View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Repositório Aberto da Universidade do Porto

José Luís Mourão Ferreira

Modelos de Regressão para Previsão de Sinistros

Departamento de Matemática
Faculdade de Ciências da Universidade do Porto
setembro de 2013

José Luís Mourão Ferreira

Modelos de Regressão para Previsão de Sinistros

Tese submetida à Faculdade de Ciências da Universidade do Porto para obtenção do grau de Mestre em Engenharia Matemática

Orientador: Prof. Doutor Joaquim Fernando Pinto da Costa
Coorientador: Dr. Luís Maranhão
Coorientadora: Prof.ª Doutora Ana Rita Gaio

Departamento de Matemática
Faculdade de Ciências da Universidade do Porto
setembro de 2013

Agradecimentos

Agradeço ao Professor Doutor Joaquim Fernando Pinto da Costa, ao Dr. Luís Maranhão e à Professora Doutora Ana Rita Gaio pela orientação nesta dissertação, a qual não poderia ter sido realizada sem os conselhos, a paciência, disponibilidade e atenção mostrada por todos.

Agradeço à minha família, em especial aos meus pais, por todo o apoio disponibilizado ao longo da minha educação.

Agradeço aos meus amigos, por todas as conversas e desabafos que aliviaram as angústias existentes.

Resumo

No ramo dos seguros automóveis, as companhias de seguros necessitam de identificar o risco de cada um dos seus clientes sofrer um acidente, de forma a calcular o prémio adequado que o cliente deve pagar. Se o valor do prémio for inferior aos custos que a companhia terá de suportar em caso de sinistro, esta incorre em perdas financeiras. Surge assim a necessidade de estudar as características que mais influenciam os acidentes.

Neste contexto, a companhia de seguros AXA S.A. tem utilizado três variáveis ligadas aos automóveis para determinar o risco de acidentes a eles associado: cilindrada, potência e o quociente entre o peso e a potência. O objetivo desta dissertação passa pela identificação de outras variáveis de natureza automóvel que possam ter um efeito significativo sobre o número de acidentes.

O facto de a variável resposta se tratar de contagens remeteu a orientação do estudo para modelos lineares generalizados para dados de contagem, tendo sido consideradas as distribuições de Poisson, Binomial Negativa e Binomial, bem como distribuições de Poisson com zeros inflacionados e de Poisson com barreira.

O mau ajustamento das duas primeiras distribuições às contagens positivas levou a uma redefinição da variável resposta apenas em ausência de acidentes versus presença de acidentes de forma a aplicar-se um modelo de regressão logística. Ainda assim, não se verificaram melhorias. Pensamos que esta má qualidade do ajustamento é devida à enorme presença de zeros nos dados recolhidos.

De seguida, foram utilizados modelos adequados para uma amostra deste tipo. Foram aplicados dois modelos: um modelo de regressão de Poisson com zeros inflacionados e um modelo de regressão de Poisson com barreira. Ambos os modelos revelaram ajustar bem os zeros mas não as contagens positivas.

Palavras-chave: MODELO COM BARREIRA, MODELO COM ZEROS INFLACIONADOS, REGRESSÃO DE POISSON, REGRESSÃO BINOMIAL NEGATIVA, REGRESSÃO LOGÍSTICA.

Abstract

In the field of car insurance, insurance companies need to identify the risk for each of its customers to have an accident, in order to calculate the appropriate premium that the customer must pay. If the value of the prize is less than the costs that the company will have to bear in the event of a claim, financial losses will incur. Thus arises the need to study the characteristics that influence accidents.

In this context, the insurance company AXA S.A. has used three car related variables to determine the risk of accidents associated with them: displacement, power and the ratio between weight and power. The aim of this work involves the identification of other car-related variables that may have a significant effect on the number of accidents.

The fact that the response variable is in the form of count data geared the study towards generalized linear models for count data, having been considered Poisson, Negative Binomial and binomial distributions, as well as a zero-inflated Poisson and a Poisson hurdle distributions.

The poor adjustment of the first two distributions regarding the positive counts led to a redefinition of the response variable as presence versus absence of car accidents so that a logistic regression model could be applied. Still, improvements were not verified. We think that the poor quality of the adjustment is due to the massive presence of zeros in the data collected.

Then, appropriate models were used for a sample of this type. We applied two models: a zero-inflated Poisson regression model and a Poisson hurdle model. Both models showed a good fit for the zero counts but not the positive counts.

Keywords: HURDLE MODEL, ZERO INFLATED MODEL, NEGATIVE BINOMIAL REGRESSION, POISSON REGRESSION, LOGISTIC REGRESSION.

Para a minha avó Alice e o meu tio Carlos.

Conteúdo

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Esta tese tem como objectivo principal a aplicação de modelos de regressão para dados de contagem a dados de acidentes de automóveis da seguradora AXA S.A..
Um seguro define-se como um contrato (apólice) através do qual uma entidade (um indivíduo ou uma empresa) recebe, de uma companhia de seguros, proteção financeira ou reembolso por perdas sofridas. A companhia de seguros garante assim uma compensação monetária que cobre total ou parcialmente os custos decorrentes de acidentes, mortes ou catástrofes naturais. À entidade que celebra o contrato com a companhia é imputado o pagamento de uma prestação designada por prémio de seguro. O ramo dos seguros divide-se principalmente em 2 vertentes: seguros de vida e seguros não vida. Os primeiros englobam seguros de vida e produtos financeiros, dos quais fazem parte seguros de capitalização e planos poupança reforma. Os seguros não vida abrangem a responsabilidade civil automóvel, acidentes de trabalho e pessoais, doenças, seguros multirriscos habitação, entre outros. Tal como noutros países, em Portugal o seguro de responsabilidade civil automóvel é obrigatório. Segundo o Instituto de Seguros de Portugal, o seguro obrigatório de automóvel assegura o pagamento de indemnizações por danos corporais e materiais causados a terceiros e às pessoas transportadas, com excepção do condutor do veículo. Dada a obrigatoriedade da aquisição deste serviço e o elevado número de indivíduos e empresas com veículos automóveis ao seu dispor, torna-se importante para uma companhia de seguros avaliar o risco inerente a cada cliente, de forma a minimizar os custos decorridos. Se um cliente tiver um risco superior ao normal, a companhia terá interesse então em aumentar o valor do prémio para poder compensar os eventuais custos que terá com o cliente.
Atualmente, a AXA utiliza como fatores de risco as variáveis cilindrada, potência e o quociente entre o peso e a potência. Pretende-se averiguar se existem outros fatores relacionados com as viaturas que aumentem o risco de um cliente sofrer um sinistro. No entanto, note-se que existem variáveis que influenciam a ocorrência de acidentes que não estão ligadas às viaturas. Fatores humanos como a idade e o sexo ou fatores ambientais como a localização, as condições atmosféricas e período do dia são considerados determinantes na ocorrência de acidentes rodoviários (**?**); **?**); **?**); **?**)).
Os modelos de regressão para dados de contagem são uma ferramenta útil em inúmeras áreas e advêm da necessidade de uma metodologia que consiga tratar dados onde a variável resposta só toma valores inteiros positivos ou nulos. Estes modelos também permitem ultrapassar os pressupostos dos modelos lineares clássicos, que geralmente

não serão verificados.
Neste contexto surge a teoria de modelos lineares generalizados, mencionados pela primeira vez por **?**) e posteriormente desenvolvidos por **?**), **?**) e **?**), com o objetivo de possibilitar o tratamento de respostas que não seguissem distribuições normais. A distribuição mais utilizada em dados de contagem é a distribuição de Poisson e, consequentemente, o modelo de regressão de Poisson, que pressupõe a igualdade entre a média e a variância da variável resposta. Em muitas situações, este pressuposto é desrespeitado já que a variância tende a ser superior à média. Este fenómeno é denominado de sobredispersão e pode ser solucionado recorrendo a um modelo de regressão associado à distribuição Binomial Negativa (**?**); **?**)).
É também frequente observar-se um número excessivo de valores nulos na variável resposta. Esta situação impossibilita o ajustamento dos dados através da distribuição de Poisson e muitas vezes também através da distribuição Binomial Negativa. Surge assim a necessidade de utilizar modelos que possam lidar com este problema, como por exemplo, modelos com zeros inflacionados e modelos com barreira. Os primeiros modelos consideram que os valores nulos podem ser verdadeiros, ou seja, provenientes de um processo de contagem, ou falsos, quando resultam de uma massa pontual em zero. Pelo contrário, os modelos com barreira não fazem distinção entre os tipos de zeros; o processo de contagem não pode originar valores nulos, dividindo assim a variável resposta em ausências e presenças. Exemplos da utilização destes modelos podem ser encontrados em vários fontes: número de defeitos num processo de fabricação (**?**)); número de crianças com cáries (**?**)); dados de violência doméstica (**?**)); análise do desenvolvimento motor de uma criança (**?**)).
A implementação de todos os modelos supra mencionados foi efetuada através do *software* **?**) e recorrendo a bibliotecas pertencentes ao diretório do R.

1.1 Estrutura da tese

O capítulo 1 é constituído por duas secções: a primeira prende-se com a introdução do tema e dos objetivos deste trabalho enquanto a segunda descreve a estrutura da tese.

No capítulo 2 descrevem-se detalhadamente os conceitos teóricos sobre os modelos lineares generalizados que foram estudados, sendo dado destaque aos modelos de regressão para dados de contagem aplicados nesta tese.

O capítulo 3 diz respeito à descrição detalhada dos modelos de regressão com um número excessivo de zeros e está dividido em duas secções. Na secção 3.1 são descritos os modelos com zeros inflacionados. Os modelos com barreira são apresentados na secção 3.2..

O trabalho envolvendo a análise dos dados está contido no capítulo 4. A primeira secção corresponde à descrição da base de dados. A implementação dos modelos de regressão é efetuada na segunda secção.

No capítulo 5 é realizada a análise dos modelos obtidos no capítulo 4 no que diz respeito

às variáveis explicativas e o seu efeito sobre o número de acidentes.

As conclusões finais obtidas referentes ao trabalho desenvolvido bem como propostas de trabalhos futuros são apresentadas no capítulo 6.

Capítulo 2

Modelos Lineares Generalizados

Desde a sua incepção, os métodos de regressão têm-se tornado a escolha padrão para analisar e descrever a relação entre uma variável resposta e um conjunto de variáveis explicativas. O objetivo da construção de modelos estatísticos é encontrar o modelo que melhor se ajuste aos dados e que respeite o princípio da parcimónia, não descurando, no entanto, o seu enquadramento e interpretação dentro da situação em estudo. O modo de regressão mais antigo começou por usar o método dos mínimos quadrados, formulado independemente no início do século XIX por **?**), **?**) e **?**). Desde então, o modelo linear normal, usando também o método da máxima verosimilhança, tornou-se no recurso mais comum para analisar a relação entre a média da variável resposta e uma combinação linear das variáveis explicativas. No entanto, este modelo tinha como principal assumpção a normalidade da variável resposta. Na realidade, é possível encontrar casos onde tal não se verifica, sendo preciso encontrar outros meios de os modelar. Ao longo do século XX vários modelos foram desenvolvidos para lidar com situações destas, sendo alguns exemplos disso o modelo complementar log-log para ensaios de diluição (**?**)), o modelo probit (**?**)) para proporções, o modelo logit (**?**)), os modelos de regressão para análise de sobrevivência (**?**); **?**)). **?**) introduziram os Modelos Lineares Generalizados, que correspondem a uma generalização dos modelos anteriormente mencionados, bem como de vários outros. Estes modelos assumem que a variável resposta segue uma distribuição pertencente à família exponencial, relacionando a média da resposta com uma combinação linear das variáveis explicativas através de uma função de ligação. Neste capítulo é apresentada a formulação teórica respeitante a estes modelos.

2.1 Família Exponencial

Seja Y uma variável aleatória. Se a sua função densidade (ou massa) de probabilidade se puder escrever na forma (**?**)):

$$f(y|\theta,\phi) \quad = \quad exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{2.1}$$

onde θ e ϕ são parâmetros escalares, a, b e c são funções reais conhecidas, então diz-se que Y tem distribuição pertencente à família exponencial e que essa distribuição está na

forma canónica.

Na fórmula anterior, θ designa-se por parâmetro canónico de localização, ϕ é denominado por parâmetro de dispersão ou parâmetro de escala. Além disso, assume-se que b é uma função diferenciável.

A função a é normalmente da forma $a(\phi) = \frac{\phi}{w}$, onde w é um peso conhecido *a priori* que varia de observação para observação (**?**)), podendo-se reescrever a fórmula acima da seguinte forma:

$$f(y|\theta,\phi,w) = exp\left\{\frac{w}{\phi}(y\theta - b(\theta)) + c(y,\phi,w)\right\} \tag{2.2}$$

De acordo com **?**):

$$E(Y) = \mu = b'(\theta)$$

$$V(Y) = b''(\theta)a(\phi) = \frac{b''(\theta)\phi}{w}$$

A variância é dada pelo produto de 2 termos, sendo o primeiro termo, $b''(\theta)$, designado por função de variância e é representado por $V(\mu)$.

De seguida apresentam-se exemplos de distribuições que pertencem à família exponencial e que serão utilizadas no decorrer desta tese.

2.1.1 Exemplos

Nesta secção, são apresentados alguns casos particulares de distribuições pertencentes à família exponencial assim como algumas das suas propriedades. Todos os exemplos apresentados pressupõem que o peso w é igual a 1.

Distribuição Normal

Seja Y uma variável aleatória seguindo uma distribuição normal de valor médio μ e variância σ^2. Então a f.d.p. de Y é dada por:

$$\begin{aligned} f(y|\mu,\sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2\sigma^2}} \\ &= exp\left\{\frac{1}{\sigma^2}\left(y\mu - \frac{\mu^2}{2}\right) - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + ln(2\pi\sigma^2\right)\right\} \end{aligned} \tag{2.3}$$

para $y \in \mathbb{R}$. Esta função é do tipo (2.1) com:

$$\begin{aligned}
\theta &= \mu, \\
a(\phi) &= \frac{\phi}{w}, \; com \; \phi = \sigma^2 \; e \; w = 1 \\
b(\theta) &= \frac{\theta^2}{2} = \frac{\mu^2}{2} \\
c(y,\phi) &= -\frac{1}{2}\left(\frac{y^2}{\sigma^2} + ln(2\pi\sigma^2)\right) \\
b'(\theta) &= \theta = \mu \\
b''(\theta) &= 1
\end{aligned}$$

Daqui se podem então retirar as já conhecidas expressões para a média e variância de Y:

$$E(Y) = b'(\theta) = \mu$$

$$V(Y) = b''(\theta)a(\phi) = \sigma^2$$

Para a distribuição normal, o parâmetro de dispersão é σ^2 e o parâmetro canónico é μ.

Distribuição de Poisson

Seja Y uma variável aleatória seguindo uma distribuição de Poisson de valor médio μ. A função de probabilidade de Y é dada por:

$$\begin{aligned}
f(y|\mu) &= \mu^y \frac{e^{-\mu}}{y!} \\
&= exp\{ylog(\mu) - \mu - log(y!)\} \qquad (2.4)
\end{aligned}$$

Esta função é do tipo (2.1) com:

$$\begin{aligned}
\theta &= log(\mu), \\
a(\phi) &= 1 \\
b(\theta) &= e^\theta = \mu \\
c(y,\phi) &= -log(y!) \\
b'(\theta) &= e^\theta = \mu = b''(\theta)
\end{aligned}$$

Assim, a média e a variância de Y são dadas por:

$$E(Y) = b'(\theta) = \mu$$

$$V(Y) = b''(\theta)a(\phi) = \mu,$$

como já é sabido para a distribuição de Poisson. Neste caso, o parâmetro de dispersão ϕ é igual a 1 e o parâmetro canónico θ é $log(\mu)$.

Distribuição Binomial

Seja Y uma variável aleatória tal que $Y \sim B(n, \pi)$, onde n é o número de experiências de Bernoulli de um certo evento e π é a probabilidade de sucesso do acontecimento em cada experiência. A função distribuição de probabilidade de Y é dada por:

$$\begin{aligned} f(y|\pi) &= \binom{n}{y}\pi^y(1-\pi)^{n-y} \\ &= exp\left\{ylog\left(\frac{\pi}{1-\pi}\right) + nlog(1-\pi) + log\left(\binom{n}{y}\right)\right\} \end{aligned} \tag{2.5}$$

Assim,

$$\begin{aligned} \theta &= \log\left(\frac{\pi}{1-\pi}\right) \\ b(\theta) &= nlog(1+e^\theta) \\ a(\phi) &= 1 \\ c(y,\phi) &= log\left(\binom{n}{y}\right) \\ b'(\theta) &= n\frac{e^\theta}{1+e^\theta} = n\pi \\ b''(\theta) &= n\frac{1}{1+e^\theta} = n(1-\pi) \end{aligned}$$

Portanto, a média e a variância de Y são dadas por:

$$E(Y) = b'(\theta) = n\pi$$

$$V(Y) = b''(\theta)a(\phi) = n\pi(1-\pi)$$

Para a distribuição binomial, ϕ é igual a 1 e o parâmetro canónico θ é $log\left(\frac{\pi}{1-\pi}\right)$.

Distribuição Binomial Negativa

Segundo **?**), a distribuição Binomial Negativa pode ser obtida através de 13 formas distintas, sendo que diversos estatísticos acreditam que este número é maior, o que faz com que a parametrização da distribuição binomial negativa usada por uma pessoa possa ser diferente da parametrização usada por outra pessoa. As duas formas mais conhecidas e largamente utilizadas são a distribuição binomial negativa tradicional, apelidada por NB2 (**?**)), derivada de uma distribuição de misturas de Poisson e Gamma, e a distribuição binomial negativa canónica, designada por NB-C.

Seja Y uma variável aleatória seguindo uma distribuição Binomial Negativa de parâmetros k e p, ou seja, $Y \sim BN(k,p)$. Y representa o número de insucessos anteriores a k sucessos, num conjunto de acontecimentos independentes e com a mesma probabilidade p de sucesso. Assim, a função de probabilidade de Y é dada por:

$$\begin{aligned} f(y|p,k) &= \binom{y+k-1}{k-1} p^k (1-p)^y \\ &= exp\left[ylog(1-p) + klog(p) + log\left(\binom{y+k-1}{k-1}\right)\right] \end{aligned} \tag{2.6}$$

Neste caso, tem-se a representação da distribuição na forma canónica. Verifica-se então que:

$$\theta = log(1-p)$$

$$b(\theta) = -klog(p)$$

$$a(\phi) = 1$$

$$c(y,\phi) = log\left(\binom{y+k-1}{k-1}\right)$$

Assim, a média e a variância de Y podem ser expressas por:

$$E(Y) = b'(\theta) = \frac{k(1-p)}{p}$$

$$V(Y) = b''(\theta)a(\phi) = \frac{k(1-p)}{p^2},$$

Logo, o parâmetro de dispersão ϕ é igual a 1 e o parâmetro canónico é dado por $log(1-p)$.

Em relação à distribuição Binomial Negativa NB2, esta pode ser derivada a partir de uma distribuição de Poisson mista. A função de distribuição de Poisson de média μ é dada por:

$$f(y|\mu) = \frac{e^{-\mu}\mu^y}{y!} \tag{2.7}$$

Considere-se agora que a média também é uma variável aleatória, ou seja, diferentes indivíduos de uma população podem estar associados a diferentes valores de μ (**?**)). Então tem-se $Y \sim P(\mu V)$, sendo V uma variável aleatória que representa a heterogeneidade não observada, satisfazendo $E(V) = 1$ (**?**)).

?) mostrou que:

$$E(Y) = E(\mu V) = \mu E(V) = \mu$$

$$V(Y) = E(\mu V) + V(\mu V) = \mu + \mu^2 V(V) = \mu + \frac{\mu^2}{k}$$

A distribuição não condicionada de Y é denominada por distribuição de Poisson mista e a sua função de probabilidade é dada por:

$$f(y|\mu, v) = \int_0^{+\infty} \frac{e^{-\mu v}(\mu v)^y}{y!} g(v) \partial v,$$

onde g é a função densidade de probabilidade de V.

Substituindo g na expressão anterior pela expressão da função densidade de probabilidade de uma distribuição $\Gamma\left(\frac{1}{\alpha}, \frac{1}{\alpha}\right)$ e aplicando manipulações algébricas (**?**)), obtém-se a expressão da função de probabilidade correspondente à distribuição Binomial Negativa NB2 de parâmetros μ e α:

$$\begin{aligned} f(y_i|\mu, \alpha) &= \frac{\Gamma(y_i + \frac{1}{\alpha})}{\Gamma(y_i + 1)\Gamma(\frac{1}{\alpha})} \left(\frac{1}{1 + \alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^{y_i} \\ &= \binom{y_i + \frac{1}{\alpha} - 1}{\frac{1}{\alpha} - 1} \left(\frac{1}{1 + \alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^{y_i} \end{aligned} \tag{2.8}$$

Através das igualdades:

$$k = \frac{1}{\alpha} \text{ e } p = \frac{1}{1 + \alpha\mu},$$

verifica-se que (2.8) pode ser reescrita como a função de probabilidade apresentada para a distribuição Binomial Negativa NB-C em (2.6).

Para a distribuição Binomial Negativa NB2 tem-se:

$$\theta = log(\tfrac{\alpha\mu}{1+\alpha\mu})$$

$$b(\theta) = \tfrac{1}{\alpha} log(1+\alpha\mu)$$

$$a(\phi) = 1$$

$$c(y,\phi) = log\left(\binom{y+\frac{1}{\alpha}-1}{\frac{1}{\alpha}-1} \right)$$

Assim, a média e a variância de Y podem ser expressas por:

$$E(Y) = b'(\theta) = \mu$$

$$V(Y) = b''(\theta)a(\phi) = \mu + \alpha\mu^2,$$

2.2 Componentes dos GLM

Todos os modelos lineares generalizados têm 3 componentes: a componente aleatória, que identifica a variável resposta Y e assume que Y tem uma função distribuição de probabilidade que faz parte da família exponencial; a componente sistemática, que especifica as variáveis explicativas do modelo, tomando uma combinação linear dessas mesmas variáveis; a função de ligação, que faz a correspondência entre as componentes aleatória e sistemática.

2.2.1 Componente Aleatória

Seja $\mathbf{X} = (X_1, \ldots, X_p)$ o vetor das variáveis explicativas de interesse. Para uma amostra aleatória de tamanho n, o vetor $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip})$ corresponde à i-ésima observação. A componente aleatória refere que as variáveis Y_i são condicionalmente independentes com distribuição pertencente à família exponencial; em particular:

$$E(Y_i|\mathbf{x}_i) = b'(\theta_i) = \mu_i, \text{ para } i = 1, \ldots, n$$

2.2.2 Componente Sistemática

A partir do conjunto das variáveis explicativas $X_1, \ldots, X_p$, considera-se uma estrutura linear η dada por:

$$\eta = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{2.9}$$

onde $\beta_i, i = 0, \ldots, p$ são os coeficientes do modelo.

Equivalentemente:

$$\eta_i = \tilde{x}_i^T \beta, i = 1, \ldots, n,$$

com $\tilde{x}_i = (1, x_i^T)^T$ e β o vetor dos parâmetros de regressão.

Igualmente, e em notação matricial:

$$\eta = X\beta,$$

onde X é a matriz de especificação com dimensão $n \times (p+1)$ que consiste de uma coluna (a primeira) populada apenas de 1's, sendo as restantes colunas respeitantes aos vetores coluna $\mathbf{x}_i$ e β é o vetor de dimensão $p+1$ dos parâmetros de regressão.

2.2.3 Função de ligação

A função de ligação g traduz a relação entre a média da resposta e o preditor linear η descrito na componente sistemática. A função de ligação é monótona e diferenciável.

$$\eta = g(\mu)$$

Quando a função de ligação torna o preditor linear igual ao parâmetro canónico θ, ou seja:

$$\eta = g(\mu) = \theta,$$

diz-se que a função g é a função de ligação canónica. Por exemplo, para a distribuição normal, a função de ligação canónica é a função identidade; para a distribuição binomial, trata-se da função *logit*; para a distribuição de Poisson, é a função log (**?**)). A função de ligação canónica tem a vantagem de apresentar toda a informação sobre os coeficientes de regressão numa função dos dados da mesma dimensão que o número de coeficientes de regressão.

2.3 Inferência

Após a formulação do modelo, supondo que este foi especificado correctamente, existe a necessidade de realizar inferências sobre esse mesmo modelo. Nos modelos lineares generalizados, este passo é feito com base na verosimilhança, não só para estimar os parâmetros de regressão mas também para efectuar testes de hipóteses sobre os mesmos e para avaliar a qualidade de ajustamento. No entanto, nem sempre os modelos estão

realisticamente bem formulados. Por exemplo, quando os dados apresentam sobredispersão, é preciso modificar a variância, como por exemplo, introduzindo um parâmetro de sobredispersão. Desse modo, o modelo não está completamente especificado já que não existe uma distribuição que tenha aqueles valor médio e variância. Esse obstáculo é ultrapassado usando modelos de quasi-verosimilhança.

2.3.1 Verosimilhança

Segundo **?**), a função de verosimilhança associada a uma amostra aleatória $Y_1, \ldots, Y_n$, é dada por:

$$\begin{aligned} L(\beta; y_i) &= \prod_{i=1}^{n} f(y_i|\theta_i, \phi, w_i) \\ &= \prod_{i=1}^{n} exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi, w_i)\right\} \\ &= exp\left\{\sum_{i=1}^{n}\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi, w_i)\right\} \end{aligned} \tag{2.10}$$

A função de verosimilhança é dada como função de β. De forma a simplificar os cálculos, tome-se o logaritmo da verosimilhança (log-verosimilhança):

$$\begin{aligned} lnL(\beta) = l(\beta) &= \sum_{i=1}^{n}\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi, w_i) \\ &= \sum_{i=1}^{n} l_i(\beta), \end{aligned} \tag{2.11}$$

sendo $l_i(\beta) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi, w_i)$ a contribuição da observação y_i para a verosimilhança.

Os estimadores de máxima verosimilhança para os parâmetros de regressão obtêm-se resolvendo o seguinte sistema de equações:

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n}\frac{\partial l_i(\beta)}{\partial \beta_j} = 0, j = 0, 1, \ldots, p$$

Pela regra da cadeia (**?**);**?**)), podemos escrever as equações anteriores na seguinte forma:

$$\frac{\partial l_i(\beta)}{\partial \beta_j} = \frac{\partial l_i(\theta_i)}{\partial \theta_i}\frac{\partial \theta_i(\mu_i)}{\partial \mu_i}\frac{\partial \mu_i(\eta_i)}{\partial \eta_i}\frac{\partial \eta_i(\beta)}{\partial \beta_j}, j = 0, \ldots, p \tag{2.12}$$

Da definição da função de log-verosimilhança e da igualdade $b'(\theta_i) = \mu_i$, tem-se:

$$\frac{\partial l_i(\theta_i)}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)} \tag{2.13}$$

Pela definição da variância de Y_i, $V(Y_i) = a(\phi)b''(\theta_i)$:

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{V(Y_i)}{a(\phi)} \tag{2.14}$$

A partir de $\eta_i = \tilde{x}_i^T \beta$, obtém-se:

$$\frac{\partial \eta_i(\beta)}{\partial \beta_j} = \tilde{x}_{ij} \tag{2.15}$$

Assim, através de (2.13), (2.14) e (2.15), a equação (2.12) pode ser reescrita da seguinte forma:

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{V(y_i)} \frac{\partial \mu_i(\eta_i)}{\partial \eta_i} \tilde{x}_{ij}, j = 1, \ldots, p \tag{2.16}$$

logo as equações de verosimilhança para β são:

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{V(y_i)} \frac{\partial \mu_i(\eta_i)}{\partial \eta_i} \tilde{x}_{ij} = 0, j = 1, \ldots, p. \tag{2.17}$$

Derivando a função log-verosimilhança em ordem a β, obtém-se a função *score*:

$$s(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} s_i(\beta) \tag{2.18}$$

onde $s_i(\beta)$ designa o vetor de componentes $\frac{\partial l_i(\beta)}{\partial \beta_j}$.
A matriz de covariância da função *score* é denominada por matriz de informação de Fisher:

$$cov(s(\beta)) = I(\beta) = E\left[-\frac{\partial s(\beta)}{\partial \beta}\right],$$

correspondendo ao simétrico do valor esperado da matriz Hessiana da função log-verosimilhança (**?**)):

$$\begin{aligned} I(\beta) = -E\left(\frac{\partial^2 l_i}{\partial\beta_j\partial\beta_k}\right) &= E\left(\frac{\partial l_i}{\partial\beta_j}\frac{\partial l_i}{\partial\beta_k}\right) \\ &= E\left[\left(\frac{Y_i-\mu_i}{V(Y_i)}\frac{\partial\mu_i(\eta_i)}{\partial\eta_i}\tilde{x}_{ij}\right)\left(\frac{Y_i-\mu_i}{V(Y_i)}\frac{\partial\mu_i(\eta_i)}{\partial\eta_i}\tilde{x}_{ik}\right)\right] \\ &= \frac{\tilde{x}_{ij}\tilde{x}_{ik}}{V(Y_i)}\left(\frac{\partial\mu_i(\eta_i)}{\partial\eta_i}\right)^2, \end{aligned}$$

sendo este o valor do elemento de ordem (j,k) da matriz de informação de Fisher.

2.3.2 Estimação dos Parâmetros

Resolvendo as equações de verosimilhança apresentadas em (2.17), obtêm-se os estimadores de máxima verosimilhança (EMV) de β. No entanto, a solução não corresponde necessariamente a um máximo global da função $l(\beta)$. Em muitos modelos, a função log-verosimilhança é côncava logo o máximo local e o máximo global coincidem. A resolução das equações de verosimilhança passa tipicamente pelo recurso a métodos iterativos, sendo os mais comuns o método iterativo dos mínimos quadrados reponderados e o método de Newton-Raphson. **?**) mostraram que os 2 métodos são assimptoticamente equivalentes já que, à medida que o tamanho amostral cresce, as propriedades dos estimadores tornam-se idênticas, em termos de distribuição.

2.3.2.1 Método iterativo dos mínimos quadrados reponderados

O método iterativo dos mínimos quadrados reponderados para resolver as equações de verosimilhança tem como fundação o método de *scores* de Fisher.

Seja $\hat{\beta}^{(0)}$ uma estimativa inicial para β. O método de *scores* de Fisher calcula as seguintes iterações:

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + [I(\hat{\beta}^{(k)})]^{-1} s(\hat{\beta}^{(k)}), \tag{2.19}$$

supondo que existe a inversa da matriz de informação de Fisher.

A diferença entre este método e o método de Newton-Raphson é que o primeiro utiliza a matriz de informação de Fisher enquanto o segundo utiliza a matriz Hessiana (**?**); **?**)). A matriz de informação de Fisher é, geralmente, mais fácil de calcular, acrescendo a

vantagem de ser sempre uma matriz semi-definida positiva.

De seguida, calcula-se o preditor linear $g(\hat{\mu}_i) = (1, x_i^T)^T \beta = \hat{\eta}_i$ para posteriormente calcular os valores ajustados $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$. Definindo uma variável dependente de trabalho T, esta toma o seguinte valor para a i-ésima observação na k-ésima iteração:

$$\begin{aligned} t_i^{(k)} &= \hat{\eta}_i^{(k)} + (y_i - \hat{\mu}_i^{(k)}) \frac{\partial \eta_i^{(k)}}{\partial \mu_i^{(k)}} \\ &= g(\hat{\mu}_i^{(k)}) + (y_i - \hat{\mu}_i^{(k)}) g'(\hat{\mu}_i^{(k)}) \end{aligned} \tag{2.20}$$

No passo seguinte, calculam-se os pesos iterativos:

$$\varpi_i^{(k)} = \left(V(y_i) \left(\frac{\partial \eta_i^{(k)}}{\partial \mu_i^{(k)}} \right)^2 \right)^{-1},$$

que, no caso da família exponencial, com a igualdade $a_i(\phi) = \frac{\phi}{w_i}$, podem ser reescritos da seguinte forma:

$$\varpi_i^{(k)} = w_i^{(k)} \left(\phi b''(\theta_i) \left(\frac{\partial \eta_i^{(k)}}{\partial \mu_i^{(k)}} \right)^2 \right)^{-1}$$

Pode-se mostrar, através de manipulações algébricas, que este peso é inversamente proporcional à variância da variável t_i, ou seja, $\varpi_i = \frac{1}{V(t_i)}$.

Assim, o estimador dos mínimos quadrados reponderados é dado por:

$$\hat{\beta}^{(k+1)} = (X^T W^{(k)} X)^{-1} X^T W^{(k)} T^k, \tag{2.21}$$

onde X representa a matriz do modelo em questão, W é a matriz diagonal dos pesos cujas entradas são dadas pela expressão de ϖ_i e T é o vetor com entradas t_i.

O método vai repetindo o procedimento tendo em conta um critério de paragem, como por exemplo (**?**)):

$$\frac{||\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)}||}{||\hat{\beta}^{(k+1)}||} \leq \epsilon,$$

para um valor de $\epsilon > 0$ previamente definido.

2.3.2.2 Estimação do parâmetro de dispersão

O parâmetro de dispersão ϕ pode ser estimado pelo método da máxima verosimilhança tal como os parâmetros de regressão, sendo no entanto preferível utilizar um método menos complexo baseado na distribuição empírica da estatística de Pearson generalizada. Assim, temos o seguinte estimador (**?**)):

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{w_i (y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}, \tag{2.22}$$

sendo $\hat{\phi}$ um estimador consistente de ϕ. O segundo factor na expressão de $\hat{\phi}$ é conhecido como a estatística de Pearson generalizada, também utilizada para avaliar a qualidade do ajustamento do modelo.

2.3.3 Testes de Hipóteses

Os testes de hipóteses são um instrumento estatístico que permitem ao investigador inferir sobre um determinado aspecto na população (parâmetro populacional, distribuição, entre outros) através da análise de uma amostra dessa mesma população. Em relação aos modelos lineares generalizados, estes são maioritariamente utilizados para testar a significância estatística dos parâmetros, a significância estatística dos parâmetros individuais de regressão do modelo ou então comparar a qualidade de ajustamento entre dois modelos.

2.3.3.1 Teste de Wald

O teste de Wald corresponde a testar um submodelo que contém todas as variáveis explicativas do modelo original à excepção de uma. A estatísica de Wald é baseada na normalidade assimtótica do EMV de β. Seja a hipótese nula:

$$H_0 : \beta_j = 0, j = 0, \ldots, p$$

que indica que o parâmetro de regressão β_j não deve constar do modelo de regressão. Neste caso, para amostras de tamanho grande, a estatística de teste é dada por:

$$W_j = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \overset{\text{a}}{\sim} N(0,1)$$

onde $se(\hat{\beta}_j) = \sigma \sqrt{(X^T X)^{-1}_{jj}}$ é o erro-padrão do coeficiente de regressão β_j. Valores "elevados", em valor absoluto, da estatística W_j correspondem à rejeição da hipótese nula H_0.

2.3.3.2 Teste da Razão de Verosimilhanças

O teste da razão de verosimilhanças permite comparar o ajustamento de dois modelos encaixados, ou seja, 2 modelos ω_1 e ω_2 tais que $\omega_1 \in \omega_2$, com p_1 e p_2 parâmetros, respetivamente, satisfazendo $p_1 < p_2$. Existem dois modelos que merecem especial atenção:

- o modelo nulo: não existe qualquer relação entre as variáveis explicativas e a variável resposta, contendo apenas um único parâmetro, μ, comum a todas as observações.
- o modelo saturado: contem n parâmetros, um por cada observação. Os valores ajustados por este modelo são iguais aos valores observados logo o modelo saturado explica completamente os dados, sendo pouco útil já que a informação obtida é igual à dos dados.

De entre todos os modelos possíveis, o modelo nulo apresenta o menor valor da função de verosimilhança e o modelo saturado apresenta o maior valor da função de verosimilhança.

Suponhamos que queremos comparar um modelo qualquer ω com o modelo saturado Ω. Sejam:

- $\hat{\mu}_i$ o valor ajustado pelo modelo ω para a i-ésima observação
- y_i o valor ajustado pelo modelo Ω para a i-ésima observação
- $\hat{\theta}_i$ os parâmetros canónicos estimados pelo modelo ω
- $\tilde{\theta}_i$ os parâmetros canónicos estimados pelo modelo Ω
- L_ω a função de verosimilhança do modelo ω
- L_Ω a função de verosimilhança do modelo Ω

O critério da razão de verosimilhanças entre os dois modelos é dado por:

$$\begin{aligned} -2log\left(\frac{L_\omega}{L_\Omega}\right) &= 2\sum_{i=1}^{n}\frac{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)}{\phi} \\ &= \frac{D(y,\hat{\mu})}{\phi} \end{aligned} \tag{2.23}$$

A quantidade $D(y,\hat{\mu}) = 2\sum_{i=1}^{n} y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)$ é denominada por desviância, sendo explicitada detalhadamente na secção seguinte.

É possível utilizar a desviância para testar:

H_0 : o modelo ω ajusta-se adequadamente aos dados,

desde que os dados se encontrem agrupados em padrões de covariáveis. Quando as variáveis explicativas são todas categóricas, é possível agregar os dados que apresentam os mesmos valores para todas as covariáveis, dando origem aos padrões de covariáveis. Obtêm-se assim tantas observações quantas combinações dos diferentes níveis das variáveis. Além disso, o número de observações n_i que constituem cada padrão de covariáveis deve ser relativamente grande. Segundo **?**), é razoável assumir-se $n_i > 10$. Quando algum dos n_i's é pequeno, um valor elevado da estatística D não aponta necessariamente para um mau ajustamento do modelo.

Assim, sob H_0, tem-se:

$$D \overset{a}{\sim} \chi^2(K-(p+1)),$$

onde K é o número de padrões de covariáveis diferentes existentes nos dados. Se o número de padrões de covariáveis, K, for aproxidamente igual ao número de observações, não existe convergência na distribuição da estatística, logo os valores-p calculados para D usando a distribuição acima referida são incorrectos.

Se $D > \chi^2_{1-\alpha}(K-(p+1))$, rejeita-se a hipótese nula com nível de significância α.

Suponhamos agora que queremos comparar dois modelos encaixados. O logaritmo da razão de verosimilhanças para os dois modelos é (Gaio, 2012):

$$\begin{aligned} -2log\left(\frac{L_{\omega_1}}{L_{\omega_2}}\right) &= -2log\left(\frac{\frac{L_{\omega_1}}{L_\Omega}}{\frac{L_{\omega_2}}{L_\Omega}}\right) \\ &= -2log\left(\frac{L_{\omega_1}}{L_\Omega}\right) + 2log\left(\frac{L_{\omega_2}}{L_\Omega}\right) \\ &= \frac{D(\omega_1)-D(\omega_2)}{\phi} \end{aligned}$$

Caso o parâmetro de dispersão ϕ seja desconhecido, este deve ser estimado a partir do modelo com o maior número de parâmetros, neste caso, o modelo ω_2.

Sob a hipótese nula:

H_0: os dois modelos têm a mesma qualidade de ajustamento

temos que a estatística de teste é dada pelo quociente anterior, seguindo este a seguinte distribuição:

$$G = \frac{D(\omega_1)-D(\omega_2)}{\phi} \overset{a}{\sim} \chi^2(p_2-p_1)$$

Se $G > \chi^2_{1-\alpha}(p_2-p_1)$, rejeita-se a hipótese nula com nível de significância α.

2.3.4 Qualidade do ajustamento

2.3.4.1 Desviância

Na secção anterior foi introduzida a desviância, medida baseada no critério da razão de verosimilhanças que permite avaliar a qualidade do ajustamento de um certo modelo tendo em conta o modelo saturado.
Como já foi referido, o critério da razão de verosimilhanças é dado por:

$$-2log\left(\frac{L_\omega}{L_\Omega}\right) = \frac{D(y,\hat{\mu})}{\phi},$$

onde

$$\begin{aligned} D(y,\hat{\mu}) &= 2\phi log\left(\frac{L_\omega}{L_\Omega}\right) \\ &= 2\sum_{i=1}^{n} y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i) \end{aligned} \tag{2.24}$$

é designada por desviância. A razão de log-verosimilhanças é denominada de desviância re-escalonada pois representa o quociente entre a desviância e o parâmetro de dispersão ϕ:

Se o parâmetro de dispersão for igual a 1, como por exemplo na distribuição binomial e distribuição de Poisson, a desviância e a desviância re-escalonada tomam o mesmo valor.

A desviância avalia o valor absoluto da diferença entre os valores observados (que coincidem com os valores ajustados pelo modelo saturado, como já foi referido) e os valores ajustados pelo modelo em estudo. Conclui-se então que a desviância é sempre maior ou igual a zero e quanto maior for a discrepância entre o modelo e os dados, maior será o valor da desviância. Para o modelo saturado, o valor da desviância é igual a zero.

2.3.4.2 Estatística χ^2 de Pearson generalizada

A estatística χ^2 de Pearson generalizada é outra medida útil para aferir a discrepância entre 2 modelos e é dada por:

$$X_P^2 = \sum_{i=1}^{n} \frac{\omega_i^2(y_i - \hat{\mu}_i)^2}{\hat{\phi}V(\hat{\mu}_i)}, \tag{2.25}$$

onde $V(\hat{\mu}_i)$ é a função de variância estimada para a distribuição do modelo em causa. No entanto, ao contrário da desviância, não é possível utilizar a diferença entre estatísticas χ^2 de Pearson para comparar modelos encaixados pois não se conhece a distribuição para a diferença das estatísticas.

A estatística de Pearson generalizada também pode ser utilizada num teste de hipóteses para testar o ajustamento do modelo, seguindo uma distribuição assimtótica $\chi^2(K-(p+1))$, onde K é o número de padrões de covariáveis diferentes existentes nos dados e p é o número de parâmetros do modelo.

2.3.5 Resíduos

Os resíduos exprimem a discrepância entre o valor observado y_i e o valor $\hat{\mu}_i$ ajustado pelo modelo. Assim, a análise dos resíduos permite avaliar a qualidade de ajustamento do modelo no que diz respeito à escolha da distribuição, da função de ligação e de termos do preditor linear mas também observações que não são bem explicadas pelo modelo. De seguida, apresentam-se 2 tipos de resíduos, os resíduos de Pearson e os resíduos da desviância.

2.3.5.1 Resíduos de Pearson

O resíduo de Pearson para a i-ésima observação corresponde à contribuição dessa mesma observação para o cálculo da estatística de Pearson generalizada e é dado por:

$$R_i = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{Y}_i)}} = \frac{(y_i - \hat{\mu}_i)\omega_i}{\sqrt{\hat{\phi}V(\hat{\mu}_i)}} \tag{2.26}$$

No entanto, os resíduos devem ser padronizados de forma a ser possível realizar uma análise adequada. De acordo com **?**), o resíduo de Pearson estandardizado é dado por:

$$R_P^* = \frac{(y_i - \hat{\mu}_i)\omega_i}{\sqrt{\hat{\phi}V(\hat{\mu}_i)(1-h_i)}}, \tag{2.27}$$

onde $h_i, i = 1,\ldots,n$ são as *leverages*[1] ou repercussões do modelo ajustado, quantidades que medem o efeito que as observações têm nos seus valores ajustados. Se as *leverages* não forem relativamente elevadas, os gráficos dos resíduos de Pearson e dos resíduos de Pearson estandardizados não apresentarão grande diferença, a menos da escala.

2.3.5.2 Resíduos da desviância

Este tipo de resíduo é baseado na função desviância. A desviância pode ser vista como a soma das contribuições d_i de todas as observações:

[1] O conceito será apresentado em maior detalhe na secção 2.3.7

$$D(y, \hat{\mu}) = \sum_i d_i \tag{2.28}$$

Os resíduos da desviância são dados então por:

$$R_i^D = \delta_i \sqrt{d_i}, \tag{2.29}$$

onde $\delta_i = sinal(y_i - \hat{\mu}_i)$. Tal como nos resíduos de Pearson, também é usual estandardizar os resíduos da desviância:

$$R_i^{*D} = \frac{\delta_i \sqrt{d_i}}{\sqrt{\hat{\phi}(1 - h_i)}} \tag{2.30}$$

$$\tag{2.31}$$

2.3.6 AIC e BIC

O critério de informação de Akaike (AIC) e o critério de informação Bayesiana (BIC) são outros dois critérios que servem de orientação na escolha de um modelo, sendo tipicamente utilizados quando dois modelos não são encaixados. Nessa situação, é preferível usar um critério que meça a quantidade de informação que o modelo recolhe dos dados em vez de um critério que meça o ruído que o modelo não consegue explicar.

A expressão para o AIC é dada por (**?**)):

$$AIC = -2log(LL) + 2p, \tag{2.32}$$

onde LL é o valor do logaritmo da verosimilhança do modelo e p é o número de parâmetros do modelo estimado.

O BIC foi desenvolvido por Gideon Schwarz em 1978 tendo como base uma argumentação bayesiana, sendo muito próximo do AIC mas tendo também em conta o número de observações. A expressao do BIC é (**?**)):

$$BIC = -2log(LL) + 2plog(n), \tag{2.33}$$

onde LL é o valor do logaritmo da verosimilhança do modelo, p é o número de parâmetros do modelo estimado e n o número total de observações.

Quanto menor for o valor dos critérios, mais preferível será o modelo. O critério BIC penaliza mais o número de parâmetros do que o critério AIC já que o número de parâmetros é multiplicado por $log(n)$, que cresce à medida que o número de observações aumenta.

Note-se que nenhum dos critérios avalia se o modelo se ajusta bem aos dados ou não, mas sim se um determinado modelo é melhor que outro.

2.3.7 Leverages

Na regressão múltipla, há observações que influenciam marcadamente a estimação dos parâmetros do modelo apesar de não serem *outliers* nem serem facilmente identificáveis através de análises gráficas dos dados. Estas observações correspondem a pontos com *leverage*[2] alta. A *leverage* de um ponto quantifica a possibilidade de essa observação não só ser um *outlier* para as variáveis explicativas mas também influenciar o seu próprio valor ajustado.

Como exemplo, consideremos o caso do modelo de regressão linear múltipla:

$$\begin{aligned} Y &= X\beta + \epsilon \\ &= \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon, \end{aligned}$$

onde os erros verificam a condição $\epsilon \sim N(0, \sigma^2 Id)$ (onde Id é a matriz identidade de dimensão n) de serem independentes e normalmente identicamente distribuídos. O estimador de máxima verosimilhança para β é

$$\hat{\beta} = (X^T X)^{-1} X^t y,$$

Em particular, o vetor de valores ajustados é

$$\hat{y} = X\hat{\beta} = Hy \tag{2.34}$$

onde $H = X(X^T X)^{-1} X^T$ é designada por matriz-chapéu. Geometricamente, esta matriz representa a projeção ortogonal de y sobre o espaço gerado por (1,X).
A equação pode ser vista da seguinte forma:

$$\hat{y}_i = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j \tag{2.35}$$

Pode-se concluir que h_{ii} traduz a contribuição da observação y_i sobre o seu próprio valor ajustado $\hat{y}_i$. $h_{ii}, i = 1, \ldots, n$ correspondem às entradas da diagonal de H. Devido à existência de várias regras práticas para determinar se uma observação tem uma *leverage* elevada, adopta-se o valor de corte dado por $\frac{2(p+1)}{n}$, onde p é o número de parâmetros do modelo (excluindo a constante) e n é o número de observações do modelo.

[2]Em português, traduz-se para repercussão ou alavanca

2.3.8 Distância de Cook

Uma observação diz-se influente se tem uma grande influência sobre os parâmetros de regressão estimados. No entanto, uma observação deste tipo pode ou não ser um outlier e pode ou não ter uma *leverage* grande mas tenderá a verificar pelo menos uma destas propriedades.
A estatística usada comummente para avaliar a influência da observação i é a distância de Cook:

$$C_i^2 = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_i)^2}{(p+1)\bar{\sigma}^2}, \tag{2.36}$$

onde $\bar{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(p+1)}$ é um estimador não enviesado da variância de Y.
De acordo com **?**), uma regra empírica para detectar pontos influentes é considerar como tais os pontos cuja distância de Cook é superior a $\frac{4}{n}$, onde n é o número de observações. Estes pontos devem ser analisados e o modelo deve ser ajustado sem estes pontos de forma a comparar as estimativas dos parâmetros obtidas.

2.4 Exemplos de Modelos Lineares Generalizados

2.4.1 Regressão Logística

A regressão logística é um dos vários tipos de modelos lineares generalizados, sendo usada quando a variável resposta é binária.

Seja $Y_1, \ldots, Y_n$ uma amostra aleatória de uma variável aleatória binária Y, codificada por 0 e 1. Dado um vector de variáveis explicativas $X = (X_1, \ldots, X_p)$ e uma observação $x_i = (x_{i1}, \ldots, x_{ip})$ do indivíduo i, assume-se que:

$$Y|X = x_i \sim B(1, \pi_i(x_i)),$$

onde $\pi_i = P(Y = 1|X = x_i)$ é a probabilidade de sucesso para Y, dada a observação x_i. Tendo em conta as propriedades da distribuição binomial, obtêm-se as seguintes fórmulas para a média e para a variância de Y dado $X = x_i$:

$$E(Y|X = x_i) = \pi_i$$
$$V(Y|X = x_i) = \pi_i(1 - \pi_i)$$

No caso em que todas as variáveis explicativas são categóricas, podemos agrupar os dados em padrões de covariáveis. Suponhamos que existem K padrões de covariáveis. Seja n_i o número de observações no padrão i, Y_i a variável aleatória que representa a frequência absoluta de sucessos no grupo i e $\bar{Y}_i$ a variável aleatória que representa a frequência relativa de sucessos no grupo i. Se as n_i observações em cada grupo forem independentes e tiverem a mesma probabilidade π_i de sucesso, então:

$$Y_i \sim B(n_i, \pi_i),$$

logo tem-se para as frequências relativas:

$$\bar{Y}_i \sim B(n_i, \pi_i)/n_i$$

Neste caso, as expressões para a média e para a variância são dadas por:

$$E(\bar{Y}_i) = \pi_i$$

$$V(\bar{Y}_i) = \frac{V(Y_i)}{n_i^2} = \frac{\pi_i(1-\pi_i)}{n_i}$$

Note-se que, como a média e a variância dependem da probabilidade de sucesso π_i, se esta for alterada, também o serão os parâmetros. Daqui se conclui, em particular, que não podemos assumir um modelo com variância constante (**?**)).

Voltando ao objectivo da regressão logística, deparamo-nos com uma situação inóspita se procedermos com uma modelação linear da média de $Y|X = x_i$, π_i:

$$\pi_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}.$$

π_i varia apenas entre 0 e 1 mas o termo do lado direito varia entre $-\infty$ e $+\infty$. Assim, é necessário escolher uma outra transformação, sendo a mais usual, a transformação *logit*.

As probabilidades π_i variam entre 0 e 1, logo o *odds*:

$$odds_i = \frac{\pi_i}{1-\pi_i}$$

varia entre 0 e $+\infty$. Aplicando a transformação logarítmica, tem-se:

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right),$$

que varia entre $-\infty$ e $+\infty$. *Logits* negativos representam probabilidades inferiores a $\frac{1}{2}$ e *logits* positivos representam probabilidades superiores a $\frac{1}{2}$.

Portanto, uma possível função de ligação é:

$$g(\pi) = logit(\pi),$$

Para a regressão logística, esta é a função de ligação canónica pois $log\left(\frac{\pi}{1-\pi}\right)$ é o parâmetro canónico da distribuição binomial.

Invertendo a função logit, obtém-se:

$$\begin{aligned} \pi(x) &= P(Y=1|X_1=x_1,\ldots,X_p=x_p) \\ &= \frac{e^{\beta_0+\beta_1x_1+\ldots+\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+\ldots+\beta_px_p}} \\ &= \frac{1}{1+e^{-(\beta_0+\beta_1x_1+\ldots+\beta_px_p)}} \end{aligned} \tag{2.37}$$

Assim, o modelo logit pode ser descrito por:

$$Y|X=x_i \sim B(1,\pi_i)$$

e

$$logit(\pi_i) = \beta_0 + \beta_1x_{i1} + \ldots + \beta_px_{ip}.$$

As probabilidades estimadas $\hat{\pi}_i$ também são designadas por $\hat{\mu}_i$, correspondendo aos valores ajustados. Para uma dada observação i, esse valor é uma estimativa da média de observações da população em causa com as características do indivíduo i que apresentam evidência de sucesso para Y.

Interpretação dos parâmetros do modelo

O *odds* é dado pelo quociente entre a probabilidade de sucesso e a probabilidade de insucesso de um certo acontecimento.

Um *odds ratio*, OR, corresponde ao quociente de dois *odds*, sendo calculado para dois indivíduos ou para dois grupos de indivíduos. Por exemplo, se Y for uma variável binária indicatriz de doença e E uma variável indicatriz de exposição a um factor de risco, o OR da exposição para a doença é o quociente entre o *odds* para a doença nos indivíduos expostos e o *odds* para a doença nos indivíduos não expostos:

$$OR = \frac{\frac{P(Y=1|E)}{P(Y=0|E)}}{\frac{P(Y=1|\bar{E})}{P(Y=0|\bar{E})}}.$$

Considere-se então que as variáveis explicativas $X_1,\ldots,X_p$ são variáveis contínuas ou categóricas e o modelo não tem interações.

No caso em que X_j toma valores contínuos, aumentando a coordenada j de $x=(x_1,\ldots,x_p)$ em c unidades, tem-se:

$$\begin{aligned} x &= (x_1,\ldots,x_j,\ldots,x_p) \\ \mu(x_{+c}) &= (x_1,\ldots,x_j+c,\ldots,x_p), \end{aligned} \tag{2.38}$$

logo, recorrendo à definição de *odds ratio*:

$$\begin{aligned} log(OR(x_{+c},x)) &= log\left(\frac{odds(x_{+c})}{odds(x)}\right) \\ &= log\left(\frac{\frac{\pi(x_{+c})}{1-\pi(x_{+c})}}{\frac{\pi(x)}{1-\pi(x)}}\right) \\ &= logit(x_{+c}) - logit(x) \\ &= (x_j+c)\beta_j - x_j\beta_j \\ &= c\beta_j. \end{aligned} \tag{2.39}$$

Assim, $c\beta_j$ representa a alteração provocada no *logit* da probabilidade por uma variação de c unidades na variável X_j, mantendo as restantes variáveis constantes. Em termos do *odds ratio*, como $OR(x_{+c},x) = e^{c\beta_j}$, mantendo as restantes variáveis explicativas e aumentando em c unidades a variável X_j, o *odds* para o sucesso varia em $e^{c\beta_j}$ vezes.

No caso em que as variáveis $X_1,\ldots,X_p$ são contínuas, dadas duas observações, $x_i = (x_{i1},\ldots,x_{ip})$ e $x_j = (x_{ji},\ldots,x_{jp})$, tem-se:

$$\begin{aligned} log(OR(x_i,x_j)) &= logit(x_i) - logit(x_j) \\ &= (x_i - x_j)^T\beta. \end{aligned} \tag{2.40}$$

Assim, $OR(x_i,x_j) = e^{\sum_{k=1}^{p}(x_{ik}-x_{jk})^T\beta_k}$ representa o *odds ratio* para o sucesso quando se passa das características x_i para x_j.

Quando a variável X_j é dicotómica, assumindo, sem perda de generalidade que os níveis da variável estão codificados em 0 e 1, tem-se então:

$$\begin{aligned} log(OR(X_j=1,X_j=0)) &= log\left(\frac{\frac{\pi(X_j=1)}{1-\pi(X_j=1)}}{\frac{\pi(X_j=0)}{1-\pi(X_j=0)}}\right) \\ &= logit(X_j=1) - logit(X_j=0) \\ &= \beta_0 + \ldots + \beta_{j-1}X_{j-1} + \beta_j + \beta_{j+1}X_{j+1} + \ldots + \beta_pX_p \\ &- \beta_0 - \ldots - \beta_{j-1}X_{j-1} - \beta_{j+1}X_{j+1} - \ldots - \beta_pX_p \\ &= \beta_j. \end{aligned} \tag{2.41}$$

Assim, $OR(X_j = 1, X_j = 0) = e^{\beta_j}$ significa que é mais (ou menos) provável e^{β_j} vezes que o sucesso ocorra nos indivíduos com $X_j = 1$ do que nos indivíduos com $X_j = 0$, mantendo as restantes variáveis explicativas constantes.

Quando X_j é politómica, contendo K categorias, a variável é representada por $K - 1$ variáveis indicatrizes sendo uma das categorias (normalmente a primeira) a categoria de referência. De seguida, estimam-se os *odds ratio* para cada uma das classes $1, 2, \ldots, K - 1$ usando a classe 0 como grupo de referência.

A constante β_0 corresponde ao log-*odds* de um indivíduos quando todas as variáveis explicativas tomam o valor nulo, ou seja, e^{β_0} representa o *odds* para o sucesso na ausência de variáveis explicativas.

Quando estamos perante uma situação de regressão simples, ou seja, existe apenas uma variável explicativas, o OR obtido designa-se por OR bruto. Se se tratar de uma regressao multivariada, o OR obtido para cada variável explicativa designa-se por OR ajustado.

Interação

Quando a associação entre uma das variáveis explicativas e a resposta é diferente conforme os valores de outra variável explicativa, estamos perante uma interação entre as respectivas variáveis explicativas.

Neste caso, a metodologia a seguir para a estimação de OR é a seguinte:

- Calcular as expressões para o logit da resposta para cada um dos níveis do factor de risco, isoladamente
- Calcular a diferença entre as expressões anteriores
- Tomar a exponencial do valor da diferença calculada, substituindo os coefientes em causa pelas suas estimativas.

Se considerarmos um modelo com um factor de risco F binário, uma variável explicativa contínua X e a sua interação $F \times X$:

$$logit(\pi(F, X)) = \beta_0 + \beta_1 F + \beta_2 X + \beta_3 (F \times X),$$

o método é:

- $logit(\pi(F = 1, X)) = \beta_0 + \beta_1 + \beta_2 X + \beta_3 X; logit(\pi(F = 0, X)) = \beta_0 + \beta_2 X$
- $logit(\pi(F = 1, X)) - logit(\pi(F = 0, X)) = \beta_1 + \beta_3 X$
- $\hat{OR}(F = 1, F = 0, X = x) = e^{\hat{\beta}_1 + \hat{\beta}_3 x}$

No caso de F ser uma variável politómica, os cálculos são análogos, levando à seguinte estimação do OR para a categoria j:

$$\hat{OR}(F = f_j, F = f_0, X = x) = e^{\hat{\beta}_1(f_j - f_0) + \hat{\beta}_3 x(f_j - f_0)}$$

Intervalos de confiança para o OR

De acordo com **?**), o intervalo com $100(1-\alpha)\%$ de confiança para o odds ratio $OR(x_{+c}, x)$ é:

$$\left(e^{\beta_j - N_{1-\frac{\alpha}{2}} se(\hat{\beta}_j)}, e^{\beta_j + N_{1-\frac{\alpha}{2}} se(\hat{\beta}_j)}\right),$$

onde $N_{1-\frac{\alpha}{2}}$ representa o $1-\frac{\alpha}{2}$-quantil da distribuição normal *standard*. A estimativa obtida para o OR será estatisticamente significativa sempre que o correspondente intervalo de confiança não contiver o valor 1.

Estimação pelo método da máxima verosimilhança

Dado que $Y|X = x_i \sim B(1, \pi_i)$, a função de probabilidade de Y condicionada por $X = x_1$ é:

$$f(y|x_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}.$$

Como as observações $y_1, \ldots, y_n$ são independentes, a função de verosimilhança é dada por:

$$L(\beta; y, x) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i},$$

e o seu logaritmo é dado por:

$$\begin{aligned} LL(\beta; y, x) &= log(L(\beta; y, x)) \\ &= \sum_{i=1}^{n}(y_i log(\pi_i) + (1-y_i) log(1-\pi_i)) \\ &= \sum_{i=1}^{n}\left(log(1-\pi_i) - y_i log\left(\frac{\pi_i}{1-\pi_i}\right)\right) \end{aligned} \tag{2.42}$$

Como $log\left(\frac{\pi_i}{1-\pi_i}\right) = X^T\beta$, a equação anterior passa a ser:

$$\begin{aligned} LL(\beta; y, x) &= \sum_{i=1}^{n}\left(-ln(1 + e^{X_i^T\beta} + y_i X_i^T\beta)\right) \\ &= \sum_{i=1}^{n}\sum_{k=0}^{p} y_i x_{k,i}\beta_k - \sum_{i=1}^{n} ln\left(1 + e^{\sum_{k=0}^{p} x_{k,i}\beta_k}\right) \end{aligned} \tag{2.43}$$

Nos modelos lineares generalizados com termo constante e função de ligação canónica, excepto o modelo com função de ligação probit, a soma dos valores observados é igual à soma dos valores previstos pelo modelo. Assim:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}_i$$

Qualidade do ajustamento

A desviância de um modelo linear generalizado com componente aleatória contendo uma distribuição pertencente à família exponencial é:

$$D = 2 \sum_{i=1}^{n} (y_i(\bar{\theta}_i - \hat{\theta}_i) - b(\bar{\theta}_i) + b(\hat{\theta}_i)).$$

Para a distribuição binomial, os parâmetros são:

$$\theta_i = log\left(\frac{\pi_i}{1-\pi_i}\right), b(\theta_i) = -n_i log(1-\pi_i), \varphi = 1.$$

Além disso, no modelo saturado temos $\pi_i = \frac{y_i}{n}$ e no modelo em estudo o valor ajustado da i-ésima observação é dado por $\hat{\mu}_i = n\hat{\pi}_i$, logo, a fórmula para a desviância pode ser reescrita como:

$$D = 2 \sum_{i} \left(y_i log\left(\frac{y_i}{\hat{\mu}_i}\right) - (n - y_i) log\left(\frac{n-y_i}{n-\hat{\mu}_i}\right)\right)$$

Outra medida que permite avaliar a qualidade do ajustamento de um modelo é a estatística χ^2 de Pearson, embora seja apenas possível utilizá-la quando os dados estão agrupados em padrões de covariáveis:

$$X_P^2 = \sum_{i=1}^{K} \frac{(y_i - \hat{\mu}_i)^2}{n_i \hat{\pi}_i (1 - \hat{\pi}_i)} \tag{2.44}$$

2.4.2 Regressão de Poisson

Quando o caso em estudo envolve dados na forma de contagem ou na forma de uma taxa de ocorrência, é comum utilizar-se um modelo linear generalizado com resposta seguindo uma distribuição de Poisson, denominado por regressão de Poisson. Exemplos destes dados incluem o número de fatalidades de anfíbios ao longo de uma estrada em Portugal (**?**)), número de chamadas telefónicas num *call center*, número de pessoas numa fila de espera, entre outros.

Seja Y uma variável aleatória que segue uma distribuição de Poisson com média μ. Então, a sua função de probabilidade é dada por

$$f(y|\mu) = \mu^y \frac{e^{-\mu}}{y!} \tag{2.45}$$

onde μ representa o número médio de ocorrências de um dado acontecimento num determinado período de tempo. Uma propriedade importante da distribuição de Poisson é a igualdade entre a média da variável aleatória e a sua variância, ou seja, $E(Y) = V(Y) = \mu$. Outra propriedade reside no facto de o somatório de variáveis aleatórias independentes, cada uma seguindo uma distribuição de Poisson, também seguir uma distribuição de Poisson:

$$Y_1 + \ldots + Y_n \sim P(\mu_1 + \ldots + \mu_n),$$

permitindo que um modelo de regressão de Poisson a partir de dados individuais retorne os mesmos resultados que um modelo de regressão de Poisson a partir de dados agrupados.

Seja $Y_1, \ldots, Y_n$ uma amostra aleatória representando o número de ocorrências de um acontecimento. Dado um vector de variáveis explicativas $X = (X_1, \ldots, X_p)$ e uma observação do indivíduo i, $x_i = (x_{i1}, \ldots, x_{ip})$,

$$Y|X = x_i \sim P(\mu(x_i)).$$

Tem-se então que

$$\mu(x_i) = E(Y|X = x_i) = V(Y|X = x_i).$$

Como o objectivo é modelar a média de $Y|X = x_i$, a opção mais comum é utilizar uma transformação logarítmica para relacionar μ_i e a combinação linear das variáveis explicativas, já que esta última pode tomar qualquer valor real enquanto que a média só pode tomar valores não negativos visto que se trata do número de ocorrências de um acontecimento. No modelo de regressão de Poisson, a função de ligação canónica é a função logaritmo. Assim, temos a seguinte relação:

$$log(\mu(x_i)) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \tag{2.46}$$

que se pode traduzir num modelo multiplicativo:

$$\begin{aligned} \mu(x_i)) &= e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}} \\ &= e^{\beta_0} e^{\beta_1 x_{i1}} \ldots e^{\beta_p x_{ip}} \end{aligned}$$

No modelo de regressão de Poisson, os coeficientes de regressão $\beta_j, j = 0, \ldots, p$ representam a variação esperada no logaritmo da média quando a variável explicativa x_j varia por unidade, no caso de a variável ser contínua, ou quando a variável explicativa muda de categoria, no caso de a variável ser categórica. Por sua vez, e^{β_j} representa o efeito multiplicativo da variável x_j na média da resposta.

Descrição do Modelo de Regressão de Poisson

- $Y_i|X = x_i \sim P(\mu(x_i))$ (por definição, segue que a variância de Y_i é igual a μ_i).
- A parte sistemática é dada por $\eta(x_{i1}, ..., x_{ip}) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$
- A função de ligação é a função logarítmica, que relaciona a média de Y_i e a função preditora η:

$$log(\mu_i) = \eta(x_{i1}, ..., x_{iq}) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \tag{2.47}$$

Risco Relativo

O risco relativo corresponde ao quociente entre o número esperado de eventos num grupo e o número esperado de eventos noutro grupo. Assumindo que não há interações entre as variáveis explicativas, o risco relativo é usado para interpretar os parâmetros de regressão do modelo. Se o risco relativo é superior a 1, observa-se uma associação positiva entre a resposta e o factor em estudo. Se o risco relativo for inferior a 1, essa associação é negativa. Quando o risco relativo toma o valor 1, a associação é inexistente. Pela própria definição do risco relativo, é impossível este tomar valores negativos.

Suponhamos que X_j toma valores contínuos. Aumentando em c unidades a componente j de $x = (x_1, \ldots, x_j, \ldots, x_p)$, tem-se:

$$\begin{aligned} \mu(x) &= e^{\beta_0 + \ldots + \beta_j x_j + \ldots + \beta_p x_p} \\ \mu(x_{+c}) &= e^{\beta_0 + \ldots + \beta_j x_j + c\beta_j + \ldots + \beta_p x_p}, \end{aligned} \tag{2.48}$$

logo

$$\frac{\mu(x_{+c})}{\mu(x)} = e^{c\beta_j}$$

Assim, o risco relativo para a resposta é dado por:

$$RR(x_{+c}, x) = e^{c\beta_j} \tag{2.49}$$

Verifica-se então que quando aumentamos a variável X_j em c unidades, mantendo as outras variáveis explicativas constantes, a média da resposta é modificada por um factor

de $e^{c\beta_j}$.

No caso em que a variável explicativa X_j toma valores dicotómicos, em que os níveis estão codificados em 0 e 1, por exemplo, temos a seguinte fórmula para o risco relativo:

$$\begin{aligned} RR(X_j = 1, X_j = 0) &= \frac{E(Y|X_j = 1)}{E(Y|X_j = 0)} \\ &= \frac{e^{\beta_0 + \ldots + \beta_{j-1}x_{j-1} + \beta_j + \beta_{j+1}x_{j+1} + \ldots + \beta_p x_p}}{e^{\beta_0 + \ldots + \beta_{j-1}x_{j-1} + \beta_{j+1}x_{j+1} + \ldots + \beta_p x_p}} \\ &= e^{\beta_j} \end{aligned} \tag{2.50}$$

Logo, o número esperado de acontecimentos entre os indivíduos em que $X_j = 1$ é e^{β_j} vezes o número esperado de acontecimentos entre os indivíduos em que $X_j = 0$, mantendo as outras variáveis explicativas constantes.

Generalizando este caso, é possivel estimar riscos relativos quando a variável X_j tem mais do que 2 níveis. Supondo que existem L níveis dentro da variável explicativa, esta é representada por $L - 1$ variáveis *dummy*. Os riscos relativos para as classes são estimados tendo como classe de referência a categoria 0.

Na ausência de variáveis explicativas, ou seja, quando todas as variáveis tomam o valor nulo, o logaritmo do número médio de acontecimentos é igual ao parâmetro de regressão β_0.

Intervalo de confiança para o risco relativo

Sob a hipótese nula

$$H_0 : RR = 1,$$

que traduz a inexistência de associação entre resposta e variável, o estimador $\hat{RR}$ segue uma distribuição aproximadamente log-normal. De acordo com **?**), os intervalos de confiança a $(1 - \alpha)\%$ para o risco relativo são dados por:

$$\left(\widehat{RR} \times e^{\pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{O_1} + \frac{1}{O_2}}} \right)$$

onde O_1 representa o número de eventos observados num grupo e O_2 representa o número de eventos observados noutro grupo. Quando o correspondente intervalo de confiança para o risco relativo contiver o valor 1, o risco relativo não é estatisticamente significativo, não se podendo concluir que os grupos são diferentes.

Estimação dos Parâmetros

No caso dos modelos lineares generalizados, a estimação dos parâmetros é feita com base no método da máxima verosimilhança. Este método consiste em especificar um critério de verosimilhança conjunta L para todos os dados $y_1, \ldots, y_n$:

$$L = P(Y_1 = y_1 \text{ e } Y_2 = y_2 \text{ e } \ldots \text{ e } Y_n = y_n),$$

e depois maximizá-lo em função dos parâmetros de regressão desconhecidos.

Assumindo que as observações são independentes, podemos utilizar a regra básica das probabilidades $P(A \text{ e } B) = P(A) \times P(B)$ para reescrever a função de verosimilhança L da seguinte forma (Zuur et al., 2009):

$$L = \prod_i \frac{\mu_i^{y_i} \times e^{-\mu_i}}{y_i!} \tag{2.51}$$

De forma a simplificar o processo, é usual prosseguir-se os cálculos utilizando o logaritmo da verosimilhança, já que podemos gozar da propriedade aditiva do logaritmo (**?**)):

$$\begin{aligned} log(L) &= \sum_i (log(\mu_i^{y_i} \times e^{-\mu_i}) - log(y_i!) \\ &= \sum_i (y_i \times log(\mu_i) - \mu_i - log(y_i!)) \\ &= \sum_i (y_i \times \mathbf{x}_i^T \times \boldsymbol{\beta} - e^{\mathbf{x}_i^T \times \boldsymbol{\beta}} - log(y_i!)) \end{aligned} \tag{2.52}$$

O termo $log(y_i!)$ pode ser deixado de parte já que não contém nenhum parâmetro de regressão. De seguida, obtêm-se as derivadas de primeira ordem em relação a β, igualam-se a 0 e resolvem-se as equações.

$$\begin{aligned} \frac{\partial log(L)}{\partial \boldsymbol{\beta}} &= \sum_i (y_i \times \mathbf{x}_i^T - \mathbf{x}_i^T \times e^{\mathbf{x}_i^T \times \boldsymbol{\beta}}) \\ &= \sum_i \mathbf{x}_i^T \times (y_i - \mu_i) \end{aligned} \tag{2.53}$$

Igualando a zero, obtém-se

$$\sum_i \mathbf{x}_i^T \times (y_i - \mu_i) = 0 \tag{2.54}$$

ou seja

$$Xy = X\hat{\mu} \tag{2.55}$$

onde X é a matriz de especificação, com uma linha por cada observação e uma coluna por cada variável explicativa.

Aplicando o método iterativo dos mínimos quadrados reponderados, obtêm-se então as estimativas para os parâmetros de regressão.

Desviância

A desviância é definida como o dobro da diferença entre o logaritmo da verosimilhança de um modelo que se ajusta perfeitamente aos dados (modelo saturado) e o modelo em estudo. Para o modelo de regressão de Poisson, a fórmula da desviância é:

$$D(y, \hat{(}\mu)) = 2\sum_i (y_i log(\frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i))), \tag{2.56}$$

onde $\hat{\mu}_i$ são os valores previstos pelo modelo.

O primeiro termo representa o dobro da soma dos valores observados multiplicados pelo logaritmo da razão entre os valores observados e os valores previstos pelo modelo. O segundo termo, que corresponde ao dobro da soma das diferenças entre os valores observados e os valores previstos, é igual a zero sempre que o modelo contiver uma constante β_0, por (2.55). De facto, suponhamos que o modelo inclui uma constante. Então, uma das colunas da matriz de especificação é uma coluna cujos elementos são todos iguais a 1. Multiplicando esta coluna pelo vector y, obtém-se a soma de todas as observações. Do mesmo modo, multiplicando essa coluna pelos valores previstos do modelo, obtém-se a soma dos valores previsto. Assim, as 2 somas são iguais, já que $Xy = X\hat{\mu}$.

Para amostras de tamanho grande, a desviância segue aproximadamente uma distribuição qui-quadrado com $n - p$ graus de liberdade, onde n é o número de observações e p é o número de pârametros. Assim, a desviância pode ser utilizada para testar a qualidade de ajustamento do modelo.

Outra medida alternativa para verificar a qualidade do modelo é através da estatística de Pearson generalizada, dada por:

$$\chi^2_p = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \tag{2.57}$$

Tal como para a desviância, para amostras de tamanho grande, a estatística de Pearson segue aproximadamente uma distribuição qui-quadrado com $n - (p + 1)$ graus de liber-

dade.

2.4.3 Regressão Binomial Negativa

A regressão Binomial Negativa é utilizada principalmente quando existe sobredispersão no modelo de regressão de Poisson.

Seja então Y uma variável aleatória seguindo a distribuição Binomial Negativa, ou seja, $Y \sim BN(\mu,\alpha)$. A função probabilidade de Y é dada por:

$$\begin{aligned} f(y;\mu,\alpha) &= \frac{\Gamma(y_i+\frac{1}{\alpha})}{\Gamma(y_i+1)\Gamma(\frac{1}{\alpha})}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \\ &= \binom{y_i+\frac{1}{\alpha}-1}{\frac{1}{\alpha}-1}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}. \end{aligned} \tag{2.58}$$

Denomina-se α por parâmetro ancilar ou de heterogeneidade. Para a distribuição Binomial Negativa, a variância é superior à média,

$$E(Y) = \mu$$

$$V(Y) = \mu + \alpha\mu^2$$

Quando α tende para zero, a distribuição Binomial Negativa tende para a distribuição de Poisson.

O modelo de regressão Binomial Negativa pode ser descrito analogamente ao modelo de regressão de Poisson, mas neste caso Y_i segue uma distribuição Binomial Negativa de média μ_i. A interpretação dos parâmetros do modelo também é análoga à do modelo de regressão de Poisson, sendo efectuada através do risco relativo.

Os parâmetros são obtidos através do método da máxima verosimilhança. A função log-verosimilhança é dada por (**?**)):

$$\begin{aligned} log(L(\beta)) &= \sum_{i=1}^{n}\left(y_i log\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)\right) - \left(\frac{1}{\alpha}\right)log(1+\alpha\mu_i) + log\left(\Gamma\left(y_i+\frac{1}{\alpha}\right)\right) \\ &\quad -log(\Gamma(y_i+1)) - log\left(\Gamma\left(\frac{1}{\alpha}\right)\right) \end{aligned} \tag{2.59}$$

Desviância

Para o modelo de regressão Binomial Negativa, a desviância é dada pela seguinte fórmula (**?**)):

$$D = 2\sum_{i=1}^{n}\left(y_i log\left(\frac{y_i}{\hat{\mu}_i}\right) - \left(\frac{1}{\alpha} + y_i\right) log\left(\frac{1+\alpha y_i}{1+\alpha \hat{mu}_i}\right)\right) \tag{2.60}$$

Tal como no modelo de regressão de Poisson, a estatística de Pearson generalizada também pode ser usada, sendo a sua expressão dada por:

$$\chi^2_P = \sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i + \alpha\hat{\mu}_i^2}. \tag{2.61}$$

Para amostras de tamanho grande, a estatística de Pearson segue aproximadamente uma distribuição qui-quadrado com $n - (p+1)$ graus de liberdade.

Capítulo 3

Modelos de contagem para dados com excesso de zeros

Em muitos casos da vida real, os dados de contagem contêm um número elevado de zeros. Esta frequência não esperada de valores nulos pode levar a que os dados não possam ser ajustados por uma distribuição de Poisson ou Binomial Negativa, distribuições normalmente utilizadas para dados desta natureza. Neste capítulo serão abordados modelos que lidam com esta situação: modelos com zeros inflacionados e modelos com barreira. Além de tratarem este fenómeno, estes modelos também são úteis para acomodar a sobredispersão muitas vezes presente neste tipo de dados.

3.1 Modelos com Zeros Inflacionados

A primeira aplicação de modelos com zeros inflacionados foi efetuada por **?**), com a utilização de um modelo de regressão de Poisson com zeros inflacionados. De acordo com estes modelos, as contagens são modeladas através de dois processos: (i) o primeiro trata do número excessivo de zeros; e (ii) o segundo trata das contagens (podendo algumas destas ser nulas) através de uma distribuição de Poisson ou Binomial Negativa. O primeiro processo ocorre com probabilidade π_i enquanto o segundo ocorre com probabilidade $1 - \pi_i$, sendo que este último gera uma contagem a partir de um modelo de Poisson ou a partir de um modelo Binomial Negativa, sendo π_i definido como a probabilidade de se observar um zero falso. Os zeros falsos são os zeros que não provêm de uma contagem de Poisson ou Binomial Negativa mas sim de uma massa pontual em zero.

3.1.1 Modelo de regressão de Poisson com zeros inflacionados

O modelo de regressão de Poisson com zeros inflacionados (ZIP), assume para o primeiro processo que

$$P(Y_i = 0) = P(\text{zero falso}) + (1 - P(\text{zero falso})) \times P(\text{processo de contagem origina um zero})$$

Como foi referido anteriormente, assumindo que a probabilidade de Y_i ser um zero falso é binomialmente distribuida com probabilidade π_i, a equação anterior traduz-se em

$$P(Y_i = 0) = \pi_i + (1 - \pi_i) \times P(\text{processo de contagem origina um zero})$$

Tendo em conta que a probabilidade de se observar um zero segundo a distribuição Poisson é dada por $P(Y_i = 0) = \frac{{\mu_i}^0 e^{-\mu_i}}{0!} = e^{-\mu_i}$, a equação anterior pode ser reescrita como:

$$P(Y_i = 0) = \pi_i + (1 - \pi_i) \times e^{-\mu_i} \tag{3.1}$$

No segundo processo, assume-se que as contagens seguem uma distribuição de Poisson com média μ_i. A função de probabilidade de uma distribuição de Poisson é dada pela seguinte fórmula:

$$f(y_i; \mu_i) = \frac{{\mu_i}^{y_i} e^{-\mu_i}}{y_i!}, \tag{3.2}$$

logo para o segundo processo tem-se

$$P(Y_i = y_i | x_i) = (1 - \pi_i) \times \frac{{\mu_i}^{y_i} e^{-\mu_i}}{y_i!}, \tag{3.3}$$

onde $\mu_i = \mu(x_i)$ e $\pi_i = \pi(x_i)$, $i = 1, \ldots, n$.

Assim, a distribuição para um modelo ZIP é dada por:

$$P(Y_i = y_i | x_i) = \begin{cases} \pi_i + (1 - \pi_i) \times e^{-\mu_i} & \text{se } y_i = 0 \\ (1 - \pi_i) \times \frac{{\mu_i}^{y_i} e^{-\mu_i}}{y_i!} & \text{se } y_i > 0 \end{cases}$$

A média μ_i das contagens positivas é modelada através de uma regressão de Poisson, ou seja:

$$log(\mu_i) = X_i \beta \tag{3.4}$$

onde X é a matriz de observações das covariáveis e o vetor β corresponde aos parâmetros de regressão.

A probabilidade de existir um zero falso é geralmente modelada através de uma regressão logística:

$$logit(\pi_i) = Z_i\gamma \tag{3.5}$$

onde Z e γ são a matriz de observações das covariáveis e o vetor dos parâmetros de regressão, respetivamente.

Utilizam-se as letras Z e γ para as covariáveis e para os parâmetros de regressão, respetivamente, já que não têm de ser necessariamente os mesmos do modelo das contagens positivas.

Utilizando propriedades sobre o valor esperado e a variância ($E(Y) = \sum y \times f(y)$ e $V(Y) = E(Y^2) - E(Y)^2$, a média e variância para este modelo são dadas por (**?**)):

$$E(Y_i) = \mu_i \times (1 - \pi_i)$$

$$V(Y_i) = (1 - \pi_i) \times (\mu_i + \pi_i \times {\mu_i}^2),$$

logo $V(Y_i) > E(Y_i)$.
Nos modelos de zeros inflacionados, as contagens nulas podem ser provenientes do processo de contagem ou de uma massa pontual em zero. Assim, a estimação dos parâmetros não pode ser feita separadamente, o que se traduz no ajustamento conjunto da regressão logística e da regressão de Poisson no que diz respeito à função de log-verosimilhança.

Segundo **?**), a função de log-verosimilhança para o modelo de regressão de Poisson com zeros inflacionados é dada por:

$$\begin{aligned} LL_{Poissonzeroinf} &= \sum_{y_i=0} log(e^{z_i\gamma} + e^{-e^{x_i\beta}}) \\ &+ \sum_{y_i>0} (y_i x_i \beta - e^{x_i\beta}) - \sum_{i=1}^{n} log(1 + e^{z_i\gamma}) \\ &- \sum_{y_i>0} log(y_i!) \end{aligned}$$

Para este modelo, a estimação dos parâmetros é efetuada recorrendo ao algoritmo de Newton-Rhapson ou ao método iterativo dos mínimos quadrados reponderados.

?) desenvolveu um algoritmo EM que maximiza a função log-verosimilhança para este modelo:

- Passo 1: Especificar valores iniciais $\beta^{(0)}$, $\gamma^{(0)}$,ϵ e ϵ_0
- Passo 2: Calcular

$$Z_i^{(0)}(\beta^{(0)}, \gamma^{(0)}) = \begin{cases} [1 + e^{-z_i\gamma^{(0)} - e^{x_i\beta^{(0)}}}]^{-1} & sey_i = 0 \\ 0 & sey_i > 0 \end{cases}$$

- Passo 3: Encontrar o valor de $\hat{\beta}$ através da maximização de $LL(\beta)$ e o valor de $\hat{\gamma}$ através da maximização de $LL(\gamma)$, onde

$$LL(\beta) = \sum_{i=1}^{n}(1 - Z_i)(y_i x_i \beta - e^{x_i \beta}$$

$$LL(\gamma)) \sum_{y_i=0} Z_i z_i \gamma - \sum_{y_i=0} Z_i log(1 + e^{z_i \gamma} - \sum_{i=1}^{n}(1 - Z_i) log(1 + e^{z_i \gamma})$$

- Passo 4: Se um dos seguintes critérios de paragem se verificar:

  - $||\hat{\beta} - \beta^{(0)}|| \geq \epsilon$
  - $||\hat{\gamma} - \gamma^{(0)}|| \geq \epsilon$
  - $|LL(\hat{\beta}, \hat{\gamma} - LL(\beta^{(0)}, \gamma^{(0)})| \geq \epsilon_0$,

  o algoritmo converge. Se não, definir $\beta^{(0)} = \hat{\beta}$ e $\gamma^{(0)} = \hat{\gamma}$ e voltar ao Passo 2 até o método convergir.

3.1.2 Modelo de regressão Binomial Negativa com zeros inflacionados

Quando os dados além de apresentarem um excesso de zeros, apresentam sobredispersão (ou seja, quando a variância é muito superior à média), utiliza-se o modelo de regressão Binomial Negativa com zeros inflacionados, já que a distribuição Binomial Negativa tem mais um parâmetro do que a distribuição de Poisson, sendo possível ajustar a variância independentemente da média. A única diferença entre este modelo e o modelo de regressão de Poisson com zeros inflacionados é a distribuição da variável Y_i, passando a seguir a distribuição Binomial Negativa de parâmetros (μ_i, α) com zeros inflacionados.

A função de probabilidade deste modelo é dada por (**?**)):

$$P(Y_i = y_i | x_i) = \begin{cases} \pi_i + (1 - \pi_i) \times (\frac{1}{1+\alpha\mu_i})^{\frac{1}{\alpha}} & \text{se } y_i = 0 \\ (1 - \pi_i) \times \frac{\Gamma(y_i + \frac{1}{\alpha})}{\Gamma(y_i+1)\Gamma(\frac{1}{\alpha})} (\frac{1}{1+\alpha\mu_i})^{\frac{1}{\alpha}} (1 - \frac{1}{1+\alpha\mu_i})^{y_i} & \text{se } y_i > 0 \end{cases},$$

onde α é uma constante não negativa denominada parâmetro de heterogeneidade (**?**)).

A média e variância do modelo de regressão Binomial Negativa com zeros inflacionados são dadas por (**?**); **?**)):

$$E(Y_i) = \mu_i \times (1 - \pi_i)$$

$$V(Y_i) = (1 - \pi_i) \times \mu_i \times (1 + \pi_i \mu_i + \alpha \mu_i)$$

Note-se que, quando α tende para zero, a distribuição Binomial Negativa com zeros inflacionados tende para a distribuição de Poisson com zeros inflacionados.

A função de log-verosimilhança do modelo de regressao Binomial Negativa com zeros inflacionados é dada por (**?**)):

$$\begin{aligned} LL_{BNzeroinf} &= \sum_{y_i=0} log\left[e^{z_i\gamma} + \left(\frac{1}{1+\alpha e^{x_i\beta}}\right)^{\frac{1}{\alpha}}\right] + \sum_{y_i>0} z_i\gamma \\ &+ log\left(\Gamma\left(y_i + \frac{1}{\alpha}\right)\right) - log(y_i!) - log\left(\Gamma\left(\frac{1}{\alpha}\right)\right) \\ &+ y_i\left[log(\alpha e^{x_i\beta}) - log(1+\alpha e^{x_i\beta})\right] \\ &+ \frac{1}{\alpha} log\left(\frac{1}{1+\alpha e^{x_i\beta}}\right) + \sum_{i=0}^{n} -log(1+e^{z_i\beta}) \end{aligned}$$

Tal como no modelo de regressão de Poisson com zeros inflacionados, a estimação dos parâmetros é efectuada recorrendo a métodos como o algoritmo de Newton-Raphson ou o método iterativo dos mínimos quadrados reponderados.

3.2 Modelos com barreira

Um modelo com barreira ou um modelo *hurdle* é, segundo **?**), "um modelo de contagem modificado onde os processos que geram os zeros e as contagens positivas não são necessariamente o mesmo". A ideia para este modelo foi desenvolvida por **?**), tendo sido posteriormente desenvolvida por **?**) e por **?**) e **?**). Utiliza-se o termo "barreira"pois interpretam-se as contagens positivas como tendo ultrapassado a barreira dos zeros. Note-se que a barreira não tem de encontrar-se necessariamente no valor zero, podendo estar em qualquer valor de acordo com o problema em análise.

Estes modelos são compostos por 2 partes:

- os dados são considerados como zeros *vs* não-zeros e um modelo binomial (regressão logística) é usado para modelar a probabilidade de um valor nulo ser observado.
- as contagens positivas (diferentes de zero) são modeladas através de uma distribuição de Poisson truncada ou Binomial Negativa truncada.

Para cada parte, é necessário considerar um conjunto de covariáveis de interesse. Usualmente, designa-se por Z a matriz das observações das covariáveis utilizadas na modelação da primeira parte e por X a matriz das observações das covariáveis escolhidas para modelar a segunda parte. Se as covariáveis selecionadas forem as mesmas, então $X = Z$. No caso particular em que a probabilidade de um valor nulo ser observado não depende de covariáveis, a matriz Z consiste apenas de uma coluna preenchida por 1's.

Formalmente, pode-se descrever um modelo com barreira através da seguinte fórmula (**?**)):

$$P_{barreira}(y;x,z,\beta,\gamma) = \begin{cases} P_{zero}(0;z,\gamma) & \text{se } y = 0 \\ (1 - P_{zero}(0;z,\gamma)) \times \frac{P_{contagem}(y;x,\beta)}{1-P_{contagem}(0;x,\beta)} & \text{se } y > 0 \end{cases}$$

onde x e z são os vetores das covariáveis e β e γ são os parâmetros de regressão do modelo de contagem truncado e do modelo binomial, respetivamente.

A primeira equação representa a probabilidade de ocorrência de uma observação nula. A segunda equação apresenta a probabilidade de ocorrência de uma observação não nula através da função de probabilidade de uma distribuição de Poisson truncada em $y_i = 0$.

3.2.1 Modelo de regressão de Poisson com barreira

A função de probabilidade do modelo de regressão de Poisson com barreira é dada por (**?**); **?**)) :

$$P(Y_i = y_i) = \begin{cases} 1 - \pi_i & \text{se } y = 0 \\ \frac{\pi_i e^{-\mu_i} {\mu_i}^y}{(1-e^{-\mu_i})y_i!} & \text{se } y > 0 \end{cases}$$

onde π_i é a probabilidade de se observar uma contagem não nula. O processo de contagem de Poisson exclui a probabilidade de existirem zeros, logo estamos perante uma distribuição de Poisson truncada. A segunda equação do sistema diz que a probabilidade de se observar uma contagem diferente de zero é igual à probabilidade de não existir um zero multiplicada pela função de probabilidade de uma distribuição de Poisson truncada.

Para este modelo, temos as seguintes fórmulas para a média e para a variância:

$$E(Y_i) = \frac{\pi_i}{1-e^{-\mu_i}} \times \mu_i$$
$$V(Y_i) = \frac{\pi_i}{1-e^{-\mu_i}} \times (\mu_i + {\mu_i}^2) - \left(\frac{\pi_i}{1-e^{-\mu_i}} \times \mu_i\right)^2$$

A estimação dos parâmetros é baseada no método da máxima verosimilhança. Como o processo de contagem não pode produzir contagens nulas, a função de verosimilhança pode ser estimada como uma soma de 2 funções log-verosimilhanças diferentes. Neste modelo, a função de log-verosimilhança pode ser escrita do seguinte modo (**?**)):

$$\begin{aligned} LL_{barreira} &= \sum_{y_i>0} z_i\gamma - \sum_{i=i}^{n} log(1 + e^{z_i\gamma}) + \sum_{y_i>0} [y_i x_i\beta - e^{x_i\beta} - log(1 - e^{e^{x_i\beta}}) - log(y_i!)] \\ &= LL(z_i\gamma) + LL(x_i\beta), \end{aligned}$$

onde $LL(z_i\gamma)$ é a função de log-verosimilhança do modelo de regressão logística e $LL(x_i\beta)$ é a função de log-verosimilhança de um modelo de regressão de Poisson truncado.

3.2.2 Modelo de regressão Binomial Negativa com barreira

A função de probabilidade do modelo de regressão Binomial Negativa com barreira é dada por (**?**); **?**)):

$$P(Y_i = y_i) = \begin{cases} 1 - \pi_i & \text{se } y = 0 \\ \pi_i [\frac{\Gamma(y_i + \frac{1}{\alpha})}{\Gamma(y_i+1)\Gamma(\frac{1}{\alpha})} (\frac{1}{1+\alpha\mu_i})^{\frac{1}{\alpha}} (1 - \frac{1}{1+\alpha\mu_i})^{y_i})] / (1 - (\frac{1}{1+\alpha\mu_i})^{\frac{1}{\alpha}}) & \text{se } y > 0, \end{cases}$$

onde π_i é a probabilidade de se observar uma contagem não nula. Tal como no modelo de regressão de Poisson com barreira, o processo de contagem é independente do processo das contagens nulas, logo a probabilidade de se observar uma contagem diferente de zero é igual à probabilidade de não existir um zero multiplicada pela função de probabilidade de uma distribuição Binomial Negativa truncada.
Para este modelo, a média e a variância são dadas por (**?**)):

$$E(Y_i) = \frac{\pi_i}{1 - \left(\frac{\frac{1}{\alpha}}{\mu_i + \frac{1}{\alpha}}\right)^{\frac{1}{\alpha}}} \times \mu_i$$

$$V(Y_i) = \frac{\pi_i}{1 - \left(\frac{\frac{1}{\alpha}}{\mu_i + \frac{1}{\alpha}}\right)^{\frac{1}{\alpha}}} \times (\mu_i + {\mu_i}^2(1+\alpha)) - \left(\frac{\pi_i}{1 - \left(\frac{\frac{1}{\alpha}}{\mu_i + \frac{1}{\alpha}}\right)^{\frac{1}{\alpha}}} \times \mu_i \right)^2$$

Tal como no modelo de regressão de Poisson com barreira, a função de verosimilhança deste modelo também é estimada como a soma de funções log-verosimilhança diferentes, podendo ser escrita do seguinte modo (**?**)):

$$\begin{aligned} LL_{barreira} &= \sum_{y_i > 0} z_i\gamma - \sum_{i=i}^{n} log(1 + e^{z_i\gamma}) \\ &+ \sum_{y_i > 0} [log(\Gamma(y_i + \frac{1}{\alpha})) - log(\Gamma(y_i + 1)) - log\left(\Gamma\left(\frac{1}{\alpha}\right)\right) - \frac{log(1 + e^{x_i\beta})}{\alpha} \\ &+ y_i log\left(\frac{\alpha e^{x_i\beta}}{1 + \alpha e^{x_i\beta}}\right) - log\left(1 - \left(1 + \alpha e^{x_i\beta}\right)^{-\frac{1}{\alpha}}\right) \\ &= LL(z_i\gamma) + LL(x_i\beta), \end{aligned}$$

onde $LL(z_i\gamma)$ é a função de log-verosimilhança do modelo de regressão logística e $LL(x_i\beta)$ é a função de log-verosimilhança de um modelo de regressão Binomial Negativa truncado.

Capítulo 4

Análise Estatística

Neste capítulo é feita uma análise descritiva da base de dados considerada neste trabalho e apresentados os vários modelos de regressão usados na sua análise.

4.1 Análise Exploratória e Descritiva da Base de Dados

Para esta tese, a base de dados foi providenciada pela empresa AXA S.A. e incide sobre 496729 indivíduos com apólices de seguros rodoviários. No total, registaram-se 88502 acidentes num período de 2 anos, com início em 1 de janeiro de 2010 e término em 31 de dezembro de 2011. A base de dados fornecida regista, para cada observação, variáveis relacionadas com a viatura, sendo o conjunto das variáveis composto por variáveis categóricas: marca, tipo de carroçaria, número de velocidades, número de portas, número de lugares, tipo de caixa, tipo de combustível e tração; e por variáveis contínuas: cilindrada, número de cavalos, distância entre eixos, peso, idade do veículo e idade do seguro. No entanto, em discussão com o co-orientador Luís Maranhão, atuário da empresa, decidiu-se proceder à transformação das variáveis contínuas em variáveis categóricas. A motivação para esta decisão prende-se com o facto de ser mais intuitivo pensar em intervalos de valores já que pequenas diferenças nos valores de variáveis contínuas não representam alterações significativas ao nível da capacidade dos veículos. Em relação às variáveis categóricas, devido à pouca representatividade de algumas classes, procedeu-se ao agrupamento de algumas categorias numa só. Relativamente à codificação das mesmas no R, por defeito, o software toma como categoria de referência a primeira categoria de cada variável, podendo esta definição ser alterada. Assim, no caso de uma variável categórica cujos factores são caracteres alfabéticos, a categoria de referência será a correspondente à primeira palavra por ordem alfabética. As variáveis cujos factores são dados por caracteres numéricos tem como categoria de referência a classe 1.
De seguida apresenta-se uma exploração sumária das variáveis explicativas e da variável resposta, acompanhada de gráficos descritivos e medidas numéricas sobre as mesmas.

- **Marca**

  A variável *Marca* diz respeito à marca do automóvel, sendo composta por 59 tipos de marcas diferentes. No entanto, um número elevado destas são pouco representativas como é o caso das marcas *Aro*, *Autobianchi*, *Cadillac* e *Galloper*. As-

sim, existiu a necessidade de agrupar classes e, a conselho do co-orientador Luís Maranhão, agruparam-se as 45 marcas menos representativas numa nova classe, *Outra*. Obteve-se assim uma variável categórica composta por 15 marcas de carro diferentes. As tabelas seguintes mostram as diferentes marcas, bem como as suas frequências absoluta e relativa.

Audi	38102 (7.7%)	Honda	4834 (1%)	Peugeot	33425 (6.8%)
BMW	18492 (3.8%)	Mercedes-Benz	48599 (9.8%)	Renault	46156 (9.3%)
Citröen	13700 (2.8%)	Nissan	28839 (5.8%)	Seat	10928 (2.2%)
Fiat	16165 (3.2%)	Opel	54364 (10.9%)	Toyota	17786 (3.6%)
Ford	18051 (3.6%)	Outra	109420 (22%)	Volkswagen	37428 (7.5%)

Tabela 4.1: Frequências absoluta e relativa da variável *Marca*

- **Carroçaria**

  A carroçaria de uma viatura é a estrutura física que a envolve, definindo a sua forma. É constituída pelo cofre do motor, habitáculo dos passageiros e mala. A tabela seguinte permite visualizar os tipos de carroçaria existentes na base de dados, bem como a sua frequência absoluta.

Carroçaria	Frequências absoluta e relativa
Cabriolet	15823 (3.2%)
Com	129069 (26%)
Coupé	18218 (3.7%)
Lim	285952 (57.6%)
Monovolume	46167(9.3%)
Roadster	1142(0.2%)
Targa	358(0.1%)

Tabela 4.2: Frequências absoluta e relativa da variável *Carroçaria*

A divisão em classes foi efectuada entre "Lim"que passou a ser designada por "normal", já que é a carroçaria correspondente ao tipo de carro mais comum, o carro ligeiro de passageiros. Todas as outras classes de carroçaria foram agrupadas na classe "outra".

Normal	Outra
285952 (57.6%)	210777 (42.4%)

Tabela 4.3: Frequências absoluta e relativa da variável *Carroçaria*

- **Cilindrada**

  A cilindrada é definida como o volume varrido pelo deslocamento de uma peça móvel numa câmara hermeticamente fechada durante um movimento unitário. É medida em litros ou centímetros cúbicos (unidade utilizada pela AXA). Na base de dados, a média da cilindrada é de 1867 cm^3, sendo o valor máximo de 6761 cm^3 e

o valor mínimo de 599 cm^3. A figura 4.1 evidencia uma predominância de veículos com cilindrada entre os 1000 e os 2000 cm^3.

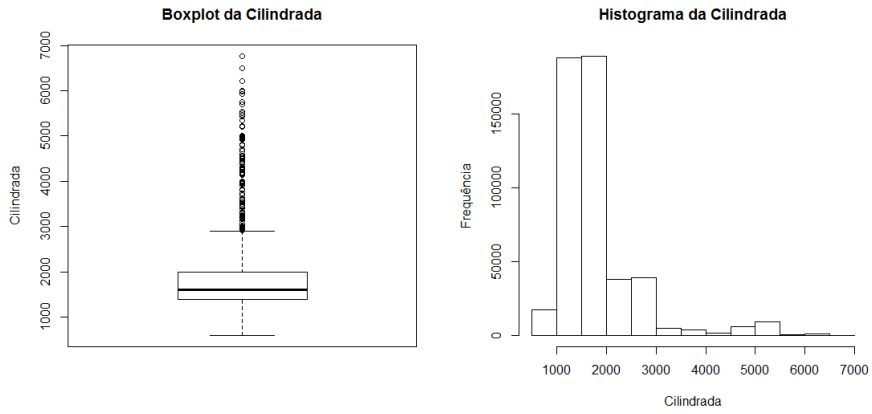


Figura 4.1: Boxplot e Histograma da variável *Cilindrada*

A cilindrada foi dividida em 3 classes: a primeira classe engloba todos os veículos cuja cilindrada é menor que 1300 cm^3, a segunda classe diz respeito aos veículos cuja cilindrada está compreendida entre os 1300 e os 1800 cm^3, inclusivé e a terceira classe os restantes, ou seja, com cilindrada superior a 1800 cm^3.

Classe 1	Classe 2	Classe 3
83650 (16.8%)	204121 (41.1%)	208958 (42.1%)

Tabela 4.4: Frequências absoluta e relativa da variável *Cilindrada*

- **Número de Cavalos**

  O número de cavalos de uma viatura é uma medida do poder da mesma. O termo foi criado no século XVII pelo engenheiro James Watt para comparar o *output* de motores a vapor com o poder de cavalos, meio de trabalho utilizado previamente à invenção dos motores a vapor. 1 cavalo corresponde a 746 *watts*, aproximadamente. Verifica-se que a média desta variável é de 124 cavalos, sendo o valor máximo de 640 cavalos e o valor mínimo de 29 cavalos. A maior parte dos veículos apresenta um número de cavalos entre os 50 e os 150.

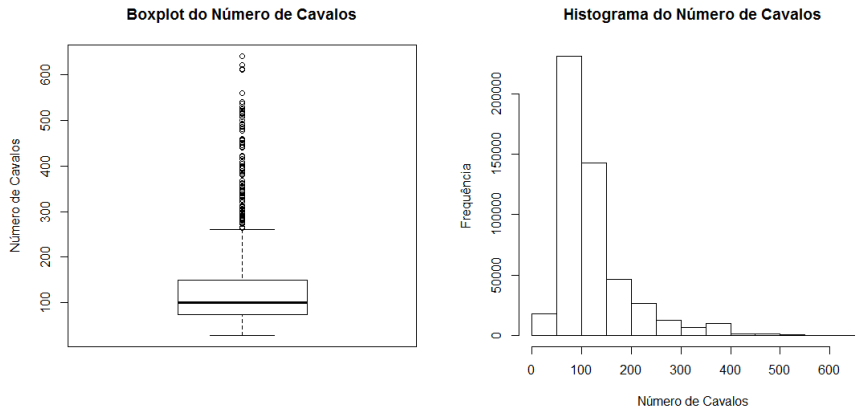


Figura 4.2: Boxplot e Histograma da variável *Cavalos*

Esta variável foi dividida em 3 classes: os veículos com número de cavalos igual ou inferior a 75 cavalos pertencem à primeira classe, aqueles com número de cavalos entre os 75 e os 100, inclusive, pertencem à segunda classe, e os veículos com mais de 100 cavalos pertencem à 3ª classe.

Classe 1	Classe 2	Classe 3
142050 (28.6%)	106886 (21.5%)	247793 (49.9%)

Tabela 4.5: Frequências absoluta e relativa da variável *Cavalos*

- **Caixa**

  A variável *Caixa* diz respeito à caixa de velocidades (ou de mudanças) de uma viatura. A caixa de velocidades tem como principal objectivo transformar a potência do motor em força ou velocidade. As caixas de velocidades podem ser manuais, ou seja, o condutor é responsável pela mudança da velocidade, automática, onde a seleção da velocidade apropriada é efectuada sem intervenção do condutor, semi-automática, existindo um sistema de controlo que, tal como numa caixa de velocidades automática, seleciona as velocidades mas com a opção de o condutor poder tomar controlo dessa seleção. A codificação da variável foi realizada em 2 níveis: o primeiro diz respeito aos veículos com caixa manual e o segundo a todos os outros veículos com caixas de mudanças automáticas ou semi-automáticas.

Caixa manual	Caixa não manual
402876 (81.1%)	93853 (18.9%)

Tabela 4.6: Frequências absoluta e relativa da variável *Caixa*

- **Velocidades**

  O número médio de velocidades das viaturas da amostra é de 5, sendo o número máximo de 8 velocidades e o número mínimo de 3 velocidades.

Velocidades	Frequências absoluta e relativa
3	5664 (1.14%)
4	27764 (5.59%)
5	289923 (58.37%)
6	157810 (31.77%)
7	15086 (3.04%)
8	482 (0.1%)

Tabela 4.7: Frequências absoluta e relativa da variável *Velocidades*

Procedeu-se à sua divisão em 2 classes: na 1ª classe encontram-se todos os veículos com 5 velocidades e na 2ª encontram-se os veículos com um número de velocidades diferente de 5.

Classe 1	Classe 2
323351 (65.1%)	173378 (34.9%)

Tabela 4.8: Frequências absoluta e relativa da variável *Velocidades*

- **Portas**

  A variável *Portas* refere-se ao número de portas de uma viatura, onde o número total de portas também inclui a porta traseira da mala.

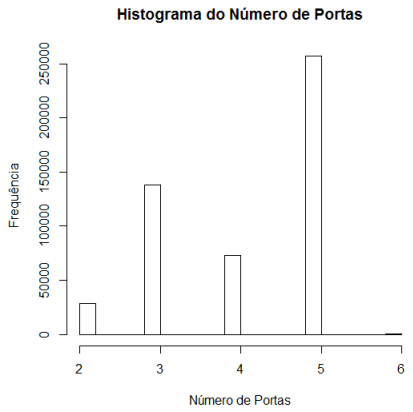


Figura 4.3: Histograma da variável *Portas*

Esta variável foi codificada em 2 níveis: o primeiro diz respeito aos veículos com menos de 5 portas e o segundo aos veículos com 5 ou 6 portas.

Classe 1	Classe 2
239566(48.2%)	257163 (51.8%)

Tabela 4.9: Frequências absoluta e relativa da variável *Portas*

- **Lugares**

  O número de lugares refere-se ao número total de passageiros que uma viatura pode legalmente transportar. O número médio de lugares na amostra é de 5 lugares, sendo o número máximo 9 lugares e o número mínimo 2 lugares.

Lugares	Frequências absoluta e relativa
2	5742 (1.2%)
4	41064 (8.3%)
5	421790 (84.9%)
6	867 (0.2%)
7	16374 (3.3%)
8	10842 (2%)
9	50 (0.01%)

Tabela 4.10: Frequências absoluta e relativa da variável *Número de Lugares*

O número de lugares foi dividido em 2 classes: na 1ª classe encontram-se todos os veículos com 5 lugares e na 2ª encontram-se os veículos com um número de lugares diferente de 5.

Classe 1	Classe 2
421790 (84.9%)	74939 (15.1%)

Tabela 4.11: Frequências absoluta e relativa da variável *Lugares*

- **Combustível**

  O tipo de combustível usado divide-se entre gasolina e gasóleo. O primeiro é obtido através do processamento de petróleo. À gasolina são geralmente adicionados aditivos que visam aumentar a *performance* do combustível. O gasóleo (ou óleo diesel) é derivado da destilação do petróleo, formado principalmente por hidrocarbonetos (carbono e hidrogénio) e, em menores concentrações, por enxofre, nitrogénio e oxigénio. Algumas das vantagens dos carros movidos a gasóleo face aos carros movidos a gasolina passam pelo preço mais reduzido e menor consumo de combustível. No entanto, a escolha de viaturas com tipos de combustível diferentes deve ser adequada à necessidade de cada condutor, dependendo se faz uma condução longa, isto é, se percorre grandes distâncias ou se utiliza o carro para deslocações curtas. Pode-se verificar na tabela 4.12 que a amostra é constituída de uma forma homogénea em relação ao tipo de combustível utilizado pelas viaturas.

Gasóleo	Gasolina
246374 (49.6%)	250355 (50.4%)

Tabela 4.12: Frequências absoluta e relativa da variável *Combustível*

- **Distância entre eixos**

  A distância entre eixos é a distância entre os eixos dianteiro e traseiros, sendo usualmente medida entre os centros das rodas dianteira e traseira do mesmo lado de cada eixo. Esta variável é determinante no que diz respeito à estabilidade de um veículo. A unidade de medida da distância entre eixos é o centímetro. Na amostra recolhida, a média da distância entre eixos é de aproximadamente 2600 *cm*, sendo a distância máxima de 3400 *cm* e a distância mínima de 1812 *cm*.

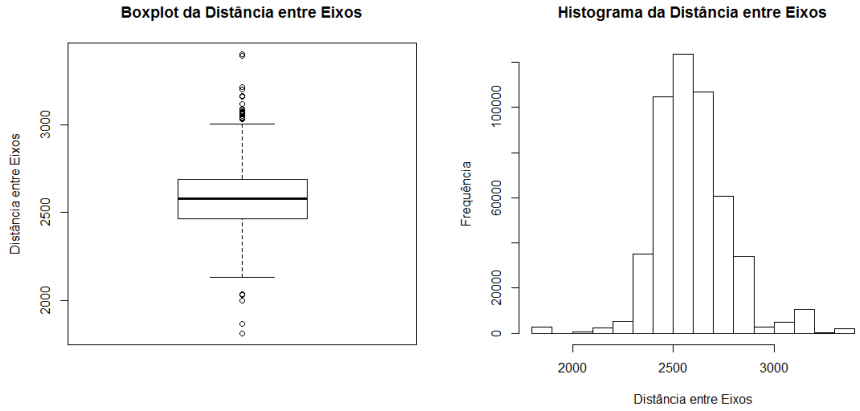


Figura 4.4: Boxplot e Histograma da variável *Distância entre Eixos*

A distância entre eixos foi dividida em 2 níveis: o primeiro engloba os veículos cuja distância entre eixos é inferior a 2600 cm^3 e o segundo todos os veículos cuja distância entre eixos é igual ou superior a 2600 cm^3.

Classe 1	Classe 2
251957 (50.7%)	244772 (49.3%)

Tabela 4.13: Frequências absoluta e relativa da variável *Distância entre Eixos*

- **Tração**

  A tração de uma viatura diz respeito à força exercida sobre o pneu, sendo necessária para deslocar a viatura. O motor de um carro produz força de torção, designada por torque, que faz o veículo mover-se. As diversas mudanças de velocidade multiplicam o torque e distribuem-no às rodas. Quando apenas as rodas da frente ou as rodas traseira puxam o carro, designa-se por tração a 2 rodas. Quando todas as rodas do carro deslocam o veículo, estamos perante tração às 4 rodas. As viaturas da amostra são compostas maioritariamente por carros com tração às 2 rodas, evidenciado pela tabela 4.14. A categoria de referência é tração a 2 rodas.

Tração a 2 rodas	Tração a 4 rodas
452238 (91%)	44492 (9%)

Tabela 4.14: Frequências absoluta e relativa da variável *Tração*

- **Peso**

  O peso do veículo é definido como a soma do peso do chassis do carro (estrutura de suporte do carro), da carroçaria, do motor, das rodas e todas as partes constituintes da viatura. Não é incluído o peso do condutor, dos passageiros ou da carga que o veículo pode eventualmente levar. Ao peso do veículo em vazio também se dá o nome de *tara*. A média do peso do carro é de 1319 *kg*. O veículo mais pesado pesa 2745 *kg* e o veículo mais leve pesa 572 *kg*.

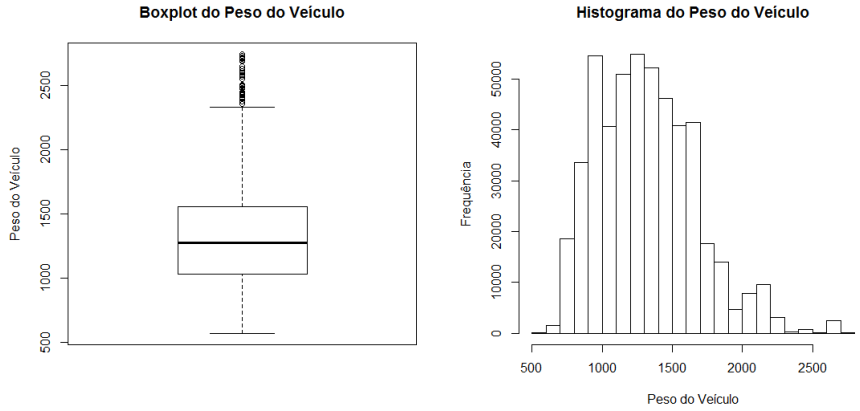


Figura 4.5: Boxplot e Histograma da variável *Peso*

A variável peso foi dividida em 3 classes: a primeira diz respeito aos veículos cujo peso é menor do que 1250 *kg*, exclusivé; a segunda classe engloba os veículos cujo peso se encontra entre os 1250 e os 1500 *kg*, inclusivé; à última classe pertencem as viaturas cujo peso é superior a 1500 *kg*.

Classe 1	Classe 2	Classe 3
224241 (45.1%)	129448 (26.1%)	143040 (28.8%)

Tabela 4.15: Frequências absoluta e relativa da variável *Peso*

- **Idade do veículo**

  A idade do veículo indica o número de anos que a viatura tem. Recorrendo à figura 4.6, verifica-se que a amostra é maioritariamente constituída por veículos recentes.

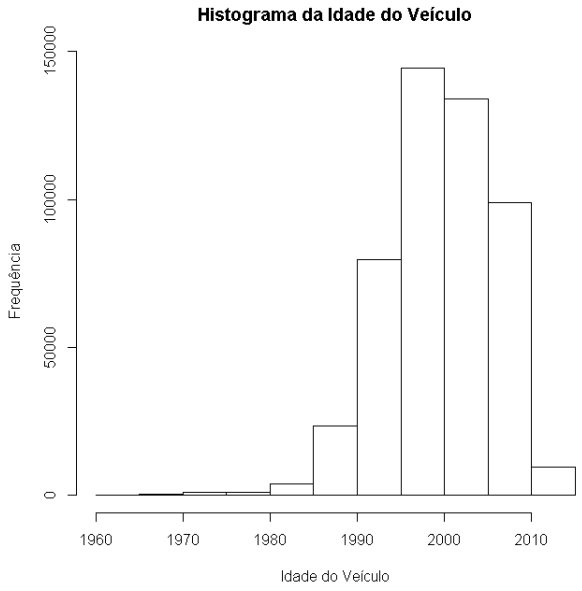


Figura 4.6: Histograma da variável *Idade do Veículo*

Esta variável foi codificada em 5 classes, que se podem observar na tabela 4.16:

Idade (em anos)	Frequências absoluta e relativa
≤ 4	85971 (17.3%)
$]4-8]$	96920 (19.5%)
$]8-12]$	127746 (25.7%)
$]12-25]$	179381 (36.1%)
> 25	6711 (1.4%)

Tabela 4.16: Frequências absoluta e relativa da variável *Idade do Veículo*

- **Idade do Seguro**

A idade do seguro prende-se com o número de anos do contrato do seguro automóvel realizado por cada indivíduo. Analisando a figura 4.7, constata-se que a amostra apresenta uma grande percentagem de veículos recentes.

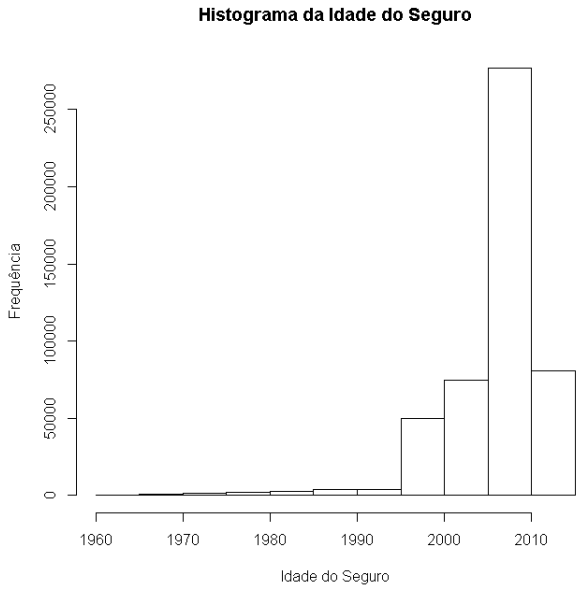


Figura 4.7: Histograma da variável *Idade do Seguro*

O número de anos da apólice foi dividido em 4 categorias, que podem ser observadas na tabela apresentada abaixo:

Idade (em anos)	≤ 1	$]1-2]$	$]2-5]$	>5
Frequência	167625 (33.7%)	68050 (13.7%)	122363 (24.6%)	138691 (27.9%)

Tabela 4.17: Frequências absoluta e relativa da variável *Idade do Seguro*

- **Número de Acidentes**

  O número de acidentes é a variável resposta e é sobre ela que se pretende analisar o efeito das variáveis explicativas descritas anteriormente. O primeiro gráfico da figura 4.8 evidencia claramente o excesso de respostas nulas, perfazendo estas 85% da amostra. O segundo gráfico permite visualizar a frequência da variável resposta quando esta não é igual a zero, concluindo-se que, dos 15% de observações que registaram pelo menos um acidente, a sua maioria diz respeito a viaturas com 1 ou 2 acidentes.

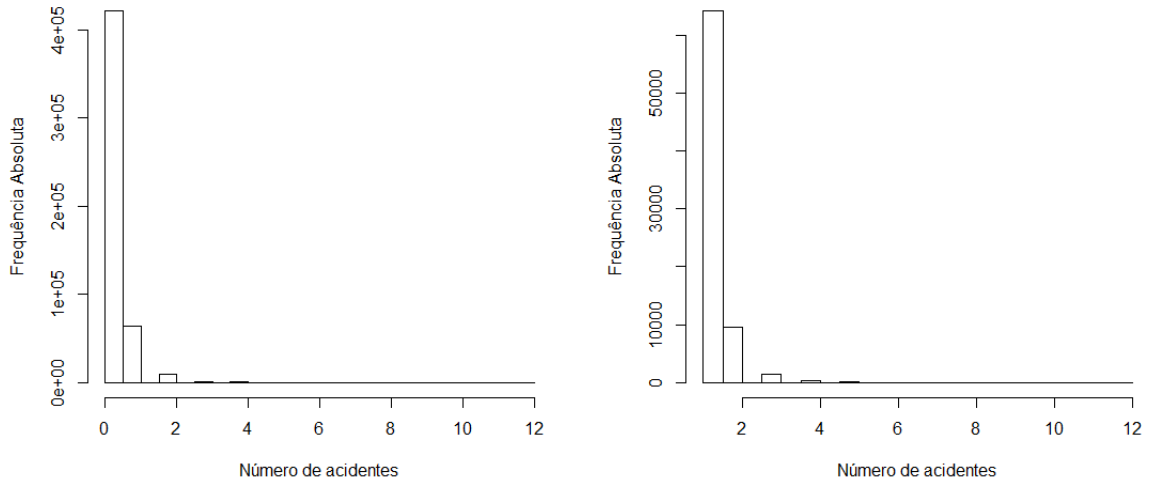


Figura 4.8: Histogramas da variável resposta *Número de Acidentes*

A tabela 4.18 apresenta a frequência absoluta das observações por número de acidentes e, entre parêntesis, a frequência relativa.

Número de Acidentes	Frequência Absoluta (Frequência Relativa)
0	421378 (84.83%)
1	64214 (12.92%)
2	9488 (1.91%)
3	1364 (0.27%)
4	228 (0.04%)
5	43 ($<$ 0.01%)
6	10 ($<$ 0.01%)
7	3 ($<$ 0.01%)
12	1 ($<$ 0.01%)

Tabela 4.18: Frequência absoluta da variável resposta *Número de Acidentes*

- **Número de dias**

  A cada viatura está associado um número de dias no qual a apólice esteve em vigor entre 1 de janeiro de 2010 e 31 de dezembro de 2011, podendo tomar um valor mínimo de 1 e um valor máximo de 730. A figura 4.9 ilustra o histograma do número de dias em que as apólices estiveram em vigor. Verifica-se que mais de metade das observações ($\approx 56\%$) estiveram em vigor durante todo o período temporal em análise. Esta variável é considerada como um *offset* no ajustamento do modelo, não se tratando de uma variável explicativa como as restantes.

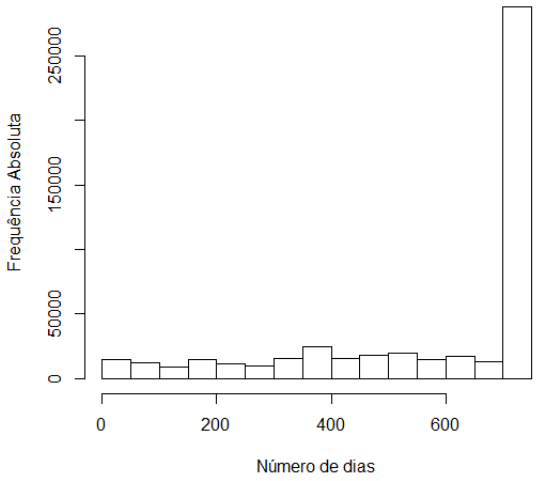


Figura 4.9: Histograma do *Número de Dias*

A tabela 4.19 representa a codificação das variáveis explicativas no R, com o objectivo de clarificar o *output* devolvido pelo R aquando do ajustamento dos modelos.

Nome da Variável	Variável no R
Marca	marca
Carroçaria	car
Cilindrada	cil
Número de Cavalos	cavalos
Caixa	caixa
Velocidades	vel
Portas	portas
Combustível	comb
Peso	peso
Idade do Veículo	iv
Idade do Seguro	is

Tabela 4.19: Codificação das variáveis no R

Do conjunto das variáveis explicativas, foram escolhidas algumas que, na opinião do co-orientador Luís Maranhão, pudessem apresentar interação. No caso de se verificar interação entre 2 variáveis, o modelo deve ser ajustado tendo em conta este fenómeno. Assim, foram selecionados 2 conjuntos de variáveis, onde o primeiro é constituído pelas variáveis *Peso*, *Número de Cavalos*, *Cilindrada*, *Combustível*, enquanto o segundo é formado pelas variáveis *Idade do Veículo* e *Idade do Seguro*, sendo que, no decorrer do ajustamento dos modelos, se testaram todas as interações entre 2 variáveis do mesmo conjunto. O número reduzido de variáveis explicativas consideradas prende-se com razões de complexidade computacional já que o número total de interações a testar é elevado. Por exemplo, para interações de 2^o grau, no caso em questão em que existem 14 variáveis

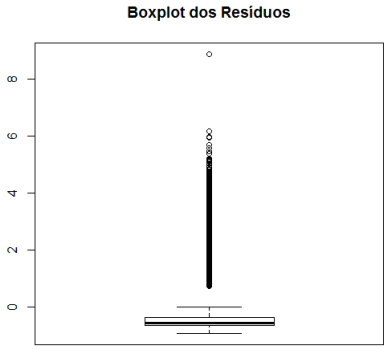
explicativas, teríamos de testar 91 interações. Além disso, apenas se verificou a existência de interações de ordem igual a 2 e não superior, já que se tornariam mais difíceis de interpretar.

4.2 Regressão de Poisson e Regressão Binomial Negativa

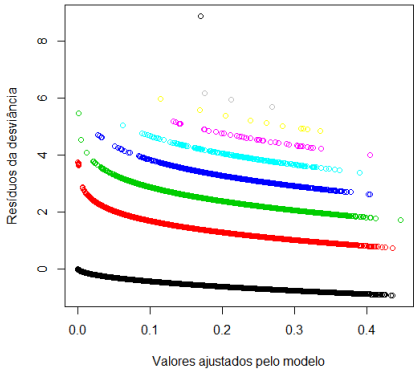
Face ao objetivo do estudo e dado que estamos a lidar com dados de contagem, recorremos ao modelo de regressão de Poisson para fazer um ajustamento aos dados. Recorrendo à função *glm* da biblioteca *stats* do software *R*, começámos por considerar um modelo de regressão de Poisson. A seleção de variáveis foi realizada através de procedimentos de seleção automática, para um nível de significância (valor-p) de 0.05. Dado que as variáveis explicativas estão divididas por fatores, cada fator terá um nível de significância diferente. Assim, no caso de um modelo apresentar uma variável em que pelo menos um dos fatores tem um nível de significância menor ou igual do que 0.05, essa variável é incluída no modelo, mesmo que outro fator da mesma variável apresente um nível de significância superior ao limite estabelecido de 0.05. É apresentada em anexo a tabela com as variáveis explicativas selecionadas, bem como o valor, erro-padrão e nível de significância dos parâmetros de regressão estimados associados às variáveis.
Para o modelo de regressão de Poisson, os valores do AIC e do BIC são iguais a 487798 e 488365, respetivamente. O parâmetro de dispersão ϕ é igual a 1.227. Como o valor de ϕ é ligeiramente superior a 1 é possível que o modelo apresenta sobredispersão. O cálculo do parâmetro de sobredispersão é efetuado dividindo o valor da desviância pelos graus de liberdade do modelo. Na prática, um valor de ϕ perto de 1 não apresenta provas de sobredispersão, utilizando-se nesta tese um valor de corte de 1.2 como sugestão de existência de sobredispersão. A figura 4.10 ilustra alguns gráficos de diagnóstico relativos ao modelo de regressão de Poisson com o objectivo de se obter uma melhor compreensão do modelo em questão.

Analisando os gráficos verifica-se que o modelo em causa não traduz um bom ajustamento aos dados. A figura 4.10(a) diz respeito ao *boxplot* dos resíduos. Tomando como valor de corte o valor 3 (valor de referência geralmente utilizado para identificar *outliers*[1]), existem 1502 observações com resíduos elevados. No segundo gráfico (figura 4.10(b)), é evidenciado o aumento do valor dos resíduos à medida que o número de acidentes aumenta. Além disso, dentro das observações com o mesmo número de acidentes, verifica-se que aquelas cujos valores ajustados pelo modelo são menores apresentam resíduos maiores, o que é esperado visto que a diferença entre valor ajustado e valor observado é maior. A figura 4.10(c) evidenciam o contraste entre os valores observados e os valores ajustados pelo modelo, sendo que estes nunca ultrapassam 0.5, traduzindo assim um mau ajustamento do modelo aos dados em questão, já que todos os valores são classificados como 0. As figuras 4.10(d) e 4.10(e) dizem respeito aos valores das distâncias de Cook e das *leverages*, respetivamente, onde as linhas horizontais de cor verde representam o valor de corte para cada uma das medidas. Existem 29341 observações com

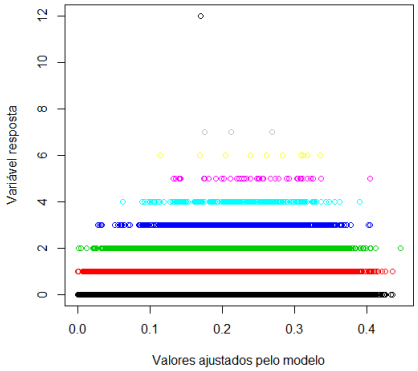
[1]sugerido pela coorientadora Prof.ª Doutora Ana Rita Gaio



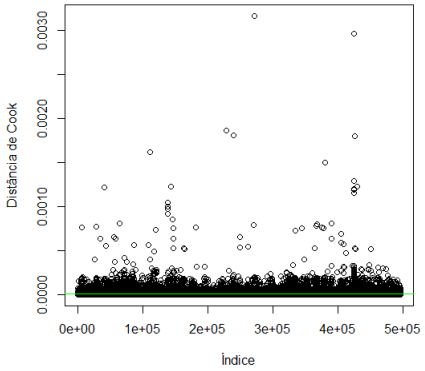
(a) Boxplot dos Resíduos



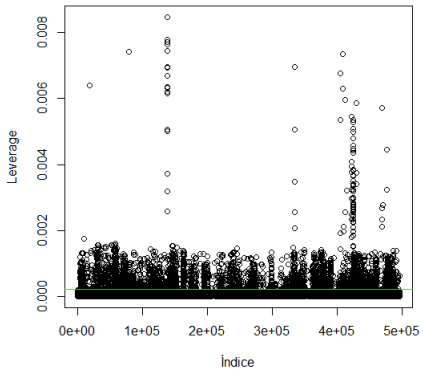
(b) Valores ajustados *versus* Resíduos



(c) Valores ajustados *versus* Valores observados



(d) Distâncias de Cook



(e) Leverages

Figura 4.10: Gráficos de diagnóstico do Modelo de Regressão de Poisson

valores de *leverage* acima do valor de corte (=0.00021) e 23370 observações com valores de distância de Cook acima do valor de corte (=8.05e-06). As tabelas 4.20 e 4.21 permitem discriminar estas observações por número de acidentes.

Número de Acidentes	0	1	2	3	4	5	6
Número de Observações	24313	4216	680	103	22	6	1

Tabela 4.20: Observações com *leverage* superior ao valor de corte (Modelo de regressão de Poisson)

Número de Acidentes	0	1	2	3	4	5	6	7	12
Número de Observações	189	12044	9488	1364	228	43	10	3	1

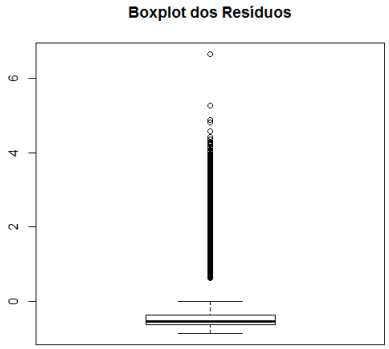
Tabela 4.21: Observações com distância de Cook superior ao valor de corte (Modelo de regressão de Poisson)

A possível existência de sobredispersão no modelo em causa, juntamente com os maus resultados retirados da análise gráfica remete-nos para o ajustamento aos dados de um modelo de regressão Binomial Negativa. Recorrendo à função *glm.nb()* da biblioteca *MASS* do R, o ajustamento aos dados em análise foi feito de forma análoga ao ajustamento do modelo de regressão de Poisson. Outra hipótese seria recorrer a um modelo Quasi-Poisson. No entanto, visto que o modelo de Quasi-Poisson não tem uma função de probabilidade específica, não existe função de verosimilhança, logo não é possível recorrer a procedimentos de seleção automática de variáveis explicativas (**?**)), o que torna o processo de ajustamento muito exigente computacionalmente. No modelo de Quasi-Poisson, é apenas estabelecida uma relação entre a média e a variância através de uma função de variância que inclui um factor multiplicativo conhecido como o parâmetro de sobredispersão.

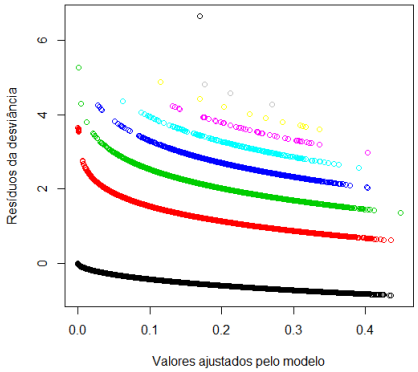
O ajustamento do modelo de regressão Binomial Negativa foi efetuado de maneira análoga ao ajustamento do modelo de regressão de Poisson. Testou-se a significância estatística das interações supramencionadas e utilizou-se um critério de seleção automática para escolher as variáveis significativas. Remete-se para anexo o resultado final do modelo de regressão Binomial Negativa.

Verificou-se que o conjunto de variáveis explicativas estatisticamente significativas coincide com o respetivo conjunto do modelo de regressão de Poisson. Verifica-se também que o valor dos parâmetros de regressão estimados são praticamente iguais em ambos os modelos. No entanto, o erro padrão é ligeiramente superior no modelo de regressão Binomial Negativa, visto que este modelo tem em conta a heterogeneidade presente e, portanto, multiplica os erros padrão pela raiz quadrada de $(1 + \alpha\mu)$ (**?**)). Por esta razão, os valores dos riscos relativos são apresentados na tabela seguinte, sendo a sua interpretação análoga à realizada para o modelo de regressão de Poisson. O modelo de regressão Binomial Negativa apresentou um AIC e um BIC de 484151 e 484729, respetivamente. Concluímos assim, com base tanto no AIC como no BIC, que este modelo sugere um melhor ajustamento aos dados. Nos modelos que se seguem, iremos de novo calcular os valores de AIC e de BIC, embora não os vamos utilizar para escolher os modelos uma vez que o conjunto das observações a que eles são aplicados nunca é o mesmo.

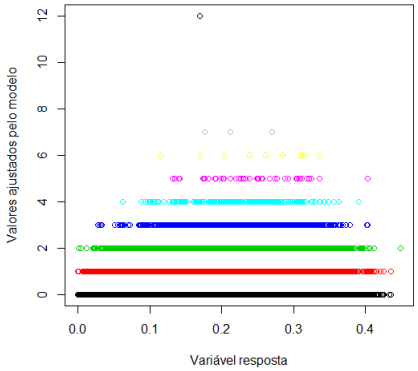
Tal como no modelo de regressão de Poisson, elaboraram-se gráficos de diagnóstico para o modelo de regressão Binomial Negativa, com o intuito de perceber se o ajustamento aos dados é bom ou mau.



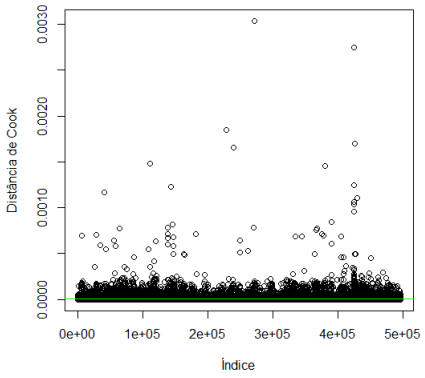
(a) Boxplot dos Resíduos



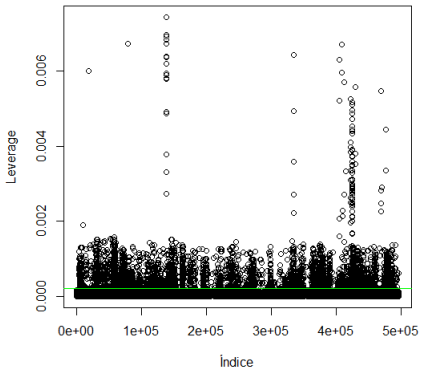
(b) Valores ajustados *versus* Resíduos



(c) Valores ajustados *versus* Valores observados



(d) Distâncias de Cook



(e) Leverages

Figura 4.11: Gráficos de diagnóstico do Modelo de Regressão Binomial Negativa

Tomando como ponto de corte o valor 3, verifica-se que existem 402 observações com resíduo acima deste valor. Além disso, estas observações apresentam todas pelo menos um acidente. A análise também revelou que todos as observações com pelo menos 4

acidentes apresentaram resíduos superiores a 3. As observações com 0 acidentes apresentaram todas um valor absoluto de resíduo da desviância inferior a 3. As conclusões retiradas da análise dos gráficos de diagnóstico são análogas às dos gráficos de diagnóstico do modelo de regressão de Poisson, o que se pode verificar pela semelhança entre as imagens. Relativamente às distâncias de Cook, 20139 observações apresentaram um valor superior ao valor de corte (=8.05e-06), sendo que todas as observações com mais de 2 acidentes fazem parte desse conjunto. No que diz respeito às *leverages*, 24067 observações têm *leverage* alta (valor de corte = 0.00021), sendo 82% indivíduos com 0 acidentes.

Número de Acidentes	0	1	2	3	4	5	6
Número de Observações	20177	3282	517	70	16	4	1

Tabela 4.22: Observações com *leverage* superior ao valor de corte (Modelo de regressão Binomial Negativa)

Número de Acidentes	0	1	2	3	4	5	6	7	12
Número de Observações	184	8818	9488	1364	228	43	10	3	1

Tabela 4.23: Observações com distância de Cook superior ao valor de corte (Modelo de regressão Binomial Negativa)

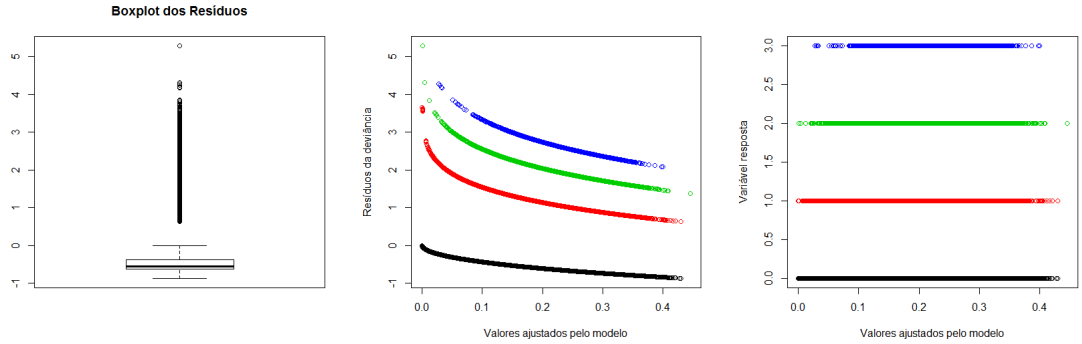
O facto de existir um número elevado de observações com elevada repercussão e influência no modelo levou à consideração de uma abordagem diferente no que diz respeito ao ajustamento dos dados. O ajustamento de um modelo sem estas observações acarreta uma perda grave de informação já que se estaria a retirar uma grande percentagem de indivíduos com pelo menos um acidente, aumentando assim a percentagem de contagens nulas. Decidiu-se então agregar o número de acidentes igual ou superior a 3 numa só classe, como se pode observar na tabela 4.24.

0	1	2	>=3
421738 (84.8%)	64214 (12.9%)	9488 (1.9%)	1649 (0.3%)

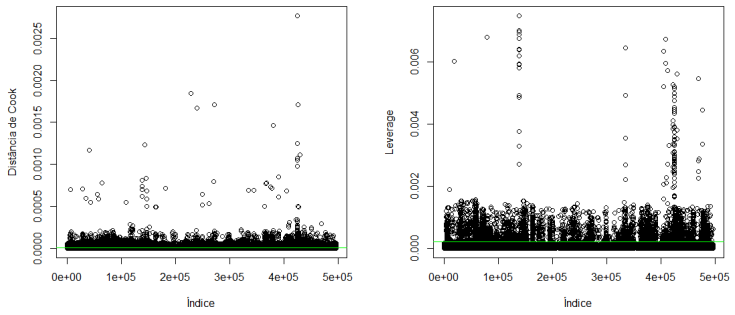
Tabela 4.24: Frequências absoluta e relativa da resposta *Número de Acidentes* com 4 classes

O ajustamento de um modelo de regressão de Poisson continuou a sugerir uma pequena sobredispersão ($\phi = 1.21$), tal como o primeiro modelo de regressão de Poisson, pelo que se prosseguiu com o ajustamento de um modelo de regressão Binomial Negativa, o qual pode ser consultado na tabela em anexo.

A análise do segundo modelo de regressão Binomial Negativa além de confirmar a igualdade de variáveis explicativas selecionadas, revela valores de parâmetros de regressão bastante semelhantes. Para este modelo obteve-se um AIC de 482751 e um BIC de 483329. A figura 4.12 ilustra os gráficos de diagnóstico relativos ao segundo modelo de regressão Binomial Negativa:



(a) Boxplot dos Resíduos (b) Valores ajustados *versus* Resíduos (c) Valores ajustados *versus* Valores observados



(d) Distâncias de Cook (e) Leverages

Figura 4.12: Gráficos de diagnóstico do Segundo Modelo de Regressão Binomial Negativa

As observações com 0 acidentes registaram um valor absoluto de resíduo da desviância inferior a 3. Das observações com mais de 0 acidentes, 264 têm correspondentes resíduos da desviância superior a 3 (em valor absoluto). Destas, 11 têm 1 acidente, 42 têm 2 acidentes e 209 têm 3 acidentes. Pelas cores da figura 4.12(b), pode-se observar que a magnitude do resíduo está associada com a magnitude da variável resposta, isto é, verifica-se que quanto maior é o número de acidentes, maior é o valor dos resíduos da desviância. Tal como nos modelos anteriores, o segundo modelo de regressão Binomial Negativa não prevê para nenhuma observação um número de acidentes maior do que 0. O número de observações com valores de distância de Cook e *leverage* superior aos valores de corte (8.05e-06 e 0.00021, respetivamente) vai de encontro ao que já foi observado nos anteriores modelos, sendo de 20561 e 24133, respetivamente.

Número de Acidentes	0	1	2	3
Número de Observações	22634	3788	599	112

Tabela 4.25: Observações com *leverage* superior ao valor de corte (Segundo Modelo de regressão Binomial Negativa)

Número de Acidentes	0	1	2	3
Número de Observações	184	9240	9488	1649

Tabela 4.26: Observações com distância de Cook superior ao valor de corte (Segundo Modelo de regressão Binomial Negativa)

Atentando às tabelas 4.25 e 4.26, verifica-se que o número de observações cujas distância de Cook e *leverage* são superiores ao valor de corte é elevado, impossibilitando a sua análise detalhada, tal como nos modelos anteriores. Assim, optou-se por apenas retirar as observações com resíduos elevados e ajustar um novo modelo. O terceiro modelo de regressão Binomial Negativa ajustado é expresso na tabela 4.27:

Tabela 4.27: Output do Terceiro Modelo de Regressão Binomial Negativa

Variável	**Coeficiente**	**Erro padrão**	**Valor-p**
constante	-7.63	0.028	$<$ 2e-16
I_{bmw}	0.073	0.021	6.11e-04
$I_{citroen}$	-0.221	0.027	3.62e-16
I_{fiat}	-0.122	0.026	4.27e-06
I_{ford}	-0.146	0.024	1.5e-09
I_{honda}	-0.144	0.04	3.63e-04
$I_{mercedes-benz}$	-0.107	0.017	8.21e-10
I_{nissan}	-0.069	0.021	1.23e-03
I_{opel}	-0.156	0.019	$<$ 2e-16
I_{outra}	-0.153	0.015	$<$ 2e-16
$I_{peugeot}$	-0.139	0.02	9.72e-12
$I_{renault}$	-0.09	0.019	6.71e-07
I_{seat}	-0.2	0.028	1.51e-13
I_{toyota}	-0.192	0.025	6.41e-16
$I_{volkswagen}$	-0.139	0.018	1.55e-15
$I_{carrocaria}$	-0.106	0.01	$<$ 2e-16
I_{cil2}	0.008	0.017	0.649
I_{cil3}	0.117	0.044	0.008
$I_{cavalos2}$	-0.014	0.032	0.68
$I_{cavalos3}$	0.425	0.138	0.004
I_{caixa}	-0.046	0.01	3.05e-05
I_{vel}	-0.075	0.01	4.17e-11
I_{portas}	0.055	0.009	7.96e-10
$I_{gasolina}$	-0.201	0.017	$<$ 2e-16
I_{peso2}	0.019	0.013	0.169
I_{peso3}	0.05	0.017	0.004
I_{iv2}	-0.023	0.02	0.28

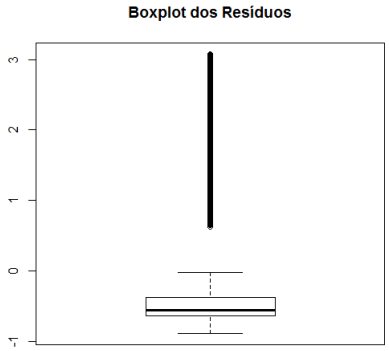
Continua na página seguinte

Tabela 4.27 – *Continuação da página anterior*

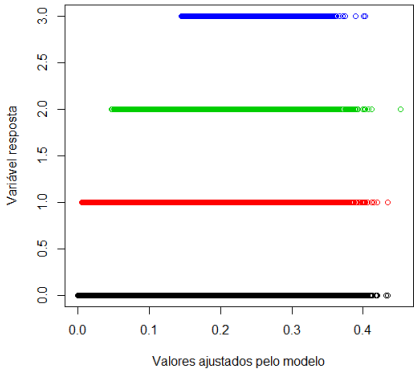
Variável	**Coeficiente**	**Erro padrão**	**Valor-p**
I_{iv3}	-0.131	0.019	1.85e-10
I_{iv4}	-0.124	0.018	1.01e-12
I_{iv5}	-1.107	0.102	$<$ 2e-16
I_{is2}	-0.03	0.021	0.187
I_{is3}	-0.11	0.019	1.22e-10
I_{is4}	-0.27	0.022	$<$ 2e-16
$I_{iv2} \times I_{is2}$	-0.067	0.033	0.043
$I_{iv3} \times I_{is2}$	-0.043	0.024	0.181
$I_{iv4} \times I_{is2}$	-0.178	0.025	3.05e-10
$I_{iv5} \times I_{is2}$	0.007	0.147	0.965
$I_{iv2} \times I_{is3}$	-0.091	0.029	0.002
$I_{iv3} \times I_{is3}$	-0.116	0.026	5.22e-05
$I_{iv4} \times I_{is3}$	-0.248	0.024	$<$ 2e-16
$I_{iv5} \times I_{is3}$	-0.192	0.138	0.191
$I_{iv2} \times I_{is4}$	-0.049	0.031	0.127
$I_{iv3} \times I_{is4}$	-0.131	0.027	3.01e-05
$I_{iv4} \times I_{is4}$	-0.359	0.03	$<$ 2e-16
$I_{iv5} \times I_{is4}$	-0.293	0.139	0.038
$I_{cil2} \times I_{cavalos2}$	0.052	0.022	0.095
$I_{cil3} \times I_{cavalos2}$	0.083	0.055	0.132
$I_{cil2} \times I_{cavalos3}$	-0.326	0.147	0.027
$I_{cil3} \times I_{cavalos3}$	-0.436	0.152	0.004
$I_{cavalos2} \times I_{gasolina}$	0.073	0.026	0.005
$I_{cavalos3} \times I_{gasolina}$	0.101	0.02	3.42e-06

A remoção dos resíduos não aparenta introduzir grandes diferenças ao nível do modelo, visto que os parâmetros das variáveis não registaram alterações substanciais. O AIC e o BIC do terceiro modelo de regressão Binomial Negativa são iguais a 478933 e a 479511, respetivamente.

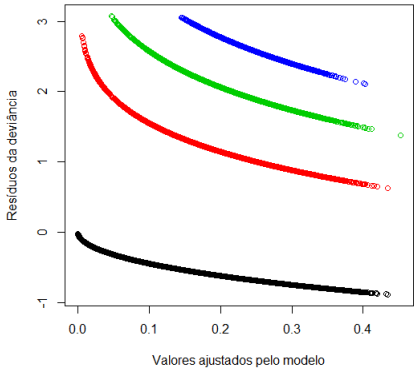
Os gráficos de diágnostico do terceiro modelo de regressão Binomial Negativa podem ser visualizados na figura 4.13:



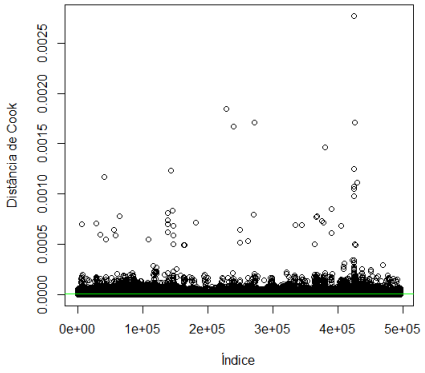
(a) Boxplot dos Resíduos



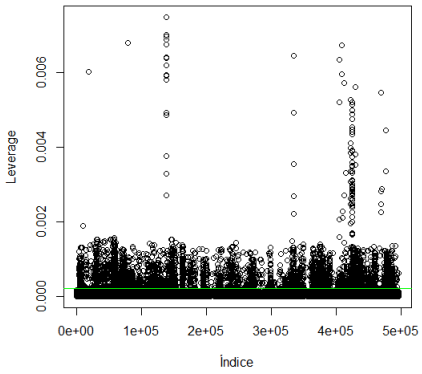
(b) Valores ajustados *versus* Resíduos



(c) Valores ajustados *versus* Valores observados



(d) Distâncias de Cook



(e) Leverages

Figura 4.13: Gráficos de diagnóstico do Terceiro Modelo de Regressão Binomial Negativa

Da análise dos gráficos, conclui-se que o modelo, tal como os anteriores, nunca regista um valor superior a 0.5 para as observações. Verifica-se também que não existem observações com resíduos superiores a 3, em valor absoluto. No entanto, e como já se tinha verificado

nas análises gráficas anteriores, as observações com um número de acidentes diferente de 0 são aquelas que apresentam resíduos mais elevados. Relativamente às observações com valores de distância de Cook e *leverage* elevada (os valores de corte obtidos são iguais a 8.057e-06 e 0.00021, respetivamente), as tabelas 4.28 e 4.29 permitem fazer a sua discriminação por número de acidentes, a partir das quais se conclui que o comportamento destas se revela semelhante às dos modelos anteriores. Assim, as observações com maior repercussão nos valores ajustados são observações com 0 acidentes, enquanto que observações com 1 e 2 acidentes aparentam ser as mais influentes sobre os parâmetros estimados.

Número de Acidentes	0	1	2	3
Número de Observações	23178	3355	510	84

Tabela 4.28: Observações com *leverage* superior ao valor de corte (Terceiro Modelo de regressão Binomial Negativa)

Número de Acidentes	0	1	2	3
Número de Observações	184	9933	9446	1440

Tabela 4.29: Observações com distância de Cook superior ao valor de corte (Terceiro Modelo de regressão Binomial Negativa)

Verifica-se assim que as tentativas realizadas para obter um modelo com um melhor ajustamentos aos dados não trouxeram resultados positivos.

4.3 Regressão Logística

Na análise exploratória do número de sinistros, observou-se que cerca de 85% dos dados dizem respeito a indivíduos sem acidentes. Os restantes 15% são observações que registaram pelo menos um acidente. Dado que os modelos anteriores não revelaram um ajustamento bom aos dados, decidiu-se utilizar um modelo de regressão logística. A variável resposta passa assim a ser classificada em 0 (ausência de acidente) e em 1 (presença de acidente). O *output* do modelo de regressão logística escolhido é expresso na tabela 4.30:

Tabela 4.30: Output do Modelo de Regressão Logística

Variável	**Coeficiente**	**Erro padrão**	**Valor-p**
constante	-1.86	0.029	$<$ 2e-16
I_{bmw}	0.04	0.024	0.067
$I_{citroen}$	-0.246	0.03	$<$ 2e-16
I_{fiat}	-0.126	0.029	1.42e-06

Continua na página seguinte

Tabela 4.30 – *Continuação da página anterior*

Variável	**Coeficiente**	**Erro padrão**	**Valor-p**
I_{ford}	-0.152	0.027	1.45e-09
I_{honda}	-0.164	0.044	2.06-04
$I_{mercedes-benz}$	-0.114	0.019	2.08e-10
I_{nissan}	-0.044	0.023	0.06
I_{opel}	-0.15	0.021	$<$ 2e-16
I_{outra}	-0.15	0.017	$<$ 2e-16
$I_{peugeot}$	-0.145	0.023	1.85e-12
$I_{renault}$	-0.106	0.021	3.18e-07
I_{seat}	-0.19	0.031	1.21e-13
I_{toyota}	-0.199	0.027	2.61e-16
$I_{volkswagen}$	-0.14	0.02	5.32e-15
$I_{carrocaria}$	-0.09	0.01	$<$ 2e-16
I_{cil2}	0.003	0.016	0.037
I_{cil3}	0.098	0.021	1.71e-06
$I_{cavalos2}$	0.073	0.019	1.67e-04
$I_{cavalos3}$	0.056	0.019	0.004
I_{caixa}	-0.046	0.012	9.00e-05
I_{vel}	-0.079	0.012	2.46e-11
I_{portas}	0.052	0.01	4.14e-10
$I_{gasolina}$	-0.215	0.019	$<$ 2e-16
I_{iv2}	0.032	0.023	0.159
I_{iv3}	-0.048	0.022	0.031
I_{iv4}	-0.098	0.02	1.43e-06
I_{iv5}	-0.974	0.11	$<$ 2e-16
I_{is2}	0.798	0.026	$<$ 2e-16
I_{is3}	0.716	0.023	$<$ 2e-16
I_{is4}	0.523	0.027	$<$ 2e-16
$I_{iv2} \times I_{is2}$	-0.108	0.038	0.004
$I_{iv3} \times I_{is2}$	-0.117	0.036	0.001
$I_{iv4} \times I_{is2}$	-0.222	0.034	5.96e-10
$I_{iv5} \times I_{is2}$	-0.15	0.162	0.353
$I_{iv2} \times I_{is3}$	-0.142	0.033	1.70e-05
$I_{iv3} \times I_{is3}$	-0.195	0.032	1.12e-05
$I_{iv4} \times I_{is3}$	-0.295	0.03	$<$ 2e-16
$I_{iv5} \times I_{is3}$	-0.342	0.15	0.023
$I_{iv2} \times I_{is4}$	-0.081	0.036	0.025
$I_{iv3} \times I_{is4}$	-0.203	0.034	5.98e-09
$I_{iv4} \times I_{is4}$	-0.359	0.033	$<$ 2e-16
$I_{iv5} \times I_{is4}$	-0.401	0.144	0.005
$I_{cavalos2} \times I_{gasolina}$	0.033	0.026	0.217
$I_{cavalos3} \times I_{gasolina}$	0.139	0.02	6.40e-10

A tabela 4.30 evidencia que, ao contrário dos modelos anteriores, a variável *peso* e a interação entre as variáveis *cilindrada* e *cavalos* não mostraram ser estatisticamente significativas. Relativamente ao AIC e BIC, o modelo de regressão logística apresentou valores de 415996 e 416495, respetivamente.

A figura 4.14(a) evidencia uma separação clara em 2 grupos. De facto, o grupo identificado como *outliers* no *boxplot* corresponde às observações em que se verificaram acidentes, enquanto o conjunto cujos resíduos são em módulo menor do que 1 diz respeito às ausências de acidentes.

Analisando a figura 4.14(b), verifica-se que as probabilidades de sucesso, ou seja, probabilidade de uma observação ter um acidente, são bastante próximas da probabilidade empírica calculada a partir da amostra ($\hat{P}(Y = 1) \approx 15\%$). Relativamente às probabilidades estimadas de sucesso, o valor máximo e o valor mínimo são iguais a 0.297 e 0.034, respetivamente. Assim, o modelo prevê a ausência de acidente para todas as observações, o que revela um mau ajustamento aos dados.

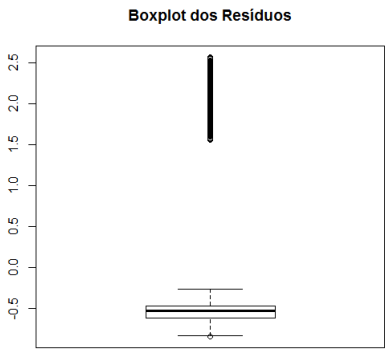
A figura 4.14(c) representa, a vermelho, os valores ajustados pelo modelos e os resíduos da desviância das observações onde se verificou um acidente. Como seria de esperar, quanto menor é o valor ajustado pelo modelo, maior é o valor do resíduo já que o valor observado é igual a 1. Os pontos assinalados a preto tratam-se das observações com 0 acidentes. Para estes pontos, quanto menor é o valor ajustado pelo modelo, menor será o resíduo.

A análise às distâncias de Cook revelou que todas as observações cujo valor era superior ao valor de corte (=8.053e-06) se tratam de casos de presença de acidentes, perfazendo 69% do total. Destas, 404 observações são facilmente destacadas no gráfico 4.14(d).

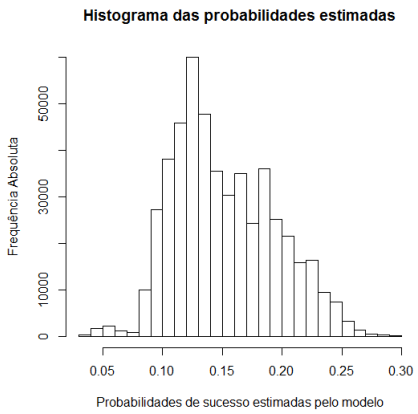
Relativamente às *leverages*, a figura 4.14(e) sugere uma divisão em 3 grupos. No entanto, o número de observações cuja *leverage* ultrapassa o valor de corte (=0.00018) é demasiadamente elevado (= 19024) para permitir uma análise mais detalhada. A tabela 4.31 discrimina o número de observações com *leverage* elevada por ausência e presença de acidentes.

Acidentes	0 (Ausência)	1 (Presença)
Número de Observações	16243	2781

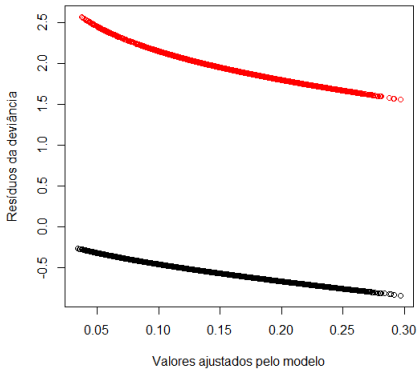
Tabela 4.31: Observações com *leverage* superior ao valor de corte (Modelo de Regressão Logística)



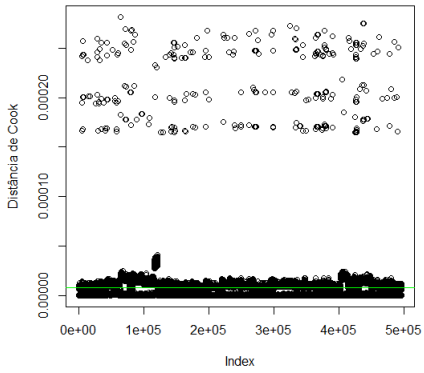
(a) Boxplot dos Resíduos



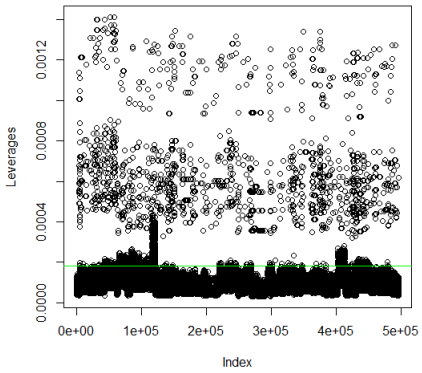
(b) Histograma das probabilidades estimadas de sucesso



(c) Valores ajustados *versus* Resíduos



(d) Distâncias de Cook



(e) Leverages

Figura 4.14: Gráficos de diagnóstico do Modelo de Regressão Logística

4.4 Modelos para dados com um número excessivo de zeros

Nesta secção vai-se proceder ao ajustamento dos dados a modelos de contagem com um número excessivo de zeros. Estes modelos são os modelos com zeros inflacionados e os modelos com barreira.

Ao contrário do que foi efetuado para os modelos anteriores, visto que os modelos de contagem com um número excessivo de zeros combinam 2 modelos, um modelo binomial e um modelo de contagem, ou seja, há um número maior de parâmetros em questão, o estudo sobre as interações existentes entre variáveis explicativas não foi efetuado pois verificou-se uma elevada exigência computacional. Além disso, verificou-se que, quando se introduziam interações no modelo, o tempo computacional necessário para obter um modelo aumentava drasticamente, não sendo possível a sua realização em tempo útil.

Devido ao elevado número de observações, a análise gráfica da previsão do modelo recai sobre 25 observações representativas dos diferentes números de acidentes. Em anexo, é apresentada a informação relativa a estas observações.

4.4.1 Modelos com zeros inflacionados

Como já foi referido, os modelos com zeros inflacionados fazem distinção entre os zeros falsos e os zeros verdadeiros. Neste problema, um zero falso diz respeito aos acidentes que ocorrem mas que não são reportados (uma pessoa pode ter um acidente e, para evitar que o valor do prémio suba, não o reportar, por exemplo) e um zero verdadeiro corresponde à situação em que realmente não houve nenhum acidente.

A implementação do modelo de regressão de Poisson com zeros inflacionados (Modelo ZIP) é feita com recurso à função *zeroinfl()* da biblioteca *pscl*.

Tabela 4.32: Output do Modelo ZIP

Variável	**Coeficiente**	**Erro padrão**	**Valor-p**
Modelo de Contagem			
Constante	-7.331	0.036	$<$ 2e-16
I_{bmw}	0.108	0.039	0.006
$I_{citroen}$	0.002	0.052	0.965
I_{fiat}	-0.1	0.053	0.06
I_{ford}	-0.1	0.047	0.032
I_{honda}	-0.129	0.083	0.121
$I_{mercedes-benz}$	0.005	0.035	0.89
I_{nissan}	-0.132	0.04	0.001

Continua na página seguinte

Tabela 4.32 – *Continuação da página anterior*

Variável	**Coeficiente**	**Erro padrão**	**Valor-p**
I_{opel}	-0.141	0.037	0.001
I_{outra}	-0.104	0.03	0.001
$I_{peugeot}$	0.005	0.039	0.894
$I_{renault}$	-0.001	0.035	0.968
I_{seat}	-0.099	0.056	0.08
I_{toyota}	-0.036	0.05	0.467
$I_{volkswagen}$	-0.067	0.037	0.07
$I_{carrocaria}$	-0.047	0.017	0.01
$I_{cavalos2}$	0.086	0.013	8.94e-12
$I_{cavalos3}$	0.128	0.014	$<$ 2e-16
I_{lug}	-0.057	0.022	0.012
$I_{gasolina}$	-0.138	0.009	$<$ 2e-16
I_{iv2}	-0.103	0.022	3.79e-06
I_{iv3}	-0.168	0.022	3.31e-14
I_{iv4}	-0.159	0.022	2.32e-13
I_{iv5}	-0.445	0.128	0.004
I_{is2}	0.025	0.022	0.27
I_{is3}	-0.084	0.02	3.16e-05
I_{is4}	-0.16	0.022	9.31e-14
Modelo de Zeros			
Constante	-7.181	0.1	$<$ 2e-16
I_{bmw}	0.082	0.098	0.404
$I_{citroen}$	0.541	0.11	9.54e-07
I_{fiat}	0.044	0.128	0.733
I_{ford}	0.116	0.123	0.344
I_{honda}	-0.021	0.239	0.928
$I_{mercedes-benz}$	0.238	0.087	0.006
I_{nissan}	-0.176	0.115	0.123
I_{opel}	0.027	0.093	0.774
I_{outra}	0.103	0.075	0.168
$I_{peugeot}$	0.367	0.093	8.24e-05
$I_{renault}$	0.212	0.084	0.012
I_{seat}	0.211	0.143	0.141
I_{toyota}	0.401	0.114	0.002
$I_{volkswagen}$	0.163	0.09	0.07
$I_{carrocaria}$	0.137	0.045	0.003
I_{cil2}	-0.121	0.031	9.39e-05
I_{cil3}	-0.184	0.043	2.08e-05
I_{caixa}	0.128	0.025	5.01e-07
I_{vel}	0.207	0.026	5.51e-16
I_{portas}	-0.113	0.023	1.35e-06

Continua na página seguinte

Tabela 4.32 – *Continuação da página anterior*

Variável	**Coeficiente**	**Erro padrão**	**Valor-p**
I_{lug}	-0.196	0.056	0.001
I_{trac}	-0.104	0.035	0.003
I_{peso2}	-0.068	0.033	0.036
I_{peso3}	-0.041	0.043	0.347
I_{iv2}	-0.139	0.062	0.025
I_{iv3}	0.047	0.059	0.425
I_{iv4}	0.366	0.054	1.56e-11
I_{iv5}	1.366	0.183	9.15e-14
I_{is2}	-0.042	0.063	0.506
I_{is3}	0.009	0.059	0.873
I_{is4}	0.279	0.058	1.83e-06

O cálculo do parâmetro de dispersão não parece indicar sobredispersão, já que $\phi = 1.05$. O modelo ZIP apresentou um AIC de 484216 e um BIC de 484872.

Através do comando *predict(zip, type = "zero")* no R (onde *zip* diz respeito ao modelo em questão), obtém-se a probabilidade π de existir um zero falso para cada observação. A tabela seguinte apresenta o valor dessa probabilidade para as 25 observações selecionadas.

1	**2**	**3**	**4**	**5**
0.414	0.452	0.261	0.487	0.283
6	**7**	**8**	**9**	**10**
0.429	0.529	0.349	0.617	0.547
11	**12**	**13**	**14**	**15**
0.173	0.439	0.374	0.335	0.225
16	**17**	**18**	**19**	**20**
0.412	0.439	0.227	0.542	0.348
21	**22**	**23**	**24**	**25**
0.366	0.191	0.469	0.365	0.209

Tabela 4.33: Probabilidade da observação i, i = 1, ..., 25, ser um zero falso (Modelo ZIP)

Os valores obtidos pela tabela 4.33 indicam que as observações 7, 9, 10 e 19 apresentam uma probabilidade de serem um zero falso superior a 0.5. No entanto, estas observações dizem respeito a casos onde existem acidentes.

A probabilidade de existir um zero, seja ele um zero verdadeiro ou um zero falso, para cada observação é dada pelo comando *predict(zip, type = "prob")[,1]*. O valor desta probabilidade é apresentado na tabela 4.34.

1	**2**	**3**	**4**	**5**
0.826	0.831	0.825	0.844	0.849
6	**7**	**8**	**9**	**10**
0.829	0.847	0.866	0.899	0.869
11	**12**	**13**	**14**	**15**
0.918	0.838	0.808	0.777	0.829
16	**17**	**18**	**19**	**20**
0.813	0.819	0.847	0.836	0.752
21	**22**	**23**	**24**	**25**
0.749	0.831	0.828	0.788	0.834

Tabela 4.34: Probabilidade da observação i, i = 1, ..., 25, ser um zero (Modelo ZIP)

A tabela 4.32 evidencia o mau ajustamento do modelo ao prever que todas as observações apresentam uma elevada probabilidade de ser zero. Para a primeira linha da tabela, este resultado é correcto pois tratam-se de viaturas sem acidentes. No entanto, todas as outras observações registaram pelo menos um acidente, o que reflete a má adequação do modelo ZIP.
A média prevista para a distribuição condicionada $Y|x = x_i, i = 1, \ldots, 25$ pode ser obtida recorrendo ao comando *fitted(zip, type = "response")*.

1	**2**	**3**	**4**	**5**
0.207	0.203	0.199	0.187	0.169
6	**7**	**8**	**9**	**10**
0.204	0.186	0.151	0.116	0.154
11	**12**	**13**	**14**	**15**
0.086	0.191	0.231	0.272	0.192
16	**17**	**18**	**19**	**20**
0.226	0.218	0.171	0.203	0.312
21	**22**	**23**	**24**	**25**
0.321	0.189	0.209	0.258	0.186

Tabela 4.35: Média prevista para a distribuição condicionada $Y|x = x_i, i = 1, ..., 25$ (Modelo ZIP)

O resultado do modelo binomial com o modelo de contagens, ou seja, o modelo de Poisson com zeros inflacionados é apresentado na figura seguinte. Em cada gráfico está representada a função de probabilidade $Y|x = x_i$, para as 25 observações. Na figura 4.15 é possível observar um ponto representado por uma cruz vermelha que corresponde ao número de acidentes sofridos pela observação.

Figura 4.15: Função de probabilidade $Y|x = x_i, i = 1,...,25$ (Modelo ZIP)

4.4.2 Modelos com barreira

Ao contrário dos modelos com zeros inflacionados, os modelos com barreira não fazem distinção entre zeros verdadeiros e zeros falsos, considerando apenas zeros e não zeros. Assim, o processo de contagem é truncado em zero. No R, a implementação dos modelos com barreira é efetuada recorrendo à função *hurdle()*, que, tal como a função *zeroinfl()*, pertence à biblioteca *pscl*. Para uma maior clareza, entitule-se de modelo Hurdle o modelo de regressão de Poisson com barreira.

Tabela 4.36: Output do Modelo Hurdle

Variável	**Coeficiente**	**Erro padrão**	**Valor-p**
Modelo de Contagem			
Constante	-7.056	0.045	$<$ 2e-16
I_{bmw}	0.099	0.046	0.033
$I_{citroen}$	-0.058	0.063	0.365
I_{fiat}	-0.125	0.065	0.055
I_{ford}	-0.054	0.054	0.321
I_{honda}	-0.039	0.093	0.678
$I_{mercedes-benz}$	0.048	0.038	0.211
I_{nissan}	-0.058	0.047	0.226
I_{opel}	-0.096	0.044	0.028
I_{outra}	-0.098	0.035	0.006
$I_{peugeot}$	0.005	0.046	0.918
$I_{renault}$	0.032	0.041	0.439
I_{seat}	-0.133	0.068	0.051
I_{toyota}	-0.014	0.056	0.798
$I_{volkswagen}$	-0.066	0.042	0.119
$I_{cavalos2}$	0.057	0.028	0.041
$I_{cavalos3}$	0.141	0.029	2.24e-06
I_{vel}	-0.052	0.023	0.028
I_{lug}	-0.079	0.025	0.002
$I_{gasolina}$	-0.128	0.018	1.51e-11
I_{iv2}	-0.176	0.024	1.73e-12
I_{iv3}	-0.273	0.025	$<$ 2e-16
I_{iv4}	-0.32	0.025	$<$ 2e-16
I_{iv5}	-0.544	0.141	0.002
I_{is2}	-0.23	0.025	$<$ 2e-16
I_{is3}	-0.35	0.023	$<$ 2e-16
I_{is4}	-0.401	0.024	$<$ 2e-16
Modelo de Zeros			
Constante	-7.588	0.024	$<$ 2e-16
I_{bmw}	0.077	0.024	0.001

Continua na página seguinte

Tabela 4.36 – *Continuação da página anterior*

Variável	**Coeficiente**	**Erro padrão**	**Valor-p**
$I_{citroen}$	-0.244	0.03	7.23e-16
I_{fiat}	-0.116	0.029	7.52e-05
I_{ford}	-0.163	0.026	1.10e-09
I_{honda}	-0.156	0.044	4.38e-04
$I_{mercedes-benz}$	-0.118	0.019	3.18e-09
I_{nissan}	-0.095	0.023	4.68e-05
I_{opel}	-0.167	0.02	8.89e-16
I_{outra}	-0.159	0.017	$<$ 2e-16
$I_{peugeot}$	-0.163	0.022	1.17e-12
$I_{renault}$	-0.11	0.02	1.51e-07
I_{seat}	-0.188	0.031	2.97e-09
I_{toyota}	-0.238	0.027	$<$ 2e-16
$I_{volkswagen}$	-0.14	0.02	7.41e-12
$I_{carrocaria}$	-0.112	0.011	$<$ 2e-16
I_{cil2}	0.054	0.014	2.58e-04
I_{cil3}	0.094	0.02	5.32e-06
$I_{cavalos2}$	0.106	0.013	3.86e-14
$I_{cavalos3}$	0.14	0.016	$<$ 2e-16
I_{caixa}	-0.055	0.011	2.57e-06
I_{vel}	-0.089	0.012	1.08e-13
I_{portas}	0.051	0.01	1.37e-06
I_{lug}	0.031	0.013	0.019
$I_{gasolina}$	-0.139	0.01	$<$ 2e-16
I_{trac}	0.049	0.015	0.001
I_{iv2}	-0.028	0.013	0.029
I_{iv3}	-0.171	0.012	$<$ 2e-16
I_{iv4}	-0.299	0.012	$<$ 2e-16
I_{iv5}	-1.219	0.052	$<$ 2e-16
I_{is2}	-0.084	0.012	1.64e-11
I_{is3}	-0.218	0.010	$<$ 2e-16
I_{is4}	-0.443	0.011	$<$ 2e-16

O parâmetro de dispersão para o modelo Hurdle é igual a $\phi = 1.11$, o que sugere a inexistência de sobredispersão e os valores de AIC e de BIC obtidos são 484693 e 485360, respetivamente. Comparando o modelo Hurdle com o modelo ZIP, é possível notar algumas diferenças ao nível das variáveis explicativas selecionadas. O modelo de contagem do modelo Hurdle exclui a variável *carroçaria* e acrescenta a variável *lugares*. Por sua vez, a variável *peso* não figura no modelo binomial do modelo Hurdle, sendo adicionadas as variáveis *combustível* e *tração*.

Tal como no modelo ZIP, foram calculadas a probabilidade de existir um zero e a média da distribuição condicionada $Y|x = x_i, i = 1,...,25$, sendo os comandos utilizados para a sua obtenção análogos aos utilizados para o modelo ZIP. Pela definição de modelos com barreira, não existe a probabilidade de se observar zeros falsos.

1	**2**	**3**	**4**	**5**
0.821	0.824	0.842	0.838	0.869
6	**7**	**8**	**9**	**10**
0.819	0.832	0.885	0.888	0.861
11	**12**	**13**	**14**	**15**
0.931	0.838	0.805	0.766	0.823
16	**17**	**18**	**19**	**20**
0.818	0.822	0.865	0.829	0.747
21	**22**	**23**	**24**	**25**
0.747	0.837	0.825	0.787	0.848

Tabela 4.37: Probabilidade da observação i, i = 1, ..., 25, ser um zero (Modelo Hurdle)

1	**2**	**3**	**4**	**5**
0.211	0.204	0.179	0.193	0.145
6	**7**	**8**	**9**	**10**
0.209	0.199	0.129	0.127	0.162
11	**12**	**13**	**14**	**15**
0.074	0.188	0.228	0.303	0.208
16	**17**	**18**	**19**	**20**
0.223	0.216	0.149	0.208	0.314
21	**22**	**23**	**24**	**25**
0.319	0.188	0.212	0.259	0.181

Tabela 4.38: Média prevista para a distribuição condicionada $Y|x = x_i, i = 1,...,25$ (Modelo Hurdle)

As tabelas 4.37 e 4.38 apresentam resultados semelhantes ao modelo ZIP. Para as 25 observações, a probabilidade de se verificar ausência de acidentes é superior a 0.5 e, em concordância, os valores das médias previstas para o número de acidentes encontram-se perto de 0. Assim, verifica-se que o modelo Hurdle prevê os zeros corretamente enquanto que o modelo de contagem, à semelhança do modelo ZIP, não faz um bom ajustamento aos dados. A figura 4.16, que representa a função de probabilidade $Y|x = x_i, i = 1,...,25$, ilustra o mau ajustamento do modelo de contagem.

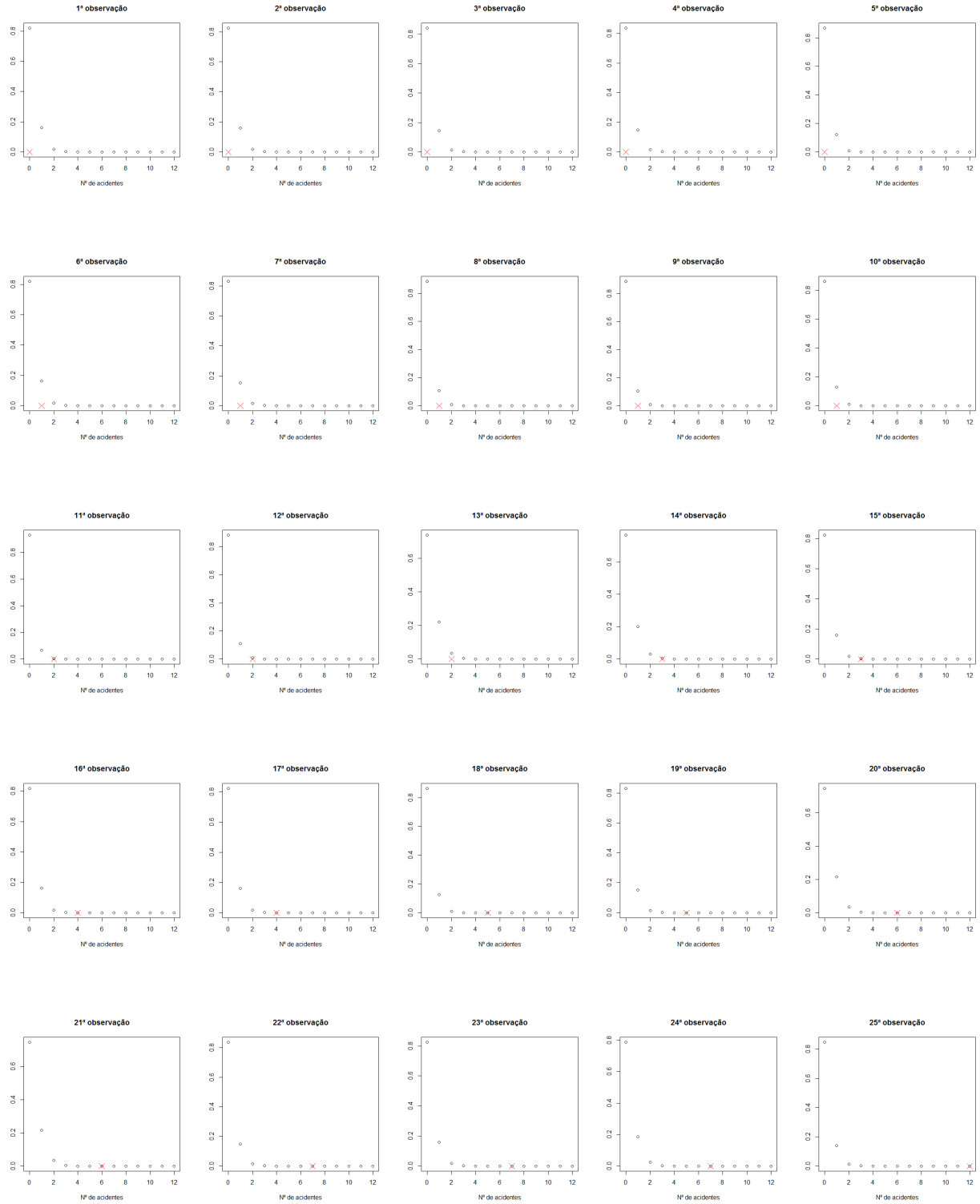


Figura 4.16: Função de probabilidade $Y|x = x_i, i = 1,...,25$ (Modelo Hurdle)

Dado o mau ajustamento do modelo com zeros inflacionados e do modelo com barreira, ajustaram-se novamente estes modelos aos dados onde o número de acidentes está dividido em 4 classes: 0, 1, 2 e >=3. No entanto, a análise dos modelos considerados revelou um mau ajustamento aos dados.

Capítulo 5

Análise do Efeito das Variáveis

Após o ajustamento dos modelos de regressão estudados, neste capítulo procede-se à interpretação do efeito das variáveis explicativas sobre o número de acidentes. Esta interpretação tem como base o cálculo do *odds ratio* e do risco relativo definido nas secções 2.4.1 e 2.4.2, respetivamente. Os modelos escolhidos são o terceiro modelo de regressão Binomial Negativa (para facilitar a leitura, este modelo é intitulado de modelo de regressão Binomial Negativa), o modelo de regressão logística, o modelo ZIP e o modelo Hurdle.

É importante notar que a interpretação realizada tem em conta que existem outros fatores associados à ocorrência de acidentes que não foram tidos em conta nesta tese, como é o caso de fatores humanos e ambientais. Assim, nem sempre é possível analisar o efeito das variáveis de uma forma clara. A conselho do coorientador Dr. Luís Maranhão, é necessária uma certa abstração de fatores externos às viaturas que influenciam os acidentes para que se possam tirar conclusões sobre as características do carro.

5.1 Modelo de Regressão Binomial Negativa

Para o modelo de regressão Binomial Negativa, o número médio de acidentes para um veículo cujas variáveis explicativas se encontram na classe de referência é igual $e^{-7.63} \approx 0$ acidentes.

O risco relativo é primeiro fornecido e interpretado para cada uma das variáveis explicativas na ausência de interação, comparando o efeito de uma categoria com a categoria de referência, pressupondo que as restantes variáveis explicativas se encontram na categoria de referência. São também apresentados os intervalos de confiança a 95% correspondentes.

- Relativamente ao efeito das marcas dos carros sobre o número de acidentes, o resultado obtido foi inesperado pois apenas uma marca (*BMW*) apresentou um efeito positivo significativo sobre o número de acidentes quando comparada com a marca *Audi*. No entanto, os valores dos riscos relativos não aparentam ser muito díspares entre si, sendo bastante próximos de 1.

BMW	Citroën	Fiat	Ford	Honda	Mercedes-Benz	Nissan
1.075	0.802	0.885	0.864	0.866	0.899	0.934
Opel	Outra	Peugeot	Renault	Seat	Toyota	Volkswagen
0.855	0.859	0.871	0.914	0.818	0.826	0.869

Tabela 5.1: Risco Relativo da Variável *marca* (Modelo de Regressão Binomial Negativa)

BMW	Citroën
(1.032,1.121)	(0.761,0.846)
Fiat	Ford
(0.841,0.933)	(0.824,0.906)
Honda	Mercedes-Benz
(0.8,0.937)	(0.868,0.93)
Nissan	Opel
(0.896,0.973)	(0.825,0.887)
Outra	Peugeot
(0.825,0.887)	(0.833,0.885)
Renault	Seat
(0.882,0.948)	(0.774,0.865)
Toyota	Volkswagen
(0.787,0.866)	(0.839,0.901)

Tabela 5.2: Intervalos de Confiança a 95% da Variável *marca* (Modelo de Regressão Binomial Negativa)

- Para os carros com outro tipo de carroçaria são esperados, em média, menos 10% de acidentes do que os carros com uma carroçaria de carro ligeiro de passageiros ($RR = 0.899$ e $IC_{95\%} = (0.882, 0.917)$).

- Para carros com cilindrada superior a 1800 cm^3 esperam-se, em média, mais acidentes quando comparando com carros com cilindrada inferior a 1300 cm^3 ($RR = 1.008$ e $IC_{95\%} = (1.03, 1.226)$). Este efeito positivo e significativo era esperado pois carros com maior capacidade de cilindrada permitem ao condutor deslocar-se a maior velocidade, incorrendo assim num maior risco de ter um acidente.

- Para veículos com mais de 100 cavalos esperam-se, em média, mais 50% de acidentes do que os veículos com menos de 75 cavalos ($RR = 1.53$ e $IC_{95\%} = (1.147, 2.041)$). O segmento da população que adquire carros com menor cavalagem é maioritariamente inexperiente e essa mesma falta de prática de condução leva os condutores desses veículos a serem mais cuidadosos, o que, aliado às grandes velocidades atingidas por carros com mais de 100 cavalos, pode explicar o valor do risco relativo obtido.

- Para viaturas com caixa automática ou semi-automática esperam-se, em média, menos 4.5% de acidentes do que veículos com caixa manual ($RR = 0.955$ e $IC_{95\%} = (0.935, 0.976)$)).

- Para carros que não têm 5 velocidades esperam-se, em média, menos 7.3% de acidentes do que veículos com 5 velocidades ($RR = 0.928$ e $IC_{95\%} = (0.908, 0.948)$). Devido à codificação da varíavel *número de velocidades*, verifica-se que existem mais veículos com 5 velocidades do que qualquer outro número de velocidades, como se pode confirmar na tabela 4.6. Esta diferença pode estar na origem do número médio esperado de acidentes ser menor para carros com um número de velocidades diferente de 5.

- Para os veículos com 5 ou 6 portas esperam-se, em média, mais 5% de acidentes do que veículos com menos de 5 portas ($RR = 1.056$ e $IC_{95\%} = (1.037, 1.076)$). A diferença registada no número médio de acidentes entre veículos com 5 ou 6 portas e veículos com menos de 5 portas é de difícil interpretação sendo uma possível explicação o facto de veículos comerciais de transporte de mercadorias serem classificados como veículos com menos de 5 portas. Pela sua natureza, estas viaturas percorrem grandes distâncias durante longos períodos de tempo, ao qual está associado um maior risco de acidentes.

- Para carros movidos a gasolina são esperados, em média, menos 19% de acidentes do que os carros movidos a gasóleo ($RR = 0.818$ e $IC_{95\%} = (0.788, 0.848)$). Como o preço do gasóleo é inferior ao preço da gasolina, os carros movidos a gasóleo são adquiridos com o objectivo de percorrer grandes distâncias, o que implica mais tempo dispendido na condução, o que pode incorrer num maior número de acidentes.

- Para carros com mais de 1500 *kg* são esperados, em média, mais acidentes do que em carros com menos de 1250 *kg*. Apesar de significativa, essa diferença revela-se pouco marcante ($RR = 1.053$ e $IC_{95\%} = (1.017, 1.092)$). O aumento do peso de uma viatura permite supôr que estas não se tratam de automóveis ligeiros mas provavelmente carros comerciais e de mercadorias que passarão mais tempo num percurso, sendo esse percurso demorado e a uma velocidade maior.

- Quanto à idade do veículo, à medida que esta aumenta, são esperados, em média, menos acidentes quando comparando com veículos com menos de 4 anos. Para as segunda, terceira e quartas categorias da variável, os riscos relativos são iguais a 0.978, 0.878 e 0.883, respetivamente, e os intervalos de confiança são (0.938, 1.019), (0.843,0.914) e (0.851,0.916), respetivamente. De acordo com o co-orientador Luís Maranhão, e com base em análises efetuadas na AXA S.A., o aumento da idade do veículo deveria apresentar um efeito negativo no número médio de acidentes, o

que vai de encontro com os riscos relativos obtidos. Analisando com mais detalhe o valor obtido do risco relativo para veículos com mais de 25 anos, constata-se que este é de 0.331 ($IC_{95\%} = (0.267, 0.409)$), ou seja, veículos com mais de 25 anos apresentam um número médio esperado de acidentes menor em aproximadamente 67% do que veículos com menos de 4 anos. Uma explicação para este fenómeno provém da natureza dos carros com mais de 25 anos, que serão carros de exposição e coleção e cujos seguros tendem a ser contra furtos ou roubos.

- Para a antiguidade do seguro, os riscos relativos são 0.972, 0.896 e 0.763, para as segunda, terceira e quarta categorias da variável, o que vai de encontro ao inicialmente conjeturado, tal como na variável *Idade do Veículo*. Os intervalos de confiança são iguais a (0.928,1.015), (0.861,0.933) e (0.728,0.8), respetivamente.

O risco relativo para as interações pode ser calculado tendo em conta 3 situações. Tome-se como exemplo a interação entre as variáveis *Idade do Veículo* e *Idade do Seguro*. É possível analisar o efeito de uma viatura com idade entre os 4 e os 8 anos e cujo seguro tem entre 1 a 2 anos comparando com uma viatura cuja idade é inferior a 4 anos e cujo seguro ainda não completou um ano. A outra alternativa será comparar com uma viatura em que apenas uma destas variáveis se encontra na categoria de referência, ou seja, comparar com uma viatura cujo veículo tem entre 4 a 8 anos e a idade do seguro é inferior a 1 ano ou com uma viatura com menos de 4 anos e cujo seguro tem entre 1 a 2 anos. Focou-se a análise apenas no primeiro caso referido. Caso se quisesse analisar as outras 2 hipóteses, o cálculo seria realizado de forma análoga.

- Caso 1:

$$RR = \frac{E(N_{acidentes}|IV = 2 \text{ e } IS = 2)}{E(N_{acidentes}|IV = 1 \text{ e } IS = 1)} = exp(-0.023 - 0.03 - 0.067) = 0.887$$

- Caso 2

$$RR = \frac{E(N_{acidentes}|IV = 2 \text{ e } IS = 2)}{E(N_{acidentes}|IV = 2 \text{ e } IS = 1)} = exp(-0.03 - 0.067) = 0.907$$

- Caso 3

$$RR = \frac{E(N_{acidentes}|IV = 2 \text{ e } IS = 2)}{E(N_{acidentes}|IV = 1 \text{ e } IS = 2)} = exp(-0.023 - 0.067) = 0.914$$

É esperado que, à medida que as idades do veículo e do seguro aumentam, o número médio de acidentes diminua em relação ao número médio esperado de acidentes para veículos cujas idade do veículo e idade do seguro são inferiores a 4 anos e a 1 ano, respetivamente. Este resultado vai de encontro ao que foi inicialmente conjeturado, isto é, que o aumento das idades do veículo e do seguro apresentem um efeito negativo significativo sobre o número médio de acidentes.

Para veículos com valores de cilindrada superior a 1300 cm^3 e número de cavalos superior a 75, espera-se um número médio de acidentes maior do que para veículos de cilindrada inferior a 1300 cm^3 e número de cavalos inferior a 75. Este efeito positivo significativo sobre o número de acidentes vai de encontro ao que foi inicialmente conjeturado, já que estes veículos são capazes de atingir velocidades elevadas, o que pode estar na origem deste aumento.

A interação entre número de cavalos e combustível utilizado apresenta resultados de uma interpretação menos clara, já que para carros movidos a gasolina e com 75 a 100 cavalos é esperado um número médio de acidentes menor em 14% do que carros movidos a gasóleo e com menos de 75 cavalos. No entanto, para carros cujo combustivel é gasolina e o número de cavalos é superior, o número médio de acidentes é superior em aproximadamente 40%.

Interação	Risco Relativo	$IC_{95\%}$
iv = 2 e is = 2	0.887	(0.876,0.898)
iv = 3 e is = 2	0.816	(0.616,1.017)
iv = 4 e is = 2	0.717	(0.689,0.745)
iv = 5 e is = 2	0.323	(0.216,0.431)
iv = 2 e is = 3	0.799	(0.72,0.878)
iv = 3 e is = 3	0.701	(0.578,0.824)
iv = 4 e is = 3	0.618	(0.495,0.741)
iv = 5 e is = 3	0.244	(0.162,0.326)
iv = 2 e is = 4	0.711	(0.366,1.056)
iv = 3 e is = 4	0.588	(0.374,0.802)
iv = 4 e is = 4	0.471	(0.414,0.528)
iv = 5 e is = 4	0.188	(0.091,0.285)
cil = 2 e cavalos = 2	1.047	(0.991,1.103)
cil = 3 e cavalos = 2	1.204	(0.976,1.432)
cil = 2 e cavalos = 3	1.114	(1.047,1.181)
cil = 3 e cavalos = 3	1.112	(1.017,1.207)
cavalos = 2 e comb = gasolina	0.867	(0.85,0.884)
cavalos = 3 e comb = gasolina	1.38	(1.283,1.477)

Tabela 5.3: Risco Relativo e Intervalos de Confiança a 95% para as Interações (Terceiro Modelo de Regressão Binomial Negativa)

5.2 Modelo de Regressão Logística

Para o modelo de regressão logística, o cálculo do *odds ratio* indica o quanto a chance de acidente varia entre uma categoria e a categoria de referência da variável explicativa em análise. Neste caso, o sucesso corresponde a ter um acidente.

- De acordo com a tabela 5.4, apenas se espera que a marca *BMW* apresente uma maior chance de ter acidente em comparação com a marca *Audi*. No entanto, este efeito não é significativo, o que se pode confirmar na tabela 5.5. O resultado obtido é análogo ao do modelo de regressão Binomial Negativa.

BMW	Citroën	Fiat	Ford	Honda	Mercedes-Benz	Nissan
1.045	0.782	0.882	0.859	0.849	0.893	0.957
Opel	Outra	Peugeot	Renault	Seat	Toyota	Volkswagen
0.861	0.861	0.865	0.9	0.827	0.82	0.87

Tabela 5.4: *Odds Ratio* da Variável *marca* (Modelo de Regressão Logística)

BMW	Citroën
(0.997,1.095)	(0.738,0.829)
Fiat	Ford
(0.833,0.933)	(0.815,0.905)
Honda	Mercedes-Benz
(0.779,0.926)	(0.86,0.926)
Nissan	Opel
(0.914,1.002)	(0.827,0.897)
Outra	Peugeot
(0.832,0.89)	(0.828,0.905)
Renault	Seat
(0.864,0.937)	(0.777,0.879)
Toyota	Volkswagen
(0.777,0.865)	(0.836,0.905)

Tabela 5.5: Intervalos de Confiança a 95% da Variável *marca* (Modelo de Regressão Logística)

- O *odds* para o sucesso diminui cerca de 8.6% para carros com uma carroçaria diferente da dos carros ligeiros de passageiros ($OR = 0.914$ e $IC_{95\%} = (0.895, 0.932)$).

- Com o aumento da cilindrada, o *odds* para o sucesso é superior ao *odds* de sucesso para carros com cilindrada inferior a 1300 cm^3 ($OR = 1.034$ com um $IC_{95\%} = (1.002, 1.067)$ e $OR = 1.103$ com um $IC_{95\%} = (1.06, 1.149)$). Este efeito positivo significativo era esperado pois carros com maior capacidade de cilindrada permitem ao condutor deslocar-se a maior velocidade, incorrendo assim num maior risco de ter um acidente.

- Para os veículos com um número de cavalos entre 75 e 100 espera-se um *odds* para o sucesso superior em 7.5% ao *odds* para o sucesso dos veículos com menos de 75 cavalos ($OR = 1.075$ e $IC_{95\%} = (1.035, 1.117)$). O *odds* para o sucesso aumenta 5.8% para veículos com mais de 100 cavalos em relação ao *odds* dos veículos com menos de 75 cavalos ($OR = 1.058$ e $IC_{95\%} = (1.018, 1.099)$).

- O *odds* para o sucesso em viaturas com caixa automática ou semi-automática é 0.955 vezes o *odds* para o sucesso para viaturas com caixa manual ($OR = 0.955$ e $IC_{95\%} = (0.934, 0.977)$).

- O *odds* para o sucesso em carros que não têm 5 velocidades é aproximadamente 8.6% menor do que o *odds* para o sucesso em carros com 5 velocidades ($OR = 0.924$ e $IC_{95\%} = (0.903, 0.946)$).

- Para viaturas com 5 ou 6 portas o *odds* para o sucesso é 1.055 vezes o *odds* para o sucesso em viaturas com menos de 5 portas ($OR = 1.055$ e $IC_{95\%} = (1.033, 1.075)$). Note-se que há uma concordância com o efeito desta mesma variável no modelo anterior.

- O *odds* para o sucesso em viaturas movidas a gasolina é 20% inferior ao *odds* para o sucesso em viaturas cujo combustível é o gasóleo ($OR = 0.806$ e $IC_{95\%} = (0.777, 0.837)$). A interpretação do *odds ratio* obtido é análoga à efetuada para o modelo de regressão Binomial Negativa.

- Quandos se consideram veículos com mais de 25 anos o *odds* para o sucesso diminui 62% em relação ao *odds* para o sucesso em veículos com menos de 4 anos ($OR = 0.377$ e $IC_{95\%} = (0.304, 0.468)$). Para as segunda, terceira e quartas categorias da variável, os valores dos *odds* para o sucesso são dados por $1.033, 0.953$ e 0.906, respetivamente, com intervalos de confiança a 95% iguais a (0.987,1.08), (0.912,0.996) e (0.87,0.943), respetivamente. O resultado obtido para a segunda categoria não está de acordo com o esperado e, apesar, de os valores para as terceira e quarta categoria serem inferiores a 1, não apresentam uma distância marcante.

- Para a idade do seguro, os *odds* para o sucesso são 2.222 em seguros com idade entre 1 e 2 anos ($IC_{95\%} = (2.111, 2.338)$), 2.046 em seguros com idade entre 2 e 5 anos ($IC_{95\%} = (1.954, 2.142)$) e 1.687 em seguros com idade superior a 5 anos ($IC_{95\%} = (1.599, 1.781)$). Era de esperar que as probabilidades de ter um acidente para estes veículos fossem inferiores à probabilidade de veículos cujo seguro tem menos de 1 ano terem acidentes, o que não se verificou.

Tal como o risco relativo para as interações, o *odds ratio* pode ser calculado de forma análoga tendo em conta as 3 situações apresentadas anteriormente. Para a primeira interação, o cálculo das 3 hipóteses é apresentado a título de exemplo informativo.

- Caso 1:

$$OR = \frac{Odds(Y = 1|IV = 2 \text{ e } IS = 2)}{Odds(Y = 1||IV = 1 \text{ e } IS = 1)} = exp(0.032 + 0.798 - 0.108) = 2.059$$

- Caso 2

$$OR = \frac{Odds(Y = 1||IV = 2 \text{ e } IS = 2)}{Odds(Y = 1||IV = 2 \text{ e } IS = 1)} = exp(0.798 - 0.108) = 1.994$$

- Caso 3

$$OR = \frac{Odds(Y = 1||IV = 2 \text{ e } IS = 2)}{Odds(Y = 1|IV = 1 \text{ e } IS = 2)} = exp(0.032 - 0.108) = 0.927$$

Os resultados obtidos não vão de encontro ao que é esperado, visto que indicam que o *odds* para o sucesso nos carros cuja idade do veículo e idade do seguro são superiores a 4 e 1 anos, respetivamente, é superior ao *odds* para o sucesso nas viaturas onde estas variáveis se encontram na classe de referência. No entanto, verifica-se que para veículos cuja idade é superior a 25 anos e com idade do seguro nas categorias que não a de referência, o *odds* para o sucesso é sempre inferior ao *odds* para veículos com menos de 4 anos e seguro celebrado há menos de 1 ano. Relativamente à interação entre o número de cavalos e o combustível, o *odds* para o sucesso em viaturas com número de cavalos acima dos 100 e que consomem gasolina é inferior a viaturas com menos de 75 cavalos movidas a gasóleo. Tal como se verificou na interação anterior, o resultado obtido é inesperado.

Interação	*Odds Ratio*	$IC_{95\%}$
iv = 2 e is = 2	2.059	(1.981,2.137)
iv = 3 e is = 2	1.884	(1.779,1.989)
iv = 4 e is = 2	1.612	(1.489,1.735)
iv = 5 e is = 2	0.722	(0.266,1.178)
iv = 2 e is = 3	1.833	(1.756,1.91)
iv = 3 e is = 3	1.604	(1.477,1.732)
iv = 4 e is = 3	1.38	(1.281,1.479)
iv = 5 e is = 3	0.549	(0.511,0.587)
iv = 2 e is = 4	1.607	(1.373,1.841)
iv = 3 e is = 4	1.312	(1.225,1.399)
iv = 4 e is = 4	1.068	(1.024,1.112)
iv = 5 e is = 4	0.426	(0.335,0.517)
cavalos = 2 e comb = gasolina	0.896	(0.402,1.39)
cavalos = 3 e comb = gasolina	0.981	(0.968,0.994)

Tabela 5.6: *Odds Ratio* e Intervalos de Confiança a 95% para as Interações (Modelo de Regressão Logística)

5.3 Modelo ZIP

O modelo ZIP é composto por um modelo de contagem e por um modelo de zeros. Para o modelo de contagem, o efeito das variáveis é medido através do risco relativo enquanto que para o modelo de zeros, o *odds ratio* representa a chance de se observar um zero falso num contexto por oposição a outro.

5.3.1 Modelo de Contagem

A tabela seguinte representa o risco relativo para todos os níveis da variável *marca*, com a exceção do nível *Audi* (marca que representa a categoria de referência), quando as restantes variáveis explicativas tomam os valores da categoria de referência.

- Da análise da tabela, é esperado que os carros das marcas *BMW*, *Citroën*, *Mercedes-Benz* e *Peugeot* tenham, em média, mais acidentes do que os veículos da marca *Audi*, enquanto que as restantes marcas tenham, em média, menos acidentes do que os carros da marca *Audi*.

BMW	Citroën	Fiat	Ford	Honda	Mercedes-Benz	Nissan
1.114	1.002	0.906	0.905	0.879	1.005	0.877
Opel	Outra	Peugeot	Renault	Seat	Toyota	Volkswagen
0.869	0.901	1.005	0.999	0.906	0.965	0.935

Tabela 5.7: Risco Relativo da Variável *Marca* (Modelo ZIP)

BMW	Citroën
(1.031, 1.203)	(0.906,1.109)
Fiat	Ford
(0.817,1.004)	(0.826,0.992)
Honda	Mercedes-Benz
(0.747,1.035)	(0.938,1.077)
Nissan	Opel
(0.811,0.948)	(0.808,0.934)
Outra	Peugeot
(0.85,0.956)	(0.931,1.085)
Renault	Seat
(0.932,1.069)	(0.811,1.012)
Toyota	Volkswagen
(0.875,1.063)	(0.87,1.005)

Tabela 5.8: Intervalos de Confiança a 95% da Variável *marca* (Modelo ZIP)

- Para a variável *carroçaria*, espera-se que os carros que se encontram na segunda classe tenham um número médio de acidentes 5% inferior ao número médio de acidentes esperado para carros com uma carroçaria normal ($RR = 0.954$ e $IC_{95\%} = (0.923, 0.986)$).

- De acordo com o risco relativo para a variável *cavalos*, espera-se que carros com uma cavalagem superior tenham, em média, um maior número de acidentes do que o número médio esperado de acidentes para veículos com menos de 75 cavalos.. Obteve-se um risco relativo de 1.09 para carros com 75 a 100 cavalos ($IC_{95\%} = (1.064, 1.118)$)e um risco relativo de 1.136 para carros com mais de 100 cavalos ($IC_{95\%} = (1.105, 1.169)$).

- Para a variável *lugares*, espera-se que os carros que não têm 5 lugares tenham, em média, menos 5% de acidentes do que carros com 5 lugares ($RR = 0.945$ e $IC_{95\%} = (0.904, 0.988)$).

- Em média, para carros a gasolina esperam-se menos 13% acidentes do que para carros a gasóleo ($RR = 0.871$ e $IC_{95\%} = (0.856, 0.887)$), resultado que vai de encontro aos obtidos nos modelos anteriores.

- À medida que a idade do veículo aumenta, espera-se que o número médio de acidentes diminua quando se compara com veículos com menos de 4 anos. Por ordem de categoria, os riscos relativos obtidos foram de 0.902 ($IC_{95\%} = (0.863, 0.942)$), 0.845 ($IC_{95\%} = (0.809, 0.883)$), 0.853 ($IC_{95\%} = (0.817, 0.89)$) e 0.641 ($IC_{95\%} = (0.499, 0.822)$) e são concordantes com o que a conjetura inicial.

- Para veículos cujo seguro tem entre 1 a 2 anos, espera-se que o número médio de acidentes aumente 2.5% ($RR = 1.025$ e $IC_{95\%} = (0.981, 1.071)$) em relação ao número médio de acidentes esperado para veículos cujo seguro tem menos de 1 ano. Para veículos nas categorias 3 e 4, o número médio de acidentes esperado é menor em 8%

e 15%, respetivamente ($RR = 0.919$ e $IC_{95\%} = (0.883, 0.956)$ e $RR = 0.852$ e $IC_{95\%} = (0.817, 0.889)$, respetivamente).

5.3.2 Modelo de Zeros

- Verifica-se pela tabela 5.9 que o *odds* para a existência de zeros falsos aumenta quando se passa da marca *Audi* para qualquer outra marca, com excepção de carros *Honda* e *Nissan*. Para a *Citröen*, o *odds* para o sucesso aumenta em 70%, efeito positivo significativo.

BMW	Citroën	Fiat	Ford	Honda	Mercedes-Benz	Nissan
1.086	1.717	1.045	1.123	0.979	1.265	0.838
Opel	Outra	Peugeot	Renault	Seat	Toyota	Volkswagen
1.027	1.108	1.443	1.236	1.235	1.494	1.177

Tabela 5.9: *Odds Ratio* da Variável *Marca* (Modelo ZIP)

BMW	Citroën
(0.895,1.317)	(1.383,2.132)
Fiat	Ford
(0.812,1.344)	(0.883,1.43)
Honda	Mercedes-Benz
(0.613,1.564)	(1.069,1.504)
Nissan	Opel
(0.67,1.05)	(0.856,1.232)
Outra	Peugeot
(0.957,1.283)	(1.202,1.732)
Renault	Seat
(1.048,1.458)	(0.933,1.636)
Toyota	Volkswagen
(1.195,1.867)	(0.987,1.405)

Tabela 5.10: Intervalos de Confiança a 95% da Variável *marca* (Modelo ZIP)

- Para carros cuja carroçaria se encontra na categoria *Outra*, o *odds* para o sucesso é aproximadamente 15% maior do que o *odds* para o sucesso em carros cuja carroçaria se encontra na categoria de referência ($OR = 1.147$ e $IC_{95\%} = (1.049, 1.245)$).

- Veículos com cilindrada entre 1300 e 1800 cm^3 e veículos com cilindrada superior a 1800 cm^3 registam um *odds* para o sucesso inferior ao *odds* para o sucesso de veículos com cilindrada inferior a 1300 cm^3 ($OR = 0.886$ e $IC_{95\%} = (0.834, 0.942)$ e $OR = 0.832$ e $IC_{95\%} = (0.764, 0.905)$, respetivamente). Como foi referido anteriormente, carros com uma cilindrada alta incorrem num maior risco de acidente devido à possibilidade de atingirem maiores velocidades. Assim, a probabilidade de não

se registar um acidente diminui, o que implica a diminuição do número de zeros falsos.

- O *odds* para o sucesso em viaturas com caixa automática ou semi-automática é superior em 13.6% ao *odds* para o sucesso para viaturas com caixa manual ($OR = 1.136$ e $IC_{95\%} = (1.081, 1.194)$).

- O *odds* para sucesso em carros que não têm 5 velocidades é aproximadamente 23% maior do que o *odds* para o sucesso em carros com 5 velocidades ($OR = 1.23$ e $IC_{95\%} = (1.17, 1.294)$).

- O *odds* para o sucesso em viaturas com 5 ou 6 portas é 0.893 vezes o *odds* para o sucesso para viaturas com menos de 5 portas ($IC_{95\%} = (0.853, 0.933)$).

- O *odds* para o sucesso em viaturas com um número de lugares diferente de 5. é 0.822 vezes o *odds* para o sucesso para viaturas com 5 lugares ($IC_{95\%} = (0.736, 0.908)$).

- O *odds* para o sucesso em viaturas com tração a 4 rodas é inferior ao *odds* para sucesso para viaturas com tração a 2 rodas em 10% ($OR = 0.901$ e $IC_{95\%} = (0.842, 0.965)$).

- Para veículos na segunda categoria da variável *peso*, o valor do *odds* obtido foi de 0.934 ($IC_{95\%} = (0.876, 0.996)$), enquanto que para os veículos na terceira categoria este foi de 0.96 ($IC_{95\%} = (0.882, 1.045)$). Carros mais pesados tendem a tratar-se de viaturas comerciais que, como já foi referido, incorrem num maior risco de acidente. Assim, o número de zeros diminui, o que implica a diminuição do número de zeros falsos.

- Os valores obtidos para os *odds ratio* da variável *idade do veículo* não são concordantes com o que se esperava obter, visto que indicam que o *odds* para o sucesso de veículos nas categorias $]8-12]$,$]12-25]$ e > 25 registarem um zero falso é maior do que o *odds* para o sucesso em veículos com menos de 4 anos. Os intervalos de confiança obtidos para as categorias são (por ordem de categoria) (0.771,0.982), (0.934,1.176), (1.297,1.604) e (2.737,5.105).

IV = 2	IV = 3	IV = 4	IV = 5
0.87	1.048	1.442	3.921

Tabela 5.11: *Odds Ratio* da Variável *Idade do Veículo* (Modelo ZIP)

- Para a variável *idade do seguro*, obtiveram-se resultados semelhantes aos da variável *idade do veículo* para os quais não parece existir uma razão aparente pela qual se verifica que os carros cujo seguro tem entre 2 a 5 anos ou mais de 5 anos esperam um *odds* para o sucesso superior ao *odds* para o sucesso dos carros cujo seguro tem menos de 1 ano. Os intervalos de confiança obtidos para as categorias são (por ordem de categoria) (0.846,1.086), (0.898,1.134) e (1.179,1.482).

IS = 2	IS = 3	IS = 4
0.959	1.009	1.322

Tabela 5.12: *Odds Ratio* da Variável *Idade do Seguro* (Modelo ZIP)

5.4 Modelo Hurdle

A interpretação do efeito das variáveis é realizada de forma análoga à do modelo ZIP, tendo em conta que para este modelo, o *odds* indica a chance de se observar um zero e não um zero falso.

5.4.1 Modelo de Contagem

- Através da tabela 5.13, verifica-se um efeito positivo sobre o número médio de acidentes das marcas *BMW*, *Mercedes-Benz*, *Peugeot* e *Renault* quando se compara com veículos da marca *Audi*. Pelo contrário, para as restantes marcas é esperado um número médio de acidentes inferior ao observado para os carros da marca *Audi*.

BMW	Citroën	Fiat	Ford	Honda	Mercedes-Benz	Nissan
1.104	0.944	0.882	0.947	0.962	1.05	0.944
Opel	Outra	Peugeot	Renault	Seat	Toyota	Volkswagen
0.908	0.906	1.005	1.032	0.875	0.986	0.936

Tabela 5.13: Risco Relativo para a variável *Marca* (Modelo Hurdle)

BMW	Citroën
(1.009,1.208)	(0.832,1.07)
Fiat	Ford
(0.777,1.002)	(0.852,1.053)
Honda	Mercedes-Benz
(0.8,1.156)	(0.972,1.133)
Nissan	Opel
(0.859,1.037)	(0.833,0.99)
Outra	Peugeot
(0.845,0.973)	(0.917,1.102)
Renault	Seat
(0.953,1.119)	(0.766,1)
Toyota	Volkswagen
(0.882,1.103)	(0.862,1.016)

Tabela 5.14: Intervalos de Confiança a 95% da Variável *marca* (Modelo Hurdle)

- Os riscos relativos da variável *cavalos* são de 1.059 para a segunda categoria ($IC_{95\%} = (1.002, 1.118)$) e de 1.151 para a terceira categoria ($IC_{95\%} = (1.086, 1.221)$), o que indica um aumento do número médio de acidentes em relação a carros com menos de 75 cavalos. Verifica-se uma concordância com os resultados obtidos até agora.

- Para carros que não têm 5 velocidades esperam-se, em média, menos 5% de acidentes do que veículos com 5 velocidades ($RR = 0.949$ e $IC_{95\%} = (0.906, 0.995)$).

- Para os carros que não têm 5 lugares, em média, esperam-se menos 7.6% de acidentes do que carros com 5 lugares ($RR = 0.924$ e $IC_{95\%} = (0.878, 0.972)$).

- O risco relativo correspondente à variável *combustível* é igual a 0.88, logo conclui-se que para os carros movidos a gasolina esperam-se, em média, menos 12% de acidentes por oposição aos carros movidos a gasóleo ($IC_{95\%} = (0.848, 0.913)$).

- Para a variável *idade do veículo*, a tabela 5.10 evidencia um decréscimo no número médio de acidentes esperados para cada categoria em oposição à categoria de referência (veículos com menos de 4 anos). Os intervalos de confiança a 95% (por ordem de categoria) são iguais a (0.799,0.881), (0.725,0.799), (0.691,0.763) e (0.44,0.765).

IV = 2	IV = 3	IV = 4	IV = 5
0.839	0.761	0.726	0.58

Tabela 5.15: Risco Relativo para a variável *Idade do Veículo* (Modelo Hurdle)

- Para a variável *idade do seguro*, os riscos relativos obtidos permitem concluir que o número médio de acidentes esperados diminui à medida que a idade do seguro aumenta, por oposição a seguros com menos de 1 ano. Os intervalos de confiança a 95% (por ordem de categoria) são iguais a (0.757,0.834), (0.674,0.737) e (0.638,0.703).

IS = 2	IS = 3	IS = 4
0.795	0.705	0.669

Tabela 5.16: Risco Relativo para a variável *Idade do Seguro* (Modelo Hurdle)

5.4.2 Modelo de Zeros

- De acordo com a tabela 5.17, apenas para a marca *BMW* é prevista uma probabilidade de sucesso maior do que a marca *Audi*. Estes valores são bastante semelhantes aos obtidos nos modelos anteriores

BMW	Citroën	Fiat	Ford	Honda	Mercedes-Benz	Nissan
1.08	0.784	0.891	0.85	0.856	0.889	0.909
Opel	Outra	Peugeot	Renault	Seat	Toyota	Volkswagen
0.846	0.853	0.85	0.896	0.829	0.788	0.869

Tabela 5.17: *Odds Ratio* para a variável *Marca* (Modelo Hurdle)

BMW	Citroën
(1.03,1.132)	(0.739,0.831)
Fiat	Ford
(0.841,0.943)	(0.806,0.896)
Honda	Mercedes-Benz
(0.785,0.933)	(0.855,0.924)
Nissan	Opel
(0.869,0.951)	(0.812,0.882)
Outra	Peugeot
(0.823,0.884)	(0.812,0.889)
Renault	Seat
(0.86,0.933)	(0.778,0.882)
Toyota	Volkswagen
(0.748,0.831)	(0.836,0.904)

Tabela 5.18: Intervalos de Confiança a 95% da Variável *marca* (Modelo Hurdle)

- Para os carros com uma carroçaria que não a de um ligeiro de passageiros esperam-se um *odds* para o sucesso inferior em 10% ao *odds* para o sucesso em carros que se encontram na categoria de referência desta variável ($OR = 0.894$ e $IC_{95\%} = (0.873, 0.915)$).

- Os valores obtidos para os riscos relativos da variável *cilindrada* são concordantes com os obtidos até agora ($OR = 1.056$ e $IC_{95\%} = (1.025, 1.087)$ e $OR = 1.099$ e $IC_{95\%} = (1.054, 1.145)$, para as segunda e terceira categorias, respetivamente).

- Para a variável *cavalos*, obteve-se um *odds* de 1.111 para a segunda categoria e um *odds* de 1.15 para a terceira categoria, o que indica um *odds* para o sucesso superior ao *odds* para o sucesso em relação a carros com menos de 75 cavalos. Os intervalos de confiança a 95% são dados por (1.082,1.143) e (1.115,1.187), respetivamente.

- Para carros cuja caixa é automática ou semi-automática, o *odds ratio* é igual a 0.947, ou seja, o *odds* para o sucesso diminui em 6% por oposição a carros com caixa manual ($IC_{95\%} = (0.924, 0.969)$).

- Para carros com um número de velocidades diferente de 5 espera-se um *odds* para o sucesso 0.914 vezes o *odds* para o sucesso de carros em 5 velocidades ($IC_{95\%} = (0.894, 0.937)$).

- O *odds* para o sucesso em veículos com 5 ou 6 portas é 1.053 vezes o *odds* para o sucesso em veículos com menos de 5 portas ($OR = 1.053$ e $IC_{95\%} = (1.030, 1.075)$).

- Para os carros que não têm 5 lugares espera-se um *odds* para o sucesso superior em 3.2% ao *odds* para o sucesso em carros com 5 lugares ($OR = 1.032$ e $IC_{95\%} = (1.006, 1.058)$).

- Para carros movidos a gasolina, espera-se um *odds* para o sucesso superior em 13% ao *odds* para o sucesso em carros movidos a gasóleo ($OR = 0.87$ e $IC_{95\%} = (0.853, 0.887)$).

- O *odds* para o sucesso em viaturas com tração a 4 rodas é superior ao *odds* para sucesso para viaturas com tração a 2 rodas em 5% ($OR = 1.051$ e $IC_{95\%} = (1.02, 1.082)$).

- Os valores da tabela 5.19 evidenciam o decréscimo da chance de acidente com o aumento da idade do veículo por oposição aos veículos com menos de 4 anos. Para veículos com mais de 25 anos, a chance de sofrer um acidente é 70% menor do que a probabilidade de sucesso de veículos com menos de 4 anos. Os intervalos de confiança a 95% (por ordem de categoria) são iguais a (0.948,0.997), (0.822,0.865), (0.724,0.759) e (0.266,0.328).

IV = 2	IV = 3	IV = 4	IV = 5
0.972	0.843	0.742	0.295

Tabela 5.19: *Odds Ratio* para a variável *Idade do Veículo* (Modelo Hurdle)

- A tabela 5.20 permite concluir que a chance de ter acidente diminui com o aumento da idade do seguro quando se compara com veículos cujo seguro tem menos de 1 ano. Os intervalos de confiança a 95% (por ordem de categoria) são iguais a (0.898,0.941), (0.787,0.822) e (0.628,0.656).

IS = 2	IS = 3	IS = 4
0.92	0.804	0.642

Tabela 5.20: *Odds Ratio* para a variável *Idade do Seguro* (Modelo Hurdle)

Capítulo 6

Conclusões

Esta dissertação teve como objectivo o estudo de modelos de regressão para contagens e a sua aplicação a dados de acidentes de automóveis fornecidos pela seguradora AXA S.A.. O problema proposto centrou-se na identificação de variáveis relacionadas com os automóveis que influenciem o número de acidentes sofridos. As variáveis explicativas consideradas foram a *marca*, a *carroçaria*, a *cilindrada*, o *número de cavalos*, a *caixa de velocidades*, o *número de velocidades*, o *número de portas*, o *número de lugares*, o *combustível*, a *distância entre eixos*, a *tração*, o *peso*, a *idade do veículo* e a *idade do seguro*. Consideraram-se também interações entre algumas das variáveis que, no entender do problema, se justificavam relevantes.

Primeiramente, os dados fornecidos pela AXA S.A. foram ajustados através de um modelo de regressão de Poisson. No entanto, a presença de alguma sobredispersão no modelo e os resultados de diagnóstico (resíduos, *leverages* e distâncias de Cook) menos bons levaram à consideração de um modelo de regressão Binomial Negativa. Dado o mau ajustamento deste, reajustou-se novamente o modelo aos dados em que o número de acidentes constava apenas de 4 classes: 0 acidentes, 1 acidente, 2 acidentes e 3 ou mais acidentes. O modelo obtido não trouxe melhorias já que o modelo ajustava, para todas as observações, um valor inferior a 0.5 para a variável resposta.

Numa segunda fase, foi utilizado um modelo de regressão logística para analisar os dados. Verificou-se que, tal como no caso do modelo de regressão Binomial Negativa, a totalidade das observações registou valores ajustados pelo modelo inferiores a 0.5, o que leva a uma classificação de ausência de acidentes para toda a base de dados.

Dado o número excessivo de zeros, consideraram-se modelos adequados a este problema, nomeadamente um modelo de regressão de Poisson com zeros inflacionados e um modelo de regressão de Poisson com barreira. Para ambos os modelos, verificou-se que as contagens não apresentavam um bom ajustamento já que a análise da previsão do conjunto de observações selecionadas denunciava uma probabilidade elevada de se trataram de contagens nulas. Foi efetuado o agrupamento da variável resposta em 4 acidentes, não tendo sido verificada nenhuma melhoria.

Apesar de, em alguns modelos se verificarem resultados inesperados, de uma forma ge-

ral, observou-se que as variáveis com uma associação positiva com o número de acidentes foram o *número de cavalos*, a *cilindrada* e o *peso*. O *número de velocidades*, o *combustível*, a *idade do veículo* e a *idade do seguro* mostraram ter uma associação negativa com o número médio de acidentes.

A primeira limitação deste estudo passou pelo excessivo número de zeros, já que todos os modelos previam as observações como ausência de acidente. A segunda limitação passou pela insuficiência de informação relativa aos acidentes. Apesar de o problema proposto pela AXA se prender apenas com variáveis de natureza automóvel, é sabido que uma pessoa se pode envolver num acidente devido a razões de natureza humana (idade, embriaguez, cansaço) ou ambientais (condições metereológicas, estado de conservação da via onde circula, trânsito). Note-se também que não havia registo sobre a culpa de um indivíduo no acidente.

Como trabalho futuro, uma sugestão para resolver os problemas encontrados ao longo desta dissertação passa pela consideração de outros métodos para identificar variáveis que influenciem o número de acidentes, como é o caso de métodos de classificação automática como método dos k-vizinhos, redes neuronais e máquinas de suporte vectorial.

Referências

Adrain, R. (1808). Research concerning the probabilities of the errors which happen in making observations. *The Analyst, or, Mathematical Museum*, 1(4):93–109.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.

Al-Ghamdi, A. S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. *Journal of Accident Analysis and Prevention*, 34(6):729–741.

Ascensão, F. and Mira, A. (2006). Factors affecting culvert use by vertebrates along two streches of road in southern portugal. *Ecological Research*, 1:57–66.

Berkson, J. (1944). Application of the logistic function to bio-assay. *Annals of Applied Biology*, 39(227):357–365.

Böhning, D., Dietz, E., and Schlattmann, P. (1997). Zero-inflated count models and their applications in public health and social science. *Applications of Latent Trait and Latent Class Models in the Social Sciences*, pages 333–344.

Bliss, C. (1935). The calculation of the dosage-mortality curve. *Annals of Applied Biology*, 22:134–167.

Bollen, K. A. and Jackman, R. W. (1985). Political democracy and the size distribution of income. *American Sociological Review*, 50:438–457.

Boswell, M. T. and Patil, G. P. (1970). Chance mechanisms generating negative binomial distributions. *Random Counts in Scientific Work*, 1.

Cameron, A. and Trivedi, P. (1998). *Regression Analysis of Count Data*. Cambridge University Press, Cambridge, first edition.

Cheung, Y. (2002). Zero-inflated models for regression analysis of count data: a study of growth and development. *Statistics in Medicine*, 21(10).

Costa, J. O., Freitas, E. F., Pereira, P. A. A., and Jacques, M. A. P. (2011). Acidentes rodoviários das estradas nacionais de portugal : estudo da associação entre as variáveis recolhidas. *C-TAC - Comunicações a Conferências Nacionais*.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, 39(5):829–844.

Dalrymple, M. L., Hudson, I., and Ford, R. (2003). Finite mixture, zero-inflated poisson and hurdle models with application to sids. *Computational Statistics and Data Analysis*, 41:491–504.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC, USA.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, New York, USA.

Famoye, F. and Singh, K. P. (2006). Zero-inflated poisson regression model with an application to domestic violence data. *Journal of Data Science*, 4:117–130.

Feller, W. (1943). On a general class of "contagious"distributions. *Annals of Mathematical Statistics*, 16:319–329.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society*, 222:309–368.

Gaio, A. R. (2011-2012). Apontamentos escritos da disciplina eace: Mestrado em engenharia matemática.

Gauss, C. (1809). *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium Frid*. Perthes et I. H. Besser, Hamburg, Germany.

Glaser, B. G. and Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine Publishing Company, Chicago, USA.

Grandell, J. (1997). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chapman and Hall, Great Britain, first edition.

Hilbe, J. M. (2011). *Negative Binomial Regression*. Cambridge University Press, United Kingdom, second edition.

Hoef, J. M. V. and Boveng, P. L. (2007). Quasi-poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology*, 88(11):2766–2772.

Ismail, N. and Jemain, A. A. (2007). Handling overdispersion with negative binomial and generalized poisson regression models. *Casualty Actuarial Society Forum*, Winter:103–158.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Lawless, J. F. (1987). Negative binomial and mixed poisson regression. *The Canadian Journal of Statistics*, 15(3).

Legendre, A. M. (1805). *Nouvelles Méthodes pour la Détermination des Orbites des Comètes*. Coursier, Paris, France.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall/CRC, London, United Kingdom, second edition.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–365.

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

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ridout, M., Demétrio, C., and Hinde, J. (1998). Models for count data with many zeros. *International Biometric Conference*, pages 179–192.

Ridout, M., Hinde, J., and Demétrio, C. (2001). A score test for testing a zero-inflated poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, 57(1):219–223.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Soemitro, R. A. A. and Bahat, Y. S. (1005). Accident analysis assessment to the accident influence factors on traffic safety improvement. *Buku Proceedings of the Eastern Asia Society for Transportation Studies*, 5:2091–2105.

Turkman, M. A. and Silva, G. L. (2000). *Modelos Lineares Generalizados - da Teoria à Prática*. Edições SPE, Lisboa, Portugal.

Wedagama, D. M. P. and Dissanayake, D. (2010). The influence of accident related factors on road fatalities considering bali province in indonesia as a case study. *Journal of the Eastern Asia Society for Transportation Studies*, 8:1905–1917.

Zeileis, A., Kleiber, C., , and Jackman, S. (2008). Regression models for count data in rr. *Journal of Statistical Software*, 27(8).

Zippin, C. and Armitage, P. (1966). Use of concomitant variables and incomplete survival information in the estimation of an exponential survival parameter. *Biometrics*, 22:665–672.

Zuur, A., Ieno, E., Walker, N., Saveliev, A., and Smith, G. (2009). *Mixed Effects Models and Extensions in Ecology with R*. Springer Science+Business Media, New York, USA, first edition.

Anexos

Anexo A

Modelos de Regressão

Tabela A.1: Output do Modelo de Regressão de Poisson

Variável	Coeficiente	Erro padrão	Valor-p
constante	-7.612	0.026	$<$ 2e-16
I_{bmw}	0.075	0.02	1.55e-04
$I_{citroen}$	-0.22	0.025	$<$ 2e-16
I_{fiat}	-0.117	0.025	2.55e-06
I_{ford}	-0.143	0.023	3.25e-10
I_{honda}	-0.139	0.038	2.24e-04
$I_{mercedes-benz}$	-0.101	0.016	5.94e-10
I_{nissan}	-0.067	0.02	6.96e-04
I_{opel}	-0.153	0.018	$<$ 2e-16
I_{outra}	-0.15	0.015	$<$ 2e-16
$I_{peugeot}$	-0.133	0.019	3.11e-12
$I_{renault}$	-0.087	0.017	5.31e-07
I_{seat}	-0.193	0.027	3.71e-13
I_{toyota}	-0.187	0.023	5.36e-16
$I_{volkswagen}$	-0.136	0.017	1.47e-15
$I_{carrocaria}$	-0.103	0.009	$<$ 2e-16
I_{cil2}	0.005	0.016	0.773
I_{cil3}	0.102	0.042	0.015
$I_{cavalos2}$	-0.015	0.032	0.626
$I_{cavalos3}$	0.405	0.138	0.003
I_{caixa}	-0.044	0.01	1.52e-05
I_{vel}	-0.076	0.01	5.75e-14
I_{portas}	0.054	0.009	6.44e-10
$I_{gasolina}$	-0.2	0.017	$<$ 2e-16
I_{peso2}	0.017	0.013	0.202
I_{peso3}	0.051	0.017	0.003

Continua na página seguinte

Tabela A.1 – *Continuação da página anterior*

Variável	Coeficiente	Erro padrão	Valor-p
I_{iv2}	-0.027	0.02	0.168
I_{iv3}	-0.135	0.019	3.35e-12
I_{iv4}	-0.122	0.018	7.73e-12
I_{iv5}	-1.044	0.102	$<$ 2e-16
I_{is2}	-0.043	0.021	0.041
I_{is3}	-0.121	0.019	2.97e-10
I_{is4}	-0.28	0.023	$<$ 2e-16
$I_{iv2} \times I_{is2}$	-0.066	0.031	0.034
$I_{iv3} \times I_{is2}$	-0.041	0.03	0.169
$I_{iv4} \times I_{is2}$	-0.181	0.028	9.52e-11
$I_{iv5} \times I_{is2}$	0.037	0.146	0.799
$I_{iv2} \times I_{is3}$	-0.088	0.028	0.001
$I_{iv3} \times I_{is3}$	-0.112	0.027	3.04e-05
$I_{iv4} \times I_{is3}$	-0.252	0.025	$<$ 2e-16
$I_{iv5} \times I_{is3}$	-0.202	0.138	0.144
$I_{iv2} \times I_{is4}$	-0.049	0.030	0.105
$I_{iv3} \times I_{is4}$	-0.129	0.029	1.16e-05
$I_{iv4} \times I_{is4}$	-0.347	0.028	$<$ 2e-16
$I_{iv5} \times I_{is4}$	-0.28	0.133	0.035
$I_{cil2} \times I_{cavalos2}$	0.049	0.029	0.096
$I_{cil3} \times I_{cavalos2}$	0.096	0.052	0.063
$I_{cil2} \times I_{cavalos3}$	-0.301	0.138	0.029
$I_{cil3} \times I_{cavalos3}$	-0.402	0.143	0.005
$I_{cavalos2} \times I_{gasolina}$	0.079	0.025	0.001
$I_{cavalos3} \times I_{gasolina}$	0.099	0.02	1.21e-06

Tabela A.2: Output do Modelo de Regressão Binomial Negativa

Variável	Coeficiente	Erro padrão	Valor-p
constante	-7.612	0.028	$<$ 2e-16
I_{bmw}	0.073	0.021	6.12e-04
$I_{citroen}$	-0.22	0.027	5.73e-16
I_{fiat}	-0.117	0.027	1.06e-05
I_{ford}	-0.142	0.024	5.98e-09
I_{honda}	-0.138	0.041	6.61e-04
$I_{mercedes-benz}$	-0.102	0.018	6.60e-09
I_{nissan}	-0.068	0.021	1.57e-03
I_{opel}	-0.153	0.019	3.55e-16

Continua na página seguinte

Tabela A.2 – *Continuação da página anterior*

Variável	**Coeficiente**	**Erro padrão**	**Valor-p**
I_{outra}	-0.15	0.016	$<$ 2e-16
$I_{peugeot}$	-0.13	0.019	5.74e-11
$I_{renault}$	-0.09	0.019	1.78e-06
I_{seat}	-0.193	0.029	1.54e-11
I_{toyota}	-0.188	0.025	3.00e-14
$I_{volkswagen}$	-0.137	0.019	8.84e-14
$I_{carrocaria}$	-0.102	0.01	$<$ 2e-16
I_{cil2}	0.003	0.017	0.858
I_{cil3}	0.102	0.044	0.023
$I_{cavalos2}$	-0.015	0.035	0.662
$I_{cavalos3}$	0.408	0.149	0.006
I_{caixa}	-0.045	0.011	4.80e-05
I_{vel}	-0.076	0.011	3.05e-12
I_{portas}	0.055	0.009	5.50e-09
$I_{gasolina}$	-0.199	0.019	$<$ 2e-16
I_{peso2}	0.016	0.014	0.24
I_{peso3}	0.05	0.019	0.007
I_{iv2}	-0.024	0.021	0.248
I_{iv3}	-0.133	0.021	1.23e-10
I_{iv4}	-0.116	0.019	8.09e-10
I_{iv5}	-1.038	0.104	$<$ 2e-16
I_{is2}	-0.04	0.023	0.077
I_{is3}	-0.118	0.021	1.61e-08
I_{is4}	-0.278	0.024	$<$ 2e-16
$I_{iv2} \times I_{is2}$	-0.067	0.034	0.05
$I_{iv3} \times I_{is2}$	-0.041	0.032	0.201
$I_{iv4} \times I_{is2}$	-0.183	0.03	1.42e-09
$I_{iv5} \times I_{is2}$	0.029	0.151	0.846
$I_{iv2} \times I_{is3}$	-0.09	0.03	0.002
$I_{iv3} \times I_{is3}$	-0.114	0.029	7.73e-05
$I_{iv4} \times I_{is3}$	-0.258	0.027	$<$ 2e-16
$I_{iv5} \times I_{is3}$	-0.209	0.142	0.141
$I_{iv2} \times I_{is4}$	-0.051	0.033	0.119
$I_{iv3} \times I_{is4}$	-0.13	0.032	3.57e-05
$I_{iv4} \times I_{is4}$	-0.352	0.03	$<$ 2e-16
$I_{iv5} \times I_{is4}$	-0.286	0.136	0.036
$I_{cil2} \times I_{cavalos2}$	0.051	0.031	0.104
$I_{cil3} \times I_{cavalos2}$	0.09	0.055	0.088
$I_{cil2} \times I_{cavalos3}$	-0.303	0.149	0.042
$I_{cil3} \times I_{cavalos3}$	-0.405	0.154	0.009
$I_{cavalos2} \times I_{gasolina}$	0.076	0.026	0.004
$I_{cavalos3} \times I_{gasolina}$	0.099	0.023	5.08e-06

Tabela A.3: Output do Segundo Modelo de Regressão Binomial Negativa

Variável	**Coeficiente**	**Erro padrão**	**Valor-p**
constante	-7.62	0.028	$<$ 2e-16
I_{bmw}	0.07	0.021	8.74e-04
$I_{citroen}$	-0.221	0.025	$<$ 2e-16
I_{fiat}	-0.116	0.024	3.14e-06
I_{ford}	-0.146	0.023	1.37e-10
I_{honda}	-0.136	0.038	7.67e-04
$I_{mercedes-benz}$	-0.104	0.018	2.83e-10
I_{nissan}	-0.066	0.021	1.83e-03
I_{opel}	-0.151	0.019	$<$ 2e-16
I_{outra}	-0.148	0.015	$<$ 2e-16
$I_{peugeot}$	-0.134	0.019	4.02e-12
$I_{renault}$	-0.089	0.019	5.48e-07
I_{seat}	-0.192	0.026	4.93e-13
I_{toyota}	-0.187	0.023	8.79e-16
$I_{volkswagen}$	-0.137	0.017	1.55e-15
$I_{carrocaria}$	-0.102	0.009	$<$ 2e-16
I_{cil2}	0.004	0.016	0.833
I_{cil3}	0.103	0.041	0.021
$I_{cavalos2}$	-0.015	0.032	0.664
$I_{cavalos3}$	0.407	0.138	0.006
I_{caixa}	-0.044	0.01	1.93e-05
I_{vel}	-0.074	0.01	3.55e-13
I_{portas}	0.055	0.009	8.87e-10
$I_{gasolina}$	-0.197	0.017	$<$ 2e-16
I_{peso2}	0.017	0.013	0.218
I_{peso3}	0.05	0.017	0.003
I_{iv2}	-0.022	0.02	0.289
I_{iv3}	-0.13	0.019	1.38e-11
I_{iv4}	-0.114	0.018	2.76e-11
I_{iv5}	-1.043	0.102	$<$ 2e-16
I_{is2}	-0.041	0.021	0.071
I_{is3}	-0.12	0.019	1.71e-10
I_{is4}	-0.28	0.022	$<$ 2e-16
$I_{iv2} \times I_{is2}$	-0.067	0.031	0.045
$I_{iv3} \times I_{is2}$	-0.042	0.029	0.191
$I_{iv4} \times I_{is2}$	-0.183	0.028	1.2e-10
$I_{iv5} \times I_{is2}$	0.041	0.146	0.786

Continua na página seguinte

Tabela A.3 – *Continuação da página anterior*

Variável	**Coeficiente**	**Erro padrão**	**Valor-p**
$I_{iv2} \times I_{is3}$	-0.09	0.028	0.001
$I_{iv3} \times I_{is3}$	-0.115	0.027	2.91e-05
$I_{iv4} \times I_{is3}$	-0.254	0.025	$<$ 2e-16
$I_{iv5} \times I_{is3}$	-0.196	0.139	0.167
$I_{iv2} \times I_{is4}$	-0.049	0.03	0.13
$I_{iv3} \times I_{is4}$	-0.129	0.029	1.51e-05
$I_{iv4} \times I_{is4}$	-0.349	0.028	$<$ 2e-16
$I_{iv5} \times I_{is4}$	-0.273	0.133	0.045
$I_{cil2} \times I_{cavalos2}$	0.053	0.029	0.088
$I_{cil3} \times I_{cavalos2}$	0.092	0.052	0.096
$I_{cil2} \times I_{cavalos3}$	-0.305	0.148	0.039
$I_{cil3} \times I_{cavalos3}$	-0.407	0.143	0.008
$I_{cavalos2} \times I_{gasolina}$	0.073	0.025	0.006
$I_{cavalos3} \times I_{gasolina}$	0.1	0.02	1.04e-06

Anexo B

Observações selecionadas para análise de previsão de modelos com um número excessivo de zeros

	marca	carroc	cil	cavalos	caixa	vel	portas	lug	comb	dee	trac	peso	pot	iv	is	acidentes
1	outra	outra	3	3	outra	2	2	1	gasolina	2	1	3	3	3	2	0
2	audi	normal	2	3	manual	2	1	1	gasolina	1	1	2	3	4	3	0
3	outra	normal	3	3	manual	1	2	1	gasóleo	2	1	2	2	2	4	0
4	renault	normal	1	1	manual	1	1	1	gasolina	1	1	1	2	4	2	0
5	volkswagen	normal	3	3	manual	2	1	1	gasóleo	1	1	2	2	3	3	0
6	outra	normal	2	2	manual	1	1	1	gasolina	1	1	1	2	4	2	1
7	bmw	outra	3	3	outra	2	1	1	gasolina	2	1	3	3	4	3	1
8	outra	normal	1	2	manual	1	1	1	gasolina	1	1	1	2	1	4	1
9	citroen	normal	2	1	manual	1	1	1	gasolina	1	1	1	3	4	4	1
10	outra	outra	3	3	outra	2	2	1	gasóleo	2	2	3	2	4	4	1
11	outra	normal	2	2	manual	1	2	1	gasolina	1	1	1	2	4	1	2
12	outra	normal	3	3	manual	1	1	1	gasóleo	1	1	2	2	2	3	2
13	outra	normal	3	3	manual	1	1	1	gasóleo	1	1	2	2	2	4	2
14	bmw	outra	3	3	outra	2	2	1	gasolina	2	2	3	3	1	1	3
15	honda	normal	1	2	manual	1	2	1	gasolina	1	1	1	2	2	1	3
16	ford	outra	2	2	manual	1	2	1	gasóleo	2	1	2	1	1	4	4
17	peugeot	normal	2	2	manual	1	2	1	gasolina	1	1	2	2	1	4	4
18	outra	normal	3	3	manual	1	1	1	gasóleo	1	1	2	2	3	3	5
19	citroen	normal	3	3	manual	2	1	1	gasóleo	2	1	3	2	1	4	5
20	bmw	outra	3	3	manual	2	2	1	gasóleo	2	1	3	2	2	3	6
21	mercedes-benz	normal	3	3	manual	2	1	1	gasóleo	2	1	3	3	2	2	6
22	audi	outra	3	3	manual	2	2	1	gasóleo	2	1	2	2	2	1	7
23	mercedes-benz	normal	3	3	outra	1	1	1	gasóleo	2	1	3	2	3	4	7
24	outra	outra	3	3	manual	2	1	2	gasóleo	2	1	3	2	1	3	7
25	ford	outra	2	3	manual	1	2	1	gasóleo	2	1	2	2	1	1	12

Figura B.1: Observações selecionadas para análise de previsão de modelos com um número excessivo de zeros