Universidade do Minho
Escola de Ciências

Jair Monteiro dos Santos

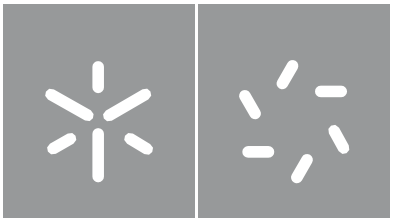
Modelos Lineares Generalizados Mistos: Aplicação a Dados de Acidentes Rodoviários em Autoestradas

outubro 2023

Modelos Lineares Generalizados Mistos:
Aplicações a Dados de Acidentes Rodoviários em Autoestradas

Jair Monteiro dos Santos

UMinho | 2023



Universidade do Minho
Escola de Ciências

Jair Monteiro dos Santos

Modelos Lineares Generalizados Mistos: Aplicação a Dados de Acidentes Rodoviários em Autoestradas

Dissertação de Mestrado em Estatística

Trabalho efetuado sob a orientação das Professoras
Doutora Susana de Sá Faria
Doutora Elisabete Fraga de Freitas

Outubro de 2023

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.

Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

Licença concedida aos utilizadores deste trabalho

Guimarães, outubro de 2023

_____

(Jair Monteiro dos Santos)

Agradecimentos

À professora Susana Faria, pela paciência e dedicação durante todo o desenvolvimento da dissertação. Muito obrigado professora.

À professora Elisabete Freitas pela disponibilidade e paciência.

À Leidy Barón pela disponibilidade e paciência na organização dos dados.

À minha mãe, por toda confiança e incentivo que permitiram alcançar esse objetivo.

Aos amigos Nilson Moreira e Nito Furtado pela colaboração, incentivo, amizade e momentos de descontração proporcionados.

Um muito obrigado a todos os Professores do Mestrado em Estatística.

Finalmente, um agradecimento à todas as pessoas que de uma forma ou de outra contribuíram para essa caminhada.

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Guimarães, outubro de 2023

_____

(Jair Monteiro dos Santos)

"Em Deus nós confiamos; todos os outros devem trazer dados."

William Edwards Deming

Resumo

Os acidentes rodoviários são considerados um importante problema de saúde pública a nível mundial. Segundo o relatório da *Organização Mundial da Saúde*, os acidentes rodoviários são a principal causa de morte dos jovens em todo mundo, especialmente entre jovens do sexo masculino.

A modelação do número de acidentes rodoviários tem como objetivo analisar e compreender os fatores que contribuem para a ocorrência de acidentes de trânsito. A investigação sobre esses fatores que influenciam a ocorrência de acidentes rodoviários visa reduzir o seu número, bem como, identificar medidas eficazes de prevenção.

Este trabalho tem como objetivo principal aplicar os *Modelos Lineares Generalizados Mistos* (em inglês: GLMM), para estudar o efeito de diferentes variáveis que definem o estado de pavimento, a geometria da estrada, o volume do tráfego e entre outras variáveis na sinistralidade rodoviária. Os dados utilizados são referentes a segmentos da autoestrada A4 da Região Grande Porto, Portugal. O modelo desenvolvido, Modelo Linear Generalizado Misto para resposta sob a forma de contagem, teve como variável resposta o número total de acidentes, num período de oito anos (2014 a 2021).

Os resultados obtidos com o modelo ajustado indicam que a taxa de acidentes na autoestrada A4 é influenciada por fatores como o atrito mínimo, a profundidade média em perfil, a extensão da via em curva côncava e a presença de vias de aceleração e desaceleração,

Palavra-Chave: Modelos Lineares Generalizados Mistos, Dados de Contagem, Acidentes Rodoviários, Autoestrada A4

Abstract

Road accidents are considered a major public health problem worldwide. According to the World Health Organization report, road accidents are the leading cause of death for young people worldwide, particularly in young men.

The modeling of the number of road accidents aims to analyze and understand the factors that contribute to the occurrence of traffic accidents. The research on these factors that influence the occurrence of road accidents aims to reduce their number, as well as to identify effective prevention measures.

This work aims to apply the Mixed Generalized Linear Models - GLMM, to study the effect of different variables that define the pavement state, the geometry of the road, traffic volume and among other variables in road accidents. The data used are related to the A4 highway segments of the Greater Porto Region, Portugal. The model developed, Generalized Linear Mixed Model for response in the form of counting, had as variable response the total number of accidents in a period of eight years (2014 to 2021).

The results obtained with the adjusted model indicate that the rate of accidents on the A4 highway is influenced by factors such as minimum friction, average depth in profile, the extension of the track in concave curve and the presence of acceleration and deceleration roads.

Keyword: Mixed Generalized Linear Models, Count Data, Road Accidents, A4 Highway

Conteúdo

Acrónimos

AIC *Critério de Informação Akaike.*

ANSR *Autoridade Nacional de Segurança Rodoviária.*

AR *Alinhamento Reto.*

BIC *Critério de Informação Bayesiano.*

CC *Curva Circular.*

CCV *Curva Côncava.*

CCX *Curva Convexa.*

CL *Clotóide.*

GLM *Modelos Lineares Generalizados.*

GLMM *Modelos Lineares Generalizados Mistos.*

LMM *Modelos Lineares Mistos.*

ML *Machine Learning.*

MQL *Quasi-Verosimilhança Marginal.*

OMS *Organização Mundial da Saúde.*

PQL *Quasi-Verosimilhança Penalizada.*

T *Trainel.*

UE *União Europeia.*

VAC *Via Acelaração.*

VACD *Via de Acelaração Direita.*

VACE *Via de Acelaração Esquerda.*

VC *Via Central.*

VD *Via Direita.*

VDC *Via Desaceleração.*

VDCD *Via Desaceleração Direita.*

VDCE *Via Desaceleração Esquerda.*

VE *Via Esquerda.*

Lista de Figuras

Lista de Tabelas

1 Introdução

1.1 Enquadramento e motivação

Anualmente, os acidentes de viação acarretam custos significativos em termos humanos, económicos e sociais. A avaliação precisa destes custos é fundamental para informar o debate sobre as políticas públicas de segurança rodoviária (Silva et al., 2021).

O relatório global da Organização Mundial da Saúde (OMS) sobre a segurança na estrada de 2018, indica que as mortes por acidente de viação continuam a aumentar, situando-se nos 1,35 milhões de vítimas anuais, e são a principal causa de mortalidade de crianças e jovens com idades compreendidas entre 5 e 29 anos.

Segundo os dados do Eurostat, entre 2009 e 2019, o número de mortes nas estradas na *União Europeia* (UE) diminuiu cerca de 31%. Em 2009, foram registradas 33 mil mortes, enquanto que em 2019 foram registadas 22756, ou seja, uma diferença superior a 10 mil vidas perdidas. Em 2018, 12% das pessoas que morreram nas estradas europeias tinham entre 18 e 24 anos de idade. Três quartos do total de vítimas que faleceram nas estradas europeias eram do sexo masculino, uma tendência que permaneceu praticamente inalterada desde 2010 e que é consistente em todos os países da UE.

Em Portugal, em 2022, a *Autoridade Nacional de Segurança Rodoviária* (ANSR), responsável pela gestão da segurança rodoviária, contabilizou 34275 acidentes de viação com vítimas, 473 vítimas mortais, 2436 feridos graves e 40123 feridos leves. Os anos de 2020 e 2021 registraram quebras significativas da circulação rodoviária em comparação com 2019, e consequentemente, reduções nos principais indicadores de sinistralidade face àquele ano. Por conseguinte a Comissão Europeia considerou 2019 como o ano base de referência para avaliar a evolução da sinistralidade rodoviária durante a presente década.

Assim, e face a 2019, em 2022 registraram-se reduções em todos os principais indicadores: menos 2976 acidentes (-8,0%), menos 47 vítimas mortais (-9,0%)(ver a Tabela 1.1).

Tabela 1.1: Sinistralidade em Portugal em 2019 e 2022

Jan - Dez	Acidentes com vítimas			Vítimas mortais		
	2019	2022	Δ(%) 22/19	2019	2022	Δ(%) 22/19
Continente	35704	32788	-8,2%	474	462	-2,5%
Açores	611	612	0,2%	7	6	-14,3%
Madeira	936	875	-6,5%	39	5	-87,2%
Total	37251	34275	-8,0%	520	473	-9,0%

1.2 Breve revisão da literatura

Devido aos enormes custos dos acidentes de viação para a sociedade, o conhecimento dos fatores que contribuem para a ocorrência dum acidente, é fundamental para as autoridades rodoviárias identificar medidas eficazes de prevenção.

A nível mundial, a segurança rodoviária é afetada por uma série de fatores, que podem ser divididos essencialmente em três categorias:

- **Comportamento do condutor**: A forma como os condutores se comportam nas estradas é crucial. O respeito às regras de trânsito, evitar distrações ao volante, não dirigir sob a efeito de álcool ou drogas, e manter uma velocidade adequada são todos aspetos importantes.
- **Infraestrutura Viária**: A qualidade das estradas, sinalização adequada, iluminação e condições gerais das vias desempenham um papel significativo na segurança rodoviária.
- **Veículos**: A condição dos veículos, incluindo a manutenção regular, influencia diretamente a segurança.

Esses fatores estão interligados, e abordagens integradas que contemplam todos esses aspetos são essenciais para melhorar a segurança rodoviária.

Na maioria de estudos existentes sobre os fatores que influenciam a segurança rodoviária (Rolison et al., 2018), o comportamento do condutor figura como responsável por mais de 90% dos acidentes de viação. As principais falhas humanas identificadas nos condutores jovens,

prende-se com a inexperiência e comportamentos de risco, incluindo a velocidade excessiva e o consumo de drogas e álcool. Por outro lado, à medida que a idade avança, o aumento da prevalência de deficiências visuais e cognitivas, bem como o uso de medicamentos, têm sido associados às colisões de condutores mais velhos.

Modelos de previsão de acidentes são essenciais para identificar padrões e avaliar fatores de risco e podem ser utilizados para estimar o número esperado de acidentes num local e, por sua vez, identificar corretamente um local como perigoso.

Para modelar a frequência esperada de acidentes, bem como para compreender os fatores que afetam o risco de um acidente, os investigadores usam vários modelos estatísticos.

De acordo com Mannering & Bhat (2014) os modelos estatísticos mais utilizados para prever a frequência esperada de acidentes são: o modelo de Regressão de Poisson, o modelo de Regressão Binomial Negativa, o modelo Regressão Inflacionada de Zero, o modelo Multinomial Negativo, modelos de efeitos aleatórios, modelos de correlação espacial e temporal, Equações de Estimação Generalizadas, o Modelo hierárquico e modelos de *Machine Learning*.

Os modelos de contagem como o modelo de regressão de Poisson e de Binomial Negativa (Lord & Mannering, 2010), bem como as suas diversas extensões, são considerados modelos estatísticos mais adequados para modelar dados de contagem de acidentes rodoviários que são geralmente caracterizados por baixos valores médios, dispersão elevada e heterocedasticidade.

Em muitas situações, para se obter um número suficiente de observações, os dados de acidentes de viação são frequentemente agregados ao longo do tempo (por exemplo, acidentes por mês) e/ou espaço (acidentes ao longo de um segmento de estrada). Isso pode gerar problemas adicionais de heterogeneidade não observada que podem depender do tempo ou do espaço. Estudos mais recentes têm controlado a heterogeneidade não observada, como correlação espacial e temporal, aplicando modelos de efeitos aleatórios. (Mannering et al., 2016)

Elvik & Katharina Høye (2023) estudaram o efeito da variável tempo na relação entre os acidentes rodoviários e os fatores que os influenciam nas estradas da Noruega, comparando quatro modelos de previsão de acidentes recorrendo aos modelos de regressão Binomial Negativa. O primeiro modelo foi desenvolvido em 2002, com base em dados de estradas norueguesas para os anos de 1993 a 2000. Este modelo foi atualizado três vezes, sendo a primeira atualização feita com dados de 2000 a 2005. Em 2014, um novo modelo foi desenvolvido para incluir as estradas

municipais, utilizando dados de 2006 a 2011. O quarto e mais recente modelo baseou-se em dados de 2010 a 2015.

Os resultados evidenciam a estabilidade das estimativas dos coeficientes nos quatro modelos ao longo do tempo. Ficou ainda demonstrado que o volume de tráfego é a variável mais influente no número de acidentes, ou seja, o aumento do volume de tráfego está diretamente associado ao aumento no número de acidentes.

Modelos multivariados de Poisson e Poisson Lognormal são propostos para modelar simultaneamente diferentes tipos de acidentes (por exemplo, modelar o nível de gravidade e a frequência). Esta abordagem permite controlar a heterogeneidade não observada resultante das correlações entre esses fatores (Imprialou et al., 2016).

Recentemente tem havido, um interesse crescente em modelos de *Machine Learning* (ML) para a previsão de acidentes. Os modelos de ML são implementados para resolver algumas limitações associadas aos modelos estatísticos tradicionais. Enquanto que os modelos tradicionais podem ser limitadas pelo fato de exigirem algumas restrições sobre a distribuição dos dados e assumirem uma relação linear entre a variável resposta e as variáveis explicativas, os modelos de ML são mais flexiveis em relação a distribuição dos dados. Além disso a revisão da literatura sugere que os modelos ML evitam os processos de tentativa e erro na especificação do modelo. Por último, os modelos de ML parecem mais adequados para aplicações que envolvem grandes dimensões de dados ou dados em tempo real, pois são capazes de lidar com conjuntos de dados muito grandes. Para mais detalhe ver (Dragomanovits et al., 2022).

Neste estudo são apresentados os Modelos Lineares Generalizados Mistos (GLMM) para prever a frequência de acidentes. Estes modelos que são uma extensão dos *Modelos Lineares Generalizados* (GLM) são geralmente utilizados quando há necessidade de modelar efeitos aleatórios, ou seja, a variabilidade não explicada por variáveis explicativas fixas. São particularmente relevantes na presença de heterogeneidade nos dados e usados nas análises longitudinais, em que as mesmas unidades são medidas repetidamente ao longo do tempo (os GLMMs podem levar em consideração a correlação entre as medidas repetidas de uma mesma unidade).

1.3 Objetivos

Neste trabalho pretende-se recorrer a Modelos Lineares Generalizados Mistos para identificar a influência de vários fatores (humanos, infraestruturais, ambientais e circunstanciais) na ocorrência de acidentes rodoviários, com base num conjunto de dados reais de acidentes rodoviários ocorridos na autoestrada A4 da região de Grande Porto. Pretende-se assim, desenvolver modelos que permitam explorar e modelar adequadamente a variabilidade dos acidentes rodoviários nos diferentes segmentos de estrada e ajustar um modelo adequado à realidade do problema.

1.4 Estrutura

Este trabalho está estruturado em 4 capítulos. O primeiro capítulo introduz o tema central do estudo, que é modelação de acidentes de viação, delineando os objetivos propostos e a organização do trabalho.

No Capítulo 2 encontra-se um breve resumo do Modelo Linear Misto (LMM), introduzindo simultaneamente os Modelos Lineares Generalizados Mistos (GLMM). Neste capítulo são abordados aspectos como a interpretação do modelo, assim como a estimação e inferência associados.

O Capítulo 3 é dedicado ao estudo do caso em questão. Inicia-se com uma descrição detalhada da base de dados, seguida de uma análise exploratória dos dados. Posteriormente, são apresentados os resultados obtidos da aplicação dos modelos descritos no Capítulo 2, bem como a respetiva interpretação.

No Capítulo 4 são discutidos os resultados obtidos no Capítulo 3 destacando-se as limitações inerentes. Além disso, são apontadas considerações sobre como essas limitações podem ser aprimoradas em trabalhos futuros.

2 Modelos Lineares

Em muitos estudos, o objetivo principal é estudar a relação entre várias variáveis, isto é, analisar a influência que uma ou mais variáveis (variáveis explicativas) têm sobre uma variável de interesse (variável resposta).

O modelo linear, criado no inicio do século XIX por Legendre e Gauss dominou a modelação estatística até meados do século XX.

Devido ao grande número de modelos que englobam e à facilidade de análise associada ao rápido desenvolvimento computacional que se tem verificado nas ultimas décadas, os modelos lineares generalizados têm vindo a desempenhar um papel cada vez mais importante na análise estatística.

Os Modelos Lineares Generalizados (GLM) propostos por Nelder & Wedderburn (1972) unificam os modelos de regressão para diferentes tipos de respostas, como por exemplo, modelos lineares para respostas contínuas, modelos logísticos para respostas binárias e modelos log-lineares para dados de contagem. Uma das limitações dos Modelos Lineares Generalizados é a exigência de que os erros sejam independentes, isso significa que, estes modelos não são capazes de modelar dados com estruturas de dependência entre as observações. Nestas situações pode-se usar Modelos Lineares Generalizados Mistos.

Modelos Lineares Generalizados Mistos (GLMM) propostos por Breslow & Clayton (1993), são muito aplicados na modelação de dados com estruturas longitudinais. São uma combinação natural de *Modelos Lineares Mistos* (LMM) e Modelos Lineares Generalizados (GLM).

Modelos Lineares Mistos propostos por Harville (1977) e Laird & Ware (1982) são modelos de regressão linear que incluem efeitos aleatórios e efeitos fixos. Estes efeitos aleatórios permitem descrever, ao londo do tempo, as alterações dentro de cada indivíduo e flexibilizar a representação da estrutura de variância-covariância induzida pelo agrupamento de dados. (Cabral & Gonçalves, 2011)

2.1 Modelos Lineares Mistos

Inicialmente introduz-se os Modelos Lineares Mistos que incorporam quer efeitos fixos, isto é, parâmetros associados a toda a população, quer efeitos aleatórios isto é, parâmetros associados aos indivíduos selecionados aleatoriamente da população.

Um Modelo Linear Misto é dado por,

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i,\ i = 1, ..., n \tag{2.1}$$

em que

- $\mathbf{Y}_i$ representa um vetor $(m_i \times 1)$ das variáveis resposta da $i-$ ésima unidade experimental;
- $\boldsymbol{\beta}$ é o vetor $(p \times 1)$ de efeitos fixos (parâmetros desconhecidos);
- $\mathbf{X}_i$ é uma matriz $(m_i \times p)$ de covariáveis de efeitos fixos;
- $\mathbf{b}_i$ é o vetor $(q \times 1)$ dos efeitos aleatórios;
- $\mathbf{Z}_i$ é uma matriz $(m_i \times q)$ de covariáveis de efeitos aleatórios;
- $\boldsymbol{\epsilon}_i$ é o vetor $(m_i \times 1)$ dos erros aleatórios.

Pode-se escrever a variável resposta para a i-ésima unidade experimental no j-ésimo instante como:

$$\mathbf{Y}_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{b}_i + \boldsymbol{\epsilon}_{ij},\ i = 1, ..., n,\ j = 1, ..., m_i \tag{2.2}$$

As condições subjacentes ao modelo são $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{G})$ e $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \mathbf{R}_i)$ e $\mathbf{b}_i$ e $\boldsymbol{\epsilon}_i$ são independentes para diferentes unidades experimentais i e entre si, em que, $\mathbf{G}_{(q\times q)}$ é a matriz de variância-covariância dos efeitos aleatórios presente no vector $\mathbf{b}_i$ e $\mathbf{R}_{i(m_i\times m_i)}$ a matriz de variância-covariância dos erros aleatórios.

O vetor da variável resposta associado à i - ésima unidade experimental segue uma distribuição Normal Multivariada com vetor de médias e matriz de variância-covariância dados, respectivamente, por

$$E[\mathbf{Y}_i] = \mathbf{X}_i\beta \tag{2.3}$$

$$V = V[\mathbf{Y}_i] = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^T + \mathbf{R}_i \tag{2.4}$$

isto é,

$$\mathbf{Y}_i \sim \mathbf{N}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^T + \mathbf{R}_i) \tag{2.5}$$

O primeiro termo da decomposição de $V[Y_i]$ em (2.4) modela a dispersão dos perfis individuais da resposta e o segundo termo está relacionado com a dispersão da resposta em torno dos perfis individuais.

O modelo (2.1) também pode ser visto como um modelo linear em dois estágios, de modo que, no primeiro estágio considera-se fixos os efeitos aleatórios $\mathbf{b}_i$, de forma que, a distribuição de $\mathbf{Y}_i$ condicional ao efeito aleatório $\mathbf{b}_i$ é Gaussiana multivariada com valor médio $\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\mathbf{b}_i$ e matriz de variância-covariância $\mathbf{R}_i$ e escreve-se

$$\mathbf{Y}_i|\mathbf{b}_i \sim \mathbf{N}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \mathbf{R}_i) \tag{2.6}$$

No segundo estágio supomos que os vetores $\mathbf{b}_i$ são independentes com distribuição $N(0, \mathbf{G})$ e consequentemente, a função distribuição probabilidade marginal de $\mathbf{Y}_i$ é

$$Y_i \sim \mathbf{N}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^T + \mathbf{R}_i) \tag{2.7}$$

2.2 Modelos Lineares Generalizados Mistos

Os Modelos Lineares Generalizados Mistos (GLMM) são uma extensão dos Modelos Lineares Generalizados (GLM), que incluem efeitos aleatórios no preditor linear além dos efeitos fixos. A introdução de efeitos aleatórios permite modelar a estrutura de correlação entre observações que pertencem à mesma unidade experimental.

Os GLMM têm como objectivo descrever as alterações da resposta média de cada unidade experimental e a relação destas com as covariáveis, ou seja, têm como finalidade realizar inferências para cada unidade experimental e não para a população (Cabral & Gonçalves, 2011).

2.2.1 Especificação do GLMM

Considere que, a distribuição condicional da variável resposta Y_{ij}, em que $i = 1, ..., n$ e $j = 1, ..., m_i$, dado os efeitos aleatórios $\mathbf{b}_i$, pertence a família exponencial, isto é,

$$Y_{ij}|\mathbf{b}_i \sim FE(Y_{ij}|\mathbf{b}_i, \theta) \tag{2.8}$$

Assim, a densidade y_{ij} condicionada aos efeitos aleatórios $\mathbf{b}_i$ é dada

$$f_i(y_{ij}|\mathbf{b}_i, \boldsymbol{\beta}, \phi) = exp\left[\frac{y_{ij}\theta_{ij} - a(\theta_{ij})}{\phi} + c(y_{ij}, \phi)\right] \tag{2.9}$$

em que $a(.)$ e $c(.)$ são funções conhecidas, e ϕ é um parâmetro de dispersão que pode ou não ser conhecido. Tem-se ainda que, o parâmetro canónico, θ_{ij}, está associado à média condicional $\mu_{ij} = E(y_{ij}|\mathbf{b}_i)$, que, por sua vez, está associada a um preditor linear

$$\eta_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i, \tag{2.10}$$

onde $\mathbf{x}_{ij}^T$ e $\mathbf{z}_{ij}^T$ são vetores conhecidos e $\boldsymbol{\beta}$ é um vetor de parâmetros desconhecidos (os efeitos fixos), através de uma função de ligação $g(.)$ tal que

$$g(\mu_{ij}) = \eta_{ij} \tag{2.11}$$

Além disso, suponha que $\mathbf{b}_i \sim \mathbf{N}(\mathbf{0}, \mathbf{G})$, onde a matriz de variância covariância $\mathbf{G}$ possa depender de um vetor θ de componentes de variância desconhecidos.

Note-se que, uma vez que $Y_{ij}|\mathbf{b}_i$ pertence à família exponencial, tem-se

$$E[y_{ij}|\mathbf{b}_i] = \mu_{ij} = a'(\theta_{ij}) \tag{2.12}$$

e

$$V[y_{ij}|\mathbf{b}_i] = \phi a''(\theta_{ij}) = \phi V(\mu_{ij}) \tag{2.13}$$

Em particular, sob a função ligação canónica, tem-se

$$\theta_{ij} = \eta_{ij} \tag{2.14}$$

isto é, $g(.) = h^{-1}(.)$, onde $h(.) = a'(.)$

De seguida, apresentam-se alguns dos casos mais aplicados de GLMM :

- Modelo Linear Generalizado Misto para Resposta Contínua:

  O Modelo linear misto normal, onde $\mathbf{R} = \sigma^2 I$, é um caso especial do GLMM, no qual a família exponencial é Gaussiana com valor médio μ_i e variância σ^2, e a função de ligação é $g(\mu_i) = \mu_i$. O parâmetro de dispersão $\phi = \sigma^2$ é desconhecido.

- Modelo Linear Generalizado Misto para Resposta Binária:

  Dado os efeitos aleatórios $\mathbf{b}_i$, a distribruição condicional $Y_{ij}|\mathbf{b}_i$ tem distribuição com

  $$var(\mathbf{Y}_{ij}|\mathbf{b}_i) = E\left(\mathbf{Y}_{ij}|\mathbf{b}_i\right)\left(1 - E\left(\mathbf{Y}_{ij}|\mathbf{b}_i\right)\right) \tag{2.15}$$

  e função ligação

  $$logit(\mu_{ij}) \equiv log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \eta_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i \tag{2.16}$$

- Modelo Linear Generalizado Misto para Resposta sob a forma de Contagem:

  Dado os efeitos aleatórios $\mathbf{b}_i$, a distribuição condicional $\mathbf{Y}_{ij}|\mathbf{b}_i$ tem a distribuição de Poisson

  $$var(\mathbf{Y}_{ij}|\mathbf{b}_i) = E\left(\mathbf{Y}_{ij}|\mathbf{b}_i\right) \tag{2.17}$$

  e a função ligação

  $$log\left(\mu_{ij}\right) = \eta_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i \tag{2.18}$$

2.2.2 Interpretação dos Parâmetros do Modelo

A introdução de efeitos aleatórios quer no modelo linear quer no modelo linear generalizado permite modelar a estrutura de correlação entre observações que pertencem à mesma unidade experimental. No entanto, enquanto no modelo linear os efeitos aleatórios não afetam a interpretação dos parâmetros fixos, o mesmo não acontece no modelo linear generalizado.

Assim no GLMM têm-se as seguintes interpretações para os parâmetros fixos e para a componente aleatória do modelo.

- **Parte fixa** Os parâmetros $\boldsymbol{\beta}$ não tem interpretação populacional, têm uma interpretação em termos específicos do indivíduo, ou seja, representam o efeito das covariáveis na resposta média de um indivíduo específico. A sua interpretação depende dos efeitos aleatórios do i-ésimo indivíduo assumirem um valor fixo (não observado).

- **Parte aleatória** Relativamente à componente aleatória uma maneira de interpretar as estimativas das variâncias dos efeitos aleatórios é produzir percentis dos efeitos baseados na hipótese Gaussiana. Por exemplo, para um modelo com efeito aleatório na interseção, $\widehat{\beta}_0 - 1.96\sqrt{\widehat{d_{11}}}$ e $\widehat{\beta}_0 + 1.96\sqrt{\widehat{d_{11}}}$ dão-nos os percentis aproximados 2.5 e 97.5, respetivamente. Por outro lado, com base nos percentis assim obtidos e aplicando a inversa da função de ligação, obtém-se os limites de variação dos valores esperados condicionais μ_{ij}. Uma outra maneira é a representação gráfica das trajetórias (perfis) condicionadas.

2.2.3 Estimação Máxima Verosimilhança

A estimação dos parâmetros do modelo é baseada no método de máxima verosimilhança, ou seja, maximizando a função verosimilhança marginal, através da integração da distribuição condicional de $y_{ij}|\mathbf{b}_i$.

No GLMM a contribuição do i-ésimo indivíduo para a função verosimilhança marginal é

dado por

$$L_i(\boldsymbol{\beta}, \phi, \mathbf{G}) = \int \prod_{j=1}^{m_i} f_{ij}\left(y_{ij}|\mathbf{b}_i, \boldsymbol{\beta}, \phi\right) f(\mathbf{b}_i|G) d\mathbf{b}_i \tag{2.19}$$

onde a integração é sobre a distribuição q-dimensional de $\mathbf{b}_i$. A função de verosimilhança conjunta sobre n indivíduos é dada por

$$\begin{aligned} L\left(\boldsymbol{\beta}, \phi, \mathbf{G}\right) &= \prod_{i=1}^{n} L_i(\boldsymbol{\beta}, \phi, \mathbf{G}) \\ &= \prod_{i=1}^{n} \int \prod_{j=1}^{m_i} f_{ij}\left(y_{ij}|\mathbf{b}_i, \boldsymbol{\beta}, \phi\right) f(\mathbf{b}_i|G) d\mathbf{b}_i \end{aligned} \tag{2.20}$$

A função log-verosimilhança (Liu, 2016) pode ser obtida aplicando logaritmo a ambos os membros da equação (2.20)

$$\begin{aligned} logL\left(\boldsymbol{\beta}, \mathbf{G}, \phi\right) &= log\left[\prod_{i=1}^{n} \int \prod_{j=1}^{m_i} f_{ij}\left(y_{ij}|\mathbf{b}_i, \boldsymbol{\beta}, \phi\right) f(\mathbf{b}_i|G) d\mathbf{b}_i\right] \\ &= \sum_{i,j}\left[\frac{y_{ij}\theta_{ij} - a(\theta_{ij})}{\phi} + c(y_{ij}, \phi)\right] + \sum_{i} log \int_{-\infty}^{+\infty} f(\mathbf{b}_i|G) d\mathbf{b} \end{aligned} \tag{2.21}$$

A obtenção dos estimadores de máxima verosimilhança de $\boldsymbol{\beta}$ e $\mathbf{G}$ envolve a maximização de (2.21) que exige o cálculo de n integrais que, em geral, não tem uma solução analitica, sendo necessário recorrer a métodos numéricos de integração, como Método de Laplace, Método de Quadratura de Gauss-Hermite ou, ainda, Integração Monte Carlo. O cálculo númerico possui também várias limitações, uma vez que envolve integrais complexos cuja dimensão depende da estrutura dos efeitos aleatórios. A seguir, apresenta-se as duas metodologias mais utilizadas, nomeadamente, Integração Numérica (Método de Laplace e Método da Quadratura de Gauss-Hermite) e Quase-Verosimilhança Penalizada (PQL).

a) Método de Aproximação da Integral

Conforme Singer et al. (2018), um método geralmente utilizado para a aproximação de integrais de funções é aquele baseado no método númerico da **quadratura de Gauss-Hermite**. Nesse contexto, integrais de funções do tipo $g(x)\exp\left(-x^2\right)$ podem ser aproximadas por uma soma pesada

$$\int g(x)\exp\left(-x^2\right)dx \approx \sum_{k=1}^{K} g(x_k)w_k \tag{2.22}$$

em que x_k e w_k, $k = 1, ..., K$ são respectivamente os m nós e pesos da quadratura de Gauss-Hermite. Quanto maior for K mais precisa é a aproximação. Em particular, a função de verosimilhança marginal corresponde à i-ésima unidade amostral em (2.20) é aproximada como

$$L_i(\beta, \mathbf{G}(\theta), \phi|y) \approx \sum_{k_i=1}^{K} ... \sum_{k_q=1}^{K} f(\mathbf{y}_i|\beta, [\mathbf{G}(\theta)]^{1/2}\mathbf{x}, \phi) w_{k_1} ... w_{k_q} \tag{2.23}$$

em que $\mathbf{x} = (x_{k_1}, ..., x_{k_q})$ e $w_{k_1}, ..., w_{k_q}$ são, respectivamente, nós e pesos apropriados.

Alternativamente (2.20) pode ser aproximada pelo método de Monte Carlo. Neste contexto, a i-ésima componente dessa expressão, nomeadamente

$$L_i(\beta, \mathbf{G}(\theta), \phi|\mathbf{y}_i) = K \prod_{i=1}^{n} \int f(\mathbf{y}_i|\beta, \mathbf{b}_i) f[\mathbf{b}_i|0, \mathbf{G}(\theta)] d\mathbf{b}_i = E_{f[0,G(\theta)]}[f(\mathbf{y}_i|\beta, \mathbf{b}_i)] \tag{2.24}$$

é aproximada por

$$L_i(\beta, \mathbf{G}(\theta), \phi|\mathbf{y}_i) \approx \frac{1}{K} \sum_{k=1}^{K} f(\mathbf{y}_i|\beta, \mathbf{b}_i^{(k)}) \tag{2.25}$$

em que $\mathbf{b}_i^{(1)}, ..., \mathbf{b}_i^{(k)}$ é uma amostra aleatória gerada de uma distribuição $f[0, G(\widehat{\theta})]$.

b) Aproximação do integrando

Conforme Molenberghs & Verbeke (2005), o objectivo da aproximação do integrando é obter um integral para o qual existam soluções que facilitam a maximização da função de verosimilhança. Neste contexto, uma alternativa é o **Método de Laplace**, apropriado para obter aproximações de integrais da forma

$$I = \int e^{Q(\mathbf{b})} d\mathbf{b} \tag{2.26}$$

onde $Q(\mathbf{b})$ é uma função unimodal e limitada de uma variável q-dimensional, $\mathbf{b}$. O método é baseado na seguinte expansão de Taylor de segunda ordem de $Q(\mathbf{b})$ em torno do ponto $\widehat{\mathbf{b}}$ que maximiza essa função

$$Q(\mathbf{b}) \approx Q(\widehat{\mathbf{b}}) + \frac{1}{2}(\mathbf{b} - \widehat{\mathbf{b}})^T Q''(\widehat{\mathbf{b}})(\mathbf{b} - \widehat{\mathbf{b}}) \tag{2.27}$$

em que $Q''(\widehat{\mathbf{b}})$ corresponde à matriz Hessiana de Q, em relação a $\widehat{\mathbf{b}}$. Substituindo $Q(\mathbf{b})$ em (2.26) por sua aproximação em (2.27), obtemos

$$I = \int e^{Q(\mathbf{b})} d\mathbf{b} \approx e^{Q(\widehat{\mathbf{b}})} \int -\frac{1}{2} e^{(\mathbf{b}-\widehat{\mathbf{b}})^T [-Q''(\widehat{\mathbf{b}})](\mathbf{b}-\widehat{\mathbf{b}})} d\mathbf{b} \tag{2.28}$$

simplificando

$$I \approx (2\pi)^{q/2} | - Q''(\widehat{\mathbf{b}})|^{-1/2} e^{Q(\mathbf{b})} \tag{2.29}$$

considerando que o integrando corresponde ao núcleo de uma distribuição Normal q-variada com vetor de médias nulo e matriz de variância - covariância $| - Q''(\widehat{\mathbf{b}})|^{-1}$.

O integral da contribuição do i-ésimo indivíduo para a função da verosimilhança (2.20) é proporcional ao integral (2.26), para funções $Q(\mathbf{b})$ com a forma

$$Q(\mathbf{b}) = \phi^{-1} \sum_{j=1}^{n_i} [y_{ij}(\mathbf{x}_{ij}^T \beta + \mathbf{z}_{ij}^T \mathbf{b}_i) - c(\mathbf{x}_{ij}^T \beta + \mathbf{z}_{ij}^T \mathbf{b}_i)] - \frac{1}{2} \mathbf{b}^T G^{-1} \mathbf{b} \tag{2.30}$$

onde o método de Laplace pode ser aplicado. O facto de $Q(\mathbf{b})$ depender de $\boldsymbol{\beta}$, ϕ e G, todos parâmetros desconhecidos, faz com que em cada iteração da maximização da função de

verosimilhança, $\widehat{\mathbf{b}}$ seja recalculado de forma condicional às estimativas destes parâmetros. A aproximação de Laplace é exacta quando $Q(\mathbf{b})$ é uma função quadrática de $\mathbf{b}$.

c) Quasi-Verosimilhança Penalizada (PQL)

Ainda segundo Singer et al. (2018), outra alternativa para a obtenção dos estimadores de máxima verosimilhança de $\boldsymbol{\beta}$ e $G(\theta)$ são os chamados de **quasi-verosimilhança**, a saber, *Quasi-Verosimilhança Penalizada* (PQL) e *Quasi-Verosimilhança Marginal* (MQL). Os dois métodos PQL e MQL em muitos casos produzem resultados semelhantes, sendo o estimador MQL, em geral, mais enviesado do que o estimador PQL correspondente. Um ponto de partida para esses métodos é a decomposição dos dados como

$$y_{ij} = \mu_{ij} + \epsilon_{ij} = h(\eta_{ij}) + \epsilon_{ij} \tag{2.31}$$

em que h é a inversa da função de ligação g e ϵ_{ij} é um erro aleatório.

A fundamentação do método PQL é baseada na expansão de Taylor de primeira ordem de (2.31) em torno da estimativa atual $\widehat{\beta}$ de β e do preditor atual $\widehat{\mathbf{b}_i}$ do efeito aleatório $\mathbf{b}_i$, nomeadamente

$$\widehat{\eta_{ij}} = \mathbf{x}_{ij}^T\widehat{\boldsymbol{\beta}} + \mathbf{z}_{ij}^T\widehat{\mathbf{b}_i} \tag{2.32}$$

$$\begin{aligned} Y_{ij} &\approx h(\widehat{\eta_{ij}}) + h^{'}(\widehat{\eta_{ij}})x_{ij}^T(\beta - \widehat{\beta}) + h^{'}(\widehat{\eta_{ij}})z_{ij}^T(\mathbf{b_i} - \widehat{\mathbf{b_i}}) + \epsilon_{ij} \\ &= \mu_{ij}^* + v(\mu_{ij}^*)x_{ij}^T(\beta - \widehat{\beta}) + v(\mu_{ij}^*)z_{ij}^T(\mathbf{b_i} - \widehat{\mathbf{b_i}}) + \epsilon_{ij} \end{aligned} \tag{2.33}$$

em que $\mu_{ij}^* = h(\widehat{\eta_{ij}})$ corresponde à média de y_{ij} calculada em termos de $\widehat{\beta}$ e $\widehat{b_i}$ e $v(\mu_{ij}^*) = h^{'}(\widehat{\eta_{ij}})$ corresponde à variância aproximada de ϵ_{ij}.

Organizemos as observações (y_{ij}) e os erros (ϵ_{ij}) em vetores $\mathbf{y}$ e ϵ, respectivamente, e os vetores x_{ij} e z_{ij} em matrizes $\mathbf{X}$ e $\mathbf{Z}$. Em seguida, consideremos o vetor μ^* cujos elementos são os termos de $\mu_{\mathbf{ij}}^*$ e a matriz diagonal $\mathbf{V}^*$, cujos elementos não nulos são os termos $v(\mu_{ij}^*)$ com organização similar. Então podemos escrever

$$\mathbf{y} \approx \mu^* + \mathbf{V}^*\mathbf{X}(\beta - \widehat{\beta}) + \mathbf{V}^*\mathbf{Z}(\mathbf{b} - \widehat{\mathbf{b}}) + \epsilon \tag{2.34}$$

subtraindo μ^* de ambos os membros e multiplicando-os à esquerda por $[V^*]^{-1}$, obtemos

$$[\mathbf{V}^*]^{-\mathbf{1}}(\mathbf{y} - \mu^*) \approx \mathbf{X}(\beta - \widehat{\beta}) + \mathbf{Z}(\mathbf{b} - \widehat{\mathbf{b}}) + [\mathbf{V}^*]^{-\mathbf{1}}\epsilon \tag{2.35}$$

e, consequentemente,

$$\begin{aligned} \mathbf{y}^* &= [V^*]^{-1}(y - \mu^*) + X\widehat{\beta} + Z\widehat{b} + [V^*]^{-1}\epsilon \\ &\approx X\beta + Zb + \epsilon^* \end{aligned} \tag{2.36}$$

com $\epsilon^* = [\mathbf{V}^*]^{-\mathbf{1}}\epsilon$.

A expressão (2.36) tem a forma de (2.1) e pode ser encarada como um modelo linear misto tendo as "pseudo-variáveis"$\mathbf{y}^*$ como respostas. Notemos que $V(\epsilon) = \mathbf{V}^*$, o que implica $V(\epsilon^*) = [\mathbf{V}^*]^{-\mathbf{1}}\mathbf{V}^*[\mathbf{V}^*]^{-\mathbf{1}}$ e $\mathbf{V}(\mathbf{Y}^*) = \mathbf{ZGZ^T} + [\mathbf{V}^*]^{-\mathbf{1}}$.

Esse contexto sugere o seguinte algoritmo para ajustar o modelo marginal:

1. Inicie o processo com valores para $\boldsymbol{\beta}$ e $\mathbf{G}$;
2. Obtenha os preditores de $\mathbf{b_i}$ com base nos valores correntes de $\boldsymbol{\beta}$ e $\mathbf{G}$ e calcule os valores das pseudo-variáveis Y^*;
3. Reajuste o modelo linear misto (2.36) obtendo novos valores para $\boldsymbol{\beta}$ e $\mathbf{G}$;
4. Repita os passos 2 e 3 até que algum critério de convergência seja atingido.

Os estimadores resultantes desse processo são conhecidos como estimadores de quasi-verosimilhança penalizados pois dependem apenas dos dois primeiros momentos condicionais penalizados relativamente aos efeitos aleatórios.

2.2.4 Inferência em GLMM

Como o ajuste do GLMM é baseado no método de máxima verosimilhança, toda a inferência se baseia na função verosimilhança. Com efeito, assumindo que o modelo ajustado é apropriado, os estimadores obtidos seguem assintoticamente a distribuição Normal com o valor médio igual ao parâmetro a estimar e a variância igual à inversa da matriz de informação de Fisher. Assim, testes de Wald, comparando estimativas padronizadas com a distribuição Normal Padrão, podem ser facilmente realizados. As hipóteses compostas podem ser igualmente testadas usando a formulação mais geral da estatística de Wald, que neste caso é comparada com a distribuição Qui-quadrado.

Os testes de Razão de Verosimilhança, são em geral, os utilizados para comparar a estrutura fixa de dois modelos encaixados que tenham os mesmos efeitos aleatórios. Assim para testar a nulidade de um subvetor de r componentes de $\boldsymbol{\beta}$

$$H_0 : \beta_r = 0 \quad vs \quad H_1 : \beta_r \neq 0 \tag{2.37}$$

utiliza-se a estatística de teste

$$2\left(logL_1 - logL_0\right) \sim \chi^2_{k_1-k_0} \ , \tag{2.38}$$

em que L_1 corresponde à verosimilhança do modelo mais geral, com k_1 parâmetros e L_0 corresponde a verosimilhança do modelo encaixado, com k_0 parâmetros. Sob a hipótese nula de que o modelo restrito é mais adequado (ou seja de que os $r = k_1 - k_0$ parâmetros adicionais são iguais a zero) a estatística de teste tem distribuição assintótica Qui-quadrado com $k_1 - k_0$ graus de liberdade.

Quando os modelos não estão encaixados usa-se o *Critério de Informação Akaike* (AIC) dado por

$$AIC = -2logL + 2k \ , \tag{2.39}$$

ou *Critério de Informação Bayesiano* (BIC) dado por

$$BIC = -2logL + 2klog(N) \tag{2.40}$$

sendo L o valor máximo da função de verosimilhança, k representa o número de parâmetros e N o número de observações.

Tanto em modelos encaixados como em modelos não encaixados, deve-se ter em conta qual o método utilizado na maximização da função verosimilhança. A função verosimilhança tem de se basear nos dados e não na aproximação dos dados.

Quando o objetivo é testar a presença de efeitos aleatórios, tem-se um problema de fronteira e, a distribuição assintótica sob a hipótese nula é uma mistura de distribuições Qui-quadrado. Detalhes podem ser obtidos em Molenberghs & Verbeke (2005).

2.2.5 Análise de Resíduos

Após a estimação de um modelo é necessário verificar os pressupostos do modelo, ou seja, verificar se os pressupostos feitos para o desenvolvimento do modelo estão satisfeitos.

Nos modelos de regressão, um dos métodos mais utilizados para verificar os pressupostos do modelo consiste na análise dos resíduos.

Os resíduos são nada mais do que a diferença entre valores observados e ajustados pelo modelo. O i - ésimo resíduo é dado por

$$\hat{e_i} = y_i - \hat{y_i}, \quad i = 1, 2, ..., n \tag{2.41}$$

Por exemplo, nos modelos lineares clássicos, utilizam-se os resíduos padronizados para verificar homocedasticidade, existência de pontos discrepantes, normalidade e independência dos erros.

Nos GLMM a análise de resíduos é realizada recorrendo a métodos gráficos. Florian Hartig (2022) apresenta um método baseado em simulações para criar resíduos escalonados (também chamados resíduos quantílicos) facilmente interpretáveis para modelos de efeitos mistos. Os resíduos resultantes são padronizados para valores entre 0 e 1 e podem ser interpretados de forma tão intuitiva quanto aos resíduos de uma regressão linear. Esta metodologia encontra-se implementada no package `DHARMa` do `R` que fornece uma série de funções gráficas e testes para problemas de especificação incorreta do modelo, como sobredispersão, subdispersão, autocorrelação espacial e temporal.

Os detalhes estatísticos e teóricos para realizar estas simulações e criar resíduos escalonados encontram-se nas referências do package `DHARMa` (Hartig, 2022).

Neste trabalho, recorre-se ao ambiente de desenvolvimento integrado para cálculos estatísticos e gráficos `R` (R Core Team, 2023), sendo essencialmente utilizada a função `glmer` do package `lme4` (Bates et al., 2015) e package `DHARMa` (Hartig, 2022).

3 Dados

Neste capítulo, num primeiro momento, realiza-se uma descrição detalhada do trabalho exaustivo que foi feito na organização dos dados e, de seguida realiza-se a descrição das variáveis do conjunto de dados.

3.1 Organização dos dados

Inicialmente, realiza-se uma descrição detalhada do tratamento exaustivo que foi aplicado aos dados disponibilizados pela empresa.

Os dados iniciais foram fornecidos em 22 arquivos distintos no formato `.excel`, organizados em 9 vias (*Via Esquerda* - VE, *Via Central* - VC, *Via Direita* - VD, *Via Acelaração* - VAC, *Via de Acelaração Direita* - VACD , *Via de Acelaração Esquerda* - VACE, *Via Desaceleração* - VDC, *Via Desaceleração Direita* - VDCD, *Via Desaceleração Esquerda* - VDCE) e 2 direções (C - crescente e D - decrescente).

A base de dados refere-se a informações relacionadas à autoestrada A4 do Grande Porto, dividida em 6 sublanços, e é constituída por variáveis como o estado do pavimento, o volume de tráfego, a geometria da estrada, entre outras, abrangendo o período de 2014 a 2021.

Nos ficheiros fornecidos, cada observação consiste nos dados de um segmento de 10 metros de extensão numa dada via e num dado sentido, com a exceção dos dados do atrito, em que cada observação corresponde a um segmento de 5 metros de extensão. Foi, portanto, necessário agrupar os dados referentes ao atrito, onde cada observação passa a corresponder a um segmento de 10 metros de extensão numa dada via e num dado sentido. De seguida, procedeu-se à correção de observações que foram recolhidas incorretamente.

Numa primeira análise dos dados, observou-se que em muitos segmentos de 10 metros, a frequência de acidentes rodoviários era zero. Assim, tornou-se necessário reagrupar novamente os dados e cada observação passa agora a corresponder a um segmento de 500 metros de extensão de estrada. Para realizar essa modificação, foi necessário transformar todas as variáveis numéricas, originalmente definidas para segmentos de 10 metros para segmentos de 500 metros, calculando os valores médios, máximos e mínimos nos novos segmentos. É importante realçar que o comprimento padrão do segmento é geralmente de 500 metros, mas pode ser menor, como por exemplo, num dos sublanços com comprimento total de 666 metros, constrói-se um segmento de 500 metros e outro segmento de 166 metros.

Em relação à variável categórica, que indica a caraterística geométrica em planta (alinhamento reto, clotóide e curva circular) no segmento de 10 metros, foram criadas três variáveis: AR, CC e CL. Estas representam a extensão em *Alinhamento Reto*, em *Clotóide* e em *Curva Circular*, respetivamente, no segmento de 500 metros. Do mesmo modo, em relação à variável categórica, que indica a caraterística geométrica em perfil (trainel, curva côncava e curva convexa) no segmento de 10 metros, foram criadas três variáveis: T, CCV e CCX. Estas representam a extensão em *Trainel*, *Curva Côncava* e *Curva Convexa*, respetivamente, no segmento de 500 metros.

Foi ainda necessário construir a variável *ENT_S* com base nos dados das vias de aceleração (VAC, VACE e VACD) e desaceleração (VDC, VDCE e VDCD). Assim, essa variável foi categorizada em E – quando há vias de aceleração no segmento de 500 metros, S - quando há vias de desaceleração no segmento de 500 metros; ES – quando há vias de aceleração e desaceleração no segmento de 500 metros e 0 – para os restantes casos.

Dessa forma, os dados ficaram assim organizados em 3 vias (VE - via esquerda, VC - via central e VD - via direita) e dois sentidos (C- crescente e D - decrescente).

Este processo de limpeza e transformação da base de dados foi muito trabalhoso e demoroso.

No final a base de dados é constituída por 784 observações e por 27 variáveis.

A autoestrada A4 em estudo é constituída por seis sublanços, que correspondem a uma extensão de $8,55$ km.

Na Tabela 3.2 pode-se observar os respetivos sublanços, a sua extensão em quilómetros.

Tabela 3.2: Sublanços de A4

Sublanço	*Pk_Inicial*	*Pk_Final*	*Extensão (km)*
Matosinhos - Sendim	0	0,66	0,66
Sendim - Guifões	0	0,5	0,5
Guifões - Custoias	0,5	2,39	2
Custoias - Via Norte	2,4	5,4	3
Via Norte - Ponte Pedra	5,4	6,4	1
Ponte Pedra - Águas Santas	6,4	7,9	1,5

3.2 Descrição das variáveis

Na Tabela 3.3 observa-se as variáveis consideradas no estudo, e uma breve descrição das mesmas.

Tabela 3.3: Descrição das variáveis da base de dados final

	Variável dependente	Descrição
	Acd_t	Número de total de acidentes
	Variáveis independentes	Descrição
Estado do pavimento	*IRI_md*	Indíce de Rugosidade Internacional médio (m/km)
	IRI_max	Indíce de Rugosidade Internacional máximo (m/km)
	RD_md	Profundidade das Rodeiras - valor médio (mm)
	RD_max	Profundidade das Rodeiras - valor máximo (mm)
	MPD_md	Profundidade Média de Perfil - valor médio (mm)
	MPD_min	Profundidade Média de Perfil - valor mínimo (mm)
	Atrito_md	Coeficiente de Atrito - valor médio (mm)
	Atrito_min	Coeficiente de Atrito - valor mínimo (mm)
Tráfego	*TMDA_lig*	Trafégo médio diário anual de veículos ligeiros
	TMDA_pes	Trafégo médio diário anual de veículos pesados
	TMDA_t	Trafégo médio diário anual de veículos ligeiros e pesados
Características geométrica em planta	*AR*	Extensão em alinhamento reto (m)
	CL	Extensão em clotóide (m)
	CC	Extensão em curva circular (m)
Características geométrica em perfil	*T*	Extensão em trainel (m)
	CCV	Extensão em curva côncava (m)
	CCX	Extensão em curva convexa (m)
Outras variáveis	*Sub*	Sublanço do segmento
	Vias	Via do segmento
	sentido	Sentido do segmento
	Ano	Ano
	Pk_inicial	Valor do quilómetro inicial do segmento
	Pk_final	Valor do quilómetro final do segmento
	N_vias	Número de vias
	ENT_S	Vias de entrada e saída do segmento

A seguir faz-se uma descrição detalhada de cada uma das variáveis:

- **Variáveis referentes ao estado do pavimento**

  O estado do pavimento refere-se às condições funcionais de uma superfície de estrada, incluindo a textura, a regularidade e o atrito, sendo que estas afetam a segurança e o conforto dos utilizadores da estrada.

  - **Índice de Irregularidade Internacional** (**IRI**):

    O IRI é uma medida que quantifica a irregularidade da superfície do pavimento ao longo de um determinado segmento. É amplamente utilizado como um indicador global de conforto e qualidade da estrada. O IRI é expresso em unidades de comprimento, geralmente em metros por quilômetro (m/km) ou em polegadas por milha (in/mi). Quanto menor o valor do IRI, mais regular e suave é a superfície do pavimento. Na base de dados inicial estava presente o valor de IRI no segmento de 10 metros de extensão e construiram-se duas novas variáveis associadas a este índice:

    IRI_md: média do índice de irregularidade internacional no segmento de 500 m.

    IRI_max: máximo do índice de irregularidade internacional no segmento de 500 m.

  - **Profundidade das Rodeiras (RD)**:

    A Profundidade das Rodeiras é uma medida que descreve a profundidade das irregularidades na superfície do pavimento. Ela representa a amplitude das variações verticais ao longo de uma distância específica. A RD é geralmente expressa em milímetros (mm) e é obtida medindo a distância vertical entre o ponto mais alto e o ponto mais baixo dentro de uma amostra de pavimento. Quanto maior o valor da RD, mais irregular é a superfície do pavimento. Na base de dados inicial estava presente o valor de RD no segmento de 10 metros de extensão e construiram-se duas novas variáveis associadas a este índice:

    RD_md: média da profundidade das rodeiras no segmento de 500 m.

    RD_max: máximo da profundidade das rodeiras no segmento de 500 m.

  - **Profundidade Média do Perfil (MPD)**:

    O MPD é uma medida da textura superficial do pavimento e está relacionado com as propriedades de drenagem e atrito. Refere-se à profundidade média das irregularidades numa escala macroscópica, geralmente na gama de 0,3 mm a 5 mm. Na base de dados inicial estava presente o valor de MPD no segmento de 10 metros de extensão e construiram-se duas novas variáveis associadas a este índice:

    MPD_md: média da profundidade média do perfil no segmento de 500 m.

MPD_min: mínimo da profundidade média do perfil no segmento de 500 m.

- **Coeficiente de Atrito** (**Atrito**):

  O atrito é uma medida da resistência ao movimento relativo de dois corpos em contacto. A força de atrito é a força que atua tangencialmente na área de contacto. É fundamental para garantir a segurança e a estabilidade dos veículos, especialmente em condições de travagem ou curvas. O atrito é geralmente avaliado medindo-se a resistência ao deslizamento de um pneu padronizado em relação à superfície do pavimento. É expresso como um coeficiente de atrito, variando de 0 a 1, sendo que valores mais altos indicam uma superfície com melhor aderência. Na base de dados inicial estava presente o valor do Atrito no segmento de 10 metros de extensão e construiram-se duas novas variáveis associadas a este coeficiente:

  Atrito_md: média do coeficiente de atrito no segmento de 500 m.

  Atrito_min: mínimo do coeficiente de atrito no segmento de 500 m.

As medições para o estado do pavimento são feitas de quatro em quatro anos. Como este estudo engloba os anos de 2014 a 2021, as medições realizadas em 2014 são as mesmas dos anos 2015, 2016 e 2017 e as realizadas em 2018 são as dos anos 2019, 2020 e 2021.

- **Variáveis referentes às características geométricas em planta**

  O traçado em planta de uma estrada é composto por curvas circulares de raio constante, curvas de transição de raio variável e alinhamentos retos de raio infinito. As curvas são introduzidas no traçado para realizar a concordância entre alinhamentos retos consecutivos e abrangem dois tipos de curvas: as circulares e as de transição, habitualmente definidas por um troço de clotóide.

  - **Alinhamentos Retos** (em metros):

    É uma secção reta de uma estrada, onde não há curvatura. É usada para permitir uma condução direta e sem desvios em trechos onde não há necessidade de curvas. Geralmente, o alinhamento reto é usado em segmentos de estradas em áreas planas ou para conectar diferentes elementos de curvatura. Na base de dados inicial estava presente o valor de AR no segmento de 10 metros de extensão e construiu-se nova variável *AR_md* que representa a extensão média do alinhamento reto no segmento de 500 m.

  - **Clotóide (ou espiral de transição)**:

É uma curva gradualmente variável usada para fazer a transição suave entre um trecho reto e um trecho curvo numa estrada. A clotóide tem um raio de curvatura que aumenta ou diminui progressivamente ao longo do seu comprimento. Ela ajuda a reduzir as mudanças abruptas de direção, permitindo que os motoristas se ajustem gradualmente à nova curvatura da estrada. A clotóide é frequentemente usada em autoestradas e estradas de alta velocidade. Na base de dados inicial estava presente o valor de CL no segmento de 10 metros de extensão e construiu-se nova variável *CL_md* que representa a extensão média da clotóide no segmento de 500 m.

- **Curva Circular (em metros)**:

  É uma curva com um raio de curvatura constante ao longo de seu comprimento. Em uma curva circular, a estrada segue uma trajetória curva constante, permitindo acomodar curvas mais fechadas ou alterações de direção mais pronunciadas. O raio da curva circular é determinado com base nos critérios de projeto e nas características da estrada, como velocidade de projeto, visibilidade e tipo de veículos esperados. Na base de dados inicial estava presente o valor de CC no segmento de 10 metros de extensão e construiu-se nova variável *CC_md* que representa a extensão média da Curva Circular no segmento de 500 m.

- **Variáveis referentes às características geométricas em perfil**

  A definição altimétrica de uma infraestrutura rodoviária é concretizada a partir de uma linha contínua localizada ao longo da respetiva plataforma, a qual se designa por rasante e corresponde ao perfil longitudinal da via. O perfil longitudinal é constituído por trainéis, que são elementos retos ascendentes (quando apresentam uma inclinação positiva) ou descendentes (quando apresentam uma inclinação negativa), e por concordâncias verticais, que se subdividem em curvas côncavas (de raio negativo) ou convexas (de raio positivo). As curvas verticais são elementos da rasante que permitem realizar a concordância entre dois trainéis de diferentes inclinações. As variáveis que resultam destes elementos são:

  - **Trainel (em metros)**:

    Os trainéis são os elementos mais simples da rasante, onde a inclinação é constante. No caso de ser positiva, é designada por rampa, e, caso seja negativa é designada por declive. Na base de dados inicial estava presente o valor de T no segmento de 10 metros de extensão e construiu-se nova variável *T_md* que representa a extensão média dos trainéis no segmento de 500 m.

- **Curva Convexa (em metros)**:

  As curvas convexas são utilizadas devido à necessidade de assegurar uma distância de visibilidade adequada, de forma a garantir a segurança na circulação da estrada. Estas são usualmente designadas como lombas. Na base de dados inicial estava presente o valor de CCX no segmento de 10 metros de extensão e construiu-se nova variável *CCX_md* que representa a extensão média da CCX no segmento de 500 m.

- **Curva Concâva (em metros)**:

  As curvas concâvas são introduzidas no traçado com o intuito de assegurar a visibilidade noturna e a comodidade na circulação da via. Usualmente designadas como depressões. Na base de dados inicial estava presente o valor de CCV no segmento de 10 metros de extensão e construiu-se nova variável *CCV_md* que representa a extensão média da CCV no segmento de 500 m.

Como seria de esperar, as variáveis referentes às características geométricas, tanto em planta como em perfil, são características da infraestrutura da autoestrada, não se alteram com o tempo.

- **Volume do tráfego**

  O volume de tráfego é uma variável fundamental na modelação de acidentes, uma vez que sem a circulação de veículos não ocorrem acidentes.

  O Tráfego Médio Diário Anual (TMDA) em estradas é uma medida que representa a quantidade diária média de veículos que passam por uma determinada secção de estrada num período de um ano. Essa métrica é usada para estimar o volume de tráfego e é muito importante no projeto e na gestão de estradas. As variáveis referentes ao volume de tráfego são:

  - *TMDA_lig*: Tráfego médio diário anual de veículos ligeiros.
  - *TMDA_pes*: Tráfego médio diário anual de veículos pesados.
  - *TMDA_t*: Tráfego médio diário anual de veículos ligeiros e pesados.

- **Outras variáveis**

  - *Sub*: sublanço da estrada
  - *Sentido*: indica o sentido do segmento de estrada. Está codificada como C - Crescente, D - Decrescente.

- *Vias*: identifica a via do segmento de estrada. Está codificada em VD - via direita, VC - via central, VE - via esquerda
- *PK_inicial*: indica o *km* em que inicia o segmento em estudo.
- *PK_final*: indica o *km* em que acaba o segmento em estudo.
- *N_vias*: indica o número máximo de vias do segmento.
- *Ano*: indica o ano em que foram observados os dados.
- *Acd_t*: número total de acidentes no segmento em estudo.
- *ENT_S*: Entradas e saídas nas autoestradas.

4 Aplicação dos Modelos Lineares Generalizados Mistos

Neste capítulo, são aplicadas as metodologias descritas no segundo capítulo à base de dados fornecida pela empresa Ascendi. O objetivo é desenvolver um modelo para prever o número de acidentes rodoviários na autoestrada A4, com base nos dados fornecidos. Num primeiro momento, é realizada uma análise exploratória dos dados; e, por fim, é estudado o efeito de várias variáveis, tais como o estado do pavimento, a geometria da estrada e o volume de tráfego, sobre a sinistralidade rodoviária.

4.1 Análise Exploratória

Nesta secção pretende-se realizar uma análise exploratória da base de dados.

4.1.1 Variável resposta: número total de acidentes

Inicialmente será apresentada a análise descritiva da variável resposta, número total de acidentes, consoante ano, sublanço e sentido.

A Tabela 4.4 é referente à distruibuição do número acidentes ao longo dos anos.

Em 2016, não houve acidentes num maior número de segmentos de estrada (49 segmentos), enquanto que em 2019, foi o ano em que não houve acidentes num número menor de segmentos.

Tabela 4.4: Distribuição de frequências do número de acidentes por ano

Nº de Acidentes	2014	2015	2016	2017	2018	2019	2020	2021	Nº de Observações
0	36	38	49	29	45	27	36	29	289
1	34	21	25	28	16	27	24	28	203
2	14	19	9	16	14	21	20	17	130
3	6	9	7	14	10	7	5	13	71
4	0	9	2	9	5	6	4	2	37
5	10	5	5	2	1	1	1	2	27
6	1	0	0	0	2	0	0	0	3
7	0	0	2	3	0	0	1	0	6
9	0	2	4	0	0	2	2	0	10
10	0	0	0	0	0	2	0	0	2
11	0	0	0	0	0	0	0	2	2
12	0	0	0	2	0	0	0	0	2
19	2	0	0	0	0	0	0	0	2

Na Figura 1 estão representados o total de acidentes registrados por ano. Observando o gráfico, vê-se que o número total de acidentes, apresenta uma tendência descrescente nos três primeiros anos (2014 a 2016) do estudo, atinge o valor máximo no ano de 2017 e o valor mínimo em 2018.

No ano 2020, apesar do menor volume de tráfego de veículos (Tabela 4.10), por causa das medidas de confinamento, não foi o ano com menor número de acidentes.

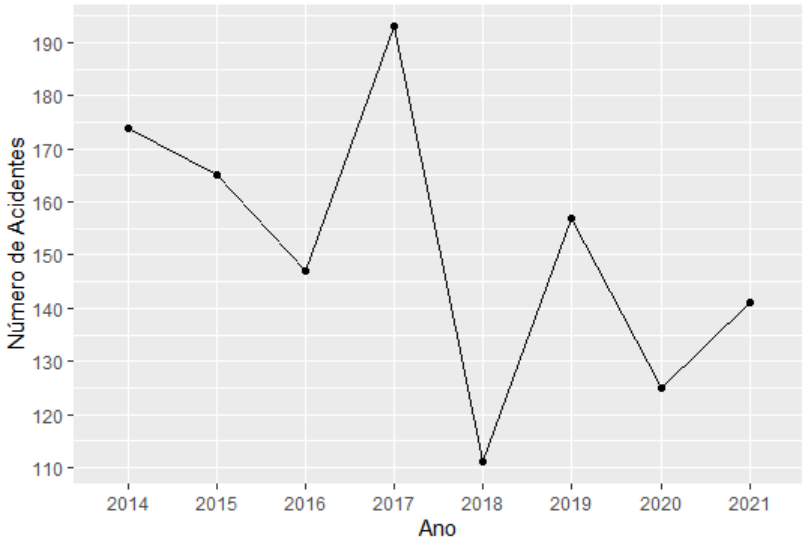


Figura 1: Gráfico de linhas para o número de acidentes por ano

Na Tabela 4.5 observa-se que o número total de acidentes ao longo dos 8 anos foi de 1213. Embora o sublanço Guifões - Custóias não seja o mais extenso, este sublanço teve maior número de acidentes ao longo dos anos e Via Norte - Ponte da Pedra é o sublanço onde houve menor número de acidentes não sendo o sublanço menos extenso. Realça-se que em Matosinhos - Sendim, o número de acidentes diminui ao longo dos anos.

Tabela 4.5: Distribuição de frequências do número de acidentes por ano e sublanço

		Ano							
	Número	*2014*	*2015*	*2016*	*2017*	*2018*	*2019*	*2020*	*2021*
Matosinhos - Sendim	143	55	17	21	16	10	14	6	4
Sendim - Guifões	187	21	18	32	30	12	24	22	28
Guifões - Custóias	289	35	33	61	57	24	39	21	19
Custóias - Via Norte	259	33	45	9	52	28	27	24	41
Via Norte - Ponte Pedra	93	3	15	6	6	10	20	14	19
Ponte Pedra - Águas Santas	242	27	37	18	32	27	33	38	30
Total	1213	174	165	147	193	111	157	125	141

Na Figura 2 pode-se observar o número total de acidentes em cada sublanço, por sentido. No sublanço Guifões - Custóias, no sentido descrescente houve mais 85 acidentes do que no sentido contrário. Nos sublanços Via Norte - Ponte da Pedra, Sendim - Guifões e Ponte da Pedra - Águas Santas houve sempre mais acidentes no sentido crescente que no sentido decrescente. Houve sempre mais acidentes no sentido decrescente nos sublanços Matosinhos - Sendim, Guifões

- Custóias e Custóias - Via Norte.

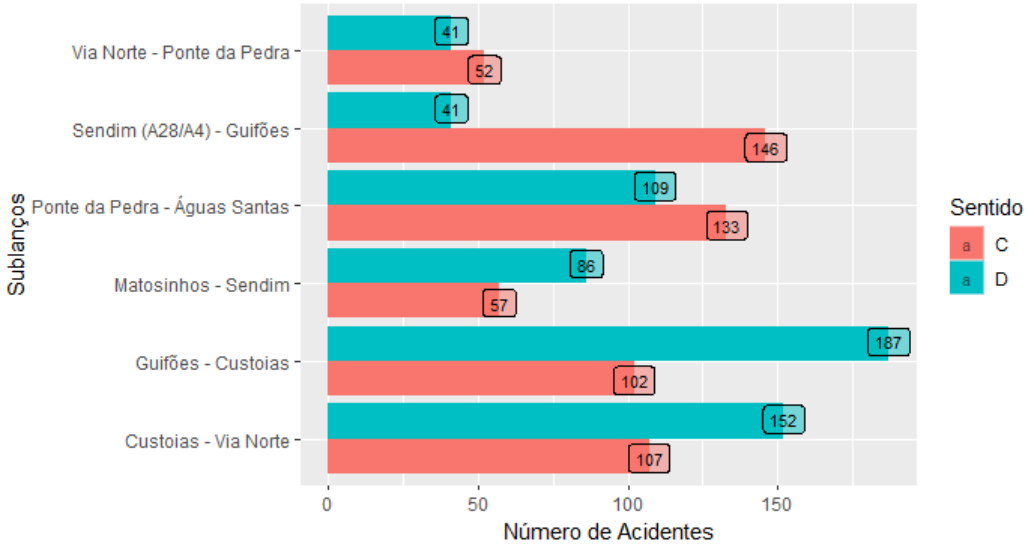


Figura 2: Gráfico de barras do número de acidentes por sublanço e sentido

4.1.2 Variáveis referentes ao estado do pavimento

A Tabela 4.6 é refere-se às estatísticas descritivas da variável IRI. Observa-se que há aumento no valor médio das variáveis *IRI_md* e *IRI_max* de 2014 para 2018, o que indica um aumento da irregularidade na superfície do pavimento. Também se observa que a autoestrada é mais irregular no sentido decrescente.

Tabela 4.6: Estatísticas descritivas da variável IRI

			Média	Mediana	Desvio padrão	Mínimo	Máximo
IRI_md	2014	C	1,35	1,25	0,32	0,944	2,41
		D	1,35	1,28	0,37	0,817	2,69
	2018	C	1,39	1,31	0,31	0,952	2,23
		D	1,45	1,39	0.37	0,953	2,60
IRI_max	2014	C	1,74	1,58	0,559	1,02	3,47
		D	1,75	1,56	0,63	1,06	3,9
	2018	C	1,86	1,67	0,558	1,12	3,5
		D	1,92	1,79	0,558	1,10	3,06

Na Figura 3, pode-se observar que cerca de 50% dos segmentos de estrada apresentam valores de *IRI_md* inferiores a $1,3$ e que existe a presença de valores atípicas (*outliers*), o que indica que existem alguns segmentos com uma irregularidade consideravelmente elevada e com valores

máximos que se destacam dos restantes valores.

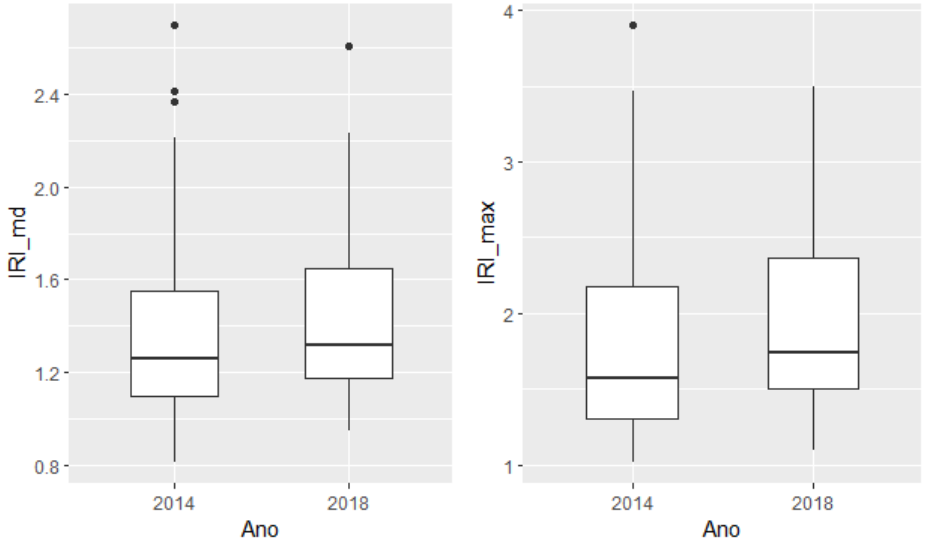


Figura 3: *Boxplot* da variável IRI

Em relação à variável *RD_md*, de 2014 para 2018, houve um decréscimo no seu valor médio (ver Tabela 4.9), o que indica uma maior regularidade na superfície do pavimento. Apesar disso, a diferença é muito pequena, o que indica que na prática não houve alteração deste indicador de qualidade. Observa-se também presença *outliers* (ver Figura 4).

Tabela 4.7: Estatísticas descritivas da variável RD

			Média	Mediana	Desvio padrão	Mínimo	Máximo
RD_md	2014	C	1,24	1,11	0,488	0,479	2,31
		D	1,25	1,12	0,558	0,399	3,28
	2018	C	1,16	1,06	0,499	0,343	2,75
		D	1,22	1,18	0,512	0,478	2,50
RD_max	2014	C	2,45	2,17	0,916	1	4,35
		D	2,42	2,05	0,965	0,7	4,45
	2018	C	2,46	2,28	0,987	0,85	5,55
		D	2,71	2,75	1,06	0,95	4,95

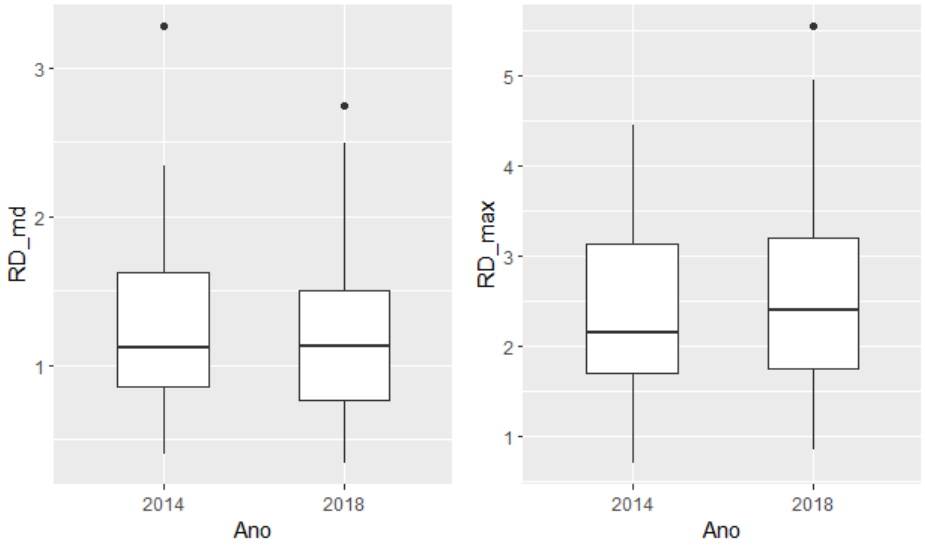


Figura 4: *Boxplot* da variável RD

Relativamente a variável MPD (ver Tabela 4.8), houve um pequeno aumento nos seus valores médios, o que indica uma agravamento natural no estado do pavimento ao longo dos anos.

Relativamente a variável *MPD_min*, observa-se que, no sentido crescente, os valores médios são inferiores comparando com o sentido crescente.

Tabela 4.8: Estatísticas descritivas da variável MPD

			Média	Mediana	Desvio padrão	Mínimo	Máximo
MPD_md	2014	C	1,56	1,58	0,135	1,09	1,78
		D	1,54	1,54	0,109	1,20	1,76
	2018	C	1,60	1,59	0,163	1,19	1,93
		D	1,59	1,56	0,145	1,26	2,05
MPD_min	2014	C	1,14	1,27	0,328	0,55	1,66
		D	1,15	1,21	0,285	0,722	1,6
	2018	C	1,20	1,27	0,303	0,687	1,80
		D	1,23	1,30	0,296	0,570	1,87

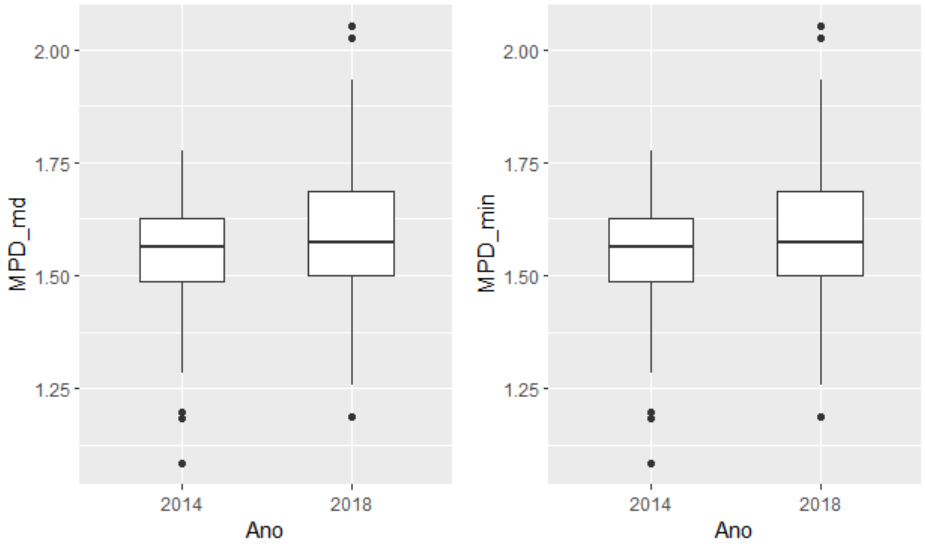


Figura 5: *Boxplot* da variável MPD

Relativamente a variável atrito (ver Tabela 4.9), observa-se um aumento no valor médio da variável *Atrito_md* de 2014 para 2018, o que indica que o desgaste provocado pela passagem dos veículos foi benéfico para este indicador do estado do pavimento. No sentido crescente, a superfície tem melhor atrito, visto que os valores médios são sempre maiores do que no sentido contrário.

Tabela 4.9: Estatísticas descritivas da variável Atrito

			Média	Mediana	Desvio padrão	Mínimo	Máximo
Atrito_md	2014	C	0,443	0,441	0,0406	0,340	0,525
		D	0,433	0,434	0,0344	0,366	0,505
	2018	C	0,472	0,466	0,0517	0,378	0,609
		D	0,452	0,446	0,0421	0,353	0,560
Atrito_min	2014	C	0,362	0,375	0,0543	0,215	0,475
		D	0,355	0,365	0,0492	0,24	0,46
	2018	C	0,416	0,405	0,0510	0,315	0,59
		D	0,394	0,385	0,0397	0,31	0,495

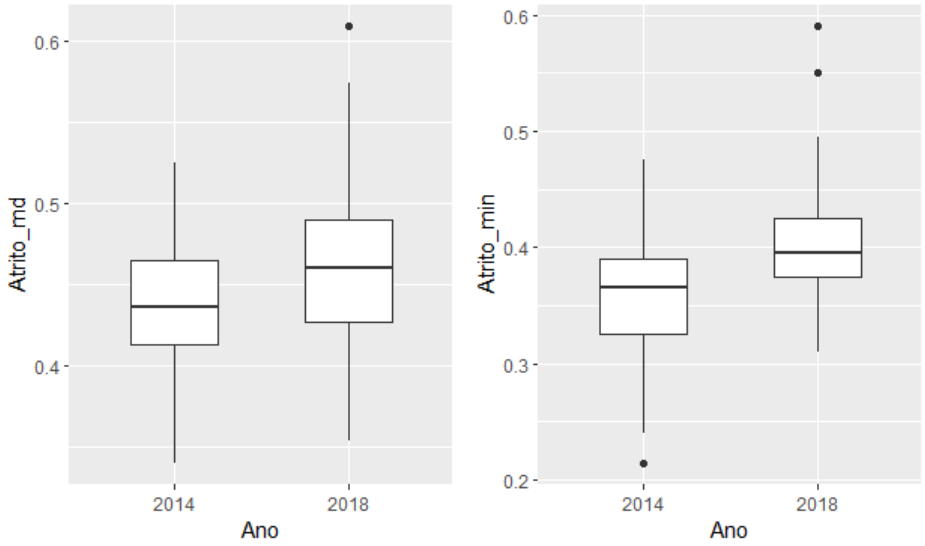


Figura 6: *Boxplot* da variável Atrito

4.1.3 Volume do tráfego

A Tabela 4.10 refere-se aos valores do tráfego médio diário anual de veículos ligeiros e pesados, nos seis sublanços, ao longo dos anos. Observa-se que:

- Em todos os sublanços, o ano de 2019 registou o maior tráfego médio diário de veículos ligeiros. Por outro lado, o ano de 2014, foi o ano com o menor tráfego médio diário de veículos pesados, com exceção no sublanço Matosinhos - Sendim;
- Nos sublanços Sendim - Guifões, Custóias - Via Norte, Via Norte - Ponte da Pedra e Ponte da Pedra - Águas Santas, 2021 foi o ano de maior tráfego médio diário de veículos pesados;
- Nos sublanços Matosinhos - Sendim, Sendim - Guifões e Guifões - Custóias, 2020 foi o ano de menor tráfego médio diário de veículos ligeiros e nos sublanços Custóias - Via Norte, Via Norte - Ponte da Pedra e Ponte da Pedra - Águas Santas, 2014 foi o ano de menor tráfego médio diário de veículos ligeiros;
- No sublanço Custóias - Via norte, houve uma tendência crescente do volume de tráfego das duas categorias de veículos nos seis primeiros anos (2014 a 2019) e no ano 2020 ocorreu um decréscimo do volume do tráfego;
- Do mesmo modo, nos sublanços Via Norte - Ponte da pedra e Ponte da pedra - Águas Santas, houve um crescimento do volume do tráfego médio de veículos ligeiros e pesados de 2014 a 2019, com uma queda em 2020;

- É de realçar que no ano 2020, foram aplicadas medidas de confinamento a nível nacional, resultando numa redução do volume do tráfego de veículos ligeiros e pesados em todos os sublanços.

Tabela 4.10: Volume de tráfego de veículos ligeiros e pesados

Sublanços	*Veículo|Ano*	*2014*	*2015*	*2016*	*2017*	*2018*	*2019*	*2020*	*2021*
Matosinhos-Sendim	*Ligeiros*	17353	17345	19020	19458	20338	22454	15662	19918
	Pesados	320	332	358	361	372	406	284	360
Sendim-Guifões	*Ligeiros*	21768	23707	26878	27348	28292	29655	21569	27329
	Pesados	1028	1394	1187	1133	1162	1220	1118	1426
Guifões-Custoias	*Ligeiros*	21745	22726	26878	27348	28292	29655	21569	27329
	Pesados	1051	1523	1187	1133	1162	1220	1119	1426
Custoias-Via Norte	*Ligeiros*	15513	18907	20288	22030	23734	25090	18039	24243
	Pesados	596	641	738	856	894	895	795	1017
Via Norte-Ponte Pedra	*Ligeiros*	15527	18267	19990	21389	23396	24816	18314	24081
	Pesados	733	769	904	984	1073	1115	1055	1268
Ponte Pedra-Aguas Santas	*Ligeiros*	18374	20765	22668	23839	25560	26874	20030	25920
	Pesados	1003	1092	1138	1197	1287	1384	1346	1560

A Tabela 4.11 apresenta as estatísticas descritivas para as variáveis referentes ao volume do tráfego. Observa-se que o tráfego médio diário de veículos ligeiros e pesados é 23203, com um máximo de 30875 veículos.

Tabela 4.11: Estatísticas descritivas do volume de tráfego

Volume do Tráfego	Média	Desvio Padrão	Mediana	Mínimo	Máximo
TMDA_lig	22252	3779	22288	14186	29655
TMDA_pes	950	322	1003	284	1560
TMDA_t	23203	4009	23140	14792	30875

4.1.4 Variáveis referentes à características geométricas

Na Tabela 4.12 apresenta-se as estatísticas descritivas relativamente às variáveis referentes à características geométricas da estrada em cada segmento de 500 m.

Em termos médios, a extensão dos alinhamentos retos, das curvas circulares das clotóides é

semelhante, o que já não se passa com as características geométricas em perfil, uma vez que a extensão média em curva côncava é mais pequena.

Tabela 4.12: Estatísticas descritivas para as variáveis referentes à características geométricas

	Mínimo	Máximo	Média	Desvio Padrão
Características geométricas em planta				
AR	0	500	164,3	147,046
CC	0	500	163,112	147,025
CL	0	330	141,7	100,8
Características geométricas em perfil				
CCV	0	490	74,49	133,2
CCX	0	500	151,122	175,5
T	0	500	243,469	180,093

4.2 Ajustamento do modelo

Após a análise descritiva, iniciou-se a estimação do modelo linear generalizado misto (GLMM), usando as técnicas adequadas para dados com estruturas de dependência.

A estimação dos modelos no `R`, foi através da função `glmer()` da biblioteca `lme4`. Esta função ajusta os dados a um modelo linear generalizado misto aplicando o método adaptativo da quadratura de Gauss-Hermite para realizar a inferência com base na máxima verosimilhança.

Na construção de um modelo GLMM, é importante estudar quais os efeitos aleatórios que devem ser considerados. Foram estudados dois modelos:

O primeiro modelo com um efeito aleatório na interseção e o segundo com dois efeitos aleatórios um na interseção e outro no declive da variável tempo.

Na construção dos modelos, a variável volume de tráfego total ($TMDA\_t_{ij}$) foi considerada como um `offset`, para ter em consideração os diversos volume de tráfego nos diferentes sublanços e anos.

Considere a variável resposta, Acd_t_{ij} - número total de acidentes no segmento i e *ano* j, $TMDA\_t_{ij}$ - o volume do tráfego total no segmento i no ano j, com $i = 1, ..., 98$ e $j = 1, ..., 8$, tem-se:

$$Acd\_t_{ij} \sim \texttt{Poisson}(\mu_{ij}) \tag{4.42}$$

e o modelo com um efeito aleatório na interseção (GLMM I)

$$log(\mu_{ij}|b_i) = \beta_0 + b_{0i} + log(TMDA\_t_{ij}) \tag{4.43}$$

e o modelo com dois efeitos aleatórios, na interseção e no declive (GLMM II)

$$log(\mu_{ij}|b_i) = \beta_0 + b_{0i} + b_{1i} \times Ano_{ij} + log(TMDA\_t_{ij}) \tag{4.44}$$

A comparação de modelos não encaixados com diferentes estruturas de efeitos aleatórios, foi realizada através dos critérios de informação AIC e BIC. Foi selecionado o modelo com dois efeitos aleatórios, que corresponde ao modelo com menor valor AIC (Tabela 4.13).

Tabela 4.13: Comparação de modelos com diferentes efeitos aleatórios

Modelo	AIC	BIC	logLik
GLMM I	2634,4	2643,7	-1315,2
GLMM II	2587,7	2606,3	-1289,8

Na formação da base de dados foram construidas duas variáveis para cada variável corresponde ao estado de pavimento, *IRI_md* e *IRI_max*, *RD_md* e *RD_max*, *MPD_md* e *MPD_min*, e, *Atrito_md* e *Atrito_min*. De seguida, procede-se à análise de qual a variável a usar no modelo, construindo o modelo generalizado misto simples com cada uma destas variáveis.

Tabela 4.14: GLMM simples das variáveis relativas ao estado do pavimento

Variável	AIC	Valor p
IRI_md	2586,9	0,0946
vs		
IRI_max	2587,9	0,1488
RD_md	2589,4	0,6235
vs		
RD_max	2589,5	0,6727
MPD_md	2587,6	0,1446
vs		
MPD_min	2582,7	0,0083
Atrito_md	2587,5	0,1400
vs		
Atrito_min	2583	0,0098

Relativamente a variável *IRI* observa-se que o valor de AIC dos dois modelos GLMM simples são praticamente iguais e ambas as variáveis não são estatisticamente significativas. Neste caso resolveu-se escolher a variável *IRI_md* uma vez que o modelo apresenta valor de AIC ligeiramente inferior.

Quanto a variável *RD*, observa-se que os valores de AIC são muito próximos e as variáveis não são estatisticamente significativas e optou-se pela variável *RD_md*.

Relativamente a variável *MPD* selecionou-se a variável *MPD_min* uma vez que o GLMM simples com a variável *MPD_min* apresenta menor valor de AIC do que o GLMM simples com a variável *MPD_md*.

Quanto a variável *Atrito*, escolheu-se a variável *Atrito_min* cujo modelo apresenta menor AIC do que o modelo com a variável *Atrito_min*.

Da mesma forma, será realizada a seleção das variáveis associadas às características geométricas. É de referir que as variáveis relacionadas às características geométricas em planta (AR, CC e CL) e às características geométricas em perfil (T, CCV e CCX) apresentam uma forte correlação entre si. Isso ocorre porque, no caso das características geométricas em planta, a soma da extensão em AR com a extensão em CC e CL é igual à extensão total do segmento

de estrada, geralmente de 500 metros e o mesmo acontece com as caraterísticas geométricas em perfil.

Deste modo, com as variáveis referentes às características geométricas em planta (ou às características geométricas em perfil) contruíram-se três modelos com duas das três variáveis possíveis (Tabela 4.15).

Tabela 4.15: GLMM para as variáveis referentes a características geométricas

Combinação de variáveis	AIC	Valor p
$AR + CC$	2578,7	0,00154
$AR + CL$	2582,9	0,01245
$CC + CL$	2586,2	0,06346
$CCV + CCX$	2586,2	0,06589
$CCV + T$	2586,3	0,06745
$CCX + T$	2591,3	0,81570

Relativamente a combinação das variáveis referentes as características geométricas em planta, selecionou-se o modelo com as variáveis $AR + CC$ uma vez que este modelo apresenta menor valor de AIC (AIC = 2578,7) do que as outras duas combinações (Tabela 4.15).

Quanto a combinação das variáveis referentes a características geométricas em perfil, observa-se que os valores de AIC dos vários modelos são próximos e que as combinações não são estatisticamente significativas. Neste caso resolveu-se escolher, o modelo com as variáveis $CCV + CCX$ que apresenta valor de AIC ligeiramente inferior (Tabela 4.15).

Após a escolha das variáveis, apresenta-se na Tabela 4.16 o modelo completo.

Tabela 4.16: Estimativas do GLMM completo

Efeitos Fixos			
Variáveis explicativas	Estimativa dos coeficientes (β)	Erro padrão	Valor P
(intercept)	- 9,517	0,577	$<$ 2e-16 ***
ENT_SE	- 0,564	0,165	0,000609 ***
ENT_SES	- 0,734	0,171	1.71e-05 ***
ENT_SS	- 0,887	0,149	2.79e-09 ***
SentidoD	- 0,059	0,067	0,3851
ViasVD	0,353	0,129	0,00609 **
ViasVE	0,443	0,104	2.10e-05 ***
N_vias	0,227	0,079	0,004383 **
IRI_md	0,281	0,224	0,2096
RD_md	- 0,029	0,104	0,778006
Atrito_min	- 3,079	0,740	3.19e-05 ***
MPD_min	- 0,306	0,167	0,0666 .
AR	0,008	0,008	0,3685
CC	0,008	0,011	0,4228
CCX	- 0,007	0,006	0,2298
CCV	0,013	0,006	0,04337 *

Retirando-se sucessivamente as variáveis estatisticamente não significativas, obtém-se as estimativas dos coeficientes do modelo final e as estimativas da variância dos efeitos aleatórios b_{0i} e b_{1i} (Tabela 4.17).

O modelo generalizado misto de Poisson é dado por:

$$Acd\_t_{ij}|\mathbf{b}_i \sim \texttt{Poisson}(\mu_{ij}) \tag{4.45}$$

$$\begin{aligned} log(\mu_{ij}|\mathbf{b}_i) = log(TMDA\_t_{ij}) + \beta_0 + b_{0i} + \beta_1 Atrito\_min_{ij} + \beta_2 MPD\_min_{ij} \\ + \beta_3 CCV_{ij} + \beta_4 ENT\_S + b_{1i} Ano_{ij} \end{aligned} \tag{4.46}$$

Tabela 4.17: Estimativas do GLMM final

Efeitos Fixos				
Variáveis explicativas	Estimativa dos coeficientes (β)	Exp (β)	Erro padrão	Valor P
(intercept)	- 9,145	0,0001	0,411	< 2e-16 ***
ENT_SE	- 0,465	0,6281	0,143	0,00111 **
ENT_SES	- 0,645	0,5246	0,155	3.04e-05 ***
ENT_SS	- 0,822	0,4395	0,137	1.78e-09 ***
VíasVD	0,418	1,5189	0,095	1.18e-05 ***
VíasVE	0,443	1,5573	0,091	1.12e-06 ***
N_vias	0,235	1,2649	0,076	0,00203 **
CCV	0,017	1,0171	0,006	0,00496 **
Atrito_min	-2,975	0,0510	0,721	3.65e-05 ***
MPD_min	-0,365	0,6941	0,160	0,02236 *
Efeitos Aleatórios	Desvio Padrão			
b_{0i}	0,729			
b_{1i}	0,092			
corr(b_{0i}, b_{1i}) = - 0,70				

Na interpretação das estimativas dos coeficientes obtidas do modelo final (Tabela 4.17) e mantendo constante as outras variáveis, verifica-se:

- Com o aumento de uma unidade no valor mínimo do atrito (*Atrito_min*), a taxa esperada de acidentes diminui cerca de 95%.
- Com o aumento de uma unidade na profundidade média em perfil (*MPD_min*), a taxa esperada de acidentes diminui cerca de 31%.
- Com o aumento de uma unidade na extensão em curva côncava *CCV*, a taxa esperada de acidentes aumenta cerca de 1,7%.
- A presença de vias de aceleração e desaceleração num segmento diminui a taxa esperada de acidentes em cerca de 47,5% em relação a segmentos sem vias de aceleração e desaceleração.
- Com o aumento de uma unidade no número de vias, a taxa esperada de acidentes aumenta cerca de 26,5%.

- Na via da direita ou da esquerda de um segmento, a taxa esperada de acidentes aumenta cerca de 50% em relação à via central do segmento.

a) Diagnóstico do modelo

Neste trabalho será aplicado a análise de diagnóstico proposto por Florian Hartig (Hartig, 2022) usando o package `DHARMa` no `R`, que é baseada em simulações para criar resíduos escalonados (também chamados de resíduos quantílicos), com valores entre 0 e 1, interpretáveis no modelo final estimado.

Na Figura 7, apresenta-se o gráfico de quantil-quantil. Observa-se que os pontos estão alinhados em cima da linha vermelha, o que nos indica que no modelo ajustado não parece haver desvios relevantes dos valores esperados.

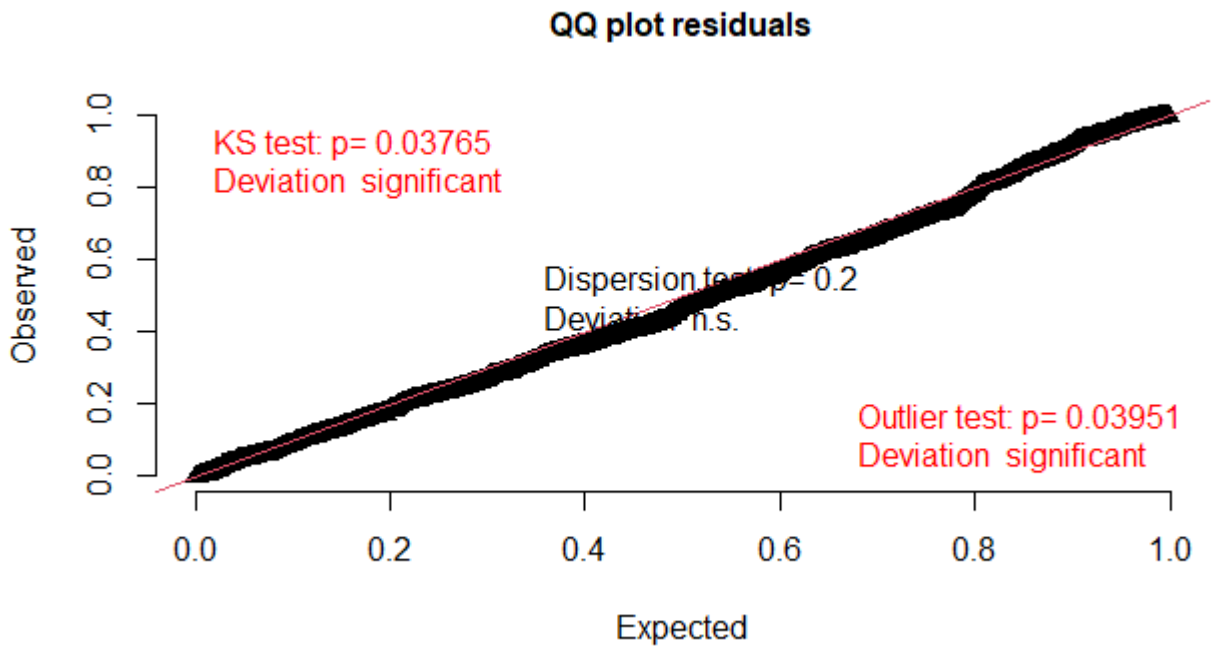


Figura 7: Gráfico QQ dos Resíduos Quantílicos

Na Figura 8, apresenta-se o gráfico dos resíduos *versus* os valores preditos pelo modelo final. Este gráfico permite detectar desvios da uniformidade na direção y. As linhas são regressões quantílicas que mostram os quantis 0,25, 0,50 e 0,75. Estas linhas deveriam ser retas horizontais para cada quantil. As linhas observadas no gráfico obtido para o modelo não são retas horizontais, mas muito próximas destas.

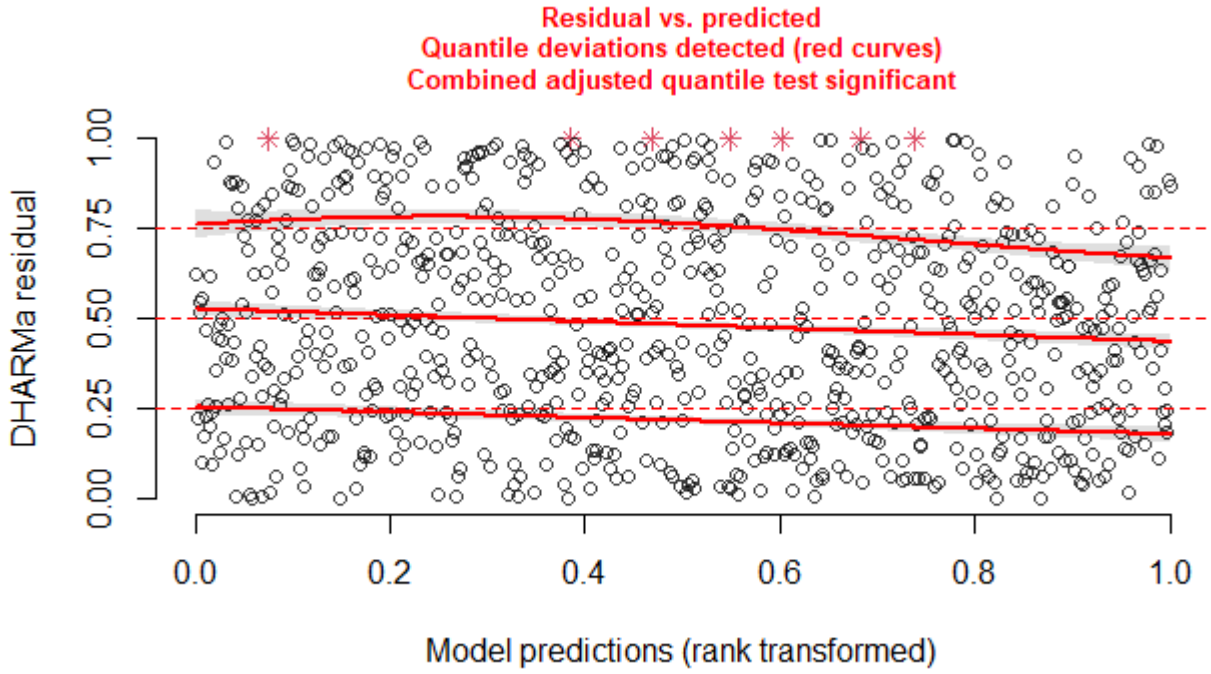


Figura 8: Resíduos Quantílicos versus Valores Preditos

5 Conclusão

A segurança viária é um problema de saúde pública com impactos significativos em termos humanos e materiais. Apesar da diminuição do número de acidentes nos países desenvolvidos ao longo das últimas décadas, a compreensão dos fatores subjacentes que os influenciam, com o objetivo de reduzir a gravidade dos danos causados às vítimas, permanece incompleta.

Neste trabalho, o objetivo principal foi identificar os fatores que influenciam a frequência de acidentes rodoviários na autoestrada A4. Para alcançar esse objetivo, foram utilizados dados relativos a segmentos da autoestrada A4, cobrindo uma extensão de 8,55 quilómetros disponibilizados pela Concessionária Ascendi, abrangendo um período de oito anos, de 2014 a 2021.

Inicialmente foi realizada uma análise exploratória dos dados para compreender os dados e investigar se há indícios de associação entre a variável resposta e as variáveis explicativas, associação essa que pode posteriormente ser melhor estudada e compreendida com a análise de regressão. De seguida, utilizaram-se Modelos Lineares Generalizados Mistos para modelar a variável resposta sob a forma de contagem, recorrendo à distribuição de Poisson.

A análise exploratória dos dados permitiu concluir que:

- Houve um total de 1213 acidentes ao longo dos oito anos;
- 2017 foi o ano com um maior número de acidentes;
- Nos sublanços Matosinhoss- Sendim, Guifões - Custóias e Custóias - Via Norte, houve mais acidentes no sentido decrescente;
- Houve mais acidentes na via esquerda;
- No sublanço Custóias - Via Norte houve maior volume de tráfego médio diário de veículos ligeiros e pesados. No entanto foi o sublanço Guifões - Custóias que houve maior número de acidentes;
- No sublanço Sendim - Guifões houve menor volume de tráfego médio diário de veículos ligeiros e no sublanço Matosinhos - Sendim houve menor volume de tráfego de veículos pesados.

Um modelo linear generalizado misto para contagens com efeitos aleatórios na intersecção e no declive da variável *Ano* foi ajustado aos dados.

Os resultados do modelo ajustado demonstraram que as variáveis *Atrito_min*, *MPD_min*, *CCV*, *ENT_S*, *Vias* e *N_vias* influenciam o número de acidentes na A4.

As estimativas dos coeficientes da variável *CCV*, *Vias* e *N_vias* são positivas, o que significa o aumento da taxa esperada de acidentes com o aumento da extensão em curva côncava e com o aumento número de vias.

Relativamente às variáveis *Atrito_min* e *MPD_min* os coeficientes estimados são negativos o que indica uma diminuição na taxa esperada de acidentes com o aumento do valor do Atrito mínimo e com o aumento da profundidade média em perfil.

A estimativa dos coeficientes da variável categórica *ENT_S* é negativa o que significa que a presença de vias de aceleração e desaceleração diminui a taxa esperada de acidentes.

Para trabalho futuro, seria interessante aplicar Modelos Poisson Inflacionados de Zeros com efeitos aleatórios a este conjunto de dados, visto que em muitos segmentos de estrada a frequência de acidentes é zero.

Bibliografia

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, *88*(421), 9–25.

Cabral, M. S., & Gonçalves, M. H. (2011). Análise de dados longitudinais. *Sociedade Portuguesa de Estatística, Lisboa*.

Dragomanovits, A., Basta, O., Deliali, A., & Yannis, G. (2022). A state-of-practice review on crash prediction modelling. 8*th Road Safety Simulation International Conference*.

Elvik, R., & Katharina Høye, A. (2023). Changes over time in the relationship between road accidents and factors influencing them: The case of norway. *Accident Analysis Prevention*, *183*, 106989. Retrieved from https://www.sciencedirect.com/science/article/pii/S0001457523000362 doi: https://doi.org/10.1016/j.aap.2023.106989

Hartig, F. (2022). Dharma: Residual diagnostics for hierarchical (multi-level / mixed) regression models [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=DHARMa (R package version 0.4.6)

Harville, D. (1977). The use of linear-model methodology to rate high school or college football teams. *Journal of the American Statistical Association*, *72*(358), 278–289.

Imprialou, M.-I. M., Quddus, M., Pitfield, D. E., & Lord, D. (2016). Re-visiting crash–speed relationships: A new perspective in crash modelling. *Accident Analysis Prevention*, *86*, 173-185. Retrieved from https://www.sciencedirect.com/science/article/pii/S000145751530083X

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*(4), 963–974.

Liu, X. (2016). Generalized linear mixed models on nonlinear longitudinal data. In *Methods and applications of longitudinal data analysis* (p. 243-279). Oxford: Academic Press.

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, *44*(5), 291-305. Retrieved from https://www.sciencedirect.com/science/article/pii/S0965856410000376

Mannering, F. L., & Bhat, C. R. (2014). Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research*, *1*, 1-22. Retrieved from https://www.sciencedirect.com/science/article/pii/S2213665713000031 doi: https://doi.org/10.1016/j.amar.2013.09.001

Mannering, F. L., Shankar, V., & Bhat, C. R. (2016). Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic Methods in Accident Research*, *11*, 1-16. Retrieved from https://www.sciencedirect.com/science/article/pii/S2213665716300100 doi: https://doi.org/10.1016/j.amar.2016.04.001

Molenberghs, G., & Verbeke, G. (2005). The generalized linear mixed model (glmm). In *Models for discrete longitudinal data* (pp. 265–280). New York, NY: Springer New York.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370–384.

Rolison, J. J., Regev, S., Moutari, S., & Feeney, A. (2018). What are the factors that contribute to road accidents? an assessment of law enforcement views, ordinary drivers' opinions, and road accident records. *Accident Analysis Prevention*, *115*, 11-24. Retrieved from https://www.sciencedirect.com/science/article/pii/S0001457518300873 doi: https://doi.org/10.1016/j.aap.2018.02.025

Silva, C. M. P., Bravo, J. M., & Gonçalves, J. M. (2021). Impacto económico e social da sinistralidade rodoviária em portugal.

Singer, J. M., Nobre, J. S., & Rocha, F. M. M. (2018). Análise de dados longitudinais.