ISSN: 2317-0840

Regressão logística: o que leva um acidente rodoviário a ser uma tragédia?

Fernanda V. Roquim[1†], Luiz Ricardo Nakamura[2], Thiago G. Ramires[3], Renato R. Lima[4]

[1] *Universidade Federal de Lavras, Programa de Pós-Graduação em Estatística e Experimentação Agropecuária.*
[2] *Universidade Federal de Santa Catarina, Departamento de Informática e Estatística. E-mail:* luiz.nakamura@ufsc.br.
[3] *Universidade Tecnológica Federal do Paraná, Departamento de Matemática.*
E-mail: thiagogentil@gmail.com.
[4] *Universidade Federal de Lavras, Departamento de Estatística. E-mail:* rrlima@des.ufla.br.

Resumo: *Os acidentes rodoviários são uma importante questão para o Brasil, uma vez que as vias rodoviárias são a maneira mais utilizada de transporte no país. A Organização Mundial de Saúde estima que as lesões causadas nesses acidentes são a principal causa de morte entre jovens em todo o mundo, o que justifica o desenvolvimento de estudos sobre o tema. Nesse sentido, a ideia central deste trabalho foi descobrir, por meio de um modelo de regressão logística, quais são os fatores e como eles influenciam um acidente rodoviário a ter vítimas feridas ou fatais. Para tal, foram utilizados dados públicos da Polícia Rodoviária Federal sobre acidentes ocorridos no Brasil no ano de 2018. Com o modelo final apresentado foi possível realizar predições a partir das características das covariáveis, o que é muito relevante para, por exemplo, o cálculo de seguros. Ademais, foi possível identificar quais covariáveis influenciam positiva ou negativamente a probabilidade de haverem vítimas, o que pode auxiliar na criação de políticas públicas para prevenção dos mesmos.*

Palavras-chave: Modelagem; Modelos lineares generalizados; Razão de chance; Vítimas.

Abstract: *Road traffic accidents are an important issue in Brazil since road transport is a key factor in the country. The World Health Organisation estimates that road traffic injuries are the leading cause of death for children and young adults worldwide, which justifies the development of studies on the subject. Hence, the main idea of this paper was to find out, through a logistic regression model, which factors may contribute to a road traffic accident present a non-fatal or fatal injury outcome. We used a public dataset from the Polícia Rodoviária Federal about road traffic accidents in 2018 in Brazil. The final fitted model able us to perform predictions, being a great feature, for instance, to calculate insurance values. Furthermore, it was possible to identify which covariates contribute positively or negatively to the probability of non-fatal or fatal injury outcomes, which may help the development of public policies in the subject.*

Keywords: Generalised linear models; Modelling; Odds ratio; Victims.

[†]Autora correspondente: fer_venturato@yahoo.com.br

1 Introdução

De acordo com Confederação Nacional do Transporte (CNT), um acidente de trânsito pode ser definido como um evento não intencional, mas evitável, ocorrido em vias terrestres que resulte em danos à veículos e/ou à sua carga, lesão em pessoas e/ou animais (CNT, 2018). Os acidentes que ocorrem nas vias rodoviárias devem ser considerados de grande importância por dois principais motivos. O primeiro é que, atualmente, o transporte terrestre por malhas rodoviárias é a forma mais utilizada no Brasil para deslocamento de pessoas e cargas. O segundo é que acidentes rodoviários estão entre as maiores causas de perda humana em todo o mundo. A Organização Mundial de Saúde estima que as lesões causadas por estes acidentes sejam a principal causa de morte entre jovens de 15 a 29 anos, e está entre as três principais entre pessoas de 5 a 44 anos (OMS, 2015). Esses impactos sociais e econômicos podem chegar a custar até 3% do PIB de um país (CNT, 2018).

Neste cenário, a Polícia Rodoviária Federal (PRF), desde de 2007, adotou uma política de banco de dados abertos, fornecendo informações sobre os acidentes rodoviários brasileiros que podem ser utilizadas e redistribuídas livremente pelas pessoas, sem restrições de licenças ou patentes (PRF, 2017). Essa política é extremamente interessante, no sentido de que cria uma ferramenta que facilita as instituições, pesquisadores e órgãos públicos a desenvolverem pesquisas que subsidiem a criação de políticas públicas e/ou privadas de segurança, por exemplo.

O objetivo geral deste trabalho é extrair informações relevantes do banco de dados da PRF sobre todos acidentes rodoviários ocorridos no Brasil em 2018 por meio de modelos estatísticos. A questão central deste estudo é descobrir quais são os fatores associados e como eles influenciam um acidente rodoviário a ter vítimas feridas e/ou fatais, ao invés de simplesmente um acidente sem vítimas. Ainda, baseado nesses fatores, qual a chance de um determinado acidente possuir vítimas feridas e/ou fatais. Assim, utilizamos aqui o modelo de regressão logística, pertencente aos modelos lineares generalizados (MLG) (NELDER; WEDDERBURN, 1972), uma vez que a variável resposta em estudo é dicotômica, isto é, pode assumir valores zero (vítimas ilesas) ou um (vítimas feridas e/ou fatais).

2 Banco de dados

Os dados referentes aos acidentes rodoviários no Brasil estão disponíveis no banco de dados oficial da Polícia Rodoviária Federal, na seção Dados Abertos, os quais podem ser obtidos em: *https://www.prf.gov.br/portal/dados-abertos/acidentes*. No total, foram registradas 69.206 ocorrências de acidentes entre as datas de 01/01/2018 à 31/12/2018, das quais foram eliminadas 114 observações (0,16% do total dos dados) da análise por terem dados faltantes.

2.1 Variável resposta

A variável *classificação do acidente* foi selecionada dentro do banco de dados supracitado. Inicialmente três níveis foram observados na variável: vítimas fatais, vítimas feridas e sem vítimas. Como o objetivo principal do trabalho é estudar quais fatores potencializam a chance de ocorrência de um acidente com vítimas feridas e/ou fatais, uma recodificação

foi realizada. Assim, as classificações vítimas fatais e vítimas feridas observadas foram alocadas em um único nível ($Y = 1$ ou $Y =$ `sim`) e acidentes sem vítimas em outro nível ($Y = 0$ ou $Y =$ `não`), gerando então uma nova variável dicotômica denominada *Vítimas* (Y).

Do total de acidentes rodoviários ocorridos em 2018, em todo território nacional, 53.905 tiveram vítimas feridas ou fatais, enquanto que em apenas 15.187 as pessoas saíram completamente ilesas, isto é, em aproximadamente 78% dos acidentes computados pela PRF foram observadas vítimas feridas ou fatais.

2.2 Variáveis explicativas

Foram selecionadas nove variáveis explicativas do banco de dados da PRF, das quais algumas foram recodificadas a fim de simplificar a estimação e interpretação dos modelos ajustados. As variáveis explicativas são descritas detalhadamente a seguir.

- **Causa do acidente:** variável categórica que identifica qual foi a causa principal do acidente, possuindo quatro diferentes níveis. O primeiro é *falha humana*, que engloba condições de carga excessiva e/ou mal acondicionada, desobediência das normas de trânsito pelo condutor ou pedestres, falta de atenção do condutor ou do pedestre, condutor dormindo, ingestão de álcool e/ou substancias ilícitas pelo condutor ou pelo pedestre, ultrapassagem indevida, velocidade incompatível, não guardar distância de segurança e não acionamento do sistema de iluminação do veículo. O segundo é *falha mecânica*, que identifica como causa do acidente algum defeito mecânico no veículo ou avarias e desgaste excessivo no pneu. O terceiro é *falha na via*, que agrega os acidentes causados por algum defeito na via, algum objeto estático sobre a via, pista escorregadia e sinalização da via insuficiente ou inadequada. Por último tem-se o nível *outros*, que recebe as causas de acidente por animais na pista, fenômenos da natureza, restrição de visibilidades e outros. No modelo esta variável foi identificada como `CausaAcidente`. Na Tabela 1 são disponibilizadas as quantidade de acidentes ocorridos por cada causa. Cabe destacar que mais de 80% dos acidentes são causados por falha humana.

- **Condição meteorológica:** variável categórica que identifica qual era a condição meteorológica no local do acidente. Possui, ao todo, cinco níveis. O primeiro, denominado *chuva*, que engloba chuva, garoa, chuvisco e granizo. O segundo é *sol*, que indica se o céu estava claro ou ensolarado. O terceiro é *neblina*, que indica se havia neblina ou nevoeiro no local. O quarto nível é *nublado* e, finalmente, o quinto, *outros*. No modelo esta variável foi identificada como `Clima`. A Tabela 1 apresenta as condições meteorológicas observadas nos locais dos acidentes. Aqui, aproximadamente 63% dos acidentes ocorreram em condições de *sol*.

- **Dia da semana:** variável categórica com sete níveis que indica em qual dia da semana ocorreu o acidente. No modelo esta variável foi identificada como `DiaSemana`. Na Tabela 1 temos a quantidade de acidentes que ocorreram em cada dia da semana. Note que os valores são relativamente maiores para os dias de final de semana.

- **Fase do dia:** variável categórica com quatro níveis que indica qual era a fase do dia, dentre *amanhecer*, *anoitecer*, *pleno dia* ou *plena noite*, em que o acidente ocorreu. No modelo esta variável foi identificada como `FaseDia`. Na Tabela 1 podemos ver

Tabela 1: Número absoluto de acidentes (e seus respectivos percentuais) pelos níveis de cada variável explicativa categórica em estudo

Variável	Níveis				
`CausaAcidente`	Falha Humana 56.425 (81,67%)	Falha Mecânica 4.718 (6,83%)	Falha na Via 4.472 (6,47%)	Outros 3.477 (5,03%)	
`Clima`	Sol 43.676 (63,21%)	Chuva 11.087 (16,04%)	Neblina 595 (0,86%)	Nublado 12.536 (18,14%)	Outros 1.198 (1,73%)
`DiaSemana`	Domingo 11.253 (16,28%)	Segunda 9.398 (13,60%)	Terça 8.629 (12,48%)	Quarta 8.804 (12,74%)	Quinta 9.195 (13,30%)
	Sexta 10.601 (15,34%)	Sábado 11.212 (16,22%)			
`FaseDia`	Amanhecer 3.311 (4,79%)	Anoitecer 3.741 (5,41%)	Plena Noite 23.765 (34,39%)	Pleno dia 38.275 (55,39%)	
`Pista`	Simples 35.241 (51,00%)	Dupla 28.263 (40,90%)	Múltipla 5.588 (8,08%)		
`Regiao`	Norte 4.227 (6,11%)	Nordeste 14.815 (21,44%)	Centro-oeste 8.466 (12,25%)	Sudeste 20.739 (30,01%)	Sul 20.845 (30,16%)
`Solo`	Rural 38.803 (56,16%)	Urbana 30.289 (43,83%)			
`TipoAcidente`	Atropelamento 4.598 (6,65%)	Capotamento 11.211 (16,22%)	Colisão 40.841 (59,11%)	Incêndio 926 (1,34%)	Saída De Leito 11.305 (16,36%)
	Outros 211 (0,30%)				

a distribuição dos acidentes a partir da fase do dia. Mais da metade (55,39%) dos acidentes acontecem em pleno dia, o que é relativamente esperado uma vez que o fluxo de carros nesta fase do dia é maior.

- **Pista:** variável categórica com três níveis, indicando qual era o tipo de pista no local do acidente. Os níveis são *simples*, *dupla* ou *múltipla*. No modelo esta variável foi identificada como `Pista`. Como pode ser observado na Tabela 1, a maioria dos acidentes ocorre em pista simples (51,00%).

- **Região:** variável categórica com cinco níveis, indicando em qual região geográfica do país ocorreu o acidente, isto é, regiões *norte*, *nordeste*, *sul*, *sudeste* ou *centro-oeste*. No modelo esta variável foi identificada como `Regiao`. Na Tabela 1 temos a quantidade de acidentes que ocorreram em cada região geográfica. Note que os maiores valores observados são das regiões sul (30,16%) e sudeste (30,01%), ao passo que a região norte apresentou o menor percentual de acidentes (apenas 6,11% do total). Vale ressaltar que no presente artigo não foi considerada a quantidade de rodovias existentes em cada uma das regiões.

- **Solo:** variável binária, que indica se o acidente ocorreu em perímetro *urbano* ou *rural*. No modelo esta variável foi identificada como `Solo`. Na Tabela 1 podemos observar que a maioria dos acidentes aconteceram em zonas rurais (56,16%).

- **Tipo de acidente:** variável categórica que identifica qual foi o tipo de acidente. Possui seis diferentes níveis. O primeiro é *atropelamento*, que inclui atropelamento de pedestres ou animais. O segundo é *colisão*, que inclui colisão com objeto estático ou em movimento; colisão lateral, traseira, frontal ou transversal; e engavetamento. O terceiro é *capotamento*, que engloba capotamentos (automóveis), tombamentos e

derramamento de carga (caminhões) e queda de ocupante de veículo (motocicletas). O quarto é *incêndio*. O quinto é *saída de leito*. E o último *outros*. No modelo esta variável foi identificada como `TipoAcidente`. Na Tabela 1 podemos observar que a maioria dos acidentes ocorrem por colisão (16,22%).

- **Veículos:** a única variável explicativa do conjunto de dados que não é categórica, indicando a quantidade total de veículos envolvidos no acidente. De posse das observações, temos que, em média, 1,65 veículos participaram dos acidentes em 2018, com um desvio padrão de 0,75 veículos. Ainda, o número mínimo de veículos envolvidos em acidentes é um, ao passo que o número máximo é igual a 16.

Na Figura1 são disponibilizados os relacionamentos entre cada uma das variáveis explicativas e a variável resposta por meio dos gráficos de mosaico (HARTIGAN; KLEINER, 1984). A largura das barras representa a proporção de observações daquele fator dentre o total. Os retâgulos cinza claro e cinza escuro representam a proporção de acidentes com vítimas (feridas e fatais) e sem vítimas, respectivamente.

Em relação à causa do acidente, observa-se que, falha mecânica é a causa menos perigosa dentre todas. Uma relação possivelmente controversa pode ser observada na figura referente às condições climáticas (painel no extremo superior à direita). Segundo o gráfico, a maior proporção (com exceção do nível outros) de acidentes com vítimas feridas ou fatais se dá quando está sol ou tempo limpo no dia do acidente. Esse resultado, apesar de inusitado, se tratado com maior cautela pode fazer sentido, uma vez que, em condições climáticas adversas (como dias de chuva), em geral, o condutor de um veículo dirige com maior cautela e, logo, se ele se envolve em algum acidente, as chances desse acidente serem graves diminui.

Adicionalmente, ainda baseado na Figura 1, a proporção de acidentes com vítimas feridas ou fatais parece ser a mesma em todos os dias da semana. A fase do dia com mais ocorrência de acidentes com vítimas é anoitecer e o tipo de pista mais seguro é pista dupla. A região centro-oeste, dentre todas, é a que possui, proporcionalmente, maior contagem de acidentes com vítimas ilesas. Por fim, nas zonas urbanas ocorrem mais acidentes com vítimas feridas ou fatais, e o tipo de acidente maior risco de vítimas nesta classe, como esperado, é o atropelamento.

3 Regressão logística

Como mencionado na Seção 2.1, a variável resposta em estudo *Vítimas* possui duas possíveis respostas, isto é, trata-se de uma variável binária ($Y = 0$ ou 1). Para identificar quais e como os fatores descritos na seção anterior influenciam a ocorrência de vítimas feridas ou fatais em acidentes utilizamos o modelo de regressão logístico, um caso particular dos modelos lineares generalizados descritos por Nelder e Wedderburn (1972). Matematicamente, seja $P(Y = 1|\boldsymbol{X}) = \boldsymbol{\mu}$, o modelo logístico pode então ser escrito como

$$\boldsymbol{\eta} = \text{logit}\boldsymbol{\mu} = \boldsymbol{X\beta}, \tag{1}$$

em que $\text{logit}(\boldsymbol{\mu}) = \log(\boldsymbol{\mu}/(1-\boldsymbol{\mu}))$, $\boldsymbol{X}$ é a matriz representando as variáveis explicativas e $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^\top$ é o vetor de parâmetros associado às variáveis explicativas. Se estivermos interessados na razão de chances referente ao incremento de uma unidade em

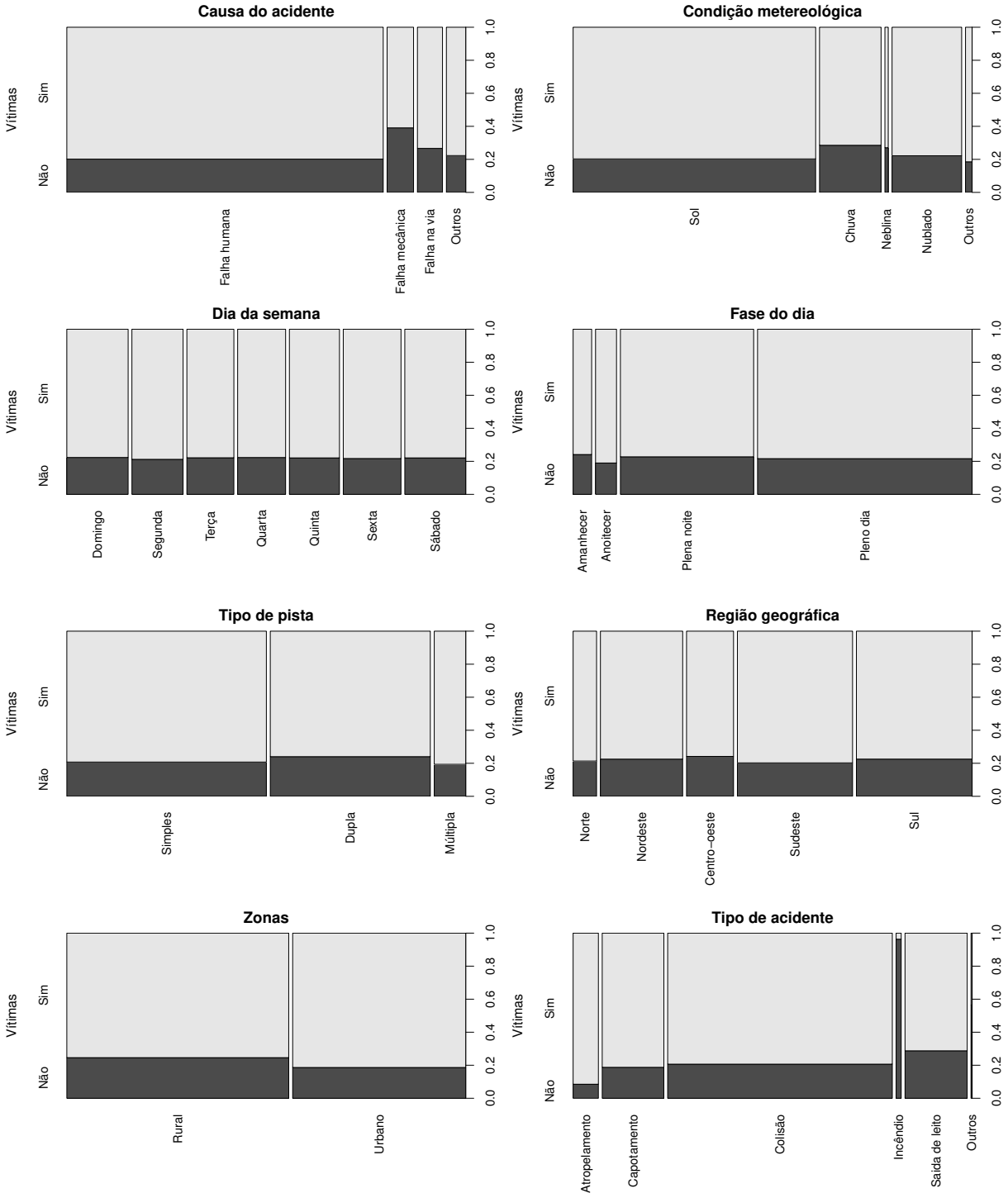

Figura 1: Gráficos de mosaico das variáveis explanatórias categóricas *vs* variável resposta.

uma variável explicativa específica, digamos X_r, considerando todas as outras como fixas, basta calcularmos $\exp\{\beta_r\}$.

A estimação do vetor de parâmetros $\boldsymbol{\beta}$ é realizada por meio do método de máxima verossimilhança. Computacionalmente, foram utilizadas as funções disponíveis no pacote `gamlss` (STASINOPOULOS; RIGBY, 2007) no R (R CORE TEAM, 2018). O pacote em questão foi utilizado pois possibilita o cálculo direto dos resíduos quantílicos aleatórios

normalizados (DUNN; SMYTH, 1996) dados por

$$\hat{r}_i = \Phi^{-1}(\hat{u}_i),$$

em que Φ^{-1} é a inversa da função acumulada da distribuição normal padrão e $\hat{u}_i$ é o resíduo quantílico estimado, sendo, nos casos em que distribuições discretas são utilizadas, um valor aleatório de uma distribuição uniforme com suporte $[\hat{u}_1, \hat{u}_2] = [F(y-1)|\hat{\boldsymbol{\mu}}, F(y|\hat{\boldsymbol{\mu}})]$ ($F(\cdot)$ corresponde à função acumulada da distribuição binomial). O grande atrativo para a utilização deste tipo de resíduo é que, independentemente da distribuição associada à variável resposta, se o modelo ajustado é adequado ou razoável, os resíduos, necessariamente, seguirão uma distribuição normal padrão. Observe que, os resíduos devem ser calculados algumas vezes para se verificar possíveis padrões ou problemas no ajuste do modelo, uma vez que os mesmos são aleatorizados.

Finalmente, como ferramente gráfica, os resíduos podem ser exibidos por meio de um *worm plot* (VAN BUUREN; FREDRIKS, 2001). O interessante deste gráfico em específico é que ele é capaz de nos fornecer uma série de características sobre o ajuste (para informações completas, consulte Stasinopoulos et al., 2017). O resultado desejado é que não haja pontos fora das bandas de 95% de confiança e que não haja nenhum tipo de tendência no gráfico, seja ela linear, quadrática ou cúbica.

4 Resultados e discussão

Para a estimação do modelo mais adequado ao conjunto de dados, prosseguiu-se com a aplicação do método de seleção de variáveis *stepwise*. Constatou-se que, com exceção da variável referente ao dia da semana em que o acidente ocorreu, todas as outras foram incluídas no modelo. Assim, o modelo logístico final ajustado pode ser escrito como

$$\begin{aligned} \boldsymbol{\eta} = \text{logit}(\boldsymbol{\mu}) &= \beta_0 + \texttt{CausaAcidente} + \texttt{Clima} + \texttt{FaseDia} + \texttt{Pista} + \texttt{Regiao} \\ &+ \texttt{Solo} + \texttt{TipoAcidente} + \texttt{Veiculos}. \end{aligned} \tag{2}$$

Na Tabela 2 estão disponíveis as estimativas e respectivos erros-padrão dos coeficientes ajustados para o modelo (2). Ressaltamos que apenas os níveis `Centro-Oeste` e `Sul` da variável Região e o nível `Plena Noite` da variável Fase do dia não foram significativos ao nível de 5% de significância.

Os sinais dos coeficientes apresentados na Tabela 2 influenciam diretamente a probabilidade de haver vítimas feridas ou fatais em um determinado acidente. O sinal positivo do coeficiente aumenta essa probabilidade, ao passo que um sinal negativo implica na diminuição da mesma. Por exemplo, os sinais dos coeficientes da covariável pista dupla e múltipla foram negativos, indicando que em pista simples acontecem mais acidentes com vítimas feridas ou fatais, justificando a necessidade de obras para duplicação de rodovias no sentido de aumentar a segurança na estrada.

Sendo assim, baseado ainda no sinal dos coeficientes, podemos dizer que a região Sudeste é a que apresenta os maiores riscos de acidente com vítimas feridas ou fatais. Mais precisamente, se considerarmos todas as outras variáveis explicativas como fixas, se um acidente ocorrer na região Sudeste, espera-se que a razão de chances aumente por um fator de $\exp\{0,22\} = 1,25$, utilizando a região Norte como base.

Algo um tanto quanto inusitado são os coeficientes calculados para cada nível da variável Clima, em que o dia de Sol foi utilizado como base. Com exceção do nível Outros

Tabela 2: Coeficientes estimados para o modelo logístico e seus respectivos erros-padrão (em parênteses)

Variável	Nível	Estimativa	Variável	Nível	Estimativa
Intercepto	–	2,31 (0,08)	Regiao	Norte	0,00
CausaAcidente	Falha humana	0,00		Nordeste	-0,08 (0,04)
	Falha mecânica	-0,28 (0,04)		Centro-oeste	-0,03 (0,05)
	Falha na via	-0,12 (0,04)		Sudeste	0,22 (0,04)
	Outros	-0,28 (0,05)		Sul	0,01 (0,04)
Clima	Sol	0,00	Solo	Rural	0,00
	Chuva	-0,39 (0,03)		Urbano	0,27 (0,02)
	Neblina	-0,26 (0,10)	TipoAcidente	Atropelamento	0,00
	Nublado	-0,12 (0,03)		Capotamento	-0,88 (0,06)
	Outros	0,17 (0,08)		Colisão	-1,22 (0,06)
FaseDia	Amanhecer	0,00		Incêndio	-5,47 (0,19)
	Anoitecer	0,25 (0,06)		Saída de leito	-1,37 (0,06)
	Plena noite	0,02 (0,05)		Outros	-2,63 (0,15)
	Pleno dia	0,13 (0,04)	Veiculos	–	0,11 (0,02)
Pista	Simples	0,00			
	Dupla	-0,29 (0,02)			
	Múltipla	-0,17 (0,04)			

dessa variável, todos os outros possuem sinal negativo, o que implica que há maior chance de um acidente com vítimas feridas ou fatais em um dia de sol. Por exemplo, se fixarmos todas as demais covariáveis, em dias de chuva, as chances de haver uma vítima ferida ou fatal diminui por um fator de $\exp\{-0,39\} = 0,68$, isto é, diminui aproximadamente 32%. Como apresentado na Seção 2.2, esse resultado apesar de inusitado parece fazer sentido, uma vez que, em condições climáticas adversas (como dias de chuva), em geral, o condutor de um veículo dirige com maior cautela e, logo, se ele se envolve em algum acidente, as chances desse acidente serem graves diminui.

Em relação à variável veículos, considerando todas as outras variáveis explicativas como fixas, espera-se que a razão de chances aumente por um fator de $\exp\{0,11\} = 1,12$, isto é, a cada aumento de um veículo envolvido no acidente, espera-se que a chance de ocorrerem vítimas feridas ou fatais aumente em 12%. No que tange às outras variáveis explicativas, aquelas que apresentam maior probabilidade de um dado acidente possuir vítimas feridas ou fatais são acidentes causados por falha humana, ao anoitecer, em perímetro urbano e que envolvam um atropelamento.

Para facilitar o cálculo da probabilidade do acidente apresentar vítimas feridas ou fatais, obtemos a inversa da equação(1). Isto é,

$$\boldsymbol{\eta} = \log\left(\frac{\boldsymbol{\mu}}{1-\boldsymbol{\mu}}\right) = \boldsymbol{X\beta} \quad \Longrightarrow \quad \boldsymbol{\mu} = \frac{\exp(\boldsymbol{\eta})}{1+\exp(\boldsymbol{\eta})} = \frac{\exp(\boldsymbol{X\beta})}{1+\exp(\boldsymbol{X\beta})}.$$

Assim, dadas diferentes características apresentadas durante o dia de um acidente, podemos calcular diretamente a probabilidade do acidente apresentar vítimas feridas ou fatais. Consideremos o seguinte exemplo: o tipo de acidente ocorrido foi uma colisão em uma pista dupla devido à uma falha mecânica envolvendo dois veículos em pleno dia, durante uma chuva em perímetro urbano na região sudeste. Logo, para nosso exemplo,

baseado nos coeficientes obtidos na Tabela 2 temos que $\eta = 0,97$, logo,

$$\mu = \frac{\exp(0,97)}{1+\exp(0,97)} = 0,73.$$

Isto é, a probabilidade de haverem vítimas feridas ou fatais em um acidente considerando essas características é de 73%, ou seja, a cada 100 acidentes ocorridos em tais condições, espera-se que em 73 deles hajam vítimas feridas ou fatais. Observe que é simples calcular esta probabilidade para quaisquer combinações de características de um dado acidente.

Finalmente, em relação à adequação do modelo, construímos o *worm plot* do modelo ajustado utilizando seis aleatorizações (para mais informações, consultar Stasinopoulos et al., 2017) dos resíduos quantílicos aleatórios normalizados (Figura 2). Os pontos em cinza representam os resíduos obtidos em cada uma das aleatorizações, ao passo que os de cor preta são as médias destes pontos. Podemos observar que o gráfico apresenta o comportamento desejado, isto é, aparentemente não há nenhuma tendência no gráfico e os pontos encontram-se, em sua maioria, dentro das bandas de 95% de confiança, indicando assim que o modelo ajustado é razoável.

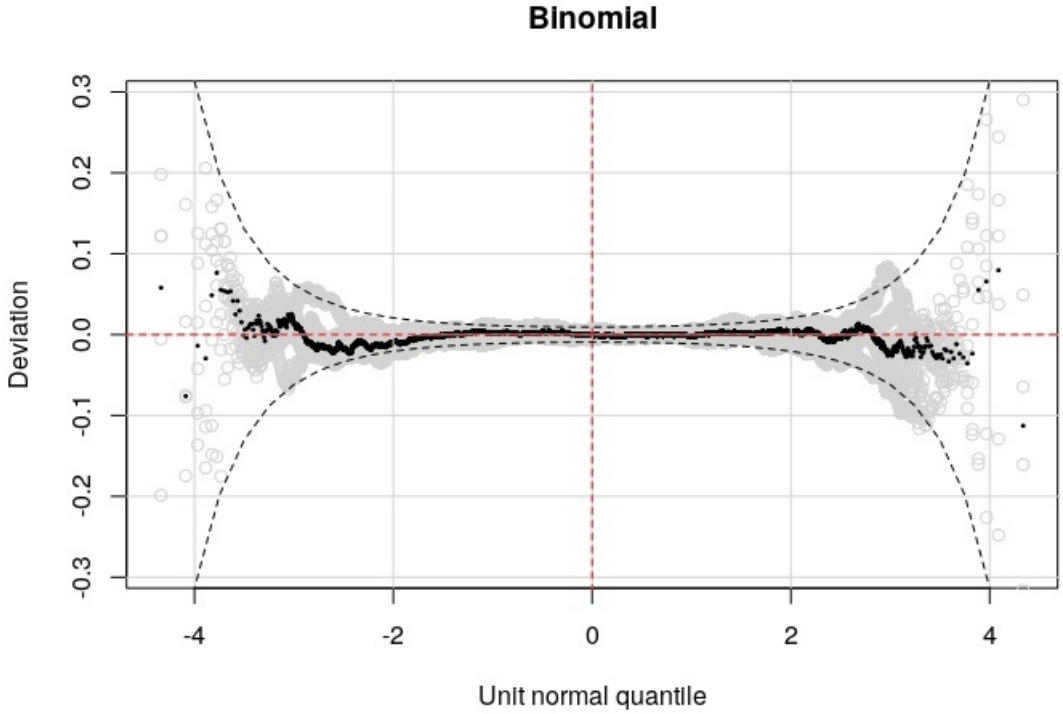

Figura 2: *Worm plot* do modelo binomial ajustado aos dados

5 Considerações finais

A utilização de modelos de regressão pode ser extremamente útil, pois, com um modelo bem ajustado, podemos captar uma série de informações sobre o fenômeno. Com o modelo final apresentado é possível realizar predições a partir das características das covariáveis, o que é muito relevante para cálculos de seguros. Ademais, foi possível identificar se as covariáveis influenciam positiva ou negativamente a probabilidade de haverem vítimas, o que pode auxiliar na criação de políticas públicas e/ou privadas ligadas ao assunto.

Agradecimentos

A primeira autora agradece à Coordenação de Aperfeiçoamento Pessoal de Nível Superior – CAPES pelo auxílio financeiro.

Referências Bibliográficas

CNT, Confederação Nacional do Transporte. *Acidentes rodoviários e a infraestrutura.* Brasília : CNT, 2018. 132 p.

DUNN, P. K.; SMYTH, G. K. *Randomized quantile residuals.* Journal of Computational and Graphical Statistics, Taylor e Francis, v. 5, n. 3, p. 236?244, 1996.

HARTIGAN, J.A., KLEINER, B. *A mosaic of television ratings.* The American Statistician, 38, 32?35. 1984.

NELDER, J. A.; WEDDERBURN, R. W. M. *Generalized linear models.* Journal of the Royal Statistical Society: Series A (General), v. 135, p. 370–384, 1972.

OMS, Organização Mundial da Saúde. *Relatório global sobre o estado da segurança viária.* 2015. 16p.

PRF, Polícia Rodoviária Federal. *Dicionário de Variáveis – Acidente.* 2017. 4p. Disponível em: https://www.prf.gov.br/portal/dados-abertos/acidentes/dicionario-de-variaveis. Acesso em: 11 de março de 2019.

R Core Team. *R: A Language and Environment for Statistical Computing.* Vienna, Austria, 2018. Disponível em: http://www.R-project.org.

STASINOPOULOS, D. M.; RIGBY, R. A. *Generalized additive models for location scale and shape (GAMLSS) in R.* Journal of Statistical Software, v. 23, p. 1–10. 2007.

STASINOPOULOS, M. D.; RIGBY, R. A.; HELLER, G. Z.; VOUDOURIS, V.; DE BASTIANI, F. *Flexible Regression and Smoothing: Using GAMLSS in R.* [S.l.]: CRC Press, 2017.

VAN BUUREN, S.; FREDRIKS, M. *Worm plot: a simple diagnostic device for modelling growth reference curves.* Statistics in medicine, Wiley Online Library, v. 20, n. 8, p. 1259–1277, 2001.