Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística

Caio Gomes Alves

Daniel Krügel

Estimação do Índice de Gini a partir de Indicadores Sociais Municipais do Estado do Paraná em 2010

Curitiba
2022

Caio Gomes Alves
Daniel Krügel

Estimação do Índice de Gini a partir de Indicadores Sociais Municipais do Estado do Paraná em 2010

Trabalho apresentado à disciplina Análise de Regressão do Curso de Graduação em Estatística da Universidade Federal do Paraná, como requisito para obtenção parcial de nota.

Professor: Prof. Dr. César Augusto Taconeli

Curitiba
2022

Sumário

1 Resumo

O presente estudo utiliza-se de métodos de regressão linear múltipla para explicar o Índice de Gini dos municípios paranaenses em 2010 a partir da População Estimada, o IDHM-Educação, o PIB per capita, a Taxa de Pobreza e a Taxa de Analfabetismo de cada município. Foi ajustada uma regressão linear múltipla, considerando todas as covariáveis, que se mostrou pouco explicativa. A partir de uma análise mais profunda, optou-se por utilizar o logaritimo de base 10 da população, devido à diferença de escala muito aparente entre Curitiba (capital do estado) e as demais cidades. Utilizando-se do método *stepwise*, optou-se por eliminar a covariável PIB, e utilizando a transformação de Box-Cox, aplicou-se o logaritmo natural da variável resposta, para melhoria na normalidade da mesma O ajuste final, apesar de melhor do que o inicial, ainda se mostrou pouco efetivo, e melhorias seriam possíveis com métodos que fogem ao escopo da matéria de Regressão Linear.

2 Introdução

O Índice de Gini, criado pelo matemático italiano Conrado Gini, é um indicador da distribuição de renda em um determinado grupo. Calculado a partir da área da Curva de Lorenz, o valor de tal índice varia de 0 a 1, onde 0 representa um grupo onde todos possuem a mesma quantia de riqueza (igualdade perfeita de renda), e em contrapartida 1 representa um grupo onde apenas 1 indivíduo possui toda a riqueza disponível.

O presente estudo tem como objetivo estimar o Índice de Gini a partir de cinco variáveis explicativas: Populacao = População (estimada) do Município, IDHM = Índice de Desenvolvimento Humano Médio da Educação do Município, PIB = PIB (Produto Interno Bruto) per capita, Tx.Pobreza = Proporção de habitantes do Município com renda mensal de menos de R$145,00 e Tx.Analfabetismo = Proporção de habitantes maiores de 15 anos que não são alfabetizadas, variáveis essas que foram consideradas "lógicas" na explicação do Gini.

Tais dados foram coletados das bases de dados do IPARDES (Instituto Paranaense de Desenvolvimento Econômico e Social), disponibilizadas em <http://www.ipardes.gov.br/imp/> , considerando o ano de 2010 (Ano do último CENSO). As dez primeiras observações (cidades, em ordem alfabética) da base de dados são as seguintes:

Tabela 1 – Dados dos 10 primeiros municipios (organizados em ordem alfabética)

Cidade	Gini	Populacao	IDHM	PIB	Tx.Pobreza	Tx.Analfabetismo
Abatiá	0.44	7727	0.596	10414	0.0966	0.1676
Adrianópolis	0.53	6328	0.563	10680	0.2259	0.1682
Agudos do Sul	0.48	8351	0.543	10535	0.1720	0.0884
Almirante Tamandaré	0.43	104350	0.575	6850	0.0489	0.0606
Altamira do Paraná	0.58	4100	0.571	8683	0.2506	0.1757
Alto Paraíso	0.52	3162	0.556	11468	0.1659	0.1466

3 Metodologia

3.1 Análise Exploratória dos Dados

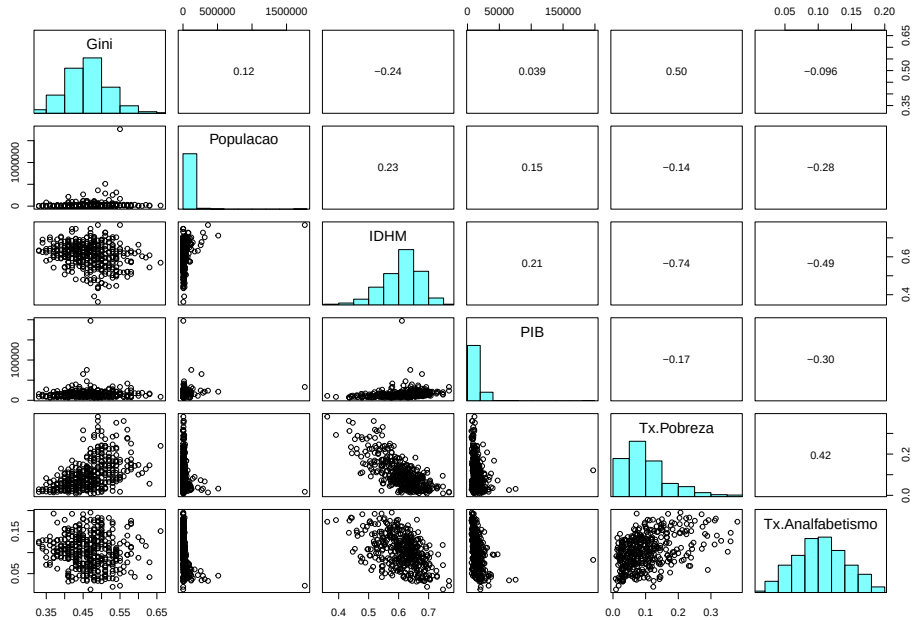
Os dados extraídos passaram por uma análise exploratória inicial, de modo a conhecer melhor o comportamento das variáveis.

Tabela 2 – Sumário dos Dados

Gini	Populacao	IDHM	PIB	Tx.Pobreza	Tx.Analfabetismo
Min. :0.3300	Min. : 1401	Min. :0.362	Min. : 6305	Min. :0.0084	Min. :0.0124
1st Qu.:0.4300	1st Qu.: 5038	1st Qu.:0.576	1st Qu.: 10285	1st Qu.:0.0518	1st Qu.:0.0761
Median :0.4700	Median : 9068	Median :0.621	Median : 12919	Median :0.0843	Median :0.1023
Mean :0.4657	Mean : 26347	Mean :0.611	Mean : 15074	Mean :0.1027	Mean :0.1036
3rd Qu.:0.5000	3rd Qu.: 17334	3rd Qu.:0.655	3rd Qu.: 16864	3rd Qu.:0.1340	3rd Qu.:0.1303
Max. :0.6600	Max. :1764541	Max. :0.768	Max. :197335	Max. :0.3811	Max. :0.1953

Como é possível verificar, existem assimetrias aparentes nas variáveis Populacao e PIB, que podem acarretar em falta de ajuste ou pontos de alavancagem/influência.

Para conferir a relação entre as covariáveis e a variável resposta, será utilizado um correlograma. Além disso, tal correlograma será utilizado para conferir as assimetrias mencionadas e se há covariáveis que são altamente relacionadas, indicando colinearidade entre elas.



O correlograma apresenta na diagonal principal os histogramas das variáveis, o que nos confirma a forte assimetria à esquerda das variáveis População e PIB. Com isso, propõe-se o uso da transformação $\log_{10}$(População), para diminuir a escala das observações. Na parte triangular inferior do correlograma estão os gráficos de disperção das variáveis duas a duas, que indicam que são poucos pontos de alavancagem nessas varáveis.

Na parte triangular superior estão indicados os coeficientes de correlação entre as variáveis. Pode-se perceber que há uma alta correlação negativa (-0.74) entre as covariáveis Tx.Pobreza e IDHM, o que pode indicar colinearidade entre elas. Além disso, há uma correlação quase nula entre a covariável PIB e a variável resposta, indicando que essa covariável pode ser não-significativa na regressão linear.

3.2 Modelo de Regressão Inicial

Como modelo inicial de regressão linear multivariada, será ajustado um modelo aditivo, sem efeitos de interação entre as variáveis e sem considerar nenhuma transformação, utilizando a fórmula:

$$\hat{Gini} = \hat{\beta}_0 + \hat{\beta}_1 \cdot Populacao + \hat{\beta}_2 \cdot IDHM + \hat{\beta}_3 \cdot PIB + \hat{\beta}_4 \cdot Tx.Pobreza + \hat{\beta}_5 \cdot Tx.Analfabetismo$$

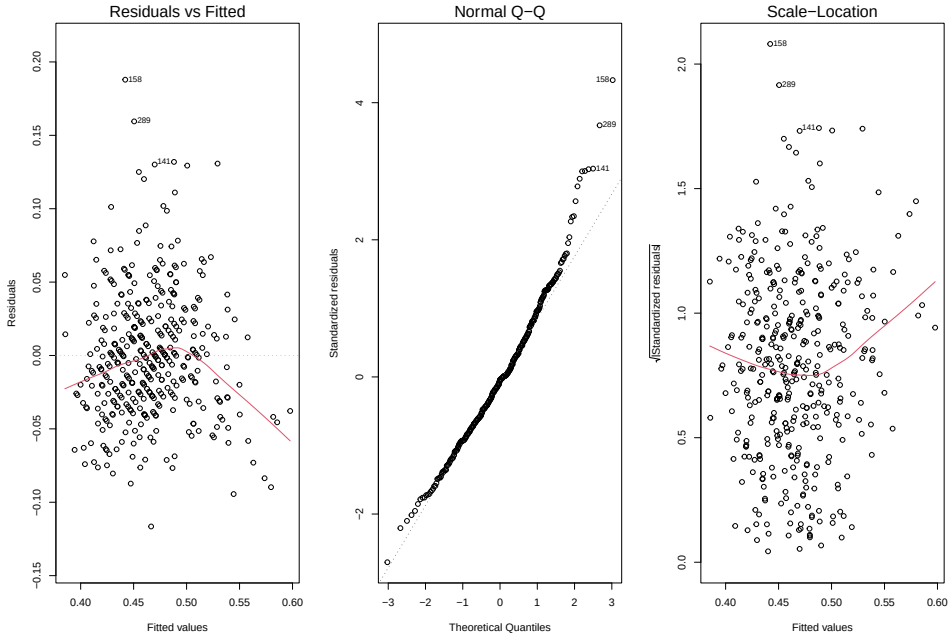
Onde os betas serão estimados a partir dos estimadores de mínimos quadrados dos dados obtidos, utilizando a função lm(Gini ~ .) do RStudio. A partir desse ajuste inicial, obtemos os seguintes estimadores para as variáveis explicativas:

```
##      (Intercept)         Populacao              IDHM               PIB
##       0.37103902        0.00000006        0.12603782        0.00000014
##       Tx.Pobreza Tx.Analfabetismo
##       0.59674477       -0.45709632
```

Realizando o diagnóstico do ajuste, foi considerado utilizar a transformação $\log_{10}$ da covariável Populacao, de forma a reduzir a escala dos dados. O segundo ajuste, considerando tal transformação é o seguinte:

$$\hat{Gini} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \log_{10}(Populacao) + \hat{\beta}_2 \cdot IDHM + \hat{\beta}_3 \cdot PIB + \hat{\beta}_4 \cdot Tx.Pobreza + \hat{\beta}_5 \cdot Tx.Analfabetismo$$

Agora, utilizando-se dos gráficos para ajuste do modelo, é possível avaliar se não há fuga dos pressupostos da regressão linear:



Como é possível verificar, há uma fuga clara do pressuposto de normalidade dos erros, a partir da estimação dos resíduos. No primeiro gráfico, de comparação de resíduos x valores ajustados, é possível perceber que a média dos erros não é constante. Com base no segundo, é possível verificar que a variável resposta provavelmente não segue uma distribuição normal, a partir da comparação dos quantis teóricos com os resíduos padronizados. Por fim, percebe-se que a variância dos resíduos não é constante, baseando-se no terceiro gráfico, que compara os valores ajustados com a raiz quadrada dos resíduos, que é um estimador não-viciado para a variância do erro.

3.3 Seleção de covariáveis

Um dos motivos que podem influenciar na falta de ajuste de um modelo de regressão linear é o excesso de covariáveis pouco explicativas no modelo. Para remediar isso, será utilizado o algoritmo *step-wise* para seleção de variáveis explicativas, que mede o impacto no ajuste do modelo com a inclusão/exclusão das covariáveis, a partir do AIC *(Akaike Information Criterion)* o Critério de Informação de Akaike, calculado da seguinte maneira:

$$AIC = -2\log(\hat{L}) + 2p, \begin{cases} \hat{L} = \text{Máxima Verossimilhança Estimada} \\ p = \text{Número de Covariáveis} \end{cases}$$

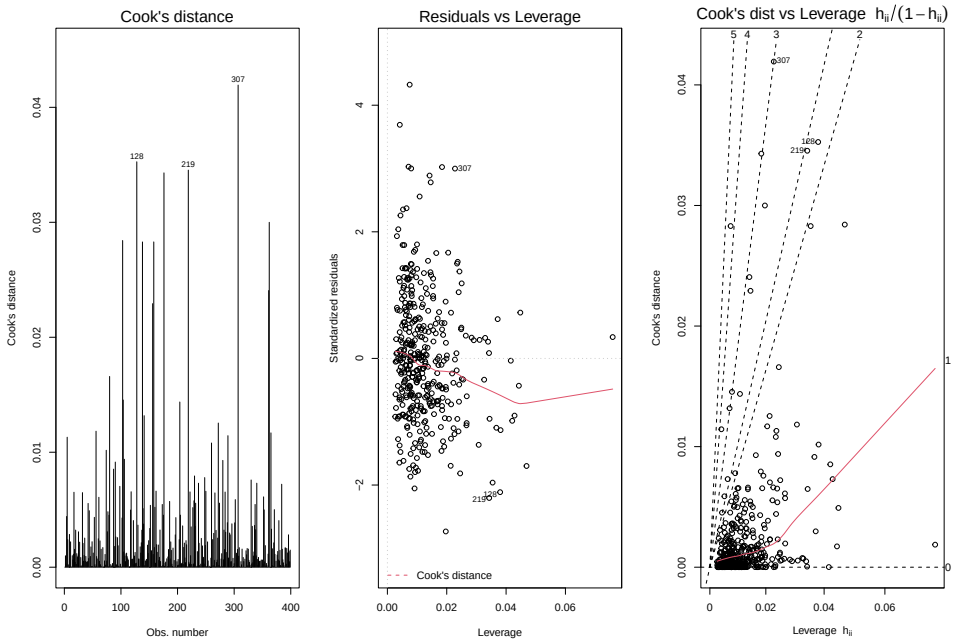
O objetivo do AIC é calcular o critério de informação de forma que, quanto melhor o ajuste aos dados, menor será, porém há um fator de penalização $2p$ que aumenta o AIC para modelos muito complexos. De tal forma, o AIC é minimizado para modelos simples, que explicam bem a correlação entre as covariáveis e a variável resposta.

Utilizando o método *step-wise* implementado pelo pacote MASS com a função stepAIC(), temos o seguinte resultado:

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## Gini ~ log10(Populacao) + IDHM + PIB + Tx.Pobreza + Tx.Analfabetismo
##
## Final Model:
## Gini ~ log10(Populacao) + IDHM + Tx.Pobreza + Tx.Analfabetismo
##
##
##     Step Df    Deviance Resid. Df Resid. Dev       AIC
## 1                              393  0.7459418 -2494.546
## 2 - PIB  1 0.001181132       394  0.7471229 -2495.914
```

Como visto, a variável PIB foi considerada não-significativa, portanto será descartada. Por isso, será proposto um novo modelo, considerando apenas as variáveis log(População), IDHM Educação, Taxa de Pobreza e Taxa de Analfabetismo municipais.

Realizando o diagnóstico do ajuste, há uma melhora no R^2 ajustado e todas as variáveis restantes se apresentaram significativas. Passando para a análise de pontos influentes, temos os seguintes resultados:



É possível perceber que não há dados muito influentes (a distância de Cook apresentada no primeiro gráfico não excede 1, que é um indicativo de outliers), e que

não há pontos de alavancagem (indicado pelo segundo e terceiro gráficos, que modelam a Distância de Cook x Alavancagem de cada ponto).

3.4 Análise de normalidade

Há ainda uma considerável falta de ajuste na variável resposta, que se deve à não-normalidade dos dados observados. Para testar a hipótese de normalidade dos resíduos, será utilizado o teste de Shapiro-Wilk.

O resultado do teste retorna um p-valor 1.138×10^{-6}, logo é rejeitada a hipótese nula de que o Índice de Gini no estado do Paraná é normalmente distribuído. Para remediar essa situação, será aplicada uma transformação do tipo potência, no caso, a de Box-Cox, para transformação da variável resposta, que funciona da seguinte maneira:

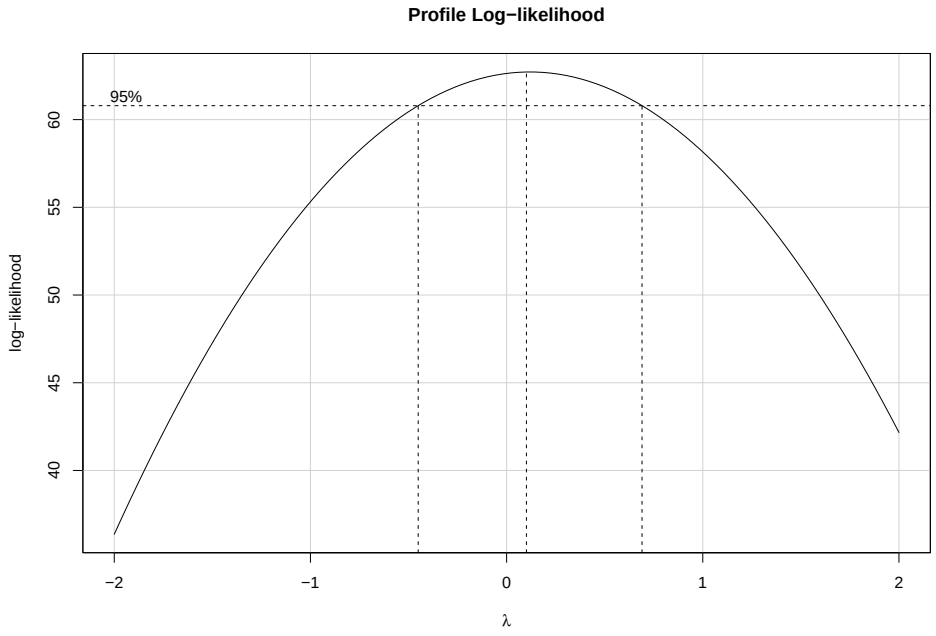
Seja Y a variável resposta (neste caso, o Índice de Gini). A transformação de Box-Cox para Y é da forma:

$$f_\lambda(Y) = \frac{Y^\lambda - 1}{\lambda Y^{\lambda-1}}$$

O objetivo aqui é encontrar o λ que maximizará a função de verossimilhança:

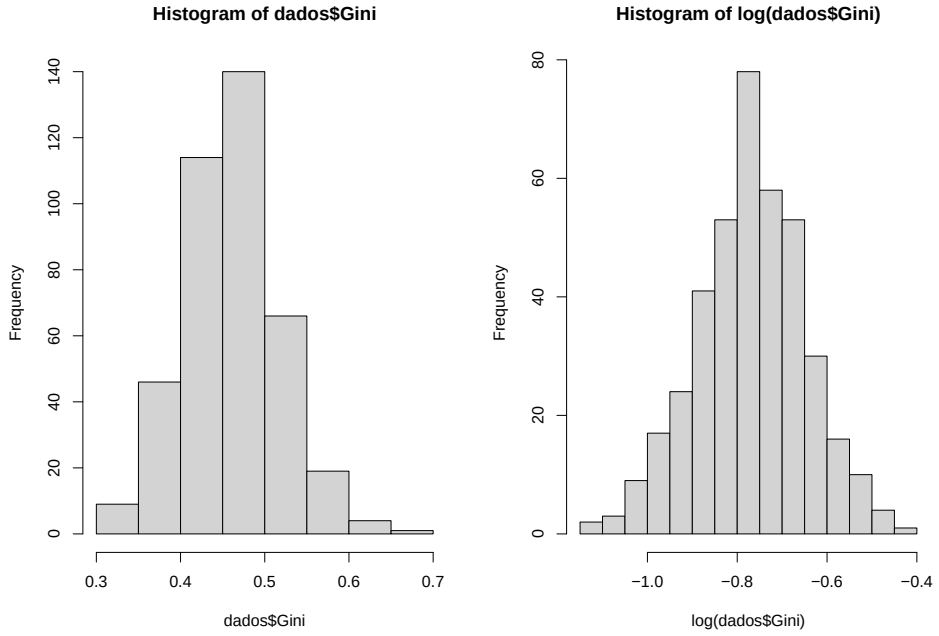
$$L(\lambda) = -\frac{n}{2}\log[SQRes(\lambda)]$$

Onde $SQRes(\lambda)$ é a soma de quadrados dos resíduos obtida pela transformação. Assim, utilizando a função boxCox(), do pacote car no modelo ajustado teremos:



Percebe-se que o ponto de máxima-verossimilhança para λ é próximo de 0, portanto aplicaremos a transformação ln na variável resposta (Gini). Com isso, podemos perceber a

diferença na normalidade diretamente na distribuição dos dados, com auxílio dos seguintes histogramas:



3.5 Modelo final

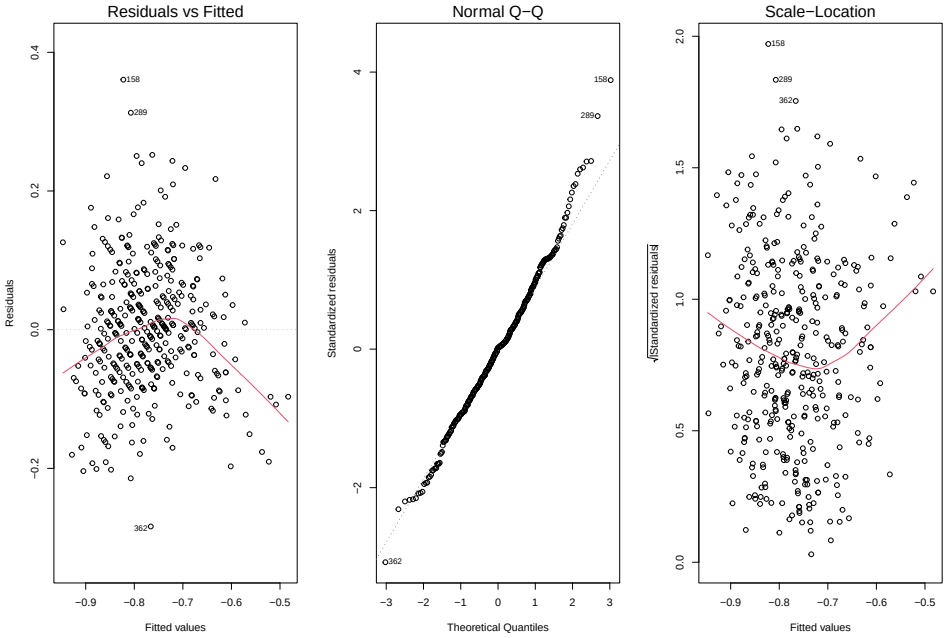
Para o modelo final, teremos:

$$\ln(\hat{Gini}) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \log_{10}(Populacao) + \hat{\beta}_2 \cdot IDHM + \hat{\beta}_3 \cdot Tx.Pobreza + \hat{\beta}_4 \cdot Tx.Analfabetismo$$

Com ele obtemos os seguintes valores ajustados:

Estimate	Std. Error	t value	Pr(>|t|)
-1.3632	0.1040	-13.11	0.0000
0.0720	0.0120	5.98	0.0000
0.3749	0.1163	3.22	0.0014
1.3065	0.0973	13.43	0.0000
-0.5961	0.1642	-3.63	0.0003

um coeficiente de determinação ajustado $R^2_{Adj} = 0.4279$, e os seguintes gráficos para diagnóstico do ajuste:


Residuals vs Fitted
Residuals
Fitted values
158
289
362
Normal Q−Q
Standardized residuals
Theoretical Quantiles
158
289
362
Scale−Location
√|Standardized residuals|
Fitted values
158
289
362

4 Conclusões

Com a análise final, foi possível perceber que, ainda que tenham sido feitas transformações para as variáveis de forma a amenizar os problemas percebidos durante o processo de modelagem, desde a seleção de covariáveis, análise de normalidade da variável resposta e transformação de Box-Cox, os pressupostos necessários para a regressão linear não foram atendidos, pois a média dos resíduos, bem como sua variâcia não foram constantes, além de ainda haver fuga da normalidade nas caudas do Q-Q plot mostrado durante a sessão de metodologia.

Dessa forma, maneiras de melhorar a modelagem seriam:

- Selecionar outras covariáveis que possam explicar o modelo, que não as selecionadas para este estudo;
- Utilizar-se de métodos de regressão não-linear, como GLMs (Modelos Lineares Generalizados) e GAMLSS (Modelos Aditivos Generalizados para Locação, Escala e Forma), que fogem do escopo da matéria atual.