Guilherme de Sousa Almeida

Análise do Número de Vítimas em Acidentes em Rodovias Federais no Ceará nos anos de 2007 a 2024

Fortaleza
2025

Introdução

O conjunto de dados contém os anos de avaliação, os tipos de acidente que são: atropelamento, capotamento, colisão e saída da pista, o número de acidentes e o número de vítimas. O conjunto de dados foi retirado do Painel CNT de Acidentes Rodoviários TRANSPORTE (2023). "O Painel CNT de Consultas Dinâmicas de Acidentes Rodoviários é uma ferramenta desenvolvida pela Confederação Nacional do Transporte que reúne dados da Polícia Rodoviária Federal sobre acidentes ocorridos em rodovias federais brasileiras no período de 2007 a 2023. Todas as buscas podem ser feitas com dados nacionais, por região e por Unidade da Federação.

O objetivo desta análise é utilizar regressão quase-binomial logística via quase-verossimilhança afim de modelar o número de acidentes.

Metodologia

A teoria da quase-verossimilhança surge como uma alternativa flexível à modelagem estatística tradicional baseada em distribuições de probabilidade completamente especificadas. Proposta por WEDDERBURN (1974), essa abordagem permite a inferência estatística mesmo quando a distribuição completa da variável resposta é desconhecida ou não segue uma forma canônica da família exponencial, baseando-se apenas nos dois primeiros momentos da variável resposta: a média e variância. Essa técnica é especialmente útil em contextos em que se observa superdispersão nos dados, ou seja, quando a variância observada excede a variância esperada sob o modelo clássico. O modelo binomial tem a característica de ter variância menor que a média. Portanto, os modelos de quase-verossimilhança é uma maneira bastante útil de se contornar essa característica.

Definição: Seja $Y : \Omega \to \mathbb{R}$ uma variável aleatória definida sobre um espaço de probabilidade $(\Omega, \mathcal{F}, \mathbb{P})$. Adicionalmente, suponha que $\mathrm{E}(Y) = \mu$ e $\mathrm{Var}(Y) = \sigma^2 \mathrm{V}(\mu)$ em que $\sigma^2 > 0$ e $V(.) > 0$ é uma função conhecida. O logaritmo da função de quase-verossimilhança é definido por

$$Q(\mu; y) = \frac{1}{\sigma^2} \int_y^\mu \frac{y - t}{\mathrm{V}(t)} dt.$$

Note que para $Y \sim \mathrm{Bin}(m, \pi)$ temos que $\mathrm{E}(Y) = m\pi = \mu$ e $\mathrm{Var}(Y) = m\pi(1-\pi) = \mu(m-\mu)/m$. Logo, o logaritmo da função de quase-verossimilhança quase-binomial é dada por

$$Q(\pi; y) = \begin{cases} \frac{m}{\sigma^2} \ln(1 - \pi), \text{ se } y = 0; \\ \frac{1}{\sigma^2} \left[y \ln\left(\frac{m\pi}{y}\right) + (m - y) \ln\left(\frac{m(1-\pi)}{m-y}\right) \right], \text{ se } 0 < y < m; \\ \frac{m}{\sigma^2} \ln(\pi), \text{ se } y = m. \end{cases}$$

Suponha então que $Y_1, ..., Y_n$ é uma amostra aleatória e que temos o modelo $g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}, i = 1, ..., n$ em que $g(.)$ é uma função de ligação conveniente, $\boldsymbol{x}_i^\top$ é a i-ésima linha da matriz de especificação do modelo, $\boldsymbol{X}_{(n \times p)}$ e $\boldsymbol{\beta}_{(p \times 1)}$ é um vetor de parâmetros de regressão.

Seja $Q(\mu_i; y_i), i = 1, ..., n$ o logaritmo da função de quase-verossimilhança conjunta é definida como

$$Q(\boldsymbol{\mu}; \boldsymbol{y}) = \sum_{i=1}^{n} Q(\mu_i; y_i).$$

A estimação de $\boldsymbol{\beta}$ é feita utilizando o método escore de Fisher. O processo iterativo é dado por

$$\boldsymbol{\beta}^{(m+1)} = \left(\boldsymbol{X}^\top \boldsymbol{W}^{(m)} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{W}^{(m)} \boldsymbol{z}^{(m)},$$

em que $\boldsymbol{z} = \boldsymbol{\eta} + \boldsymbol{W}^{-1/2} \boldsymbol{V}^{-1/2} (\boldsymbol{y} - \boldsymbol{\mu})$.

Já para σ^2, um estimador é dado por

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

A estatística do teste individual é dada por

$$z_j = \frac{\hat{\beta}_j - \beta_j^0}{\sqrt{\hat{V}\text{ar}\left(\hat{\beta}_j\right)}}.$$

Sob a hipótese nula, para n suficientemente grande, a estatística $z \sim N(0, 1)$.

Um intervalo de confiança com $(1 - \alpha)\%$ de confiança é dado por

$$\text{IC}(\beta_j; 1 - \alpha) = \left[\hat{\beta}_j \pm z_{1-\alpha/2} \sqrt{\hat{V}\text{ar}\left(\hat{\beta}_j\right)}\right],$$

em que $z_{1-\alpha/2}$ é o quantil $1 - \alpha/2$ da distribuição normal padrão.

Para diagnóstico e validação do modelo serão utilizados os métodos a seguir.

A matriz de projeção é dada por

$$\hat{\boldsymbol{H}} = \hat{\boldsymbol{W}}^{1/2} \boldsymbol{X} \left(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \hat{\boldsymbol{W}}^{1/2}.$$

Para se detectar pontos de alavanca, pode-se plotar os elementos da diagonal principal da matriz $\hat{\boldsymbol{H}}$, $h_{ii}, i = 1, ..., n$, contra os índices das observações.

O resíduo de Pearson para a i-ésima observação é definido como

$$\hat{r}_{p_i} = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}\sqrt{V(\hat{\mu}_i)}}.$$

Em geral, plota-se os valores do resíduo de Pearson contra alguma função de $\hat{\mu}$. Uma distribuição aleatória dos resíduos em torno de zero indica uma boa especificação do modelo. Além disso, a dispersão dos resíduos permanecer constante para todos os valores indica homoscedasticidade.

Para os modelos de quase-verossimilhança, uma distância de Cook é dada por

$$\mathrm{LD}_i = \frac{\hat{h}_{ii}}{\left(1-\hat{h}_{ii}\right)^2}\hat{r}^2_{p_i}.$$

Plotar os valores da distância de Cook contra os índices das observações pode indicar possíveis pontos influentes.

Análise dos Dados

A Tabela 1 contém um recorte do conjunto de dados a ser analisado.

Tabela 1 – Evolução Anual dos Acidentes, Vítimas e Proporção de Vítimas por Tipo de Acidente (2007–2024)

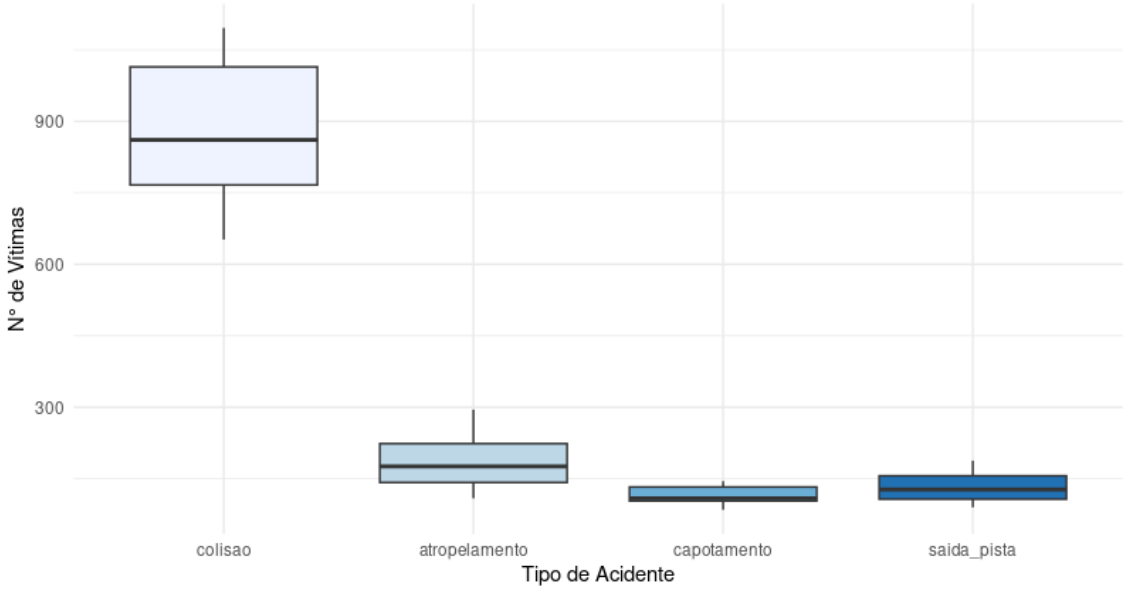
Ano	Acidente	N° de Acidentes	Vítimas	Não-Vítimas	Prop. de Vítimas
2007	Atropelamento	328	223	105	0,68
2008	Atropelamento	356	235	121	0,66
2009	Atropelamento	319	217	102	0,68
2010	Atropelamento	436	295	141	0,68
...	...	...	...	...	...
2022	Saída da Pista	136	103	33	0,76
2023	Saída da Pista	114	90	24	0,79
2024	Saída da Pista	117	94	23	0,80

De acordo com a Tabela 2, observa-se que, em média, os acidentes do tipo Colisão apresentam o maior número de vítimas, além de registrarem os maiores valores mínimo e máximo. Por outro lado, os acidentes do tipo Capotamento exibem os menores valores para a média, o mínimo e o máximo. Destaca-se ainda que as variâncias observadas são significativamente superiores às respectivas médias, o que indica a presença de superdispersão nos dados. A Figura 1 reforça essas conclusões.

Tabela 2 – Medidas Descritivas para o Número de Vítimas por Tipo de Acidente

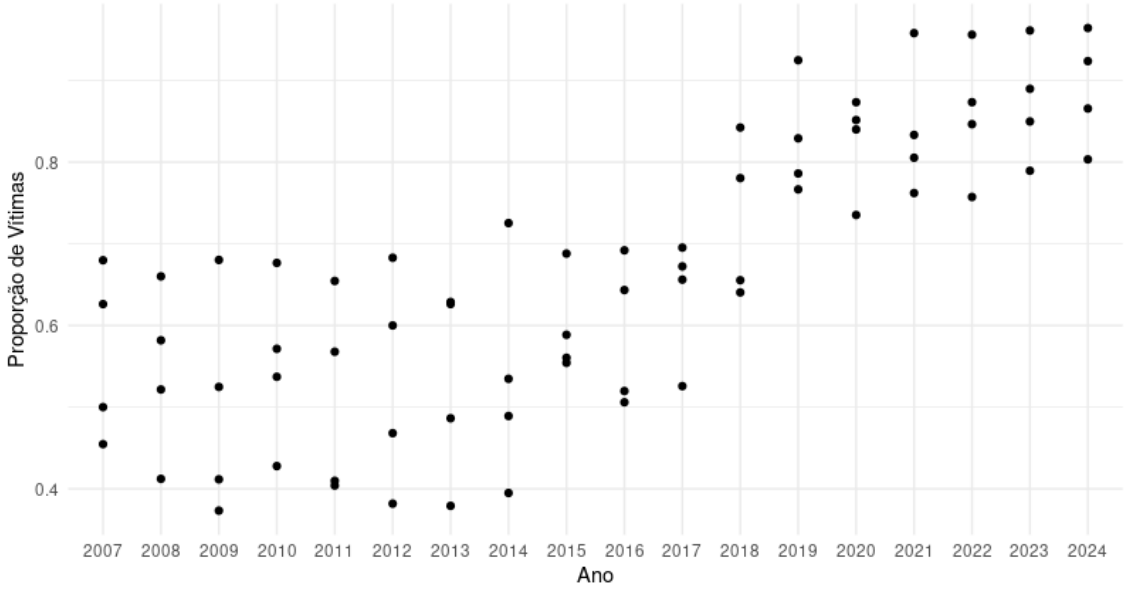
Acidente	n	Mín.	Máx.	Média	Var.	Coef. Var. (%)
Colisão	18	652	1096	882,89	20344,10	16,16
Atropelamento	18	109	295	186,00	2602,35	27,43
Capotamento	18	85	145	115,11	312,22	15,35
Saída da Pista	18	90	188	132,50	977,21	23,59

Figura 1 – Boxplot do Número de Vítimas por Acidente



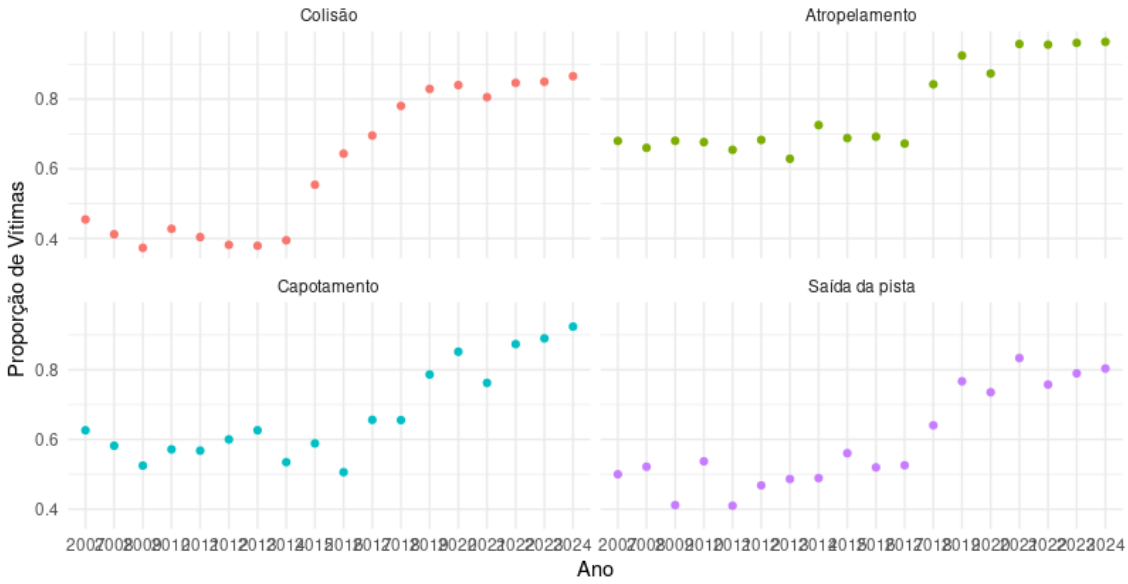
A Figura 2 revela que a proporção de vítimas tem aumentado ao longo dos anos. Esse crescimento é um sinal de alerta e merece atenção por parte das autoridades e formuladores de políticas públicas.

Figura 2 – Proporção de Vítimas ao Longo dos Anos



Conforme ilustrado na Figura 3, os acidentes do tipo Colisão apresentam um crescimento mais acentuado ao longo dos anos. Em contraste, os tipos Capotamento e Saída de Pista registram um aumento mais gradual. Já os Atropelamentos iniciam o período com proporções relativamente elevadas.

Figura 3 – Proporção de Vítimas ao Longo dos Anos por Tipo de Acidente



Para a modelagem, utiliza-se o seguinte modelo

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{3}\beta_j(x_{ij}-2007)^i + \sum_{j=4}^{6}\beta_j I_j + \sum_{j=7}^{9}\beta_j(x_{ij}-2007)I_j,$$

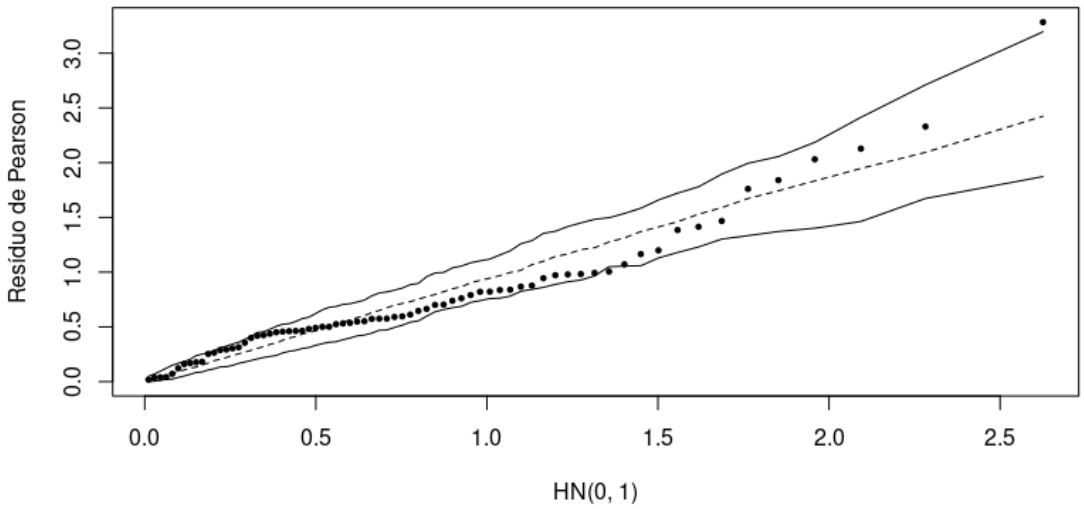
em que π_i é a probabilidade de ocorrer vítimas e x_i é o ano de avaliação. Para considerar o efeito dos acidente, são criadas varáveis indicadoras para os tipos de acidente em que Colisão é a casela de referência. Temos então que I_4 e I_7 são para atropelamento, I_5 e I_8 são para capotamento e I_6 e I_9 são para saída de pista. A subtração do ano por 2007 tem como intuito ganhar interpretação do modelo.

Tabela 3 – Estimativa dos Coeficientes do Modelo Quase-Binomial Testes Individuais para a Proporção de Vítimas

Coeficiente	Estimativa	Erro Padrão	Intervalo de Confiança	z	Nível Descritivo
β_0	-0,1164	0,0956	$[-0{,}30; 0{,}07]$	-1,22	0,2280
β_1	-0,2850	0,0486	$[-0{,}38; -0{,}19]$	-5,87	$<$ **0,01**
β_2	0,0534	0,0073	$[0{,}04; 0{,}07]$	7,31	**< 0,01**
β_3	-0,0017	0,0003	$[-0{,}00; -0{,}00]$	-5,65	$<$ **0,01**
β_4	1,1414	0,1672	$[0{,}82; 1{,}47]$	6,83	$<$ **0,01**
β_5	0,8161	0,1950	$[0{,}44; 1{,}20]$	4,18	$<$ **0,01**
β_6	0,4785	0,1622	$[0{,}16; 0{,}80]$	2,95	$<$ **0,01**
β_7	-0,0346	0,0246	$[-0{,}08; 0{,}01]$	-1,41	0,1635
β_8	-0,0658	0,0236	$[-0{,}11; -0{,}02]$	-2,79	$<$ **0,01**
β_9	-0,0641	0,0210	$[-0{,}10; -0{,}02]$	-3,05	$<$ **0,01**

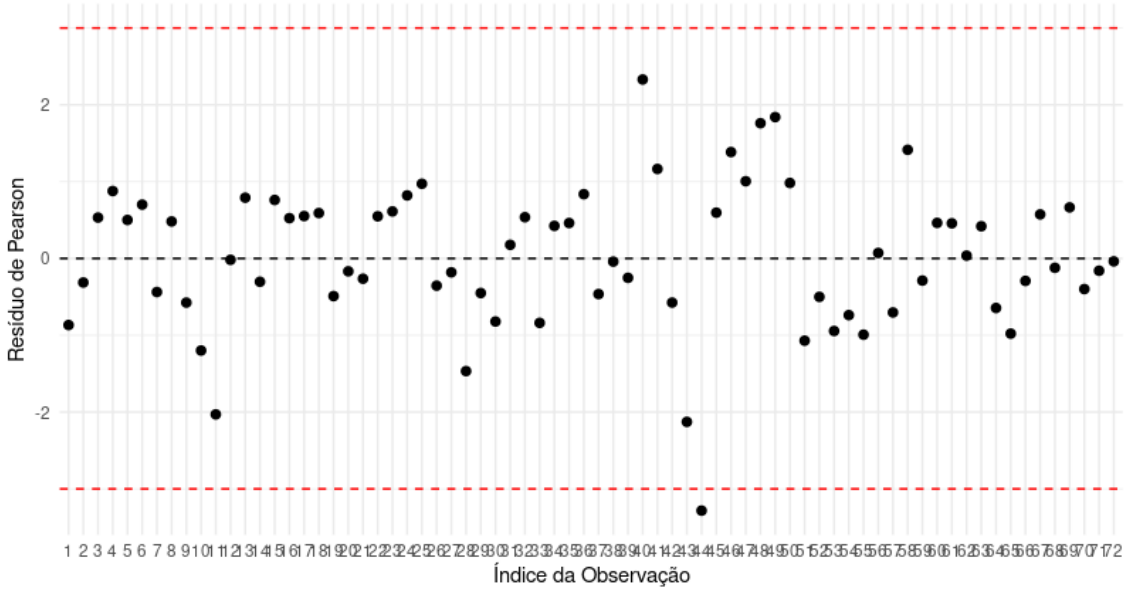
De acordo com a Figura 4, podemos observar que houve aderência dos dados em relação ao modelo quase-binomial.

Figura 4 – Gráfico Meio-Normal de Probabilidade



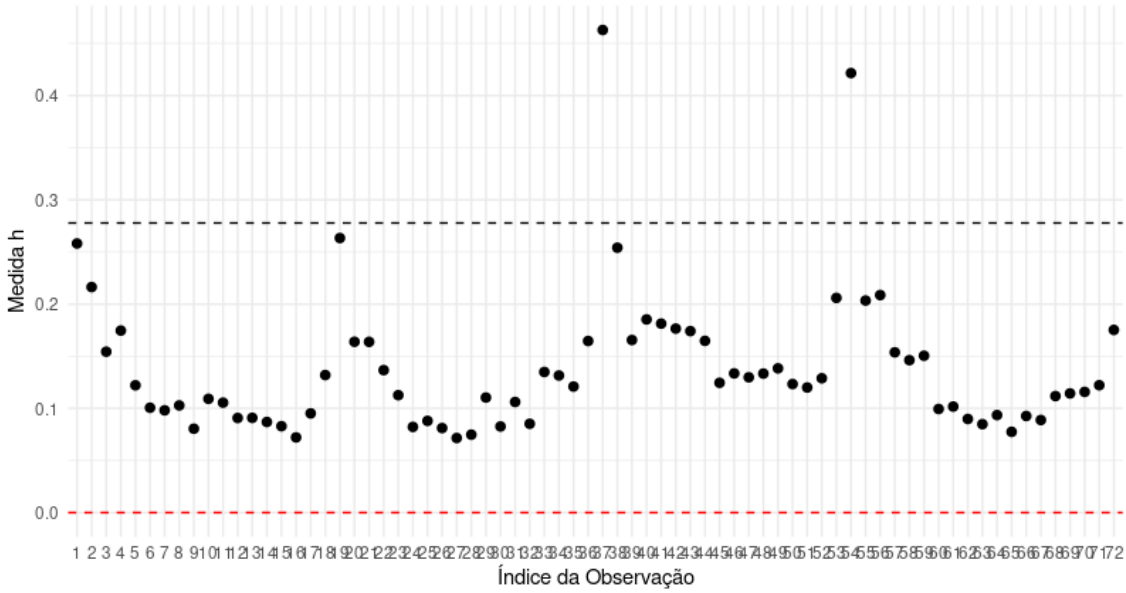
De acordo com a Figura 5, observa-se um padrão aleatório nos resíduos, o que sugere uma boa aderência do modelo aos dados. Além disso, a variabilidade dos resíduos se mantém, em geral, constante, indicando que a variância foi adequadamente acomodada. Por fim, há apenas uma observação fora do intervalo $[3, 3]$, mas ela não se encontra muito distante dos limites, não representando um ponto de preocupação.

Figura 5 – Gráfico do Resíduo de Pearson contra o Índice das Observações



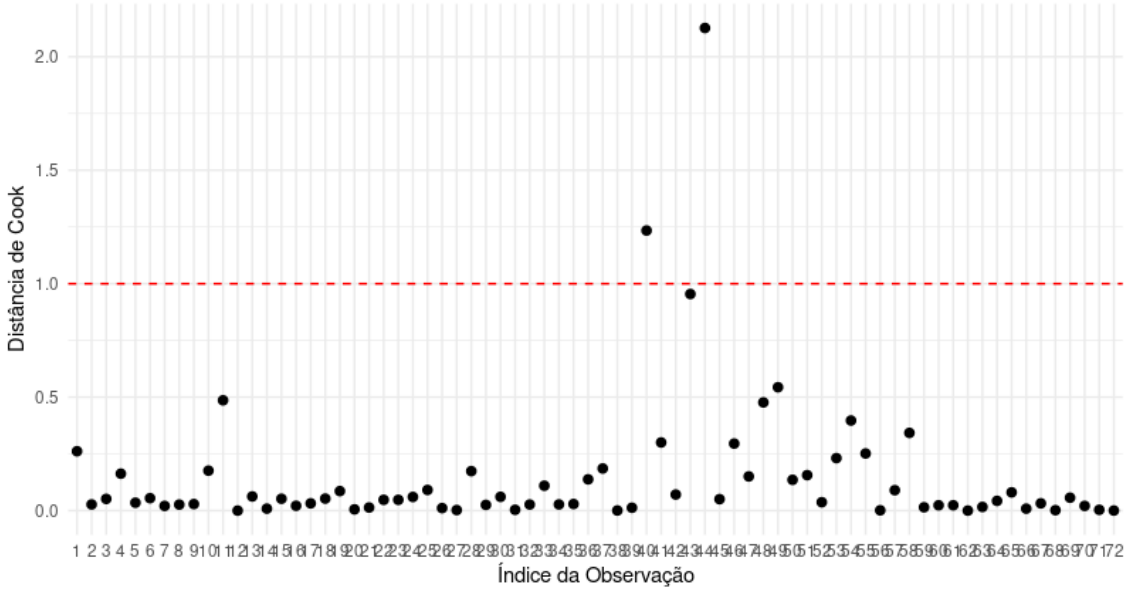
De acordo com a Figura 6, podemos notar que 2 pontos estão fora da reta, indicando possíveis pontos de influência.

Figura 6 – Gráfico da Medida h contra o Índice das Observações



De acordo com a Figura 7, pode-se também notar que também há duas observações acima da reta, porém, somente a observação mais distante deve ser analisada.

Figura 7 – Gráfico da Distância de Cook contra o Índice das Observações



Note que, em geral, obtemos um bom modelo com boas capacidades de inferência e predição.

Referências

CORDEIRO, G. M.; DEMÉTRIO, C. G. Modelos lineares generalizados e extensões. **Piracicaba: USP**, p. 31, 2008.

MCCULLAGH, P. **Generalized linear models**. [s.l.] Routledge, 2019.

MCCULLOCH, C. E.; SEARLE, S. R. **Generalized, linear, and mixed models**. [s.l.] John Wiley & Sons, 2004.

PAULA, G. A. **Modelos de regressão: com apoio computacional**. [s.l.] IME-USP São Paulo, 2024. v. 1

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria: R Foundation for Statistical Computing, 2024.

TRANSPORTE, C. N. DO. **Painel CNT de Acidentes Rodoviários**https://www.cnt.org.br/painel-acidente, 2023.

WEDDERBURN, R. W. Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. **Biometrika**, v. 61, n. 3, p. 439–447, 1974.