UNIVERSIDADE FEDERAL DO PARANÁ

Modelos Lineares Generalizados para Dados de Contagem

Ananda Bordignon, Brendha Lima, Giovanna Lazzarin

28 de novembro de 2018

# 1. Introdução

# 2. Dados

Os dados a respeito do número de acidentes no estado de Alagoas no ano de 2016 foram extraídos do Departamento de Informática do Sistema Único de Saúde (DATASUS) e suas características sociodemográficas do último senso de 2010 foram extraídas no portal do Instituto de Pesquisa Econômica Aplicada (IPEADATA). Segundo o Batalhão de Polícia de Trânsito (BPTRAN), todo evento ocorrido na via pública (incluindo calçadas), decorrente do trânsito de veículos e pessoas, que resulta em danos humanos e/ou materiais é definido como acidente de trânsito. Compreende: colisões entre veículos, choque com objetos fixos, capotamentos, tombamentos, atropelamentos, queda de pedestres e ciclistas, etc. Além disso toda ocorrência fortuita ou danosa, envolvendo veículos em circulação, ou parados, respectivos ocupantes, pedestres e objetos móveis ou fixos.

Os dados a serem trabalhados neste estudo referem-se à acidentes de trânsito ocorridos nas vias municipais, sem incluir as Rodovias Estaduais e Federais.

Cada linha da base corresponde a 1 dos 102 municípios do estado do Alagoas, as características sociodemográficas selecionadas como variáveis explicativas foram estas:

*frota* - Frota total de veículos.

*pib* - Pib per capita a preços correntes.

*populacao* - População residente.

*emergencia* - Estabelecimentos de saúde com atendimento de emergência total.

o interesse deste trabalho é modelar o número de acidentes de trânsito em funções das demais.

As primeiras 6 linhas da base de dados tem a seguinte forma:

Tabela 1 - Primeiras observações da base de dados

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| codigos | municipios | obitos | pib | Populacao | frota | emergencia |
| 270010 | Água Branca | 5 | 24 | 19377 | 2370 | 1 |
| 270020 | Anadia | 1 | 49 | 17424 | 1718 | 1 |
| 270030 | Arapiraca | 172 | 101 | 214006 | 68913 | 9 |
| 270040 | Atalaia | 7 | 81 | 44322 | 4164 | 1 |
| 270050 | Barra de Santo Antônio | 1 | 54 | 14230 | 1134 | NA |
| 270060 | Barra de São Miguel | 1 | 98 | 7574 | 935 | NA |

# 3. Análise Descritiva

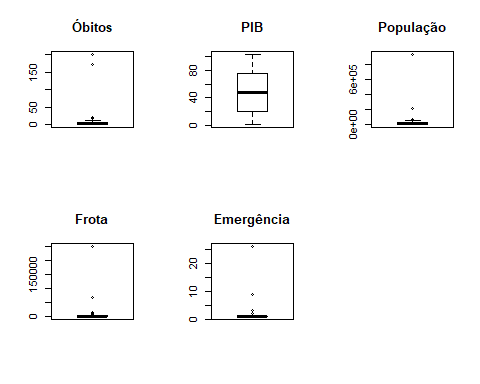
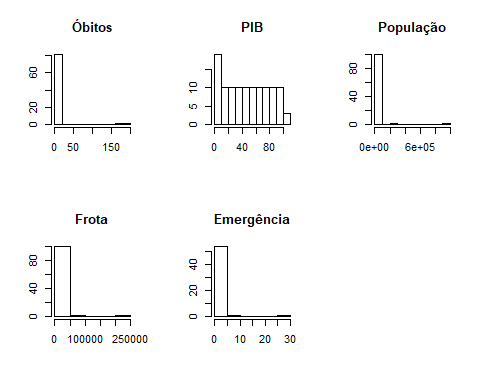
## 3.1 Estatísticas básicas

Para iniciar uma breve análise descritiva, podemos usar a função “summary”" para visualizar o mínimo, máximo, mediana e quartis das varíaveis explicativas do estudo.

Na variável “município” são elencados todos os 102 municípios do Alagoas. Para a variável resposta, nota-se um total de 23 de dados faltantes, ou seja, não foram registrados acidentes de trânsito nestes municípios.

Observamos alguns histogramas e box-plots que facilitam a visualização dos dados. Nota-se que algumas variáveis têm um ou mais pontos discrepantes. Em alguns casos pode ser conveniente trabalhar com o log da variável para obter uma maior simetria, ou, outra alternativa é remover os valores discrepantes.

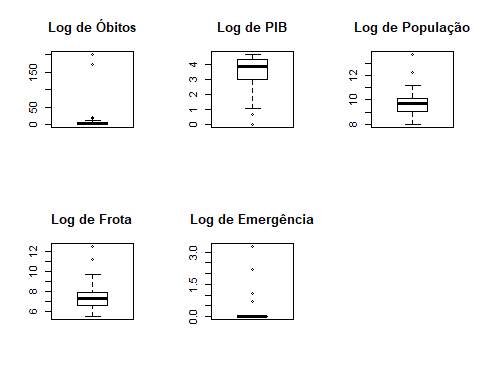
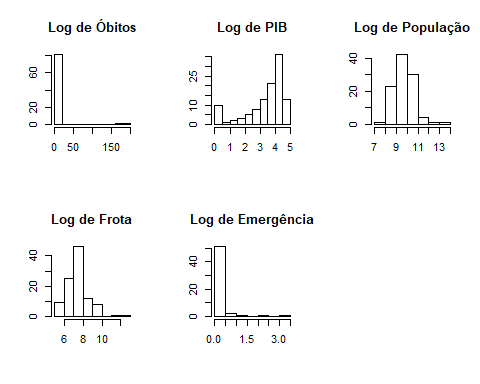
Gráfico 1 – Gráficos descritivos não ajustados



## 3.2 Transformação das covariáveis

Aplicaremos uma transformação logarítmica nas variáveis observadas como mais assimétricas nas análises descritivas e verificaremos novamente a forma da distribuição das transformações.

Gráfico 2 – Gráficos descritivos ajustados



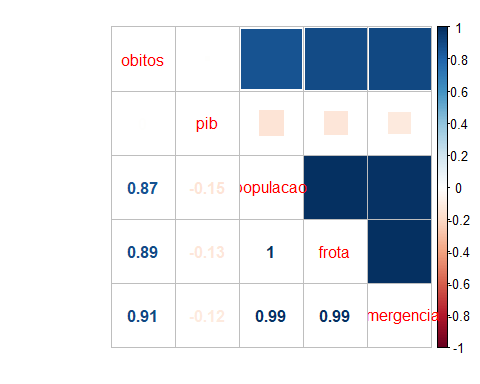
Agora podemos observar uma simetria maior nas distribuições das variáveis transformadas.

## 3.3 Correlação

Agora, verificaremos a correlação entre as variáveis em estudo, nesta etapa vamos considerar as variáveis transformadas no tópico anterior.

Como já visto, a variável resposta do estudo possui alguns dados faltantes, para resolver este problema é possível obter a correlação desta com as demais utilizando o argumento “use” da função cor.

Gráfico 3 – Gráfico de correlação



Pode ser observado que a correlação é muito forte entre log da população e log da frota, estando em 100% de proporcionalidade direta (aumento de uma implica no aumento da outra, pois valor do coeficiente de correlação é positivo) e também de população com emergência (99%) e frota com emergência (99%).

Optou-se por utilizar uma variável nova melhor aproveitando o efeito de interação entre duas, que é a combinação entre população e frota, já que são completamente correlacionadas e de forma positiva, assim criou-se a variável frota média. A matriz de correlação das variáveis melhorou muito com este ajuste e pode ser observada a nova distribuição das variáveis testadas no scaterplot.

# 4. Ajuste dos Modelos de Regressão

Neste trabalho, queremos modelar uma variável de contagem, ou seja, uma variável discreta com suporte no conjunto dos inteiros não negativos. Nossa resposta é o número de acidentes de trânsito, para problemas deste tipo, comumente a primeira alternativa de modelagem via modelo linear generalizado faz uso da distribuição Poisson com função de ligação logarítmica.

Existe a possibilidade da Poisson não se ajustar bem aos dados e à resposta, assim uma alternativa é usar uma modelagem com Binomial Negativa (usualmente nos casos em que a variância da distribuição não for igual a média, mas sim superior, a chamada superdispersão ou sobredispersão).

Segue abaixo os ajustes dos GLM com log-linear de Poisson e com a distribuição Binomial Negativa.

## 4.1 Escolhendo o Modelo

A escolha do modelo deve levar em consideração algumas métricas. Neste caso foi utilizada a verossimilhança e o AIC de cada um dos modelos testados com distribuição diferente.

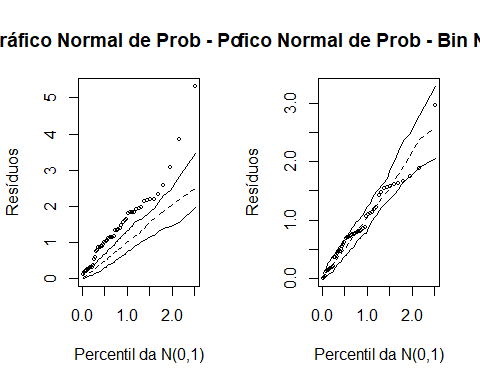
Tabela 2 - Escolha do modelo adequado à base de dados

|  |  |  |
| --- | --- | --- |
| ajuste | aic | verossimilhanca |
| m1 | 322.9249 | -157.4624 |
| m2 | 286.7599 | -138.3800 |

Acima observa-se o resultado do Critério de Informação de Akaikke (AIC), que mede a “perda de informação” ao se usar o modelo ajustado em relação aos dados reais (modelo hipotético real), portanto o que fornecer menor valor é mais indicado. A log-verossimilhança possui o mesmo propósito, porém o melhor modelo a ser escolhido é aquele que tiver o maior valor. Para este conjunto de dados a distribuição e modelo mais aderentes são da Binomial Negativa.

A seguir, obtivemos os gráficos envelopes dos dois ajustes, para confirmar qual modelo está melhor ajustado, através de seus resíduos. Deve-se observar se há presença de pontos fora dos limites ou se há pontos dentro dos limites porém apresentando padrões sistemáticos.

Gráfico 3 – Envelope simulado



## 4.2 Modelo Escolhido

O modelo escolhido é o que foi ajustado através da Binomial Negativa, que é uma distribuição de probabilidades discreta, que conta o número de tentativas necessárias para se obter k sucessos, em n ensaios de Bernoulli com probabilidade p em cada ensaio, conforme imagem abaixo.

## 4.3 Reajuste do modelo

Por conta das covariáveis correlacionadas deve-se testá-las nos modelos, inserindo uma a uma sem a presença da outra. Buscou-se não incluir ao mesmo tempo o log da população e log da frota com log da frota média, pois esta é uma combinação delas. O modelo que apresentou melhor ajuste, menor deviance residual e mais sentido prático foi este citado a seguir, indexado como “m2.1”, que utiliza informação de estabelcimentos de saúde com emergência e frota média para predizer o número médio de óbitos. Se comparado ao modelo anterior, que tinha o acréscimo do PIB (que neste caso não foi significante), o AIC e verossimilhança são muito parecidos, portanto optou-se por usar um modelo mais simples com menos variáveis.

Tabela 3 - Comparativo de AIC e Verossimilhança entre ajustes

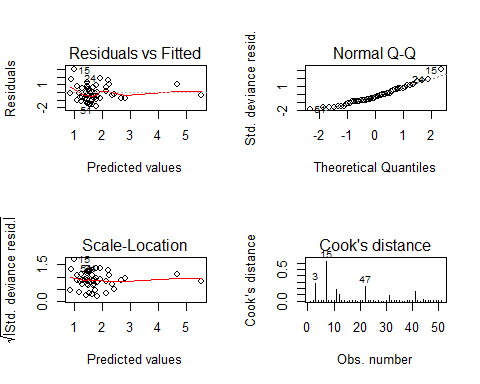
|  |  |  |
| --- | --- | --- |
| ajuste | aic | verossimilhanca |
| m2 | 286.7599 | -138.3800 |
| m2.1 | 284.9111 | -138.4555 |

Tabela 4 - Estimativas dos coeficientes do modelo ajustado

|  |  |
| --- | --- |
|  | x |
| (Intercept) | 3.5227978 |
| emergencia\_log | 0.9585852 |
| frota\_media\_log | 0.8487391 |

A seguir temos algumas considerações a cerca do ajuste do modelo, para garantir homocedasticidade, normalidade dos resíduos e adequação das estimativas.

Gráfico 4 – Gráficos de resíduos



## 5. Medidas de influência

Acima pode-se verificar que existe um ponto candidato a influente, que é a observação 15, por ter valor de distância de cook aproximadamente 0,6. Fizemos um teste retirando-a da análise para observar se havia diferença significativa entre modelos, para garantir que ela não está dando o efeito de superestimação ou subestimação de coeficientes. Concluímos não haver diferença com a retirada da amostra, e assim a mantivemos no modelo.

# 6. Interpretação

A interpretação deste modelo envolve a transformação das variáveis brutas (escala do preditor linear) para escala da resposta, em termos de suas probabilidades. Segue um exemplo abaixo de como é feita a leitura do acréscimo de uma variável, fixando outras variáveis, e qual seu impacto direto na probabilidade de óbito por acidente de trânsito.

# Conclusão

Após análise de distribuição, verificando que a Poisson não é aderente a esta população e escolhendo a Binomial Negativa como melhor aproximação, foram feitos alguns testes nas variáveis, avaliando correlação entre elas, criando uma nova variável que melhor descrevia o comportamento de duas, chegou-se a um modelo candidato. Foi necessário realizar uma “Limpeza” em sua composição, retirando um covariável pouco significativa e chegando a um modelo com intercepto, log do número de estabelecimentos de saúde e log da frota média, tendo indicativos para afirmar que ele prediz bem o número de acidentes de trânsito no Estado do Alagoas.