UNIVERSIDADE FEDERAL DO PARANÁ
CURSO DE ESTATÍSTICA

Andressa Luiza Cordeiro GRR:20160218
Jayme Gomes dos Santos Junior GRR:20160210
Luciana Helena Kowalski GRR:20160231

MODELAGEM DE DADOS DE MORTALIDADE POR ACIDENTES DE TRANSPORTE NO PARÁ

CURITIBA
2019

1 RESUMO

Com o aumento da violência no trânsito e fatalidades, cresce a necessidade de ações de prevenção de acidentes e cuidados posteriores aos mesmos. Com isso, objetivou-se realizar a modelagem de dados de acidentes fatais no estado do Pará para predição do número de acidentes. Para isso, utilizou-se modelos lineares generalizados, com auxílio do software *R*. Inicialmente foi realizada análise descriva, para então ajuste de modelos e escolha do mais adequado. Com o modelo escolhido foi realizada um reajuste e verificação de predição. O modelo escolhido manteve duas covariáveis que se mostraram significativas, sendo elas relacionadas à frota e urbanização.

2 INTRODUÇÃO

A violência no trânsito é um fator bastante preocupante devido ao aumento da população assim como do número de veículos circulantes. As baixas taxas e juros facilitados permitem que mais pessoas tenham acesso à compra de veículos e esses números tem como consequência elevados casos de acidentes, muitos deles fatais. Estudos indicam que as maiores taxas de mortalidade estão associadas à fatores de risco, que podem ser imprudência, falta de educação, insegurança, falta de atendimento médiço posterior ao acidente e até mesmo estar relacionada ao gênero do condutor. Identificar as causas de óbitos no trânsito permite a execução de ações mitigatórias, tais como planejamentos rodoviários, gestões políticas e outras ações preventivas.

Sendo assim, o trabalho tem como objetivo realizar a modelagem de dados de acidentes fatais no estado do Pará visando a predição do número de acidentes.

3 MATERIAL E MÉTODOS

A base de dados principal com a variável resposta utilizada no estudo foi extraída do site **DATASUS** (http://tabnet.datasus.gov.br). As covariáveis população total, taxa de população urbana, renda per capta foram extraídas do site **Atlas Brasil** (http://www.atlasbrasil.org.br/2013/pt/consulta/), já o número de veículos foi extraído do site **IBGE** (https://cidades.ibge.gov.br/brasil/pa/pesquisa/22/28120?ano=2017&localidade1=150690). Também foi realizada a subdivisão dos dados em macroregiões, estas foram divididas de acordo com as macroregiões de saúde do estado (4 macroregiões) (Tabela 1). AS variáveis utilizadas na modelagem foram:

1. **Município -** 143 municípios do estado do Pará;
2. **Fatalidades -** Número de acidentes de transporte por municipio (variável resposta);
3. **Macrorregião -** 4 níveis divididas segundo macrorregiões de saúde do estado;
4. **n_veiculos -** número de veículos por minicípio;
5. **pop_total -** população total por município;
6. **pop_urb -** taxa de população urbana por município;
7. **per_capta -** renda per capta por município.

Tabela 1: Primeiras linhas da base de dados usada para o ajusta do modelo.

Município	Fatalidades	Macrorregião	n_veiculos	pop_total	pop_urb	per_capta
Abaetetuba	22	1	31363	141100	0.59	293.01
Abel Figueiredo	1	4	1334	6780	0.89	390.12
Acará	14	2	4347	53569	0.24	199.34
Afuá	0	1	16	35042	0.27	163.98
Água Azul do Norte	4	4	2803	25057	0.19	266.02
Alenquer	1	3	9159	52626	0.53	215.33
Almeirim	4	3	3761	33614	0.59	484.16
Altamira	67	3	60625	99075	0.85	492.05
Anajás	0	1	282	24759	0.38	186.88
Ananindeua	195	1	129756	471980	1.00	564.76

A análise estatística foi realizada com o software R.

Para realizar o ajuste do modelo foi utilizado GLM *(Generalized Linear Models)*, foram testados ajustes para a família poisson e binomial negativa ambos com função de ligação logarítmica. A escolha do modelo foi baseada no critério de informação de Akaike (AIC) e na verossimilhança.

4 RESULTADOS E DISCUSSÕES

4.1 Análise descritiva

Para avaliação da distribuição dos dados foi realizado boxplot das variáveis (Figura 1). As variáveis n_veiculos, pop_total e per_capta apresentaram dados com elevada variação e com pontos discrepantes, para melhor verificação foi avaliado em conjunto o histograma.

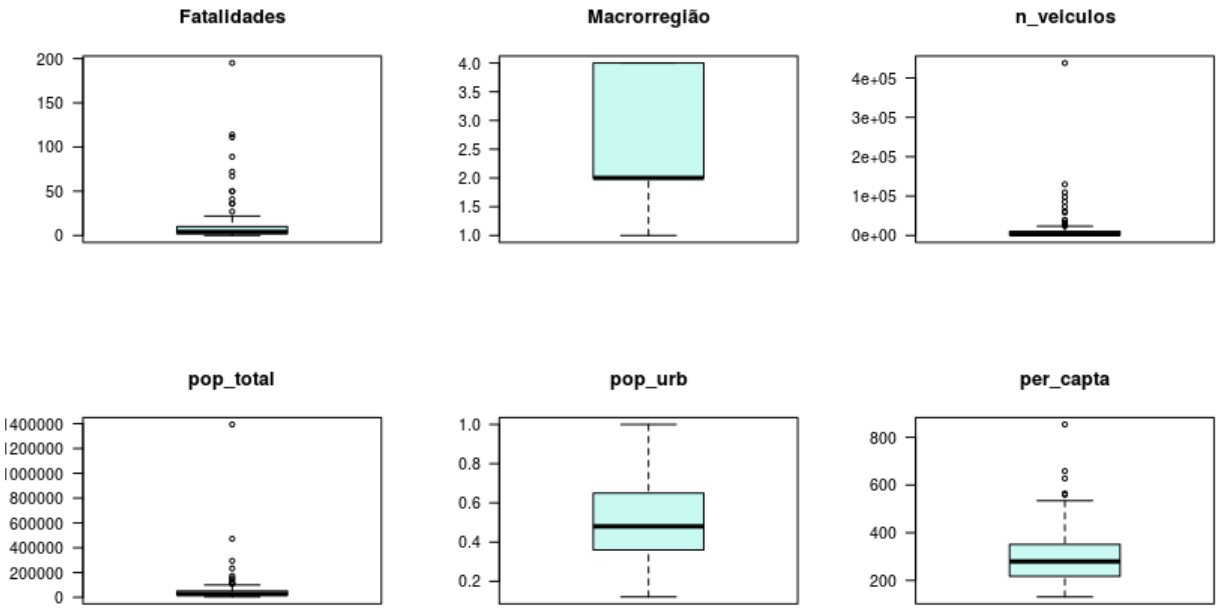

Figura 1: Boxplot das variáveis utilizadas na análise.

Pelos histogramas ficou evidente que as variáveis número de carros, população e renda per capta são assimétricas (Figura 2). A variável resposta (Fatalidades) também se mostrou assimétrica devido a alguns municípios terem muito mais casos que outros.

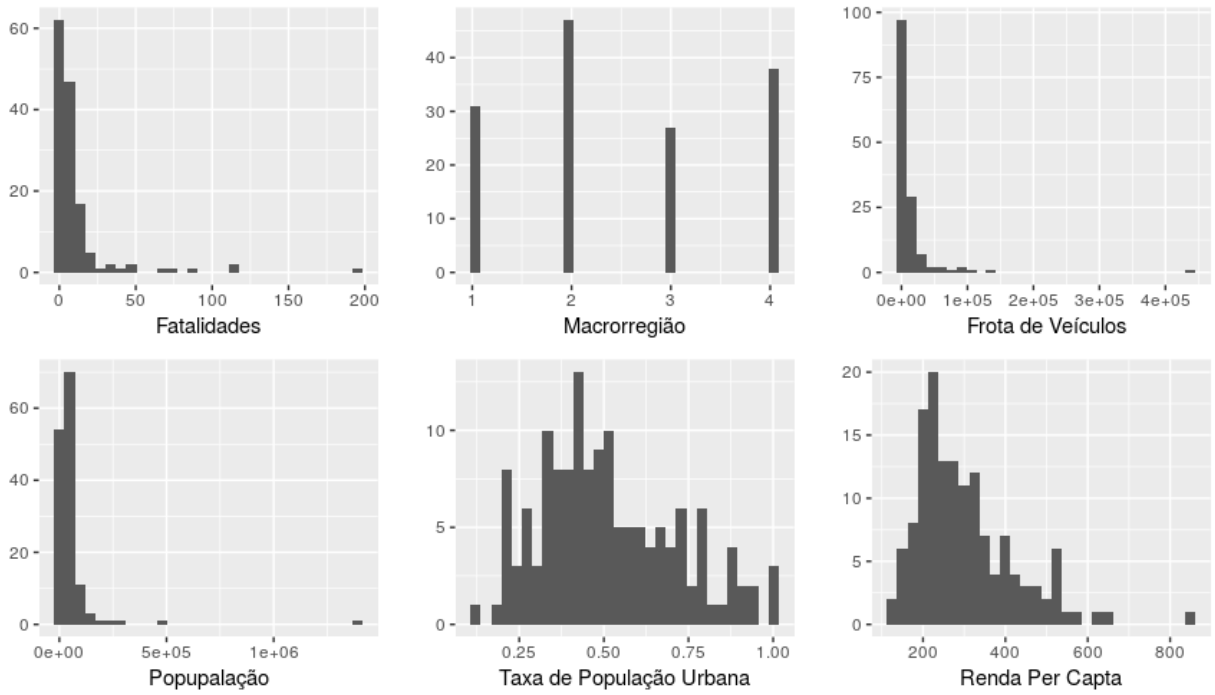

Figura 2: Histograma das variáveis utilizadas na análise.

Sendo assim, foi realizada a transformação logarítmica destas variáveis e foi verificada que esta transformação foi efetiva para correção da assimetria (Figura 3).

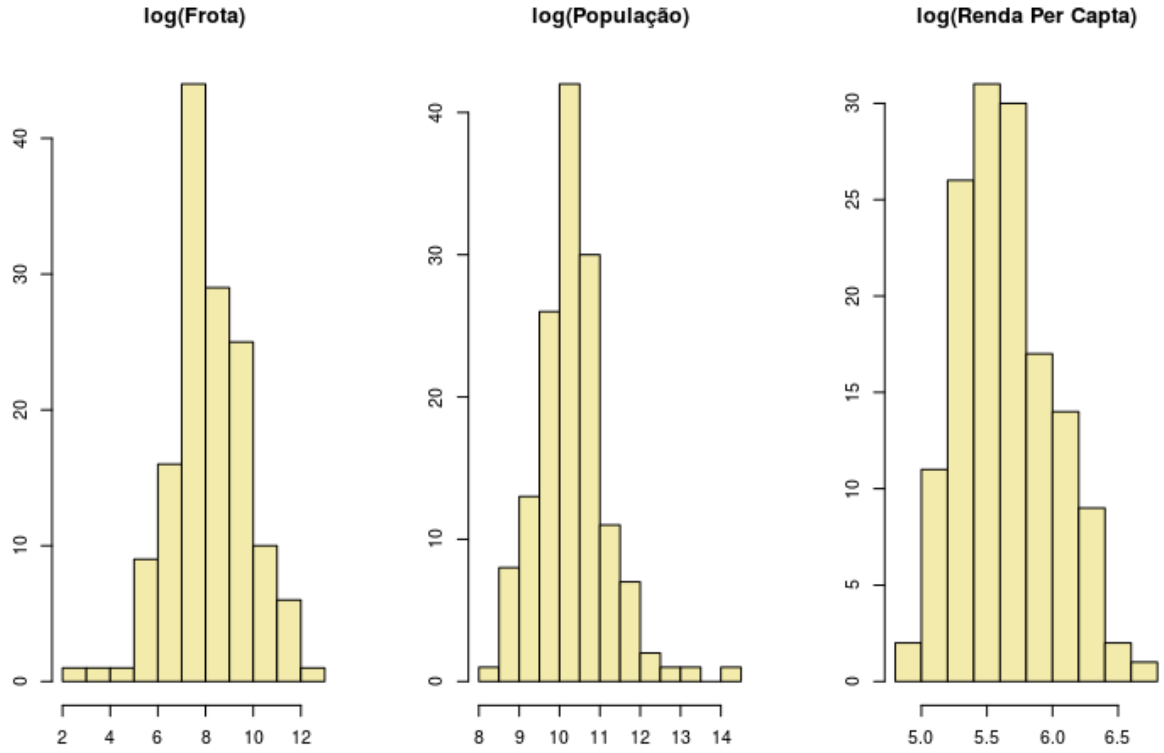

Figura 3: Histograma das variáveis com transformação logarístmica.

Foi verificada a correlação entra as variáveis através do gráfico de correlograma (Figura 4). Baseado nestes rersultados foi verificado que não houveram variáveis correlacionadas que poderiam trazer problemas para o ajuste do modelo.

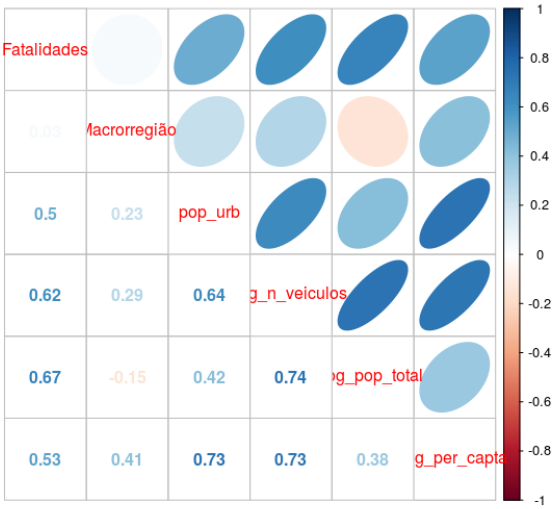

Figura 4: Correlograma das variáveis.

Foi realizado a matriz de gráficos de dispersão (Figura 5) para confirmar que não houve forte correlação das variáveis com a resposta ou entre si, sendo assim possível confirmar a ausência de corelação entre as mesmas.A variável resposta apresentou pontos discrepantes, longe das nuvens de pontos.

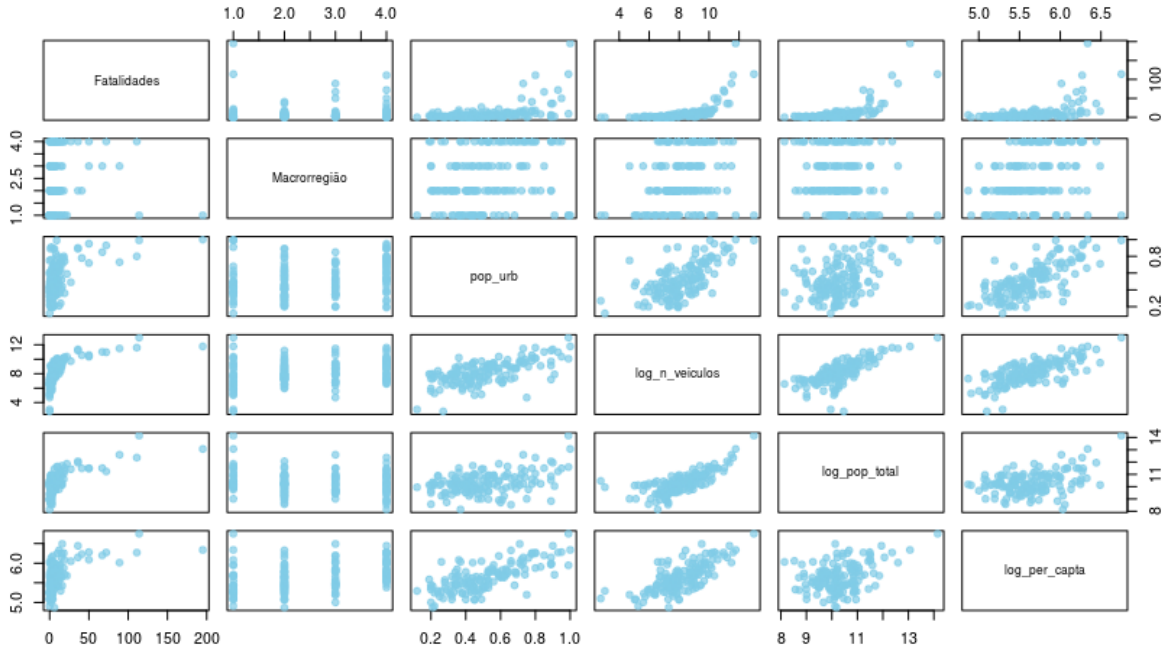

Figura 5: Matriz de dispersão.

4.2 Ajuste do Modelo de Regressão

Foram testados os modelos da família **Poisson** e **Binomial Negativa**, ambos com função de ligação logarítmica. Na tabela 2 são apresentados os resultados do ajuste para a escolha do modelo.

Tabela 2 - Ajuste dos MLG avaliados.

ajuste	aic	verossimilhança
Poisson	970.8606	-479.4303
Binomial Negativa	715.3398	-350.6699

Com base nos resultados da tabela acima, foi possível verificar que o modelo Binomial negativo apresentou melhor ajuste uma vez que que AIC foi menor e a verossimilhança foi maior. Para confirmar este resultado foram plotado os gráficos de envelope simulados (Figura 6), sendo confirmado que o modelo Binomial negativo foi o que melhor ajustou os dados uma vez que os mesmos se mantiveram dentro do limite do envelope.

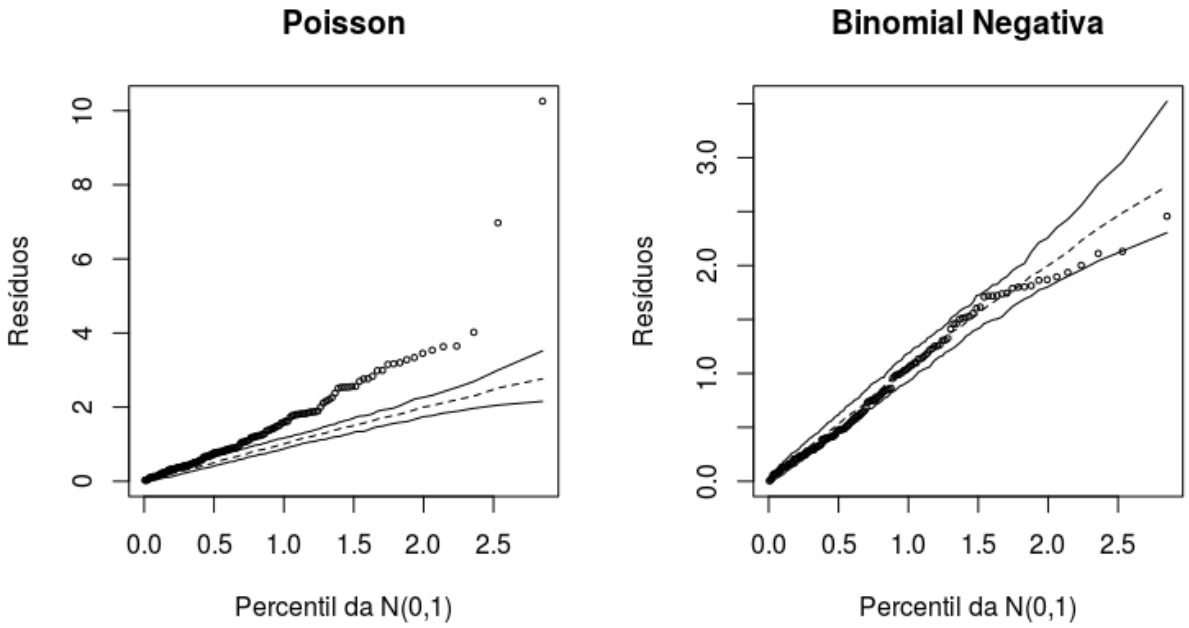

Figura 6: Gráfico de envelopes simulados dos dois modelos avaliados para ajuste.

Sendo assim, através do método *stepwise* foi realizada a seleção das covariáveis para o modelo ajustado, o mesmo é apresentado na Tabela 3.

Tabela 3 - Seleção das covariáveis do modelo.

	Estimativa	Erro Padrão
(Intercept)	-5.383	0.368
log_n_veiculos	0.881	0.054
pop_urb	-0.593	0.355

O algoritmo indica que as variáveis log do número de veículos e taxa de população urbana são significativas. A variável log_n_veículos tem relação positiva com o número de acidentes de trânsito já a variável pop_urb tem relação negativa.

Após foi realizado o teste da razão de verossimilhança a fim de comparar o modelo saturado e do reduzido (Tabela 4), onde foi verificado que os modelos foram similares, sendo assim utilizado o modelo reduzido (menos complexo) no ajuste.

Tabela 4 - Teste da razão de verossimilhança entre modelo reduzido e saturado.

Modelo	Theta	GL Residual	Verossimilhança	Estatística de teste
Binomial Negativa(Restrito)	5.6165	140	-704.5713	
Binomial Negativa(Saturado)	5.9568	137	-701.3398	0.3573

Portanto o modelo final ajustado foi:

$$y_i|\underline{x}_i \sim Binomial\ Negativa(\mu_i, \phi)$$

$$log(\mu_i) = -5.383 + 0.881 log(n\_veiculos_i) - 0,593 pop\_urb_i$$

Com os gráficos de medidas de influência, foi possível verificar que não há indicativos fortes de outliers ou observações influentes (Figura 7).

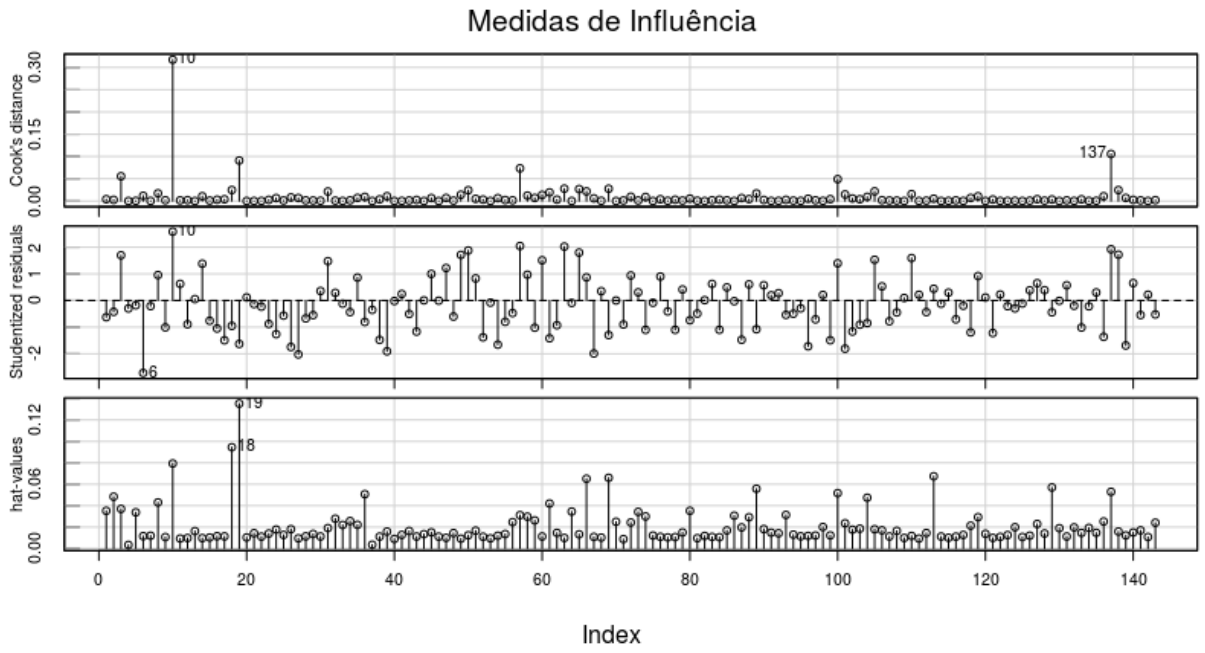

Figura 7: Medidas de influência.

Também, através dos resíduos quantílicos aleatorizados, que verifica a qualidade do ajuste. Foi possível observar que o modelo está satisfatóriamente ajustado, porém com as caudas levemente mais pesadas (Figura 8).

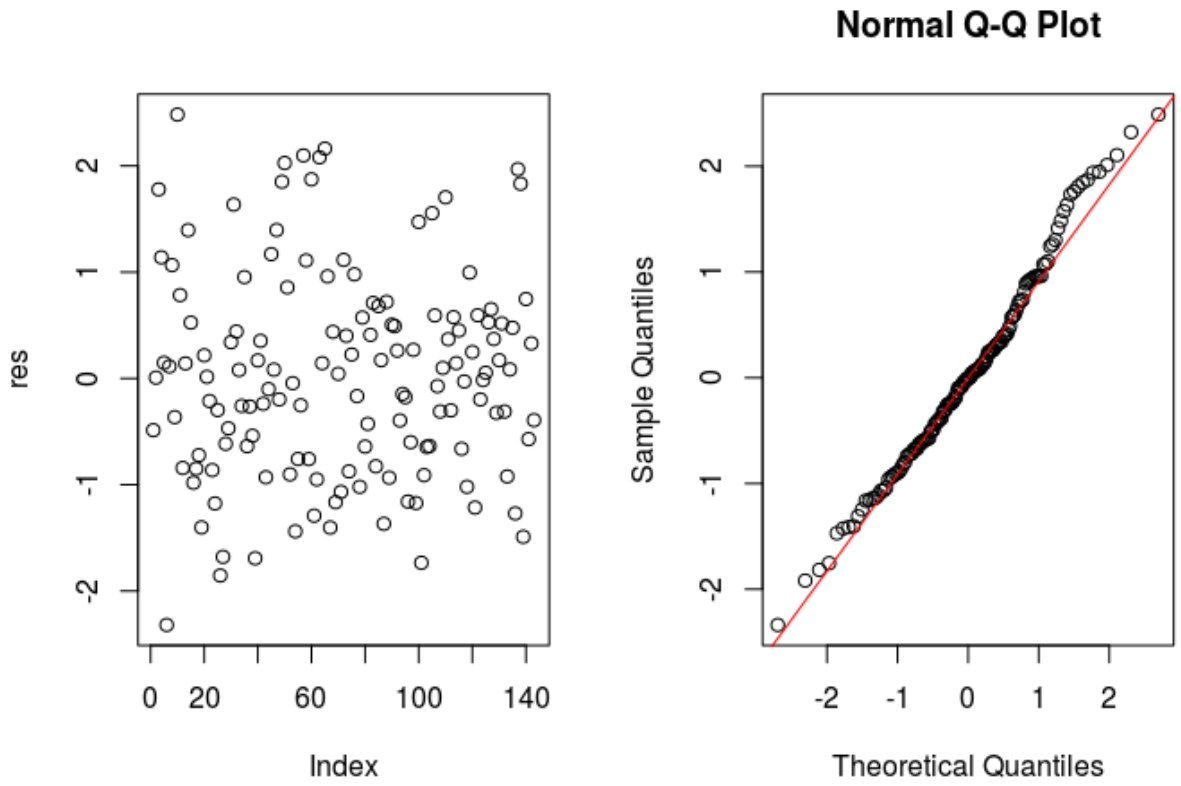

Figura 8: Gráfico dos resíduos quantílicos aleatorizados.

4.3 Gráfico de Efeitos

O gráfico de efeitos, que fornece uma vizualização do efeito das variáveis explicativas no preditor. Na Figura 9, foi observado que o comportamento de todas está dentro das bandas e do esperado para cada uma. Para o log da frota de veículos foi observado efeito positivo e para a taxa de população urbana, efeito negativo. O número de óbitos cresce para municípios com maior quantidade de veículos e menos urbanizadas.

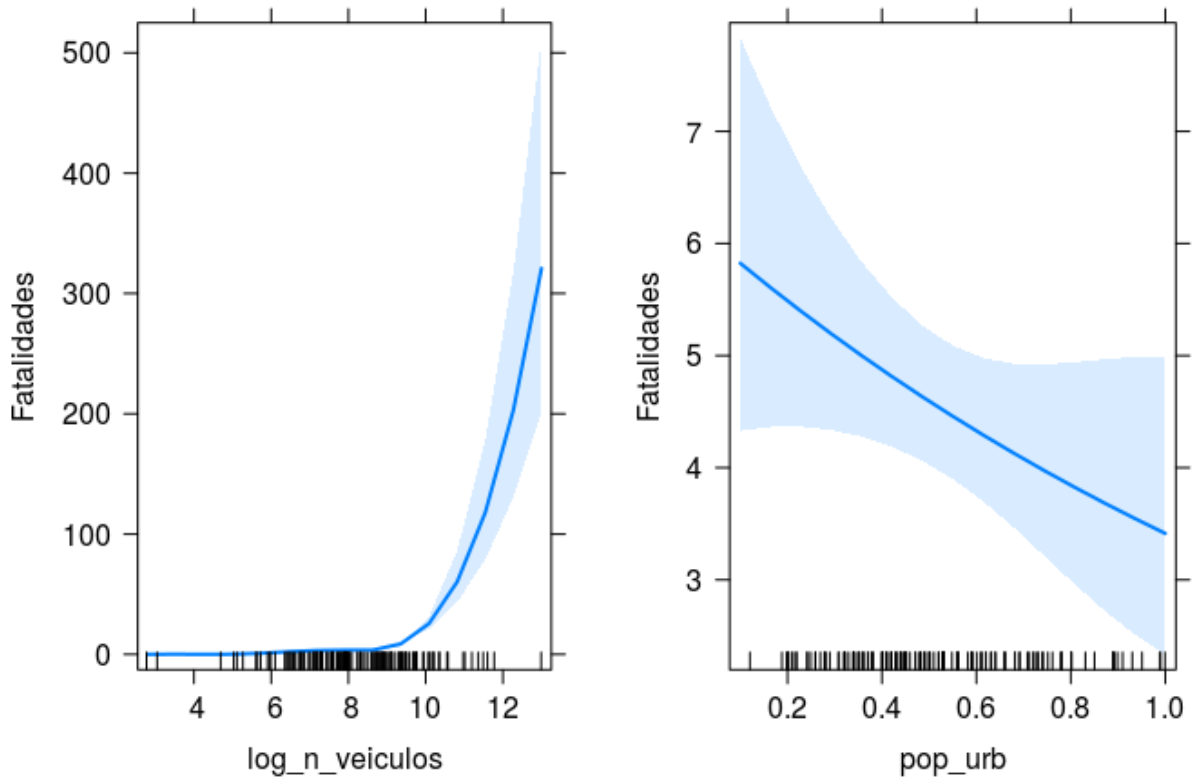

Figura 9: Gráfico de efeitos.

4.4 Predição

Após a conclusão da modelagem e ajuste, foram testados alguns perfis fictícios de municípios a fim de ilustrar quais poderiam ser as respostas para cada um deles. Para o perfil 1 foi selecionado frota com valor alto e taxa de população urbana intermediária, para o perfil 2 foi selecionado frota com valor intermediário e taxa de população urbana baixa e para o perfil 3 foi selecionado frota com valor baixo e taxa de população urbana alta. Os valores e as respostas para cada um dos perfis podem ser observados na Tabela 5. Cada uma das respostas significa o valor predito de óbitos no trânsito para caso houvesse algum município que se enquadrasse em cada perfil.

Tabela 5 - Predições baseadas em perfis de municípios.

	Número de Veículos (log)	Taxa de População Urbana	Predição
Perfil 1	12.5	0.5	208
Perfil 2	7.0	0.1	2
Perfil 3	4.0	0.9	0