Universidade de Brasília

DEPARTAMENTO DE ESTATÍSTICA

21 de fevereiro de 2025

Lista 2

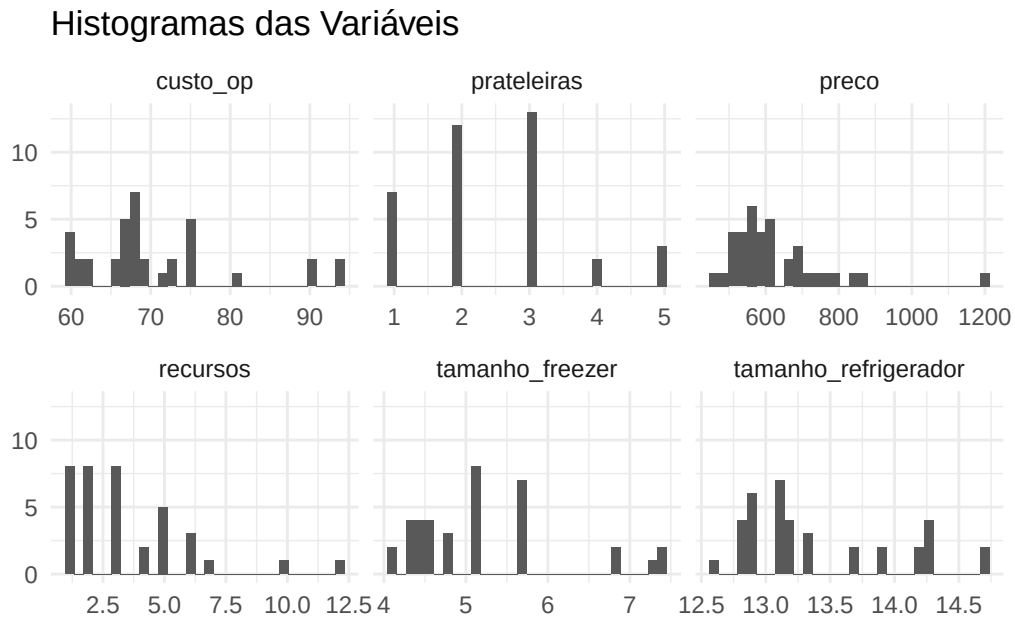
Prof. Drª. Terezinha Kessia de Assis Ribeiro

Modelos Lineares Generalizados

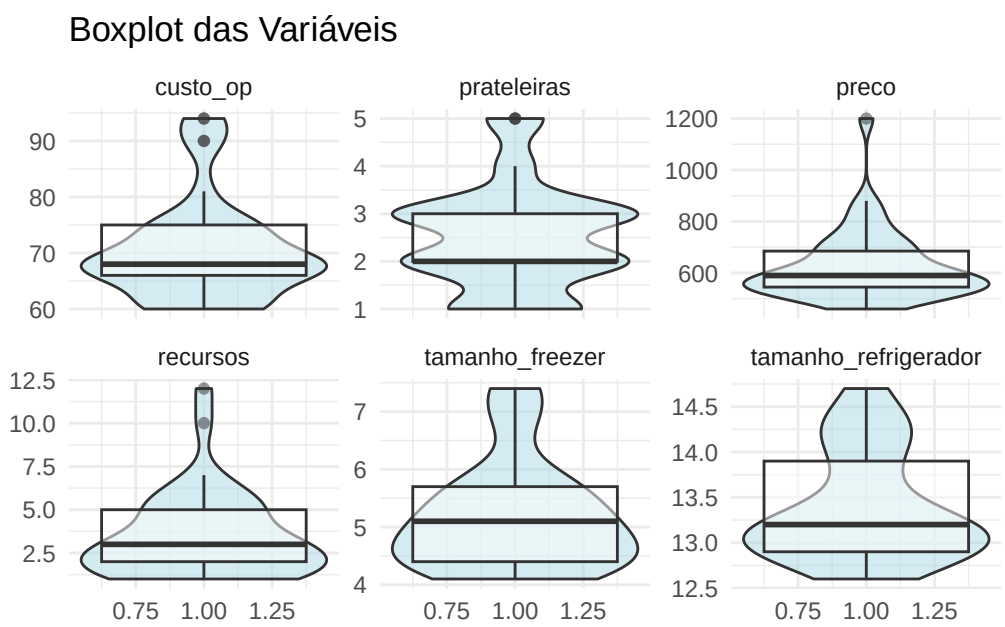
Aluno: Bruno Gondim Toledo | Matrícula: 15/0167636

a)

Neste problema, iremos estudar a relação do preço de refrigeradores com características pertinentes destes eletrodomésticos. Nossa variável resposta, o preço do refrigerador (preco) será explicada por características do refrigerador, sendo elas: o tamanho do refrigerador em pés cúbicos (tamanho_refrigerador), o tamanho do compartimento do freezer (tamanho_freezer), a quantia média de dinheiro gasta por ano para operar o refrigerador (custo_op) e o número de recursos do refrigerador (recursos).



Observando os histogramas das variáveis, podemos observar que, naturalmente, todas elas são quantitativas positivas, algumas sendo contagens e outras de perfil mais contínuo. Deste gráfico, podemos dizer que possivelmente iremos necessitar de uma distribuição com suporte nos positivos, $Y \sim (0, \infty]$.



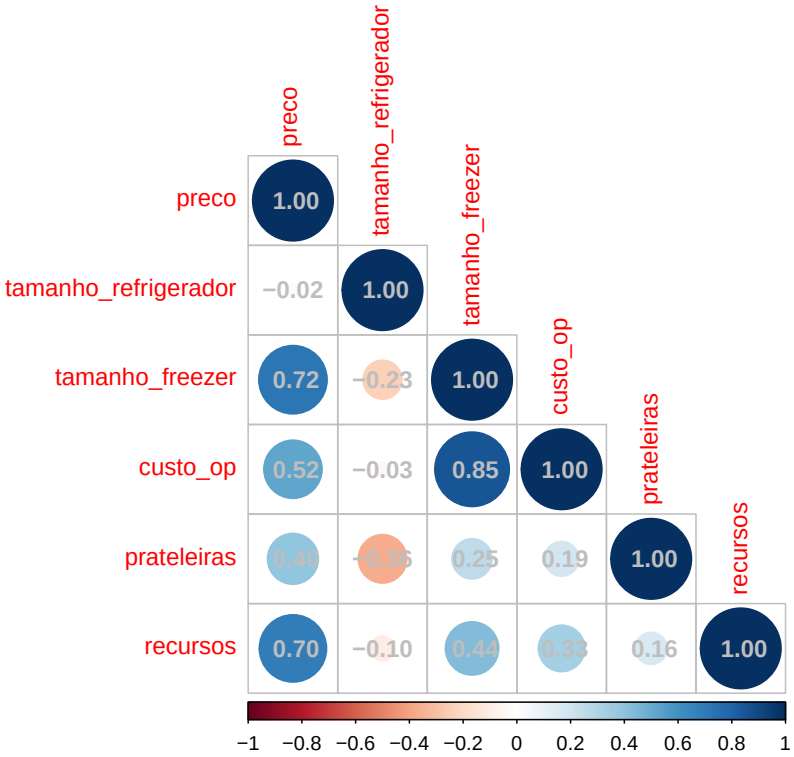
Os boxplots sobrepostos aos violinos são interessantes para visualizar nossas observações. Nosso conjunto de dados é pequeno (37 observações), e em geral as variáveis apresentam assimetria à direita, sendo possível também observar alguns pontos candidatos a outlier no conjunto.

Variável	Média	D.P.	Min.	q25	Mediana	q75	Max.
preco	626.35	139.79	460.0	545.0	590.0	685.0	1200.0
tamanho_refrigerador	13.40	0.60	12.6	12.9	13.2	13.9	14.7
tamanho_freezer	5.18	0.94	4.1	4.4	5.1	5.7	7.4
custo_op	70.51	9.14	60.0	66.0	68.0	75.0	94.0
prateleiras	2.51	1.12	1.0	2.0	2.0	3.0	5.0
recursos	3.46	2.51	1.0	2.0	3.0	5.0	12.0

Pela Tabela acima, podemos observar quantitativamente as relações percebidas no gráfico anterior. A variável preço por exemplo, apresenta um alto desvio padrão comparada as demais variáveis. Isto se dá também por uma questão de escala, mas podemos notar que o valor máximo é quase o dobro dos valores presentes até o terceiro quartil. Demais variáveis apresentam média e mediana próximas, também sendo possível observar possíveis valores de máximo discrepantes para as covariáveis recursos, custo de operação e prateleiras.

Mín.	1º Quartil	Mediana	Média	3º Quartil	Máx.	Assimetria	Curtose
460	545	590	626.35	685	1200	2.13	8.86

Analisando com maior atenção a variável preço pela Tabela acima, notamos que o coeficiente de curtose é bastante elevado, apresentando um excesso de curtose de 5,86; indicando um possível problema com a modelagem gaussiana.



É bastante pertinente calcular correlações antes de realizar uma modelagem de regressão. Pelo gráfico acima, podemos observar que as covariáveis tamanho do freezer, custo de operação, número de prateleiras e recursos do refrigerador tem correlação positiva com a variável preço. A covariável tamanho do refrigerador apresenta uma baixíssima correlação negativa com o preço do refrigerador (o que é um pouco estranho). Além disso, notamos que as covariáveis custo de operação e tamanho do freezer apresentam forte correlação positiva, indicando um possível problema de multicolinearidade se ambas as variáveis forem inseridas no modelo. As covariáveis tamanho do freezer e quantidade de recursos também apresentam uma correlação positiva considerável, e também devem ser avaliadas com cautela.

Portanto, de todo o observado, notamos que possivelmente as covariáveis tamanho do freezer, custo de operação, número de prateleiras e recursos do refrigerador apresentam

tendência de impacto positivo no preço do refrigerador, numa primeira análise simplificada. Estas relações serão melhor quantificadas nas análises seguintes, isto é, na modelagem regressiva.

b)

Farei a primeira modelagem como manda a práxis, tentando ajustar um modelo gaussiano homocedástico saturado.

Estimativas do modelo

	Estimativa	Erro padrão	Valor t	Pr(>|t|)
(Intercept)	-797.81	271.41	-2.94	0.006
tamanho_refrigerador	76.50	19.44	3.93	0.000
tamanho_freezer	137.38	23.76	5.78	0.000
custo_op	-6.96	2.28	-3.06	0.005
prateleiras	37.94	9.89	3.84	0.001
recursos	23.76	4.51	5.27	0.000

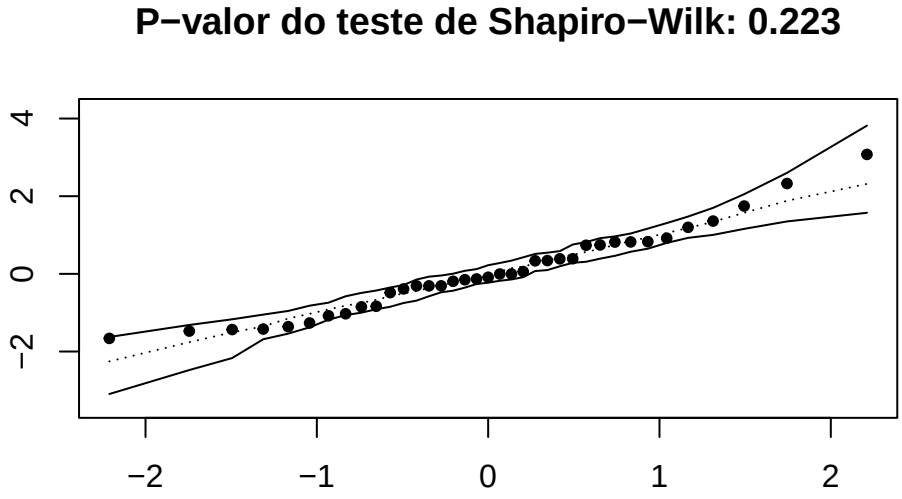
A Tabela acima mostra as estimativas dos parâmetros produzidas para o modelo gaussiano homocedástico. Foi realizado também o procedimento stepAIC, que indica pela manutenção do modelo saturado.

Observando os resultados, notamos que todos os coeficientes foram marginalmente significativos sob (quase) qualquer nível de significância. O modelo apresentou também valor $R^2 = 0{,}838$ e R^2 ajustado = 0,812; e estes resultados sob uma interpretação ingênua sugeririam que já temos um modelo excelente que explica bem os dados. Entretanto, é importante observar com cautela os resultados deste modelo. Notamos que a covariável custo de operação apresentou uma estimativa negativa, o que não faz sentido se comparado ao gráfico de correlações produzido na análise descritiva. Possivelmente, como esta covariável tem correlação alta com a covariável tamanho do freezer, isto esteja atrapalhando as estimativas do modelo. O intercepto negativo numa alta escala também pode ser problemático, visto que não faz sentido para este dado um preço negativo: ele precisa ser estritamente positivo! E para um novo dado não presente neste conjunto, uma combinação de covariáveis poderia levar a modelagem do preço desta observação como sendo um preço negativo, o que não faria sentido algum. Além disso, a covariável tamanho do refrigerador, que na análise descritiva também aparecia não correlacionada/fracamente negativamente correlacionada apresenta uma significância altíssima, bem como um alto valor positivo na estimativa do parâmetro.

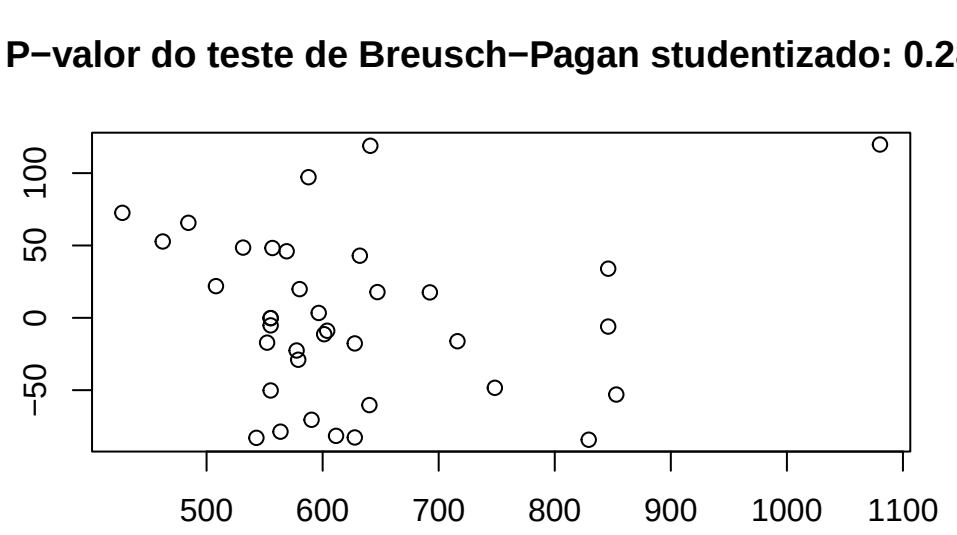
Para este modelo, mantido constante todos os outros coeficientes, o aumento de um pé cúbico do refrigerador aumentaria a média do preço em 76,5 U.M. Analogamente, o aumento de uma unidade no tamanho do freezer aumenta em 137,38 U.M. a média do preço. O aumento de 1 U.M. no custo de operação diminuiria em 7 U.M. o preço médio do refrigerador. O acréscimo de 1 prateleira aumenta em 37,94 U.M. o preço médio do refrigerador. O acréscimo de 1 recurso no refrigerador aumenta seu preço médio em 23,76 U.M.

c)

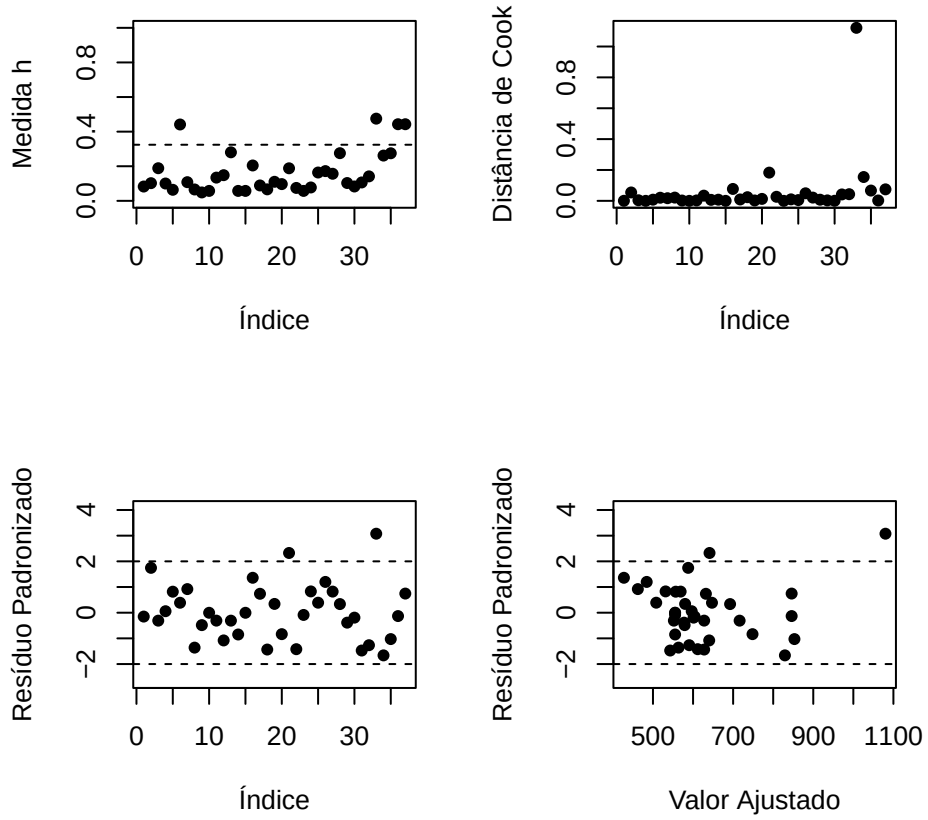
Podemos também realizar algumas etapas de diagnóstico para observar outros possíveis problemas ou qualidades neste modelo.



Analisando os resíduos sob envelope simulado, notamos que todos estão contidos nos limites do envelope simulado. Além disso, quanto ao pressuposto de normalidade dos resíduos, o modelo parece atender.



Observando o gráfico dos resíduos contra os valores ajustados, notamos que o pressuposto de homocedasticidade aparenta ser respeitado, porém este resultado deve ser observado com cautela, visto a quantidade reduzida de observações neste conjunto de dados, o que pode prejudicar o poder do teste de Breusch-Pagan, bem como a interpretação visual.



Realizando um diagnóstico mais aprofundado, notamos que existem possíveis pontos de alavancagem no modelo, sendo eles os pontos 6,33,36 e 37. Voltando ao conjunto de dados, notamos que o ponto 6 apresenta uma grande quantidade de recursos no refrigerador: 10 recursos, enquanto as demais variáveis estão próximas à média. O ponto 33 se destaca como o ponto de maior preço (1200 U.M.), enquanto os pontos 36 e 37 também apresentam valor alto no preço (840 e 880 U.M., respectivamente), indicando serem possíveis outliers.

Analisando o gráfico de influência pelas distâncias de Cook, notamos que existe um ponto significativamente aberrante ante aos demais, sendo ele justamente o ponto 33 — o de maior preço (1200 U.M.). Isso praticamente o define como sendo um ponto aberrante, visto ser detectado em ambas as análises. Caso a intenção fosse manter o modelo normal heterocedástico, o ideal seria remover este ponto e observar o ajuste sem ele.

Os outros dois gráficos indicam um possível problema de sobredispersão, ainda que um tanto difícil de visualizar pela pouca quantidade de pontos, mas aparenta haver uma leve tendência de aumento da dispersão. Indicam também uma aparente independência, o que não parece ser realmente o maior problema deste modelo.

Portanto, vemos que não aparentam haver afastamentos sérios das suposições do modelo normal homocedástico para este modelo. Ainda assim, considerando os coeficientes observados, elesapresentam um padrão de problema interpretativo, visto que as interpretações são conflituosas com o observado na análise descritiva, e seria esperado alguns resultados diferentes para este modelo. Além disso, notamos que existem problemas de valores influentes e pontos aberrantes. Também existe um problema conceitual, quanto a natureza da variável resposta, que é na prática estritamente posisitiva, ante a modelagem gaussiana que sempre retornará resultados contidos nos reais.

Concluindo a avaliação deste modelo, podemos dizer que a dependender do problema específico, este modelo até poderia ser utilizado. Se considerarmos que esta é uma boa amostra, representativa da população, e não seja do interesse do possuidor deste modelo interpretar a rigor os parâmetros deste modelo, considerando também uma baixa existência de outliers na população — visto que este modelo apresenta problemas de robustez — isto é, se quiséssemos por exemplo embarcar este modelo numa aplicação leve, onde valeria mais uma previsão aproximada, com um modelo simples e de baixa complexidade computacional — como por exemplo em um chip, ou rodando no backend de um aplicativo web como indicador para uma propaganda para um perfil de consumidor, por exemplo — não haveria um grande problema preditivo para a maioria dos casos em média. Entretanto, o objetivo aqui é justamente explorar uma série de distribuições que poderiam se ajustar melhor a estes dados, levando mão também da possibilidade de modelar a dispersão destes dados e não só a média, portanto não faria sentido analisar todas as possibilidades de melhorar o ajuste normal homocedástico, como por exemplo realizar várias possibilidades de remoção de covariáveis conflituosas de interpretação confusa, bem como a repetição da modelagem com a remoção de pontos aberrantes.

Partiremos então para uma modelagem mais sofisticada, utilizando do pacote *gamlss*

d)

MLG Duplo: Modelo normal heterocedástico

Dada as tentativas anteriores com o modelo normal homocedástico, é possível que apenas modelando a precisão conjuntamente, o modelo gaussiano já entregue resultados mais satisfatórios, ainda que as etapas de diagnóstico não tenham indicado uma clara necessidade de modelagem da precisão.

Para isso, será utilizada a ligação canônica para o parâmetro μ e ligação log para o parâmetro σ^2.

Estimativas para μ

	Estimativa	Erro padrão	Valor t	Pr(>|t|)
X.Intercept.	-1413.68	194.47	-7.27	0.000
tamanho_refrigerador	112.58	11.84	9.51	0.000

	Estimativa	Erro padrão	Valor t	Pr(>|t|)
tamanho_freezer	135.51	12.24	11.07	0.000
custo_op	-4.47	0.84	-5.31	0.000
prateleiras	52.43	6.49	8.08	0.000
recursos	-4.48	5.66	-0.79	0.436

Estimativas para σ^2

	Estimativa	Erro padrão	Valor t	Pr(>|t|)
X.Intercept..1	40.58	11.19	3.63	0.001
tamanho_refrigerador.1	-2.38	0.84	-2.82	0.009
tamanho_freezer.1	0.12	0.68	0.18	0.858
custo_op.1	-0.02	0.07	-0.26	0.796
prateleiras.1	-0.78	0.34	-2.31	0.029
recursos.1	0.43	0.12	3.54	0.002

Notamos que ao modelar conjuntamente o MLG duplo gaussiano, existem parâmetros com significância para a variância, bem como os parâmetros fornecidos para a média também são diferentes do modelo homocedástico. Existem, entretando, covariáveis que marginalmente não estão apresentando significância neste modelo saturado, tanto para μ quanto para σ^2. Realizando o procedimento stepAIC e também removendo coeficientes manualmente, chegamos a um modelo com mais parcimônia e significância:

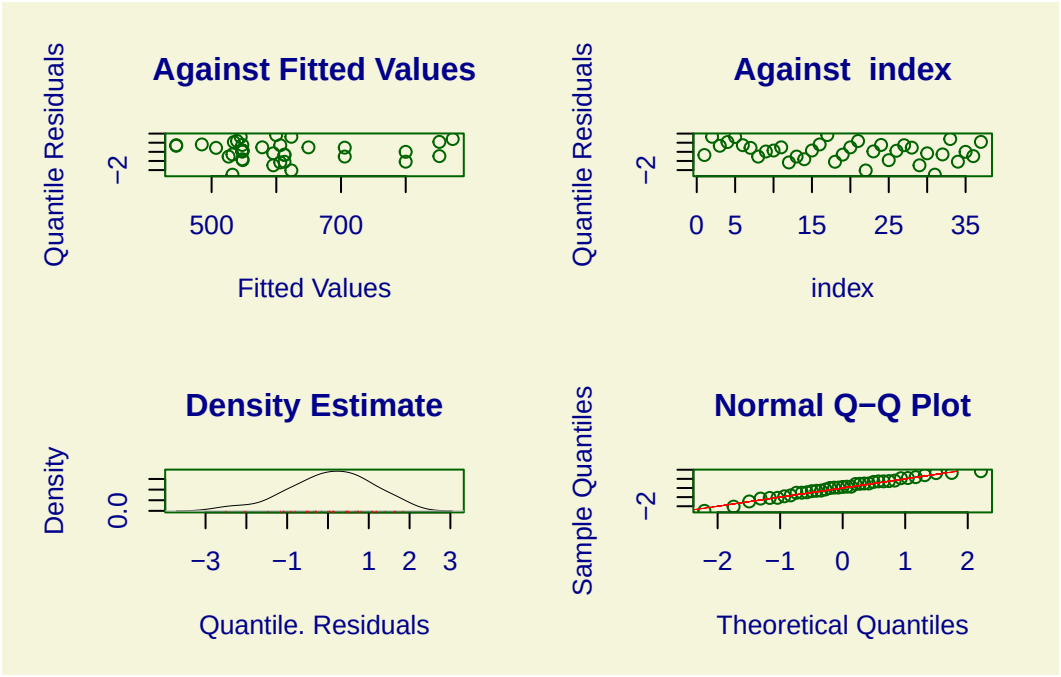
Estimativas para μ

	Estimativa	Erro padrão	Valor t	Pr(>|t|)
X.Intercept.	-1306.43	141.23	-9.25	0
tamanho_refrigerador	106.15	9.02	11.77	0
tamanho_freezer	129.93	9.85	13.20	0
custo_op	-4.40	0.93	-4.73	0
prateleiras	49.85	4.96	10.04	0

Estimativas para σ^2

	Estimativa	Erro padrão	Valor t	Pr(>|t|)
X.Intercept..1	35.84	7.54	4.76	0.000
tamanho_refrigerador.1	-2.09	0.53	-3.96	0.000
prateleiras.1	-0.70	0.30	-2.33	0.027
recursos	0.40	0.10	4.01	0.000

Notamos que este modelo, com as covariáveis tamanho do refrigerador, tamanho do freezer, custo de operação e quantidade de prateleiras explicando a média, e tamanho do refrigerador, quantidade de prateleiras e quantidade de recursos modelando a variância, obtém significância marginal para todas as covariáveis, e aparenta ser um modelo parcimonioso. Ainda não interpretando os coeficientes das estimativas dos parâmetros, mas simplesmente comparando com os coeficientes obtidos no modelo normal homocedástico, ainda existem valores que não fazem sentido, como o custo de operação com valor negativo. Portanto, acredito que os resultados obtidos com o MLG duplo gaussiano heterocedástico ainda não é o melhor possível de ser obtido para este conjunto de dados. Ainda assim, farei algumas etapas de diagnóstico deste modelo, afim de identificar possíveis pontos positivos e negativos deste.

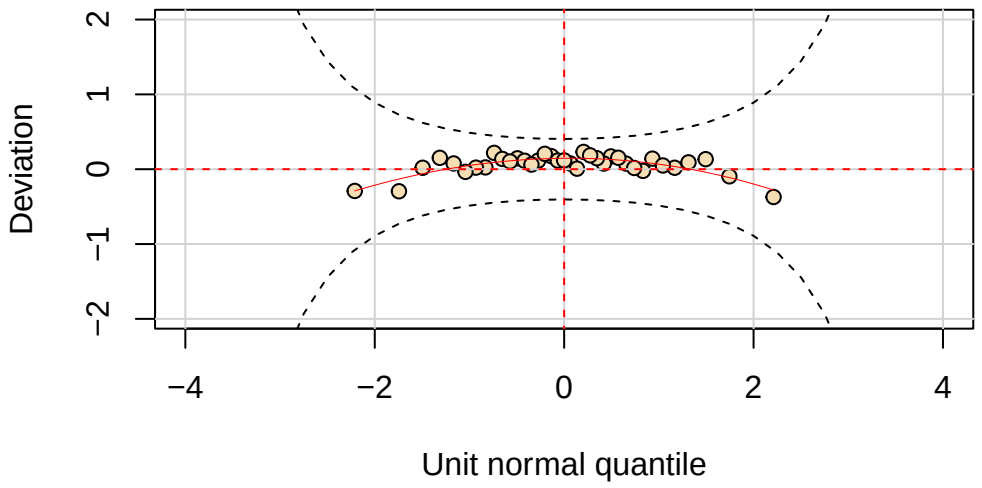


Destes gráficos de diagnósticos, o maior problema aparenta ser assimetria do modelo. Para avaliar melhor esta questão, irei utilizar o wormplot, juntamente com a seguinte tabela, de Gonçalves (2024), para avaliar o modelo:

Forma	Resíduos	Interpretação
Concentração Acima da Origem	Média Muito Alta	Parâmetro de Localização Subestimado
Concentração Abaixo da Origem	Média Muito Baixa	Parâmetro de Localização Sobrestimado
Inclinação Positiva	Variância Muito Alta	Parâmetro de Escala Muito Baixo
Inclinação Negativa	Variância Muito Baixa	Parâmetro de Escala Muito Alto

Forma	Resíduos	Interpretação
Forma de U	Assimetria à Direita	Excesso de Assimetria à Esquerda
Forma de U Invertido	Assimetria à Esquerda	Excesso de Assimetria à Direita
Forma de S com Esquerda para Baixo	Leptocúrticos	Caudas Muito Leves
Forma de S com Esquerda para Cima	Platicúrticos	Caudas Muito Pesadas

E este é o wormplot deste ajuste:



Comparando a Figura com as indicações da Tabela, notamos que o formato de U invertido, o que indica problemas com excesso de assimetria à direita. Este diagnóstico faz todo sentido, visto que estou tentando modelar dados com imagem em $(0, \infty)$ utilizando uma distribuição com imagem nos reais $(-\infty, \infty)$. Portanto, o ideal é buscar uma distribuição que se encaixe melhor nestas especificidades, ao invés de seguir tentando o modelo normal

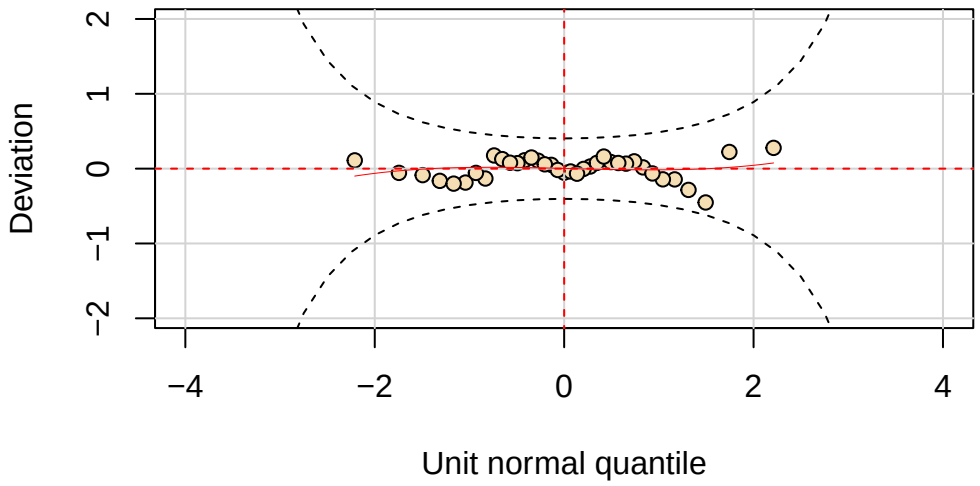
Modelo gama

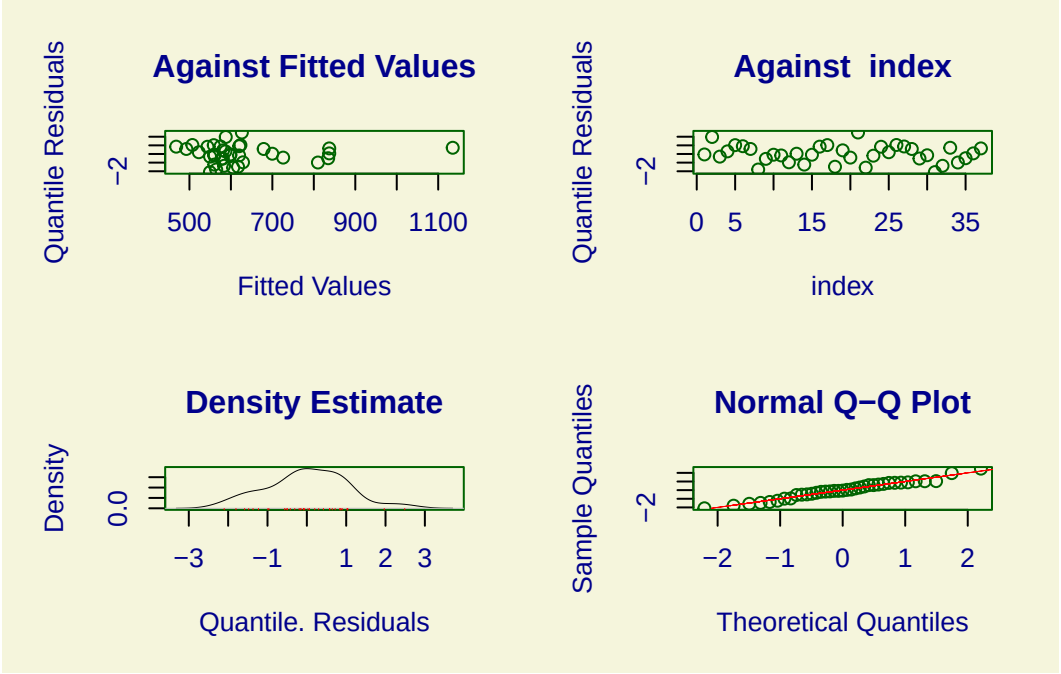
O modelo gama é uma possibilidade interessante, visto que esta distribuição têm suporte nos positivos, o que é compatível com a natureza destes dados.

Tentarei uma modelagem MLG gama, utilizando o pacote gamlss, com ligação log para a média — e sem tentar adicionar parâmetros para a precisão num primeiro momento. Irei iniciar o ajuste com o modelo saturado.

	Estimativa	Erro padrão	Valor t	Pr(>|t|)
X.Intercept.	4.40	0.36	12.24	0.000
tamanho_refrigerador	0.11	0.03	4.25	0.000
tamanho_freezer	0.20	0.03	6.17	0.000
custo_op	-0.01	0.00	-3.29	0.003
prateleiras	0.05	0.01	4.06	0.000
recursos	0.03	0.01	5.07	0.000
X.Intercept..1	-2.51	0.12	-21.02	0.000

Observando as significâncias marginais, notamos que a modelagem saturada utilizando a gama indica significância marginal para todas as covariáveis. Seguimos com a interpretação estranha do coeficiente relacinado a covariável custo de operação com sinal negativo. O procedimento stepAIC indica também pela manutenção deste modelo saturado. Como na gama estamos naturalmente modelando a precisão ainda que sem adicionar parâmetros para σ, notamos que o intercepto também é significativo. Este pode ser um indício de que seja interessante adicionar covariáveis na precisão para modelá-la também de forma a melhorar o ajuste do modelo.





Os gráficos de diagnóstico mostram que já temos uma melhora em relação a utilização do modelo gaussiano. o wormplot não detecta um padrão significativo dente os que constam na Tabela de interpretação deste gráfico, enquanto os demais gráficos de diagnóstico mostram que ainda existem problemas na forma da distribuição, mas que no geral já é um modelo melhor ajustado aos dados.

MLG Duplo: Gama

Tentarei uma abordordagem MLG duplo, adicionando covariáveis para melhor modelar a precisão afim de observar se o ajuste do modelo melhora ainda mais.

Estimativas para μ

	Estimativa	Erro padrão	Valor t	Pr(>|t|)
X.Intercept.	5.09	0.62	8.26	0.000
tamanho_refrigerador	0.07	0.04	2.00	0.057
tamanho_freezer	0.13	0.04	3.43	0.002
custo_op	-0.01	0.00	-4.78	0.000
prateleiras	0.08	0.01	6.51	0.000
recursos	0.02	0.01	2.06	0.049

Estimativas para σ^2

	Estimativa	Erro padrão	Valor t	Pr(>|t|)
X.Intercept..1	-1.55	4.99	-0.31	0.759
tamanho_refrigerador.1	-0.13	0.34	-0.40	0.695
tamanho_freezer.1	0.89	0.36	2.46	0.021
custo_op.1	-0.04	0.03	-1.32	0.199
prateleiras.1	-0.44	0.13	-3.33	0.003
recursos.1	0.02	0.08	0.27	0.786

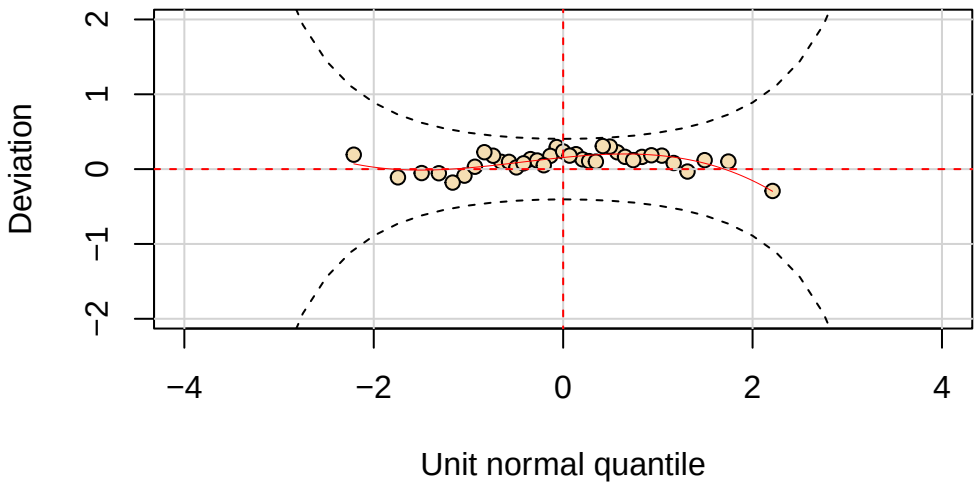
Notamos que existem covariáveis significativas e não significativas marginalmente quando testamos o modelo com todas as covariáveis modelando tanto μ quanto a precisão. O procedimento stepAIC indica pela remoção da covariável recursos para modelar o parâmetro μ, e deixar todas as covariáveis para modelar a precisão. Neste caso, chegaríamos a estimativas de parâmetros que ainda assim contaria com significâncias marginais bastante precárias, e seguiríamos com um modelo de difícil interpretação, principalmente pela covariável custo de operação seguir com valor negativo. A análise descritiva já havia indicado possíveis problemas de multicolinearidade, portanto acredito que este seja o passo em que devemos forçar mão de obter um modelo parcimonioso e interpretável, portanto testarei colocar covariáveis que acredito que sejam mais pertinentes para os parâmetros, levando em consideração tudo que fora analisado até agora. Portanto, utilizarei as covariáveis tamanho do refrigerador, tamanho do freezer e número de prateleiras para modelar o parâmetro μ, enquanto que para o parâmetro de precisão utilizarei as covariáveis tamanho do refrigerador, quantidade de prateleiras e quantidade de recursos, utilizando a distribuição gama com função de ligação log para ambos os parâmetros. Veremos como fica este ajuste

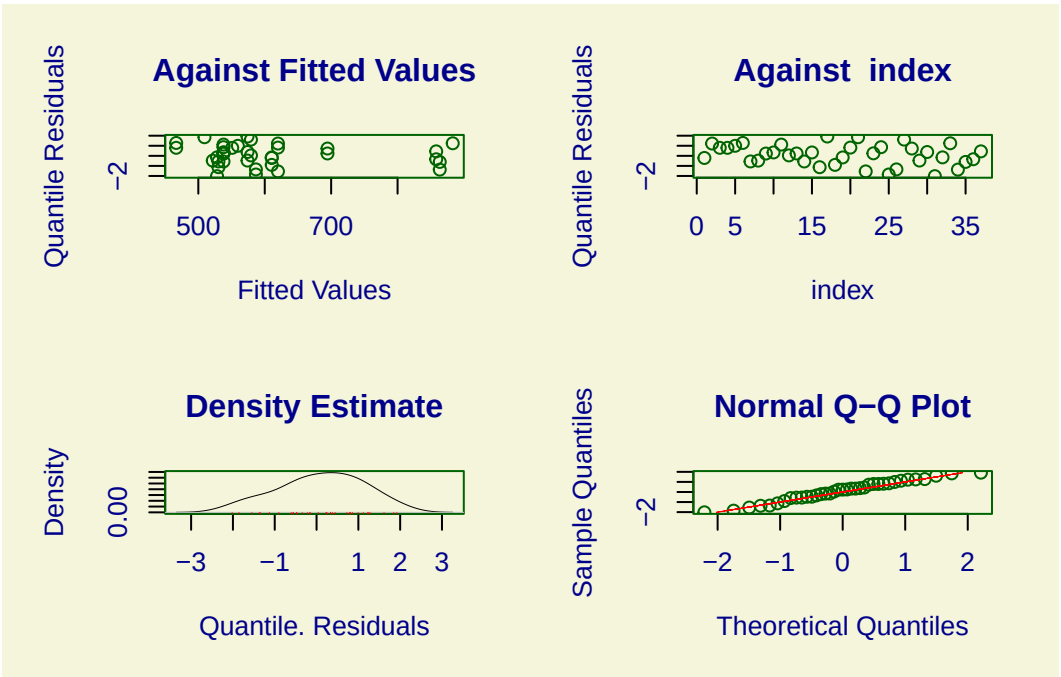
Estimativas para μ

	Estimativa	Erro padrão	Valor t	Pr(>|t|)
X.Intercept.	3.41	0.36	9.46	0
tamanho_refrigerador	0.15	0.02	6.33	0
tamanho_freezer	0.14	0.01	9.82	0
prateleiras	0.09	0.01	6.88	0

Estimativas para σ^2

	Estimativa	Erro padrão	Valor t	Pr(>|t|)
X.Intercept..1	7.67	4.28	1.79	0.083
tamanho_refrigerador.1	-0.74	0.30	-2.45	0.021
prateleiras.1	-0.23	0.13	-1.82	0.078
recursos	0.13	0.05	2.62	0.014





Analisando este modelo, notamos que não existem grandes fugas as suposições nas etapas de diagnóstico, e este parece ser um modelo aderente e coerente para estes dados, além de ainda contar com interpretação relativamente fácil. É do objetivo desta lista percorrer diversas possibilidades de modelagem para estes dados, e por este motivo seguirei testando ajustes em seguida. Entretanto, já é possível interpretar este modelo, pois numa situação real acredito que seria possível optar já por este modelo.

Portanto, como utilizamos ligação log e para a gama o parâmetro μ é a média mesmo,

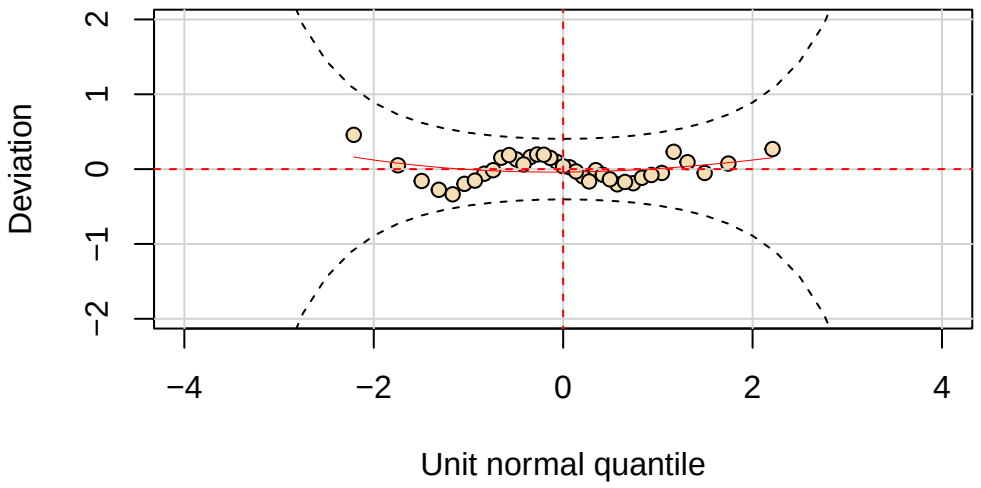
então a interpretação para estes coeficientes está associado a um aumento percentual na variável resposta mesmo. Desta forma, podemos dizer que, mantido constante todas as demais covariáveis, existe um aumento percentual da média do preço em $e^{0,15} \approx 1,1618 \rightarrow 16,18\%$ para cada pé cúbico extra do refrigerador. Existe um aumento percentual da média do preço em $e^{0,14} \approx 1,1503 \rightarrow 15,03\%$ para cada aumento no tamanho do freezer. Existe também um aumento percentual da média do preço em $e^{0,09} \approx 1,0942 \rightarrow 9,42\%$ para cada prateleira extra do refrigerador.

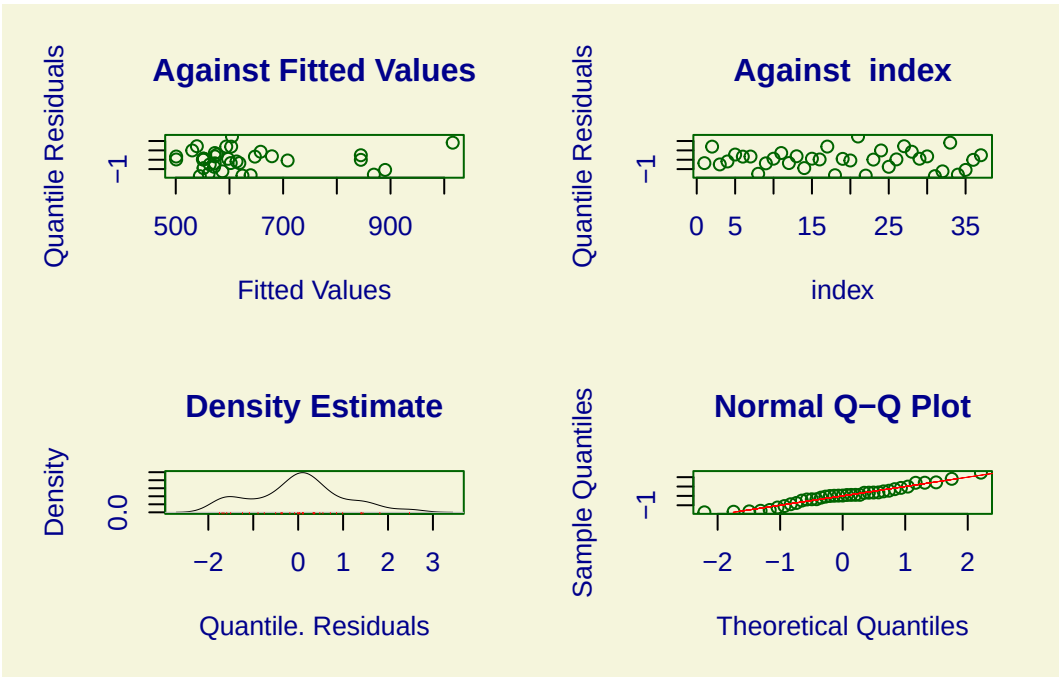
Modelo Birnbaum-Saunders reparametrizado pela média

Resolvi trazer aqui a possibilidade de utilizar o modelo RBS (Birnbaum-Saunders reparametrizado pela média), pois é um modelo que estudei recentemente para meu seminário desta disciplina. Este modelo é adequado para esta natureza de dados (suporte nos positivos), e a reparametrização pela média e implementação utilizando o framework do gamlss facilita bastante tanto a modelagem quanto a interpretação dos parâmetros. Tentarei primeiramente a modelagem somente da média utilizando a RBS, utilizando todas as covariáveis menos custo de operação (pois num primeiro teste, a estimativa do parâmetro assiciada à esta covariável seguia com sinal negativo como nos modelos anteriores, o que dificulta a interpretação. Portanto, se for possível obter um ajuste satisfatório sem utilizar esta covariável, seria preferível).

	Estimativa	Erro padrão	Valor t	Pr(>|t|)
X.Intercept.	-716.09	289.98	-2.47	0.019
tamanho_refrigerador	58.02	18.90	3.07	0.004
tamanho_freezer	76.27	14.19	5.37	0.000
prateleiras	38.36	10.38	3.69	0.001
recursos	21.03	4.58	4.59	0.000
X.Intercept..1	5.42	0.23	23.31	0.000

Os coeficientes apresentam significância marginal para todos os casos, e os valores dos parâmetros são de fácil interpretação, considerando que neste caso estamos modelando a média com função de ligação identidade — portanto a interpretação seria direta. Entretando, é importante observar os diagnósticos deste modelo antes de qualquer coisa.





O wormplot mostra que os resíduos tem um formato de U, problema inverso ao observado no modelo gaussiano heterocedástico. Neste caso, o gráfico está indicando problema de assimetria à esquerda.

Os outros gráficos de diagnóstico sugerem também o problema da assimetria, bem como menor aderência dos resíduos quantílicos ao quantis teóricos da distribuição assintótica (normal).

Ou seja, este modelo até poderia ser utilizado, mas o modelo anterior (gama duplo) cumpriu com maior rigor os diagnósticos.

Modelo duplo: Birnbaum-Saunders reparametrizado pela média

Podemos ainda tentar modelar a precisão, para tentar obter um ajuste melhor utilizando a RBS.

Para este modelo, foi testada a utilização de todas as covariáveis para modelagem de ambos os parâmetros, porém observando a significância marginal e utilizando o procedimento stepAIC, selecionou-se um modelo mais parcimonioso, contendo as covariáveis tamanho do freezer, número de prateleiras e quantidade de recursos para modelagem de μ, e as covariáveis tamanho do refrigerador, tamanho do freezer e custo de operação para modelagem do parâmetro de precisão.

Além disso, utilizou-se aqui a função de ligação log tanto para o parâmetro μ quanto para a precisão, afim de obter um melhor modelo segundo testes realizados anteriormente.

Ao fim, estes foram os coeficientes das estimativas dos parâmetros associados à cada uma das covariáveis:

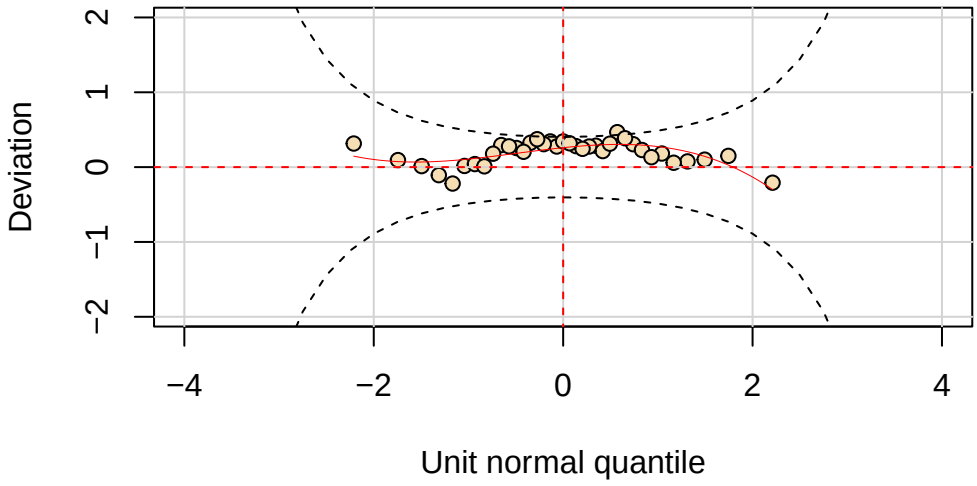
Estimativas para μ

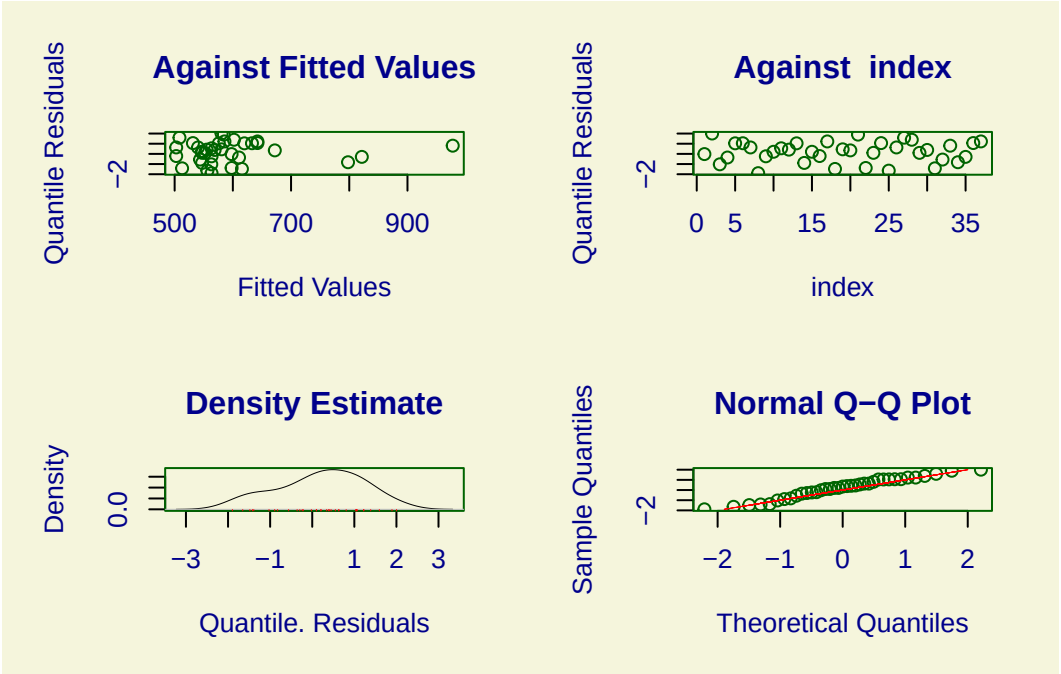
	Estimativa	Erro padrão	Valor t	Pr(>|t|)
X.Intercept.	5.85	0.08	73.20	0.000
tamanho_freezer	0.06	0.02	2.52	0.017
prateleiras	0.06	0.02	2.99	0.006
recursos	0.03	0.01	5.43	0.000

Estimativas para σ^2

	Estimativa	Erro padrão	Valor t	Pr(>|t|)
X.Intercept..1	18.79	5.86	3.21	0.003
tamanho_refrigerador	-1.02	0.43	-2.36	0.025
tamanho_freezer.1	-2.57	0.65	-3.95	0.000
custo_op	0.20	0.07	2.71	0.011

Entretanto, é sempre imperativo realizar uma análise dos diagnósticos do modelo antes de qualquer interpretação:





Os gráficos de diagnóstico sugerem que a modelagem dupla da RBS não foi suficiente para resolver os problemas da utilização desta distribuição para este conjunto de dados. O wormplot não apresenta um padrão problematico claro descrito na tabela, entretanto existem oscilações e fuga das margens sugeridas do gráfico. Os demais gráficos de diagnóstico mostram problemas na aderência dos resíduos quantílicos próximos à média, formando um padrão hiperbólico. Os problemas relacionados à assimetria estão menos acentuados se comparado ao modelo RBS simples, entretando ainda inferiores se com-

parado com o modelo gama duplo.

Modelagem indicada pelo gamlss

Podemos, por último, verificar a sugestão automática calculada via critério de seleção de Akaike do pacote gamlss, utilizando a função fitDist.

	AIC
GG	452.3450
BCCGo	452.3611
BCCG	452.3611
GB2	454.0389
BCTo	454.2678
BCT	454.2678
BCPE	454.6212
BCPEo	454.6212
IGAMMA	458.6159
GIG	460.6159
LOGNO2	461.4291
LOGNO	461.4291
IG	461.6200
GA	464.9191
exGAUS	475.5580
WEI	481.3826
WEI3	481.3826
WEI2	491.8776
EXP	552.5534
PARETO2o	554.5535
GP	554.5535
PARETO2	554.5536

A tabela acima apresenta as famílias com suporte nos reais positivos, ordenadas do menor para o maior valor do critério de informação de Akaike se ajustado o modelo saturado. Neste caso, quanto menor, melhor o modelo. Notamos que a gama generalizada apresenta o menor AIC, com valor de 452,34; seguida de perto pelos modelos box cox cole green, beta generalizada tipo 2, box cox t, box cox exponenciada e gama inversa. Destes, salvo a gama generalizada, o parâmetro μ se trata ou da mediana, ou de um parâmetro de localização sem interpretação clara. Para a gama generalizada, o parâmetro μ é a média mesmo.

Note que no modelo MLG duplo gama utilizado anteriormente, o critério de Akaike observado foi de 405, portanto ainda é melhor que qualquer um destes modelos sugerido pelo gamlss. Entretanto, podemos testar o ajuste fazendo seleção de covariáveis com um

destes modelos afim de verificar se obtemos um ajuste ainda melhor. Para isto, irei selecionar o mais parcimonioso, de fácil interpretação e menor critério de Akaike entre os listados, que por acaso é o de menor critério absoluto, a gama generalizada.

Modelagem gama generalizada

Como o objetivo é tentar obter o melhor ajuste final possível, farei direto a modelagem dupla, dos parâmetros de média e precisão.

Testou-se aqui diversas combinações de covariáveis para cada um dos parâmetros μ e precisão. Observando significâncias marginais, qualidade do ajuste, valor AIC, sugestões do procedimento stepAIC, optou-se pelo seguinte modelo: Gama generalizado, com função de ligação log para ambos os parâmetros, sendo a média μ modelada utilizando as covariáveis tamanho do refrigerador, tamanho do freezer, quantidade de prateleiras e quantidade de recursos. A precisão foi modelada utilizando a covariável quantidade de prateleiras.

Desta forma, obteve-se as seguintes estimativas de parâmetros:

Estimativas para μ

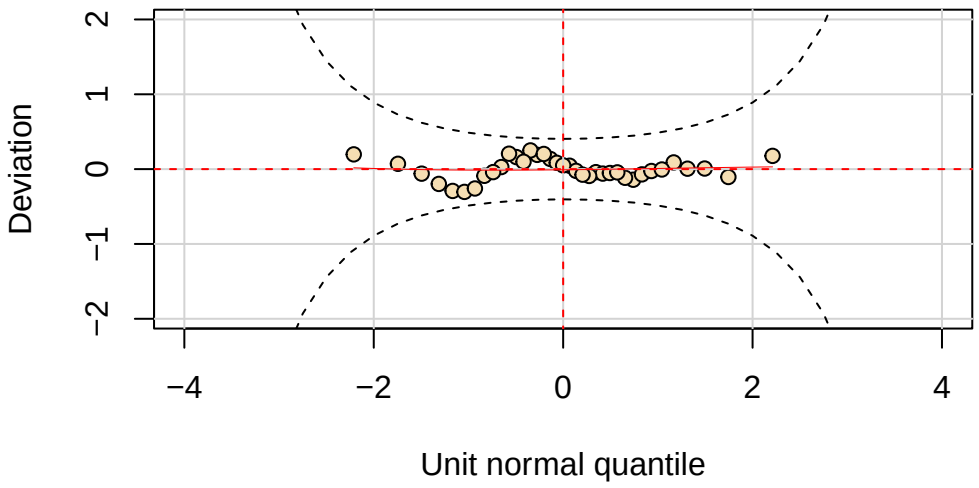
	Estimativa	Erro padrão	Valor t	Pr(>|t|)
X.Intercept.	4.46	0.43	10.42	0.000
tamanho_refrigerador	0.09	0.03	2.98	0.006
tamanho_freezer	0.10	0.02	4.54	0.000
prateleiras	0.05	0.02	3.47	0.002
recursos	0.03	0.01	4.38	0.000

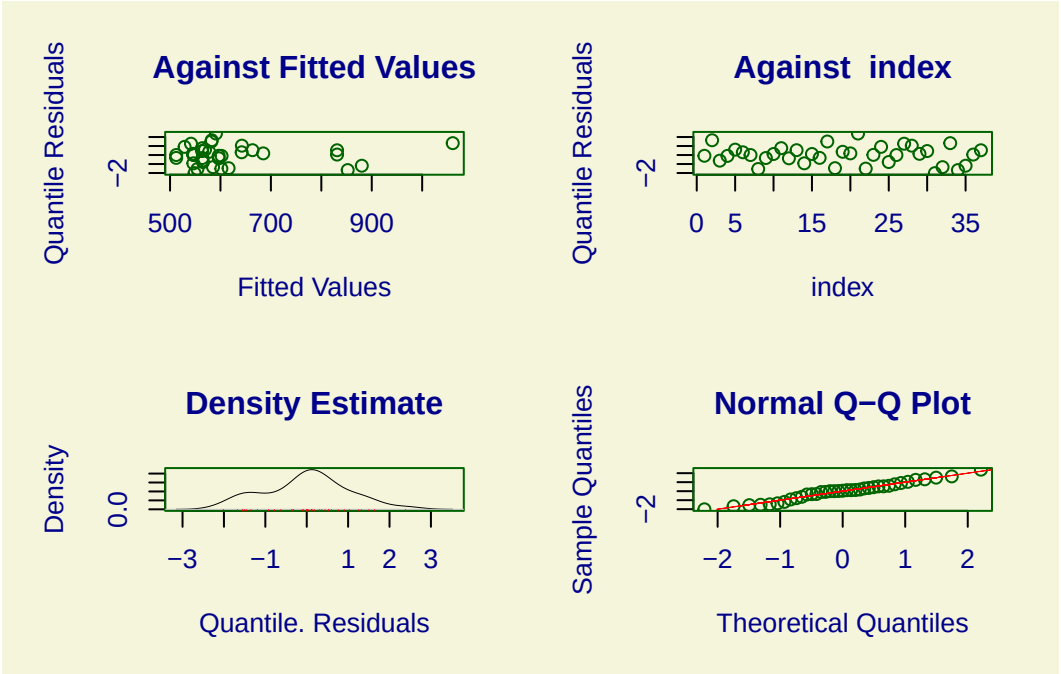
Estimativas para σ^2

	Estimativa	Erro padrão	Valor t	Pr(>|t|)
X.Intercept..1	-2.33	0.28	-8.19	0.000
prateleiras.1	-0.02	0.10	-0.23	0.817

Notamos que todas as estimativas para o parâmetro μ são positivos e consistentes com o esperado para o conjunto de dados. A covariável associada a precisão não apresenta significância marginal, entretando a escolha deu-se pela tentativa manual sucessiva de obter o melhor ajuste com os diagnósticos mais aceitáveis, e esta foi a combinação que forceneu estes resultados. Portanto, a interpretação da precisão pode não ser confiável, mas o ajuste global do modelo aparenta estar adequado desta forma.

Isso pode ser observado nos gráficos de diagnóstico:





Este modelo têm o wormplot mais “comportado” de todas as tentativas realizadas no decorrer deste trabalho. Além disso, os demais gráficos de diagnóstico não apresentam grandes fugas aos pressupostos do modelo. Portanto, este modelo é candidato a ser utilizado também.

Entretando, na minha opinião, o ajuste fornecido pela gama ainda foi superior e mais parcimonioso do que este, desta forma, irei manter a escolha pelo ajuste utilizando o mlg duplo gama, realizado anteriormente.

Referências

RIGBY, Robert A. et al. Distributions for modeling location, scale, and shape: Using GAMLSS in R. Chapman and Hall/CRC, 2019.

Gonçalves, Stefan Zurman. Estudo sobre Modelos Aditivos Generalizados para Localização, Escala e Forma: Teoria e Aplicações sob Modelos Paramétricos. UnB, 2024.