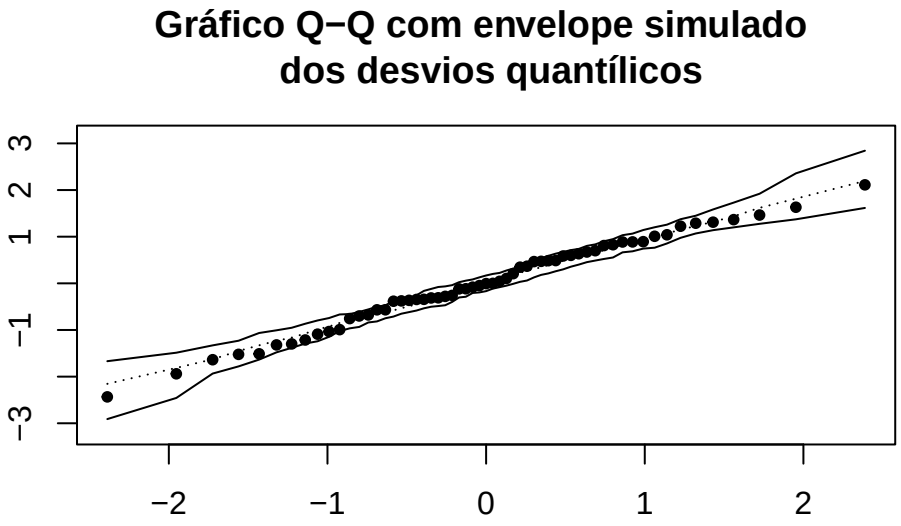
Para o modelo GLM gama com ligação identidade, notamos que a significância dos parâmetros é análoga a significância observada para a modelagem gaussiana com ligação identidade, sendo tempo e diâmetro significativas, e algumas categorias da variável tipo significativas também ante ao fator de referência sob um $\alpha = 0,05$ fixado.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-22.1457	3.5253	-6.2819	0.0000
diametro	4.9811	0.5416	9.1975	0.0000
tempo	1.6042	0.1820	8.8128	0.0000
tipoCassDish	8.0129	7.7931	1.0282	0.3086
tipoDish	0.6000	4.1518	0.1445	0.8857
tipoPlate	-11.8241	2.7140	-4.3566	0.0001
tipoTray	-24.2059	3.2266	-7.5020	0.0000

A abordagem *stepwise* indica para esta família que também devemos ficar com o modelo saturado.

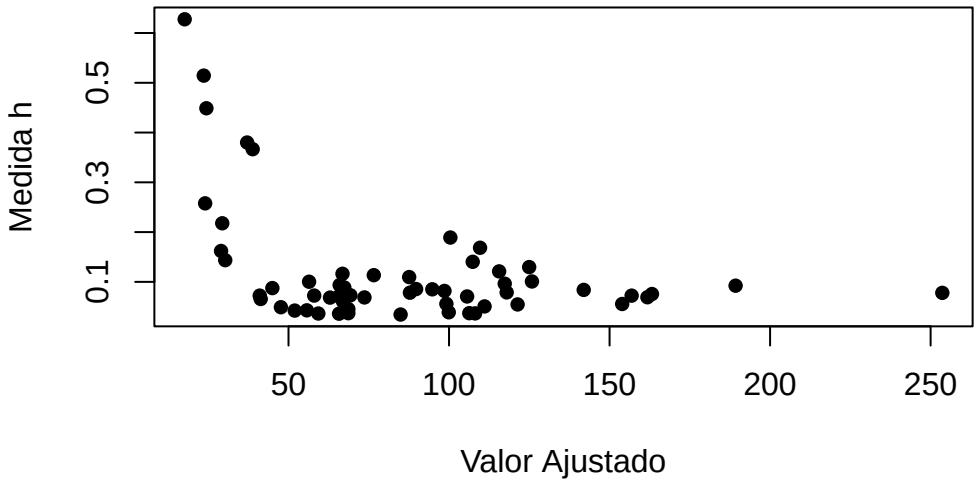
1.6.1 Diagnósticos

1.6.1.1 Envelope simulado



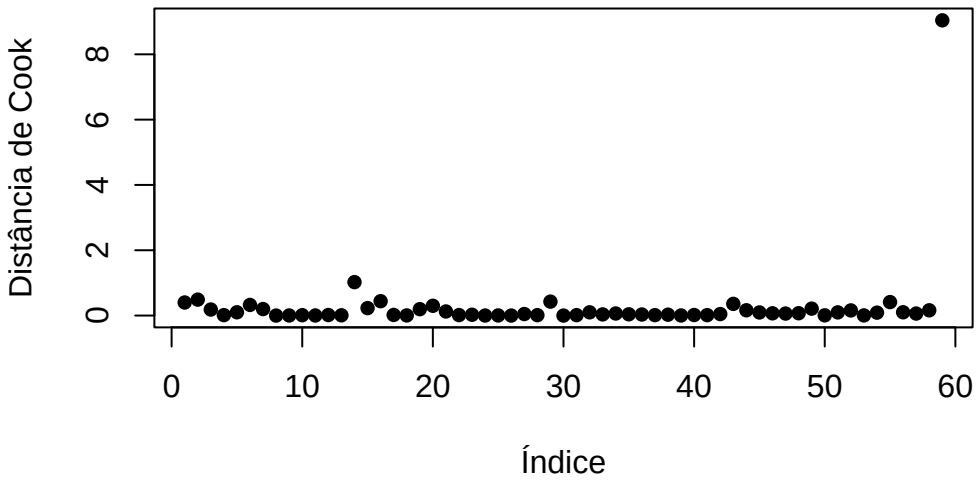
Realizando o envelope simulado dos desvios quantílicos sobre o erro do modelo, notamos que não aparenta haver grande fuga dos erros aos limites produzido pelo envelope simulado, indicando que o modelo se adequa bem aos dados

1.6.1.2 Alavancagem



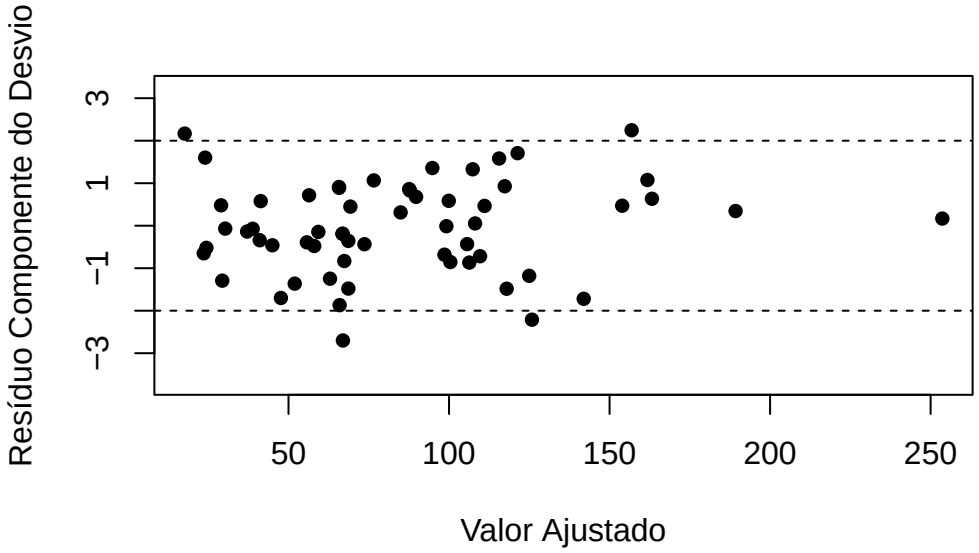
Da Figura acima, notamos que existem alguns pontos de alavancagem, ainda que aparentemente sob uma magnitude inferior ao comparado no modelo Gaussiano. O ponto 11 segue sendo o principal ponto de alavancagem.

1.6.1.3 Distância de Cook



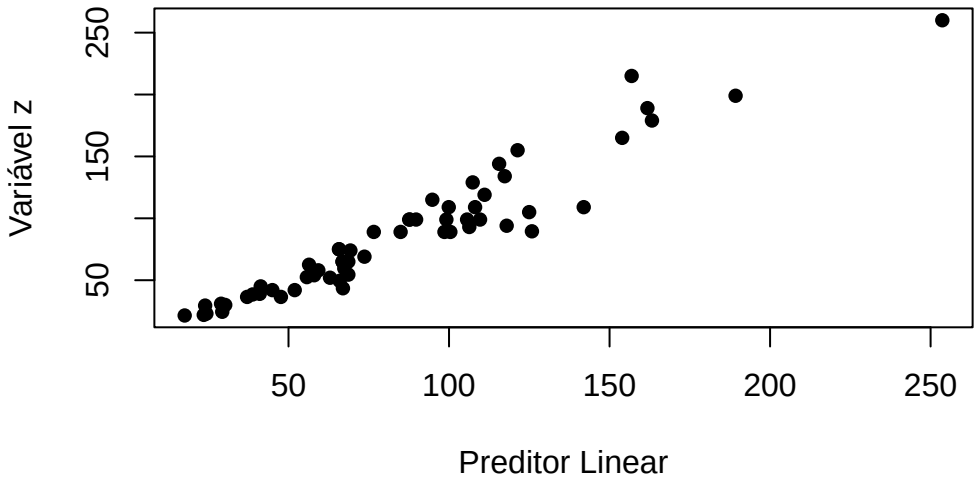
O gráfico da distância de Cook mostra um resultado interessante. Ainda que o ponto 11 seja ainda considerado um ponto influente, com medida de cook alta, o ponto com maior medida de Cook é o ponto 59: o último ponto dos dados, que contém o menor valor na variável preço: 21,5; sendo este um prato com também o menor diâmetro do conjunto de dados.

1.6.1.4 Independência



O gráfico do resíduo componente do desvio contra os valores ajustados mostra que grande parte dos pontos se adequa aos limites espedados, com apenas 4 pontos ultrapassando a margem usualmente utilizada, ainda que sem grande fuga da referência. Não aparenta haver um padrão claro nos dados, sugerindo independência.

1.6.1.5 Qualidade do ajuste



Observando a Figura acima, o modelo aparenta ter obtido um bom ajuste, em especial para os pontos de maior e menor valor, porém com certa fuga nos pontos centrados ao redor da média, indicando que é um modelo útil, mas que pode ser tentado algum modelo mais sofisticado

1.7 Modelo 4: gama com link inversa

A ligação canônica do GLM gama é a inversa, portanto irei testar também a modelagem com esta função de ligação.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0242	0.0019	12.5998	0.0000
diametro	-0.0008	0.0002	-3.7823	0.0004
tempo	-0.0001	0.0000	-1.4808	0.1447
tipoCassDish	-0.0038	0.0015	-2.4769	0.0165
tipoDish	-0.0025	0.0020	-1.2199	0.2280
tipoPlate	0.0038	0.0024	1.6005	0.1156
tipoTray	0.0035	0.0023	1.5376	0.1302

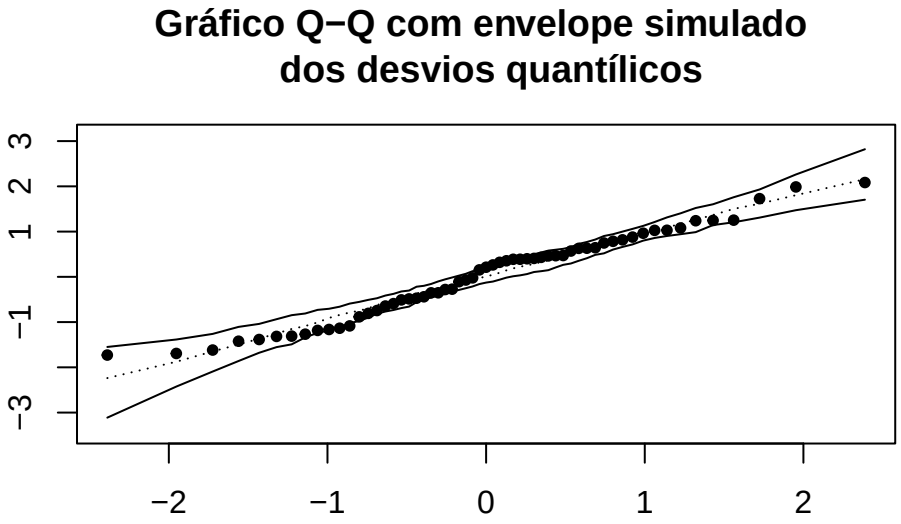
Para este modelo, as covariáveis tempo e diâmetro seguem sendo extremamente significativas sob qualquer nível de significância, enquanto que para a covariável tipo, apenas a categoria cassarola apresenta significância ante a categoria de referência; a cumbuca.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0242	0.0019	12.5998	0.0000
diametro	-0.0008	0.0002	-3.7823	0.0004
tempo	-0.0001	0.0000	-1.4808	0.1447
tipoCassDish	-0.0038	0.0015	-2.4769	0.0165
tipoDish	-0.0025	0.0020	-1.2199	0.2280
tipoPlate	0.0038	0.0024	1.6005	0.1156
tipoTray	0.0035	0.0023	1.5376	0.1302

O procedimento *stepwise* indica continuar com todas as covariáveis no modelo. Ainda que a significância de algumas categorias tenha diminuido, covariável ainda é percebida como útil e parcimoniosa para bom ajuste deste modelo.

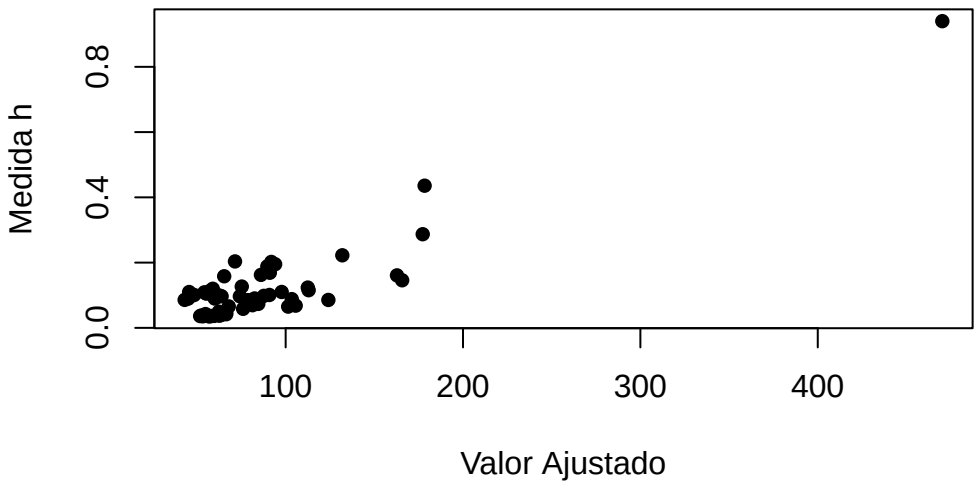
1.7.1 Diagnósticos

1.7.1.1 Envelope simulado



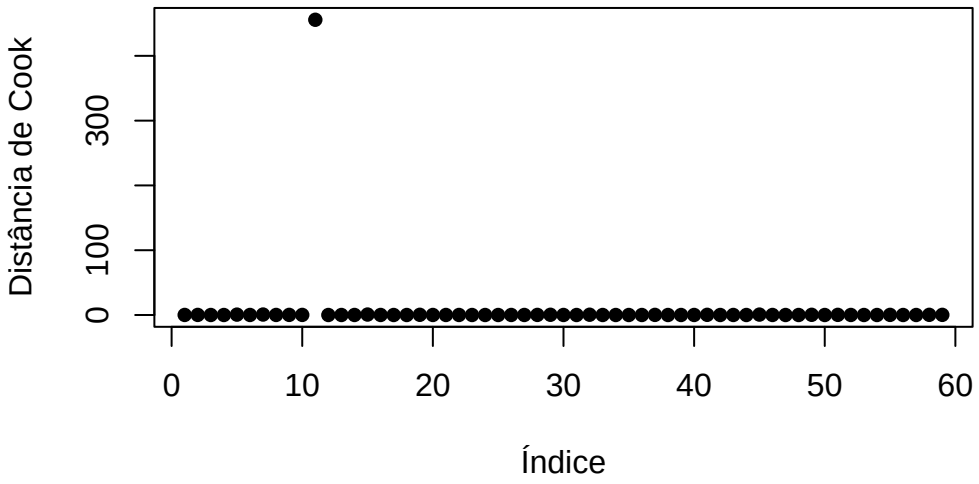
Realizando o envelope simulado dos desvios quantílicos sobre o erro do modelo, notamos que existe uma fuga dos valores esperados em torno do zero, o que pode comprometer a qualidade do ajuste deste modelo.

1.7.1.2 Alavancagem



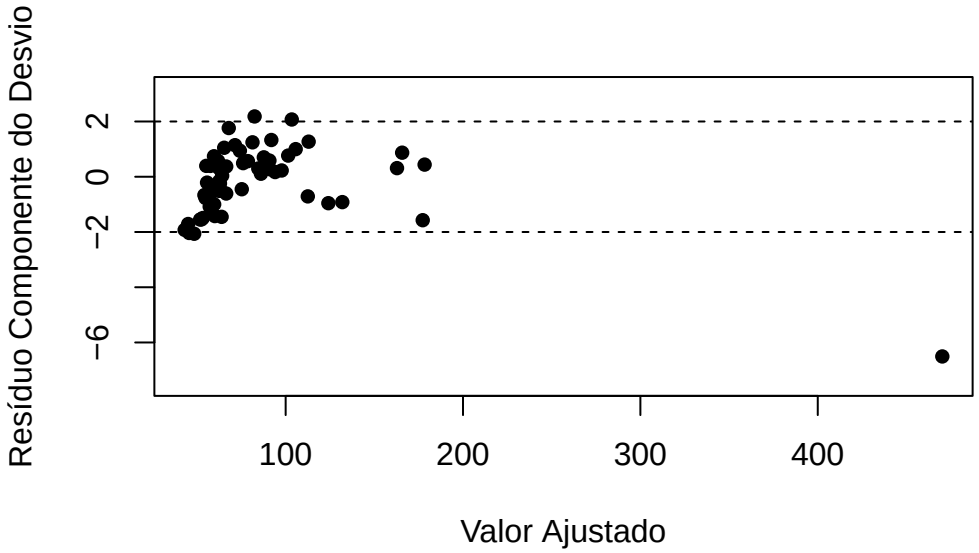
Da Figura acima, notamos que existem pontos com alavancagem significativamente alta, incluindo um outlier bastante significativo — sendo este justamente o ponto 11, de maior valor na variável preço.

1.7.1.3 Distância de Cook



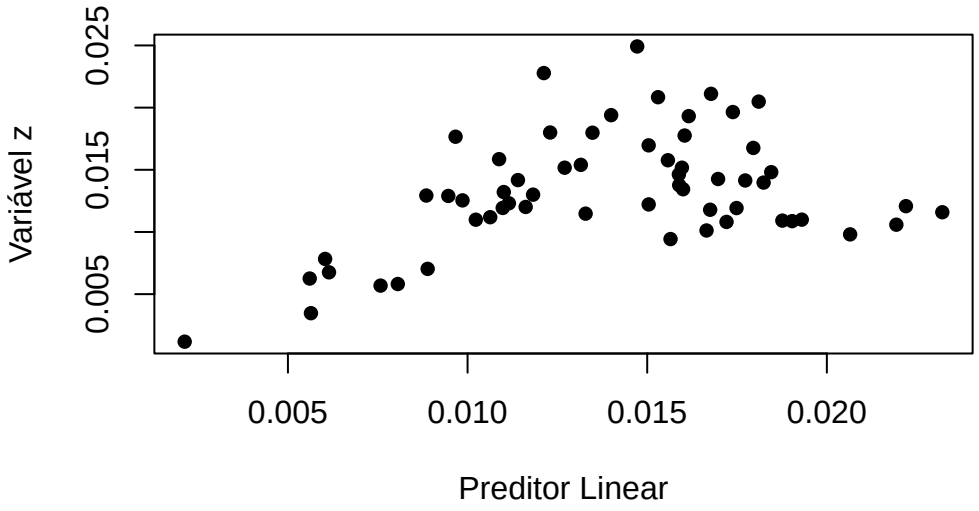
O gráfico da distância de Cook mostra um resultado análogo ao observado nos modelos gaussianos, porém em uma magnitude ainda maior: o ponto 11 é extremamente influente.

1.7.1.4 Independência



O gráfico do resíduo componente do desvio contra os valores ajustados mostra que, ainda que a maioria dos pontos esteja dentro do esperado, existem alguns valores fora e um valor extremamente aberrante: o ponto 11.

1.7.1.5 Qualidade do ajuste



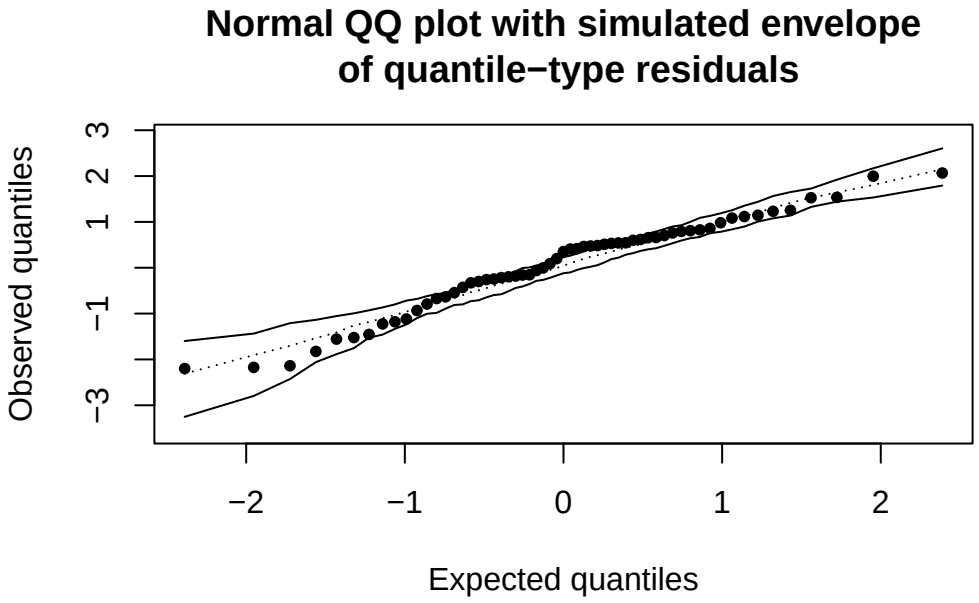
Observando a Figura acima, o modelo aparenta ter obtido um ajuste prejudicado em relação aos demais. Possivelmente o ponto 11 foi o maior responsável por este resultado, por tudo observado nos diagnósticos anteriores.

1.8 Modelo 5: Normal inversa com link canônico

Será testada também a modelagem GLM utilizando a distribuição Normal inversa com algumas funções de ligação. Inicialmente será utilizada a função canônica para esta distribuição, $\frac{1}{\mu^2}$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4e-04	1e-04	6.3801	0.0000
diametro	0e+00	0e+00	-2.8275	0.0066
tempo	0e+00	0e+00	-0.6921	0.4920
tipoCassDish	-1e-04	0e+00	-2.0041	0.0503
tipoDish	-1e-04	1e-04	-1.0557	0.2960
tipoPlate	1e-04	1e-04	1.2780	0.2069
tipoTray	1e-04	1e-04	1.0695	0.2898

Os resultados desta modelagem devem ser observados com calma. Este foi o primeiro modelo que indica a variável tempo como não significativa, possivelmente por problemas de multicolinearidade antecipados pela análise descritiva. Também este modelo não encontra significância em nenhuma categoria da variável tipo ante a categoria de referência, restando apenas o intercepto e a variável diâmetro. Estes resultados **NÃO** são confiáveis, visto que o modelo apresentou problemas de convergência quando executada a modelagem no *software* R.



A abordagem *stepwise* não funcionou, e o gráfico dos resíduos quantílicos com envelope simulado também apresentou problemas de convergência.

Por todos estes problemas, este modelo não é possível de analisar mais profundamente.

1.9 Modelo 6: Normal inversa com link identidade

Dado todos os problemas com a ligação canônica, podemos dar um passo atrás e testar a modelagem com a normal inversa utilizando a função de ligação identidade

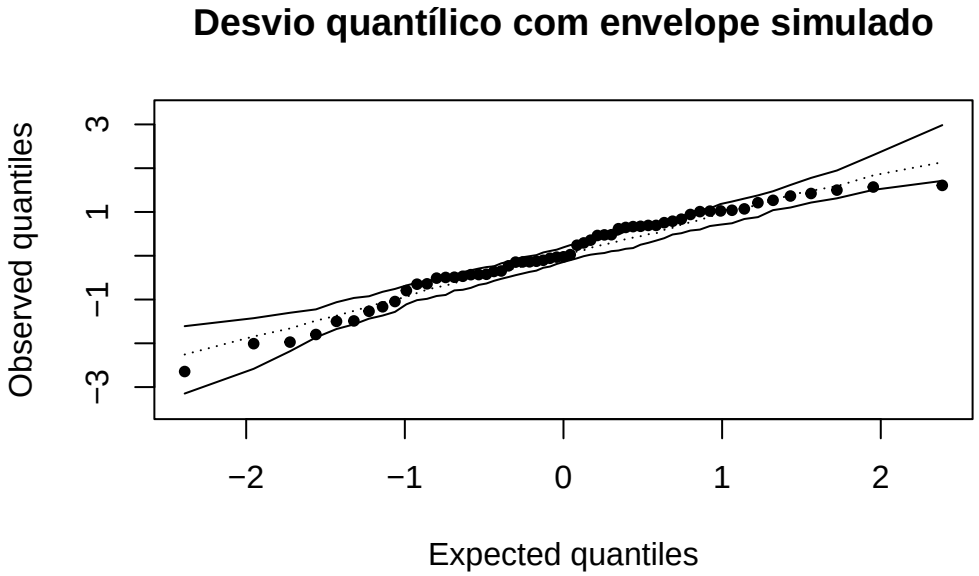
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-18.1020	2.9593	-6.1170	0.0000
diametro	4.9128	0.4615	10.6453	0.0000
tempo	1.4358	0.1696	8.4641	0.0000
tipoCassDish	12.2935	9.2822	1.3244	0.1912
tipoDish	0.5117	3.2710	0.1564	0.8763
tipoPlate	-11.1100	1.9112	-5.8131	0.0000
tipoTray	-24.6332	2.3000	-10.7100	0.0000

As estimativas produzidas pelo modelo são mais comparáveis com todos os modelos testados até então. As variáveis diâmetro e tempo são acusadas como bastante significativas, e duas categorias da variável tipo são acusadas como significativa ante a categoria de referência, sendo elas o prato e a travessa.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-18.1020	2.9593	-6.1170	0.0000

	Estimate	Std. Error	t value	Pr(>|t|)
diametro	4.9128	0.4615	10.6453	0.0000
tempo	1.4358	0.1696	8.4641	0.0000
tipoCassDish	12.2935	9.2822	1.3244	0.1912
tipoDish	0.5117	3.2710	0.1564	0.8763
tipoPlate	-11.1100	1.9112	-5.8131	0.0000
tipoTray	-24.6332	2.3000	-10.7100	0.0000

Neste caso também, o procedimento *stepwise* recomenda a manutenção do modelo saturado para melhor ajuste do modelo.



Analisando os desvios quantílicos do modelo com envelope simulado, notamos que existem diversos valores centrais ultrapassando o limite do envelope. Ainda que a magnitude deste desvio não seja aparentemente tão grande, isto é um significativo pertinente que este modelo não obteve boa qualidade de ajuste comparado com todos os modelos testados até agora.

1.10 Modelo 7: Normal inversa com link log

Apesar dos dois últimos resultados, é possível que a distribuição normal inversa seja útil para modelar estes dados, bastando apenas escolher uma função de ligação melhor para este caso. Testarei a função de ligação log

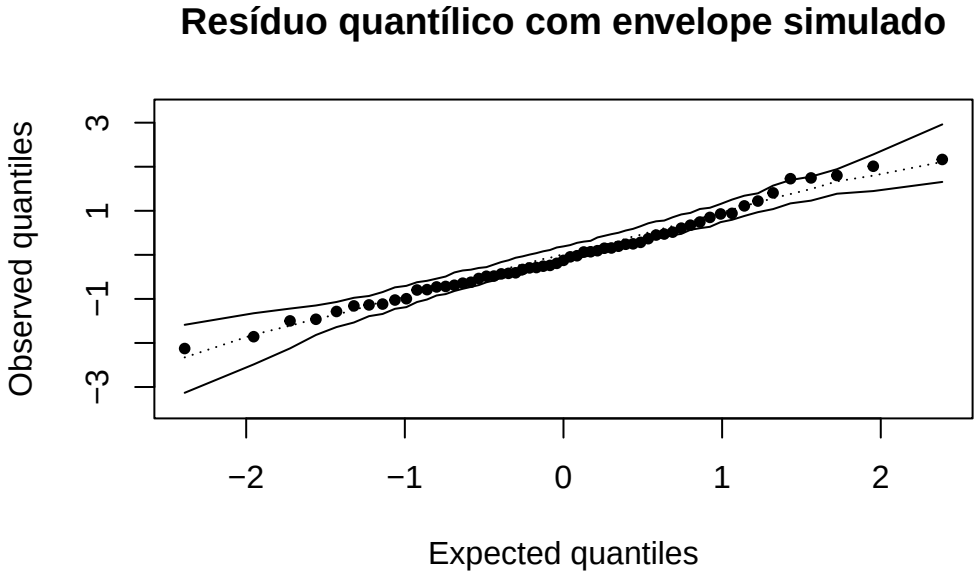
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.5234	0.0667	37.8220	0.0000
diametro	0.1186	0.0088	13.4912	0.0000

	Estimate	Std. Error	t value	Pr(>|t|)
tempo	0.0180	0.0026	7.0296	0.0000
tipoCassDish	-0.0127	0.0957	-0.1331	0.8946
tipoDish	0.0808	0.0714	1.1317	0.2630
tipoPlate	-0.3318	0.0557	-5.9563	0.0000
tipoTray	-0.4944	0.0625	-7.9111	0.0000

Os parâmetros acusados como significativos para este modelo são análogos aos encontrados anteriormente, com as covariáveis tempo e diâmetro apresentando alta significância, e algumas categorias da variável tipo apresentando significância ante a categoria de referência.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.5234	0.0667	37.8220	0.0000
diametro	0.1186	0.0088	13.4912	0.0000
tempo	0.0180	0.0026	7.0296	0.0000
tipoCassDish	-0.0127	0.0957	-0.1331	0.8946
tipoDish	0.0808	0.0714	1.1317	0.2630
tipoPlate	-0.3318	0.0557	-5.9563	0.0000
tipoTray	-0.4944	0.0625	-7.9111	0.0000

O procedimento *stepwise* indica o mesmo observado para os outros modelos — isto é, indica pela opção ao modelo saturado.



Diferente do observado anteriormente, o gráfico dos resíduos quantílicos com envelope simulado para este modelo não apresenta grande fuga dos valores observados, com ape-

nas um ou dois pontos ultrapassando o limite estabelecido pelo envelope. Isso indica que este é um dos modelos que também pode ser escolhido para realizar esta modelagem.

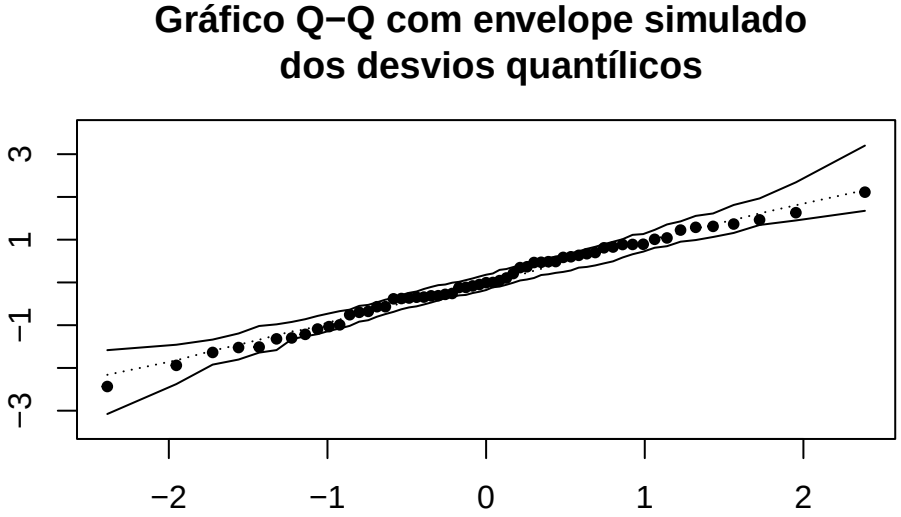
1.11 Modelo final

Diversas famílias de distribuições e funções de ligação foram testadas afim de escolher aquela que obtive melhor ajuste, e também parcimônia do modelo e interpretação. Desta forma, acredito que uma boa escolha para este conjunto seja o modelo 3, isto é, o MLG Gama com ligação identidade, pois forneceu bom ajuste, e com qualidade de ajuste aceitável. Levando em consideração também a análise exploratória realizada anteriormente à modelagem, podemos observar que a variável resposta é positiva e assimétrica, uma característica melhor suportada pelo modelo gama do que pelo modelo gaussiano.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-22.1457	3.5253	-6.2819	0.0000
diametro	4.9811	0.5416	9.1975	0.0000
tempo	1.6042	0.1820	8.8128	0.0000
tipoCassDish	8.0129	7.7931	1.0282	0.3086
tipoDish	0.6000	4.1518	0.1445	0.8857
tipoPlate	-11.8241	2.7140	-4.3566	0.0001
tipoTray	-24.2059	3.2266	-7.5020	0.0000

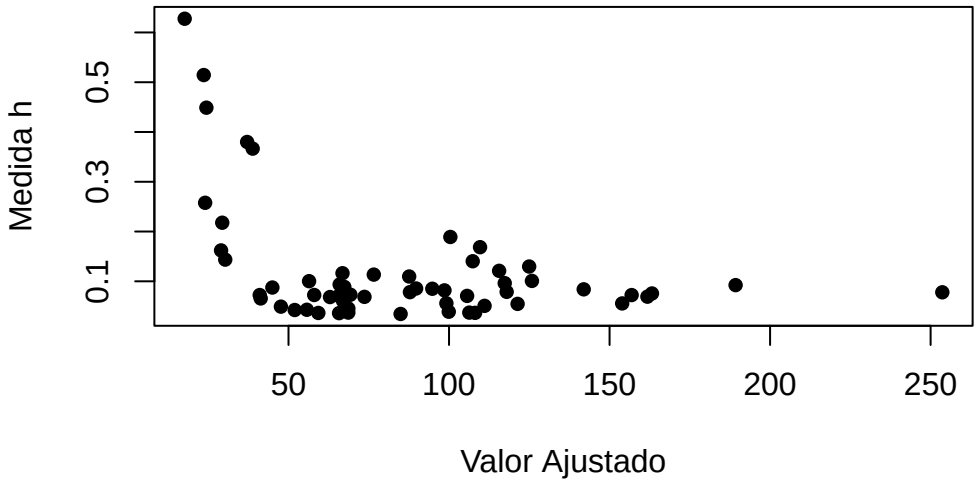
1.11.1 Diagnósticos

1.11.1.1 Envelope simulado



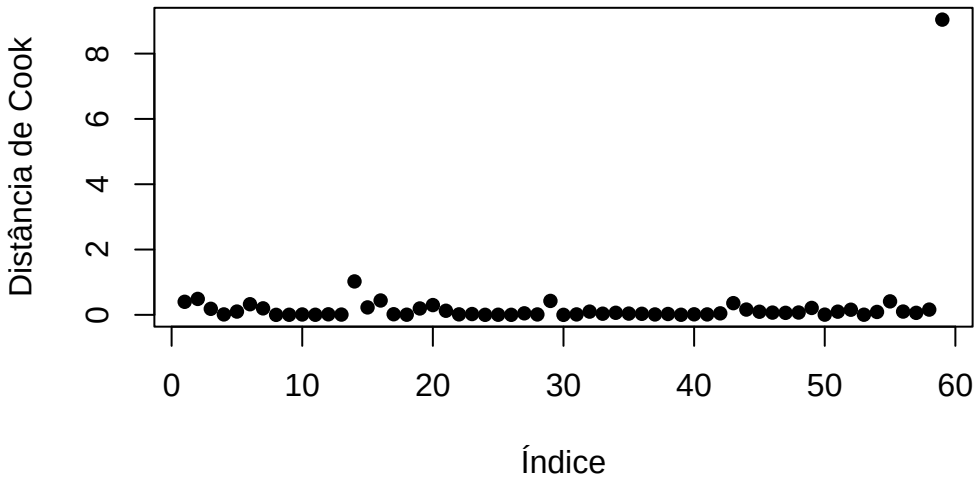
Realizando o envelope simulado dos desvios quantílicos sobre o erro do modelo, notamos que não aparenta haver grande fuga dos erros aos limites produzido pelo envelope simulado, indicando que o modelo se adequa bem aos dados

1.11.1.2 Alavancagem



Da Figura acima, notamos que existem alguns pontos de alavancagem. Os pontos 29 e 59 são os principais pontos de alavancagem.

1.11.1.3 Distância de Cook



O gráfico da distância de Cook mostra que os pontos 14 e 59 são de maior influência, com medida de cook alta > 1, sendo o ponto com maior medida de Cook o 59, com valor maior do que 9: o último ponto dos dados, que contém o menor valor na variável preço: 21,5; sendo este um prato com também o menor diâmetro do conjunto de dados.

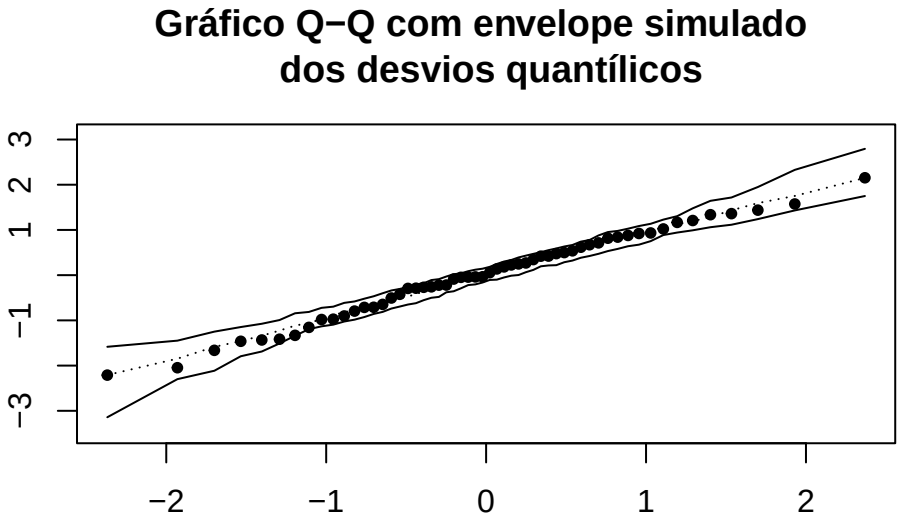
Dito isso, irei realizar o ajuste removendo os pontos 14, 29 e 59.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-27.6730	4.0039	-6.9114	0.0000
diametro	5.5033	0.6034	9.1199	0.0000
tempo	1.6182	0.1826	8.8598	0.0000

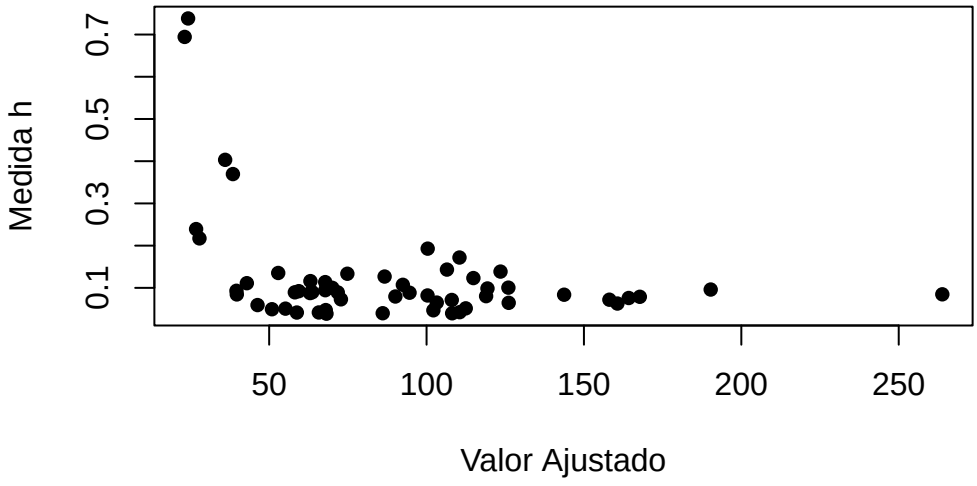
	Estimate	Std. Error	t value	Pr(>|t|)
tipoCassDish	6.5534	7.4783	0.8763	0.3851
tipoDish	2.1290	3.9754	0.5355	0.5947
tipoPlate	-15.9277	3.1553	-5.0479	0.0000
tipoTray	-23.0624	3.6432	-6.3302	0.0000

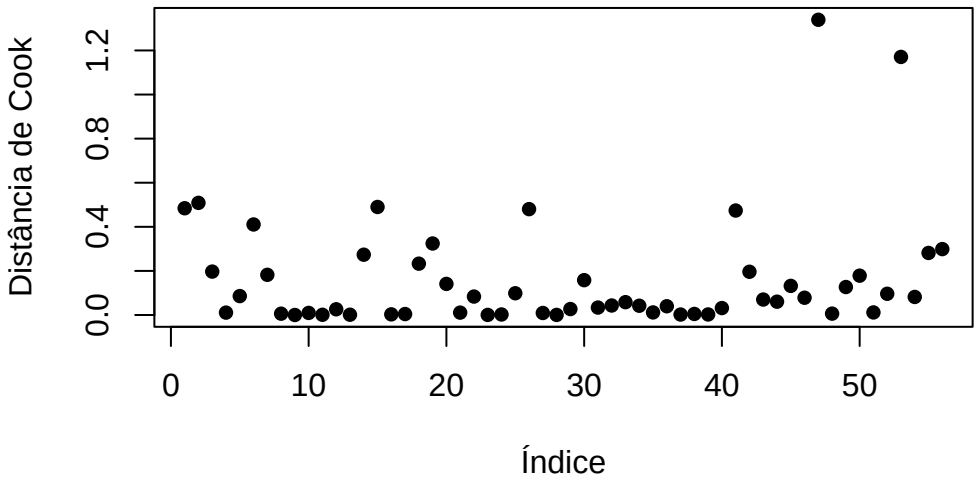
Podemos ver uma leve diferença nos parâmetros com a remoção destes pontos.

Realizando o diagnóstico sem estes 3 pontos



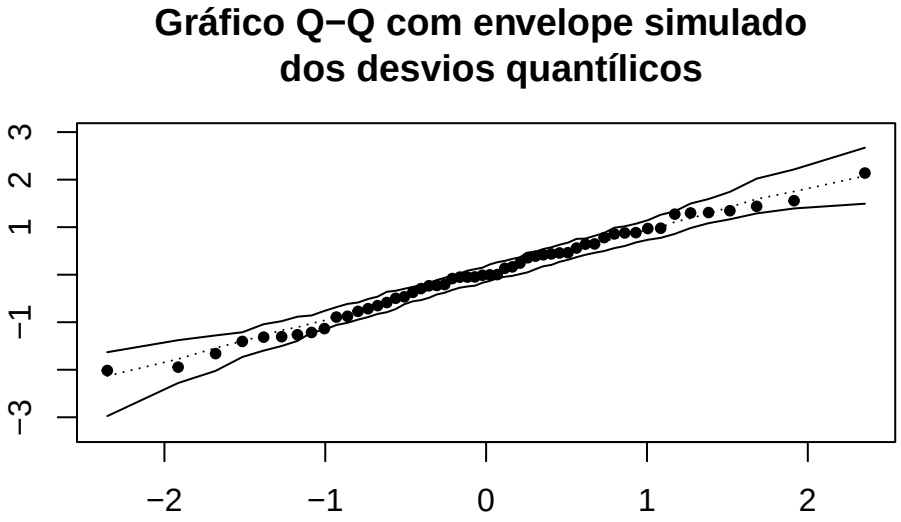
Os desvios aparentam estar comportados, seguindo a tendência esperada



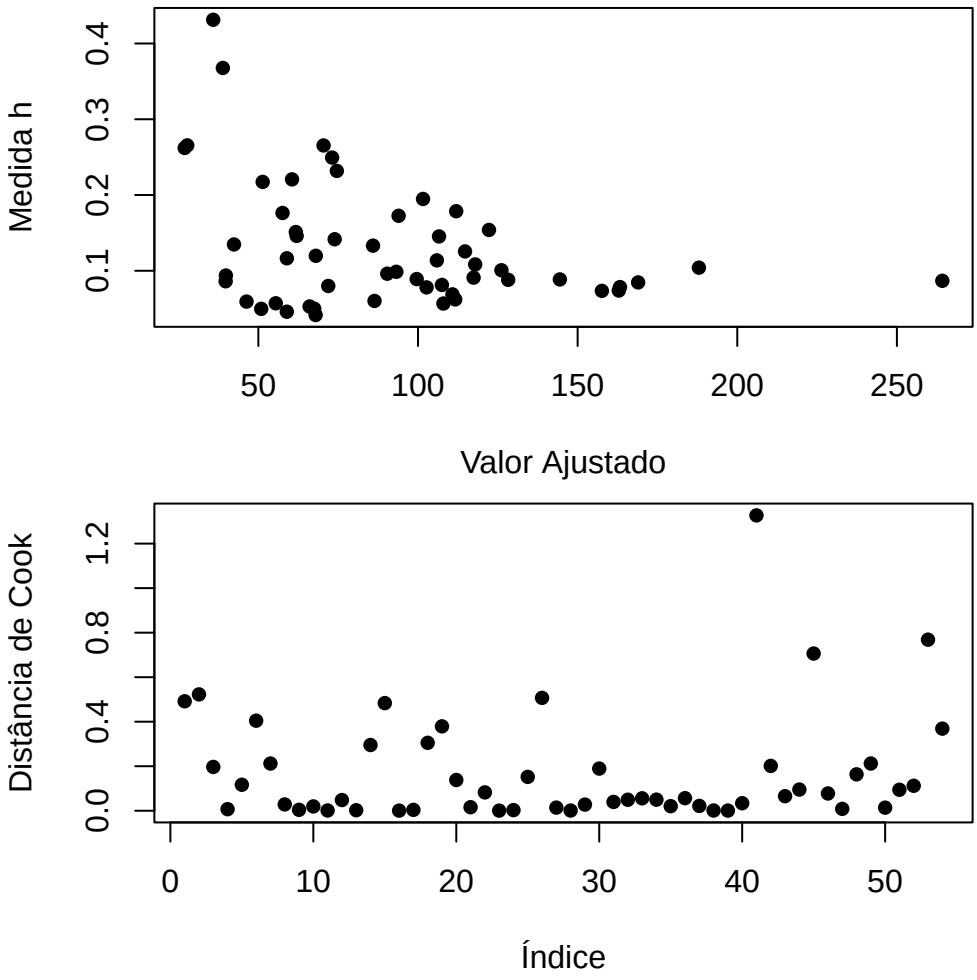


Agora para este modelo, dois pontos se destacam tanto em alavancagem quanto em influência, sendo estes os pontos 47 e 53. Irei testar removê-los para verificar o ajuste.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-27.9118	5.2296	-5.3372	0.0000
diametro	5.7160	0.8052	7.0989	0.0000
tempo	1.5504	0.1979	7.8360	0.0000
tipoCassDish	7.6051	7.8608	0.9675	0.3383
tipoDish	2.5741	4.0949	0.6286	0.5327
tipoPlate	-18.0868	5.0499	-3.5816	0.0008
tipoTray	-20.3505	7.0200	-2.8990	0.0057



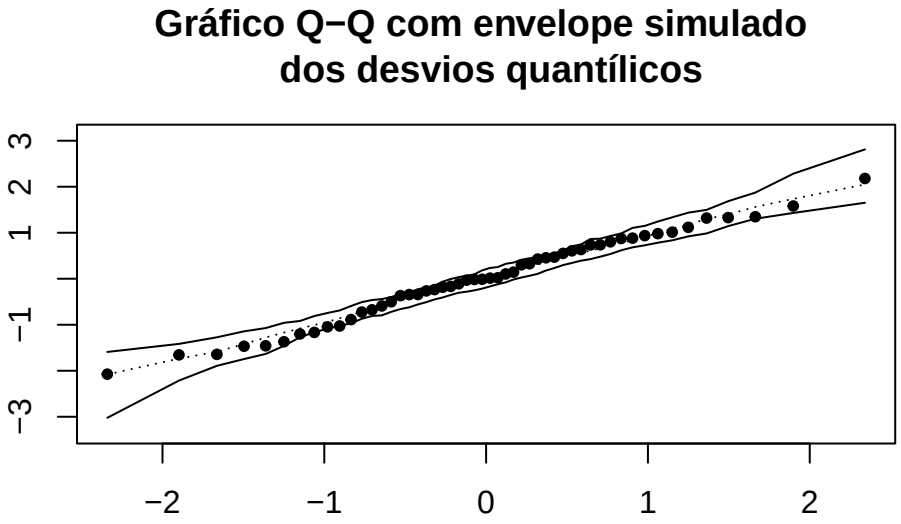
Os desvios aparentam estar comportados, seguindo a tendência esperada



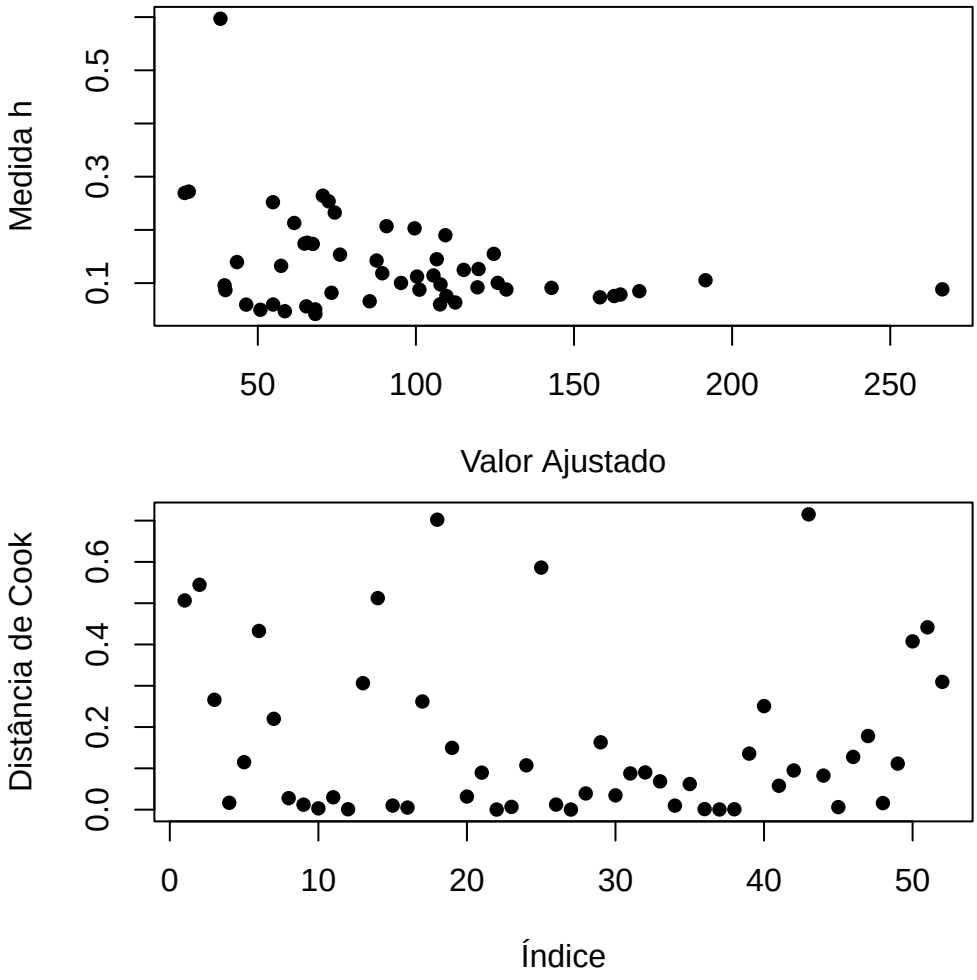
Agora, os novos pontos de alavancagem e influência foram os pontos 10 e 41. Podemos testar remove-los também

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-26.6802	5.2933	-5.0404	0.0000
diametro	5.2605	0.8327	6.3172	0.0000
tempo	1.6601	0.1991	8.3388	0.0000
tipoCassDish	6.4164	7.6192	0.8421	0.4042
tipoDish	1.6904	4.6866	0.3607	0.7200
tipoPlate	-13.2339	5.6078	-2.3599	0.0227
tipoTray	-19.9512	6.8521	-2.9117	0.0056

Vemos novamente uma leve alteração na estimativa dos parâmetros do modelo



Os desvios aparentam estar comportados, seguindo a tendência esperada

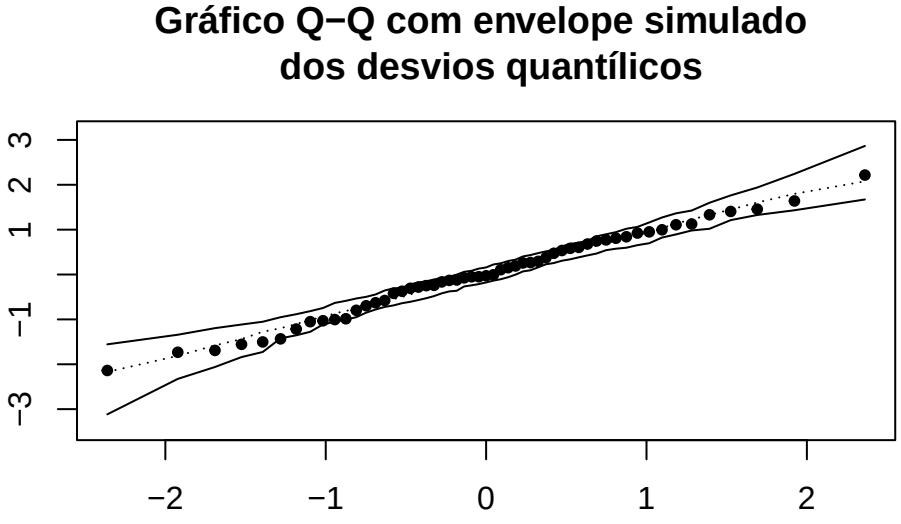


E novamente temos pontos de influência e alavancagem. Analisando estes pontos, não aparentam ser outliers do conjunto de dados, levando a crer que ainda que eu siga re-

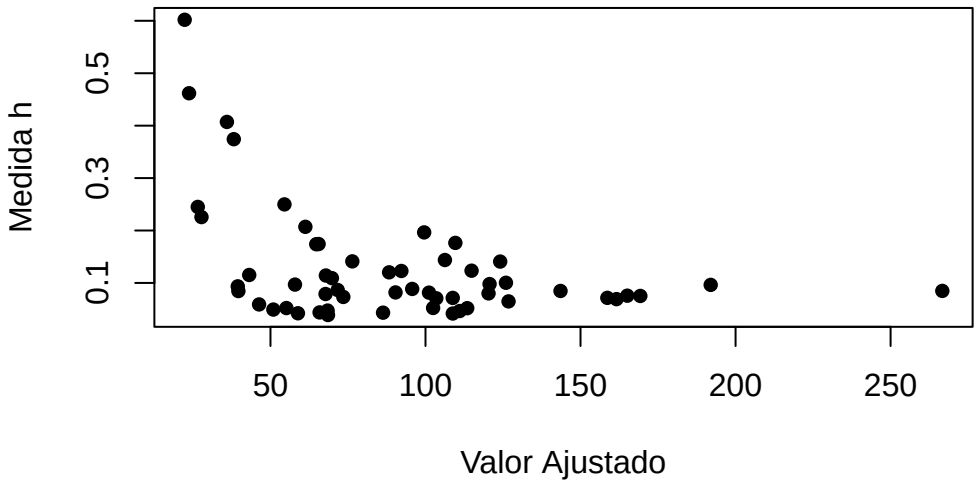
movendo estes pontos, numa próxima análise haverão novos pontos de influência e alavancagem. Como as estimativas dos parâmetros não estão sendo significativamente alteradas, acredito que o melhor procedimento seja re-inserir todos os pontos, e fixar remover apenas aqueles com magnitude > 1 após cada re-ajuste, tanto para alavancagem tanto para distancia de Cook. Desta forma, optei por remover os pontos 14, 43, 55 e 59, e interpretar o ajuste final sem esses pontos

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-28.2474	4.0239	-7.0199	0.0000
diametro	5.4543	0.6223	8.7652	0.0000
tempo	1.6656	0.1784	9.3339	0.0000
tipoCassDish	5.3785	7.1803	0.7491	0.4575
tipoDish	1.8792	3.8140	0.4927	0.6245
tipoPlate	-14.2164	4.9821	-2.8535	0.0064
tipoTray	-23.6129	2.9421	-8.0259	0.0000

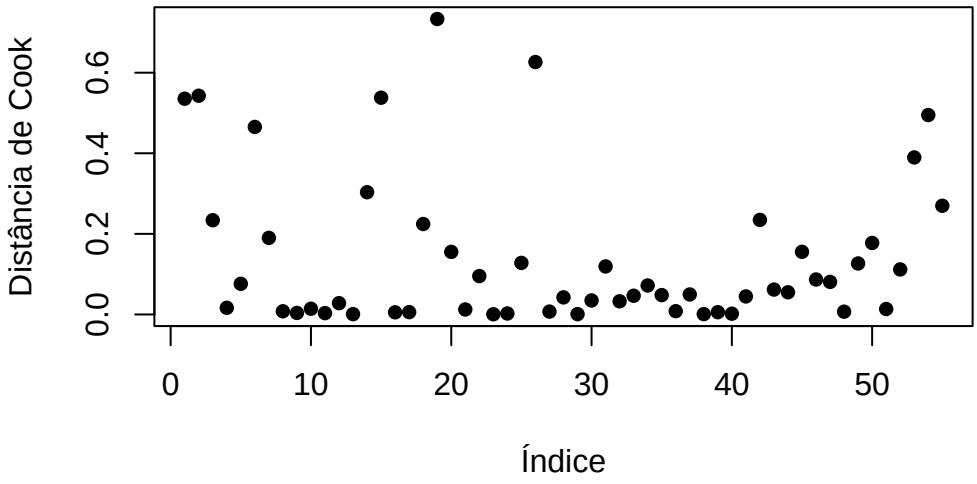
Como o modelo com a ligação identidade foi utilizado, a interpretação dos coeficientes é feita de forma direta, isto é, temos um intercepto negativo, e para as covariáveis tempo e diâmetro, o aumento de uma unidade de tempo implica no aumento pontual de aproximadamente 1,66 unidades de preço, assim como o aumento de uma unidade no diâmetro, aumenta em aproximadamente 5,45 unidades o preço. Fixada a categoria de referência para a variável tipo, podemos dizer que existe significância para afirmar que um prato de mesma dimensão e tempo de polimento seria em média 14,21 U.M. mais barato que uma cumbuca, e uma travessa de mesmo diametro e tempo de polimento que uma cumbuca seria 23,61 U.M. mais barato que uma cumbuca. Para a cassarola e as louças, este coeficiente seria positivo porém insignificante a $\alpha = 0,05$, portanto não devem ser considerados.



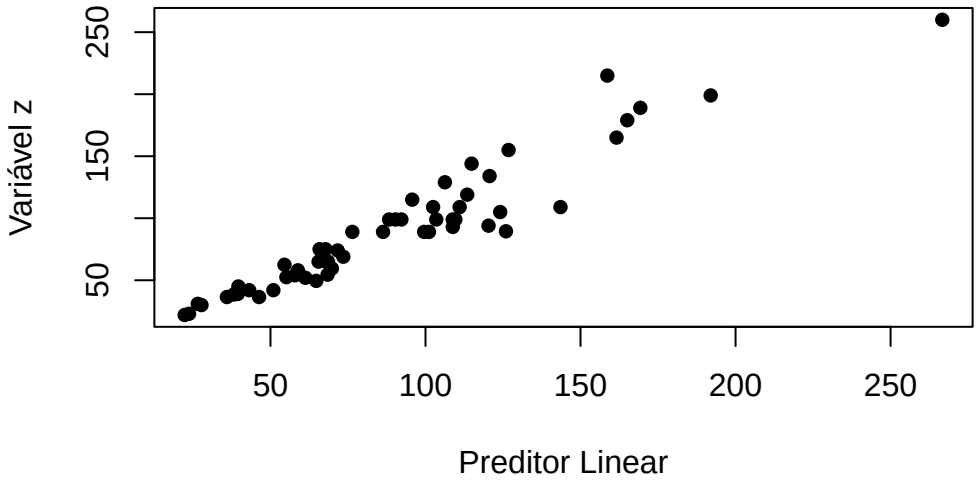
Todos os pontos se encontram dentro do envelope simulado construido para os resíduos quantílicos, com apenas um ponto na fronteira mas ainda contido.



Existem alguns pontos de alavancagem aumentada em relação aos demais, mas conforme discutido nos textos anteriores, optei por considerar esta magnitude aceitável (abaixo de 1).



Para a distância de cook, em geral os pontos estão em magnitudes comparáveis, com dois pontos ultrapassando a distância de 0,6 mas todos inferiores a 1. Conforme discutido anteriormente, acredito que esta seja uma magnitude aceitável para este modelo, e irei considerar normal estes pontos.



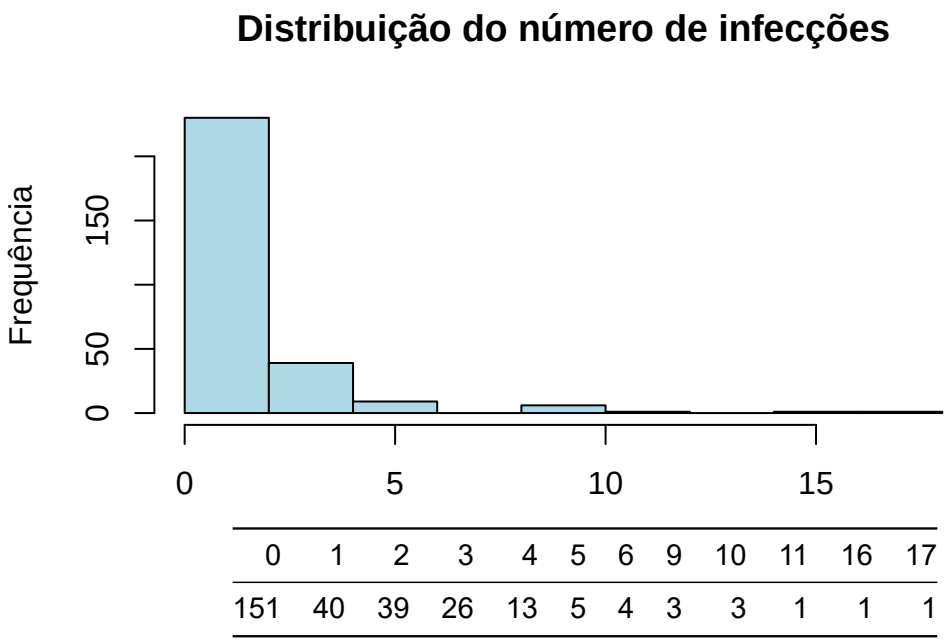
Finalmente, analisando o preditor linear, podemos ver que temos bastante confiança das estimativas dos parâmetros fornecidas anteriormente, com grande parte dos valores sendo aproximadamente corretamente preditos em média se comparado ao valor real, formando uma tendência de reta crescente e semelhante a um gráfico Q-Q normal, com uma boa aderência do preditor linear em relação à variável z.

2 Questão 2

Neste estudo de caso, estamos trabalhando com dados relacionados a contagem de infecção de ouvido em recrutas. Nossa variável resposta será o número de infecções de ouvidos de um recruta, suportado pelas covariáveis hábito (se o recruta tem hábito de nadar ocasional — Occas; ou frequente — Freq), local (local onde o recruta nada, praia — Beach; ou piscina — NonBeach), fetaria (faixa etária do recruta, podendo ser 15 a 19 anos, 20 a 24 anos, 25 a 29 anos) e genero (masculino — Male ou feminino — Female).

2.1 Análise descritiva

2.1.1 Número de infecções



0	1	2	3	4	5	6	9	10	11	16	17
151	40	39	26	13	5	4	3	3	1	1	1

Observando o Histograma e a Tabela da frequência do número de infecções de ouvido nos recrutas contidos no conjunto de dados, observamos que o valor varia de 0 a 17 infecções por recruta, com uma considerável inflação de zeros, bem como uma quantidade relevante de recrutas com 1 a 4 infecções de ouvido.

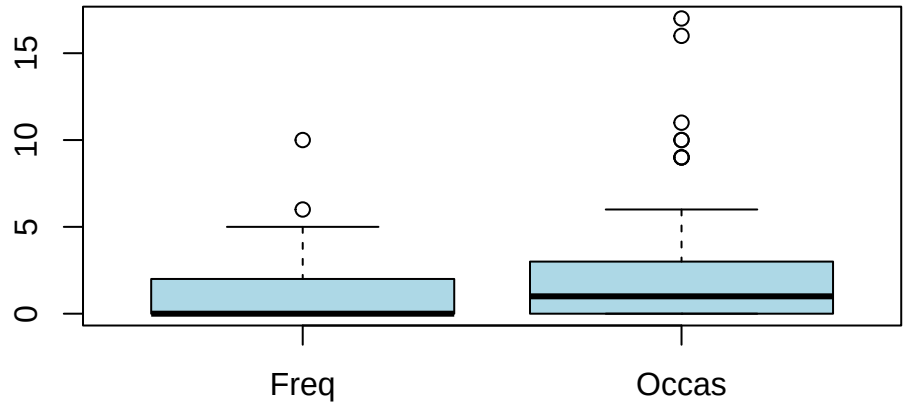
2.1.2 Análise bivariada

Realizando uma análise bivariada, podemos ver para cada categoria das covariáveis o número de infecções de ouvido por meio de boxplots.

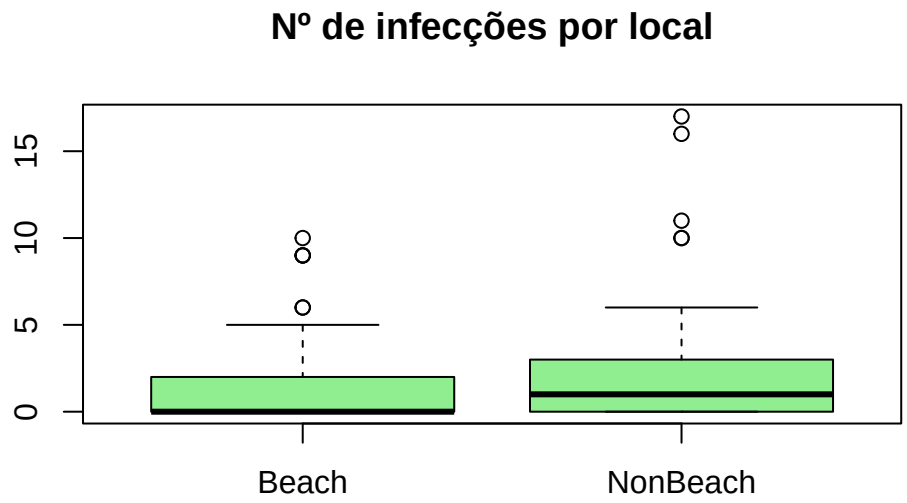
Podemos produzir também tabelas bivariadas, fixando as covariáveis e vendo a frequência de cada uma das demais categorias, bem como observando a diferença nas médias de número de infecções para a categoria fixada.

2.1.2.1 Número de infecções por habito; número de infecções por local

Nº de infecções por habito



	Freq	**Occas**
	N=143	***N=144***
genero:		
Female	48 (33.6%)	51 (35.4%)
Male	95 (66.4%)	93 (64.6%)
local:		
Beach	72 (50.3%)	75 (52.1%)
NonBeach	71 (49.7%)	69 (47.9%)
fetaria:		
15-19	70 (49.0%)	70 (48.6%)
20-24	40 (28.0%)	39 (27.1%)
25-29	33 (23.1%)	35 (24.3%)
ninfec	0.98 (1.52)	1.79 (2.88)

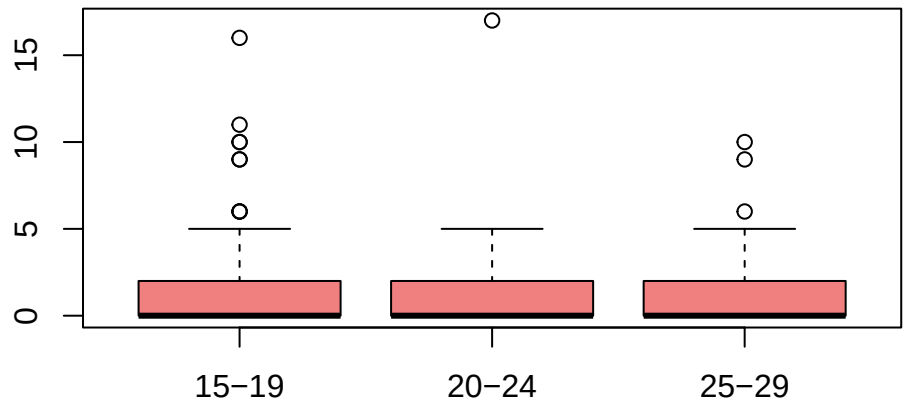


	Beach ***N=147***	**NonBeach** ***N=140***
habito:		
Freq	72 (49.0%)	71 (50.7%)
Occas	75 (51.0%)	69 (49.3%)
genero:		
Female	62 (42.2%)	37 (26.4%)
Male	85 (57.8%)	103 (73.6%)
fetaria:		
15-19	69 (46.9%)	71 (50.7%)
20-24	33 (22.4%)	46 (32.9%)
25-29	45 (30.6%)	23 (16.4%)
ninfec	1.05 (1.92)	1.74 (2.67)

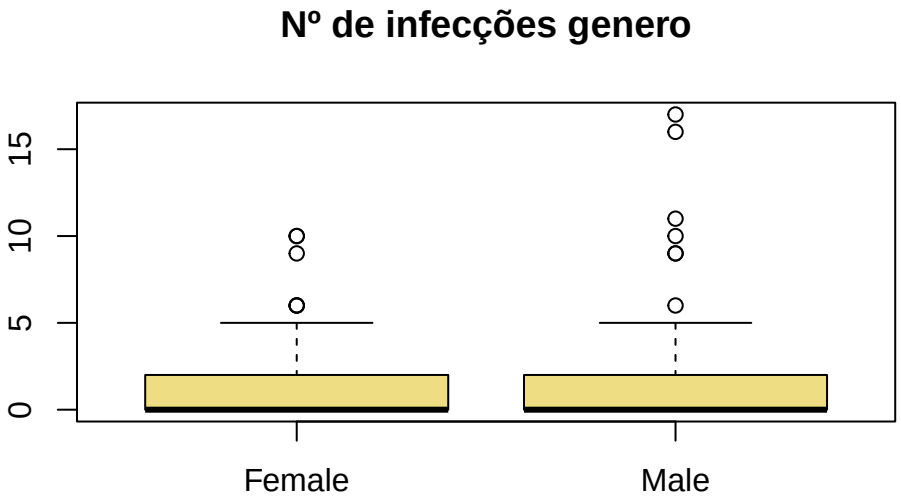
Dos boxplots e tabelas acima, podemos observar que aparenta haver uma incidência maior de número de infecções entre os recrutas que nadam ocasionalmente ante aos que nadam frequentemente, bem como uma maior prevalência de infecções nos que nadam em piscinas ante aos que nadam na praia.

2.1.2.2 Número de infecções por faixa etária; número de infecções por gênero

Nº de infecções fetaria



	15-19 ***N=140***	**20-24** ***N=79***	**25-29** ***N=68***
habito:			
Freq	70 (50.0%)	40 (50.6%)	33 (48.5%)
Occas	70 (50.0%)	39 (49.4%)	35 (51.5%)
local:			
Beach	69 (49.3%)	33 (41.8%)	45 (66.2%)
NonBeach	71 (50.7%)	46 (58.2%)	23 (33.8%)
genero:			
Female	37 (26.4%)	40 (50.6%)	22 (32.4%)
Male	103 (73.6%)	39 (49.4%)	46 (67.6%)
ninfec	1.59 (2.54)	1.16 (2.22)	1.22 (2.02)



	Female *N=99*	Male *N=188*
habito:		
Freq	48 (48.5%)	95 (50.5%)
Occas	51 (51.5%)	93 (49.5%)
local:		
Beach	62 (62.6%)	85 (45.2%)
NonBeach	37 (37.4%)	103 (54.8%)
fetaria:		
15-19	37 (37.4%)	103 (54.8%)
20-24	40 (40.4%)	39 (20.7%)
25-29	22 (22.2%)	46 (24.5%)
ninfec	1.32 (2.11)	1.42 (2.45)

Pelos gráficos e tabelas acima, a faixa etária e o gênero numa análise bivariada aparentam não serem muito significantes se analisadas isoladamente quanto ao número de infecções, mas esta relação pode estar mascarada pela análise bivariada, e ser importante quanto consideradas mais variáveis — o paradoxo de Simpson.

2.1.3 Conclusões

Da análise descritiva, temos um indicativo de que numa modelagem, as covariáveis local e habito serão significativas, enquanto gênero e faixa etária possivelmente não serão tão

significantes. Por se tratar de um problema de contagem, é natural pensar em MLGs que abarcam este tipo de problema, sendo o mais simples o modelo Poisson.

2.2 Modelagem

No contexto de dados de contagem, a modelagem poisson em geral se mostra especialmente útil para modelagem, entretando observamos que neste problema existe uma inflação de zeros, o que pode levar a um problema de ajuste para o modelo Poisson, especialmente se considerarmos a função de ligação log. Neste caso, podemos explorar outras funções de ligação, bem como pensar em um modelo que se comporta melhor ante a dados inflacionados de zero, como por exemplo a modelagem utilizando a binomial negativa.

2.2.1 Modelo 1: MLG Poisson com função de ligação canônica

Apesar da previsão de problema com o MLG Poisson com ligação log (canônica), faz parte da praxis testar este que seria o mais simples para dados de contagem.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.1226	0.1371	-0.8946	0.3710
habitoOccas	0.6115	0.1050	5.8234	0.0000
localNonBeach	0.5345	0.1067	5.0106	0.0000
fetaria20-24	-0.3744	0.1284	-2.9169	0.0035
fetaria25-29	-0.1897	0.1301	-1.4584	0.1447
generoMale	-0.0899	0.1123	-0.8000	0.4237

Observando as significâncias dos parâmetros no modelo saturado, podemos observar que de fato a covariável gênero não foi significativa sob um $\alpha = 0,05$. Temos entretando forte significância das covariáveis habito e local, conforme indicado na análise descritiva. Além disso, existe significância na subcategoria da faixa etaria 20 a 24 anos, ante a categoria de referência 15 a 19 anos — uma associação que não pode ser observada na análise descritiva.

Podemos realizar a escolha manual de covariáveis ou, idealmente, uma escolha automatizada, utilizando a função stepAIC

O procedimento stepAIC indica pelo modelo utilizando apenas as covariáveis local, hábito e faixa etária, não considerando o gênero significativo para a modelagem.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.1814	0.1165	-1.5563	0.1196
habitoOccas	0.6130	0.1050	5.8387	0.0000
localNonBeach	0.5171	0.1044	4.9520	0.0000

	Estimate	Std. Error	z value	Pr(>|z|)
fetaria20-24	-0.3487	0.1241	-2.8093	0.0050
fetaria25-29	-0.1830	0.1299	-1.4084	0.1590

Para este novo modelo com estas covariáveis, temos outras estimativas para os parâmetros, ainda que seja necessária etapas de diagnóstico antes da interpretação destes.

Estatística	Valor	p.valor	H1
z	3.7828	< 0,001	Sobredispersão

O modelo Poisson abarca a suposição de que $\mathbb{E}(Y) = Var(Y)$. Realizando um teste de equidispersão, notamos que a hipótese nula de equidispersão é rejeitada sob qualquer nível de significância, indicando que este não é um modelo adequado para estes dados.

2.2.2 Modelo 2: MLG Poisson com link raiz quadrada

Podemos testar outra função de ligação, afim de conseguir um modelo que cumpra os pressupostos da família selecionada.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.9311	0.0601	15.4965	0.0000
habitoOccas	0.3448	0.0590	5.8399	0.0000
localNonBeach	0.2903	0.0600	4.8367	0.0000
fetaria20-24	-0.2179	0.0705	-3.0909	0.0020
fetaria25-29	-0.1403	0.0746	-1.8812	0.0599

Realizando o ajuste sob o modelo saturado e posteriormente realizando o procedimento stepAIC, notamos que neste modelo também a covariável gênero não é significativa sob um $\alpha = 0,05$. Para este, para além das covariáveis significativas observadas no modelo anterior, também o subfator de faixa etária 25-29 se mostra significativo para um $\alpha = 90\%$.

Estatística	Valor	p.valor	H1
z	3.7785	< 0,001	Sobredispersão

Realizando novamente um teste de equidispersão, notamos que novamente rejeitamos a hipótese nula de $\mathbb{E}(Y) = Var(Y)$. Isso indica que o modelo Poisson de fato não deve ser utilizado para estes dados.

2.2.3 Modelo 3: Binomial negativa com função de ligação canônica

O modelo binomial negativo é geralmente utilizado para problemas de contagem em que não é possível obter um bom ajuste com o modelo Poisson, seja por inflação de zeros, seja por sobredispersão.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.0608	0.2391	-0.2543	0.7992
habitoOccas	0.6030	0.1897	3.1797	0.0015
localNonBeach	0.5455	0.1961	2.7826	0.0054
fetaria20-24	-0.4292	0.2343	-1.8318	0.0670
fetaria25-29	-0.2625	0.2405	-1.0915	0.2751
generoMale	-0.1400	0.2079	-0.6735	0.5006

O modelo binomial negativo com função de ligação log (canônico) traz resultados diferentes dos modelos poisson experimentados anteriormente. As covariáveis hábito e local seguem significantes, enquanto que a covariável faixa etária perde a significância sob $\alpha = 0,05$ para todas as categorias ante a de referência. A covariável gênero, bem como o intercepto, também não são significativos.

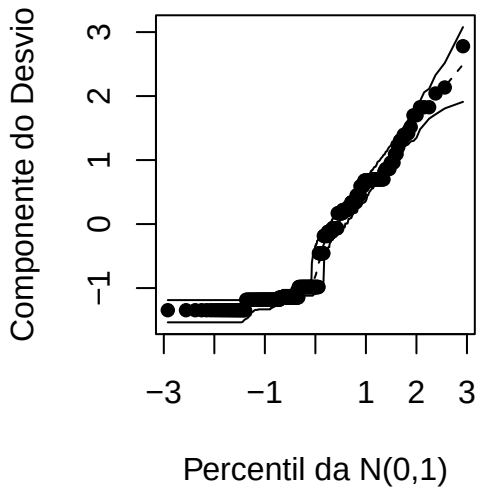
Pela análise anterior e também realizando o procedimento stepAIC, existe a indicação do modelo contendo apenas as covariáveis hábito e local para um ajuste mais adequado e parcimonioso.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.2777	0.1743	-1.5933	0.1111
habitoOccas	0.5902	0.1906	3.0960	0.0020
localNonBeach	0.4813	0.1900	2.5328	0.0113

Devemos realizar os diagnósticos antes de seguir com a interpretação das estimativas dos parâmetros.

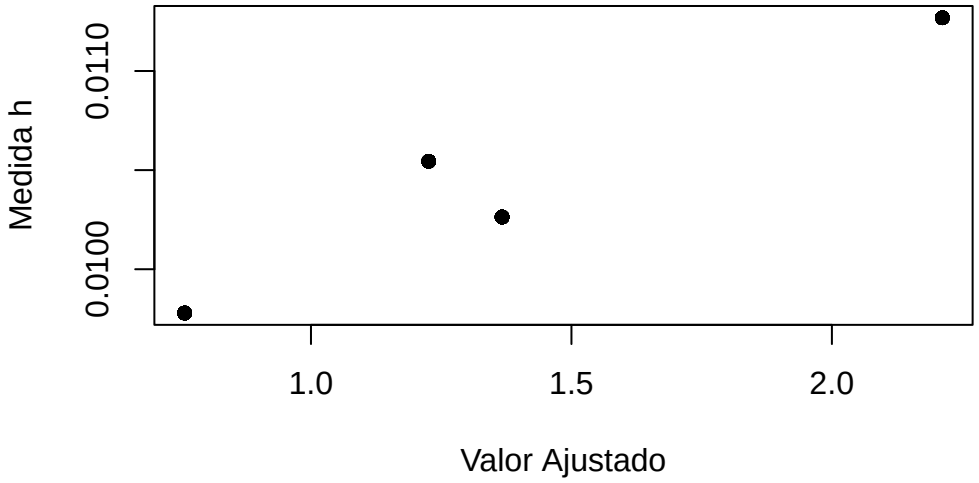
2.2.4 Diagnósticos do modelo 3: Binomial negativa com ligação log

2.2.4.1 Erros sob invelope simulado



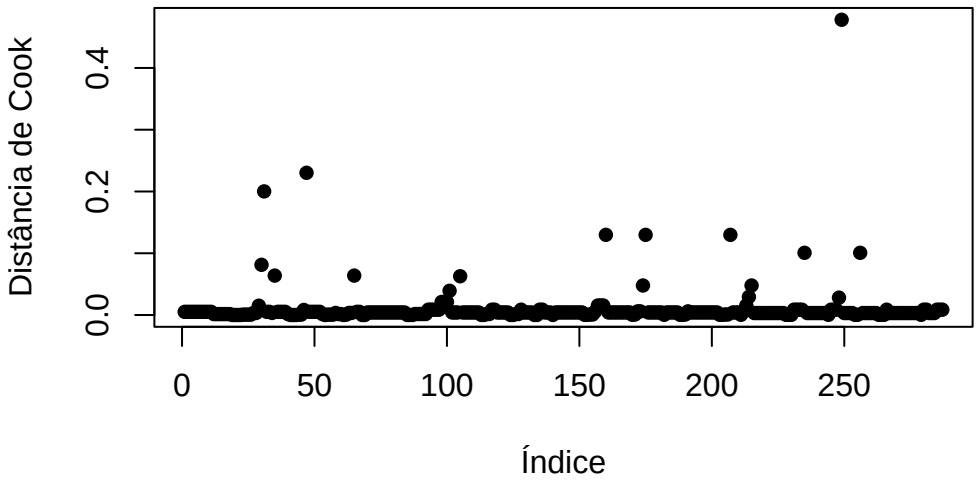
Analisando os erros do modelo utilizando um envelope simulado, o modelo binomial negativo aparenta ter ajustado bem os dados, com todos os valores contidos ou próximos do intervalo simulado construído pelo envelope

2.2.4.2 Medida h



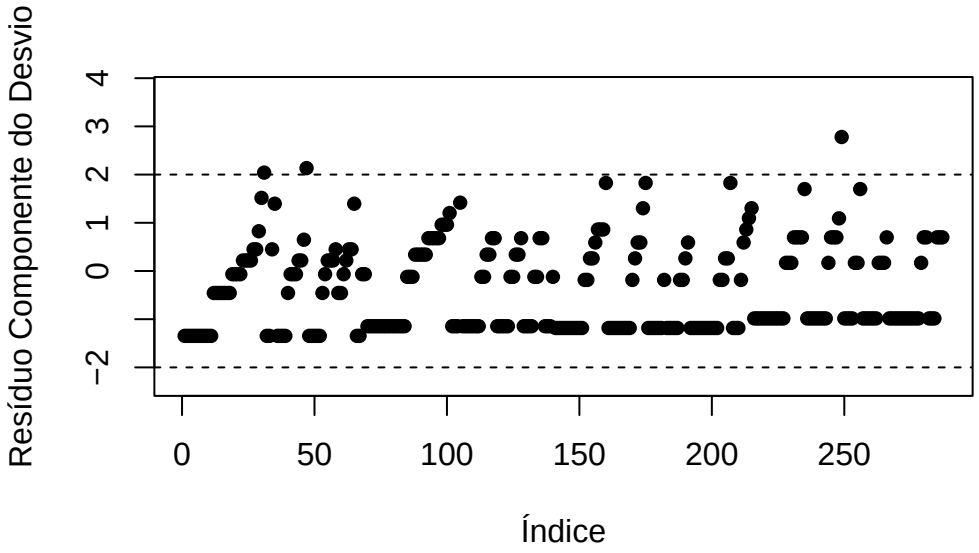
Observando os gráficos da medida h sobre os valores ajustados, temos o indicativo de que não existem pontos dominando as estimativas do modelo. No gráfico aparecem apenas 4 pontos, justamente porque todas as 287 observações obtiveram uma dessas 4 medidas, o que era esperado dado que trabalhamos apenas com duas covariáveis explicativas cada uma com apenas 2 fatores.

2.2.4.3 Distâncias de Cook



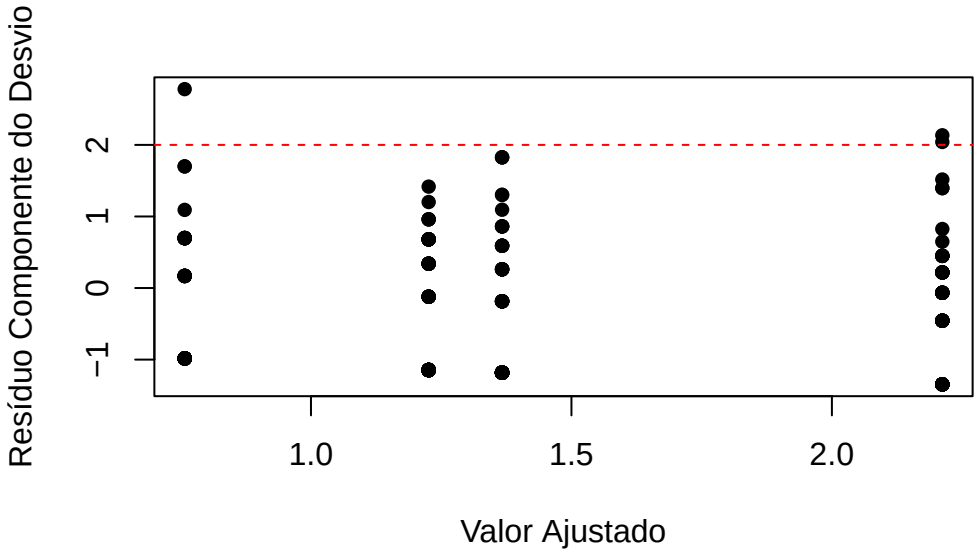
O gráfico das distâncias de Cook também mostra que em geral não existem pontos que estejam inflando as estimativas do modelo. O ponto com maior valor de distância de cook é o ponto 249, com um valor de 0,48, o que em geral não é considerado tão aberrante.

2.2.4.4 Resíduo componente desvio



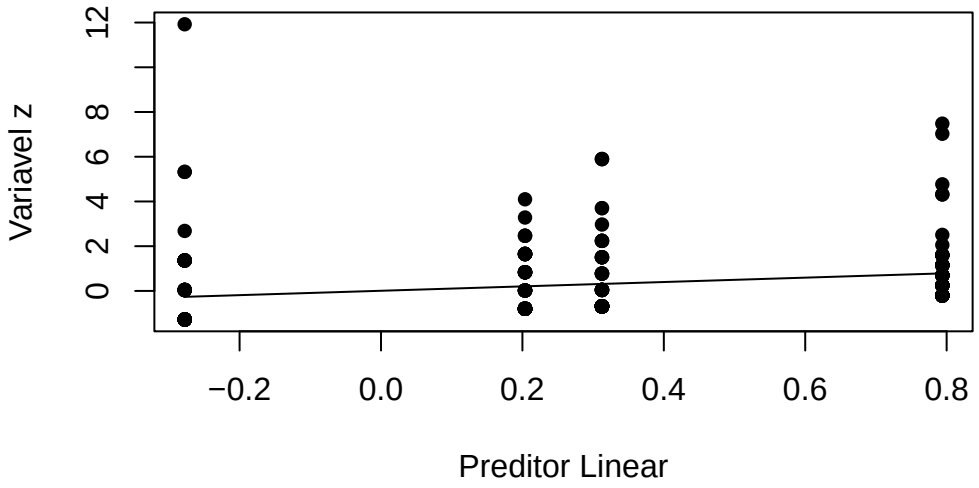
Observando o gráfico do resíduo componente desvio, notamos que o modelo se ajustou bem para a maioria dos dados, com apenas 3 pontos ultrapassando a banda desejável de |2|. Ainda assim, estes pontos não ultrapassam sob uma magnitude muito alta, estando abaixo de 3. Desta forma, seguimos confirmando o diagnóstico de bom ajuste do modelo até aqui.

2.2.4.5 Resíduo componente desvio X valores ajustados.



Pelo gráfico do resíduo componente desvio versus o vajor ajustado, podemos observar que o modelo ajusta bem para a maioria das observações, salvo 3 que estão fora do esperado. Estes pontos são as observações 31, 47 e 249, em que foram observadas 16, 17 e 10 infecções de ouvido, respectivamente. Isso indica que para estes valores aberrantes, o modelo não consegue fazer boas previsões, porém conseguindo realizar boas previsões para as demais observações, e sem que estes outliers comprometam o ajuste do modelo.

2.2.4.6 Qualidade do ajuste



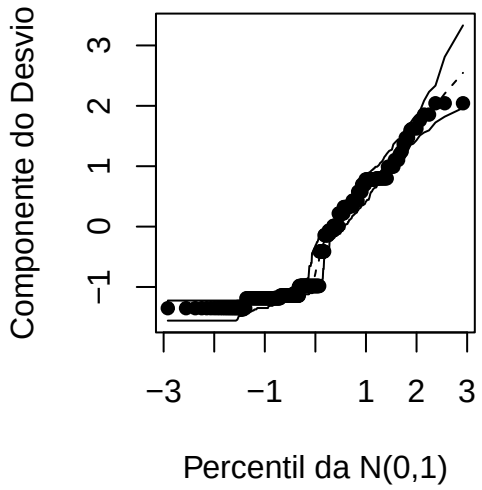
Notamos que a relação não é suave e nem linear, mas pela natureza dos dados contidos no conjunto de dados, não era esperado este comportamento, visto se tratar de um dado de contagem.

Podemos remover os pontos mais aberrantes para melhorar este ajuste — os pontos 31, 47 e 249.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.3420	0.1706	-2.0049	0.0450
habitoOccas	0.5678	0.1856	3.0595	0.0022
localNonBeach	0.4506	0.1849	2.4368	0.0148

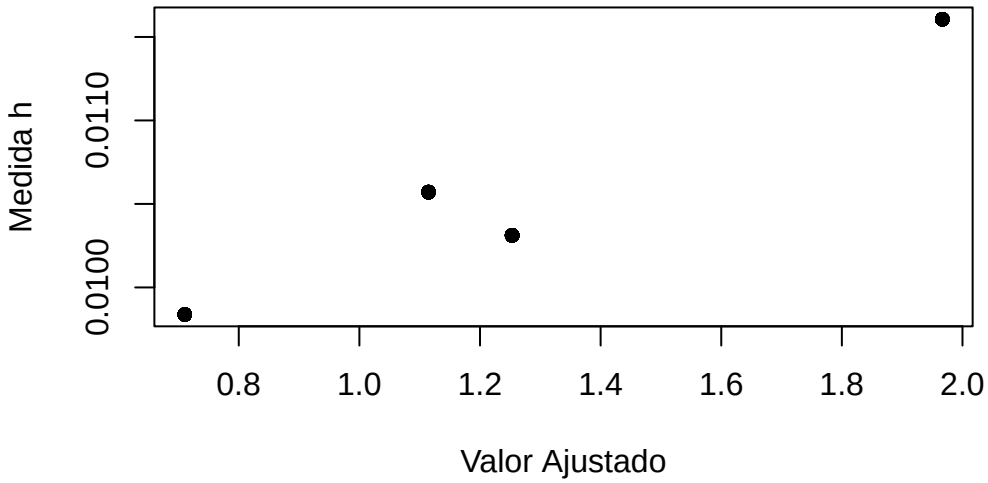
Removendo os pontos, temos novas estimativas para os parâmetros, levemente ajustadas.

2.2.4.7 Erros sob invelope simulado



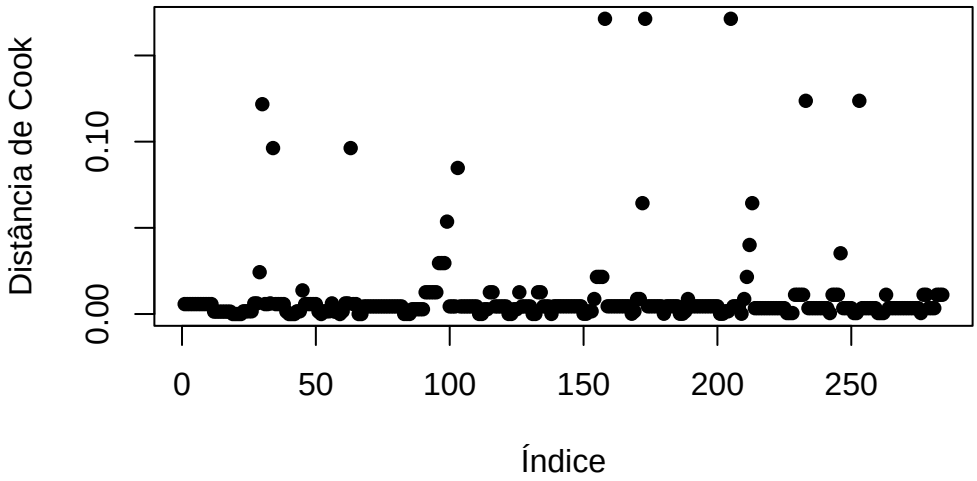
Os resíduos seguem contidos no envelope simulado, dando indicativo de boa qualidade de ajuste

2.2.4.8 Medida h



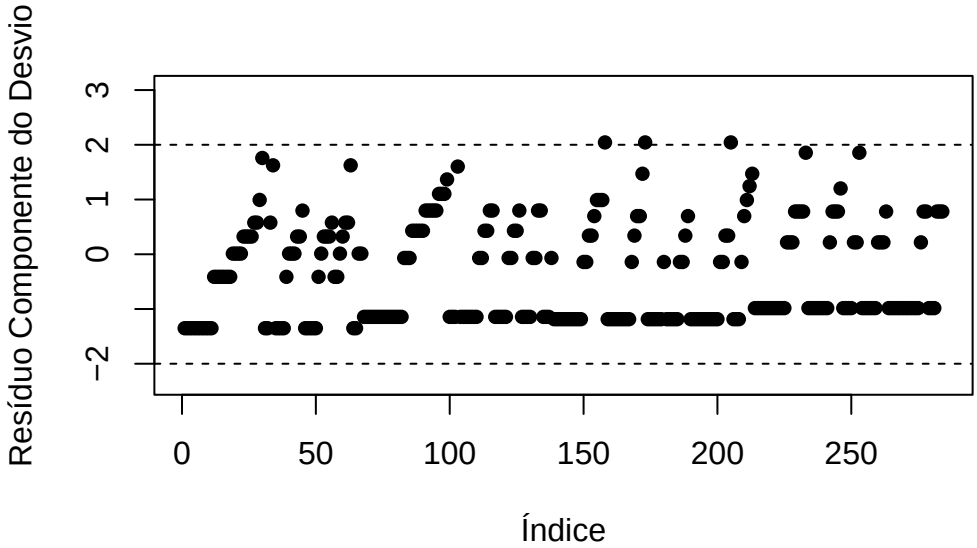
A medida h tem comportamendo análogo a anterior, sem grandes pontos de alavancagem.

2.2.4.9 Distâncias de Cook



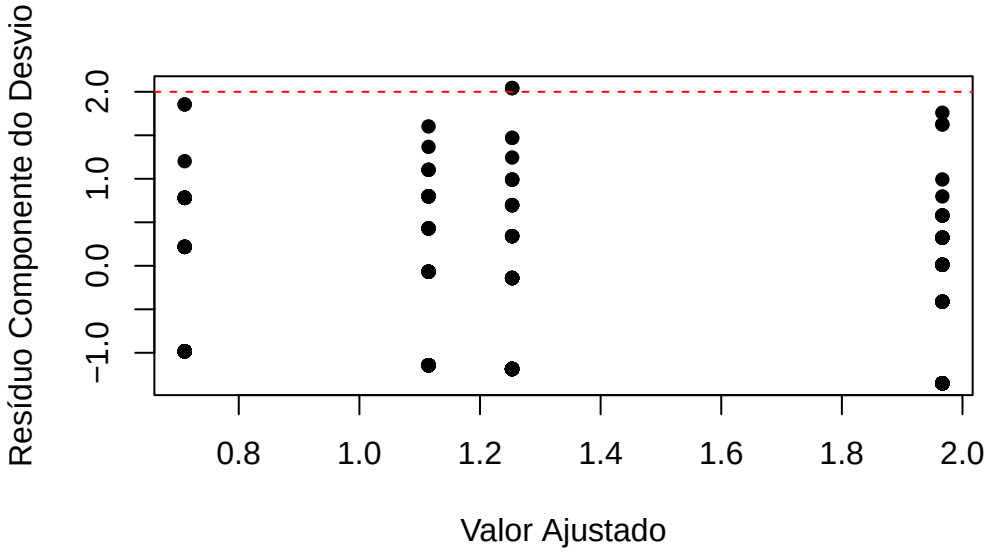
O gráfico das distâncias de cook também não apresenta uma grande fuga do esperado dada a magnitude.

2.2.4.10 Resíduo componente desvio



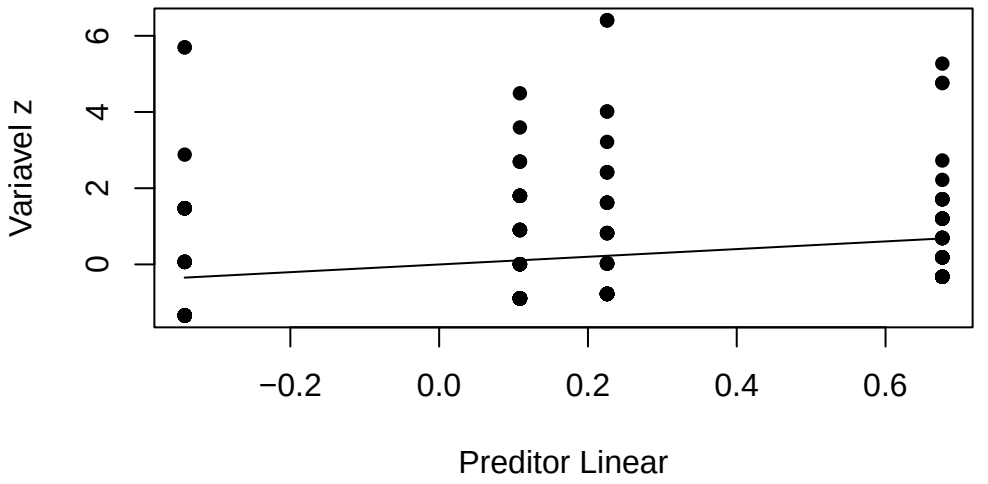
Apesar de padrões de escada, como são dados de contagem talvez fosse esperado algo do tipo. E, ainda assim, não é possível inferir um padrão claro.

2.2.4.11 Resíduo componente desvio X valores ajustados.



O resíduo componente desvio ainda tem um padrão estranho, porém contido no limite desejável.

2.2.4.12 Qualidade do ajuste



Aparentemente, houve uma ligeira melhora na qualidade do ajuste ao remover os três pontos supracitados.

2.2.5 Interpretação do modelo

Visto que o modelo binomial negativo com função de ligação canônico utilizando as covariáveis preditoras hábito e local não teve grandes fugas aos pressupostos, irei escolher este como o melhor modelo para interpretação

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.3420	0.1706	-2.0049	0.0450
habitoOccas	0.5678	0.1856	3.0595	0.0022
localNonBeach	0.4506	0.1849	2.4368	0.0148

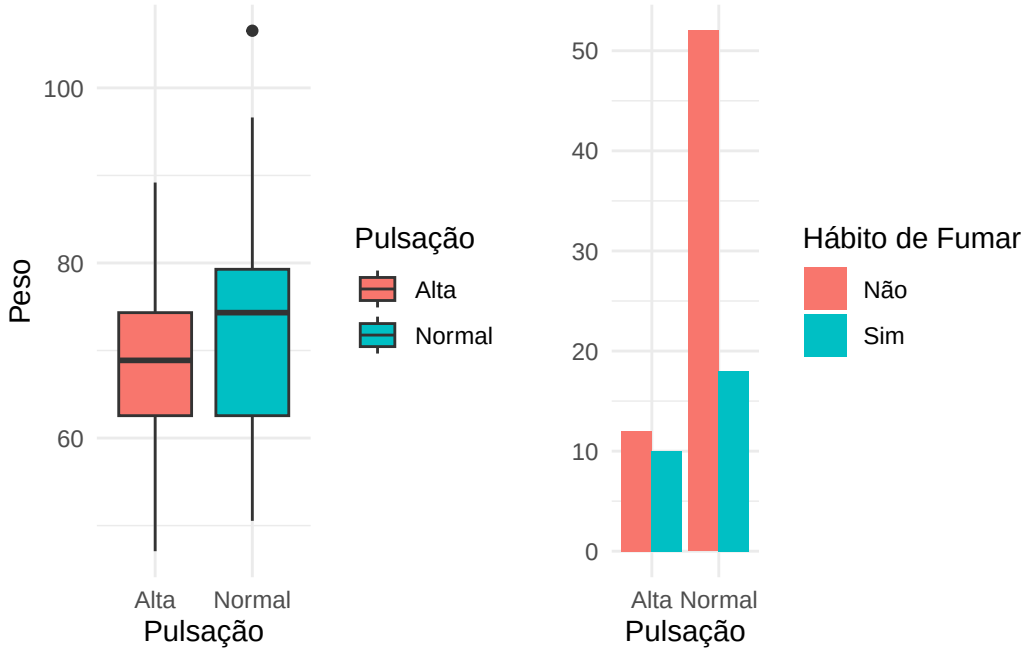
Analisando as estimativas dos parâmetros do modelo, notamos que para um $\alpha = 0,05$ o intercepto do modelo torna-se significativo, ainda que na fronteira da insignificância. Temos então as covariáveis significativas hábito e local. Destas, notamos que, fixada a categoria de referência hábito frequênte de nadar, o hábito ocasional de nadar aumenta a taxa de ocorrência de infecção de ouvido em $e^{0,5678} \approx 1,76 \rightarrow 76\%$ em relação aos recrutas com hábito de natação frequente, enquanto que nadar em piscina aumenta a taxa de ocorrência de infecção de ouvido em $e^{0,4506} \approx 1,57 \rightarrow 57\%$ em relação aos recrutas que nadam na praia.

3 Questão 3

3.1 Análise descritiva

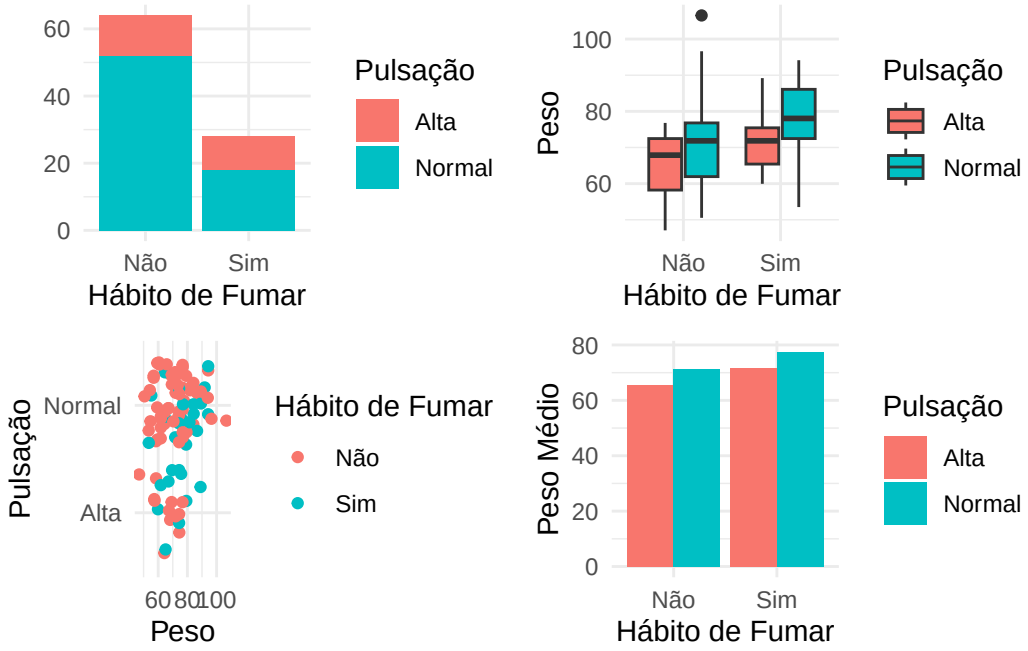
Neste estudo de caso, temos um conjunto de dados de 93 observações, representando um indivíduo por observação, dos quais temos um registro da pulsação em repouso (normal ou alta), do hábito de fumar (sim ou não), e o peso do indivíduo. Buscaremos explicar a relação das covariáveis peso e hábito de fumar com a variável resposta pulsação em repouso.

3.1.1 Análise bivariada



Dos gráficos acima, podemos observar algumas relações bivariadas entre a variável resposta e as covariáveis. Para os indivíduos com pulsação alta, o peso aparenta ser ligeiramente mais baixo. A proporção de indivíduos com hábito de fumar e com pressão alta aparenta ser bem maior que indivíduos sem o hábito de fumar.

3.1.2 Todas as variáveis



Avaliando conjuntamente, vemos que temos uma contagem maior de indivíduos com pulsação normal do que com pulsação alta, mas começa a aparecer uma relação entre um maior peso e hábito de fumo em relação a pulsação. O peso médio dos indivíduos com hábito de fumar é ligeiramente maior, tanto para os com alta pulsação quanto para os com baixa pulsação.

3.2 Modelagem

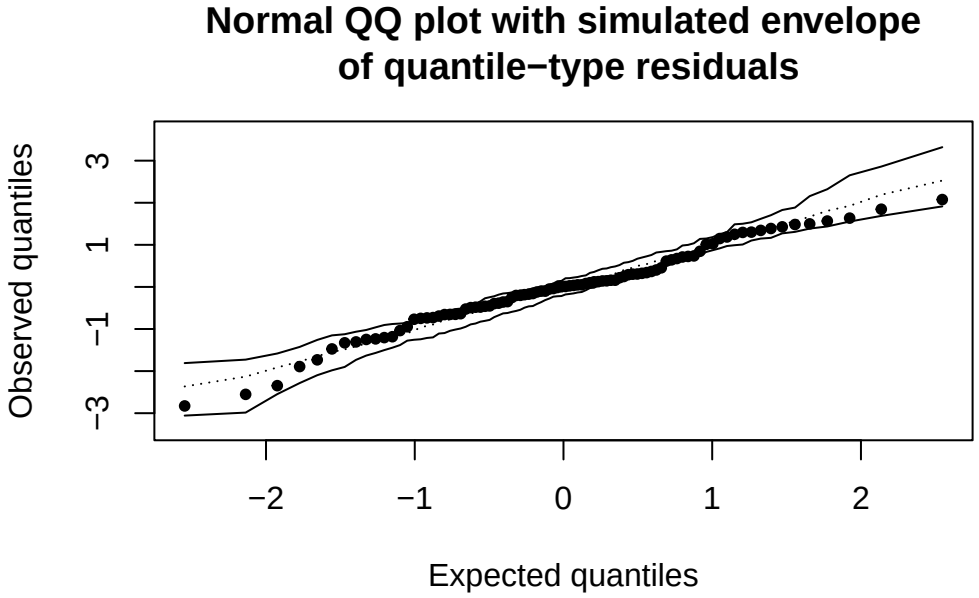
Como temos uma variável resposta binária, podemos ajustar um MLG Bernoulli para explicar a pulsação pelas covariáveis candidatas peso e hábito de fumar

3.2.1 MLG Binomial com função de ligação Probit

Uma das possíveis funções de ligação para o modelo binomial é a ligação probit, que ainda resguarda alguma interpretação, mas não tão boa quanto a interpretação da função de ligação logit.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.9049	1.0844	1.7566	0.0790
peso	-0.0304	0.0142	-2.1473	0.0318
habitof2	-0.7038	0.3250	-2.1653	0.0304

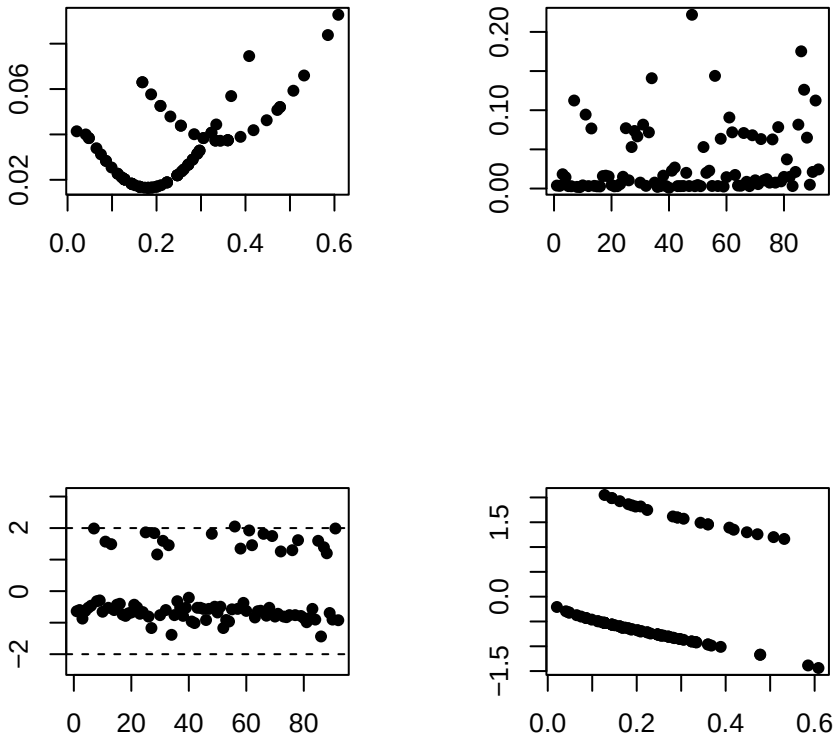
Observando os p-valores marginais dos parâmetros, notamos que ambas as covariáveis são significativas sob $\alpha = 0,05$, e o intercepto somente é significativo sob $\alpha = 0,10$. O procedimento stepAIC recomenda a observação dos p-valores marginais, isto é, manter o modelo saturado.



O gráfico dos resíduos quantílicos com envelope simulado mostra que existem valores próximos da borda do envelope, mas num geral não existe um grande afastamento da distribuição assintótica, sem algum desvio grave perceptível.

3.2.1.1 Diagnósticos

Pontos de alacanca, Pontos influentes (Distância de Cook), Resíduos X Índice, Resíduos X Ajustado.

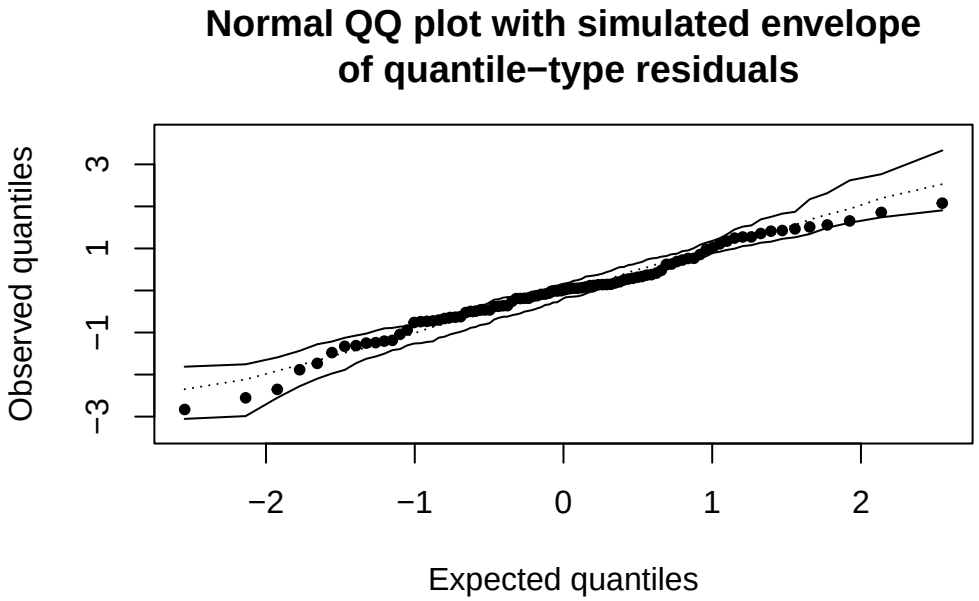


Observando superficialmente os gráficos de diagnóstico do modelo, não existem grandes afastamentos das suposições, nem grandes pontos de alavancagem ou influência, logo este é um modelo aparentemente adequado para estes dados.

3.2.2 MLG Binomial com função de ligação complemento log-log

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.1168	1.5021	1.4092	0.1588
peso	-0.0398	0.0204	-1.9469	0.0515
habitof2	-0.9690	0.4485	-2.1608	0.0307

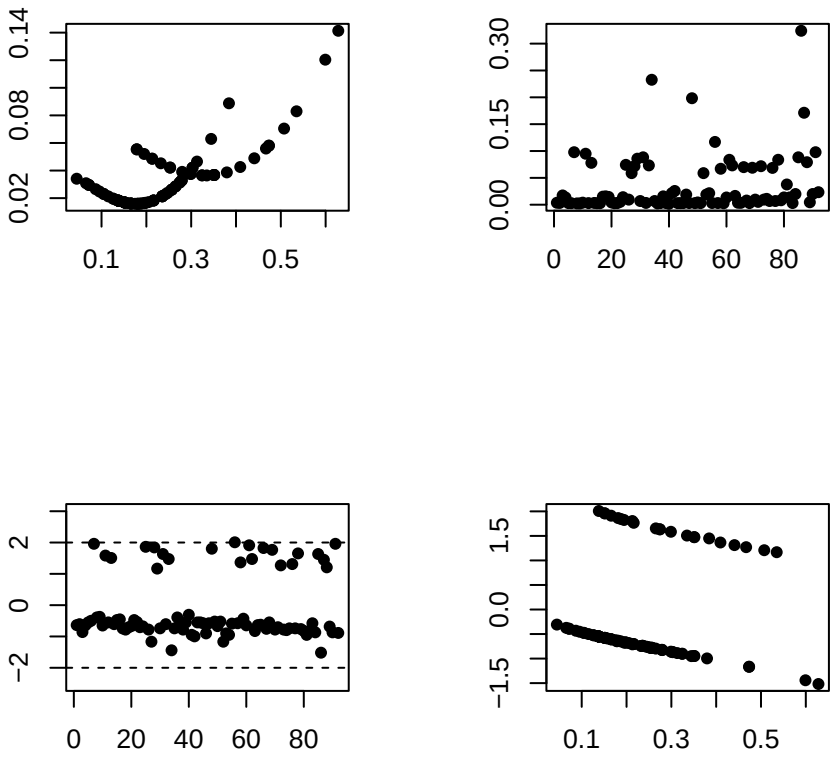
Observando os p-valores marginais dos parâmetros, notamos que apenas a covariável hábito de fumar é significativa sob $\alpha = 0,05$, sendo o intercepto e o peso somente significativos sob $\alpha = 0,10$. O procedimento stepAIC recomenda manter o modelo saturado para melhor explicabilidade.



O gráfico dos resíduos quantílicos com envelope simulado mostra que existem valores próximos da borda do envelope, mas num geral não existe um grande afastamento da distribuição assintótica, sem algum desvio grave perceptível.

3.2.2.1 Diagnósticos

Pontos de alacanca, Pontos influentes (Distância de Cook), Resíduos X Índice, Resíduos X Ajustado.

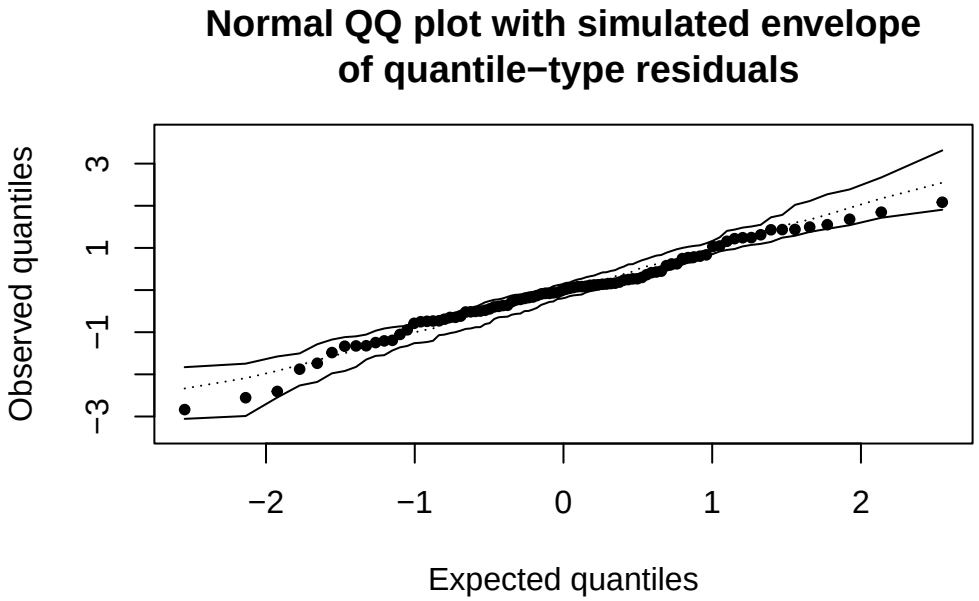


Observando superficialmente os gráficos de diagnóstico do modelo, não existem grandes afastamentos das suposições. Existem alguns pontos com indício de alavancagem e influência, ainda que sob uma magnitude baixa. Caso este modelo fosse escolhido, o ideal seria remover estes pontos e refazer o ajuste para verificar novamente os gráficos, porém irei seguir testando outras funções de ligação ao invés.

3.2.3 MLG Binomial com função de ligação Cauchy

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.4109	2.2377	1.5243	0.1274
peso	-0.0521	0.0310	-1.6813	0.0927
habitof2	-1.3541	0.6477	-2.0905	0.0366

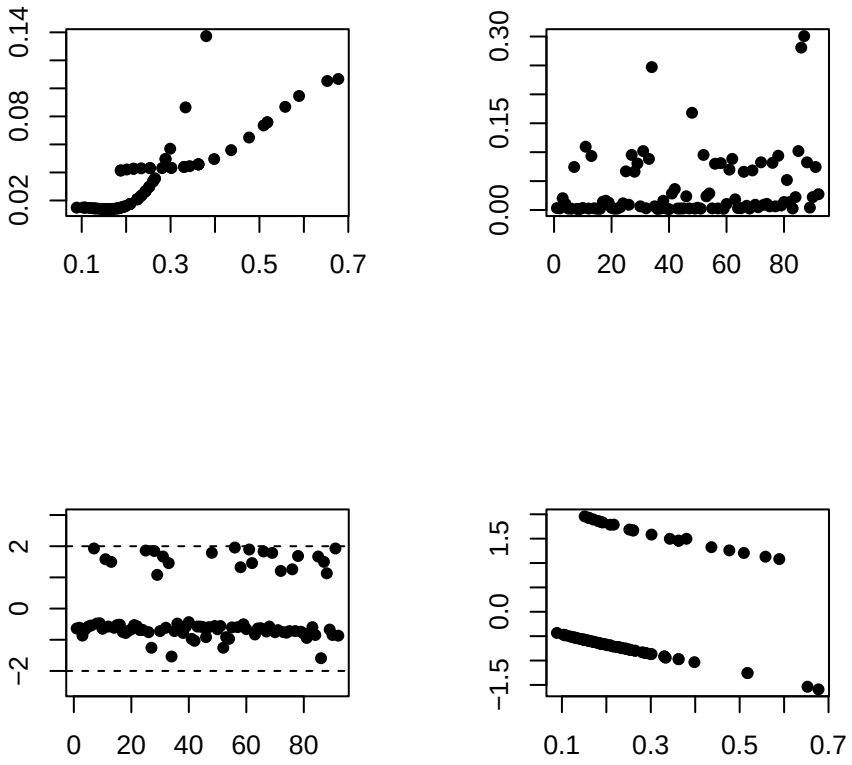
Observando os p-valores marginais dos parâmetros, notamos que apenas a covariável hábito de fumar é significativa sob $\alpha = 0,05$, o peso somente significativo sob $\alpha = 0,10$. O intercepto não é mais significativo neste modelo nem para $\alpha = 0,10$. O procedimento stepAIC recomenda manter o modelo saturado para melhor explicabilidade.



O gráfico dos resíduos quantílicos com envelope simulado mostra que existem valores próximos da borda do envelope, mas num geral não existe um grande afastamento da distribuição assintótica, sem algum desvio grave perceptível.

3.2.3.1 Diagnósticos

Pontos de alacanca, Pontos influentes (Distância de Cook), Resíduos X Índice, Resíduos X Ajustado.



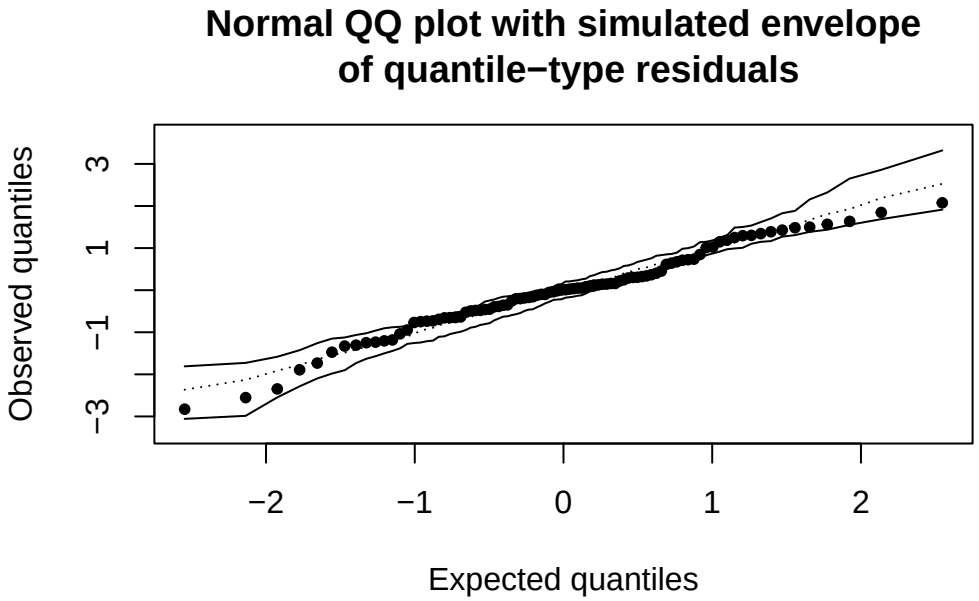
Observando superficialmente os gráficos de diagnóstico do modelo, não existem grandes afastamentos das suposições. Existem um ponto com indício de alavancagem, e dois ou três pontos com indício de influência, ainda que sob uma magnitude baixa. Caso este modelo fosse escolhido, o ideal seria remover estes pontos e refazer o ajuste para verificar novamente os gráficos, porém irei seguir testando outras funções de ligação ao invés.

3.2.4 Modelo logístico (binomial com função de ligação logit)

Um modelo que sempre devemos testar neste tipo de situação é o modelo logístico, visto que é adequado para dados com 2 fatores possíveis, e seus parâmetros são passíveis de serem interpretados de forma elegante.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.1802	1.8714	1.6994	0.0892
peso	-0.0505	0.0247	-2.0418	0.0412
habitof2	-1.1930	0.5530	-2.1574	0.0310

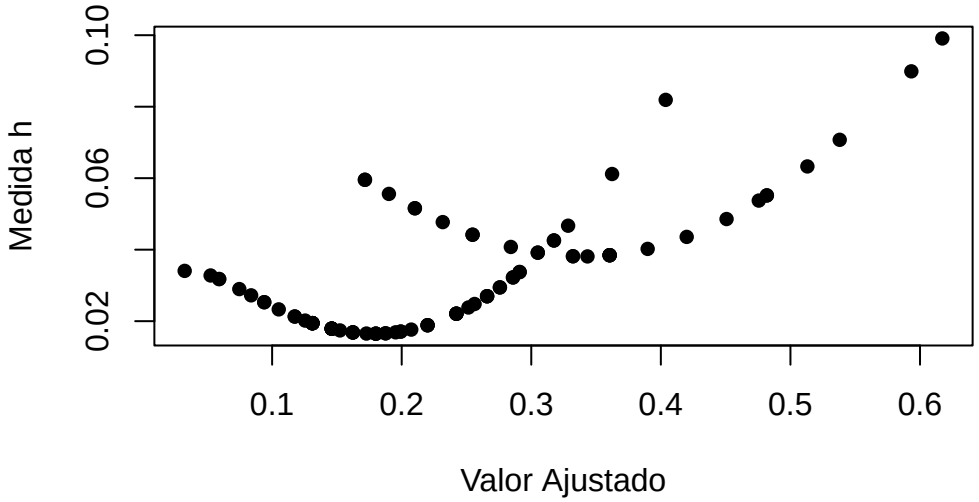
Observando os p-valores marginais dos parâmetros, notamos que ambas as covariáveis são significativas sob $\alpha = 0,05$, e o intercepto somente é significativo sob $\alpha = 0,10$. O procedimento stepAIC recomenda a observação dos p-valores marginais, isto é, manter o modelo saturado.



O gráfico dos resíduos quantílicos com envelope simulado mostra que existem valores próximos da borda do envelope, mas num geral não existe um grande afastamento da distribuição assintótica, sem algum desvio grave perceptível.

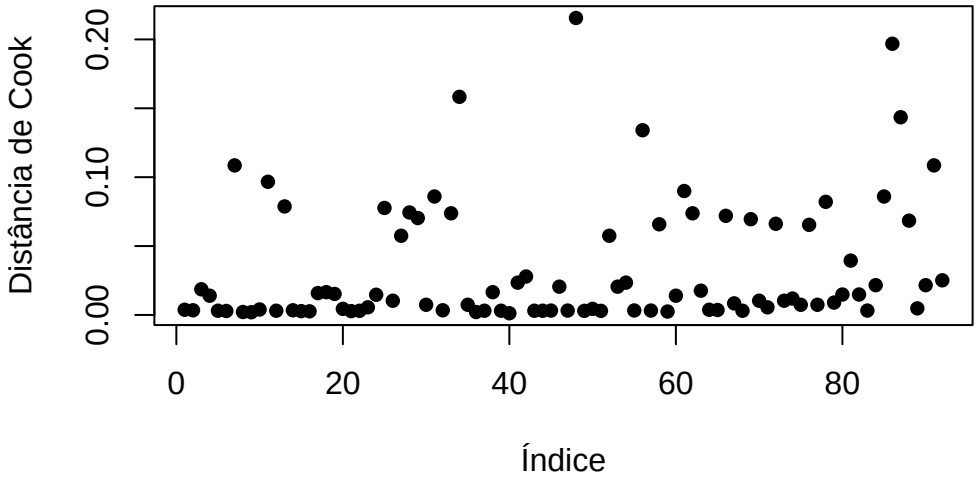
3.2.4.1 Diagnósticos

3.2.4.1.1 Pontos de alacanca



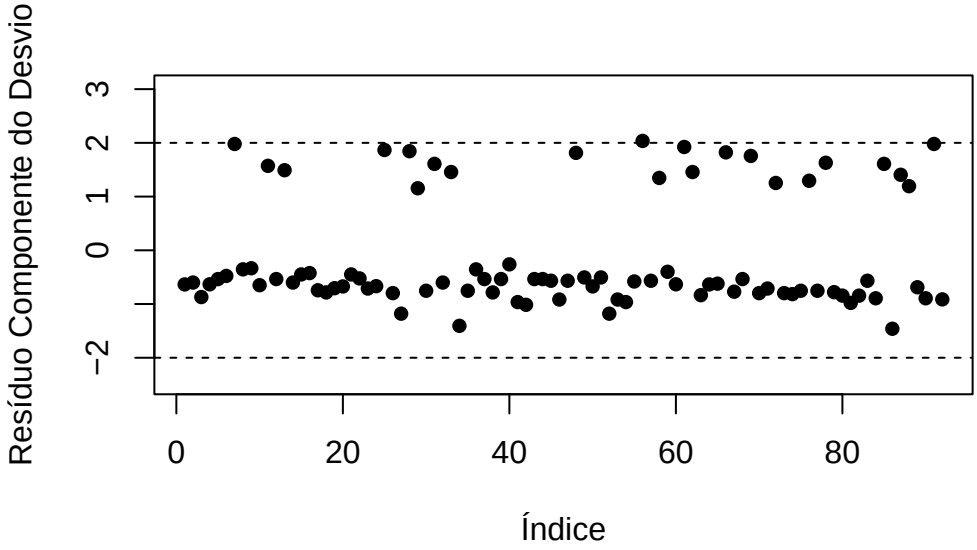
Avaliando o gráfico de alavancagem, existem dois ou três pontos que se destacam um pouco em relação aos demais, mas em escala de magnitude não aparentam estar distorcendo muito as estimativas do modelo, portanto não julgo necessária a remoção destes.

3.2.4.1.2 Pontos influentes (Distância de Cook)



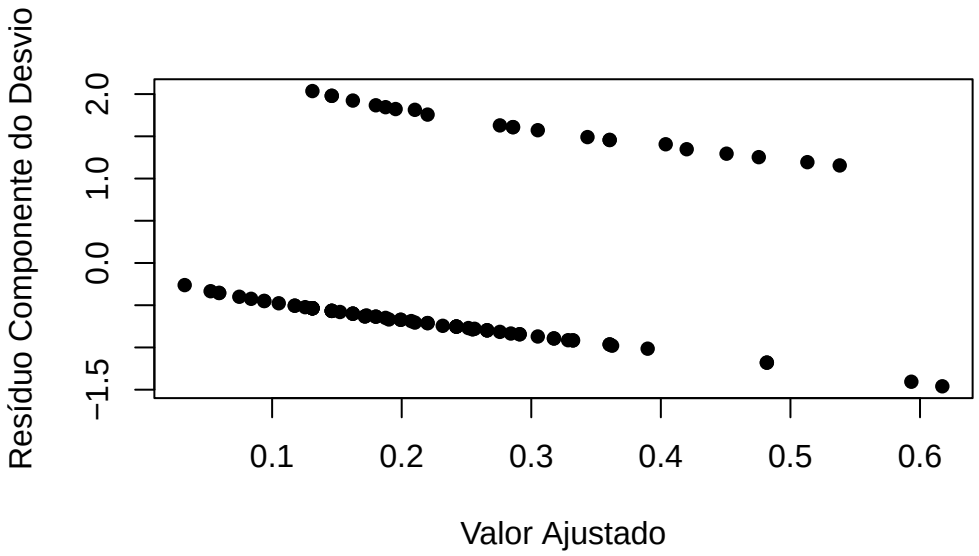
Analisando o gráfico de influência pela distância de Cook, a interpretação que farei será análoga a anterior, ou seja, existem alguns pontos que se destacam levemente em relação aos demais, mas como em escala de magnitude não aparentam ser muito aberrantes, acredito que não seja necessário fazer a remoção destes pontos para o ajuste.

3.2.4.1.3 Resíduos X Índice



Claramente podemos ver algum tipo de relação de dependência neste gráfico, com as observações próximas ao limite superior provavelmente sendo dos com hábito de fumar, e as próximas do limite inferior referente aos indivíduos sem hábito de fumo em maioria. Ainda assim, aparenta estar “ok” os pressupostos.

3.2.4.1.4 Resíduos X Ajustado



O preditor aparenta ter obtido um bom ajuste.

3.2.4.2 Conclusão

Dentre todos os modelos observados, nenhum apresentou grande fuga aos pressupostos, tendo levado a ajustes semelhantes e até parâmetros parecidos em algum sentido. Neste caso, devemos adotar pelo princípio da parcimônia, e aceitar o modelo mais simples e com maior explicabilidade para modelar estes dados, que neste caso é este último modelo apresentado: o logístico. Neste caso, partirmos para interpretação dos coeficientes do modelo

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.1802	1.8714	1.6994	0.0892
peso	-0.0505	0.0247	-2.0418	0.0412
habitof2	-1.1930	0.5530	-2.1574	0.0310

Interpretando os parâmetros significativos a $\alpha = 0,05$ deste modelo, o peso e o hábito de fumar, vemos que os indivíduos hábito de fumar tem $1 - exp^{-1.1930} \approx 0,70 \rightarrow 70\%$ mais chance de apresentar pulsação em repouso alta em comparação com indivíduos sem o hábito de fumar, mantendo todo o restante constante. Assim como existe um aumento de $1 - exp^{-0.0505} \approx 0,05 \rightarrow 5\%$ na chance do indivíduo apresentar pulsação alta em repouso para cada aumento de 1kg do indivíduo, mantendo todo o restante constante.

Essas conclusões são interessantes, visto que essas relações eram bastante difíceis de observar na análise descritiva exploratória, e somente com a modelagem foi possível ter mais credibilidade nesta interpretação, e na magnitude destas relações.

4 Questão 4

4.1 Dose 1

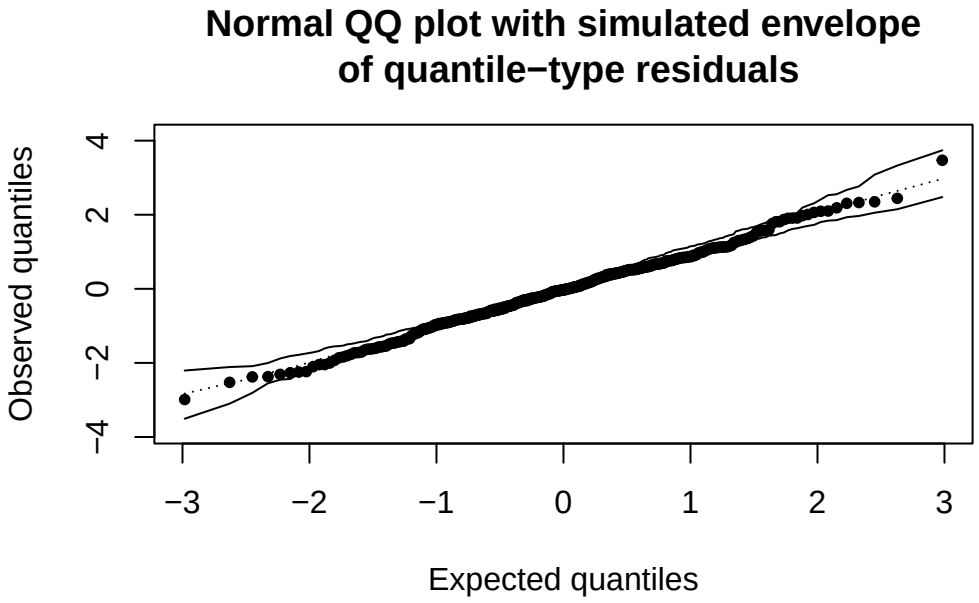
Neste estudo de caso, iremos investigar um experimento de dose-resposta conduzido para avaliar a influência do extrato vegetal aquoso frio de folhas na morte de um determinado tipo de caramujo utilizando modelagem GLM Binomial com diversas funções de ligação para modelar o status (vivo/morto) de acordo com a dose do extrato aplicada ao caramujo.

	morto	vivo
0	4	46
15	5	45
20	14	36
25	29	21
30	38	12
35	41	9
40	47	3

Pela tabela, é fácil observar que o aumento da dose leva a morte dos caramujos. Estamos interessados em quantificar esta relação, portanto, irei testar diversos modelos GLM Binomial e, baseado no critério definido pelo enunciado da questão (envelope), irei interpretar o modelo que for o melhor ao final.

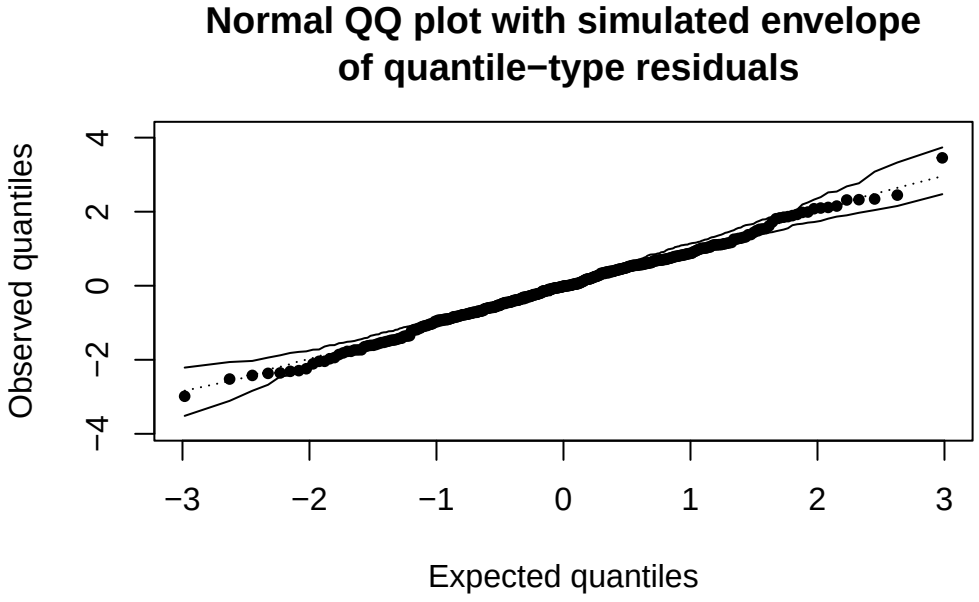
4.1.1 Modelo 1: Modelo logístico (binomial com função de ligação logit)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.8067	0.4519	8.4238	0
dose	-0.1571	0.0171	-9.1991	0



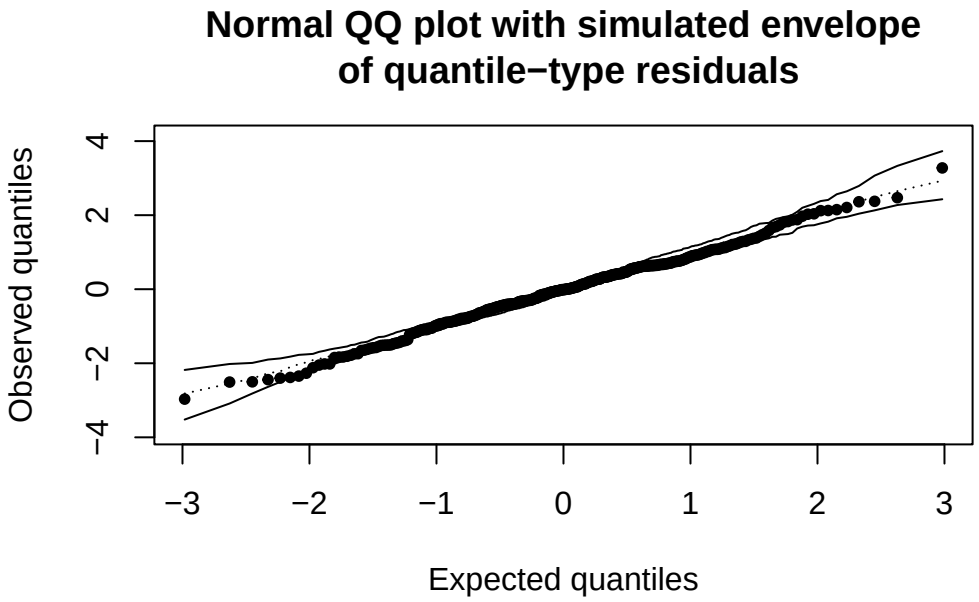
4.1.2 Modelo 2: MLG Binomial com função de ligação Probit

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.0689	0.2325	8.8989	0
dose	-0.0864	0.0086	-9.9942	0



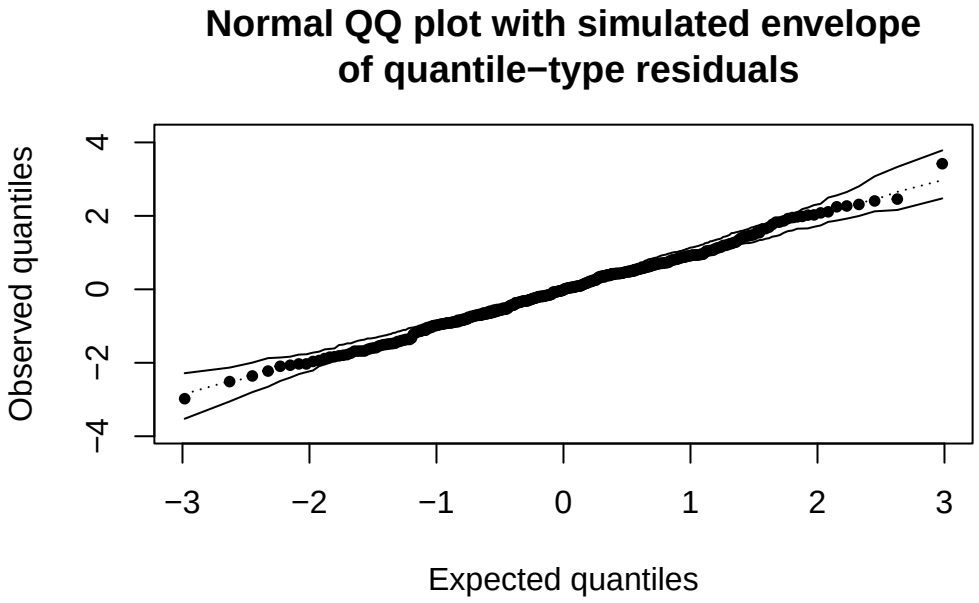
4.1.3 Modelo 3: MLG Binomial com função de ligação complemento log-log

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.4321	0.2016	7.1025	0
dose	-0.0799	0.0085	-9.3724	0



4.1.4 Modelo 4: MLG Binomial com função de ligação Cauchy

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	5.0740	0.8926	5.6844	0
dose	-0.2092	0.0358	-5.8469	0



4.1.5 Escolha do modelo e interpretação

Pelos gráficos e tabelas acima, vimos que para todas as funções de ligação houve forte significância da dose em explicar a morte do caramujo. Observando os envelopes simulados, qualquer um dos modelos poderia ser utilizado, visto que todos estão aderentes ao pressuposto. Desta forma, a escolha será pelo mais parcimonioso e de melhor interpretabilidade, que é o modelo GLM Binomial com função de ligação logito — O modelo logístico.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.8067	0.4519	8.4238	0
dose	-0.1571	0.0171	-9.1991	0

Para este modelo, observamos que para cada aumento em 1 unidade da dose, aumenta em $1 - exp^{-0.1571} \approx 0,1454 \rightarrow 14,54\%$ a chance de morte do caramujo.

Podemos também obter a estimativa da dose letal DL50, podendo ser calculada neste caso por $DL50 = \frac{\beta_0}{\beta_1} = \frac{3,8067}{-0,1571} \approx 24,23$. Logo, esta é a dose letal estimada que mata 50% dos caramujos!

4.2 Dose 2

Neste estudo de caso, iremos investigar um experimento de dose-resposta conduzido para avaliar a influência do extrato vegetal aquoso frio de frutos na morte de um determinado tipo

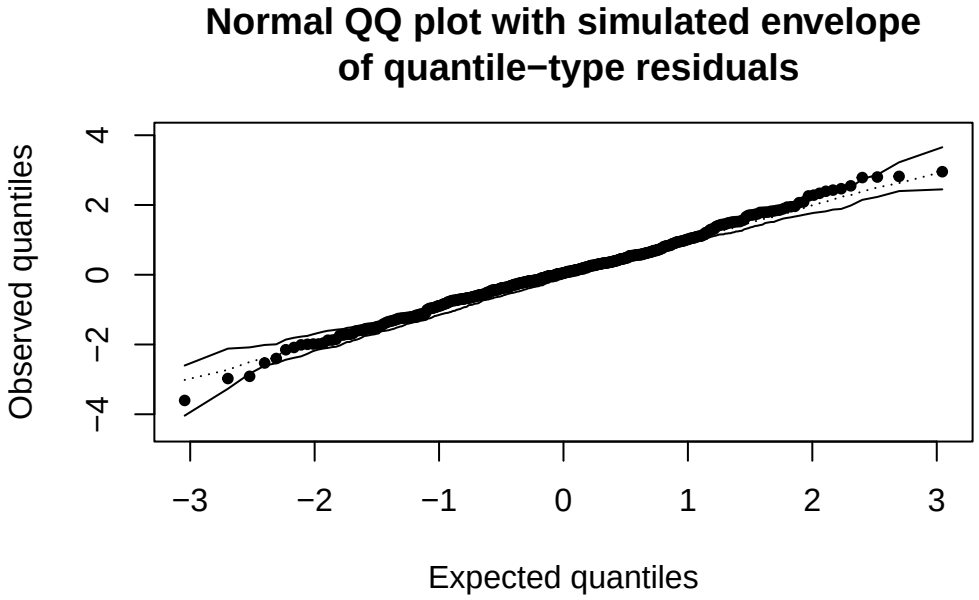
de caramujo utilizando modelagem GLM Binomial com diversas funções de ligação para modelar o status (vivo/morto) de acordo com a dose do extrato aplicada ao caramujo.

	morto	vivo
0	2	63
100	2	50
150	8	44
200	20	32
250	41	11
300	48	4
350	51	1
400	52	0

Pela tabela, é fácil observar que o aumento da dose leva a morte dos caramujos. Estamos interessados em quantificar esta relação, portanto, irei testar diversos modelos GLM Binomial e, baseado no critério definido pelo enunciado da questão (envelope), irei interpretar o modelo que for o melhor ao final.

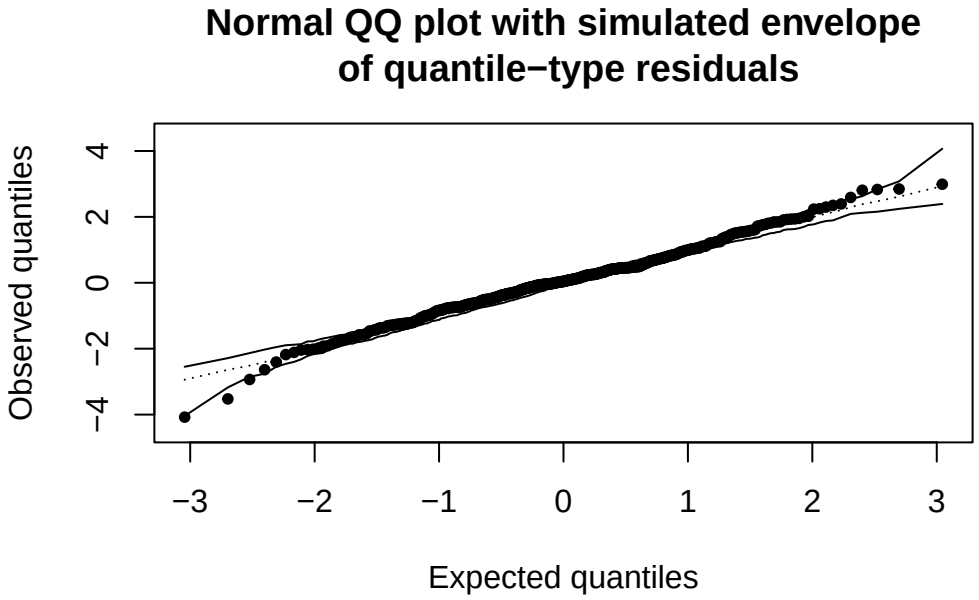
4.2.1 Modelo 1: Modelo logístico (binomial com função de ligação logit)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	5.5462	0.5653	9.8111	0
dose	-0.0265	0.0026	-10.3626	0



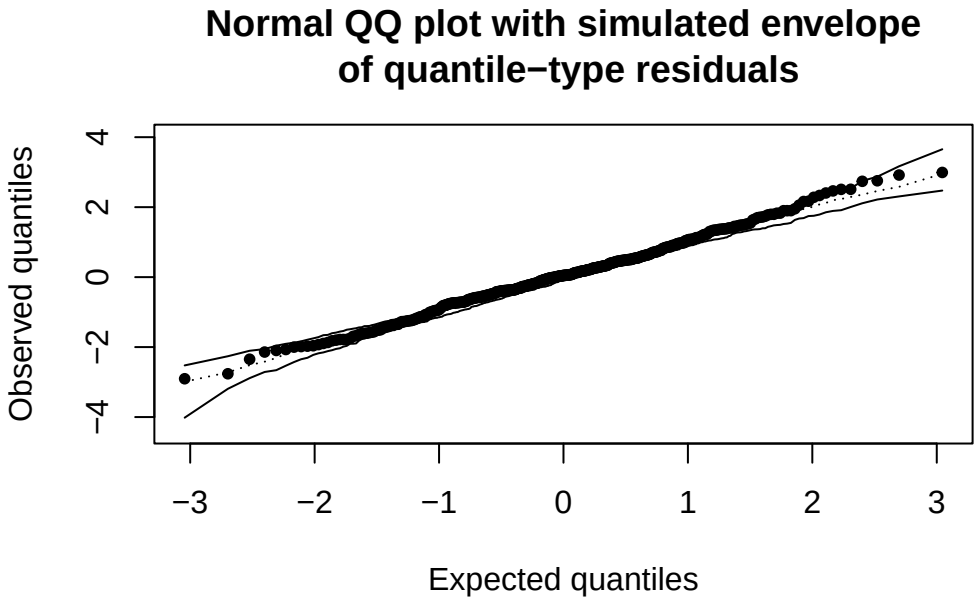
4.2.2 Modelo 2: MLG Binomial com função de ligação Probit

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.9061	0.2690	10.8043	0
dose	-0.0140	0.0012	-11.7180	0



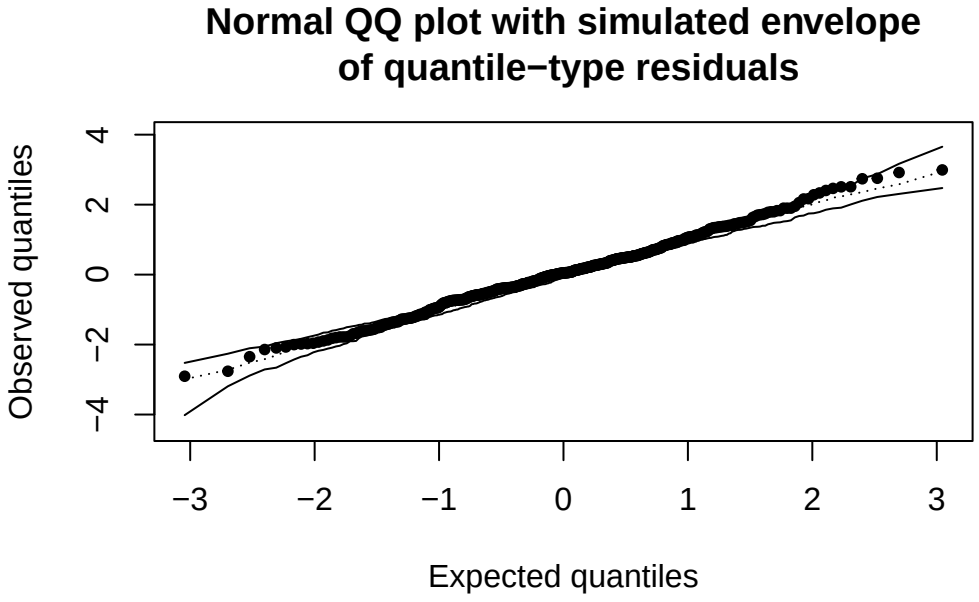
4.2.3 Modelo 3: MLG Binomial com função de ligação complemento log-log

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.1080	0.2332	9.0385	0
dose	-0.0128	0.0012	-10.8651	0



4.2.4 Modelo 4: MLG Binomial com função de ligação Cauchy

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	9.4315	1.7391	5.4233	0
dose	-0.0451	0.0084	-5.3399	0



4.2.5 Escolha do modelo e interpretação

Pelos gráficos e tabelas acima, vimos que para todas as funções de ligação houve forte significância da dose em explicar a morte do caramujo. Observando os envelopes simulados, qualquer um dos modelos poderia ser utilizado, visto que todos estão aderentes ao pressuposto. Desta forma, a escolha será pelo mais parcimonioso e de melhor interpretabilidade, que é o modelo GLM Binomial com função de ligação logito — O modelo logístico.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	5.5462	0.5653	9.8111	0
dose	-0.0265	0.0026	-10.3626	0

Para este modelo, observamos que para cada aumento em 1 unidade da dose, aumenta em $1 - exp^{-0.0265} \approx 0,0261 \rightarrow 2,61\%$ a chance de morte do caramujo.

Podemos também obter a estimativa da dose letal DL50, podendo ser calculada neste caso por $DL50 = \frac{\beta_0}{\beta_1} = \frac{5,5462}{-0,0265} \approx 209,29$. Logo, esta é a dose letal estimada que mata 50% dos caramujos!

4.3 Dose 3

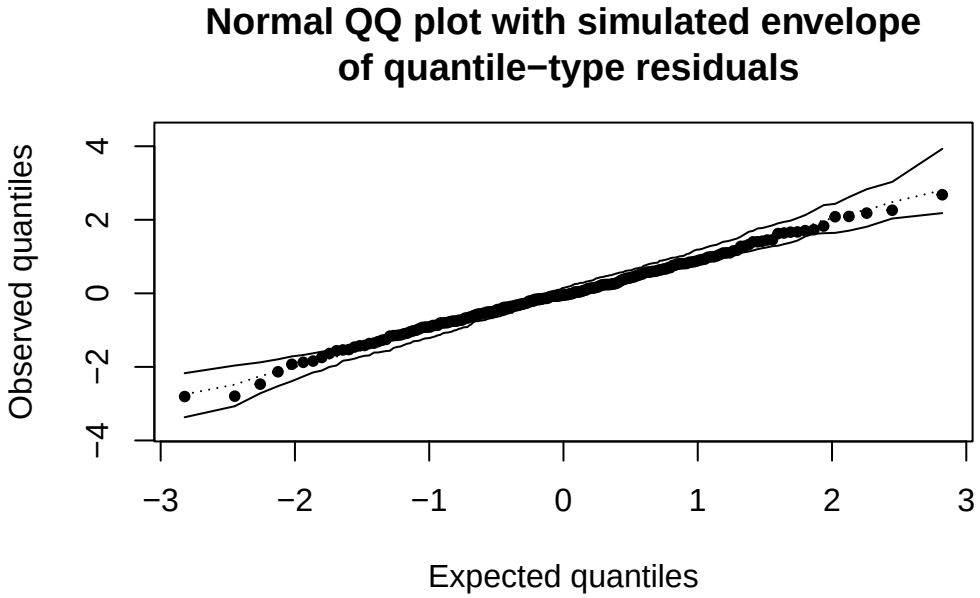
Neste estudo de caso, iremos investigar um experimento de dose-resposta conduzido para avaliar a influência de um extrato químico na morte de um determinado tipo de caramujo utilizando modelagem GLM Binomial com diversas funções de ligação para modelar o status (vivo/morto) de acordo com a dose do extrato aplicada ao caramujo.

	morto	vivo
0	2	28
0.01	4	26
0.015	4	26
0.02	11	19
0.03	9	20
0.04	14	16
0.05	24	6

Pela tabela, é fácil observar que o aumento da dose leva a morte dos caramujos. Estamos interessados em quantificar esta relação, portanto, irei testar diversos modelos GLM Binomial e, baseado no critério definido pelo enunciado da questão (envelope), irei interpretar o modelo que for o melhor ao final.

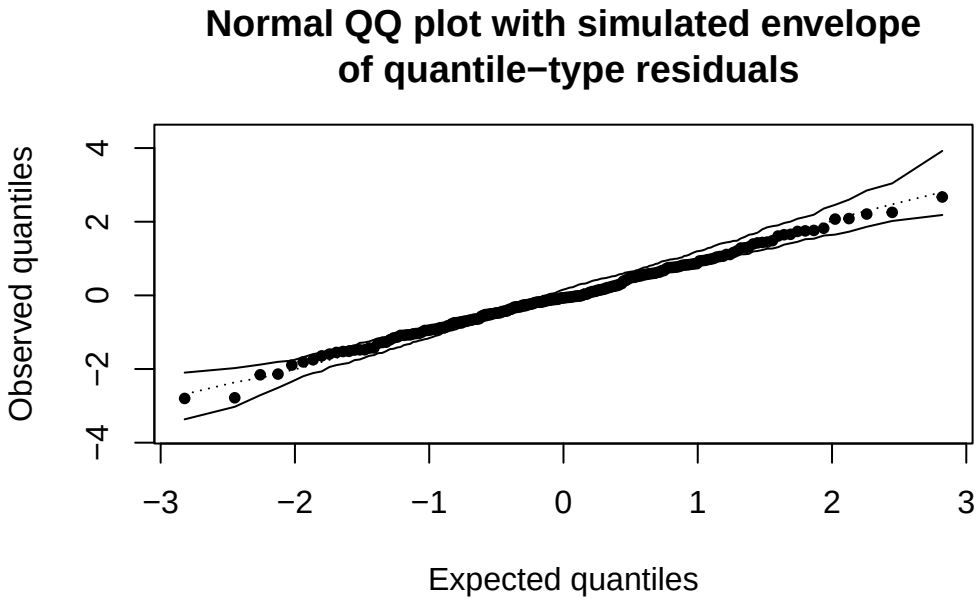
4.3.1 Modelo 1: Modelo logístico (binomial com função de ligação logit)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.6016	0.3662	7.1049	0
dose	-71.0948	11.3951	-6.2390	0



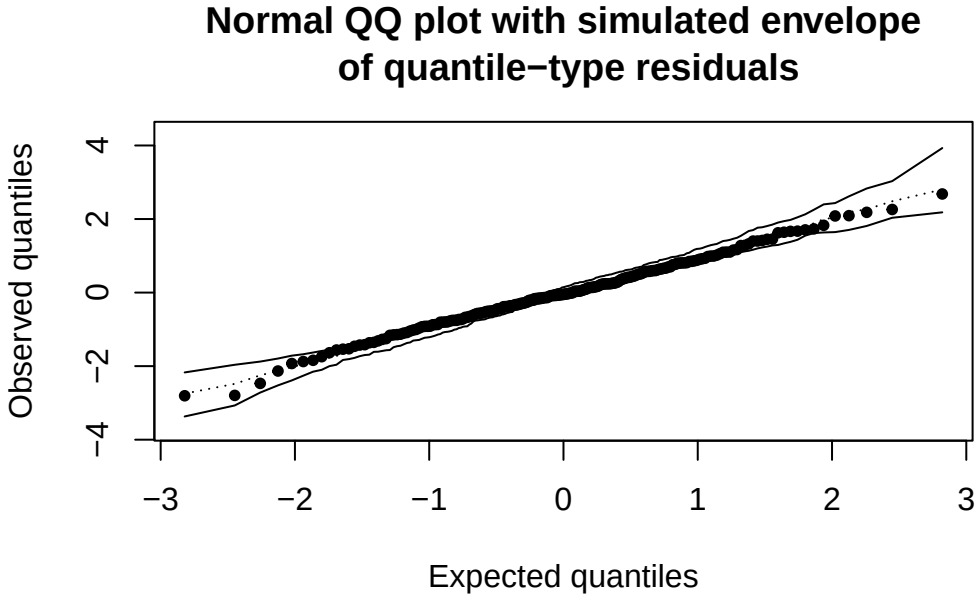
4.3.2 Modelo 2: MLG Binomial com função de ligação Probit

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.5453	0.1997	7.7388	0
dose	-42.4226	6.4375	-6.5899	0



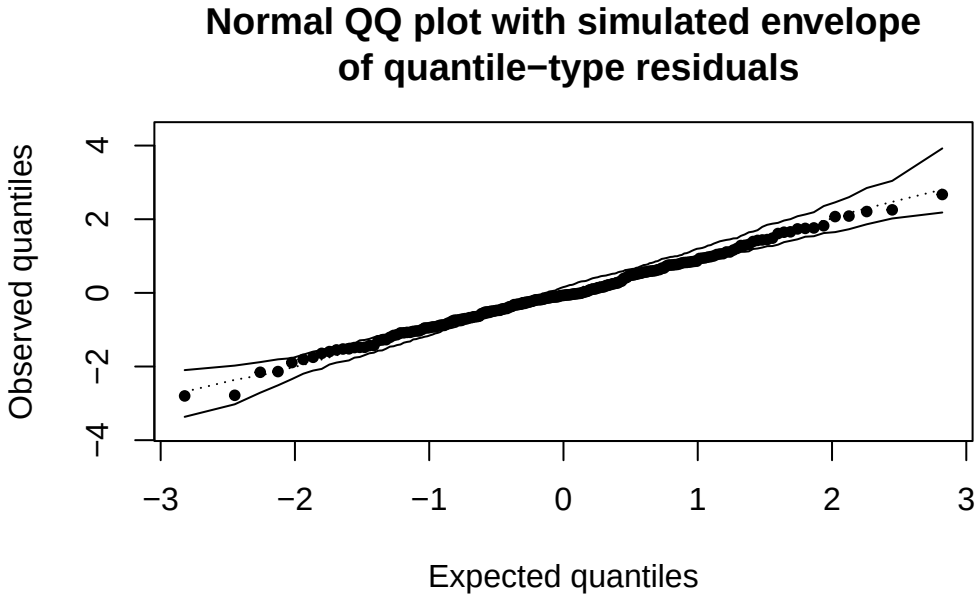
4.3.3 Modelo 3: MLG Binomial com função de ligação complemento log-log

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.1099	0.1694	6.5504	0
dose	-41.0982	6.5767	-6.2491	0



4.3.4 Modelo 4: MLG Binomial com função de ligação Cauchy

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.7688	0.5906	4.6883	0
dose	-73.2051	16.2180	-4.5138	0



4.3.5 Escolha do modelo e interpretação

Pelos gráficos e tabelas acima, vimos que para todas as funções de ligação houve forte significância da dose em explicar a morte do caramujo. Observando os envelopes simulados, qualquer um dos modelos poderia ser utilizado, visto que todos estão aderentes ao pressuposto. Desta forma, a escolha será pelo mais parcimonioso e de melhor interpretabilidade, que é o modelo GLM Binomial com função de ligação logito — O modelo logístico.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.6016	0.3662	7.1049	0
dose	-71.0948	11.3951	-6.2390	0

Para este modelo, observamos que para cada aumento em 1 unidade da dose, aumenta em $1 - exp^{-71.0948} = 1 \rightarrow 100\%$ a chance de morte do caramujo. Para este modelo, a magnitude da dose necessária para matar o caramujo é muito menor que para as soluções vegetais, visto que no exemplo a maior dose observada é 0,05

Podemos também obter a estimativa da dose letal DL50, podendo ser calculada neste caso por $DL50 = \frac{\beta_0}{\beta_1} = \frac{2,601569}{-71,094836} = 0,04$. Logo, esta é a dose letal estimada que mata 50% dos caramujos!

5 Referências

Materiais de aula — Modelos lineares generalizados. UnB, 2º/2024. Prof.ª Dr.ª Terezinha Kessia de Assis Ribeiro.

https://www.ime.usp.br/~giapaula/textoregressao.htm

Obs: Lista produzida utilizando R e Quarto documents. Os códigos se encontram disponíveis em https://github.com/penasta/modelos-lineares-generalizados/blob/main/rdocs/lista1.qmd