MEGI

MESTRADO
Estatística e Gestão de Informação

MODELO DE AVALIAÇÃO DE RISCO EM ACIDENTES NO RAMO AUTOMÓVEL

Sandra de Jesus Alves Martins

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Estatística e Gestão de Informação

UNIVERSIDADE NOVA ISEGI

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Modelo de Avaliação de Risco em Acidentes no Ramo Automóvel

por

Sandra de Jesus Alves Martins

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Estatística e Gestão de Informação, especialização em Análise de Informação

Orientador: Professor Doutor José António Rui Amaral Santos

Agosto, 2012

AGRADECIMENTOS

Ao meu Orientador, Professor Doutor José António Rui Amaral Santos, por ter aceite orientar o meu trabalho, pelo seu apoio permanente, pelas suas criticas e preciosas sugestões, pela paciência e dedicação que foram fundamentais para concretização desta dissertação.

À CaixaSeguros por ter autorizado a disponibilização dos dados sem os quais não teria sido possível a realização deste trabalho, em particular ao Dr. Paulo Figueiredo e ao Sr. Carlos Pereira, por toda a disponibilidade, compreensão e empenho demonstrados.

Aos docentes e aos colegas do ISEGI pelos conhecimentos transmitidos, pela amizade e apoio sempre presentes.

Aos meus pais José Augusto e Maria Luísa, à minha irmã Susana, ao meu cunhado José Miguel, e aos meus sobrinhos Miguel e Tiago, por todo o amor, carinho, apoio, incentivo e por estarem incondicionalmente ao meu lado.

Por último, a todos aqueles que de alguma forma contribuíram para a concretização deste projecto pessoal e profissional, o meu sincero agradecimento.

RESUMO

O objectivo deste trabalho é desenvolver um modelo econométrico, no âmbito da análise de regressão, que permita a uma companhia de seguros avaliar o risco de cada condutor em potência, ou efectivo, em função da sua propensão para o acidente de viação.

A taxa de sinistralidade rodoviária foi, desde sempre, elevada em Portugal e as seguradoras, apesar de disporem de vários instrumentos para a avaliação das características dos riscos seguráveis, são confrontadas diariamente com a gestão da certeza dos eventos, em vez do carácter aleatório das ocorrências de sinistros automóveis.

Assim, numa primeira fase importa proceder à análise da composição da sinistralidade para melhor compreender o fenómeno, nomeadamente estabelecer e modelar relações de causa-efeito, identificar factores de risco com efeitos significativos, tipificar eventuais acidentes e comportamentos de aversão/apetência pelo risco.

Posteriormente, com base na selecção de uma amostra aleatória de indivíduos da população em estudo de clientes de uma dada seguradora, num determinado conjunto de variáveis regressoras a eleger, e utilizando metodologias estatísticas adequadas, designadamente a família dos modelos lineares generalizados, de que se destaca a análise da regressão logística e a análise de regressão de dados de contagem (modelos de regressão de poisson e binomial negativo), propõe-se uma correcta identificação, estimação e quantificação da propensão de cada indivíduo para o acidente automóvel.

O modelo desenvolvido para a análise do risco será estimado, testado e validado, de modo a vir a ser implementado numa companhia de seguros, como instrumento de gestão no âmbito da sinistralidade dos seus clientes actuais e potenciais.

PALAVRAS-CHAVE

Seguros, sinistralidade automóvel, risco, risco moral, informação assimétrica, selecção adversa, modelo econométrico, modelos lineares generalizados, propensão ao risco.

ABSTRACT

The purpose of this work is to develop an econometric model, according to the regression analysis, allowing an insurance company to assess the risk of each driver either in potency power or effective, depending on their propensity to traffic accident.

The rate of road accidents has always been high in Portugal and insurance companies, despite having several tools to evaluate the characteristics of insurable risks are daily confronted with the management of certain events, instead the random nature of the occurrences of drive insurance claims.

Thus, initially it will de done the examination of the composition of the car accident to better understand better the phenomenon, namely establishing risk factors, criminalizing accidents and behaviours of risk aversion/risk appetite.

Subsequently, based on the selection of a random sample of individuals in the population under study belonging to a specific insurance company, a given set of explanatory variables elected, and using proper statistical methodological tools, such as the family of generalized linear models, that stands out the logistic regression analysis and the regression analysis of count data (poisson regression model and negative binomial models), proposes an accurate identification, quantification and estimation of the propensity of individuals to the automobile accident.

The model developed for the analysis of risk will be estimated, tested and validated in order to eventually be implemented in an insurance company as a management tool within the framework of customers claims and potential customers.

KEY-WORDS

Insurance, automobile claim, risk, moral hazard, asymmetric information, adverse selection, econometric models, generalized linear models, risk proneness.

ÍNDICE

ÍNDICE DAS FIGURAS

ÍNDICE DAS TABELAS

LISTA DE SIGLAS E ABREVIATURAS

CEA – European Insurance and Reinsurance Federation

C. S. – Companhia de Seguros

GLM – Generalized Linear Models

LPM – Linear Probability Models

NB1 – Modelo Binomial Negativo de ordem 1

NB2 – Modelo Binomial Negativo de ordem 2

OLS – Ordinary Least Squares

UE – União Europeia

v.a – variável aleatória

1. INTRODUÇÃO

Neste capítulo será apresentado o tema da sinistralidade automóvel, focando a sua relevância na sociedade contemporânea, e delimitando a sua abordagem à necessidade da avaliação da propensão individual para o acidente.

1.1. Enquadramento

Em Portugal, até final dos anos 70 o seguro automóvel não era obrigatório (Vieira & Quintero, 2008), o custo de aquisição de uma viatura era elevado, a sua utilização restrita, e o acidente era considerado um azar, sendo encarado como um preço inevitável a pagar pela modernização do país.

Com o crescimento económico da década de 70, verificou-se um aumento na circulação automóvel (Schmidt & Nave, 2004). Contudo, a generalização do uso do automóvel não trouxe apenas benefícios: associado a ela surgiu o fenómeno da sinistralidade rodoviária.

O aumento da exposição dos cidadãos aos riscos decorrentes da circulação automóvel esteve na origem de um aumento dos instrumentos de protecção, tendo o Dec.-Lei nº 408/79 transformado o seguro automóvel em seguro obrigatório.

Apesar do risco passar a estar, obrigatoriamente, coberto por uma companhia de seguros mediante o pagamento de um prémio de seguro pelo tomador, a exposição ao risco associada à circulação automóvel não diminuiu e actualmente existe com um grave problema de segurança pública que tem como origem a sinistralidade automóvel.

1.2. Objectivo

É objectivo desta dissertação desenvolver um modelo econométrico, no âmbito da análise de regressão, que permita avaliar o risco de cada indivíduo para o acidente.

Tabela 1.1 - *Causas de morte na União Europeia em 2008 (por 100.000 habitantes)*

Países da EU	Cancro	Sistema Circulatório	Doenças Cardíacas	Doenças Respiratórias	Acidentes Automóvel
UE-27	**173**	**227,2**	**84,1**	**44,7**	**8,3**
Bélgica	147,5	196,2	67,5	68,9	10,6
Bulgária	171,6	611,3	126	41,7	13,3
Republica Checa	201	335,8	176,2	40,2	10,3
Dinamarca	208	193,7	71,6	60,6	5,8
Alemanha	162,6	229,2	86,4	37,7	5,4
Estónia	100,3	451,4	224,4	26,5	11,4
Irlanda	176,7	190,7	102,3	64,8	6,2
Grécia	157,2	258,9	67,3	53,5	14,1
Espanha	154,6	151,4	47,4	52,8	7,2
França	166	124,7	33,8	27,3	6,9
Itália	163,7	179,1	62	29,6	9,2
Chipre	121,6	208,6	73,9	36,3	11,6
Letónia	191	505,9	263,5	25	15,9
Lituânia	195	520,1	321,3	39,5	16,8
Luxemburgo	167,7	210,8	63,8	43,4	8,7
Hungria	241,7	428,6	216,9	43,4	11,7
Malta	155	231,5	119,9	52,2	3,6
Holanda	184,4	159,3	46,8	53,4	4,1
Áustria	161,6	212,7	97,4	28,6	7,4
Polónia	201,3	356,4	102,2	40	14,6
Portugal	**155,6**	**184,9**	**44,4**	**62**	**9,1**
Roménia	170,7	557,9	140,1	49,5	16,66
Eslovénia	201,9	234,9	67,4	36,4	11,5
Eslováquia	201,7	455	290,5	49,9	13,3
Finlândia	137	224	128,8	22,3	6,9
Suécia	149,1	200,9	93	30,8	5
Reino Unido	178,1	188,7	93	73,7	5,3

Fonte: Eurostat

Da tabela 1.1 (Eurostat, 2010) conclui-se que a sinistralidade rodoviária é uma das principais causas de morte na Europa, e Portugal é um dos países onde este fenómeno toma dimensão mais apreciável.

Apesar das campanhas de prevenção rodoviária para sensibilização da população e da legislação portuguesa ser mais rigorosa e punitiva com as infracções ao Código da Estrada, a redução da sinistralidade nas estradas portuguesa pouco diminui (Oliveira, 2007).

Nestes termos e porque o cliente de risco é definido pela sua probabilidade para o acidente (Rothschild & Stiglitz, 1976), importa conhecer a propensão dos indivíduos para os acidentes de viação. Trata-se de uma questão de elevada importância no sector segurador, quer na concepção de novos produtos, (ajustando coberturas e valores de capitais às características de cada cliente), quer na correcta tarifação dos riscos que uma seguradora se proponha garantir.

A nível mundial, os prejuízos para a sociedade causados por acidentes de viação são consideráveis (Lord, 2000). Após um acidente, as vítimas e respectivas famílias ficam muitas vezes com as vidas transformadas, e tal verifica-se quer a nível emocional quer a nível económico (Blincoe et al., 2002), dependendo das companhias de seguros para, de algum modo, refaze-las.

1.3. Estrutura

Com vista ao cumprimento do objectivo proposto, este trabalho será estruturado em 7 capítulos organizados da seguinte forma:

O capítulo 1 é introdutório, segue-se o capítulo 2 de revisão bibliográfica onde será apresentada a problemática da origem da propensão para o acidente, a necessidade da avaliação do risco, bem como alguns estudos que efectuam a abordagem do tema e onde é efectuada uma apresentação e caracterização do sector segurador automóvel.

O capítulo 3 descreve detalhadamente a metodologia adoptada para a concretização do objectivo do trabalho, a saber, a modelação do risco de acidente automóvel: a análise de regressão logística e os modelos de regressão de dados de contagem (Poisson e Binomial negativo).

O capítulo 4 incide sobre a análise e modelação de dados, com vista ao encontro de variáveis explicativas para estimação dos modelos.

No capítulo 5 serão apresentados os modelos econométricos de regressão.

No capítulo 6 serão apresentados e discutidos os resultados estatísticos dos modelos.

Por último, no capítulo 7 apresentam-se as conclusões obtidas e serão apresentadas propostas para trabalho futuro.

2. REVISÃO DA LITERATURA

Neste capítulo apresentam-se os principais trabalhos e os autores mais relevantes que se dedicaram ao estudo do fenómeno da sinistralidade rodoviária. Essas referências provêm das mais diversas áreas do conhecimento, desde a psicologia à sociologia e à literatura estatística e econométrica. Procura-se extrair de todas os aspectos mais relevantes que permitam alcançar atingir o objectivo proposto por esta dissertação.

2.1. Contexto histórico

No início do século XX com a sociedade a tornar-se cada vez mais industrializada verifica-se um aumento da frequência com que os acidentes ocorrem. A falha/erro industrial rapidamente se torna uma área de estudo da psicologia de investigação, e é nessa época e nesse contexto que se conclui acerca da existência de causas que dão origem a distúrbios funcionais de menor importância no quotidiano de pessoas saudáveis, os quais levam ao erro (Freud, 1901).

O conceito de propensão para o acidente surgiu mais tarde, nos anos 20 do século passado (Burnham, 2009, p. 32), sendo considerada uma característica pessoal de cada indivíduo, a qual estabelece uma relação de causa-efeito entre o número excessivo de acidentes e a propensão para o acidente.

No seu trabalho, Burnham (2009) revelou o resultado de um estudo efectuado durante 1929-1931 por uma empresa americana de transportes, a Dayton Power & Light Company, a qual utilizando o registo de acidentes com os motoristas da própria empresa trabalhou a teoria de que um limitado grupo de condutores seria responsável por um larga percentagem de acidentes rodoviários. Com a publicação dos resultados do estudo em 1931 surgiu o termo “propensão para o acidente” e os condutores foram divididos em dois grupos “os bons condutores”, ou condutores não propensos para o acidente, e os “maus condutores”, condutores com propensão para o acidente.

Com base nestes resultados, surgiram posteriormente vários estudos a desenvolverem técnicas de selecção e identificação dos indivíduos com baixa propensão para acidentes (os mais desejados pelas empresas).

Ainda de acordo com Burnham (2009), foi no pós II Guerra Mundial que estes trabalhos e estudos se consolidaram e implementaram no mercado. Termos como "selecção" e "exclusão" de pessoas passaram a fazer parte do quotidiano empresarial como forma garantir a minimização de perdas económicas devido a acidente.

No entanto, as escolhas dos "bons condutores" nem sempre se mostrou uma decisão acertada o que juntamente com as consequências de que os "maus condutores" passaram a ser alvo, causaram impactos sociais e económicos dentro e fora das organizações levando muitos a por em causa se a propensão para o acidente realmente existiria.

De facto, o conceito de "propensão para o acidente" foi exaustivamente utilizado para enfatizar que alguns condutores configuram um risco desproporcional na estrada e deveriam, por isso, ser impedidos de circular de automóvel (Hoffman, 2005), mas com o aumento da circulação rodoviária a que se tem assistido nas últimas décadas esta posição tornou-se cada vez menos sustentável. Porém, o foco da questão da propensão para o acidente deixou de se centrar nas capacidades relacionadas com o desempenho, e foi transferido para factores sociais relacionados com a disposição para o risco.

2.2. Factores potenciadores da sinistralidade

A generalização da circulação automóvel ampliou as condições humanas de locomoção, um aspecto fundamental quer da individualidade quer da sociabilidade do Homem, uma vez que redefine a sua autonomia em relação ao tempo, mas principalmente em relação ao espaço (Schor, 1999).

As transformações nos meios de transporte, especialmente no automóvel, com as suas novas práticas e necessidades sociais carregam consigo formas de conduta social (e.g. como conduzir, como comportar-se ao volante, etc.), porque o automóvel tornou-se em algo mais do que num meio de deslocação: ele representa o poder e o *status* social de um tipo de individuo (Gordon, 2004 e Schor, 1999).

Recentemente, aprofundaram-se estudos (e.g. Harvey, 2004; Oliveira, 2007) que têm vindo a confirmar que o principal factor que dá origem aos acidentes é o erro humano, acrescido de factores externos (e.g. condições da estrada, condições climatéricas, condições do veículo).

Os resultados de alguns estudos analisados por Hoffmann (2005) apontam no sentido de que o envolvimento em acidentes de viação está relacionado com o comportamento social divergente, com a motivação social, e com um conjunto de determinadas variáveis demográficas.

Constatou-se (Petridou & Moustaki, 2000) que 3 em cada 5 acidentes resultam de factores comportamentais do condutor e que são os condutores mais qualificados para a condução que registam mais acidentes do que os outros condutores.

Uma revisão concisa por alguns estudos e literatura contemporânea ajudam na compreensão da análise do risco perante a ocorrência de um evento adverso. Destacam-se investigações à componente humana nos acidentes de viação publicadas por Bailey & Simon (1960) e Petridou & Moustaki (2000), que concluem pela existência de inúmeros factores que contribuem para a redução da capacidade de condução dos indivíduos: idade, doença, inexperiência, alcoolismo, fadiga, *stress*, conhecimento das regras de trânsito e vícios, influenciam a sua propensão para o acidente.

De acordo com Visser et al. (2007), mesmo que não seja identificável a componente de propensão para o acidente, ao observar-se a distribuição dos acidentes de viação pela população em geral verifica-se haver um número de indivíduos, maior do que seria de esperar, no qual se observa o fenómeno de "repetição de acidentes".

Tal situação decorre do facto dos indivíduos terem diferentes atitudes perante o risco (Mulligan, 1989), podendo ser divididos em 3 grupos: os que têm apetência pelo risco, os avessos ao risco e os indiferentes ao risco.

Vários estudos na área da sociologia sobre acidentes rodoviários realçam o facto de nestes eventos ser tido em conta o sistema de valores e crenças do sujeito interveniente (atitude face ao risco), na tentativa de compreender o nível de risco que o próprio considera como aceitável (Kouabenan, 1998).

Existem modelos motivacionais que focam a reacção dos indivíduos perante o risco: a Teoria Homeostática de Risco (Wilde, 1982, citado por Hoffmann, 2005), a Teoria do Risco Zero (Summala, 1988, citado por Hoffmann, 2005) e a Teoria para Evitar a Ameaça (Fuller, 1984, citado por Hoffmann, 2005). Estas teorias defendem que os condutores adaptam o seu comportamento tendo por base o risco percebido (ou antecipado) em comparação com o nível de risco aceitável.

Apesar desta característica individual sobre a atitude perante o risco, o facto da mesma estar directamente relacionada com a experiência partilhada transporta-a de um nível individual para o universo colectivo (Dake, 1991), sendo possível extrair da

sociedade uma ideia de "cultura de segurança" reportando-se esta ao conjunto de crenças, normas, atitudes e práticas sociais através das quais os indivíduos identificam e evitam situações potencialmente perigosas. Assim, não surpreende que Mulligan (1989) tenha constatado que no âmbito das subscrições de apólices de seguro (coberturas e capitais) a maioria das pessoas é avessa ao risco.

No entanto, apesar da maioria da população ser avessa ao risco, a propensão para o acidente por grupos existe (Visser et al., 2007, p. 562) e o seu estudo é muitas vezes dificultado dado o grande número de variáveis que a determinam, mas não só.

No âmbito da circulação rodoviária (Peden et al., 2004) o risco para o principal factor que dá origem aos acidentes - o factor humano - é constituído por quatro elementos fundamentais:

- A exposição (elemento constituído pela quantidade de deslocações/viagens) efectuadas por um indivíduo;
- A probabilidade implícita de acidente tendo em conta uma exposição particular;
- A probabilidade de lesões por acidente;
- Consequências das lesões.

Face ao exposto, não só o conhecimento de características do segurado, mas também o histórico dos seus acidentes é importante para avaliação do risco no momento da contratação de um seguro.

2.3. A importância da informação disponível

A identificação da propensão para ao acidente baseia-se, assim, na informação recolhida sobre determinado indivíduo. No entanto, essa informação pode ser parcialmente omitida pelo próprio e, quando tal acontece, o mesmo é incluído numa classe de risco à qual não pertence (Dionne, 2000). Assim, para além das variáveis observáveis *à priori* (idade, sexo, morada, anos de carta de condução, etc.), há um conjunto de atributos (variáveis não observáveis à partida, mas sim *à posteriori*) e que definem e classificam um indivíduo em matéria de avaliação do seu nível de risco (Denuit et al., 2007).

Quando a assimetria de informação é significativamente desigual entre o indivíduo e a seguradora (possuindo o indivíduo mais informação do que aquela que

muitas vezes transmite) a probabilidade de ocorrência de sinistro pode perder o seu carácter puramente aleatório (Dionne & Laberge-Nadeau, 1999).

Há fortes razões para pensar que a assimetria de informação tem um peso muito significativo na indústria seguradora automóvel (Chiappori & Salanié, 2000) e consequentemente as empresas seguradoras têm dificuldade em julgar o grau de risco de quem pretende comprar uma apólice (Denuit et al., 2007), o que conduz a uma selecção adversa, ou seja os "maus condutores" procuram vantagens no seguro do seu risco. No âmbito da selecção adversa, aumenta-se a probabilidade de um contrato de seguro ser celebrado a um "mau condutor".

Nestes termos, uma companhia de seguros deverá ter um conjunto de instrumentos que permitam distinguir os diferentes níveis de risco entre os vários indivíduos no momento da celebração do contrato de seguro. Sabendo, desde logo (Chiappori & Salanié, 2000) que a partir do momento em que o contrato é celebrado o segurado passa a ter comportamentos mais arriscados do que os que tinha antes de ter o seguro por saber que transferiu a sua responsabilidade para outrém; no fundo trata-se de um "risco moral" que tem de ser assumido pela seguradora.

Torna-se claro que o trabalho prévio da seguradora é fundamental para garantir uma correcta avaliação do risco sob pena de perda de recursos e de mau funcionamento do mercado como um todo.

2.4. A actividade seguradora

Apesar do seguro, na sua concepção formal e técnica ser instrumento de gestão de risco recente, a verdade é que o Homem desde a antiguidade desenvolve e aperfeiçoa formas de minimizar os seus prejuízos (Gilberto, 2010).

O princípio de mutualidade, desde a antiguidade, que envolve a partilha de fundos de diversas entidades (exposições) para pagar os prejuízos/perdas sofridos por alguns. No entanto, foi durante o período da revolução industrial inglesa que o seguro recebeu o seu grande impulso. Actualmente, o seguro funciona como uma transferência de risco de uma entidade exposta ao risco (segurado) para uma companhia de seguros que passa a assumir o pagamento dos prejuízos, caso se verifique a ocorrência de um

sinistro. Esta transferência de risco é feita mediante o pagamento de um prémio que o segurado faz à companhia de seguros (Op. cit).

De acordo com Maslow (1954) no seu estudo sobre as necessidades humanas, foram definidos grupos de cinco necessidades que este autor hierarquizou numa pirâmide que ficou conhecida como "Pirâmide de Maslow", e segundo a qual se defende que o Homem é motivado pelas suas necessidades às quais atribui diferentes graus de importância, e em que as necessidades de nível mais baixo são de realização prioritária face as necessidade de nível mais elevado.

As necessidades fisiológicas constam da base da pirâmide e representam necessidades relacionadas com o nosso organismo (e.g. alimentação), depois de garantidas estas necessidades surgem no Homem as necessidades de segurança (e.g. segurança, protecção). É neste contexto que a actividade seguradora se insere, pois sem o seguro – o mesmo seria dizer sem segurança – não haveria estabilidade individual, social ou económica, sem os quais o crescimento e desenvolvimento económicos ocorreriam muito lentamente (Gilberto, 2010), o seguro como instrumento que garante a diminuição da incerteza é essencial para o desenvolvimento económico global, não sendo concebível o desenvolvimento económico sem uma indústria seguradora forte, conforme tabela 2.1 (CEA, 2007), a qual possa garantir a cobertura de riscos que as empresas e/ou particulares não têm capacidade para garantir pela sua conta.

Tabela 2.1 - *Dados do Sector Segurador no ano de 2007*

	Total de prémios (*)	Prémios médios *per capita* (*)	Emprego gerado	Nº de Seguradoras
CEA	1.059.319	2.026	996.686	5.124
União Europeia (27)	1.001.812	2.271	914.693	4.741
Zona Euro (15)	674.456	2.152	618.249	2.843
Portugal	15.332	1.296	11.295	81

(*) em milhões de Euros

Fonte: CEA

2.5. O mercado segurador em Portugal

Em Portugal, o sector segurador teve um desenvolvimento notável nas últimas década, profissionalizou-se, tornou-se mais eficiente, consolidou-se financeiramente, tornou-se mais forte, mais solvente e ganhou importância na actividade económica portuguesa (Gilberto, 2010).

De acordo com elementos divulgados pelo Instituto de Seguros de Portugal (2011), o mercado segurador contabilizou nesse um valor de 16,3 mil milhões de Euros (o que corresponde a cerca de 10% do valor do PIB), correspondendo a um acréscimo de 12,5% face a 2009. Ainda, de acordo com o Instituto de Seguros de Portugal, o ramo Não Vida (onde se inclui o ramo automóvel) evidenciou um crescimento de 0,7% face ao ano anterior. Este modesto crescimento é influenciado pelo contexto de redução/moderação salarial que está a conduzir a uma efectiva perda de poder de compra. Assim, se o parque automóvel não cresce, consequentemente vendem-se menos apólices (Op. cit.) e uma melhor selecção de riscos assume maior importância.

Não existem seguros que indemnizem todo o tipo de sinistros, mas ao passarem a ser partilhados, passam a ser minimizados ou mesmo anulados, e existem inúmeros instrumentos disponíveis no mercados para o poder fazer (Jorion, 2000).

O "segredo" do negócio segurador consiste em fazer um trabalho de selecção de riscos superior ao das seguradoras concorrentes, calculando correctamente o preço dos riscos aceites de forma a garantir o pagamento dos prejuízos decorrentes do sinistro, a suportar os custos operacionais e a remunerar o capital do accionista (Gilberto, 2010).

2.6. A importância do ramo automóvel

O ramo automóvel faz parte dos oito mercados relevantes do ramo Não Vida na sua componente geográfica e de produto da actividade seguradora, caracterizando-se pela existência de um número considerável de companhias de seguros, verificando-se, uma concentração do volume de negócios num número relativamente reduzido de seguradoras (Gata, 2006). O número de empresas de seguros a actuar em regime de estabelecimento, reduziu-se de 87 em 2009 para 83 em 2010. Contudo essa concentração não se reflectiu na produção que atingiu os 16.427 milhões de euros, traduzindo-se num crescimento de 12,7% face a 2009, (Instituto de Seguros de Portugal, 2011).

Tendo em conta o número de automóveis em circulação, actualmente cerca 5 milhões, conforme informação da Associação Portuguesa de Seguradores (2010), o agrupamento Automóvel continua a ser o mais representativo segmento da actividade seguradora no ramo Não Vida.

Cada seguradora é livre de fixar os seus preços incluindo o do seguro, obrigatório por lei, de responsabilidades civil automóvel. Factores como a idade do veículo, a idade do condutor e número de anos da carta de condução influenciam o nível de risco e consequentemente, o preço final a pagar por cada cliente, de acordo com a tabela de penalizações/bonificações específica de cada segurador (Instituto de Seguros de Portugal, 2007).

O sistema utilizado pelas seguradoras é o sistema de *bonus-malus*, que regula o prémio de seguro, segundo o qual, de acordo com o número de sinistros participados com culpa o prémio no ano seguinte será diminuído (*bonus*) caso não se registe sinistro, ou será agravado (*malus*) caso se registem sinistros (Lemaire, 1995).

Os preços das apólices constituem o factor determinante na escolha do consumidor particular nos "seguros de massas", assim designados por terem uma forte componente de seguro obrigatório, e do qual se destaca o seguro automóvel (Instituto de Seguros de Portugal, 2007).

Importa referir que no período em análise (de 2000 a 2010), apesar da componente obrigatória que o seguro automóvel possui, o crescimento apresentado por este ramo deve-se, também, à subscrição de coberturas complementares com comportamento crescente (Op. cit.).

O seguro automóvel constitui um dos mercados de maior peso em termos de volume de negócio (Gata, 2006), e reveste-se de uma enorme importância porque constitui desta forma uma "porta de entrada" para a oferta de outro tipo de produtos.

Assim, a angariação do cliente pela via do seguro automóvel é uma estratégia muito utilizada pelas seguradoras e seus mediadores, e como o preço é um factor decisivo na escolha do cliente, apresentar o preço mais atractivo (mais baixo) faz a diferença para conseguir concretizar a venda.

Perante a grande concorrência existente no mercado segurador os clientes facilmente mudam de companhia de seguros à procura de maiores descontos que se traduzam num preço mais reduzido para o seu seguro. É neste cenário concorrencial que o sistema de *bonus-malus* se torna ineficiente, pelo menos no mercado português, (Guerreiro & Mexia, 2003) dado que o preço que segurado paga não traduz o seu nível de risco. Várias razões levam a que isso aconteça, fundamentalmente são:

- Razões comerciais, como o seguro automóvel é determinante na angariação e fidelização de cliente, por isso o *bonus-malus* muitas vezes não reflecte o nível de risco do cliente, mas sim o seu peso comercial na carteira de clientes;

- Razões de informação assimétrica, dado que é prática comum entre os segurados portugueses logo após um sinistro (de preferência antes de definidas as responsabilidades no acidente) mudarem de seguradora, declarando na nova seguradora não terem tido sinistros.

2.7. O risco no ramo automóvel

Nos dias de hoje tornou-se extremamente difícil para as Companhias de Seguros, manterem as diferenças de preços entre as diversas categorias de risco num mercado cada vez mais competitivo. A projecção de uma estrutura tarifária que permita distribuir equitativamente o custo dos sinistros pelos seus clientes obriga à procura de factores de classificação adicional, que muitas vezes não são exigidos por lei, nem pela teoria actuarial, mas pela competição entre seguradoras. (Denuit et al., 2007).

O valor do prémio é decisivo como factor de compra por parte dos clientes e de primordial importância para a seguradora. O seu cálculo depende de um modelo estatístico que incorpore toda a informação relevante e disponível sobre o risco. É objectivo primordial de uma seguradora avaliar de forma tão precisa quanto possível o prémio puro para cada segurado, o que é feito por meio de técnicas de análise de regressão (Op. cit.)

Assim, sem por em causa a competitividade de uma companhia de seguros, nesta dissertação, irá procurar-se avaliar cuidadosamente o nível de risco individual dos seus actuais e potenciais clientes, garantindo o equilíbrio técnico da seguradora e permitindo o desempenho da sua contribuição para o bem estar social.

3. METODOLOGIA

Neste capítulo é apresentada a metodologia utilizada na construção dos modelos.

3.1. Conceitos Preliminares

As definições apresentadas de seguida podem ser recolhida em qualquer obra clássica ou de texto sobre análise de contagem de dados (e.g., Agresti, 1996, Balakrishnan, 1991; Cameron & Trivedi, 1998; Hosmer & Lemeshow, 2000; McCullagh & Nelder, 1989; Winkelmann, 2000).

Definição 1 (Processo de contagem)

Um processo estocástico $\{N(t), t \in T\}$ *é um processo de contagem se N(t) representa o número total de acontecimentos que ocorrem até t, com* $t \geq 0$, *verificando:* $N(t) \geq 0$; *N(t) toma valores inteiros; para* $0 \leq s \leq t$, $N(s) \leq N(t)$ *e para* $0 \leq s \leq t$, $N(t) - N(s)$ *é o número de acontecimentos que ocorrem no intervalo (s,t).*

Definição 2 (Processo de Poisson)

O processo de contagem, $\{N(t), t \in T\}$, *é um processo de Poisson com uma taxa* $\lambda > 0$ *se N(t)* $t \geq 0$, *satisfaz: P (N(0) = 0) =1;* $N(t) - N(s)$ *tem distribuição de Poisson com valor médio* $\lambda (t - s)$, $0 \leq s < t$, *e* $N(t_1)$, $N(t_2)$ - $N(t_1)$, …, $N(t_n)$ - $N(t_{n-1})$ *são independentes para* $0 \leq t_1 < \ldots < t_n$.

Teorema 1 (Distribuição Exponencial)

Se um acontecimento ocorre de acordo com um processo de Poisson com taxa λ *constante, sendo os tempos decorridos entre dois acontecimentos consecutivos independentes, a variável aleatória* $X_i = t_i - t_{i-1}$ *tem distribuição exponencial com valor médio* $1/[\lambda(t_i - t_{i-1})]$, *isto é, tem função distribuição cumulativa dada por:*

$$F(x) = P(X_i \leq x) = 1 - \exp(-\lambda x), \quad x \geq 0.$$

Teorema 2 (Distribuição de Poisson)

Se acontecimentos ocorrem de acordo com um processo de Poisson com taxa λ*, independentemente do tempo, então a distribuição do número de acontecimentos* Y_i *que ocorrem no intervalo de tempo* (t_{i-1}, t_i) *tem distribuição de Poisson com valor médio e variância* $\lambda(t_i - t_{i-1})$*, isto é, tem uma função massa de probabilidade dada por:*

$$P(Y_i = y) = \frac{\exp[-\lambda(t_i - t_{i-1})][\lambda(t_i - t_{i-1})]^y}{y!}, \; y = 0,1,2,\ldots$$

Teorema 3 (Distribuição de Binomial Negativa)

Seja Y *o número de acontecimentos que ocorrem de acordo com uma distribuição de Poisson com valor médio* λ*. Se* λ *for aleatório com distribuição gama parâmetros* γ *e* δ*, isto é, com função densidade de probabilidade:*

$$f(\lambda) = \left[\gamma^{\delta}\Gamma(\delta)\right]^{-1} \lambda^{\delta-1}\delta^{\delta} \exp\left(\frac{-\lambda\delta}{\gamma}\right), \lambda > 0, \;\; \gamma > 0, \;\; \delta > 0,$$

então Y tem distribuição binomial negativa de parâmetros γ *e* δ*, isto é, tem a seguinte função massa de probabilidade:*

$$P(Y = y \mid \gamma, \delta) = \frac{\Gamma(y+\delta)}{\Gamma(y+1)\Gamma(\delta)}\left(\frac{\delta}{\delta+\gamma}\right)^{\delta}\left(\frac{\gamma}{\delta+\gamma}\right)^{y}, \; y = 0,1,2,\ldots$$

em que $\Gamma(.)$ *é a função gama.*

O valor médio e a variância de Y são:

$E[Y] = \gamma$

$Var[Y] = \gamma + \frac{1}{\delta}\gamma^2$

Note-se que $Var[Y] > E[Y]$.

Teorema 4 (Distribuição Logística)

Considerando a existência de "n" variáveis aleatórias independentes $Y_i \sim B(1, \pi_i)$*:*

$f(y_i \mid \pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$*, onde* $y_i = 0,1$

e cada individuo i, está associado a um vector de especificação z_i*, resultante do vector de covariáveis* x_i*,* $i = 1,\ldots,n$.

Como E(Y $_i$)= π_i e, se tem para este modelo, $\theta_i = ln\left(\frac{\pi_i}{1-\pi_i}\right)$, *ao fazer* $\theta_i = \eta_i = z_1^t \beta$, *conclui-se que a função de ligação canónica é a função "logit". Assim, a probabilidade de sucesso, ou seja P(*Y_i *=1)=* π_i, *está relacionada com o vector* z_i *através de* $\pi_i = \frac{\exp(z_i^t \beta)}{1+\exp(z_i^t \beta)}$.

Assim, a função de distribuição logística é definida por F(x)= $\frac{\exp(x)}{1+\exp(x)}$.

Considere-se um vector de *k* variáveis explicativas $\mathbf{x}^t = (x_1, x_2, ..., x_k)$. Seja cada observação *i* deste vector representada por $\mathbf{x}_i^t = (x_{i1}, x_{i2}, ...x_{ik})$, *i* = 1,…, *n*. Podem ser especificadas várias funções de ligação dos parâmetros γ e δ às covariáveis **x**. De seguida apresentam-se as especificações mais comuns.

Definição 3 (Função de ligação do parâmetro γ às covariàveis)

A especificação mais comum para γ *é:* $E[Y_i] = \gamma_i = \exp(\mathbf{x}_i^t \boldsymbol{\beta})$, *em que* $\boldsymbol{\beta}^t = (\beta_1, \beta_2, ..., \beta_k)$ *é um vector de k parâmetros.*

A especificação para δ é importante, na medida em que pode decidir a forma de heteroscedasticidade que se assume. Seguem-se as funções de ligação do parâmetro δ às covariáveis.

Definição 4 (Regressão de Poisson)

É o caso mais simples, em que se toma $\delta \to +\infty$, *resultando, no limite, a média igual à variância:* $E[Y_i] = Var[Y_i] = \gamma_i$, *situação que se define como de igual dispersão.*

Definição 5 (Regressão Binomial Negativa de ordem um)

Admite-se que δ *é uma função das observações da forma* $\delta_i = (1/\alpha)\exp(\mathbf{x}_i^t \boldsymbol{\beta})$, *verificando-se que a variância é uma função linear do valor médio:*

$$Var[Y_i] = (1+\alpha) E[Y_i] = (1+\alpha)\gamma_i,$$

em que α *é uma constante.*

Note-se que a razão entre a variância e o valor médio é constante. Caso $\alpha > 0$*, verifica-se sobredispersão, isto é,* $Var[Y_i] > E[Y_i]$*. Quando* $\alpha < 0$ *verifica-se subdispersão, isto é,* $Var[Y_i] < E[Y_i]$.

Definição 6 (Regressão Binomial Negativa de ordem dois)

Para $\delta_i = 1/\alpha$*, verifica-se que a variância é uma função quadrática do valor médio:* $Var[Y_i] = E[Y_i] + \alpha(E[Y_i])^2 = \gamma_i + \alpha\gamma_i^2$

Note-se que a razão entre a variância e o valor médio é uma função linear do valor esperado.

Definição 7 (Regressão Binomial Negativa de ordem p)

Corresponde ao caso geral. Admite-se que $\delta_i = \frac{1}{\alpha[\exp(x_i^t \beta)]^{p-2}}$*, verificando-se que a variância é dada por* $Var[Y_i] = E[Y_i] + \alpha(E[Y_i])^p = \gamma_i + \alpha\gamma_i^p$

Note-se que as primeiras, segundas e terceiras especificações obtêm-se fazendo, respectivamente, p=0, p=1, p=2, isto é, cada uma das especificações anteriores são casos particulares desta última.

3.2. Os modelos Lineares Generalizados

Os modelos de regressão pretendem estudar a relação entre variáveis, ou seja, analisar a influência que uma ou mais variáveis explicativas, ou regressores, observadas em indivíduos ou objectos, têm sobre uma variável de interesse, designada de variável resposta ou dependente conforme Balakrishnan (1991), Denuit et al. (2007), Draper & Smith (1998), Johnston & Dinardo (1996), Long (1997), McCullagh & Nelder (1989) e Winkelmann (2000).

O modelo de regressão linear clássico normal surgiu no século XIX dominou a modelação estatística até meados do século XX (McCullagh & Nelder, 1989), embora alguns modelos não lineares tenham, entretanto, sido desenvolvidos para fazer face a situações não adequadamente explicadas pelo modelo de linear: o modelo complementar *log-log* para ensaios de diluição (Lindsey 1997, citado por McCullagh &

Nelder, 1989); os modelos *probit* (Bliss, 1935, citado por McCullagh & Nelder, 1989) e modelo *logit* (Berkson, 1994, citado por McCullagh & Nelder, 1989 e Rash, 1960, citado por McCullagh & Nelder, 1989) entre outros. De acordo McCullagh & Nelder (1989) todos os esses modelos apresentam em comum o seguinte:

- Uma estrutura de combinação linear dos regressores;
- O facto da variável resposta seguir uma distribuição com propriedades muito específicas: a família exponencial.

Os GLM são a uma síntese alargada dos modelos acima apresentados e pretenderam unificar a teoria da modelação estatística até então desenvolvida neste âmbito (Op. cit.).

Uma das primeiras aplicações dos modelos lineares generalizados em análise de seguros foi realizada por Brockman & Wright (1992), consideravam que a análise dos dados dos veículos e a sua melhor avaliação contribuía para uma melhor avaliação do risco e formação do prémio de seguro.

Nas últimas décadas, tem-se verificado um grande desenvolvimento de modelos estatísticos para análises estudos no âmbito dos acidentes de viação, sendo ponto comum a todos eles o facto de recorrem à utilização dos GLM como metodologia para a sua estimação (e.g. Abbess et al., 1981; Guikema & Coffelt, 2007; Hauer et al,. 1988; Kumala, 1995; Lord, 2000; Lord et al. 2005, Miaou & Lord, 2003; Miaou & Song, 2005; Poch & Mannering, 1996; Xie et al., 2007).

3.2.1. A família exponencial

Conforme foi referido anteriormente, os GLM pressupõem que a variável resposta tenha uma distribuição pertencente a uma família particular, a família exponencial. A definição apresentada por McCullagh & Nelder (1989) é a mais adequada para a variável resposta que interessa considerar.

Definição 8 (Família exponencial)

Diz-se que uma variável aleatória Y tem distribuição pertencente à família exponencial de dispersão (ou simplesmente família exponencial) se a sua função

densidade de probabilidade (f.d.p) ou função de massa de probabilidade (f.m.p) se se puder escrever na forma: $f(y \mid \theta, \phi) = \exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$,

onde θ *e* ϕ *são parâmetros escalares, a(.), b(.) e c(.) são funções reais conhecidas.*

3.2.2. A estrutura dos modelos lineares generalizados (GLM)

Os GLM possuem três componentes: a componente aleatória, que identifica a distribuição de probabilidade da variável dependente; a componente sistemática que específica a estrutura linear da variáveis independentes quantitativas e/ou qualitativas, que é utilizada como preditor linear, e a função de ligação que descreve a relação funcional entre a componente sistemática e o valor esperado da componente aleatória (Agresti, 1996; Dobson, 1990; Fahrmeir & Tutz, 1994, McCullagh & Nelder, 1989).

3.2.2.1. Componente aleatória

A componente aleatória especifica uma variável aleatória y com n observações independentes e identicamente distribuídas, um vector de médias $\mu = (\mu_1, ..., \mu_n)^T$ e uma distribuição pertencente à família exponencial (Dobson, 1990, McCullagh & Nelder, 1989). De acordo com Agresti (1996), os resultados para cada observação de y são binários, designados como "sucesso" ou "insucesso", ou mais geralmente, cada y_i poderá ser definido como o número de sucessos num certo número fixo de tentativas. Assume-se, neste caso, uma distribuição binomial para a componente aleatória.

3.2.2.2. Componente sistemática

A componente sistemática específica a estrutura linear, não aleatória, das variáveis independentes quantitativas e/ou qualitativas, que é utilizada como preditor linear (McCullagh & Nelder, 1986). Para Agresti (1996), esta componente específica as

variáveis independentes que entram linearmente à direita da equação do modelo como preditores, conforme equação: $y = \alpha + \beta_1 x_1 + ... + \beta_k x_k$.

De acordo com Cordeiro (1986), a estrutura de um GLM pode ser escrita como $\eta_i = \sum_{j=1}^{k} x_j \beta_j$, onde a função linear η_i dos parâmetros desconhecidos $\beta = (\beta_1,..., \beta_k)$ é denominada de preditor linear, x_j representa os valores de k ($k < n$) variáveis independentes que são assumidas fixas (não aleatórias) e conhecidas.

3.2.2.3. A função de ligação

A terceira componente dos GLM é a função de ligação, que descreve a relação entre a componente sistemática e o valor esperado da componente aleatória (a média da variável dependente). A estrutura da função de ligação na equação do modelo pode ser representada conforme a fórmula:

$$g(\mu_i) = \alpha + \beta_1 x_1 + ... + \beta_k x_k \quad ,$$

podendo ser reescrita como $g(\mu_i) = \eta_i$, em que a função $g(\mu_i)$ segundo vários autores (Agresti, 1996; Dobson, 1990; McCullagh & Nelder, 1989, etc.), é uma função estritamente monótona e duplamente diferenciável e que possibilita modelar directamente a média da variável dependente, conforme fórmula

$\mu_i = g^{-1}(\eta_i)$; i= 1,…, n

3.2.3. Os modelos discretos

Os modelos lineares generalizados podem ter variável resposta discreta ou contínua (Maddala, 1993). Dado que a presente dissertação centra-se na ocorrência, ou não de acidentes, e no número de acidentes que ocorrem por condutor, estamos perante variáveis de interesse de natureza discreta, ou seja, a variável dependente y_i assume valores de 0, 1 ou 0, 1, 2, 3, … e a metodologia utilizada baseia-se em modelos lineares generalizados discretos para dados binários ou para dados de contagem.

3.2.4. O modelo de probabilidade linear

Long (1997), Maddala (1993), entre outros autores, apresentam como ponto de partida o *linear probability model* (LPM) aplicado a variáveis dependentes categóricas e limitadas, o qual é definido da seguinte forma:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + ... + \beta_K x_{iK} + \varepsilon_i , \quad (3.1)$$

onde x_i é um vector de variáveis explicativas relativo à i-ésima observação, β é o vector dos parâmetros a estimar e ε_i é o erro aleatório.

Se se tiver uma única variável independente o modelo de regressão passa a ser escrito da seguinte forma:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (3.2)$$

Graficamente o LPM é representado por uma recta, na qual β_0 representa a ordenada na origem, e β_1 representa o declive da recta de regressão.

O LPM determina que para um dado valor para x_i (que é conhecido), é correspondente uma observação de y_i que é consistente com o valor de $\beta_0 + \beta_i x_i$, acrescido da grandeza ε_i que representa o incremento pelo qual qualquer valor de y_i pode sair da recta de regressão (Draper & Smith, 1998). De acordo com Maddala (1983), a expectativa condicional $E(y_i \mid x_i)$ é igual a $\beta_i x_i$, que dá a probabilidade estimada de um dado evento ocorre para um dado valor particular de x_i, sendo que essa probabilidade se encontra no intervalo [0,1].

3.2.4.1. A especificação

O LPM assenta nas seguintes cinco suposições fundamentais (Long, 1997):

Suposição 1 – A linearidade nos parâmetros.

De acordo com a recta de regressão $y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + ... + \beta_K x_{iK} + \varepsilon_i$ *, a variável dependente y está linearmente relacionada com as observações de x através dos* β *parâmetros.*

Suposição 2 – Colinearidade.

Significa que as observações de x são linearmente independentes, ou seja, nenhuma das observações de x é uma combinação linear das restantes observações de x.

Suposição 3 – A média de ε é de zero.

Esta suposição reporta-se à distribuição da componente de erro ε, os quais são intrínsecos às observações de x, mas cuja influência pode não ser observada em y, ou seja, ε pode ser visto como o efeito da exclusão de um largo número de variáveis que individualmente têm reduzido efeito sobre y.

$E(\varepsilon_i \mid \mathbf{x}_i)=0$

Em suma , supõem-se que para um dado conjunto de valores de x, é expectável que no seu conjunto os erros se anulem.

Suposição 4 – Homocedastecidade e a não correlação dos erros

$Var(\varepsilon_i \mid \mathbf{x}_i)=\sigma^2$ *para todo o i*

Assume-se que para um dado conjunto de observações de x, os erros têm variância constante, e assume-se igualmente que as observações de x não estão correlacionada, ou seja, $cov(\varepsilon_i, \varepsilon_j)=0$.

Suposição 5 – Normalidade os erros

Considerando que os erros são uma combinação de inúmeros pequenos factores, assume-se que são normalmente distribuídos.

3.2.4.2. A estimação

O método dos mínimos quadrados (*OLS – Ordinary Least Squares*) é, de acordo com vários autores (Cameron & Trivedi, 1998; Draper & Smith, 1998; Johnson & Dinardo, 1996; Long, 2007; Madalla, 1983; McCullagh & Nelder, 1989) o método que com maior frequência é utilizado para estimar o modelo de regressão linear.

O estimador do método do mínimos quadrados de β é o valor de $\hat{\beta}$ que minimiza a soma dos quadrados dos resíduos tal que:

$$\varepsilon_i = y_i - \hat{y}_i = y_i - a - bx_i \text{, para } i = 1,2,\ldots, n \quad (3.3)$$

Cada par de valores de *a,b* determina uma recta diferente no modelo de regressão linear e consequentemente um diferente conjunto de resíduos. A soma dos quadrados dos resíduos é função de *a* e de *b*. O principio do métodos dos mínimos quadrados é seleccionar *a,b* que minimizem a soma dos resíduos.

$$\sum e_i^2 = f(a,b) \quad (3.4)$$

As condições necessárias para estabilizar em zero o referido valor são:

- $$\frac{\delta\left(\sum e^2\right)}{\delta a} = -2\sum\left(Y - a - bX\right) = -2\sum e = 0 \quad (3.5)$$
- $$\frac{\delta\left(\sum e^2\right)}{\delta b} = -2\sum X\left(Y - a - bX\right) = -2\sum Xe = 0 \quad (3.6)$$

3.2.4.3. Análise do modelo de regressão linear

Pela revisão da literatura, verificamos que o modelo de regressão foi amplamente utilizado para explicar diversas crises económicas e financeiras (Dooley, 1997; Flood & Garber, 1984; Krugman, 1979; Obstfeld, 1994, 1996; Ozkan & Sutherland, 1995; Radelet & Sachs, 1998), mas não só: apresenta-se igualmente em estudos de aplicações de determinados métodos de ensino (Spector & Mazzeo, 1980).

De acordo com os referidos autores, apresenta como principal vantagem o facto de ser um modelo de estimação acessível e fácil interpretação.

O valor esperado da v.a. binária *y* é que a probabilidade de que o evento ocorra:

$$E(y_i) = (1.\Pr(y_i = 1)) + (0.\Pr(y_i = 0)) = \Pr(y_i = 1) \quad (3.7)$$

Assim, para o modelo de regressão linear temos o mesmo valor esperado

$$E(y_i \mid x_i) = (1.\Pr(y_i = 1 \mid x_i)) + (0.\Pr(y_i = 0 \mid x_i)) = \Pr(y_i = 1 \mid x_i) \quad (3.8)$$

Desta forma, o valor esperado de *y* para um dado valor de *x* é probabilidade do evento ocorrer (*y* = 1) dado o valor de *x*, o que nos permite reescrever o modelo de regressão linear da seguinte forma:

$$\Pr(y_i = 1 \mid x_i) = x_i\beta \quad (3.9)$$

No entanto, de um ponto de vista teórico, os modelos LPM apresentam alguns aspectos pouco satisfatórios (Greene, 1993).

Em primeiro lugar a linearidade dos parâmetros é condição que na maior parte das situações não se verifica. Uma variável dependente dicotómica [0,1] conduz a sérios problemas de estimação pelo método dos mínimos quadrados sobretudo porque viola as suposições fundamentais do modelo de regressão linear, como por exemplo a presença de heteroscedasticidade, os erros deixarem de apresentar distribuição normal, etc. (Op, cit.). Como y_i assume o valor de 1 ou 0, o valor residual apenas pode assumir os seguintes valores: 1- $\beta' x_i$ e $-\beta' x_i$. Também, dada a interpretação da equação e o requisito de que $E(u_i)$ =0, a respectiva probabilidade destes eventos são $\beta' x_i$ e 1- $\beta' x_i$.

Assim temos:

Var $(u_i) = \beta' x_i\ (1-\beta' x_i)^2 + (1-\beta' x_i)\ (\beta' x_i)^2 = \beta' x_i\ (1-\beta' x_i) = E(y_1)\,[1- E(y_i)]$

Assim, por causa deste problema de heteroscedastecidade, o modelo dos mínimos quadrados apresenta um estimados valores de β não é eficiente (Maddala, 1983).

Em segundo lugar, se o modelo for linear na variável X_i, e as derivadas abaixo apresentadas existirem, então,

$$\frac{\delta E(y)}{\delta x_i} = \frac{\delta \Pr(y=1)}{\delta x_i} = \beta_i \qquad (3.10)$$

ou seja, se tudo o resto permanecer igual, a probabilidade de verificação do acontecimento em análise (o sinistro) é afectada sempre de modo idêntico por uma variação de X_i (factores explicativos da sinistralidade) qualquer que seja o nível da variável.

Por último, uma vez que E(Y) é uma medida de probabilidade, que estará situada $0 \leq E(Y) \leq 1$, é difícil compatibilizar a restrição $0 \leq E(Y) \leq 1$ com a possibilidade de variações absolutas constantes em E(Y) induzidas por variações unitárias de uma variável explicativa. Se E(Y) é uma função linear de X_i, para alguns valores de X_j E(Y) virá a situar-se fora do referido intervalo (Green, 1993, Maddala, 1983), o que também conduz a sérios problemas de estimação pelo método dos mínimos quadrados, sobretudo porque viola as suposições fundamentais do modelo de regressão linear, como por exemplo a presença de heteroscedasticidade, os erros deixarem de apresentar distribuição normal, etc. Assim, será necessário optar por um modelo probabilístico que satisfaça duas condições:

- À medida que X_i aumenta, $\Pr_i = E(Y = 1 | X_i)$ aumenta sem assumir valores fora do intervalo [0,1].
- A relação entre $\Pr_i$ e X_i ser não linear.

De acordo com Maddala (1983) uma abordagem alternativa aos modelos LPM é o modelo de Regressão Logística (*Logit* e *Probit)*.

3.2.5. O Modelo de Regressão Logística (*Logit*)

O modelo de regressão logística é um modelo linear generalizado para situações em que as variáveis resposta são discretas e os erros não são normalmente distribuídos, permitindo, a partir de um dado conjunto de variáveis explicativas, modelar uma variável dependente de natureza discreta com dois resultados possíveis: a presença de uma característica, ou a ausência de mesma (Agresti, 1996; Balakrishnan, 1991). É por esse motivo adequado a muitas situações, porque permite que seja analisado o efeito de uma ou mais variáveis independentes sobre uma variável dependente dicotómica (Hosmer & Lemeshow, 1989).

A vantagem da regressão logística face à regressão linear, reside no facto não impor a existência de linearidade nos parâmetros e como tal é uma regressão que permite explicar mais aprofundadamente os resultados obtidos, dado a variável resposta ser mais precisa (Rawlings et al., 1998), e sendo menos rígida nos pressupostos, torna-se uma técnica estatística bastante robusta quando as restrições iniciais são reduzidas (Hair et al., 1998)

Estes modelos aplicam-se em duas situações importantes: quando as variáveis resposta são discretas e quando os erros são normalmente distribuídos (Long, 1997).

Hosmer & Lemeshow (1989) apontam pelo menos duas razões para a utilização do modelo logístico na análise de variáveis-resposta dicotómicas:

- De um ponto de vista matemático, é bastante flexível e fácil de ser utilizado;
- Permite uma interpretação de resultados bastante rica e directa.

Salienta-se que nesta dissertação, tal como em muitos estudos (e.g. Aldrich & Cnudde, 1975; Austin et al.,1992; Cabrera, 1994 Chuang, 1997; Janik & Kravitz, 1994; Peng & So, 2002; Ragsdale, 1984; Santos et al, 2011; Tillman & Pontell, 1995; Tolman

& Weisz, 1995), há todo o interesse em que a variável resposta tenha um resultado dicotómico podendo receber os valores 0 (zero) e 1 (um), consoante determinado fenómeno ou comportamento se verifica, ou pelo contrário, não se verifica.

3.2.5.1. Especificação

Considerando o modelo do regressão linear:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (3.11)$$

Importa referir que este modelo de regressão assenta nas seguintes suposições (Aldrich & Nelson, 1984; Agresti, 1996; Maddala, 1983, McCullagh & Nelder, 1989):

Suposição 1 – *O resultado da variável resposta* Y_i *pode assumir apenas 2 valores possíveis:* $Y_i = \begin{cases} 1 \\ 0 \end{cases}$

Suposição 2 – *Por definição de valor esperado, obtém-se:* $E(Y_i) = \pi_i \qquad (3.12)$

Suposição 3 – *Desta forma, a resposta média quando a variável dependente é binária (1 ou 0) é dada por:* $E(Y_i) = \beta_0 + \beta_1 x_i = \pi_i \qquad (3.13)$

Este modelo analisa os dados binomialmente distribuídos, onde as provas de Bernoulli n_i, são conhecidas e as probabilidades de sucesso p_i são desconhecidas.

$$Y_i \sim B(p_i, n_i), i = 1,2,3,...,m,$$

3.2.5.2. Estimação

Enquanto no modelo de regressão linear a estimação dos parâmetros é efectuada através método dos mínimos quadrados, no modelo de regressão logística a estimação é feita pelo método da máxima verosimilhança (Hosmer & Lemeshow, 2000; Maddala, 1983).

De um modo geral, este método é a base da estimação por via da regressão logística, através do qual os valores para a variável resposta que maximizam a probabilidade de ser obter resultados fiéis à realidade observada. Assim, no âmbito de uma variável resposta binária e considerando que os resultado são independentes:

$$Y_i = 1 \rightarrow P(Y_i = 1) = \pi_i$$

$$Y_i = 0 \rightarrow P(Y_i = 0) = 1 - \pi_i$$

De uma forma geral (Hosmer & Lemeshow, 2000; McCullagh & Nelder, 1989), partindo do pressuposto que existe independência dos valores observados, a função de máxima verosimilhança para resultados binários é dada por:

$$L((\beta_0, \beta_1); Y_1) = \prod_{i=1}^{n} \pi(x_i)^{Y_i} (1 - \pi_i)^{1-Y_i} \quad (3.14)$$

Onde a função de resposta é dada por:

$$\hat{\pi}_i = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} = \frac{e^{\beta_0 + \beta_1 X_i}}{1 + e^{\beta_0 + \beta_1 X_i}} \quad (3.15)$$

E onde a estimativa do *logit* é dada pela seguinte equação:

$$\hat{g}(x) = \beta_0 + \beta_1 X_i \quad (3.16)$$

Assim, a função de distribuição $F : R \rightarrow [0,1]$ definida por:

$$F(x) = \frac{\exp(x)}{1 + \exp(x)}$$

é a função de distribuição logística.

Desta forma, o GLM definido pelo modelo binomial com função de ligação canónica *logit* é designado de modelo de regressão logística.

Segundo Gelman & Hill (2007) e Hosmer & Lemeshow (2000) a diferença no *logit* para um acontecimento com $x = 1$ e $x = 0$ é: $g(1) - g(0) = [\beta_0 + \beta_1] - [\beta_0] = \beta_1$

O *logit* do valor *p* varia entre 0 e 1 e é dado por:

$$logit\ (p) = \log\left(\frac{p}{1-p}\right) = \log\ (p) - \log\ (1\text{-}p)$$

Onde *p* é a probabilidade de sucesso (ocorrência do acidente) e 1-*p* corresponde à probabilidade de insucesso (não ocorrência de acidente).

Esta transformação é chamada de "transformação *logit* da probabilidade *p*" e a razão *p*/(*p*-1) na transformação *logit* é chamada de *Odds Ratios*.

3.2.5.3. O modelo *Probit*

O modelo *probit* surgiu como uma alternativa do modelo *logit* por Bliss (1935). Devido ao facto de no modelo *logit* se ter que $E(Y_i) = \mu_i \in [0,1]$, em principio não só a função de distribuição logística, como qualquer outra função de distribuição, pode ser candidata a função inversa de função de ligação (McCullagh & Nelder; 1989).

O modelo *probit* é definido da seguinte forma:

$$\pi_x = \Phi(\alpha + \beta_x) \tag{3.17}$$

onde $\Phi(\cdot)$ é a função cumulativa de distribuição Normal de uma v.a. e α e β são os parâmetros desconhecidos a serem estimados. Este modelo respeita a propriedade de que π_x é um probabilidade e por isso situa-se no intervalo [0, 1] quaisquer que sejam os valores de *x* e os valores dos parâmetros. A função de ligação designada de função *probit* é dada por: $g(\mu_i) = \Phi^{-1}(\mu_i)$

O modelo GLM obtido pela associação do modelo binomial para as respostas, é função de ligação *probit*.

3.2.5.4. O modelo *logit* versus o modelo *probit*

É agora apresentada uma abordagem (Goldberger 1964, citado por Maddala 1983) de acordo com a qual se assume que há uma variável resposta latente fornecida y_i^* definida pela relação de regressão: $y_i^* = \beta' x_i + u_i$ (3.18)

Na prática y_i^* é inobservável, sendo observada a variável *dummy* *y* definida por

y = 1 se $y_i^* > 0$

y = 0 se caso contrário

Nesta fórmula $\beta' x_i$ não é $E(y_i \mid x_i)$ como no LPM, mas sim $E(y_i^* \mid x_i)$. Das equações anteriores retira-se que:

$$\text{Prob}(y_i = 1) = \text{Prob}\,(u_i > -\beta' x_i)$$

$$= 1 - F(-\beta' x_i) \qquad (3.19)$$

onde F é a função de distribuição cumulativa para u.

Neste caso os valores observados para y são realizações do processo binomial com probabilidades dadas pela equação anterior e variando de julgamento para julgamento (dependendo de x_i). Por isso, a função verosimilhança é

$$L = \prod_{y_i=0} F(-\beta' x_i) \prod_{y_i=1} \left[1 - F(-\beta' x_i)\right] \qquad (3.20)$$

A forma funcional de F irá depender das assunções feitas sobre u_i em (4.21). Se a distribuição cumulativa de u_i é logística , temos o modelo *logit*. Neste caso:

$$F(-\beta' x_i) = \frac{\exp(-\beta' x_i)}{1 + \exp(-\beta' x_i)} = \frac{1}{1 + \exp(\beta' x_i)}$$

Por isso, $$1 - \text{F}\,(-\beta' x_i) = \frac{\exp(\beta' x_i)}{1 + \exp(\beta' x_i)} \qquad (3.21)$$

Neste caso, diz-se que existe uma expressão fechada para F. Nem todas as distribuições permitem tais distribuições fechadas. Por exemplo, no modelo *probit* assumimos que u_i tem IN $(0, \sigma^2)$. Nesse caso,

$$F(-\beta' x_i) = \int_{-\infty}^{-\beta' x_i / \sigma} \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{t^2}{2}\right) dt \qquad (3.22)$$

Pode ser visto na equação anterior e na função verosimilhança (3.20) que podemos estimar apenas β / σ e não apenas β e σ separadamente. Por isso deveremos assumir $\sigma = 1$, por exemplo.

Por causa da distribuição cumulativa normal e da distribuição logística serem muito próximas uma da outra, excepto nas caudas, não se obtém resultados muito diferentes se utilizarmos a expressão (3.21) ou a expressão (3.22), ou seja, nos modelos *logit* ou o *probit*.

Contudo as estimativas de β não são directamente comparáveis pelos dois métodos. Dado que a distribuição logística tem uma variação de $\pi^2 / 3$, as estimativa

obtidas para β pelo modelo *logit* terão de ser multiplicadas por $3^{1/2}/\pi$ para serem comparáveis às estimativas obtidas pelos modelo probit (no qual normalizamos σ para ser igual a 1).

Sugere-se (Amemiya 1981, citado por Maddala, 1983) que as estimativas do *logit* sejam multiplicadas por 1/1.16 =0.625, em vez de por $3^{1/2}/\pi$, e defende que essa transformação produz uma aproximação entre a regressão logística e a função de distribuição Normal. Também se sugere que os coeficientes do LPM $\hat{\beta}_{LP}$ e os coeficientes do modelo logit $\hat{\beta}_{L}$ se relacionam da seguinte forma:

$\hat{\beta}_{LP} \approx 0{,}25\ \hat{\beta}_{L}$ excepto para o termo constante

$\hat{\beta}_{LP} \approx 0{,}25\ \hat{\beta}_{L} + 0{,}5$ para o termo constante

Assim, se quisermos fazer $\hat{\beta}_{LP}$ comparável aos coeficientes *probit*, basta multiplicá-lo por 2,5 e subtrair 1,25 do termo constante.

Há outras formas de comparação que podem ser feitas nos referidos modelos e bastantes estudos foram realizados para comparar as *performances* dos referidos modelos (e.g. Cragg 1971 e McFadden, 1974), As conclusões são praticamente comuns, os processos interactivos de cálculo são muito semelhantes, os estimadores ou são de máxima verosimilhança ou tendem para estimadores de máxima verosimilhança e os valores das estimativas dos coeficientes são muito próximos.

Face ao acima exposto, na presente dissertação optou-se por proceder à estimação logística apenas pela função de ligação *logit*, sendo os resultados muito semelhantes ao modelo *probit*.

3.2.5.5. Análise do modelo de regressão de Logística

O modelo de regressão logística é o mais adequado para explicar o efeitos das variáveis explicativas numa variável resposta que apenas assume dois resultados possíveis a existência de uma característica ou não (Agresti, 1996; Peng & So, 2002).

De acordo com Hosmer & Lemeshow (1989), a regressão logística tornou-se a principal regressão para variáveis medidas sob a forma dicotómica. A principal diferença entre a regressão logística e o modelo linear reside na distribuição da variável

resposta, a qual segue uma distribuição binomial no caso da regressão logística e uma distribuição normal no modelo linear.

Relativamente ao LPM (Peng et al., 2001) a regressão logística apresenta a vantagem de garantir que a correcta adequação à situação em estudo, mesmo quando não é possível estabelecer uma relação linear entre as variável resposta e os parâmetros, e quando os erros não têm tem distribuição normal, com variância constante.

Apesar da sua flexibilidade e simplicidade, existe o pressuposto de baixa correlação entre as variáveis independentes, pois este modelo é sensível à colineriedade entre as variáveis (Hair et al., 1998).

Também nos modelos de regressão logística, há conjunto de literatura que aponta as suas limitações (e.g. Berkson, 1953; Boskin, 1974; Cragg, 1971; Cragg & Uhler, 1970; McFadden, 1974; Schmidt & Strauss, 1975 e Theil, 1969), uma vez que a sua eficácia é por vezes colocada em questão pelos seguintes aspectos:

- A interpretação das estimativas dos coeficientes dos modelos não é imediata;
- Os métodos de classificação e previsão destes modelos são frequentemente questionados face a outros modelos, pois apresentam taxas de erro aparentemente assinaláveis;
- Colocam-se problemas de identificação em virtude da variável dependente ser qualitativa ou limitada.

3.2.6. O Modelo de Regressão de Poisson

O número de ocorrências (contagem) de um determinado acontecimento durante um determinado número de tempo depende da natureza do processo inerente à ocorrência de acontecimentos individuais. O modelo regressão de Poisson constitui o modelo base de análise de regressão de dados de contagem, o qual assenta no processo de distribuição de Poisson (Winkelmann, 2000).

Ainda de acordo com Winkelmann (2000) *o processo de contagem,* $\{N(t), t \in T\}$, *é um processo de Poisson com uma taxa* $\lambda > 0$ *se N(t),* $T \geq 0$, *satisfaz: P(N(0)=0)=1; N(t)-N(s) tem distribuição de Poisson com valor médio* $\lambda(t-s)$, $0 \leq s < t$, *e* $N(t_1)$, $N(t_2)$ *-* $N(t_1)$,…, $N(t_n)$ *-* $N(t_{n-1})$, *são independentes para* $0 \leq t_1 < \ldots < t_n$.

Assim, está-se perante uma distribuição de Poisson *se os acontecimentos ocorrem de acordo com um processo de Poisson com taxa* λ, *independentemente do tempo, então a distribuição do número de acontecimentos* Y_i *que ocorrem no intervalo de tempo* ($t_i - t_{i-1}$), *isto é, tem uma função densidade de probabilidade:*

$$P(Y_i = y) = \frac{\exp[-\lambda(t_i - t_{i-1})][\lambda(t_i - t_{i-1})]^y}{y!}, \; y=0,1,2,\ldots \qquad (3.23)$$

Pela revisão da literatura, verifica-se que grande parte dos estudos que visam a modelização de um conjunto de observações de uma dada variável, iniciam-se muitas das vezes, pela estimação de um modelo de Poisson (e.g., Cameron & Trivedi, 1986; Davutyan, 1989; Dionne et al.; 1997; Hausman, Hall & Griliches, 1984; Johnston & Dinardo, 1996; Kennan, 1985; Long, 1997; McCullagh & Nelder, 1989; Winkelmann, 2000; Winkelmann & Zimmermann, 1994), o qual apresenta como principais vantagens:

- Uma estrutura simples e facilmente estimada;
- O facto dos dados de contagem assumirem apenas valores inteiros não negativos, adequam-se perfeitamente à distribuição de Poisson;
- A boa estimação do modelo permite inferir acerca da probabilidade futura de um dado acontecimento.

3.2.6.1. A especificação

No modelo de regressão de Poisson (Winkelmann,2000) *considerando uma observação i de uma variável aleatória resposta de contagem,* Y_i, *e de um vector de k regressores,* $\mathbf{x'}_i = (x_{i1}, x_{i2},\ldots, x_{ik})$, *e um vector de k parâmetros,* $\boldsymbol{\beta'} = (\beta_1, \beta_2,\ldots, \beta_k)$, *o valor médio e a variância são iguais:* $\lambda = E(y \mid x) = Var(y \mid x) = \exp(x'\beta)$ *(3..27)*

Importa referir que este modelo de regressão assenta em três suposições básicas (Winkelmann, 2000):

Suposição 1 – A distribuição condicional da variável dependente.

$f(y \mid \lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$, $y = 0,1,2,\ldots$ *onde* $f(y \mid \lambda)$ *é a função de probabilidade condicional de y dado* λ, *garantindo-se que* $\lambda > 0$.

Suposição 2 – A especificação do valor médio.

$\lambda = \exp(x'\beta)$, *onde* β *é o vector dos parâmetros (kx1)e x é o vector dos regressores (kx1), ou seja admite-se que o parâmetro* $\lambda(x_i)$ *é a exponencial de uma combinação linear dos regressores* x_i.

Suposição 3 – A independência das observações

Cada par de observações (y_i, x_i), $i = 1,\ldots,n$ *são independentemente distribuídos.*

3.2.6.2. A estimação

As condições de primeira ordem da função verosimilhança constituem o sistema de *k* equações (não linear nos parâmetros): $\sum_{i=1}^{n}[y_i - \exp(x'_i\beta)]\ x_i = 0$

Com as seguintes condições:

- $\sum_{i=1}^{n} e_i = 0$ em que $e_i = y_i - \exp(x'_i\hat{\beta})$; ou seja a média dos resíduos é nula;
- $\sum_{i=1}^{n} e_i x_{ij} = 0$, j= 2,…, k; isto é, a existência de ortogonalidade entre o vectores dos resíduos e = ($e_1,...,e_n$) e os k-1 ($x_{1j},...,x_{nj}$)' vectores das n observações dos regressores não constantes, x_j , j=2,…, k.

3.2.6.3. Análise do modelo de regressão de Poisson

Apesar do modelo de regressão de Poisson ser o candidato natural para a especificação de um modelo de análise de regressão de dados de contagem, este raramente explica as observações (Chin & Quddus, 2003).

Como já foi referido o modelo de Poisson tem como principal propriedade o facto de a média e a variância condicionais serem iguais, o que se designa de "equidispersão". Esta hipótese é, com frequência, contestada empiricamente, sobretudo porque com frequência se verifica sobredispersão (Gelman & Hill, 2007).

Entre os factores que conduzem à violação desta hipótese encontra-se a heterogeneidade não observada da população (Gourieroux & Visser, 1997), o contágio entre acontecimentos (King, 1989; Zorn, 1998), os erros de medida das covariáveis (Guo & Li, 2000) e a frequência elevada de observações com valor zero (Lambert, 1992). A imposição desta restrição pode resultar num erro padrão estimado de baixo valor e estatística *t* sobrestimada (Cameron & Trivedi, 1986; Ganio & Schafer, 1992, Haynes et al., 2003). O erro padrão e a inferência sobre os parâmetros da regressão, designadamente os intervalos de confiança e os testes de hipóteses, podem ser enganadores.

Para verificar a hipótese da igual dispersão, existe um conjunto de testes estatísticos que podem ser estudados em Brännäs (1992a,b), Breslow (1990), Cameron & Trivedi (1986) , Dean (1992), Dean & Lawless (1989), 1990), Ganio & Schafer (1992), Gourieroux & Visser (1997), Gurmu (1991), Gurmu & Trivedi (1992), Lawless (1987), Lee (1986), Mullahy (1986, 1997).

Assim, devido às suas deficiências, existe um conjunto de modelos de regressão alternativos ao modelo de Poisson que têm em conta a possibilidade de haver sobredispersão. Alguns desses modelos podem vistos em diversos trabalhos científicos (e.g., Cameron & Trivedi, 1986; Hausman et al., 1984; Hinde & Demétrio, 1998 e Winkelmann, 2000), nos quais o modelo alternativo que tem em conta a possibilidade de sobredispersão ou de subdispersão é modelo Binomial Negativo.

3.2.7. O Modelo de Regressão Binomial Negativo

A diferença entre o modelo de Poisson e o modelo Binomial negativo reside na significância estatística do parâmetro α (Carrivick et al., 2003; Chin & Quddus., 2003), caso α não seja significativamente diferente de zero, o modelo binomial negativo reduz-se ao modelo de regressão de Poisson.

Este modelo constitui uma especificação que permite uma modelação mais flexível da variância no que respeita a situações onde ocorre “sobredispersão”, sendo de destacar que mesmo quando o valor médio e a variância são correctamente especificados de acordo com a distribuição binomial negativa, se esta não for a distribuição correcta, o estimador de máxima verosimilhança é inconstante (Cameron &

Trivedi, 1990). Apesar disso, o modelo de regressão binomial negativo, ao dar conta de uma possível heterogeneidade não observada da população e ao permitir a existência de "sobredispersão," é um modelo, em geral, mais adequado do que o modelo de Poisson (Chin & Quddus, 2003), obtendo-se bons ajustamentos (Hausman et al., 1984; Haynes et al., 2003; Lee et al., 2002). Contudo, é frequente ocorrer variância superior à tolerada pelo modelo binomial negativo (Dean, 1992; Gardner et al., 1995).

3.2.7.1. A especificação

A especificação mais corrente é a NB2 (Cameron & Trivedi, 1998), admite que a variância é uma função quadrática do valor médio, isto é:

$$\lambda_i = \mathrm{E}(y_i \mid \mathbf{x}_i) = \exp(\mathbf{x'}_i, \beta) \quad \text{e} \quad (3.28)$$

$$Var(y_i \mid \mathbf{x}_i) = \lambda_i + \alpha\lambda_i^2 \quad (3.29)$$

Em que $\alpha \geq 0$ é um parâmetro de dispersão a ser estimado, relativamente ao qual a inferência permite testar a sobredispersão e a adequação do modelo de Poisson, sob a hipótese do tipo $\alpha = 0$.

Verifica-se, assim, que $Var(y_i \mid \mathbf{x}_i) \geq E(y_i \mid \mathbf{x}_i)$. Deste modo, o modelo binomial negativo contempla o caso de sobredispersão dos dados. Todavia, é possível testar a hipóteses de subdispersão, isto é, $\alpha < 0$ (Lawless, 1987).

Para especificação do NB2, a função massa de probabilidade é dada por:

$$\Pr(Y_i = y_i \mid \mathbf{x}_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\lambda_i + \alpha^{-1}}\right)^{\alpha^{-1}} \left(\frac{\lambda_i}{\lambda_i + \alpha^{-1}}\right)^{y_i} \text{com } y_i = 1,2,\ldots,n$$

em que $\lambda_i = \lambda(\mathbf{x}_i)$ e $\Gamma(.)$ é a função de gama. Assume-se, ainda, que as observações são independentes.

Outra especificação, designada correntemente na literatura pelo acrónimo NB1 (Cameron & Trivedi, 1998), estabelece a variância como uma função linear do valor médio:

$$\lambda_i = \mathrm{E}(y_i \mid \mathbf{x}_i) = \exp(\mathbf{x'}_i, \beta) \quad \text{e} \quad Var(y_i \mid \mathbf{x}_i) = \lambda_i + \alpha\lambda_i, \alpha \geq 0$$

Genericamente, o modelo de regressão binomial negativo de ordem *p*, NB_p, assume a variância como uma função do valor médio, $\lambda_i = \exp(\mathbf{x'}_i \beta)$, do tipo:

$$Var(y_i \mid \mathbf{x}_i) = \lambda_i + \alpha\lambda_i^p, \alpha \geq 0, p = 1,2,...$$

O modelo de regressão binomial negativo é uma generalização do modelo de Poisson, verificando-se $\alpha = 0$, para este último. Neste sentido, a inadequação do modelo de regressão de Poisson relativamente ao modelo de regressão binomial negativo é determinada pela significância estatística do parâmetro α; se α não é significativamente diferente de zero, o modelo reduz-se ao modelo de regressão de Poisson (Carrivick et al., 2003 e Chin & Quddus, 2003).

Porém, conforme referem Cameron & Trivedi (1986), cada uma destas especificações equivale a uma determinada forma funcional de heteroscedasticidade, e a sua escolha é uma entre diferentes modelos. Salienta-se, por isso, que o modelo Binomial Negativo e o modelo de Poisson são modelos de regressão distintos.

Barron (1992) sublinha o facto de não haver uma razão sustentada para preferir uma determinada especificação do modelo de regressão binomial negativo, pelo que *p*, ou a ordem deste modelo, deve ser tomado como mais um parâmetro a ser estimado a partir dos dados (King, 1989; Winkelmann & Zimmermann, 1994 e Zorn, 1998).

3.2.7.2. A estimação

As condições de primeira ordem da função verosimilhança para a estimação de β e α, são:

$$\sum_{i=1}^{n} \frac{y_i - \exp(\mathrm{x'}_i \beta)}{1 + \alpha \exp(\mathrm{x'}_i \beta)} \mathrm{x}_i = 0$$

$$\sum_{i=1}^{n} \left\{ \frac{1}{\alpha^2} \left(\ln(1 + \alpha \exp(\mathrm{x'}_i \beta)) - \sum_{j=0}^{y_i - 1} \frac{1}{j + \alpha^{-1}} \right) + \frac{y_i - \exp(\mathrm{x'}_i \beta)}{\alpha(1 + \alpha \exp(\mathrm{x'}_i \beta))} \right\} = 0$$

Para a resolução deste sistema de equações não lineares nos parâmetros β e α pode utilizar-se o algoritmo de Newton-Raphson (Lawless, 1987; Maul et al., 1991). De salientar que, sob determinadas condições, a matriz Hessiana é definida negativa (Mukerjee & Sutradhar., 2002), o que assegura a existência de solução única.

3.3. Análise dos modelos de regressão

Pela revisão feita à literatura sobre trabalhos publicados no âmbito dos modelos de regressão, destacam-se estudos sobre acidentes de trabalho nos vários sectores de actividade económica (Bailer et al., 1997; Wang et al., 2003); sobre a mobilidade do factor força de trabalho (Winkelmann & Zimmermann, 1995; Winkelmann & Zimmermann, 1998); sobre a evolução de algumas patologias no âmbito da saúde pública (Böhning et al., 1997; Cameron & Windmeijer, 1996; Greene, 1993; Henderson & Taylor, 2003); sobre factores que determinam a procura de meios de telecomunicação (Gupta et al., 2001; Heitfield & Levy, 2001); sobre a relação entre as condições de saúde dos condutores e acidentes automóvel (Dionne et al., 1995; Lee et al., 2002); sobre relação entre acidentes automóvel e apólices de seguro (Dionne & Vanasse, 1992); sobre interacção rodoviária peão-condutor (Shankar et al., 2003; Schneider et al., 2004), entre muitos outros.

Da análise a todos estes trabalhos conclui-se ser comum a todos, a tentativa de modelação de um conjunto de observações de uma variável de contagem a partir da estimação de modelos de contagem.

De acordo com Winkelmann (2000), está-se perante modelos teóricos de concepção de risco, pelo que partindo de um mesmo modelo de risco tem-se à disposição um conjunto de modelos de análise de regressão para avaliação do risco, sendo o propósito da sua utilização, a eleição e utilização final do modelo com melhor desempenho, ou seja, o modelo que melhor aderir aos dados.

A escolha *à priori* de um destes modelos poderia ter como consequência um ajustamento não optimizado aos dados. A modelação adequada de dados de contagem deve sempre ser efectuada à luz de um conjunto de modelos de análise de regressão de dados de contagem alternativos, que se ajustem de forma diferente e em conformidade com as particularidades dos dados em questão.

4. ANÁLISE E MODELAÇÃO DE DADOS

Em qualquer trabalho que se utilize dados quantitativos é de extrema importância, a familiarização com o conjunto de dados que será utilizado posteriormente na análise empírica (Wooldridge, 2002), pelo que numa primeira fase é objectivo proceder a uma análise da base de dados, com vista a descobrir ou a reduzir a dimensão do conjunto de dados e para identificar variáveis significativas que irão compor os modelos, posteriormente será feita a definição das variáveis a utilizar nos modelos econométricos.

4.1. As empresas da base de dados

As três Seguradoras que constituem o Grupo Caixa Seguros são: Fidelidade-Mundial, Império-Bonança e Via Directa.

A história da C. S. Fidelidade-Mundial remonta a 1835, vários momentos marcaram a actividade: fusões, nacionalização e posterior privatização, e expansão para novos mercados (Companhia de Seguros Fidelidade-Mundial, 2011). O percurso da C. S. Império-Bonança é muito semelhante surgiu em 1808 e passou por iguais processos (Companhia de Seguros Império-Bonança, 2011). Salienta-se, ainda, que de acordo com os Relatórios e Contas de 2010 desta duas seguradoras, as mesmas apresentam uma organização interna, estratégica e *modus operandis* (quer na angariação de clientes, quer oferta de produtos e formação de preços) muito similares.

A C. S. Via Directa apresenta um percurso e características diferentes. Foi fundada em 1998 com vocação para a venda através dos canais directos (telefone e internet), promovendo os seus produtos através da marca “Ok!teleseguros”. Os seus clientes caracterizam-se pelo seu perfil urbano, com carro e seguro em nome próprio, com idades entre os 25 e 55 anos de idade, maioritariamente pertencentes à classe social “média”e com forte sensibilidade para a componente preço na aquisição do produto (Companhia de Seguros Via Directa, 2011).

4.2. Análise da composição da sinistralidade

Da análise da base de dados, e através da figura 4.1. verifica-se que do número total de apólices que a compõem 80% não registam qualquer sinistro, pelo que apenas 20% dessas apólices têm acidentes.

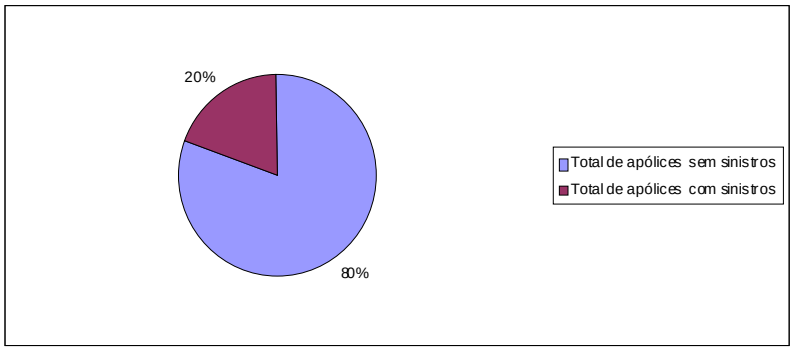

Figura 4.1– Sinistralidade das apólices que compõem a base de dados
Fonte: Elaboração própria

Das apólices que registam acidentes, verifica-se pela figura 4.2 que a grande maioria dos acidentes regista apenas danos materiais (86%), uma parte muito mais reduzida (12%) dos acidentes regista simultaneamente danos materiais e corporais e apenas 2% das apólices com acidentes têm apenas danos corporais.

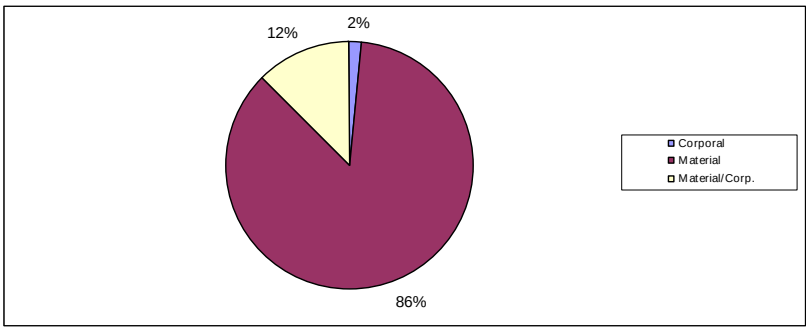

Figura 4.2 – Consequências decorrentes dos acidentes de viação
Fonte: Elaboração própria

Das apólices com observações de acidentes, que pode ser efectuada na figura 4.3, pode ver-se a grande maioria regista apenas um acidente, muito menos de metade do número de apólice regista dois acidentes. O número de apólice que regista mais do que dois acidentes não é relevante.

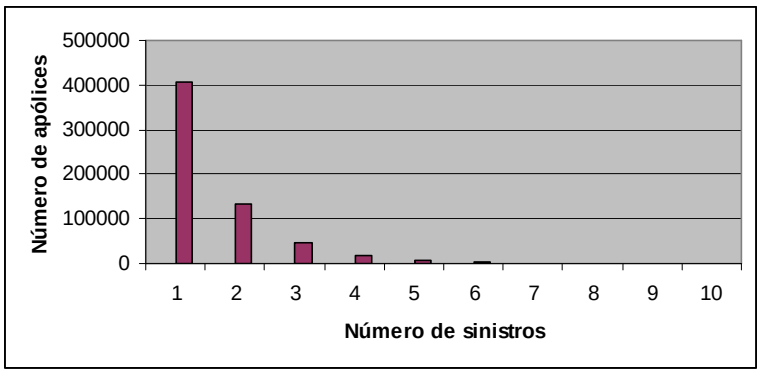

Figura 4.3 – Número de sinistros verificados por apólice
Fonte: Elaboração própria

As apólices sem acidentes representam 80% das observações, o que representa 3.103.085 observações. Considerando que se trata de um valor muito superior ao das observações de apólice com acidentes, optou-se por não incluir na figura 4.3 o número de apólice com zero sinistros para que graficamente fosse possível comparar o número de sinistros superiores a zero por apólice.

Assim, fazendo uma breve análise à sinistralidade no período de 2000 a 2010 com base na figura 4.4 verificamos que o número de observações de sinistros registou um aumento significativo em 2001, sendo por esse motivo considerado um ano atípico em termos de observações. A partir desse ano o número de aberturas de sinistros começou a diminuir, sendo a partir de 2004 se mantém praticamente constante.

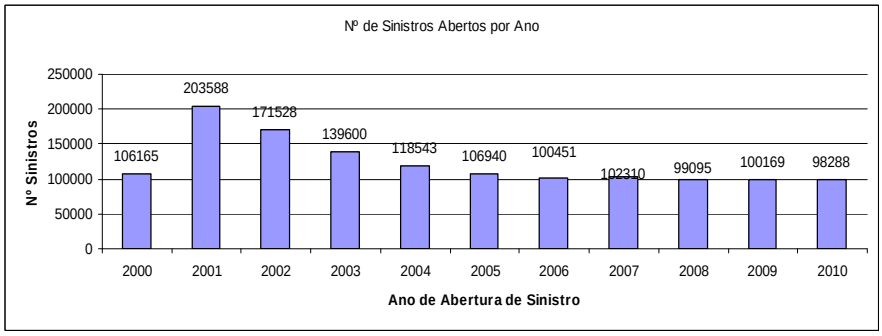

Figura 4.4 - Número de sinistros abertos por ano na Seguradora entre 2000 e 2010
Fonte: Elaboração própria

Considerando que estar-se em presença de dados numéricos e pretendendo-se descrever a evolução ao longo dos últimos 11 anos dos valores observados, então os mesmos podem ser apresentados sob a forma de uma série temporal. Assim, analisando

o cronograma que consta da figura 4.5 conclui-se que estamos perante uma série estacionária em média e em variância, ao longo do período 2000-2010.

Há uma quebra de série no ano de 2001, conforme já foi referido anteriormente. Não é perceptível a presença de tendência, mas há forte evidência de sazonalidade, onde os valores mais elevados de sinistralidade correspondem de um modo geral aos meses de condições meteorológicas mais adversas (o Inverno) de onde se salienta os meses de Janeiro dos vários anos, sempre com os valores mais elevados. As observações mais baixas verificam-se nos períodos de férias escolares/festividades: Carnaval (mês de Fevereiro), Páscoa (mês de Abril), e o mês de Agosto.

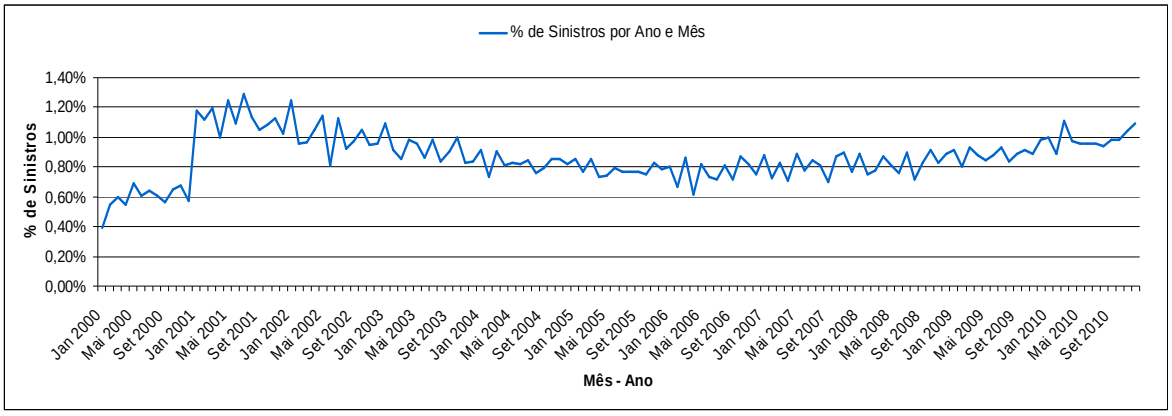

Figura 4.5 – Cronograma sobre a evolução mensal dos sinistros no período 2000-2010
Fonte: Elaboração própria

4.3. Análise das possíveis variáveis explicativas

Partido da base de dados referida no ponto 4.2. e acordo com a revisão da literatura nomeadamente Denuit et al. (2007), Harvey (2004), Hoffman (2002), Oliveira (2007), Peden e tal (2004) e Petridou & Mouskati (2000), de onde se concluiu que os factores humano, veículo, e ambiental estão na origem dos acidentes viação, procurou-se dentro da base de dados excluir informação redundante, incongruente e pouco relevante, o que levou nesta fase à selecção dos campos abaixo referidos, e cuja informação será analisadas em detalhe, com vista à escolha das variáveis para construção dos modelos: DATA_SINISTRO, HORA_SINISTRO, CONSEQUÊNCIA, REGIÃO_SINISTRO, RESP_SEGURADO, CAUSA_SIN,

SEGURADO_DATA_NASC, SEGURADO_SEXO, CATEGORIA_VIATURA, ANO_CONSTRUÇÃO e MOTIVO_REEMBOLSO.

4.3.1. DATA_SINISTRO

Analisando os sinistros na perspectiva da data em que ocorreram do acidente procedeu-se ao agrupando das observações de sinistros dos vários anos por mês, conforme se pode visualizar na figura 4.6 e da qual se conclui que os primeiros três anos do período em estudo foram aqueles onde verificou maior número de acidentes. Pode, também, observar-se que por norma a evolução do número de acidentes distribuída pelos vários meses tem um comportamento muito similar ao longo dos 11 anos. O número de acidentes de viação atinge sempre um valor mínimo em meados de Fevereiro e um valor máximo em meados de Março, volta a atingir outro valor máximo em Janeiro e em meados de Julho, sendo que a partir de meados de Novembro regista quase sempre uma tendência crescente.

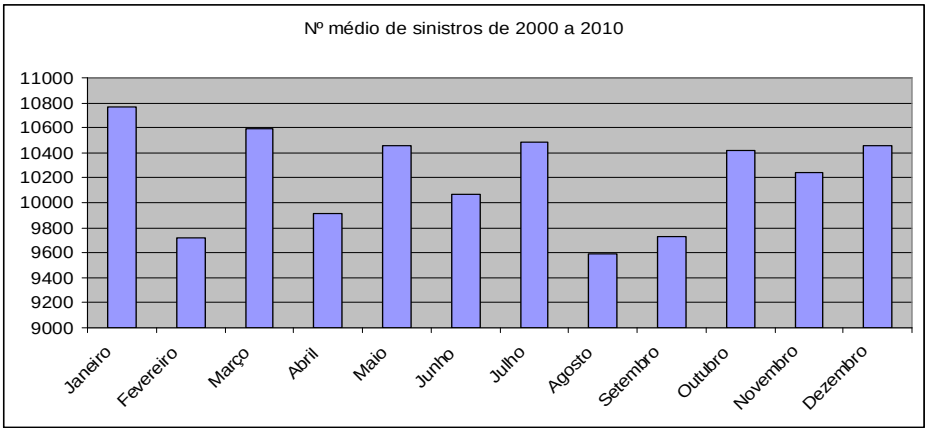

Figura 4.6 – Histograma do número médio de acidentes por mês de 2000 a 2010
Fonte: Elaboração própria

Não se verifica simetria na distribuição, é muito frequente a ocorrência de um grande número de sinistros durante todos os meses do ano. Com destaque para os meses de Janeiro e Março que são os aqueles onde se verificam sinistros com maior frequência por oposição ao meses de Fevereiro, Agosto e Setembro.

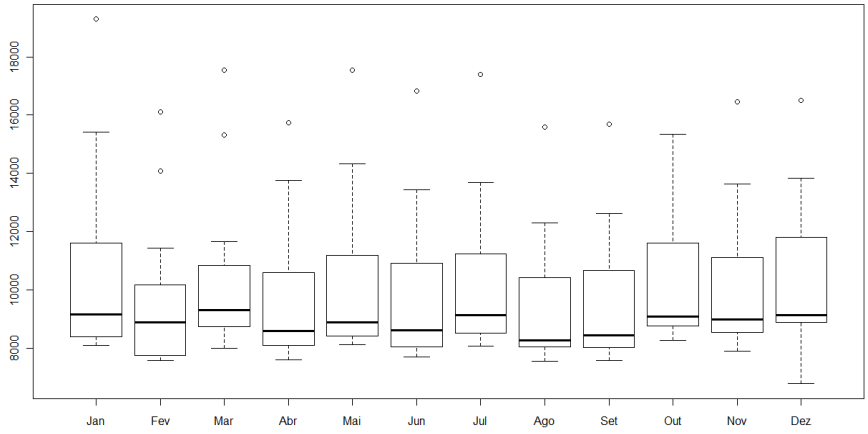

Figura 4.7– Diagrama caixa-de-bigodes do número de sinistros por mês no período 2000-2010

Fonte: Elaboração própria

A figura 4.7 confirma as conclusões anteriores e evidencia uma significativa concentração de observações do número de acidentes nos valores dos quantis inferiores, denotando-se assimetria positiva, a qual é visível pelo maior prolongamento da aba superior de praticamente todos os meses. Evidenciam-se *outliers* em quase todos os meses. Verifica-se, ainda, existir uma significativa dispersão dos gráficos pelo domínio da distribuição, sendo esta mais forte nos meses de Janeiro, Março, Maio, Julho e Outubro, o que revela uma forte variabilidade dos dados.

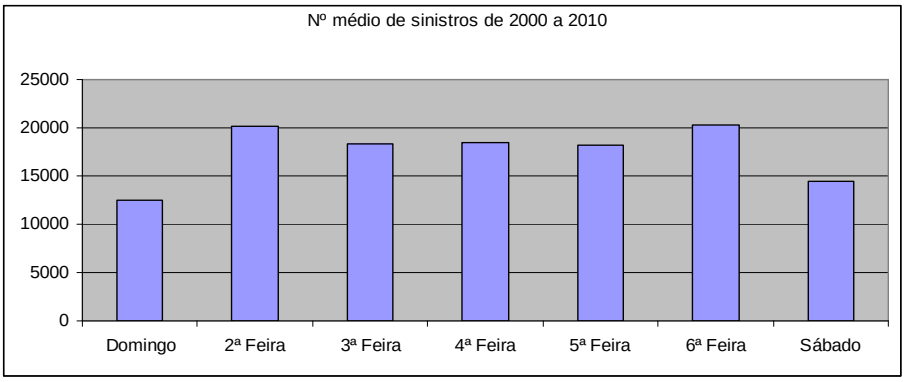

Figura 4.8 – Histograma do número médio de acidentes por dia da semana

Fonte: Elaboração própria

O número de observações dos acidentes de viação, ao longo dos 11 anos, conforme evidenciado na figura 4.8 tem um comportamento praticamente idêntico para

os vários dias úteis da semana, atingindo valores mínimos ao fim-de-semana, em especial ao Domingo.

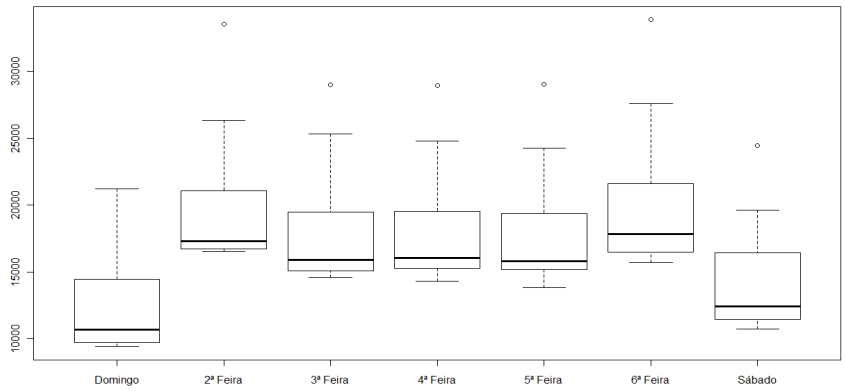

Figura 4.9 – Diagrama caixa-de-bigodes para o número de sinistros por dia da semana no período 2000-2010
Fonte: Elaboração própria

Este gráfico reforça as conclusões anteriores, a distribuição é assimétrica positiva, pois é nos quantis inferiores que se verifica maior concentração das observações. É frequente a presença de *outliers* e verifica-se a dispersão dos dados.

4.3.2. HORA_SINISTRO

Representa a hora do dia em que ocorre o sinistro (horas e minutos).

Na figura 4.10 observam-se 4 grupos distintos de frequência de acidentes, entre as 23h00 e as 06h00 é um período de quase ausência de observações. A partir das 06h00 verifica-se um aumento exponencial no número de observações de acidentes, sendo o valor máximo atingido sempre por volta das 12h00. Durante a tarde o número de observações tende a ser constante até por volta das 18h00. A partir das 18h00 regista-se um decréscimo acentuado do número de acidentes.

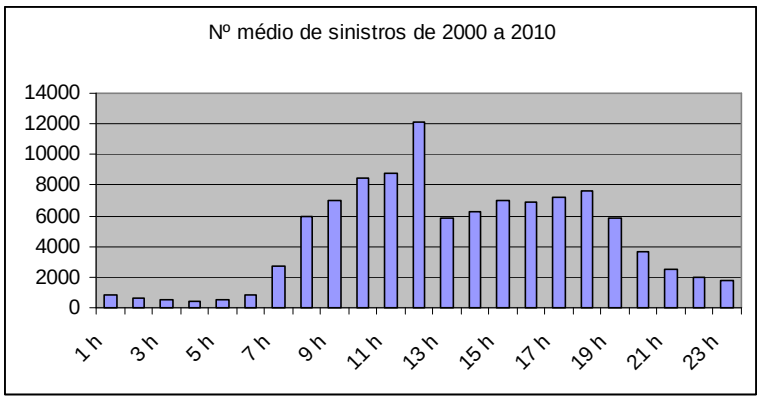

Figura 4.10 – Histograma do número médio de acidentes por hora do dia no período 2000 a 2010

Fonte: Elaboração própria

A figura 4.10 mostra haver uma distribuição quase simétrica em torno das 12h00, hora a que ocorrem o maior número de observações (atingindo-se os 12.000 sinistros).

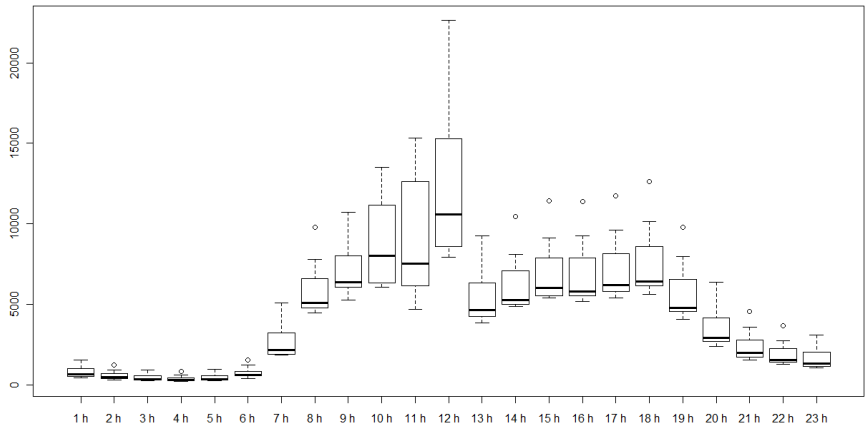

Figura 4.11 – Diagrama Caixa-de-bigodes para o número de sinistros por hora do dia, no período 2000-2010

Fonte: Elaboração própria

A figura 4.11 reforça a existência dos 4 grandes grupos de observações e evidencia uma concentração das observações nos primeiros quantis, o que torna a evidenciar assimetria positiva dos dados.

4.3.3. CONSEQUÊNCIA

É um campo no qual se avalia os resultados produzidos por um acidente.

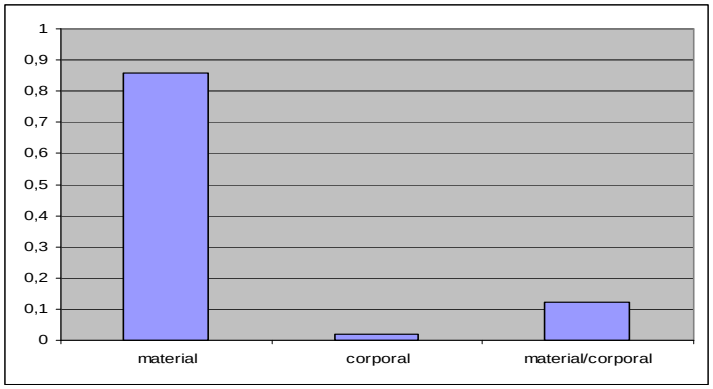

Figura 4.12 – Histograma das consequências por acidente entre 2000 e 2010 (%)
Fonte: Elaboração própria

Ressalta-se a grande frequência de acidentes pouco graves (apenas com consequências materiais) por oposição aos acidentes graves (com consequências para a vida humana) pela visualização da figura 4.12. Mais de 80% dos acidentes têm apenas consequências materiais: outros veículos e objectos (e.g.: postes, sinais de trânsito, muros, etc.)

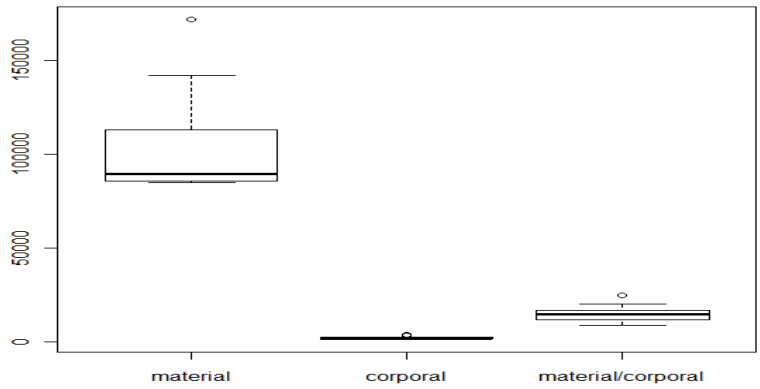

Figura 4.13 – Diagrama caixa-de-Bigodes das consequências por acidente
Fonte: Elaboração própria

Na figura 4.13 evidencia-se a presença de *outliers* nos três grupos de consequências possíveis decorrentes de acidente. Nos acidentes com danos materiais é

evidente uma acentuada concentração de observações nos quantis superiores, denotando a presença de forte assimetria positiva.

O critério para a criação de duas classes desta variável baseou-se na preservação da vida humana enquanto valor supremo, relegando, assim, para um segundo plano qualquer outro tipo de consequências que decorrem de um acidente de viação.

4.3.4. REGIÃO

É uma variável qualitativa nominal, provém da informação que consta nos campos CONCELHO_SINISTRO e SEGURADO_CONCELHO, que é praticamente coincidente, para ser agrupada por zona geográfica da ocorrência dos acidentes.

Assim, partindo da base de dados, onde geograficamente os acidentes são identificados pelo Código Postal do Concelho no qual se verifica o acidente, foi possível elaborar classes de maior dimensão e mais fácil manuseamento desta variável agrupando-a, não por Concelho, mas sim por região geográfica.

Inicialmente, o agrupamento por região baseou-se em três factores:

- Destacar as duas principais áreas metropolitanas de Portugal: Lisboa e Porto;
- Considerar as duas regiões autónomas: Açores e Madeira;
- Agrupar os restantes Concelhos de acordo com a proposta de criação de divisão regional oficializada na Lei da Criação das Regiões Administrativas (Lei nº 19/98), mas considerando os ajustamentos atrás efectuados ao destacar as duas grandes áreas metropolitanas. Assim, as restantes regiões são: Alentejo, Algarve, Beira Interior, Beira Litoral, Minho, Ribatejo e Trás-os-Montes.

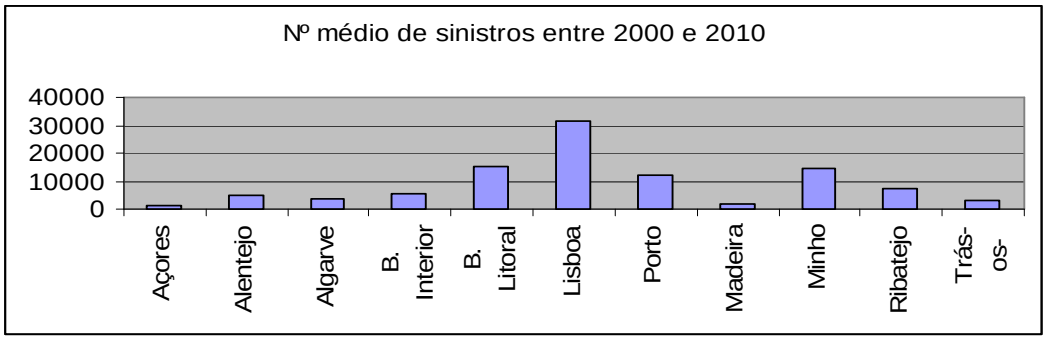

Figura 4.14 – Histograma do número médio de acidentes por região entre 2000 e 2010
Fonte: Elaboração própria

Na figura 4.14 a área metropolitana de Lisboa (a maior concentração urbana) com o maior número de acidentes verificados: 31699, seguida de outras regiões urbanas do litoral (Porto, Minho e Beira Litoral). As regiões do interior do país e as regiões autónomas são as aquelas onde ocorre menor número de observações.

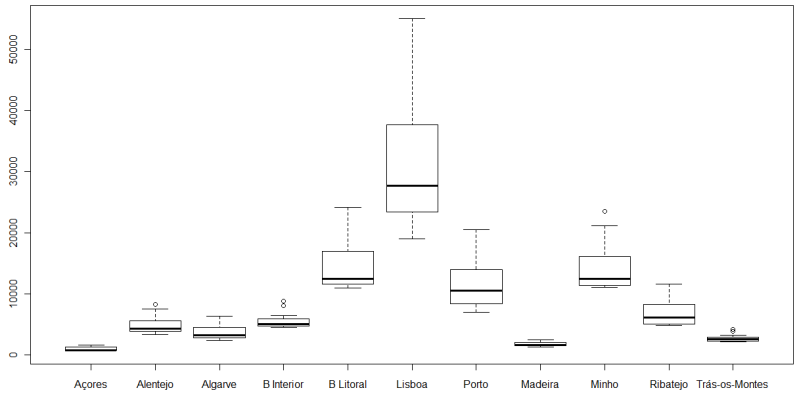

Figura 4.15 – Diagrama caixa-de-bigodes para o número de sinistros por hora do dia, no período 2000-2010

Fonte: Elaboração própria

Pela visualização da figura 4.15 verifica-se que as regiões com menos acidentes, têm dispersão reduzida. Por outro lado as regiões com maior número de acidentes registam níveis de dispersão mais elevados. Salienta-se ser comum a todas as regiões uma concentração no número de acidentes nos quantis inferiores, denotando uma assimetria positiva das observações para todas as regiões.

Optou-se por criar apenas duas categorias para esta classe: a região formada por Lisboa e Porto, por contraposição às restantes regiões do país.

4.3.5. RESP_SEGURADO

É uma variável quantitativa ordinal, é representada pelos valores ordenados por ordem crescente em percentagem de culpa que pode ser atribuída ao segurado num acidente. Varia entre 0% e 100%.

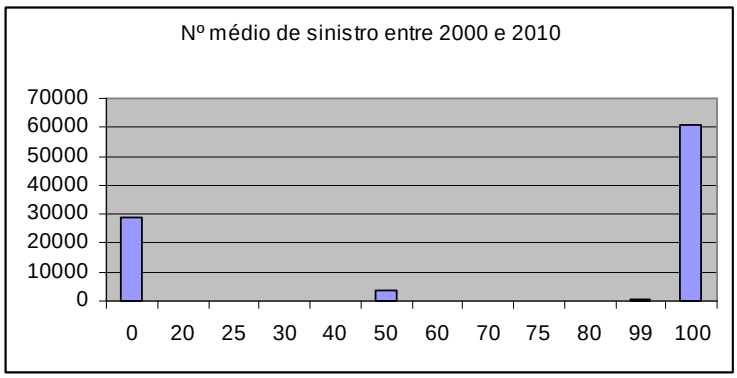

Figura 4.16 – Histograma da percentagem de responsabilidade média atribuída nos acidentes entre 2000 e 2010

Fonte: Elaboração própria

Da figura 4.16 destaca-se o facto da percentagem de responsabilidade do segurado, e consequentemente da Companhia de Seguros, se concentrar em 3 grupos:

- Ausência de responsabilidade do segurado: 0%
- Repartição equitativa de responsabilidades: 50%
- Total responsabilidade do segurado: 100%

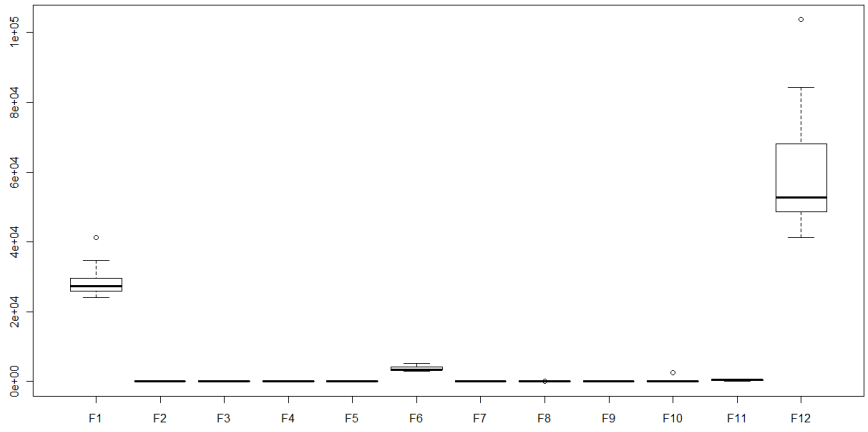

Figura 4.17 – Diagrama caixa-de-bigodes da responsabilidade no acidente no período 2000-2010

Fonte: Elaboração própria

A figura 4.17 reforça as conclusões anteriores: evidencia acentuada assimetria positiva, dado existir um significativo número de observações nos quantis inferiores.

O histórico da sinistralidade seria uma variável interessante, não fosse conter informação endógena ao estudo, o que traria problemas na estimação.

4.3.6. CAUSA_SIN

Coloca em evidência que das 26 causas possíveis para a ocorrência de acidentes, menos de metade são consideradas relevantes para que um acidente de viação ocorra.

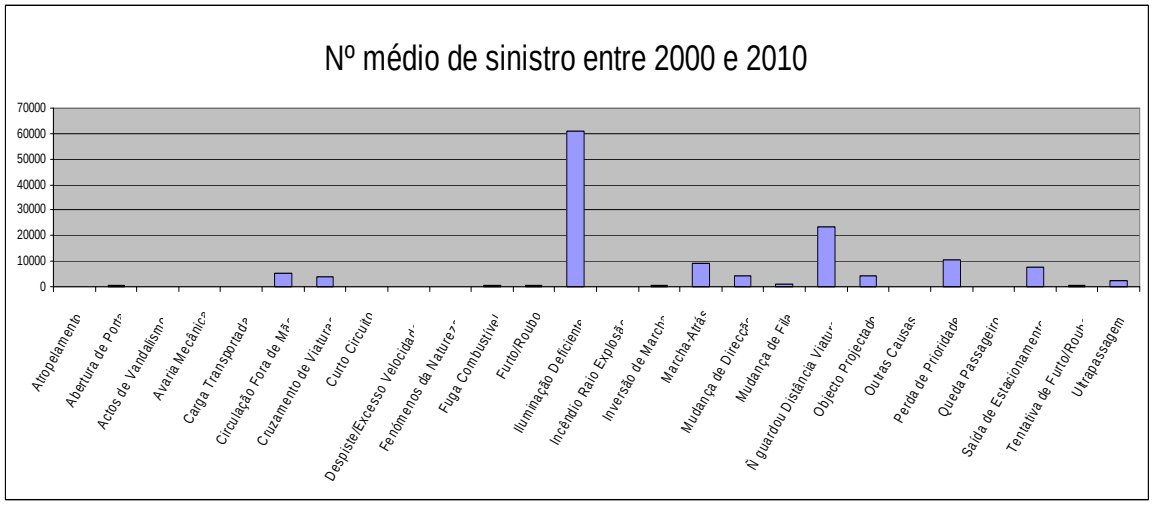

Figura 4.18 – Histograma das causas de sinistro entre 2000 e 2010
Fonte: Elaboração própria

De forma mais perceptível pela leitura da figura 4.18, verifica-se que a "Deficiente iluminação" e o facto de "Não guardar a distância da viatura de frente" registam o maior número de observações para a causa de acidente de viação.

Acidentes com a intervenção de uma única viatura, como é o caso dos "atropelamentos", "quedas de passageiros dentro de transportes de passageiros" ou "avaria mecânica" quase não têm expressão.

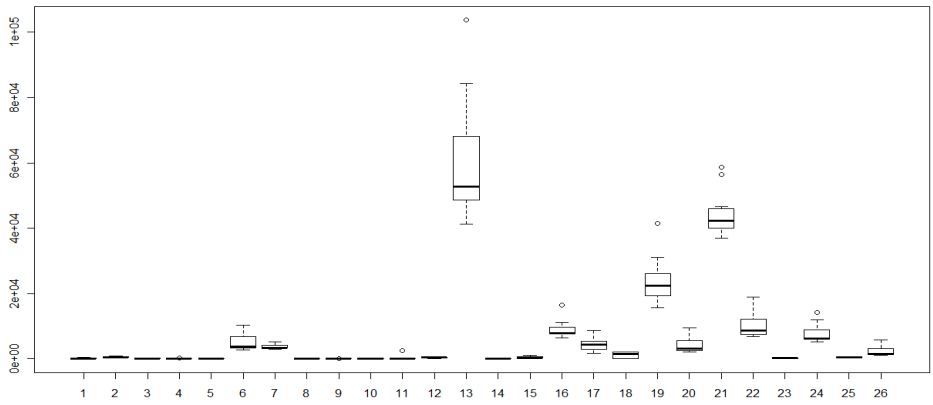

Figura 4.19 – Diagrama caixa-de-bigodes das causas de acidente, no período 2000-2010
Fonte: Elaboração própria

Ilustra-se na figura 4.19 as conclusões anteriores, acresce referir que são as causas com maior número de observações, as que apresentam maior dispersão. A presença de *outliers* é frequente. A assimetria não é semelhante para todas as variantes possíveis desta variável, se nas causas "Iluminação deficiente" e "Perda de prioridade" a assimetria à direita é óbvia, para outras causas tal não sucede.

As categorias desta classe não são mutuamente exclusivas, pelo que não será de incluir na estimação dos modelos.

4.3.7. SEGURADO_ DATA_ NASC

É uma variável quantitativa que abrange indivíduos dos 18 aos 90 anos. Apesar de tratar de uma variável numérica, optou-se nesta fase por agrupa-la por classes para a sua mais fácil compreensão.

Importa referir que este indivíduo (segurado) é aquele que celebrou o contrato com a seguradora, mas nem sempre é ele o condutor habitual da viatura.

A opção pela escolha deste campo em detrimento da variável CONDUTOR_VEIC_SEGURO_DATA_NASC, baseia-se no facto deste último ter menos quantidade de informação disponível.

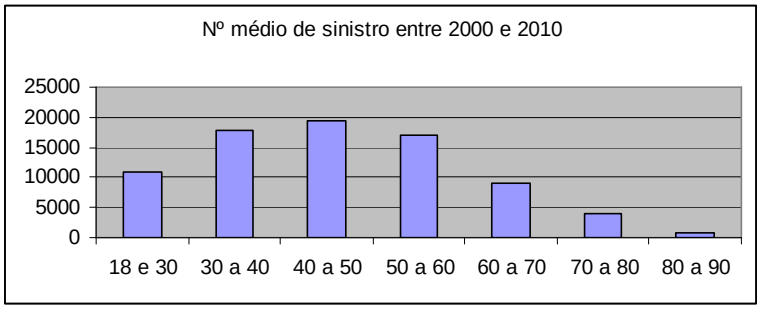

Figura 4.20 – Histograma do número médio de acidentes por idade dos segurados entre 2000 e 2010

Fonte: Elaboração própria

Denota-se, na figura 4.20, que o maior número de observações aparece na faixa etária entre os 30 e 40 anos e na faixa etária seguinte entre os 40 e 50 anos de idade; e coloca em evidência o facto de se registar uma tendência decrescente nas observações à

medida que as idades dos indivíduos aumentam até aos 40 anos de idade, a partir daí ocorre um lento decréscimo no número de observações.

É evidente uma assimetria à esquerda, em virtude da frequência com que os acidentes ocorrem ser crescente até à faixa etária dos]40, 50] e a partir dessa faixa etária passar a decrescer quase na mesma dimensão.

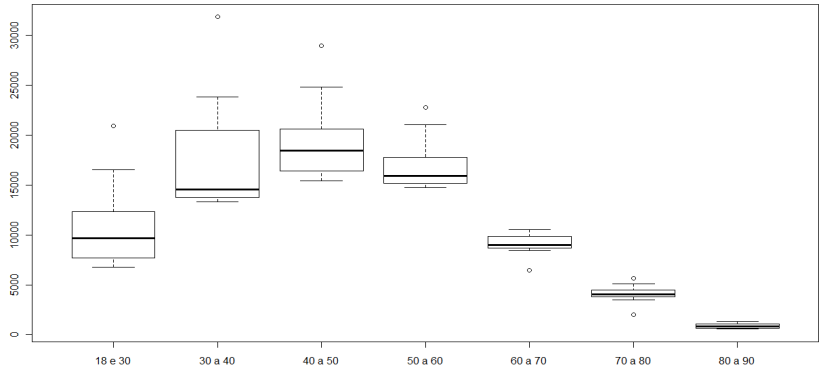

Figura 4.21 – Diagrama caixa-de-bigodes das idades dos segurados, no período 2000-2010

Fonte: Elaboração própria

A figura 4.21 demonstra ainda que a assimetria tende a ser positiva, mas de forma mais vincada em determinadas faixas etárias, como por exemplo na faixa etária]30, 40].

Destaca-se a presença de *outliers* em todos os grupos etários.

Para compreender esta variável foi importante criar classes para facilitar a análise, no entanto esta variável será incluídas nos modelos de regressão como numérica.

4.3.8. SEGURADO_SEXO

É uma variável qualitativa nominal, pode assumir três possibilidades: “Masculino”, “Feminino” ou “Empresa”.

O género “Masculino” destaca-se por ter um maior registo de observações largamente superior ao conjunto das restantes categorias, o mesmo se passa relativamente ao seu nível de dispersão. O género “Feminino” e a categoria “Empresa” têm valores mais próximos entre si, conforme figura 4.22.

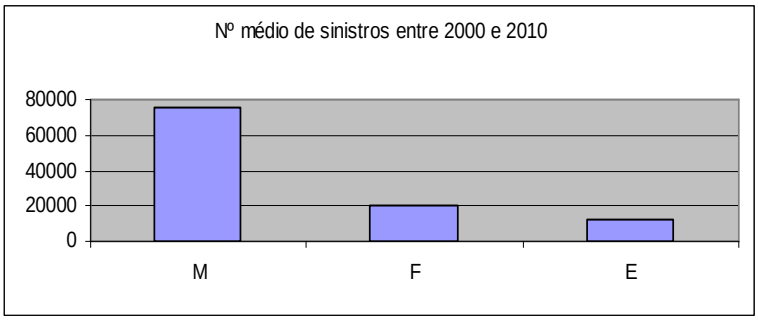

Figura 4.22 – Histograma do número médio de acidentes por género do segurado entre 2000 e 2010

Fonte: Elaboração própria

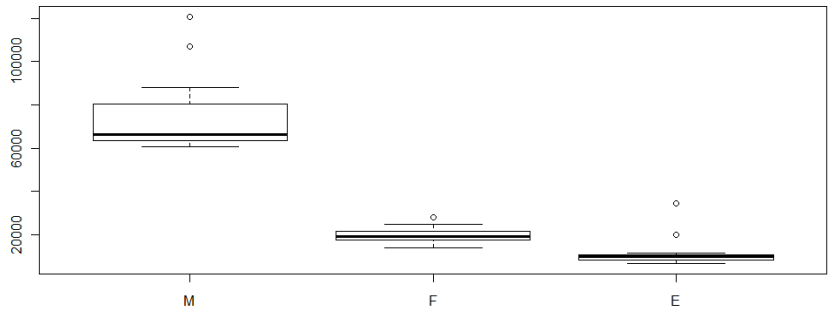

Figura 4.23 – Diagrama caixa-de-bigodes do género dos segurados, no período 2000-2010

Fonte: Elaboração própria

Confirma-se, na visualização da figura 4.23, que o maior número de observações ocorre no género masculino, onde se manifesta uma assimetria positiva acentuada face aos restantes géneros. De salientar que a distribuição não é simétrica para nenhum dos géneros. E é constante a presença de *outliers*.

4.3.9. CATEGORIA_VIATURA

É uma variável qualitativa nominal identifica as viaturas de acordo com a categoria definida pela Seguradora. Assim, as viaturas são agrupadas em 11 tipos: Ciclomotor, Galera, Jeep, Ligeiro Comercial Deriv. Turismo, Ligeiro Comercial,

Ligeiro de Passageiros, Motociclo, Pesado Especial, Pesado de Mercadorias, Pesado de Passageiros e Táxi, conforme figura 4.24.

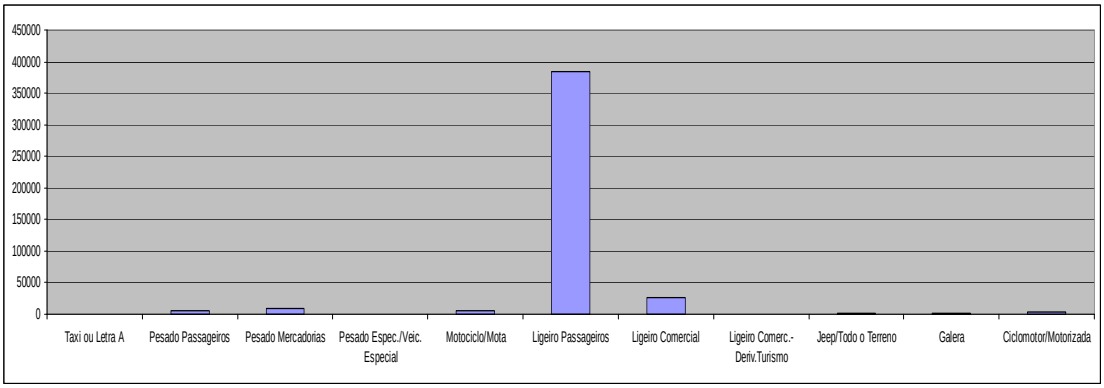

Figura 4.24 – Histograma do número médio de acidentes por categoria de viatura segura entre 2000 e 2010

Fonte: Elaboração própria

É claramente evidente que a esmagadora maioria das observações ocorre nos veículos ligeiros de passageiros. O segundo maior grupo de observações verifica-se na categoria de ligeiros comerciais, mas com valores muito abaixo. Há um terceiro conjunto de categorias (Pesados de passageiros, Pesado de mercadorias, Motociclos e ciclomotores) que apresentam alguma relevância. Salienta-se a categoria de "ligeiro de passageiro" como a categoria mais frequente no registo de acidentes. Categorias como "táxi", "galeras", entre outros quase não têm expressão.

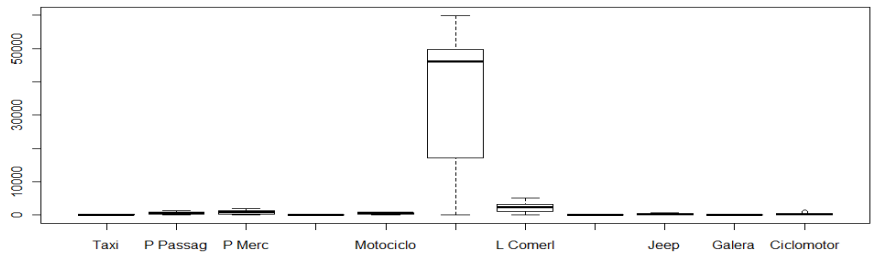

Figura 4.25 – Diagrama caixa-de-bigodes das onze categorias de viaturas seguras, no período 2000-2010

Fonte: Elaboração própria

A figura 4.25 confirma as conclusões acima expostas, há uma acentuada assimetria negativa e significativa dispersão, evidente pela distância entre o valor mínimo e o valor da mediana na categoria de "ligeiros de passageiros",

Demonstra-se, ainda, forte variabilidade dos dados.

No sentido de facilitar a análise procedeu-se a um reagrupamento de categorias, tendo em conta a similaridade técnica entre categorias. As novas categorias passam a ser:

- Ligeiros de Passageiros (veículos ligeiros de passageiros, monovolumes e táxis);
- Ligeiros Comerciais (ligeiros comerciais, veículos *pick-up*);
- Pesados (pesados de mercadorias e de passageiros, empilhadores, escavadoras, em suma viaturas com peso superior a 3500 kg que podem utilizar a via pública);
- Motas (inclui veículos de duas rodas, velocípedes com e sem motor e veículo que não carecem de carta de condução por parte do respectivo condutor).

Importa referir que existe um número de categorias "não definidas" designadas de "n.d.", são constituídas por viaturas sem classificação (e.g. seguros de carta) e representam menos de 1% dos veículos e como tal não foi tida em conta.

Analisando detalhadamente cada uma destas quatro categorias tem-se:

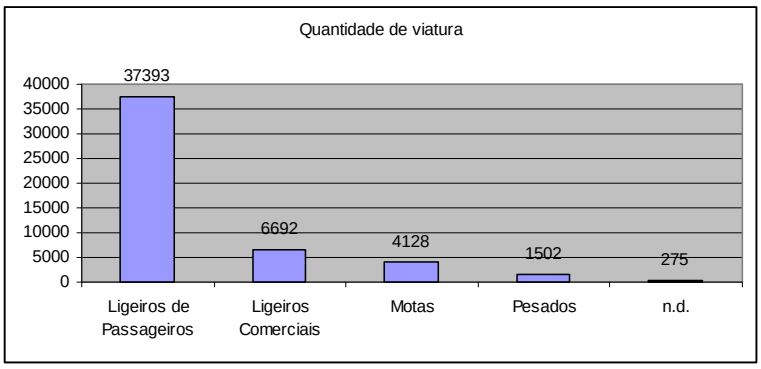

Figura 4.26 – Quantidade de viaturas existentes em cada tipo de viatura

Fonte: Elaboração própria

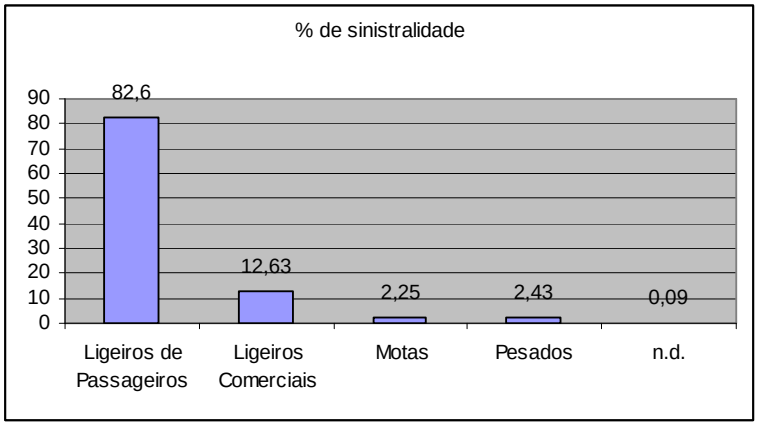

Figura 4.27 – Sinistralidade para cada um dos tipos de viatura.

Fonte: Elaboração própria

Pela análise conjunta das figuras 4.26 e 4.27 concluísse que a sinistralidade em cada categoria de viatura é proporcional à quantidade de viaturas existentes.

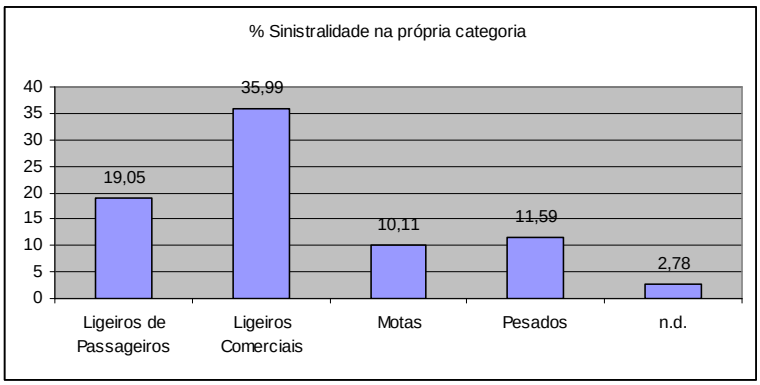

Figura 4.28 – Sinistralidade no âmbito de cada tipo de viatura
Fonte: Elaboração Própria

Quando analisada a sinistralidade no âmbito de cada categoria de veículo conforme figura 4.28, verifica-se os veículos comerciais é a classe de viatura que mais tem acidentes, 36% dos veículos comerciais já sofreu um acidente.

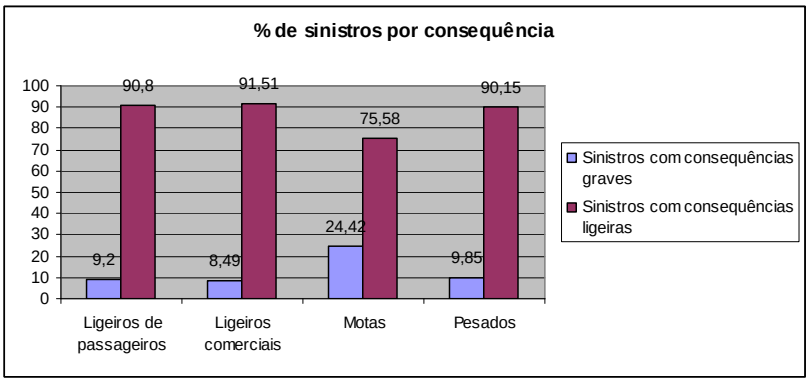

Figura 4.29 – Distribuição dos sinistros por consequência em função do tipo de viatura
Fonte: Elaboração própria

A figura 4.29 permite concluir não haver relação entre a quantidade de sinistros e as suas consequências.

4.3.10. ANO_CONSTRUÇÃO

É uma variável numérica que indica a idade da viatura, mas para uma melhor análise procedeu-se ao agrupamento por classes.

Constata-se pela observação da figura 4.30 que até 2005 as viaturas com maior número de observações de sinistros ocorria em viaturas com idade até 5 anos de idade. A partir de 2006, inclusive, o maior número de observações passa a verificar-se em viaturas que têm entre 6 e 10 anos. Denotando-se um envelhecimento do parque automóvel nos últimos anos. É evidente uma assimetria à esquerda denotando uma maior frequência de acidentes nas viaturas mais recentes.

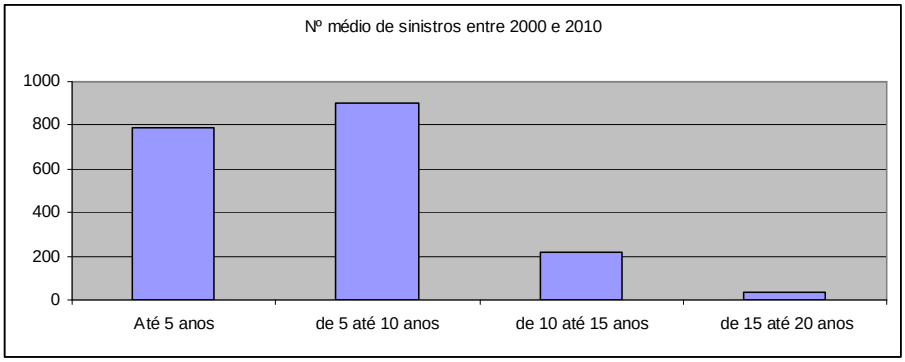

Figura 4.30 – Histograma do número médio de acidentes por idade de viatura entre 2000 e 2010

Fonte: Elaboração própria

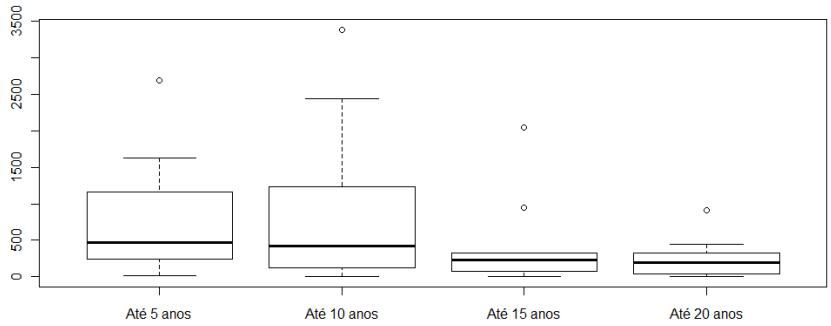

Figura 4.31– Diagrama caixa-de-bigodes das viaturas seguras com sinistros, por idade, para o período 2000-2010.

Fonte: Elaboração própria

Pela figura 4.31 é notória presença de *outliers* em todas as classes. Regista-se uma significativa concentração das observações nos valores inferiores, revelando assimetria positiva de dados. É, evidente nas classes]0, 5] e]5, 10] uma forte variabilidade dos dados. Há semelhança da idade do condutor entrará nos modelos como variável numérica.

4.3.11. MOTIVO_REEMBOLSO

É uma variável qualitativa nominal, a qual explica em que condições é que Seguradora tem a reaver os montantes indemnizados. Existem quinze situações nas quais tal situação pode ocorrer: Abandono, Alcoolemia, Apólice anulada, Carga mal acondicionada, Condutor sem carta de condução, Dolo, Factura Hospitalar, Fraude/Falsas Declarações, Franquia, Franquia em RC, Inspecção Periódica, Reembolso na cobertura CRT, Responsabilidade de Terceiros, Roubo e Sem Seguro.

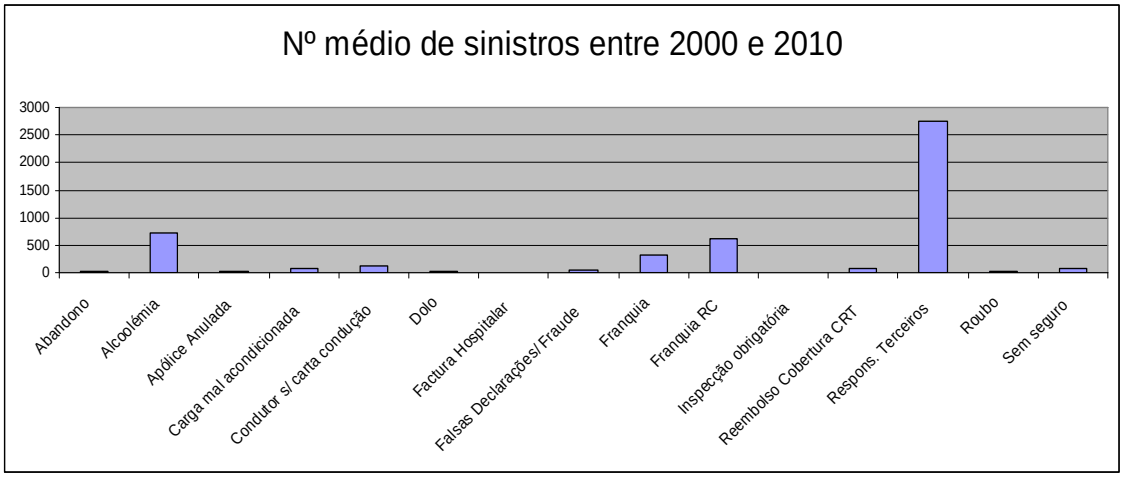

Figura 4.32 – Histograma do número médio de acidentes por tipo de reembolso entre 2000 e 2010

Fonte: Elaboração própria

De acordo com a figura 4.32 os motivos mais frequentes para reembolso são "Responsabilidade de terceiro" e a "Alcoolemia". Destaca-se que um grande número de "motivos de reembolso" apresenta valores para muito pouco significativos.

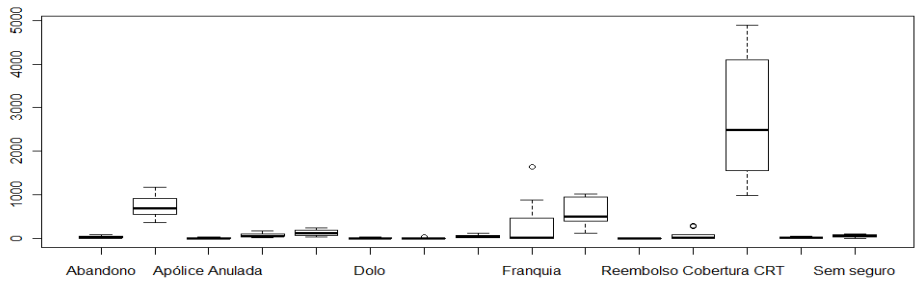

Figura 4.33– Diagrama caixa-de-bigodes das viaturas seguras com sinistros onde existe motivo de reembolso para o período 2000-2010.

Fonte: Elaboração própria

Constata-se pela figura 4.33 a existência de quatro motivos de reembolso assinaláveis, com níveis de simetria e variabilidade de dados bastante distintos. Enquanto o motivo "alcoolemia" apresenta uma distribuição quase simétrica, os motivos "Franquia" evidência uma total assimetria positiva, o motivo "Responsabilidade de terceiros" regista o maior número de observações e uma grande variabilidade dos dados.

Do ponto de vista da modelação e com base na revisão da literatura apenas a variável "alcoolemia" tem interesse em ser considerada, uma vez que influência o factor humano, aquele que é preponderante para a ocorrência do acidente (Oliveira, 2007).

4.4. A escolha das variáveis para os modelos

De acordo com Richardson (1989) e Hosmer & Lemeshow (2000) não há regras inflexíveis na escolha das variáveis, estas alteram em conformidade com a realidade ou problema em estudo. Importa que todas as variáveis apresentem aspectos do fenómeno em análise, bem como diferenciação entre si na explicação do referido fenómeno.

Há que ter alguns cuidados na construção dos modelos GLM (Agresti, 1996), por exemplo um modelo com muitas variáveis explicativas tem maior tendência em incorrer em problemas de multicolinearidade, situação que ocorre quando há variáveis fortemente correlacionadas, tal situação pode trazer problemas de estimação dos modelos e poderá ser detectada através da matriz de correlações (Johnston & Dinardo, 1996). Assim, para avaliar o grau de associação linear entre as variáveis irá proceder-se ao cálculo do coeficiente de correlação de *Pearson* ou, simplesmente, coeficiente de correlação – r , (Murteira et al., 2010) através de:

$$r = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Onde S_x e S_y representam o desvio padrão, respectivamente das variáveis x e y. Importa, ainda, referir que $-1 \le r \le 1$, quaisquer que sejam os valores assumidos pelas variáveis x e y. Nos casos em que valor de r se aproxima do valor 0 as variáveis não aparentam qualquer padrão de associação linear, nas situações em que r se afasta do valor 0 conclui-se pela inexistência um padrão não linear de associação, sendo que o

sinal do coeficiente indica se a correlação é positiva ou negativa (Murteira et al., 2010). Deverão, assim, rejeitarem-se as variáveis multicolineares, ou seja, aquelas onde se denota uma associação não linear entre si.

De acordo com Hosmer & Lemeshow (2000), a abordagem tradicional para a construção de qualquer modelo baseia-se no princípio da parcimónia, segundo o qual se deve minimizar o número de variáveis do modelo de forma a obter um modelo mais estável possível, mais facilmente generalizado e evitando problemas de sobredispersão.

A sobredispersão ocorre com frequência em dados referentes a sinistralidade automóvel (Harrel et al., 1996; Hibe, 2007; Oh et al., 2006) e pode dever-se ao facto de existir heterogeneidade entre indivíduos não explicada pelas variáveis, ou por haver correlação entre as observações.

Face ao acima exposto, partindo tabela de dados inicial, sobre a qual se efectuou uma análise exploratória dos dados, e considerando, também, a revisão da literatura efectuada no Capítulo 2, onde se verificou que em estudos efectuados no âmbito dos seguros, nomeadamente Bailey & Simon (1960) e Brockman & Wright (1992), utilizaram nas suas análises as seguintes variáveis: idade, experiência (entenda-se anos de carta do condução), género, zona geográfica e características do veículo.

Conforme já foi referido, de acordo com Harvey (2004), Hoofmann (2005), Oliveira (2007) e Penden et al. (2004) o factor predominante que está na origem dos acidentes de viação é o factor humano acrescido de factores externos (e.g. condições de segurança da viatura; condições da estrada, condições climatéricas, etc.).

Segundo Denuit et al. (2007) as variáveis que acarretam risco são divididas em dois grupos: variáveis *à priori*, e variáveis *à posteriori*. Nas primeiras, os seus valores são determinados antes da apólice ter início, nas segundas só se detectam quando o contrato de seguro já está em curso.

De acordo com Antonio & Valdez (2010), Denuit et al. (2007), Harvey (2004), Hoffman (2005), Li et al. (2003), Oliveira (2007), Penden et al. (2004) e Petridou & Moustaki (2000) são exemplos de variáveis *à priori*: idade, género, ocupação, estado civil, morada, idade do veículo, tipo de viatura, cor da viatura. E são exemplos de variáveis *à posteriori*: paciência, habilidade, agressividade, cultura de desresponsabilização, desempenho do condutor, conhecimento do Código da Estrada.

Nesta fase o objectivo é identificar as melhores variáveis explicativas *à priori* (X_i) de forma a garantir que as variáveis relevantes sejam incluídas nos modelos para

uma correcta avaliação do o risco dos indivíduos. A selecção das variáveis assentou em quatro factores:

- Existência de dados disponíveis;
- Principio de parcimónia;
- Abrangência das três principais dimensões que estão na origem dos acidentes de viação, conforme referido anteriormente: factor humano, características da viatura, condições ambientais de tempo e espaço;
- Valores das correlações entre variáveis (cujos valores constam na tabela 4.7.).

Assim, escolheram-se como variáveis explicativas para estimação dos modelos:

- Idade (do segurado/individuo);
- Género (do segurado/individuo);
- Região (do segurado/individuo);
- Tipo de veículo (do segurado/individuo);
- Idade do veículo (do segurado/individuo);
- Data (do acidente);
- Hora (do acidente);
- Alcoolémia (no acidente).

4.5. "Variáveis explicativas" não consideradas

Nem todas as possíveis variáveis possíveis apresentadas no ponto 4.4. são passíveis de serem consideradas nos modelos econométricos.

A variável "Responsabilidade" apresentada no ponto 4.3.5. e a qual representa o histórico do segurado no que respeita a responsabilidade em acidentes de viação é uma variável que caracteriza o acidente, sendo por esse motivo endógena ao mesmo, e como tal não deverá ser usada como variável explicativa, uma vez que as variáveis explicativas deverão ser exógenas. Há que garantir que as variáveis explicativas consideradas não sejam endógenas, ou seja, não sejam influenciadas pelo fenómeno em estudo, sob pena de comprometer os resultados dos modelos (Wooldridge, 2002).

Relativamente à variável "Causa" apresentada no ponto 4.3.6. as categorias apresentadas não são mutuamente exclusivas, há inúmeras causas para a ocorrência de um acidente de viação. Dos dados disponibilizados foram identificadas vinte e seis

causas possíveis, onde uma delas é designada de “Outras causas”. Face ao exposto, não sendo possível delimitar a extensão desta variável, não será de considerá-la nos modelos.

4.6. Forma como as variáveis explicativas vão entrar nos modelos

De acordo com Agresti (1996), Cameron & Trivedi (1998), Hosmer & Lemeshow (2000), Long (1997), McCullagh & Nelder (1989), Winkelmann (2000) entre outros autores, não há regras inflexíveis na escolha das variáveis.

No caso das variáveis que apresentavam características numéricas, como a “Idade” e “Idade do veículo”, optou-se por considerá-las, naturalmente, como variáveis numéricas. As restantes variáveis serão incluídas no modelo como variáveis não numéricas, ou categóricas: “Género”, “Região”, “Tipo de veículo”, “Data”, “Hora”, “Alcoolemia”. A inclusão das variáveis categóricas nos modelos de regressão efectua-se em duas fases (Gelman & Hill, 2007):

Numa primeira fase, codificam-se as variáveis em níveis (categorias):

- “Género” = 1 para Homem e 0 para Mulher;
- “Região” = 1 para a Lisboa e Porto e 0 para as restantes regiões;
- “Tipo de veículo” = 1 para ligeiros passageiros, 2 para ligeiros comerciais, 3 para motas, 4 para pesados;
- “Data” = 1 para Domingo, 2 para 2ª feira, 3 para 3ª feira, 4 para 4ª feira, 5 para 5ª feira, 6 para 6ª feira e 7 para Sábado.
- “Hora” = 1 para o período 00h às 06h, 2 para o período das 06h às 12h, 3 para o período das 12h às 18h e 4 para o período das 18h às 24h
- “Alcoolemia” = 1 para quando há alcoolemia, 0 para quando não há alcoolemia.

Numa segunda fase, recorre-se à utilização de variáveis binárias (variáveis com apenas dois resultado possíveis) para cada uma das categorias dentro da mesma variável.

Desta forma, para as variáveis “Género”, “Região” e “Alcoolémia”, tem-se:

Tabela 4.1 - *Variáveis binárias para a variável “Género”*

Variàvel: Género	**Valores possíveis**
Homem	1
Mulher	0

Fonte: Elaboração própria

A variável "Género" vai entrar nos modelos como variável binária HOMEM, sendo a variável binária de referência MULHER.

Tabela 4.2 - *Variáveis binárias para a variável "Região"*

Variàvel: Região	Valores possíveis
Lisboa e Porto	1
Restantes Regiões	0

Fonte: Elaboração própria

A variável "Região" vai entrar nos modelos como variável binária REGIAO para os casos em que o segurado pertença à região de Lisboa ou do Porto, (os dois maiores centros urbanos), sendo a variável binária de referência RESTANTES REGIÕES.

Tabela 4.3 - *Variáveis binárias para a variável "Alcoolémia"*

Variàvel: Alcoolémia	Valores possíveis
Com alcoolémia	1
Sem alcoolémia	0

Fonte: Elaboração própria

A variável "Alcoolémia" vai entrar nos modelos como variável binária ALCOOLEMIA quando no acidente o segurado acusa taxa de álcool no sangue positiva sendo a variável binária de referência SEM ALCOOLEMIA.

Nas situações em que existem várias categorias para a variável explicativa, serão criadas tantas variáveis binárias quantas as necessárias para definir todas as categorias. Assim, para as variáveis "Tipo de veículo", "Data" e "Hora" tem-se:

Tabela 4.4 - *Variáveis binárias para a variável "Tipo de Viatura"*

Variàvel: Tipo de Viatura	Apólice 1	Apólice 2	Apólice 3	Apólice 4	…	Apólice n
Ligeiros	0	1	0	0	…	0
Comerciais	0	0	0	0	…	1
Motas	0	0	0	1	…	0
Pesados	0	0	0	0	…	0

Fonte: Elaboração própria

A variável "Tipo de Viatura" vai entrar nos modelos com as variáveis binárias LIGEIROS, COMERCIAIS e MOTAS sendo a variável binária de referência

PESADOS. A decisão pela variável binária de referência teve por base o facto de ser uma das categorias com frequência de sinistralidade baixa reduzida, conforme se pode verificar no ponto 4.3.9.

Tabela 4.5 - *Variáveis binárias para a variável "Data"*

Variàvel: Data	Apólice 1	Apólice 2	Apólice 3	Apólice 4	…	Apólice n
2ª f.	0	0	1	0	…	0
3ª f.	0	1	0	0	…	0
4ª f.	0	0	0	1	…	0
5ª f.	0	0	0	0	…	1
6 f.	0	1	0	0	…	0
Sábado	0	0	0	0	…	0
Domingo	0	0	0	0	…	0

Fonte: Elaboração própria

A variável "Data" vai entrar nos modelos com as variáveis binárias dos dias da semana, sendo o dia da semana de referência o DOMINGO por ser a categoria com frequência de sinistralidade reduzida, conforme se pode verificar na figura 4.5.

Tabela 4.6 - *Variáveis binárias para a variável "Hora"*

Variàvel: Hora	Apólice 1	Apólice 2	Apólice 3	Apólice 4	…	Apólice n
Zero-Seis	0	0	0	0	…	0
Seis-Doze	0	1	0	0	…	0
Doze-Dezoito	0	0	0	1	…	0
Dezoito-vintequatro	0	0	0	0	…	1

Fonte: Elaboração própria

A variável "Hora" vai entrar nos modelos com as variáveis binárias correspondentes às seguintes categorias: ZERO-SEIS, SEIS-DOZE, DOZE-DEZOITO, DEZOITO-VINTEQUATRO, por serem períodos distintos em termos de ocorrência de sinistralidade, conforme figura 4.6., sendo o período que fica de fora como período de referência o ZERO-SEIS, por ser o que aquele onde se regista a menor frequência de acidentes. Os intervalos das categorias são fechados à esquerda e abertos à direita.

4.7. As correlações das variáveis explicativas

Depois de escolhido um grupo de potenciais variáveis explicativas importa aferir sobre o seu grau de associação linear.

Tabela 4.7 - *Matriz de correlações das variáveis explicativas dos modelos.*

	Idade	Regiao	Homem	Ligeiros	Comerciais	Motas	Idade veículo	2ªf.	3ªf.	4ªf.	5ªf.	6ªf.	Sabado	seis-doze	doze-dezoito	dezoito-v.quatro	Acodemia
Idade	1,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Regiao	0,075	1,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Homem	0,178	-0,049	1,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ligeiros	-0,031	0,149	-0,132	1,000	-	-	-	-	-	-	-	-	-	-	-	-	-
Comercial	-0,012	-0,097	0,051	-0,658	1,000	-	-	-	-	-	-	-	-	-	-	-	-
Motas	-0,018	-0,059	0,101	-0,500	-0,129	1,000	-	-	-	-	-	-	-	-	-	-	-
Idade veículo	0,170	-0,135	0,114	-0,191	0,090	0,091	1,000	-	-	-	-	-	-	-	-	-	-
2ªf.	-0,008	0,017	-0,001	0,005	0,003	-0,043	-0,076	1,000	-	-	-	-	-	-	-	-	-
3ªf.	-0,005	0,013	-0,007	0,027	0,008	-0,039	-0,065	-0,028	1,000	-	-	-	-	-	-	-	-
4ªf.	-0,008	0,025	-0,015	0,029	0,003	-0,036	-0,073	-0,029	-0,026	1,000	-	-	-	-	-	-	-
5ªf.	-0,002	0,020	-0,007	0,026	0,007	-0,037	-0,068	-0,028	-0,026	-0,026	1,000	-	-	-	-	-	-
6ªf.	-0,025	0,019	-0,005	0,034	-0,001	-0,034	-0,066	-0,029	-0,027	-0,027	-0,026	1,000	-	-	-	-	-
Sabado	-0,030	0,011	-0,003	0,021	0,006	-0,029	-0,055	-0,025	-0,023	-0,023	-0,023	-0,024	1,000	-	-	-	-
seis-doze	-0,063	0,021	-0,006	0,042	0,014	-0,063	-0,113	0,235	0,212	0,215	0,216	0,182	0,171	1,000	-	-	-
doze-dezoito	-0,035	0,024	-0,013	0,038	0,014	-0,055	-0,092	0,232	0,182	0,202	0,182	0,208	0,173	-0,065	1,000	-	-
dezoito-v.quatro	-0,038	0,030	-0,008	0,045	-0,013	-0,037	-0,077	0,116	0,141	0,131	0,141	0,163	0,115	-0,045	-0,045	1,000	-
Acodemia	0,002	0,000	0,001	0,009	-0,009	-0,003	-0,001	0,007	0,033	0,003	0,009	0,002	0,038	-0,001	0,019	0,024	1,000

Fonte: Elaboração própria

Apesar da maioria das variáveis explicativas serem categóricas, não existe problema no cálculo aproximado da correlação, uma vez que a característica fundamental de tais variáveis reside no facto de, embora sendo qualitativas, poderem ser ordenadas (Murteira et al., 2010). De acordo com a tabela 4.7 não existem correlações significativamente fortes entre as variáveis escolhidas, pelo que são garantias de estimativas não inflacionadas dos coeficientes de regressão (Hosmer & Lemeshow, 1989). Assim, será de aceitar as variáveis explicativas propostas como adequadas para estimação dos modelos econométricos de avaliação de risco.

4.8. Forma como as variáveis dependentes vão ser apresentadas nos modelos

As variáveis dependentes foram criadas em função do que é objectivo desta dissertação, ou seja, a avaliar o risco para o acidente de viação.

Esta avaliação pode ser feita sob duas perspectivas, na perspectiva do indivíduo onde se avaliará a probabilidade para o acidente e o número de acidentes, e nesse sentido serão criadas duas variáveis dependentes: “LogAcidentes” e “ContAcidentes”.

Tabela 4.8 - *Variável binária para a variável existência de acidente*

Variável: LogAcidentes	**Valores possíveis**
Não teve acidente	0
Teve acidente (um ou mais)	1

Fonte: Elaboração própria

A variável LogAcidentes na tabela 4.8 surge a partir da criação de um novo campo na base de dados, o qual é preenchido a 0 e 1 consoante as apólices não registem a presença de sinistralidade ou registe sinistralidade, respectivamente.

Tabela 4.9 *Variável binária para a variável número de acidentes*

Variável: ContAcidentes	**Valores possíveis**
Não teve acidente	0
Teve um acidente	1
Teve dois acidentes	2
Teve três acidentes	3
…	…
Teve n acidentes	n

Fonte: Elaboração própria

A variável ContAcidentes surge a partir da criação de um novo campo na base de dados, o qual é preenchido com valores iguais ou superiores a zero consoante o número de acidente que cada apólice regista.

A outra perspectiva na avaliação para o risco/propensão para o acidente, é aquela que poderá ser feita a partir dos acidentes que se vão ocorrendo, assim, poderá ser avaliada a gravidade do acidente através variável dependente: Gravidade.

Tabela 4.10 *Variável binária para a gravidade dos acidentes*

Variável: Gravidade	**Valores possíveis**
Acidente ligeiro	0
Acidente grave	1

Fonte: Elaboração própria

Foi criado na base de dados o novo campo Gravidade, a partir campo já existente designado de “CONSEQUÊNCIA”. Este novo campo foi preenchido com os valores de 0 caso a consequência fosse “danos materiais”, e com valor 1 caso a consequência fosse “danos materiais/corporais” ou “danos corporais”.

4.9. Síntese das variáveis endógenas e exógenas dos modelos

No seguimento dos três pontos anteriores, sistematizou-se na tabela 4.11 o conjunto de variáveis endógenas e exógenas dos modelos de regressão, a sua correspondente origem na tabela de dados fornecidos pela Companhia de Seguros, modelos nos quais serão utilizadas, designações, escalas e respectivos acrónimos.

Tabela 4.11 - *Resumo das variáveis dos modelos*

Campo da tabela	Modelo onde será utilizada	Variável	Descrição	Escala	Acrónimos
DATA_SINISTRO	Modelos na perspectiva do acidente	Dia	Trata-se de uma variável explicativa categórica. Indica o dia da semana na qual se verifica a ocorrência do sinistro. Vai entrar no modelo como variàvel *dummy* , onde o dia Domingo estará em oposição com as restantes.	1 - Domingo 2 - 2ª feira 3 - 3ª feira 4 - 4ª feira 5 - 5ª feira 6 - 6ª feira 7 - Sábado	SegundaFeira TercaFeira QuartaFeira QuintaFeira SextaFeira Sabado
HORA_SINISTRO	Modelos na perspectiva do acidente	Hora	Trata-se de uma variável explicativa categórica. Corresponde ao período do dia em que ocorre o acidente. As várias horas do dia foram agrupadas em conjuntos de 6 horas por serem períodos distintos em termos de frequência de sinistros. Vai entrar no modelo como variàvel *dummy* e primeiro período do dia (madrugada) estará em oposição aos restantes.	1 - [00h, 06h[2 - [06h, 12h[3 - [12h, 18h[4 - [18h, 24h[	seis_doze doze_dezoito dezoito_vintequatro
Foi criada para ser possível a estimação do modelo. Partiu-se da contagem do número de acidentes registados por apólice, sendo que quando não se registavam acidentes o campo era preenchido a 0. A ocorrência de pelo menos um acidente é sempre classificada com 1.	Modelos na perspectiva do condutor	LogAcidente	Trata-se da variàvel dependente (Binária) do modelo na perspectiva do condutor para a regressão Logistica. Indica mediante uma dada combinação de variàveis explicativas se ocorre ou não acidente.	0 - Não teve acidente 1 - Teve pelo menos acidente	Condutor_c_s_Acidentes
Foi criada para ser possível a estimação do modelo. Apresenta o número que acidente da apólice	Modelos na perspectiva do condutor	ContAcidente	Trata-se da variável dependente numérica do modelo na perspectiva do condutor para os modelos de contagem. Indica mediante uma dada combinação de variàveis explicativas o número de acidentes que ocorre.	>= 0	Condutor_n_Acidentes
CONSEQUENCIA	Modelo na perspectiva do acidente	Gravidade	Trata-se da variável dependente (Binária) do modelo na perspectiva do acidente. Indica se o acidente produziu sequelas com ou sem gravidade. Considera-se grave um acidente com danos corporais, e considera-se sem gravidade um acidente apenas com danos materiais.	0 - Danos ligeiros 1 - Danos graves	Gravidade

REGIAO_SINISTRO	Modelo na perspectiva do acidente	Região	Trata-se de uma variável explicativa correpondente à zona geográfica de Portugal onde se regista o acidente. Entra no modelo como variável dummy no qual as duas maiores regiões urbanas (Lisboa e Porto) estão em oposição às restantes regiões.	0 - Outras regiões 1- Lisboa e Porto	Regiao
SEGURADO_DATA_NASC	Modelos na perspectiva do condutor e na perspectiva do acidente	Idade	Trata-se de uma variável numérica. É a idade do segurado da apólice.	[18, 90]	Idade
SEGURADO_SEXO	Modelos na perspectiva do condutor e na perspectiva do acidente	Género	Trata-se de uma variável categórica explicativa. Corresponde ao género do condutor do segurado. O género feminino estará em oposição ao masculino.	0 - Feminino 1 - Masculino	Homem
CATEGORIA_VIATURA	Modelos na perspectiva do condutor e na perspectiva do acidente	Tipo viatura	Trata-se de uma variàvel categórica explicativa. Indica o tipo de viatura segura. Os veículo do tipo "veículos pesados" estará em oposição aos restantes.	1 - Ligeiros passageiros 2 - Ligeiros comerciais 3 - Pesados 4 - Motas	LigeiroPassageiros LigeiroComercial Motas
ANO_CONSTRUÇÃO	Modelos na perspectiva do condutor e na perspectiva do acidente	Idade da viatura	Trata-se de uma variável explicativa numérica. É a idade do veículo seguro	> 0	IdadeVeiculo
MOTIVO_REEMBOLSO	Modelos na perspectiva do acidente	Alcoolemia	Trata-se de uma variável binária explicativa. Reflecte o facto do segurado acusar taxa de alcoolémia superior a 0,5 g/l no sangue durante o acto de condução	0 - TAS < 0,5 g/l 1 - TAS >= 0,5 g/l	alcoolemia

Fonte: Elaboração própria

5. A ESTIMAÇÃO DOS MODELOS

Neste capítulo procede-se à estimação dos modelos de regressão, tendo por base a metodologia apresentada no capítulo 3 e a análise e modelação de dados efectuada no capítulo 4.

5.1. Caracterização espacial e temporal da análise

Do capítulo anterior decorre que o espaço físico considerado para este estudo é o território nacional português e a dimensão temporal abrange o período de 2000 a 2010.

Desta forma, os dados que servem de base a este estudo têm aspectos seccionais e temporais, sendo por esse motivo designados de danos seccionais combinados ou de *poolled cross sections* (Murteira et al., 2010). As observações referem-se a determinadas unidades estatísticas (apólices) com um conjunto de dados seccionais relativos aos tomadores de seguros (idade do seguro, género, registo ou não de acidente, em caso afirmativo características do acidente, etc.), num certo período de tempo (durante 11 anos), em Portugal.

5.2. A opção pela extracção de uma amostra

De acordo com inúmeros autores (e.g. Vicente et al., 2001) várias razões contribuem para a não utilização de todos os elementos da população, nesta dissertação três factores levaram à opção por uma amostra:

- Quando a dimensão da amostra aumenta os ganhos de precisão são mínimos;
- O tempo excessivo de processamento computacional (mais de um milhão de registos) para realização da estimação dos modelos;
- O enorme número de registos da base de dados pode trazer problemas de "inflação" da significância dos parâmetros;
- Mesmo a totalidade dos registos não constitui por si só a população correspondente à totalidade de (potenciais) tomadores de seguro.

O sucesso do desenvolvimento dos modelos de regressão, que serão desenvolvidos mais à frente, depende essencialmente da dimensão e representatividade [1] da amostra de dados recolhida.

No planeamento de uma amostra a questão que merece maior atenção é a decisão quanto ao número de indivíduos a considerar (n) que a amostra deve conter (Vicente et. al., 2001). Nesta dissertação, considerando que a base de dados tem 3.878.856 registos, qualquer amostra com um número de observações inferior a 1.000 traduz-se numa amostra de dimensão inferior a 0,03% do total de registos da base de dados, o que é um valor muito baixo.

Assim, a amostra não deve ser pequena, sob pena de não ser "representativa" da população e, nos modelos que irão ser utilizados, também não se deverá utilizar uma amostra suficientemente grande sob pena de trazer problemas de "inflação" da significância dos parâmetros e problemas computacionais.

Importa, ainda, referir que no âmbito dos acidentes de viação deverá existir uma amostra considerável de dados de vários anos (Denuit et al., 2007), de forma a esbater a importância de algum ano de calendário no qual as condições atmosféricas atípicas poderão ter aumentado ou diminuído significativamente o números de acidentes.

5.3. A decisão pela amostragem aleatória simples

Uma amostra pode ser obtida através de diversos critérios, mas qualquer que seja o tipo de amostragem escolhida o princípio orientador da decisão deverá ser o da eficiência, ou seja, obter a informação mais rigorosa possível com o mínimo custo (Vicente et. al., 2001).

Para se obterem conclusões válidas, a maioria os métodos estatísticos assume que a amostra é obtida de modo aleatório, ou seja, é conhecida a probabilidade com que determinado elemento da população pode ser escolhido, ou seja é igual a n/N, e a escolha de um dado elemento não influência a escolha de outro(s). Os métodos aleatórios são os que melhor asseguram a representatividade da mostra (Black, 1993).

[1] De acordo com Vicente et al. (2001) *a amostra representativa é aquela que reflecte os aspectos típicos, as características mais relevantes da população. É uma espécie de maquete, que capta, para o estudo em concreto, as características mais relevantes da população.*

A amostra aleatória simples é o tipo de amostragem probabilística mais utilizada quando todos os elementos da população estão numerados e listados numa base de dados. Garante a exactidão e eficácia à amostragem, para além de constituir um procedimento de mais fácil aplicação (Vicente et. al., 2001).

5.4. A escolha da amostra e sua dimensão

Conforme já foi referido anteriormente o Grupo Caixa Seguros cedeu os dados referentes ao período de 2000 a 2010, das três Seguradoras que o compõem: Fidelidade-Mundial, Império-Bonança e Via Directa. Essa base de dados é constituída por dois ficheiros distintos: um ficheiro com os dados das apólices (nome do segurado, morada, viatura segura, etc.) e um ficheiro com os dados dos sinistros (data, local, causa, etc.), os quais apresentam dimensões diferentes quer em número de registos, quer em número de atributos.

Considerando o objectivo da dissertação: avaliação do risco e propensão para o acidente; tal pressupõe que a informação tenha de ficar concentrada num único ficheiro, pois só desta forma será possível estimar correctamente o risco para o acidente de viação. Caso fosse utilizado apenas o ficheiro dos sinistros, seria estimado o risco dos indivíduos que já tiveram acidentes. Teve de ser feita a junção à base de dados do ficheiro das apólices para incluir na amostra indivíduos que nunca tiveram acidentes.

Assim, a informação dos dois ficheiros foi reunida numa única base de dados, sendo o campo de ligação entre os ficheiros e a ordenação da nova base de dado feitos através do campo: APÓLICE.

O primeiro registo da tabela é o número de apólice, podendo essas apólices ter, ou não, inseridos acidentes de viação. Se a apólice tiver acidentes todos os campos desse registo estão preenchidos. Se a apólice não tiver acidentes, os campos respeitantes aos dados dos acidentes estão a “null”.

Optou-se por excluir a Companhia de Seguros Via Directa uma vez que o seu *target* é bastante diferente das restantes seguradoras e os seus dados disponíveis reportam-se a um período mais restrito (2007-2010).

Considerando que o perfil de clientes, tipo de produto e canais de distribuição das Companhia de Seguros Fidelidade-Mundial e Império-Bonança são comuns, optou-

se por excluir a Companhia de Seguros Fidelidade Mundial do estudo, uma vez que os dados disponíveis desta última seguradora reportam-se a um período mais curto (2007-2010), do que aquele que é pretendido.

A nova base de dados é constituída por 53 campos e por 3.878.856 registos.

Considerando que nem todas as apólices têm registo de acidentes, nesses casos os campos que dizem respeito ao sinistro (data, local, causa, gravidade, etc,) são preenchidos a "null".

Partindo da nova base de dados procedeu-se à análise exploratória de dados, no capítulo 4, para se conhecer os dados, detectar padrões e relações, com vista à identificação das variáveis endógenas e exógenas a utilizar na estimação dos modelos.

Da exploração de dados ressaltou-se um aspecto importante: quando comparadas as idades dos segurados por género, temos os seguintes valores médios:

Tabela 5.1 - *Idade média pelo género de segurado*

Variàvel	Homens	Mulheres	Empresas
Idade média	43,3	40,8	4,9

Fonte: Elaboração própria

Tal situação decorre do facto, do género do condutor de empresa ser indefinido e como tal a sua idade corresponde em alguns casos à idade da viatura segura e noutros à data de constituição da empresa.

Assim, optou-se por excluir a classe empresa não só devido à disparidade de valores médios do género empresa face aos restantes, o que tratará problemas à estimação dos modelos, mas há também que considerar que sendo o objectivo desta dissertação a avaliação da propensão dos indivíduos para o acidente, a categoria empresa, dada a sua especificidade representa uma realidade que não se enquadra no presente estudo.

Face à situação acima exposta, dada a grande dimensão desta base de dados e considerado a sua inoperabilidade do ponto de vista informático para estimação dos modelos, entendeu-se, com base na revisão da literatura, e.g. Abbring et al. (2003); Brockman & Wright (1992); Levitt & Porter (1999); Li et al. (2001) e Visser et al. (2006) proceder à extracção de amostras aleatórias simples, apenas para o subconjunto dos particulares (Homens e Mulheres). Após análise prévia de amostras de menor dimensão, verificou-se que à medida que a dimensão da amostra aumentava o erro de

amostragem diminuía, aumentava o número de graus de liberdade e a estimação tornava-se mais robusta. Optou-se pela amostra com dimensão de 50.000 observações, por ser a dimensão que produziu resultados mais satisfatórios.

5.5. Hipóteses de Trabalho

Importa salientar que o risco é uma exposição para a possibilidade ou probabilidade de algum resultado, tipicamente pensado como uma perda ou dano (Hilbe, 2007). Neste trabalho o risco é, portanto, uma medida de probabilidade.

De acordo com revisão da literatura e conforme se tem vindo a referir ao longo deste trabalho, num sinistro automóvel o risco está presente sobretudo através do factor humano, influenciado por algumas condicionantes externas.

Assim, de acordo com a análise de dados, e com as metodologias apresentadas, será avaliado o risco através da estimação de modelos de acordo com duas perspectivas:

A. Na perspectiva do indivíduo/condutor, uma vez que o risco é um produto social e as pessoas, em função das suas características, percebem e incorrem em diferentes níveis de risco (Areosa, 2008).

   Nesta perspectiva e com base num conjunto de variáveis que caracterizam os indivíduos, tem-se como unidade estatística o segurado/apólice, e irá proceder-se à estimação de dois tipos de modelos:

   - Um modelo de regressão logística, onde a variável endógena LogAcidente tem resultado binário: Y = 0 para as situações em que o segurado tem zero acidentes e Y = 1 para quando o segurado tem um ou mais acidentes;
   - Dois modelos de contagem (regressão de Poisson e Binomial Negativo) onde a variável endógena ContAcidente tem resultado numérico e sempre superior ou igual a zero (Y= 0, 1, 2, 3, …) e no qual é aferido o número de acidente ocorridos para um dado individuo.

B. Na perspectiva do acidente, uma vez que o evento aleatório em si tem subjacentes aspectos que importam ser avaliados (Areosa, 2008).

   Nesta perspectiva será construído um modelo com base num conjunto mais alargado de variáveis, ou seja, para além de conter as variáveis que caracterizam

os indivíduos/apólice são consideradas variáveis que caracterizam o(s) acidente(s) registado(s) por esses mesmos indivíduos. Neste modelo a unidade estatística é o acidente e a variável dependente Gravidade apresenta o resultado binário (Y=0 ou Y=1) consoante a consequência do acidente seja pouco ou muito grave, respectivamente.

Para a estimação deste modelo foi extraído da amostra de dimensão de 50.000 registos um *data set* de dimensão de 8.625 observações, que corresponde ao conjunto de todas as apólices da amostra com registo de acidente.

Será utilizado o *software R Project*. Trata-se de uma linguagem de programação e de um ambiente computacional estatístico (Maindonald & Braun, 2006), o qual está disponível sob os termos da licença do Free Software Foundation GNU (General Public License). Este pacote possui uma variedade de metodologias estatísticas, incluindo vários modelos de regressão, permitindo que estes sejam testados e comparados.

5.6. Etapas para a construção dos modelos

De acordo com Agresti (1996), Hosmer & Lemeshow (2000), Long (1997), McCullagh & Nelder (1989), Wooldridge (2002), entre muitos outros autores, existem três etapas fundamentais a seguir na construção de modelos lineares generalizados:

- A formulação dos modelos;
- O ajustamento dos modelos;
- A selecção e validação de modelos.

Na fase de formulação dos modelos há que ter em consideração os seguintes aspectos:

- A escolha da distribuição para a variável resposta.

  No seguimento da metodologia apresentada no capítulo 3, os GLM pressupõem que a variável resposta tenha uma distribuição pertencente à família exponencial, sendo os dados de que dispomos de natureza discreta, as distribuições a considerar serão as distribuições Binomial, de Poisson e Binomial Negativa.

- A escolha das variáveis explicativas e formulação adequada da matriz de especificação.

  De acordo com a análise exploratória de dados efectuada no capítulo 4, que permitiu a compreensão do fenómeno da sinistralidade automóvel, teremos duas matrizes de variáveis: uma matriz de 8 variáveis para estimar os modelos na perspectiva do condutor e outra de 17 variáveis para estimar o modelo na perspectiva do acidente.

Na estimação dos modelos na perspectiva do indivíduo/condutor procurar-se determinar a propensão/risco para o acidente por parte de um condutor. Face ao acima exposto e tendo, também, em consideração os dados disponíveis na base de dados, partiu-se do seguinte conjunto de variáveis exógenas à sinistralidade e que caracterizam o segurado:

Tabela 5.2 - *Matriz de Especificação para modelos na perspectiva do individuo*

Variàvel	Descrição	Escala
Condutor_c_s_Acidentes	Variàvel Dependente	0 - Sem acidentes 1- Pelo menos um acidente
Condutor_n_Acidentes	Variàvel Dependente	>=0
Idade	Idade do segurado da apólice	[18; 90]
Género	Género do segurado seguro	0 - Feminino 1 - Masculino
Tipo viatura	Tipo de viatura segura	1 - Ligeiros passageiros 2 - Ligeiros comerciais 3 - Pesados 4 - Motas
Regiao	Região do país onde habita o segurado	0 - Outras regiões 1- Região de Lisboa e Porto
Idade da viatura	Idade do veículo seguro	> =0

Fonte: Elaboração própria

Para a estimação do modelo na perspectiva do sinistro pretende-se determinar a propensão/risco para o acidente por parte de um condutor, partindo de um conjunto de variáveis exógenas e endógenas à própria sinistralidade de seguida apresentadas.

Tabela 5.3 - *Matriz de Especificação para o modelo na perspectiva do acidente*

Variável	Descrição	Escala
Gravidade	Variàvel dependente	0 - Danos ligeiros 1 - Danos graves
Idade	Idade do segurado da apólice	[18; 90]
Género	Género do segurado seguro	0 - Feminino 1 - Masculino
Tipo viatura	Tipo de viatura segura	0 - Restantes veículos 1 - Motas
Dia	Dia da semana no qual ocorre o acidente	1 - Domingo 2 - 2ª feira 3 - 3ª feira 4 - 4ª feira 5 - 5ª feira 6 - 6ª feira 7 - Sábado
Hora	Período do dia em que ocorre o acidente	1 - [00h, 06h[2 - [06h, 12h[3 - [12h, 18h[4 - [18h, 24h[
Idade da viatura	Idade do veículo seguro	> =0
Regiao	Região do país onde habita o segurado	0 - Outras regiões 1- Região de Lisboa e Porto
Alcoolemia	Quando o segurado tem taxa de alcoolemia superior a 0,5 g/l no sangue durante o acto de condução	0 - TAS < 0,5 g/l 1 - TAS >= 0,5 g/l

Fonte: Elaboração própria

- Escolha da função de ligação.

  A função de ligação deverá ser escolhida *à priori* após análise extensiva aos dados e deverá ser compatível com a distribuição do erro proposto para esses mesmos dados (McCullagh & Nelder, 1989; Dobson, 1990). Assim, para cada distribuição há uma função de ligação específica, conforme tabela 5.4.

Tabela 5.4 - *Ligações canónicas para as distribuições da família exponencial*

Funções da familia exponencial	Função de ligação
Normal	$\eta = \mu$
Poisson	$\eta = \log \mu$
Binomial	$\eta = \log\{\pi/(1-\pi)\}$
Logística	$\eta = \log\{\pi/(1-\pi)\}$

Fonte: McCullagh & Nelder, 1989

5.7. Os Modelos na perspectiva do condutor

De acordo com as variáveis indicadas na Tabela 5.2. irá proceder-se à estimação dos modelos apresentados no capítulo da metodologia.

5.7.1. Estimação dos modelos

Esta etapa passa pela estimação dos parâmetros dos modelos, ou seja, pela estimação dos coeficientes β associados às variáveis e do parâmetro de dispersão ϕ, caso este esteja presente (Johnston & Dinardo, 1996 e Long, 1997), e pela avaliação da qualidade das referidas estimativas.

Salienta-se que o nível de significância de 0,05 é recomendado como o mínimo aceite (Hair et al., 1998), pelo que será esse valor utilizado para avaliação dos ajustes dos modelos.

Importa referir que o *software* R providencia uma implementação bastante flexível no âmbito dos modelo GLM através da função `glm( )` (Zeileis et al, 2008). Assim, o ajuste do modelo de regressão logística do programa R, versão 2.12.2, (que será designado de mod.M1), tem a seguinte sintaxe:

```
> glm (variável_resposta ~ variáveis_explicatorias, family=binomial, poisson)
```

5.7.1.1. Estimação do Modelo de Regressão Logístico

Para o modelo Logístico o *software R* produziu os resultados que constam da tabela 5.5:

Tabela 5.5 - *Resultados da Estimação do Modelo de Regressão Logística (mod.M1)*

Variàveis Explicativas	Estimativas dos coeficientes	Desvios-padrão	z-values	p-values
Constante	-0,950	0,088	-10,819	< 2e-16
Idade	-0,013	0,001	-14.915	< 2e-16
Género				
Homem	0,167	0,028	5.911	3.40e-09
Mulher	Classe de referência			
Tipo de Viatura				
LigeirosPassageiros	0,721	0,074	9.794	< 2e-16
LigeirosComerciais	0,861	0,080	10.759	< 2e-16
Motas	-0,767	0,104	-7.395	1.42e-13
Pesados	Classe de referência			
IdadeVeículo	-0,077	0,002	-38.316	< 2e-16
Região				
Regiao	0,147	0,028	5.238	1.63e-07
Restantes	Classe de referência			

Fonte: Elaboração própria

5.7.1.1.1. Análise da adequabilidade das estimativas.

As estimativas do modelo mod.M1 (Tabela 5.4) apresentam desvios padrão - que representam o nível de incerteza associado às estimativas (Gelman & Hill, 2007) -, com valores bastante reduzidos, tal como os níveis de significância (todos muito próximos do valor zero), pelo que os valores das estimativas produzidas pelo modelo mod.M1 revelam que os respectivos parâmetros são bastantes significativos.

Pode afirmar-se que as estimativas dos coeficientes são estatisticamente significativas.

Pela análise das estimativas dos coeficientes modelo mod.M1 conclui-se que a probabilidade para a ocorrência de acidente aumenta nos casos em que o segurado é homem ($\hat{\beta}_2$ = 0,167) com mais idade ($\hat{\beta}_1$ = -0.013) e conduz uma viatura ligeiro de passageiros ou comercial ($\hat{\beta}_3$ = 0,721; $\hat{\beta}_4$ = 0,861), com mais idade ($\hat{\beta}_6$ = -0,077), na zona da grande Lisboa ou do grande Porto ($\hat{\beta}_7$ =0,147).

5.7.1.1.2. Avaliação do ajustamento do Modelo Logístico

Após a estimação dos parâmetros do modelo pretende-se aferir sobre a eficácia dos resultados obtidos para se aferir da qualidade dos modelos estimados (Draper & Smith, 1998; Hair et al., 1998; Long, 1997; Hosmer & Lemeshow, 2000).

- Teste de significância dos coeficientes

De acordo com Agresti (1996), Hair et al. (2005), Long (1997), entre outros autores é sugerida a análise ao desempenho individual dos coeficientes estimados. Avalia-se a hipótese nula de que o parâmetro estimado é igual a zero, ou seja, $H_0 : \beta_j = 0$ vs $H_1 = \beta_j \neq 0$.

Caso H_0 seja rejeitada significa que a variável explicativa em questão não deve ser eliminada do modelo (Hosmer & Lemeshow, 2000; Long, 1997).

Para testar $H_0 : \hat{\beta}_1 = 0$ é utilizado z = -14,915 (*p-value* < 2e-16), conclui-se que a idade tem um impacto negativo significativo na ocorrência de acidente.

Para testar $H_0 : \hat{\beta}_2 = 0$ é utilizado z = 5,911 (*p-value* < 3.40e-09), também se conclui que o género é uma variável significativa que influência a probabilidade de ocorrência de acidente.

Para testar $H_0 : \hat{\beta}_3 = 0$ é utilizado z = 9,794 (*p-value* < 2e-16), permitindo constatar que os veículos ligeiros de passageiros é uma variável com impacto na ocorrência de acidentes, situação idêntica se passa com os veículos ligeiros comerciais, uma vez que ao testar $H_0 : \hat{\beta}_4 = 0$ se utiliza z = 10,759 (*p-value* < 2e-16).

Para testar $H_0 : \hat{\beta}_5 = 0$ é utilizado z = -7,395 (*p-value*=1.42e-13), o modelo sugere que a condução de motas tem impacto negativo significativo na probabilidade de ocorrência de acidente.

$H_0 : \hat{\beta}_6 = 0$ é utilizado z = -38,316 (*p-value* < 2e-16), sugere que a idade do veículo tem igualmente impacto significativo na probabilidade de ocorrência de acidente e o mesmo sucede para a variável Região ao testar $H_0 : \hat{\beta}_7 = 0$ é utilizado z = 5,238 (*p-value*=1.63e-07).

- Teste de significância do modelo

Uma importante medida estatística de ajuste geral de modelos de regressão é a estatística qui-quadrado (Agresti, 1996; Aldrich & Nelson 1975; Balakrishnan, 1992; Cameron & Trivedi, 1998; Hosmer & Lemeshow, 2000; Maddala, 1983; McCullagh & Nelder), a qual permite avaliar a significância das variáveis explicativas como um todo incluídas no modelo, tal avaliação efectua-se através da realização do teste χ^2, sob a hipótese nula de que todos os coeficientes são iguais a zero, $H_0 : \beta_0 = \beta_1 = ... = \beta_k = 0$ vs $H_1 = \beta_j \neq 0$ para algum j, $j = 1,...k$

Na prática são comparados dois modelos, é estimado um modelo reduzido que contém apenas o termo da constante, e é comparado com o modelo completo (mod.M1), e é feito o teste *likelihood ratio* que consiste na comparação estatística dos modelos.

Tabela 5.6 – *Análise de variância entre o modelo completo mod.M1 e o seu modelo reduzido*

	Residual DF	Residual Dev	DF	Desvio	P(>|Chi|)
Modelo Reduzido	49999	45983	-	-	-
Modelo Completo	49992	42959	7	3024,3	< 2,2e-16

Fonte: Elaboração própria

O teste estatístico *likelihood ratio* é χ^2 = 3024,3 (*p-value* = 2.2e-16). Conclui-se existir forte evidência contra H_0, ou seja, o modelo mod.M1 é fortemente significativo.

- O *Odds ratio*

Tabela 5.7 – *Odds ratio e intervalo de confiança a 95%*

Variàveis (X)	*Odds Ratio*	2,50%	97,50%
Idade	0,987	0,985	0,989
Homem	1,182	1,119	1,249
LigeirosPassageiros	2,056	1,780	2,375
LigeirosComerciais	2,364	2,021	2,766
Motas	0,464	0,379	0,569
IdadeVeículo	0,926	0,921	0,929
Região	1,158	1,096	1,223

Fonte: Elaboração própria

Verifica-se que, pelo *Odds ratio*, por exemplo num segurado homem é de 1,182, ou seja a probabilidade de ocorrência de acidente é maior que a probabilidade de não ocorrência de acidente. Para um intervalo de confiança a 95% essa probabilidade situa-se no intervalo [1,119; 1,249].

De um modo geral, para um nível de confiança de 95% o *Odds ratio* das estimativas das variáveis, os resultados são bastante precisos.

5.7.1.2. Estimação do Modelo de Regressão de Poisson

Para o modelo de Poisson o *software R*, produziu os seguintes resultados:

Tabela 5.8 - *Resultados da Estimação do Modelo de Regressão de Poisson (mod.M2)*

Variàveis Explicativas	Estimativas dos coeficientes	Desvios-padrão	z-values	p-values
Constante	-1,155	0,0749	-15,419	< 2e-16
Idade	-0,010	0,001	-13,727	< 2e-16
Género				
Homem	0,126	0,023	5,504	3.70e-08
Mulher	Classe de referência			
Tipo de Viatura				
LigeirosPassageiros	0,632	0,064	9,867	< 2e-16
LigeirosComerciais	0,748	0,069	10,848	< 2e-16
Motas	-0,712	0,093	-7,693	1.44e-14
Pesados	Classe de referência			
IdadeVeículo	-0,663	0,002	-39,978	< 2e-16
Região				
Regiao	0,116	0,022	5,169	2.36-07
Restantes	Classe de referência			

Fonte: Elaboração própria

5.7.1.2.1. Análise da adequabilidade das estimativas

A estimativa do coeficiente $\hat{\beta}_1$ indica o impacto sobre o número esperado de acidentes da idade do segurado, sendo $e^{-0,01} = 0,99$. Verifica-se que à medida que a variável idade aumenta o número de acidentes registados, em média, também aumenta.

A estimativa do coeficiente $\hat{\beta}_2$ indica o impacto sobre o número esperado de acidentes pelo facto do segurado ser homem. Desta forma $e^{0,126}$ = 1,134, constatando-se, que esta variável contribui mais para o número esperado de acidentes registados do que os casos em que o segurado é mulher.

As estimativas dos coeficiente $\hat{\beta}_3$ e $\hat{\beta}_4$ apontam o impacto sobre a diferença no número esperado de acidentes do facto do veículo seguro ser um ligeiro de passageiros ou um ligeiro comercial, respectivamente. Em ambas as situações, está-se perante um aumento considerável quando em causa está um dos tipos de viatura indicado, uma vez que $e^{0,632}$ = 1,881 e $e^{0,748}$ = 2,113, respectivamente

Contrariamente aos dois tipos de viaturas acima indicados, apesar da estimativa do coeficiente $\hat{\beta}_5$ indicar o impacto sobre o número esperado de acidentes se a viatura segura é mota, não se está perante um aumento do número de acidentes quando o veículo seguro é uma mota dado o coeficiente correspondente a esta variável é estimado com valor negativo.

A estimativa negativa do coeficiente $\hat{\beta}_6$ demonstra que o aumento da idade da viatura contribui para um decréscimo do número esperado de acidentes.

A estimativa do coeficiente $\hat{\beta}_7$ representa a diferença no número de acidentes perante a região de residência do segurado, está-se perante um efeito multiplicativo moderado quando o veículo em questão pertence às duas maiores cidades portuguesas, traduzindo num aumento de acidentes médio de $e^{0,116}$ = 1,123.

Verifica-se que os valores dos desvios-padrão das estimativas dos coeficientes são todos muito próximo do valor zero, tal como acontece com os valores dos *p-values*. Assim o nível de incerteza associado às estimativas é muito reduzido. Pode afirmar-se que as estimativas dos coeficientes do mod.M2 são estatisticamente significativas.

5.7.1.2.2. Avaliação do ajustamento do Modelo de Poisson

Tal como no modelo de Logístico, também no modelo de Poisson se pretende aferir sobre a eficácia dos resultados para se aferir da qualidade do modelo estimado, será igualmente utilizado o teste de significância dos coeficientes.

- Teste de significância dos coeficientes

De acordo com Hair et al. (2005) entre outros autores, é sugerida a análise à do desempenho individual dos parâmetros estimados. Avalia-se a hipótese nula de que o parâmetro estimado é igual a zero, ou seja, $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$. Caso H_0 seja rejeitada significa que a variável explicativa em questão não pode ser eliminada do modelo (Long, 1997). Para testar $H_0 : \hat{\beta}_1 = 0$ é utilizado z = -13,727 (*p-value* < 2e-16), conclui-se que a idade do segurado tem um impacto negativo muito forte no número de acidentes que esse segurado poderá ter.

Para testar $H_0 : \hat{\beta}_2 = 0$ é utilizado z = 5,504 (*p-value*=3.70e-08), verifica-se que o género tem uma influência negativa e significativa que a idade na contribuição para o número de acidentes.

Relativamente às restantes variáveis, para testar $H_0 : \hat{\beta}_3 = 0$, $H_0 : \hat{\beta}_4 = 0$, $H_0 : \hat{\beta}_5 = 0$, $H_0 : \hat{\beta}_6 = 0$, $H_0 : \hat{\beta}_7 = 0$, são utilizado, respectivamente z = 9,867, z = 10,848, z = -7,693, z = -39,978 e z = 5,169, todos com *p-value* próximos do valor zero, podendo-se concluir que todas as variáveis contribuem significativamente para o modelo mod.M2.

- Teste de significância do modelo

O ajustamento geral de modelos de regressão e a estatística qui-quadrado são igualmente aplicáveis ao modelo de Poisson (Cameron & Trivedi, 1998; Long, 1997, Maddala, 1983; McCullagh & Nelder, 1989). Tal avaliação efectua-se através da realização do teste χ^2, sob a hipótese nula de que todos os coeficientes são iguais a zero, $H_0 : \beta_0 = \beta_1 = ... = \beta_k = 0$ vs $H_1 = \beta_j \neq 0$ para algum *j*, $j = 1,...k$

Tal como na regressão logística é feito o teste *likelihood ratio*.

Tabela 5.9 – *Análise de variância entre o modelo completo mod.M2 e o seu modelo reduzido*

	Residual DF	Residual Dev	DF	Desvio	P(>|Chi|)
Modelo Reduzido	49999	37089	-	-	-
Modelo Completo	49992	33923	7	3166,2	< 2,2e-16

Fonte: Elaboração própria

O teste estatístico *likelihood ratio* é χ^2 = 3166,20 com *p-value* < 2,2e-16, havendo forte evidência contra H_0, logo o modelo mod.M2 é globalmente significativo.

- Teste de sobredispersão

Conforme foi referido no capítulo da metodologia, nos modelos de contagem um dos aspectos que requer maior atenção é a situação de sobredispersão. Face ao exposto, há interesse em testar a hipótese de ausência de sobredispersão (Long, 1997), testando $H_0 : \theta = 0$ contra $H_0 : \theta > 0$, onde a estatística de teste é dada por: $Z = \frac{\hat{\theta}}{S_{\hat{\theta}}} \approx N(0,1)$.

Desta forma, utilizado o *output* do mod.M3 tem-se:

Tabela 5.10 – *Análise da sobre-dispersão do modelo mod.M2*

Parâmetros	Valores
Estimativa de teta	2,619
Variância da estimativa de teta	0,242
Estatistica de teste (*z*)	5,324

Fonte: Elaboração própria

Considerando o valor da estatística de teste (*z*) de 5,324, esta aponta para a rejeição de H_0, logo há indícios da presença de sobredispersão que o modelo Binomial Negativo poderá, eventualmente, resolver (Hibe, 2007).

5.7.1.3. Estimação dos parâmetros do Modelo de Binomial Negativo

Conforme já foi apresentado no capítulo da Metodologia, o modelo Binomial Negativo é a alternativa mais utilizada ao modelo de Poisson para controlar não só o problema da sobredispersão, mas também permite melhorar a modelação em situações de contagem (Hilbe, 2007 e Lawless, 1987).

Para o modelo de Binomial Negativo o *software R*, produziu os resultados que constam da tabela 5.11.

Tabela 5.11 - *Resultados da Estimação do Modelo de Regressão Binomial Negativa (mod.M3)*

Variàveis Explicativas	Estimativas dos coeficientes	Desvios-padrão	z-values	p-values
Constante	-1,149	0,077	-14,893	< 2e-16
Idade	-0,010	0,008	-13,213	< 2e-16
Género				
Homem	0,128	0,024	5,306	1.12e-07
Mulher	Classe de referência			
Tipo de Viatura				
LigeirosPassageiros	0,629	0,066	9,592	< 2e-16
LigeirosComerciais	0,743	0,071	10,491	< 2e-16
Motas	-0,721	0,094	-7,642	2.13e-14
Pesados	Classe de referência			
IdadeVeículo	-0,066	0,002	-38,416	< 2e-16
Região				
Regiao	0,117	0,023	4,972	6.63e-07
Restantes	Classe de referência			

Fonte: Elaboração própria

5.7.1.3.1. Análise da adequabilidade das estimativas

À semelhança dos restantes modelos também se pretende aferir sobre a eficácia dos resultados obtidos para se aferir a qualidade do modelo estimados. Verifica-se que as estimativas dos coeficientes apresentam valores muito similares ao modelo de Poisson, quer em termos de valor como de tendência, pelo que a interpretação a ser feita é praticamente igual à efectuada no ponto 5.7.1.2.1. Os valores dos desvios-padrão das estimativas dos coeficientes são igualmente todos muito próximos de zero, tal como acontece os valores de todos os *p-values.* Nestes termos, pode afirmar-se que as estimativas dos coeficientes do mod.M3 são estatisticamente significativas.

5.7.1.3.2. Avaliação do ajustamento do modelo Binomial Negativo

À semelhança da regressão logística e de Poisson, também no modelo Binomial Negativo se pretende aferir sobre a eficácia dos resultados obtidos para aferir da qualidade do modelo estimado (Long, 1997; McCullagh & Nelder, 1989).

- Teste de significância dos coeficientes

A análise do desempenho individual dos parâmetros estimados é novamente efectuada. Avalia-se a hipótese nula de que o parâmetro estimado é igual a zero, ou seja, $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$. Caso H_0 seja rejeitada significa que a variável explicativa em questão não deve ser eliminada do modelo (Long, 1997).

Para testar $H_0 : \hat{\beta}_1 = 0$, é utilizado z = -13,213 (*p-value* < 2e-16); para testar $H_0 : \hat{\beta}_3 = 0$, é utilizado z = 9,592 (*p-value* < 2e-16); para $H_0 : \hat{\beta}_4 = 0$, é utilizado z = 10,491; para $H_0 : \hat{\beta}_6 = 0$, é utilizado z = 38,716 (*p-value* < 2e-16). Conclui-se que a idade do segurado, as viaturas ligeiros de passageiros e comerciais e a idade do veículo, respectivamente, têm um impacto muito significativo no número médio de acidentes que esse segurado poderá ter, à semelhança do que sucede com o modelo de Poisson.

Para testar $H_0 : \hat{\beta}_2 = 0$ é utilizado z = 5,306 (*p-value*=1,12e-07), verificando-se que o género (o facto do condutor ser homem) tem contribuição positiva e significativa para o número médio de acidentes.

Relativamente às restantes variáveis, para testar $H_0 : \hat{\beta}_2 = 0$, $H_0 : \hat{\beta}_5 = 0$, $H_0 : \hat{\beta}_7 = 0$, são utilizado, respectivamente z = 5,306, z = -7,642, z = -7,693, z = 4,972, verifica-se que as variáveis correspondentes são também significativas.

De salientar os valores de todos os *p-value* são próximos de zero, levando a concluir que todas as variáveis contribuem significativamente para o modelo mod.M3.

- Teste de significância do modelo

O ajuste geral de modelos de regressão e a estatística qui-quadrado são igualmente aplicáveis ao modelo ao modelo Binomial Negativo (Cameron & Trivedi, 1998; Long, 1997, Maddala, 1983; McCullagh & Nelder, 1989), através da realização do teste χ^2 , sob a hipótese nula de que todos os coeficientes são iguais a zero, $H_0 : \beta_0 = \beta_1 = ... = \beta_k = 0$ vs $H_1 = \beta_j \neq 0$ para algum *j*, $j = 1,...k$

Tal como nos anteriores modelos, também neste modelo será efectuando-se o teste *likelihood ratio*.

Tabela 5.12 – *Análise de variância entre o modelo completo mod.M2 e o seu modelo reduzido*

	Theta Res.	DF	2 x log-lik.	DF	Desvio	Pr(Chi)
Modelo Reduzido	1,278	49999	-54778,92	-	-	-
Modelo Completo	2,619	49992	-51920,66	7	2858,265	0

Fonte: Elaboração própria

Verifica-se que o teste estatístico *likelihood ratio* é χ^2 = 2858,265 com *p-value* = 0, evidência a rejeição de H_0, pelo que se conclui que o modelo mod.M3 é globalmente significativo.

5.8. O Modelo na perspectiva do acidente

Na perspectiva do acidente, procede-se à avaliação do risco/propensão para o acidente de viação utilizando as variáveis da estimação na perspectiva do condutor, mas acrescidas de novas variáveis explicativas relativas ao acidente em si.

Para esse efeito, e conforme foi atrás referido, da amostra de dimensão de 50.000 registos procedeu-se à extracção de um *data set* correspondente a todas as apólices com registo de acidentes, o correspondente a 8.625 registos, com base no qual se procedeu à estimação deste modelo.

Após várias experiências de modelação verificou-se que os resultados obtidos para este modelo considerando somente variáveis explicativas simples não se revelaram satisfatórios, nomeadamente:

- As variáveis binárias "Ligeirospassageiros" e "Ligeiroscomerciais" apresentaram estimativas não significativas (valores de *p-value* a exceder 10%);
- A variável "Idade" e a variável binária "Homem" também apresentaram níveis de significâncias muito reduzidas.
- As significâncias da maioria das variáveis revelam-se reduzidas.

De acordo com Gelman & Hill (2007) e Hosmer & Lemeshow (2000) é importante procurar nas variáveis explicativas comportamentos que possam ter de ser ajustados com recurso a:

- Variáveis artificiais, ou seja, variáveis criadas a partir das variáveis já existentes. Assim, com o objectivo de encontrar eventuais efeitos não lineares nas variáveis numéricas o modelo foi reestimado utilizando duas novas variáveis: a "Idade" ao quadrado e a "Idadeviatura" ao quadrado. Das experimentações efectuadas resultou que só a "Idadequadrado" se revelou significativa, pelo que a variável "Idadeviaturaquadrado" foi eliminada do modelo.
- Variáveis compostas, as quais são criadas a partir da relação entre variáveis já existentes no modelo, no estudo em concreto através do produto de duas variáveis explicativas. Assim, através de algumas combinações de variáveis (e.g. "homem_mota", "homem_região", *etc.*) reestimou-se o modelo. Considerando que nenhuma das variáveis compostas se mostrou significativa e, em alguns casos, diminuiu a significância das variáveis simples que serviram à sua criação, foram eliminadas do modelo por serem irrelevantes.

De salientar que a variável "Tipo de Viaturas" revelou-se não significativa para as categorias "ligeiros de passageiros" e "ligeiros comerciais", pelo que se experimentou efectuar uma alteração nesta variável. Assim, das quatro classes inicialmente estabelecidas e utilizadas na estimação nos modelos na perspectiva do condutor, passou-se para apenas dois níveis, conforme indicado na tabela 5.13, tendo a respectiva transformação melhorado os níveis de significância de algumas variáveis simples.

Tabela 5.13 - *Variáveis binárias para a variável "Motas"*

Variàvel: Motas	**Valores possíveis**
Tipo de viatura Motas	1
Outros tipos de viaturas	0

Fonte: Elaboração própria

5.8.1. Estimação do Modelo de Regressão Logístico

No seguimento do que vem sendo exposto, e após experimentação incluindo e retirando variáveis explicativas candidatas do modelo, os melhores resultados para o

modelo Logístico na perspectiva do acidente gerados pelo *software R*, são os que constam da tabela 5.14.

Tabela 5.14 - *Resultados da Estimação do modelo de Regressão Logística (mod.M4)*

Variàveis Explicativas	Estimativas dos coeficientes	Desvios-padrão	z-values	p-values
Constante	-1,053	0,360	-2,967	0,003
Idade	-0,029	0,014	-2,125	0,003
IdadeQuadrado	0,0003	0,0001	2,205	0,027
Genéro	Classe de referência			
Homem	-0,032	0,087	-0,369	0,712
Mulher	Classe de referência			
Tipo de Viatura				
Motas	1,243	0,171	7,256	4.00e-13
Outras	Classe de referência			
IdadeVeículo	0,011	0,006	1,863	0,062
Regiao				
Regiao	-0,286	0,089	-3,211	0,001
Restantes	Classe de referência			
Alcoolemia				
C/alcoolemia	2,593	0,455	5,701	1.19e-08
S/alcoolemia	Classe de referência			
Dia				
SegundaFeira	-0,431	0,132	-3,261	0,001
TerçaFeira	-0,641	0,145	-4,425	9.65e-06
QuartaFeira	-0,354	0,136	-2,598	0,009
QuintaFeira	-0,366	0,137	-2,677	0,007
SextaFeira	-0,432	0,135	-3,197	0,001
Sábado	-0,305	0,142	-2,153	0,031
Domingo	Classe de referência			
Hora				
Zero_seis	Classe de referência			
Seis_doze	-0,619	0,123	-5,026	5.00e-07
Doze_dezoito	-0,244	0,117	-2,082	0,037
Dezoito_vintequatro	0,069	0,126	0,543	0,587

Fonte: Elaboração própria

5.8.1.1. Análise da adequabilidade das estimativas.

De um modo geral, as estimativas produzidas pelo modelo mod.M4 são estatisticamente significativas, pois apresentam valores de desvios padrão (níveis de incerteza) e *p-values* reduzidos.

A variável artificial "IdadeQuadrado" apresenta a forma de uma parábola com a concavidade virada para cima, uma vez que o sinal da estimativa do coeficiente β_1 tem sinal positivo. Tal significa que os segurados mais jovens e os mais velhos tendem a ter

acidentes com menor gravidade, ao contrário dos condutores de "meia-idade" que propendem para acidentes mais graves.

Pela análise das estimativas dos coeficientes modelo mod.M4 conclui-se que a probabilidade para a ocorrência de acidente grave aumenta nas situações em que o segurado é de "meia idade" ($\hat{\beta}_1$ = -1,053 e β_2 = -0.029) conduz uma mota ($\hat{\beta}_4$ = 1,243), a viatura é recente ($\hat{\beta}_5$ = 0,011), na região da grande Lisboa ou do grande Porto ($\hat{\beta}_6$ =-0,286), ao fim-de-semana ($\hat{\beta}_{13}$ = -0,305), nos períodos de tempo entre as 18h00 e as 24h00 ($\hat{\beta}_{16}$ = 0,069) e tem alcoolemia ($\hat{\beta}_7$ = 2,593).

5.8.1.2. Avaliação do ajustamento do Modelo Logístico

Após a estimação dos parâmetros das estimativas do modelo pretende-se aferir sobre a eficácia dos resultados obtidos.

- Teste de significância dos coeficientes

Procedendo-se à avaliação da hipótese nula de que o parâmetro estimado é igual a zero, ou seja, $H_0 : \beta_j = 0$ vs $H_1 = \beta_j \neq 0$, caso H_0 seja rejeitada significa que a variável explicativa em questão não deve ser eliminada do modelo (Hosmer & Lemeshow, 2000; Long, 1997).

Para testar as estimativas dos parâmetros mais significativos tem-se $H_0 : \hat{\beta}_4 = 0$, $H_0 : \hat{\beta}_7 = 0$, $H_0 : \hat{\beta}_9 = 0$, $H_0 : \hat{\beta}_{14} = 0$, todos com valores de *p-value* muito próximos de zero, são utilizados z_4 = 7,256, z_7 = 5,701, z_9 = -4,425 e z_{14} = -5,026 conclui-se que as variáveis "Motas", "Alcoolemia", "TerçaFeira" e o período "Seis_ doze" contribuem de forma muito significativa para a gravidade dos acidentes.

- Teste de significância do modelo

Tal como nos modelos anteriores a estatística de ajuste geral de modelos de regressão é a estatística qui-quadrado onde são testadas as hipóteses, $H_0 : \beta_0 = \beta_1 = ... = \beta_k = 0$ vs $H_1 = \beta_j \neq 0$ para algum j, $j = 1,...k$

Será estimado o modelo reduzido que contém apenas o termo do intercepto na origem, o qual é comparado com o modelo completo (mod.M4), e é feito o teste *likelihood ratio*.

Tabela 5.15 – *Análise de variância entre o modelo completo mod.M4 e o seu modelo reduzido*

	Residual DF	Residual Dev	DF	Desvio	P(>|Chi|)
Modelo Reduzido	8608	5234,9	-	-	-
Modelo Completo	8624	5414,0	-16	-179,16	< 2,2e-16

Fonte: Elaboração própria

O teste estatístico *likelihood ratio* é χ^2 = -179,16 com *p-value* = 2.2e-16 conclui pela existência de evidência contra H_0, pelo que o modelo mod.M4 é globalmente significativo.

- O *Odds ratio*

Tabela 5.15 – *Odds ratio e intervalo de confiança a 95%*

Variàveis (X)	*Odds Ratio*	2,50%	97,50%
Idade	0,971	0,945	0,998
IdadeQuadrado	1,000	1,000	1,001
Homem	0,969	0,817	1,148
Motas	3,466	2,478	4,849
IdadeVeículo	1,011	0,999	1,024
Regiao	0,752	0,631	0,895
Alcoolemia	13,376	5,484	32,625
SegundaFeira	0,650	0,501	0,842
TerçaFeira	0,527	0,396	0,700
QuartaFeira	0,702	0,537	0,917
QuintaFeira	0,693	0,531	0,907
SextaFeira	0,649	0,498	0,846
Sábado	0,737	0,558	0,973
Seis_doze	0,538	0,428	0,685
Doze_dezoito	0,784	0,623	0,986
Dezoito_vintequatro	1,071	0,836	1,371

Fonte: Elaboração própria

Para um nível de confiança de 95% do *Odds ratio* das estimativas das variáveis, os resultados do *Odds ratio* são bastante precisos, com excepção do que acontece com a variável “alcoolemia”. De salientar que o resultado da variável “alcoolemia” revela que na presença da mesma, a probabilidade de acidente aumenta substancialmente.

De acordo com Gelmann & Hill (2007), o modelo logístico tende a funcionar menos bem para os casos de modelos com muitos preditores. De facto, o modelo logístico na perspectiva do acidente tem mais de o dobro das variáveis explicativas utilizadas na estimação na perspectiva do condutor.

No entanto, perante todos os testes estatísticos acima efectuados conclui-se que o mod.M4 é estatisticamente significativo para a avaliação do risco/propensão na gravidade dos acidentes.

5.9. Validação dos modelos

A validação dos modelos consiste em avaliar as suas capacidades de previsão, em atribuir probabilidades a valores não observados. Desta forma torna-se possível ver o funcionamento do modelo e generalizar os resultados obtidos a toda a população (Gelman & Hill, 2007, Hair et al., 1998, McCullagh & Nelder, 1989, Murteira et al., 2010).

Assim, utilizando valores da base de dados fornecida pela Seguradora e que não constam na amostra de dimensão de 50.000 registos utilizada para estimação do modelos, através da função `predict.glm` do *software R-Project* estimaram-se as probabilidades para as previsões e construíram-se intervalos de confiança para as respectivas previsões, cujos *outputs* do *R-Project* constam do anexo B.

Para o mod.M1 considerou-se que uma mulher de 62 anos, com um veículo ligeiro de passageiros de 15 anos e que não resida na região de Lisboa nem do Porto. A previsão de ocorrência de acidente para esta observação é de 10,01%. Para um intervalo de confiança a 95% essa probabilidade situa-se no intervalo [9,44%, 10,61%], um intervalo com uma precisão que pode ser considerada assinalável.

No mod.M2 considerou-se a observação de um homem de 20 anos com um ligeiro comercial com 1 ano, na região de Lisboa. A previsão do número médio de acidentes para esta observação é de 65,27%. Tendo em conta o intervalo de confiança a 95%, essa probabilidade situa-se no intervalo [37,57%, 41,45%].

Para o mod.M3, seleccionou-se a observação de uma mulher de 44 anos que conduz uma mota com menos de 1 ano na região de Lisboa. A previsão obtida para o

aumento do número de acidentes é de 11,24%. Considerando o conta o intervalo de confiança a 95%, essa probabilidade situa-se no intervalo [8,84%, 11,52%].

Para o mod.M4, teve-se em conta um homem de 31 anos que conduzia uma mota de 3 anos, na região do Porto, com acidente numa terça-feira, no período entre as 06h00 e as 12h00, e que não acusou taxa de alcoolemia positiva. A previsão obtida para o acidente ter consequências graves é de 12,40%. Considerando o conta o intervalo de confiança a 95% essa probabilidade situa-se no intervalo [8,38%, 17,96%], com uma precisão já mais baixa do que os anteriores.

Face ao exposto, e tendo em conta que para os mod.M1, mod.M3 e mod.M4 os intervalos de confiança a 95% confirmam os valores das previsões pontuais, perante tais resultados pode concluir-se pela aparente validade dos referidos modelos, no contexto dos casos apresentados.

6. CONCLUSÕES

Neste capítulo serão apresentadas as conclusões do trabalho desenvolvido ao longo da dissertação, bem com uma síntese dos principais resultados práticos.

6.1. Conclusões Gerais

Nesta dissertação foram apresentados os principais modelos de regressão para a avaliação do risco no âmbito dos acidentes automóvel. De acordo com a metodologia apresentada, privilegiou-se a utilização dos modelos lineares generalizados. Numa primeira fase, foi dada especial relevância ao modelo de regressão Logística na perspectiva do condutor (mod.M1) que permitiu avaliar a probabilidade de um indivíduo, mediante o conhecimento de determinadas características *à priori*, ter, ou não ter acidente, e qual a propensão para o acidente. Foi igualmente dada relevância ao modelo de contagem Binomial Negativo (mod.M3), que permite determinar o impacto de variáveis explicativas do condutor sobre do número médio de acidentes que esse mesmo indivíduo venha ter.

Relativamente aos modelos de contagem, o modelo de Poisson (mod.M2) revelou fraco desempenho, concluindo-se que raramente explica os dados devido a várias restrições importantes, sobretudo devido à sobredispersão, pelo que apesar de teoricamente ser um importante candidato para à avaliação da propensão para o acidente, a sua aplicação, não se mostrou adequado ao presente estudo.

No entanto, apesar do risco/propensão para o acidente poder ser estimado com base nas características individuais observadas *à priori*, pode também ser avaliado com base em características externas aos indivíduos. Assim, numa segunda fase procedeu-se à estimação de um novo modelo Logístico na perspectiva do acidente (mod.M4), no qual se incluíram novas variáveis associadas à sinistralidade com vista à modelação avaliação da gravidade dos acidentes.

Com base no desenvolvimento das metodologias apresentadas e nos resultados obtidos, obtiveram-se como principais conclusões:

- Foi possível criar um modelo de avaliação de risco/propensão para o acidente automóvel, conforme proposto no início desta dissertação.

- O modelo avaliação de risco/propensão para o acidente automóvel, no seu todo, é o resultado da junção de três partes (três modelos econométricos): o mod.M1, mod.M3 e o mod.M4.
- Os modelos mod.M1, mod.M3 e mod.M4 revelam-se, significativos, de fácil controlo e compreensão, capazes de produzir estimativas satisfatórias e de, em conjunto, permitir avaliar com robustez o risco/propensão para o acidente de um dado individuo.

6.2. Resultados Práticos

Os resultados práticos mais relevantes da aplicação dos modelos mod.M1, mod.M3 e mod.M4, e que constam do capítulo anterior, são os seguintes:

- Na perspectiva do indivíduo, concluí-se, através do modelo mod.M1, que a probabilidade de acidente aumenta quando o acto de condução é praticado por um homem, com mais idade, ao volante de um veículo ligeiro de passageiros ou comercial, também com alguma idade, na região da grande Lisboa ou do grande Porto.

  Pela aplicação do modelo mod.M3 concluiu-se que a idade do indivíduo, a idade da viatura, os veículos ligeiros de passageiros e comerciais são variáveis com impacto muito significativo no número médio de acidentes que um indivíduo poderá ter.
- Na perspectiva do acidente, modelo mod.M4, conclui-se que a probabilidade para a ocorrência de acidente grave aumenta nas situações em que um indivíduo, de “meia-idade”, conduz um motociclo recente, sob influência do álcool, na zona da grande Lisboa ou do grande Porto, entre as 18h00 e as 24h00.

Face ao exposto, identificados os indivíduos e as situações que comportam maior nível de risco, segundo os modelos apresentados, deverá uma companhia de seguros criar mecanismos que permitam a sua correcta aceitação e tarifação. Deverá, também, desenvolver instrumentos de agregação e partilha de risco, com vista à sua minimização, e constituir provisões técnicas adequadas que visem suportar o pagamento das indemnizações em caso de sinistro.

7. LIMITAÇÕES E RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Neste capítulo destacam-se as principais vantagens e limitações deste trabalho e são propostos desenvolvimentos ao trabalho efectuado.

7.1. Vantagens e Limitações

No que respeita a vantagens resultantes dos modelos seleccionados:

- Os modelos mod.M1, mod.M2 e mod.M3 são complementares na avaliação do risco, abrangendo praticamente todo o fenómeno.
- Os modelos são adaptáveis a alterações que possam surgir, com a facilidade de se retirarem e colocarem novas variáveis, e utilizarem outras amostras ou outros dados de outras origens.

Relativamente a limitações desta dissertação, importa referir que:

- Tecnicamente, não é possível colocar os três modelos escolhidos sob a forma de um único modelo, terão de funcionar separadamente;
- A base para este estudo assentou nos dados fornecidos pela seguradora, desconhecendo-se a quantidade e características dos acidentes que não passam pela seguradora, e se o seu peso poderia alterar as conclusões desta dissertação.

7.2. Propostas para trabalhos futuros

Nesta dissertação procurou-se alcançar uma vantagem competitiva para ser trabalhada e desenvolvida por uma seguradora na tomada de decisão de aceitação, ou não, dos seus riscos, no âmbito do seguro automóvel. Com vista a aperfeiçoar os resultados encontrados neste trabalho, delineiam-se como propostas para trabalhos futuros:

- A obtenção de mais informação por parte das seguradoras no momento da contratação do seguro, e.g. utilização preferencial da viatura (nos dias úteis, aos

fins-de-semana, em hora de ponta, de noite, em zonas urbanas, rurais, em trajectos curtos, em viagens longas, etc.), profissão/ocupação do segurado, multas/coimas de trânsito, entre outras. Quanto maior for a informação disponível, mais assertivos e precisos serão os resultados produzidos pelos modelos.

- A inclusão, criação, alteração, substituição de variáveis como forma de se atingirem resultados cada vez mais robustos e ajustados. Estes modelos não são estáticos, dependem de factores humanos e ambientais que estão em permanente evolução. É importante rever variáveis e fazer modelação experimental para detectar alterações nos níveis de risco.
- Um último ponto de investigação importante e interessante, seria a criação de um banco de dados, se possível em conjunto as autoridades de trânsito, com informações do histórico da sinistralidade dos segurados, o que iria permitir a criação de novos modelos mais precisos na avaliação do risco e propensão para o acidente automóvel.

ANEXOS

Anexo A – Informação dos vários campos que formam a base de dados fornecida pela Seguradora

CAMPOS DA BASE DE DADOS	DESCRIÇÃO
APOLICE	Nº de apólice
ID_SINISTRO	Nº de sinistro
ID_SUB_SINISTRO	Nº do sub-sinistro
DATA_ABERTURA	Data em que o processo de sinistro foi aberto na Companhia de Seguros
DATA_ENCERRADO	Data em que o processo foi encerrado na Companhia de Seguros (fim da regularização dos prejuízos).
SITUACAO_SINISTRO	Informa o estado do processo, o qual pode estar: - Em curso; - Fechado; - Pré-encerrado; - Reaberto; - Em reembolso.
DATA_SINISTRO	Data em que o acidente ocorreu.
HORA_SINISTRO	Hora em que o acidente ocorreu.
TIPO_SINISTRO	Indica a forma como o sinistro foi regularizado (para saber se regularizámos com o segurado e/ou com o terceiro ou com a congénere), existem 17 classificações de sub-sinistro: - CIDS Credor normal; - CIDS Devedor extra-convenção; - CIDS Devedor normal; - IDS Credor especial; - IDS Credor extra-convenção - IDS credor normal; - IDS Devedor especial; - IDS Devedor extra convenção; - IDS Devedor normal - IRT Especial; - IRT normal; - Normal; - Sinistro carta verde; - Sinistro geral; - Sinistro repetido; - Sinistro representadas; - Sinistro Vidros.

CONSEQUENCIA	Se os danos decorrentes do acidente são não corporais (viaturas, imóveis, rails, sinais de trânsito, animais), se são corporais (pessoas), ou em ambos. Desta forma os sinistros são classificados em: - Material; - Corporal; - ou Material/Corporal.
CONCELHO_SINISTRO	Concelho onde ocorre o acidente
ID_CODPOSTAL_SINISTRO	Código postal do local do acidente
COD_POSTAL_SINISTRO	Freguesia onde ocorre o acidente
RESP_SEGURADO	Indica a % de responsabilidade do segurado. Varia entre 0% e 100%. A tabela tem 15 escalões de responsabilidade do segurado: 0%, 20%, 25%, 30%, 35%, 40%, 50%, 60%, 65%, 70%, 75%, 80%, 95%, 99%, 100%.
OBJECT_REGUL	Indica se quem sofreu danos. Tem 3 categorias possíveis: - Uma pessoa; - Um objecto; - Ou uma viatura
CATEGORIASINISTRADO	Tem 3 classificações possíveis. Indica se a Companhia de Seguros regularizou os danos com: - O segurado; - O terceiro; - ou Congénere (outra Companhia de Seguros).

CAUSA_SIN	Causa do sinistro. São definidas 26 causas: - Atropelamento; - Abertura de porta; - Actos de vandalismo; - Avaria mecânica; - Carga transportada; - Circulação fora de mão; -Cruzamento de viaturas; - Curto-circuito; - Despiste; - Fenómenos da natureza; - Furto/roubo; - Ia estacionar - Iluminação deficiente; - Incêndio; - Inversão de marcha; - Marcha-atrás, - Mudança de direcção; - Mudança de fila; - Não guardou distância da viatura; - Objecto projectado; - Perda de prioridade; - Queda de passageiro; - Saída estacionamento; - Tentativa de furto; - Ultrapassagem; - Outras causas.
ID_SEGURADO	Nº de identificação do segurado
SEGURADO	Género do segurado (masculino ou feminino)
SEGURADO_DATA_NASC	Data de nascimento do segurado
SEGURADO_MORADA	Morada do segurado
ID_SEGURADO_CPOSTAL	Código postal da morada do segurado
SEGURADO_COD_POSTAL	Freguesia da morada do segurado
SEGURADO_CONCELHO	Concelho da morada do segurado
SEGURADO_PAIS	País do segurado
CONDUTOR_VEIC_SEGURO_SEXO	Género do condutor (masculino ou feminino
CONDUTOR_VEIC_SEGURO_DATA_NASC	Data de nascimento do condutor do veículo seguro
CONDUTOR_VEIC_SEGURO_MORADA	Morada do condutor do veículo seguro
ID_CONDUTOR_VEIC_SEGURO_CPOSTAL	Código postal da morada do condutor do veículo seguro
CONDUTOR_VEIC_SEGURO_COD_POSTAL	Freguesia da morada do condutor do veículo seguro
CONDUTOR_VEIC_SEGURO_CONCELHO	Concelho da morada do condutor do veículo seguro

CONDUTOR_VEIC_SEGURO_PAIS	País do condutor do veículo seguro
LESADO_SEXO	Género do condutor do sinistrado
LESADO_DATA_NASC	Data de nascimento do sinistrado
LESADO_MORADA	Morada do sinistrado
ID_LESADO_CPOSTAL	Código postal da morada do sinistrado
LESADO_COD_POSTAL	Freguesias da morada do sinistrado
LESADO_CONCELHO	Concelho da morada do sinistrado
LESADO_PAIS	País do condutor do sinistrado
CONDUTOR_VEIC_LESADO_SEXO	Género do condutor do condutor do veículo terceiro (masculino ou feminino)
CONDUTOR_VEIC_LESADO_DATA_NASC	Data de nascimento do condutor do veículo terceiro
CONDUTOR_VEIC_LESADO_MORADA	Morada do condutor do veículo terceiro
ID_CONDUTOR_VEIC_LESADO_CPOSTAL	Código postal da morada do condutor do veículo terceiro
CONDUTOR_VEIC_LESADO_COD_POSTAL	Freguesias da morada do condutor do veículo terceiro
CONDUTOR_VEIC_LESADO_PAIS	País do condutor do veículo seguro
REEMBOLSO	Indica se no final da regularização, o processo tem, ou não, de ser reembolsados à Companhia de Seguro. A situação pode ser de 7 tipos: - A (anulado); - C (Contencioso); - F (Fechado); - N (sem motivo para reembolso); - P (Pendente); - R (Recusado); - S (Há motivo para reembolso).

MOTIVO_REEMB	Motivo pelo qual um processo segue para reembolso. Existem 18 motivo: - Abandono; - Alcoolemia; - Apólice anulada; - Carga mal acondicionada; - Condutor sem carta de condução; - Dolo; - Factura Hospitalar; - Fraude; - Franquia; - Franquia em RC; - Inspecção periódica; - Penalidade DL 83/2006 nº3 Artº2; - Reembolso cobertura CRT; - Responsabilidade terceiros; - Roubo; - Seguro de Garagista Artº 7 DL 291/2007; - Sem seguro.
CATEGORIA_SINISTRADO	É definida em função do local onde o sinistrado se encontrava quando sofreu o acidente, tem 10 classificações possíveis: - Passageiro de transporte público; - Peão; - Condutor veículo seguro em viatura de 2 rodas; - Condutor do veículo seguro noutro tipo de viatura; - Passageiro veículo seguro em viatura de 2 rodas; - Passageiro veículo seguro noutro tipo de viatura; - Condutor do veículo terceiro em viatura de 2 rodas; - Condutor do veículo terceiro noutro tipo de viatura; - Passageiro veículo terceiro em viatura de 2 rodas; - Passageiro veículo terceiro noutro tipo de viatura.

SITUAÇÃO_CLÍNICA	Tipo de consolidação das lesões do ferido após regularização. São classificadas em 9 tipos: - Alta por abandono; - Assistência Hospitalar; - Curado sem desvalorização; - Desconhecido; - Incapacidade permanente parcial; - Incapacidade Temporário Absoluta; - Incapacidade Temporária Parcial; - Morte; - Sem Incapacidade.
CATEGORIA_VIATURA	Tipo da viatura sinistrada, são agrupados em 12 tipos: - Ciclomotor; - Galera; - Jeep; - Ligeiro Comercial Deriv.Turismo; - Ligeiro comercial; - Ligeiro Passageiros; - Motociclo; - Pesado Especial; - Pesado de mercadorias; - Pesado Passageiros; - Táxi.
MARCA	Marca da viatura segura
ANO_CONSTRUÇÃO	Ano de construção da viatura segura
CILINDRADA	Cilindrada da viatura segura

Fonte: Elaboração própria

Anexo B – Sintaxe e Outputs do R-Project referente à estimação dos modelos

- Regressão Logística na perspectiva do condutor (mod.M1)

> Dados <-
read.table("c://Sandra/Novo_Modelo_Final/Tbl_Dados_Wrk_50000_Modelos.txt", header=T, sep=";")
> mod.M1 <- glm (Condutor_c_s_Acidentes ~ Idade + Homem + LigeiroPassageiros + LigeiroComercial + Motas + IdadeVeiculo + Regiao, family=binomial, data = Dados)
> summary(mod.M1)

```
Call:
glm(formula = Condutor_c_s_Acidentes ~ Idade + Homem + LigeiroPassageiros +
    LigeiroComercial + Motas + IdadeVeiculo + Regiao, family = binomial,
    data = Dados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1145  -0.6810  -0.5283  -0.3072   3.2077

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.9494352  0.0877529 -10.819  < 2e-16 ***
Idade              -0.0130412  0.0008743 -14.915  < 2e-16 ***
Homem               0.1673106  0.0283044   5.911 3.40e-09 ***
LigeiroPassageiros  0.7207479  0.0735923   9.794  < 2e-16 ***
LigeiroComercial    0.8605006  0.0799807  10.759  < 2e-16 ***
Motas              -0.7672160  0.1037539  -7.395 1.42e-13 ***
IdadeVeiculo       -0.0772720  0.0020167 -38.316  < 2e-16 ***
Regiao              0.1464502  0.0279615   5.238 1.63e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 45983  on 49999  degrees of freedom
Residual deviance: 42959  on 49992  degrees of freedom
AIC: 42975

Number of Fisher Scoring iterations: 5
```

> mod.M1Reduzido <- glm (Condutor_c_s_Acidentes ~ 1, family=binomial, data = Dados)
> summary(mod.M1Reduzido)

```
Call:
glm(formula = Condutor_c_s_Acidentes ~ 1, family = binomial,
    data = Dados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6154  -0.6154  -0.6154  -0.6154   1.8748

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.56801    0.01184  -132.5   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 45983  on 49999  degrees of freedom
Residual deviance: 45983  on 49999  degrees of freedom
AIC: 45985

Number of Fisher Scoring iterations: 3
```

> anova(mod.M1Reduzido,mod.M1, test = "Chisq")

```
Analysis of Deviance Table

Model 1: Condutor_c_s_Acidentes ~ 1
Model 2: Condutor_c_s_Acidentes ~ Idade + Homem + LigeiroPassageiros +
    LigeiroComercial + Motas + IdadeVeiculo + Regiao
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1     49999      45983
2     49992      42959  7   3024.3 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

> exp(coef(mod.M1))

```
      (Intercept)             Idade             Homem       LigeiroPassageiros
        0.3869595         0.9870434           1.1821213         2.0559702

 LigeiroComercial             Motas       IdadeVeiculo             Regiao
        2.3643440         0.4643039          0.9256380         1.1577173
```

> exp(confint.default(mod.M1))

```
                      2.5 %    97.5 %
(Intercept)        0.3258142 0.4595798
Idade              0.9853534 0.9887364
Homem              1.1183283 1.2495533
LigeiroPassageiros 1.7798154 2.3749731
LigeiroComercial   2.0213008 2.7656065
Motas              0.3788671 0.5690073
IdadeVeiculo       0.9219865 0.9293040
Regiao             1.0959775 1.2229350
```

> pi.hat = predict.glm(mod.M1,data.frame (Idade = 62 , Homem = 0 , LigeiroPassageiros = 1 , LigeiroComercial = 0, Motas = 0, IdadeVeiculo = 15, Regiao = 0), type = "response", se.fit = TRUE)
> pi.hat$fit

```
        1
0.1000811
```

> l.hat = predict.glm(mod.M1,data.frame (Idade = 62 , Homem = 0 , LigeiroPassageiros = 1 , LigeiroComercial = 0, Motas = 0, IdadeVeiculo = 15, Regiao = 0) ,se.fit = TRUE)
> ci = c(l.hat$fit - 1.96*l.hat$se.fit,l.hat$fit + 1.96*l.hat$se.fit)

> exp(ci)/(1+exp(ci))

```
        1         1
0.0943555 0.1061134
```

- Regressão de Poisson (mod.M2)

> mod.M2 <- glm (Condutor_n_Acidentes ~ Idade + Homem + LigeiroPassageiros + LigeiroComercial + Motas + IdadeVeiculo + Regiao, family= poisson, data = Dados)

> summary(mod.M2)

```
Call:
glm(formula = Condutor_n_Acidentes ~ Idade + Homem + LigeiroPassageiros +
    LigeiroComercial + Motas + IdadeVeiculo + Regiao, family = poisson,
    data = Dados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1256  -0.6913  -0.5469  -0.3317   4.3351

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.1551245  0.0749149 -15.419  < 2e-16 ***
Idade              -0.0097691  0.0007117 -13.727  < 2e-16 ***
Homem               0.1261087  0.0229106   5.504 3.70e-08 ***
LigeiroPassageiros  0.6324089  0.0640950   9.867  < 2e-16 ***
LigeiroComercial    0.7480858  0.0689580  10.848  < 2e-16 ***
Motas              -0.7122297  0.0925822  -7.693 1.44e-14 ***
IdadeVeiculo       -0.0663628  0.0016600 -39.978  < 2e-16 ***
Regiao              0.1159833  0.0224392   5.169 2.36e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 37089  on 49999  degrees of freedom
Residual deviance: 33923  on 49992  degrees of freedom
AIC: 52099

Number of Fisher Scoring iterations: 6
```

> mod.M2Reduzido <- glm (Condutor_n_Acidentes ~ 1, family= poisson, data = Dados)
> summary(mod.M2Reduzido)

```
Call:
glm(formula = Condutor_n_Acidentes ~ 1, family = poisson, data = Dados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6389  -0.6389  -0.6389  -0.6389   4.7322

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.589145   0.009899  -160.5   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 37089  on 49999  degrees of freedom
Residual deviance: 37089  on 49999  degrees of freedom
AIC: 55251

Number of Fisher Scoring iterations: 6
```

> anova(mod.M2Reduzido,mod.M2, test = "Chisq")

```
Analysis of Deviance Table

Model 1: Condutor_n_Acidentes ~ 1
Model 2: Condutor_n_Acidentes ~ Idade + Homem + LigeiroPassageiros + LigeiroComercial +
    Motas + IdadeVeiculo + Regiao
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1     49999      37089
2     49992      33923  7   3166.2 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

> exp(coef(mod.M2))

```
      (Intercept)             Idade             Homem     LigeiroPassageiros
        0.3150183         0.9902785         1.1344055          1.8821390

  LigeiroComercial             Motas       IdadeVeiculo             Regiao
        2.1129515         0.4905492         0.9357913          1.1229771
```

> exp(confint.default(mod.M2))

```
                      2.5 %    97.5 %
(Intercept)        0.2719995 0.3648409
Idade              0.9888981 0.9916608
Homem              1.0845929 1.1865059
LigeiroPassageiros 1.6599459 2.1340739
LigeiroComercial   1.8458327 2.4187262
Motas              0.4091443 0.5881508
IdadeVeiculo       0.9327517 0.9388409
Regiao             1.0746589 1.1734678
```

> pi.hat = predict.glm(mod.M2,data.frame (Idade = 20 , Homem = 1 , LigeiroPassageiros = 0 , LigeiroComercial = 1, Motas = 0, IdadeVeiculo = 1, Regiao = 1), type = "response", se.fit = TRUE)
> pi.hat$fit

```
        1
0.6526652
```

> l.hat = predict.glm(mod.M2, data.frame (Idade = 20 , Homem = 1 , LigeiroPassageiros = 0 , LigeiroComercial = 1, Motas = 0, IdadeVeiculo = 1, Regiao = 1) ,se.fit = TRUE)
> ci = c(l.hat$fit - 1.96*l.hat$se.fit,l.hat$fit + 1.96*l.hat$se.fit)
> exp(ci)/(1+exp(ci))

```
        1         1
0.3756543 0.4145112
```

- Regressão Binomial Negativa (mod.M3)

> mod.M3 <- zelig (Condutor_n_Acidentes ~ Idade + Homem + LigeiroPassageiros + LigeiroComercial + Motas + IdadeVeiculo + Regiao, model = "negbin", data = Dados)

```
How to cite this model in Zelig:
Kosuke Imai, Gary King, and Oliva Lau. 2007. "negbin: Negative Binomial Regression for
Event Count Dependent Variables" in Kosuke Imai, Gary King, and Olivia Lau, "Zelig:
Everyone's Statistical Software," http://gking.harvard.edu/zelig
```

> summary(mod.M3)

```
Call:
zelig(formula = Condutor_n_Acidentes ~ Idade + Homem + LigeiroPassageiros +
    LigeiroComercial + Motas + IdadeVeiculo + Regiao, model = "negbin",
    data = Dados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0671  -0.6764  -0.5394  -0.3299   3.8412

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.1492583  0.0771682 -14.893  < 2e-16 ***
Idade              -0.0098456  0.0007451 -13.213  < 2e-16 ***
Homem               0.1275535  0.0240412   5.306 1.12e-07 ***
LigeiroPassageiros  0.6289229  0.0655654   9.592  < 2e-16 ***
LigeiroComercial    0.7426851  0.0707946  10.491  < 2e-16 ***
Motas              -0.7207526  0.0943113  -7.642 2.13e-14 ***
IdadeVeiculo       -0.0663430  0.0017270 -38.416  < 2e-16 ***
Regiao              0.1173650  0.0236060   4.972 6.63e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(2.6191) family taken to be 1)

    Null deviance: 33385  on 49999  degrees of freedom
Residual deviance: 30434  on 49992  degrees of freedom
AIC: 51939

Number of Fisher Scoring iterations: 1


             Theta:  2.619
         Std. Err.:  0.242

 2 x log-likelihood:  -51920.656
```

> mod.M3Reduzido <- zelig (Condutor_n_Acidentes ~ 1, model = "negbin", data = Dados)

> summary(mod.M3Reduzido)

```
Call:
zelig(formula = Condutor_n_Acidentes ~ 1, model = "negbin", data = Dados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6154  -0.6154  -0.6154  -0.6154   3.7225

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.58915    0.01066  -149.1   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for Negative Binomial(1.2771) family taken to be 1)

    Null deviance: 30571  on 49999  degrees of freedom
Residual deviance: 30571  on 49999  degrees of freedom
AIC: 54783

Number of Fisher Scoring iterations: 1


              Theta:  1.2771
          Std. Err.:  0.0760

 2 x log-likelihood:  -54778.9210
```

> anova(mod.M3Reduzido,mod.M3, test = "Chisq")

```
Likelihood ratio tests of Negative Binomial Models
Response: Condutor_n_Acidentes

  Model           theta Resid.     df       2 x log-lik.  Test    df   LR stat.   Pr(Chi)
1   1              1.277135       49999     -54778.92
2 Idade … Regiao     2.619117      49992     -51920.66     1 vs 2   7   2858.265    0
```

> exp(coef(mod.M3))

```
      (Intercept)             Idade             Homem       LigeiroPassageiros
        0.3168717         0.9902027         1.1360456          1.8755893
  LigeiroComercial             Motas       IdadeVeiculo             Regiao
   2.1015708                 0.4863861         0.9358099          1.1245298
```

> exp(confint.default(mod.M3))

```
                       2.5 %    97.5 %
(Intercept)        0.2723942 0.3686117
Idade              0.9887577 0.9916499
Homem              1.0837568 1.1908573
LigeiroPassageiros 1.6494091 2.1327850
LigeiroComercial   1.8292943 2.4143735
Motas              0.4042995 0.5851390
IdadeVeiculo       0.9326477 0.9389828
Regiao             1.0736865 1.1777807
```

> pi.hat = predict.glm(mod.M3,data.frame (Idade = 44 , Homem = 0 , LigeiroPassageiros = 0 , LigeiroComercial = 0, Motas = 1, IdadeVeiculo = 0, Regiao = 1), type = "response", se.fit = TRUE)

> pi.hat$fit

```
        1
0.1123819
```

> l.hat = predict.glm(mod.M3, data.frame (Idade = 44 , Homem = 0 , LigeiroPassageiros = 0 , LigeiroComercial = 0, Motas = 1, IdadeVeiculo = 0, Regiao = 1) ,se.fit = TRUE)

> ci = c(l.hat$fit - 1.96*l.hat$se.fit,l.hat$fit + 1.96*l.hat$se.fit)

> exp(ci)/(1+exp(ci))

```
         1          1
0.08841155 0.11521763
```

- Regressão Logística na perspectiva do acidente (mod.M4)

> mod.M4<- glm (Gravidade ~ Idade + IdadeQuadrado + Homem + Motas + IdadeVeiculo + Regiao + alcoolemia + SegundaFeira + TercaFeira + QuartaFeira + QuintaFeira + SextaFeira + Sabado + seis_doze + doze_dezoito + dezoito_vintequatro , family=binomial, data = Dados)
> summary(mod.M4)

```
Call:
glm(formula = Gravidade ~ Idade + IdadeQuadrado + Homem + Motas +
    IdadeVeiculo + Regiao + alcoolemia + SegundaFeira + TercaFeira +
    QuartaFeira + QuintaFeira + SextaFeira + Sabado + seis_doze +
    doze_dezoito + dezoito_vintequatro, family = binomial, data = Dados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8507  -0.4704  -0.4118  -0.3575   2.5661

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.0533029  0.3549646  -2.967  0.00300 **
Idade              -0.0292536  0.0137691  -2.125  0.03362 *
IdadeQuadrado       0.0003071  0.0001393   2.205  0.02745 *
Homem              -0.0319825  0.0866302  -0.369  0.71199
Motas               1.2430034  0.1713149   7.256 4.00e-13 ***
IdadeVeiculo        0.0113672  0.0061005   1.863  0.06242 .
Regiao             -0.2855491  0.0889337  -3.211  0.00132 **
alcoolemia          2.5934500  0.4549254   5.701 1.19e-08 ***
SegundaFeira       -0.4315032  0.1323292  -3.261  0.00111 **
TercaFeira         -0.6414045  0.1449580  -4.425 9.65e-06 ***
QuartaFeira        -0.3541918  0.1363235  -2.598  0.00937 **
QuintaFeira        -0.3657644  0.1366376  -2.677  0.00743 **
SextaFeira         -0.4320631  0.1351391  -3.197  0.00139 **
Sabado             -0.3050334  0.1416950  -2.153  0.03134 *
seis_doze          -0.6191665  0.1231873  -5.026 5.00e-07 ***
doze_dezoito       -0.2434966  0.1169510  -2.082  0.03734 *
dezoito_vintequatro  0.0685761  0.1262254   0.543  0.58694
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5414.0  on 8624  degrees of freedom
Residual deviance: 5234.9  on 8608  degrees of freedom
AIC: 5268.9

Number of Fisher Scoring iterations: 5
```

> mod.M4Reduzido <- glm (Gravidade ~ 1 , family=binomial, data = Dados)

> summary(mod.M4Reduzido)

```
Call:
glm(formula = Gravidade ~ 1, family = binomial, data = Dados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4467  -0.4467  -0.4467  -0.4467   2.1700

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.25456    0.03673  -61.38   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5414  on 8624  degrees of freedom
Residual deviance: 5414  on 8624  degrees of freedom
AIC: 5416

Number of Fisher Scoring iterations: 5
```

>anova(mod.M4,mod.M4,test= "Chisq")

```
Analysis of Deviance Table

Model 1: Gravidade ~ Idade + IdadeQuadrado + Homem + Motas + IdadeVeiculo +
    Regiao + alcoolemia + SegundaFeira + TercaFeira + QuartaFeira +
    QuintaFeira + SextaFeira + Sabado + seis_doze + doze_dezoito +
    dezoito_vintequatro
Model 2: Gravidade ~ 1
  Resid. Df Resid. Dev  Df Deviance P(>|Chi|)
1      8608     5234.9
2      8624     5414.0 -16  -179.16 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

> exp(coef(mod.M4))

```
(Intercept)              Idade       IdadeQuadrado              Homem
 0.3487839          0.9711701          1.0003071          0.9685236

Motas        IdadeVeiculo             Regiao
3.4660075          1.0114321          0.7516014

alcoolemia        SegundaFeira          TercaFeira         QuartaFeira
13.3758381          0.6495320          0.5265524          0.7017404

QuintaFeira          SextaFeira              Sabado
0.6936662          0.6491684          0.7370987

seis_doze        doze_dezoito dezoito_vintequatro
0.5383930          0.7838822          1.0709821
```

> exp(confint.default(mod.M4))

```
                     2.5 %     97.5 %
(Intercept)        0.1739442  0.6993632
Idade              0.9453117  0.9977359
IdadeQuadrado      1.0000341  1.0005802
Homem              0.8172792  1.1477570
Motas              2.4774563  4.8490090
IdadeVeiculo       0.9994107  1.0235981
Regiao             0.6313746  0.8947219
alcoolemia         5.4838787 32.6252742
SegundaFeira       0.5011430  0.8418591
TercaFeira         0.3963265  0.6995681
QuartaFeira        0.5372021  0.9166746
QuintaFeira        0.5306943  0.9066855
```

```
SextaFeira          0.4981118  0.8460343
Sabado              0.5583605  0.9730534
seis_doze           0.4229044  0.6854198
doze_dezoito        0.6233069  0.9858246
dezoito_vintequatro 0.8362555  1.3715936
```

> pi.hat = predict.glm(mod.M4,data.frame(Idade =30, IdadeQuadrado = 900, Homem = 1, Motas = 1, IdadeVeiculo = 3, Regiao = 1, alcoolemia = 0, SegundaFeira = 0, TercaFeira = 1, QuartaFeira = 0, QuintaFeira = 0, SextaFeira = 0, Sabado = 0, seis_doze =1, doze_dezoito = 0, dezoito_vintequatro = 0), type = "response", se.fit = TRUE)
> pi.hat$fit

```
        1
0.1239519
```

> l.hat = predict.glm(mod.M4,data.frame(Idade =30, IdadeQuadrado = 900, Homem = 1, Motas = 1, IdadeVeiculo = 3, Regiao = 1, alcoolemia = 0, SegundaFeira = 0, TercaFeira = 1, QuartaFeira = 0, QuintaFeira = 0, SextaFeira = 0, Sabado = 0, seis_doze =1, doze_dezoito = 0, dezoito_vintequatro = 0), se.fit = TRUE)
> ci = c(l.hat$fit - 1.96*l.hat$se.fit,l.hat$fit + 1.96*l.hat$se.fit)
> exp(ci)/(1+exp(ci))

```
         1          1
0.08377704 0.17961560
```

BIBLIOGRAFIA

Abbess, C., Jarret, D. & Wright, C. (1981). Accidents at blackspots: estimating the effectiveness of remedial treatment, with special reference to the "regression-to-mean" effect. *Traffic Engineering and Control*, 22, 535-542. Recuperado a 2 Setembro, 2011, de http://trid.trb.org/view.aspx?id=180819

Abbring, J.; Chiappori, P. & Pinquer, J. (2003). Moral hazard and dynamic insurance data. *Journal of the European Economic Association*. 4. 767-820. Recuperado a 6 Março, 2011, de http://public.econ.duke.edu/~hf14/teaching/socialinsurance/reading/fudan_hsbc/abbring_chiappori_heckman03.pdf

Agresti, A. (1996). *An introduction to categories data analysis*. New York: Wiley. ISBN: 978-0-471-22618-5

Aldrich, J. & Cnudde, C. (1975). Probing the bounds of conventional wisdom: a comparation of regression, probit and discriminant analysis. *American Journal of political science*, 19, 571-608. Recuperado a 12 de Abril, 2011, de http://www.jstor.org/stable/2110547

Aldrich, J & Nelson F. (1984). *Linear probability, logit, and probit models - Advanced quantitative techniques in the social sciences series*, 7. London: SAGE Publications Ltd. ISBN: 0-8039-2133-0

Antonio, K. & Valdez, E. (2010). Statistical concepts of a priori and a posteriori risk classification in insurance. *AStA Advances in statistical analysis, forthcoming special issue on insurance statistics*, 1-35. Recuperado a 16 de Junho, 2011 de http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1665463

Areosa, J. (2008). *O Risco no âmbito da Teoria Social*. In: VI Congresso Português de Sociologia – Mundos Sociais: saberes e práticas. Universidade Nova de Lisboa, Faculdade de Ciências Sociais e Humanas. Lisboa. Recuperado a 25 de Julho, 2011, de http://www.aps.pt/vicongresso/pdfs/323.pdf

Associação Portuguesa de Seguradores. (2010). *O sector automóvel em Portugal – evolução recente da produção, sinistralidade e resultados.* (2010), 30, 14-20. Recuperado a 10 de Abril, 2011, de http://www.apseguradores.pt/Site/Home.jsf

Austin, J., Yaffe R. & Hinkle, D. (1992). Logistic regression for research in higher Education. *Higher education: handbook of theory and research*, 8, 379-410. Recuperado a 9 de Março, 2011, http://www.google.pt/search?tbm=bks&tbo=1&q=Logistic+regression+for+research+in+higher+Education.+Higher+education%3A+Handbook+of+theory+and+research%2C+8%2C&btnG=#start=20&hl=pt.PT&sa=N&tbo=1&tbm=bks&q=editions:6dDk8QhECgC&bav=on.2,or.r_gc.r_pw.,cf.osb&Fp=1cfd2a011ba3326d&biw=1046&bih=562

Balakrishnan, N. (1992). *Handbook of the Logistic Distribution*. NY: Marcel Dekker. ISBN-13: 97-8082-4785-8-71.

Bailer, A., Reed, L. & Stayner, L. (1997). Modelling fatal injuries rates using Poisson Regression: a case study of workers in agriculture, foresty, and fishing. *Journal of Safety Research*, 28, 177-186. DOI: 10.1016/S0022-4375(97)80006-0

Bailey, A. & Simon, J. (1960). Two studies in automobile Insurance Ratemaking. *Proceedings of the Casualty Actuarial Society*, 87, 1- 19. Recuperado a 2 de Maio, 2011, de http://casualtyactuarialsociety.org/pubs/proceed/proceed60/60001.pdf

Barron, D. (1992). The analysis of count data: overdispersion and autocorrelation. In P. Marsden (ed.). *Sociological Methodology*, Blackwell, 179-220. Recuperado a 22 de Março, 2011, de http://www.mendeley.com/research/the-analysis-of-count-data-overdispersion-and-autocorrelation-2/

Berkson, J. (1953). A statistical precise and relatively simple method of estimating the Bioassay with quantal response, based on the logistic function. *Journal of theAmerican Statistical association*. 48. 565-599. Recuperado a 14 de Março, 2011, de http://www.jstor.org/pss/2281010

Black, T. (1993). *Evaluating Social Science Research*. London: Sage Publications. ISBN: 0-8039-8853-2

Blincoe, L., Seay, A., Zaloshnja, E., Miller, T., Romano, E., Luchter, S. & Spicer, R. (2012). *The economic impact of motor vehicle crashes 2000*. Washington: U.S. Department of Transportation. Recuperado a 9 de Dezembro, 2011, de http://www-nrd.nhtsa.dot.gov/Pubs/809446.pdf

Böhning, D., Dietz, E. & Schlattmann, P. (1997). Zero-inflated count models and their Applications in public health and social science. In J. Rost & R. Langeheine (eds.) *Aplications of latent traits and latent class models in the social sciences*, 333-344, Recuperado a 15 de Março, 2011, http://www.ipn.uni-kiel.deaktuell/buecher/rostbuch/c32.pdf

Boskin, M. (1974). A conditional logit model of occupational choice. *Journal of political economy*. 82. 389-398. Recuperado a 14 de Março, 2011, de http://www. jstor.org/pss/1831185

Brännäs, K. (1992a). Finite sample proprieties of estimators and tests in Poisson regression models. *Journal of Statistical Computation and Simulation*, 41, 229-241. DOI: 10.1080/00949659208811403

Brännäs, K. (1992b). Limited dependent Poisson regression. *The Statistician*, 41, 413-423. Recuperado a 22 de Março, 2011, de http://www.jstor.org/pss/2349006

Breslow, N. (1990). Tests of hypoteses in overdispersed Poisson regression and other quasi-likelihood models. *Journal of the American Statistical Association*, 85, 565-571. Recuperado a 30 de Março, 2011, de http://www.jstor.org/pss/2289799

Brockman, M. & Wright, T. (1992). Statistical Motor rating: Making effective use of your data. *Journal of Institute of Actuaries*, 119, 457-543. Recuperado a 1 de Maio, 2011, de http://www.actuaries.org.uk/research-and-resources/documents /statistical-motor-rating-making-effective-use-your-data

Burnham, J. C. (2009). *Accident prone: a history of technology, psychology, and misfits of the machine age*, 4, 67-86. Chicago: University of Chicago Press Ltd. Recuperado a 8 de Março, 2011, de http://www.press.uchicago.edu/ucp/books/book/chicago/A/bo6407645.html

Cabrera, A. (1994). Logistic regression analysis in higher education: an applied erspective. *Higher education: handbook of theory and research*, 10, 225-256. Recuperado a 9 de Março, 2011, de http://www.education.umd.edu/EDPL/faculty/cabrera/Chapter%20on%20logistic%20regression.pdf

Cameron, A. & Trivedi, P. (1986). Econometric models based on count data: Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics*, 1, 29-53. DOI: 10.1002/jae.3950010104

Cameron, A. & Trivedi, P. (1990). Regression-based tests for overdipersion in the Poisson model. *Journal of Econometrics*, 46, 347-364. DOI: 10.1016/0304-4076(90)90014-K

Cameron, A. & Windmeijer, F. (1996). R-square measures for count data regression models with applications to health-care utilization. *Journal of Business and economic statistics*, 14, 209-220. DOI: 10.1.1.169.6693

Cameron, A. & Trivedi, P. (1998). *Regression analysis of count data*. Cambridge: University Press. ISBN: 0-521-63201-3

Carrivick, P., Lee, A. & Yau, K. (2003). Zero-inflated Poisson modelling to evaluate occupational safety interventions. *Safety Science*, 41, 53-63. DOI: 10.1016/S0925-7535(01)000571

CEA (2007). *The European Motor Insurance Market.* Recuperado a 30 de Março, 2011, de http://www.cea.eu

Chiappori, P. & Salanié B. (2000). Testing for asymmetric information in Insurance Markets. *The Journal of Political Economy*, 108, 1. 56-78. Recuperado a 19 de

Março, 2011, http://links.jstor.org/sici?sici=0022-3808%28200002%29108%3A1%3C56%3ATFAIII%3E2.0.CO%3B2-J

Chin, H. e Quddus, M. (2003). Modeling count data with excess zeros. An empirical application to traffic accidents. *Statistics in Medicine,* 21, 1461-1469. Recuperado a 30 de Março, 2011, http://www.cts.cv.ic.ac.uk/documents/publications/iccts00290.pdf

Chuang, H. (1997). High school youth's dropout and re-enrollment behaviour. *Economics of educational review*, 16, 171-186. DOI: 10.1016/S0272-7757(96)0058-1

Companhia de Seguros Fidelidade-Mundial. (2011). *Relatório e Contas.* Lisboa: Grupo CGD. Recuperado a 15 de Junho, 2011, Recuperado a 15 de Julho, 2011, de http://www.cgd.pt/Corporativo/Informacao-Financeira/Seguradoras/Fidelidade-Mundial/Documents/RC-Fidelidade-Mundial-2010.pdf

Companhia de Seguros Império-Bonança. (2011). *Relatório e Contas.* Lisboa: Grupo CGD. Recuperado a 15 de Junho, 2011, de http://www.cgd.pt/Corporativo/Informacao-Financeira/Seguradoras/Imperio-Bonanca/Documents/RC-Imperio-Bonanca-2010.pdf

Companhia de Seguros Via Directa. (2011). *Relatório e Contas.* (2010). Lisboa: Grupo CGD, Recuperado a 15 de Julho, 2011, de http://www.cgd-pt/Corporativo/Informacao-Financeira/Seguradoras/Via-Directa/Documents/RC-Via-Directa-2010.pdf

Cordeiro, G. (1986). *Modelos Lineares Generalizados*. In: VII Simpósio Nacional de Probabilidades e Estatística. São Paulo. Recuperado a 23 de Julho, 2011, de http://www.bdpa.cnptia.embrapa.br/busca.jsp?baseDados=ACERVO&fraseBusca=%22CORDEIRO,%20G.M.%22%20em%20AUT

Cragg, J. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*. 39. 829-844. Recuperado a 9 de Março, 2011, de http://www.jstor.org/pss/1909582

Cragg, J. & Uhler, R. (1970). The demand for automobiles. *The Canadian journal of Econometrics*. 3. 386-406. Recuperado a 9 de Março, 2011 de http://www.jstor.org/pss/133656

Dake, K. (1991). Orienting dispositions in the perception of risk: an analysis of contemporary worldviews and cultural biases. *Journal Cross-Cultural Psycology*, 22, 61-82. DOI: 10.1177/0022022191221006

Davutyan, N. (1989). Bank failures as Poisson variates. *Economic Letters*, 29, 333-338. DOI:10.1016/0165-1765(89)90212-7

Dean, C. (1992). Testing for overdispersion in Poisson and binomial regression models. *Journal of the American Statistical Association*, 35, 451-457. Recuperado a 18 de Março, 2011, de http://www.jstor.org/pss/2290276

Dean, C. & Lawless, J. (1989). Tests for detecting overdispersion in Poisson regression. models. *Journal of the American Statistical Association*, 84, 467-472. Recuperado a 23 de Março, 2011, de http://jstor.org/pss/2289931

Decreto Lei nº 408/79 de 25 de Setembro. *Diário da República nº 222 – I Série A*. Ministério das Finanças. Lisboa. Recuperado a 12 de Abril, 2011, de http://diario.vlex.pt/vid/decreto-lei-setembro-33063029

Denuit, M., Maréchal, X., Pitrebois, S. & Walhin J. (2007). Actuarial modelling of claim counts – risk classification, credibility and Bonus-Malus systems. London: Wiley. ISBN: 13 978-0-470-02677-9

Dionne, G. (2000). *Handbook of Insurance*. Massachusetts: Kluwer Academic Publishers. Recuperado a 19 de Março, 2011, de http://www.springer.com/business+%26+management/finance/book/978-0-7923-7911-9

Dionne, G. & Vanasse, C. (1992). Automobile insurance ratemaking in the presence of asymmetrical information. *Journal of Applied Econometrics*, 7, 149-165. DOI: 10. 1002/jae.3950070204

Dionne, G., Desjardins, D., Laberge-Nadeau, C. & Maag, U. (1995). Medical conditions, risk exposure and truck driver's accidents: an analysis with count data regression models. *Accident Analysis & Prevention*, 27, 295-305. DOI: 10.1016/0001+4575(94)00071-S

Dionne, G., Gagné, R., Gagnon F. & Vanasse C. (1997) Debt, moral hazard and air line safety an empirical evidence. *Journal of Econometrics*, 79, 379-402. DOI: 10.1016/S0304-4076(97)82989-2

Dionne, G. & Laberge-Nadeau, C. (1999). *Automobile insurance: road safety, newdrivers, risks, insurance fraud and regulation*. Massachusetts: Kluwer Academic Publishers. ISBN: 978-0-7923-8394-9

Dobson, A. (1990), *An Introduction to Generalized Linear Models*. London: Chapman & Hall. Recuperado a 23 de Julho, 2011, http:www.ats.ucla.edu/stat/examples/iglm/default.htm

Dooley, M. (1997). A model of crises in emerging markets. *Economic Journal*, 110, 256-272. Recuperado a 13 de Março, 2011, de http:www.nber.org/papers/w6300

Draper, N. & Smith H. (1998). *Applied Regression Analysis*. New York: Wiley. ISBN: 0-471-17082-8

Eurostat (2010). *European statistics on death causes.* Recuperado a 26 de Fevereiro, 2011, de http://epp.eurostat.ec.europa.eu

Fahrmeir, L. & Tutz, G. (1994).*Multivariate Statistical modelling based on Generalized Linear Models*. New York: Springer. Recuperado a 23 de Julho, 2011, de http://www.library.wisc.edu/selectedtocs/bc025.pdf

Flood, R. & Garber, P. (1984). Collapsing exchanges-rate regimes: some linear examples. *Journal of International Economics*, 17, 1-13. DOI: 10.1016/0022-1996(84)90002-3

Freud, F. (1901). Psychopathology of everyday life. *Classics in the history of psychology*. Recuperado a 19 de Março, 2011, de http://www.dominiopublico. gov.br//pesquisa/DetalheObraForm.do?select_action=&co_obra=4562

Ganio, L. & Schafer, D. (1992). Diagnostics for overdespersion. *Journal of the American Statistical Association*, 87, 795-804. Recuperado a 26 de Março, 2011, de http://www.jstor.org/pss/2290217

Gardner, W., Mulvey, E. & Shaw, E. (1995). Regression analysis of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin*, 118, 392-404. DOI: 10.1037/0033-2909.118.3.392

Gata, J. (2006). Política de concorrência nos seguros: a experiência portuguesa. *Jornal de Negócios*. Recuperado a 15 de Março, 2011, de http://www2.egi.ua.pt/ cursos_2005/files/DE/JN%20Janeiro%202006%20_AdC_.pdf

Gelman, A & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: University Press. ISBN: 978-0-521-68689-1

Gilberto, F. (2010). *Os grandes desafios da indústria seguradora*. Lisboa: Lidel. ISBN:978-972-757-689-0

Gordon, H. (2004). Psychiatry, the law and death on the roads. *Advanced in psychiatric Treatment*, 10, 439-445. DOI:10.119/apt.10.6.439

Gourieroux, C. & Visser, M. (1997). A count data model with unobserved heterogeneity *Journal of Econometrics*, 79, 247-268. DOI: 10.1016/S0304-4076(97)00022-5

Greene, M. (1993). Cognitive remediation in schizophrenia: is it time yet?. *American Journal of Psychiatry*, 150, 178-187. Recuperado a 17 de Março, 2011, de http://www.ncbi.nlm.nih.gov/pubmed/8422068

Guerreiro, G. & Mexia, J. (2003). Open approach to bonus malus. *Boletim do Instituto dos Actuários Portugueses*. 42. 41-51. Recuperado a 15 de Março, 2011, de http://www.institutodosactuarios.org/docs/Boletins/BoletimIAP42.pdf

Guikema, S. & Coffelt J. (2007). A flexible count data regression model for risk analysis. *Risk Analysis*. 28, 213-221. Recuperado a 2 de Setembro, 2011, de http://dx.doi.org/10.1111/j.1539-6924.2008.01014x

Guo, J. & Li, T. (2000). Poisson regression models with errors-in-variables: implication and treatment. *Journal of Statistical Planning and Inference*, 104, 391-401. DOI: 10.1016/S0378-3758(01)00250-6

Gupta, A., Jukic, B., Li, M., Stanhl, D., & Whinston, A. (2001). Estimating internet users'demand characteristics. *Computational Economics*. 17. 203-218. DOI: 10. 1023/A:1011648714807

Gurmu, S. (1991). Tests for detecting overdispertion in the positive Poisson regression model. *Journal of Business and Economic Statistics*, 9, 215-222. Recuperado a 17 de Março, 2011, de http://www.jstor.org/discover/10.2307/1391790?uid=3738880&uid=2&uid=4&sid=21101169194007

Gurmu, S. & Trivedi, P. (1992). Overdispertion tests for truncated Poisson regression models. Journal of Econometrics, 54, 347-370. DOI: 10.1016/0.04(92)90113-6

Hair J., Anderson R., Tatham, R. & Black W. (1998). *Multivariate Data Analysis* (5ª Ed.). New York: Prentice Hall. ISBN: 0-13-017706-7

Harrel, F., Lee, K. & Mark D. (1996). Tutorial in Biostatistics: Multivariable prognostic Models: Issues in developing models, evaluating assumptions and measuring

and reducing errors. *Statistics in Medicine*, 15, 361-387. Recuperado a 5 de Julho, 2011, de http://www.unt.edu/rss/class/Jon/MiscDocs/Harrell_1996.pdf

Harvey, G. (2004). Psychiatry, the law and death on the roads. *Advances in Psychiatric Treatment*, 10, 439-445. DOI: 10.1192.apt.10.6.439

Hauer, E., Ng, J. & Lovell, J. (1988). Estimation of safety at signalized at intersections. *Transportation Research Record*. 1185. 48-61. Recuperado a 2 de Setembro, 2011, de http://pubsindex.trb.org/view.aspx?id=301420

Hausman, J., Hall, B. & Griliches, Z. (1984). Econometric models for count data with an application to the patents-R&D relationship. *Econometrica*, 52, 909-938. Recuperado a 17 de Abril, 2011, de http:links.jstor.org/sici?sici=0012-9682%28198407%2952%3A4%3C909%3AEMFCDW%3E2.0.Co%3B2-M

Haynes, M., Thompson, S. & Wright, M. (2003). The determinants of corporate divestement: evidence from a panel of UK firms. Journal of Economic Behavior & Organization, 52, 147-166. ISSN: 1360-2438

Heitfield, E. & Levy, A. (2001). Parametric, semi-parametric, and non-parametric models of telecommunications demand. An investigation of residential calling paterns. *Information economics and policy*, 13, 311-329. DOI: 10.1016/S0167-6245(01)00033-6

Henderson, J. & Taylor, B. (2003). Rural isolation and the availability of hospitalservices. *Journal of rural studies*, 19, 363-372. DOI: 10.1016/S0743-0167 (03)00007-X

Hilbe, J. (2007). *Negative Binomial regression*. (2ª ed.). New York: Cambridge University Press. ISBN: 978-0521-19815-8

Hinde, J. & Demétrio, C. (1998). Overdispersion: models and estimation. *Computational Statistics & Data Analysis*, 27, 151-170. DOI: 10.1016/S0167-9473 (98)00007-3

Hoffman, H. H. (2005). Comportamento do condutor e fenómenos psicológicos. *Psicologia: Pesquisa & Trânsito*, 1, 17-24. Recuperado a 15 de Fevereiro, 2011, de http://pepsic.bvsalud.org/scielo.php?pid=S1808-91002005000100004&script=sci_arttext

Hosmer, D. & Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics*. A10. 1043-1069. DOI: 10.1080/03610928008827941

Hosmer, D. & Lemeshow, S. (2000). *Applied Logistic Regression* (2ª Ed.). New York: Wiley. ISBN: 0-471-35632-8

Instituto de Seguros de Portugal. (2007). *Guia de Seguros e Fundos de Pensões*. 17-26. Lisboa. Recuperado a 15 de Março, 2011, de http://www.isp.pt/NR/rdonlyres/C688F77F-0CE7-4059-9C88/Guiawebuv.pdf

Instituto de Seguros de Portugal. (2011). *Relatório do Sector Segurador e Fundos de Pensões*. Lisboa. Recuperado a 8 de Dezembro, 2011, de http://www.isp.pt/NR/rdonlyres/06B0C9C2-515B-47B8-9696-22F471AB9EE9/0/E-RSSP2010AF.pdf

Janik, J. & Kravitz H. (1994). Linking work and domestic problems with police suicide. *Suicide and life threatening behaviour*, 24, 267-274. DOI: 10.1111/j.1943-278X.1994.tb00751.x

Johnston, J. & Dinardo, J. (1996). *Econometric Methods* (4ª ed.). New York: McGraw Hill International Editions. ISBN:9780071153423

Jorion, P. (2000). *Value at risk. The new benchmark for managing financial risk.* (3ª ed.) New York: McGraw-Hill International Editions. ISBN-13: 978-0-07-146495-6

Kennan, J. (1985). The duration of strikes in U.S. manufacturing. *Journal of Econometrics*, 28, 5-28. DOI: 10.106/0304-4076(85)90064-8

King, G. (1989). Variance specification in event count models: from restrictive assumptions to generalized estimators. *American Journal of Political Science*, 33, 762-784. DOI: 10.2307/2111071

Kouabenan, D. (1998). Beliefs and the perception of risk and accidents. *Risk Analysis*, 18, 243-252. DOI: 0272-4332/98/0600-0243815.00/1

Krugman, P. (1979). A model of balance-of-payments Crises. *Journal of Money, Credit and Banking.* Recuperado a 17 de Março, 2011, de http://www.kailchan.ca/wp-content/uploads/2010/10/A-model-of-balance-of-payments-crises.pdf

Kumala, R. (1995). Safety at rural three-and four-arm junctions: development and applications of the accident prediction models. *VTT Publications*, 233. Recuperado a 2 de Setembro, 2011, de http://trid.trb.org/view.aspx?id=684994

Lambert, D. (1992). Zero-inflated regression with an application to defects in manufacturing. *Tecnometrics*, 34, 1-14. Recuperado a 27 de Março, 2011, de http://www.jstor.org/pss/1269547

Lawless, J. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics*, 15, 209-225. DOI: 10.2307/3314912

Lee, A., Stevenson, M., Wang, K. & Yau, K. (2002). Modelling young drivers motor vehicle crashes: data with extra zeros. *Accident analysis & prevention*. 34. 515-521. DOI: 10.1016/S0001+4575(01)00049.5

Lee, L. (1986). Specification test for Poisson regression models. *International Economic Review*, 27, 689-706. Recuperado a 23 de Março, 2011, de http://www.jstor.org/discover/10.2307/2526689?uid=3738880&uid=2&uid=4&sid=21101169241317

Lei nº 19/98 de 28 de Abril. *Diário da República nº 98 – I Série A*. Ministério da Justiça. Lisboa. Recuperado a 1 de Maio, 2011, de http://dre.pt/pdf1s/1998/04/098A00/18881888.pdf

Lemaire, J. (1995). *Bonus-malus system in automobile insurance.* Massachusetts: Kluwer Academic Publishers. ISBN: 0-7923-9545-X

Levitt, D. & Porter, J. (1999). Sample selection in the estimation of air bag and seat belt effectiveness. *The review of economics and statistics*. 83, 603-615. Recuperado a 17 de Março, 2011, de http://pricetheory.uchicago.edu/levitt/Papers/LevittPorter2001.pdf

Li, G.; Braver, E. & Chen, L. (2001). Fragility versus excessive crash involvement as determinants of high death rates per vehicle-mile of travel among old drivers. *Accident Analysis and Prevention*. 35, 227-235. Recuperado a 6 de Março, 2011, de http://www.ncbi.nlm.nih.gov/pubmed/12504143

Lindsey, J. (1997). *Applying Generalized Linear Models*. New York: Springer. ISBN: 0-387-98218-3

Long, J. (1997), *Regression models for categorical and limited dependent variables – Advanced quantitative techniques in the social sciences series,* 7. London: SAGE Publications Ltd. ISBN: 0-8039-7374-8

Lord, D. (2000). *The predition of accidents on digital networks: characteristics and issues related to the application of accident predition models*. Dissertação de Mestrado. Department of Civil Engineering, University of Toronto, Ontario, Canada. Recuperado a 2 de Setembro, 2011, de http://www.collectionscanada.gc.ca/obj/s4/f2/dsk1/tape4/PQDD_0020/NQ53687.pdf

Lord, D., Mannar, A. & Vizioli, A. (2005). Modelling crash-flow-density and crash-flow-V/C ratio for rural and urban freeway segments. *Accident Analysis & Prevention*, 37, 35-46. Recuperado a 2 de Setembro, 2011 de http://projectwaalbrug.pbworks.com/f/sdarticle.pdf

Maddala, G. (1983), *Limited-dependent and quantitative variables in econometrics.* Cambrige: University Press. ISBN: 0521338255

Maindonald, J & Braun, W. (2006). Data Analysis and Graphics Using R – An example based approach. (2ª ed.). Cambridge: Series in statistical and probabilistic mathematics. ISBN-13: 978-0-511-24957-0

Maslow, A. (1954). *Motivation and Personality*. NY: Harper Row. Recuperado a 11 de Março, 2011, de http://www.chaight.com/Wk%2015%20E205B%20Maslow%20%-20Human%20Motivation.pdf

Maul, A., El-Shaarawi, A. & Ferard, J. (1991). Application of negative binomial regression models to the analysis of quanta bioassays data. *Environmetrics*, 2, 253-261. DOI: 10.1002/env.3770020302

McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models*, (2ª ed.). New York: Chapman & Hall. ISBN: 0-412-31760-5

McFadden, D. (1974). The measurement of urban travel demand. *Journal of public economics*. 3. 303-328. Recuperado a 15 de Março, 2011, de http://elsa.berkeley.edu/pub/reprints/mcfadden/measurement.pdf

Miaou, S. & Lord, D. (2003). Modelling traffic crash-flow relationships for Intersections: dispersion parameter, functional form, and Bayes versus Empirical Bayes. *Transportation Research Record*, 1840, 31-40. Recuperado a 2 de Setembro, 2011, de http://trb.metapress.com/content/ymj7un0267785616/

Miaou, S. & Song, J. (2005). Bayesian ranking of sites for engineering safety Improvements: decision parameter, treatability concept, statistical criterion and Spatial dependence. *Acident Analysis and Prevention*, 37, 699-720. Recuperado a 2 de Setembro, 2011, de http://www.sciencedirect.com/science/article/pii/S0001457505000497

Mukerjee, R. & Sutradhar, B. (2002). On the positive definiteness of the information matrix under binary and Poisson mixed models. *Annals of the Institute of Statistical Mathematics*, 54, 355-366. DOI: 10.1023/A:1022478119885

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33, 341-365. DOI: 10.1016/0304-4076(86)90002-3

Mullahy, J. (1997). Heterogeneity, excess zeros, and the structure of count data models. *Journal of Applied Econometrics*, 12, 337-350. DOI: 10.1002/(SICI)1099-1255(199705)

Mulligan, J. G. (1989), *Managerial economics – Strategy for profit*, 17, 493-519. Massachusetts: Allyn and Bacon. ISBN: 0-205-11711-2

Murteira, B., Ribeiro, C., Silva. J. & Pimenta, C. (2010). *Introdução à Estatística*. Lisboa: Escolar Editora. ISBN: 978-972-592-282-8

Obstfeld. M (1984). The logic of currency crises. *National Bureau of Economic Research*, 4640, 1-54 Recuperado a 8 de Março, 2011, de http://www.nber. org/papers/w4640.pdf

Oh, J., Washington, S. & Nam, D. (2006). Accident prediction model for railway-Highway interfaces. *Accident Analysis & Prevention*, 38, 346-356. Recuperado a 2 de Setembro, 2011, de http://www.sciencedirect.com/science/article/pii/S0001457505001776

Oliveira, P. (2007). *Os factores potenciadores da sinistralidade rodoviária*. Recuperado a 12 de Maio, 2011 de http://www.aca-m.org/w/images/3/3d/Factores_potenciadores_sinistralidade_rodoviaria.pdf

Ozkan, F & Sutherland, A. (1995). Policy measures to avoid a currency crisis. *The Economic Journal*, 105, 510-519. Recuperado a 12 de Março, 2011 de http://www.jstor.org/discover/10.2307/2235508?uid=3738880&uid=2&uid=4&sid=21101169261297

Peden, M., Scurfield, R., Sleet, D., Mohan, D., Hyder, A., Jarawan, E., et al. (2004). *World report on road traffic injury prevention*. Genova: Wold Health

Organization. Recuperado a 15 de Março de http://www.who.int/violence_ injury_prevention/publications/road_traffic/world_report/en/index.html.

Peng, C., Manz, B. & Keck , J. (2001). Modelling categorical variables by logistic regression. *American journal of health behaviour*, 25, 278-284. ISSN: 1087-3244

Peng, C. & So, T. (2002). Modelling strategies in logistic regression. *Journal of modernapplied statistical methods*, 14, 147-156. Recuperado a 9 de Março, 2011, de http://www.indiana.edu/~jopeng51/teaching-logistic.pdf

Petridou, E., & Moustaki M. (2000), Human factors in the causation of road traffic crashes. *European Journal of Epidemiology*, 16, 819-826. Netherlands: Kluwer Academic Publishers. DOI: 10.1023/A:1007649804201

Poch, M. & Mannering, F. (1996). Negative binomial analysis of intersection-accident frequency. *Journal of Transportation Enginnering*, 122, 105-113. Recuperado a 2 de Setembro, 2011, de http://heather.cs.ucdavis.edu/~matloff/132/Transp.pdf

Radelet, S. & Sachs, J. (1998). The onset of the East Asian financial crisis. *National Bureau of Economic Research*, 6680, 105-162. Recuperado a 4 de Abril, 2011, de http://www.nber.org/chapters/c8691

Ragsdale, L. (1984). The politics of presidential speechmaking, 1949-1980. *TheAmerican Political Science Review*, 78, 971-984. Recuperado a 12 de Março, 2011, de http://www.jstor.org/stable/1955802

Rawlings, J., Pantula, S. & Dickey, D. (1998). *Applied Regression Analysis – A Research Tool*, (2ª ed.). New York: Springer. ISBN: 0-387-98454-2

Richardson, R. J. (1999). *Pesquisa Social: métodos e técnicas* (3ª ed.). São Paulo: Atlas. Recuperado a 19 de Julho, 2011, de http://hdl.handle.net/1904/11430

Rothschild, M. & Stiglitz (1976), Equilibrium in competitive insurance market: an essay on the economics of imperfect information. *Quaterly Journal of Economics*, 90, 236-257. Recuperado a 10 de Março, 2011, de http://www.jstor.org/pss/1885326

Santos, J., Oliveira, S., & Barroso, C. (2011). An innovative statistical approach for analysing non-continuous variables in environmental monitoring: assessing temporal trends of TBT pollution. *Journal of Environmental Monitoring*, 13, 673- 680. DOI: 10.1039/c0em00435a

Schmidt, L. & Nave, J. (2004). *O Automóvel – usos e desusos do transporte individual*. Lisboa: Instituto Superior das Ciências do Trabalho e da Empresa. Recuperado a 15 de Março, 2011, de http://observa.iscte.pt/docs/relatorio%20final %20automovel.pdf

Schmidt, P. & Strauss R. (1975). The prediction of occupation using multiple logit models. *Internacional Economic Review*. 16. 471-486. Recuperado a 14 de Março, 2011, de http://www.jstor.org/pss/2525826

Schor, T. (1999). O automóvel e o desgaste social. *São Paulo em Perspectiva*. 13. DOI: 10.1590/S0102-8839199900300014

Schneider, R., Ryznar, R. & Khattak, A. (2004). An accident waiting to happen: a special approach to proactive pedestrian planning. *Accident Analysis & Prevention*, 29, 829-837. DOI: 10.1016/S0001-4575(02)00149-5

Shankar, V., Ulfarsson, G., Pendyala, R. e Nebergall, M. (2003). Modelling crashes involving pedestrians and motorized traffic. *Safety Science*, 41, 627-640. DOI: 10.1016/S0925-7535(02)00017-6

Spector, L. & Mazzeo, M. (1980). Probit analysis and economic education. *Journal of Economic Education*, 11, 37-44. Recuperado a 17 de Março, 2011 de http://www.jstor.org/discover/10.2307/1182446?uid=3738880&uid=2&uid=4&sid=21101169286247

Theil, H. (1969). A multinomial extension of the linear logit model. *International Economic review*. 10. 251-259. Recuperado a 15 de Março, 2011, de http://www.jstor.org/pss/2525642

Tillman, R. & Pontel, H. (1995). Organizations and fraud in the savings and loan industry. *Social Forces*, 73, 1439-1463. Recuperado a 20 de Março, 2011, de http://www.jstor.org/stable/2580454

Tolman, R. & Weisz A. (1995). Coordinated community intervention for domesticviolence: the effects of arrest and prosecution on recidivism of woman abuse perpetrators. *Crime and delinquency*, 41, 481-495. DOI: 10.1177/001112879504100 4007

Vicente, P., Reis, E. & Ferrão, F. (2001). *Sondagens – A amostragem como factor decisivo de qualidade*. Lisboa: Edições Sílabo. ISBN: 972-618-246-8

Vieira, D. & Quintero, A. (2008), *Aspectos práticos da avaliação do dano corporal em Direito Civil*. Coimbra: Caixaseguros. ISBN:978-989-8074-31-7

Visser, E., Pijl Y. J., Stolk, R. P., Neeleman, J., & Rosmalen. J. G. M. (2006). Accident proneness does it exist? A review and meta-analysis. *Accident Analysis and Prevention*, 39, 556-64. Elsevier Ltd. DOI:10.1016/j.aap.2006.09.012

Wang K., Lee, A., Yau, K. & Carrivick (2003). A bivariate zero-inflated Poisson regression model to analyze occupational injuries. *Accident Analysis & Prevention*, 35, 625-629. DOI: 10.1016/S0001-4575(02)00036-2

Winkelmann, R. (2000). *Econometric analysis of count data*. (3ª ed.). London: Springer. ISBN: 3-540-67340-7

Winkelmann, R. & Zimmermann, K. (1994). Count data models for demographic data. *Mathematical Population Studies*, 4, 205-21. DOI:10.1080/08898489409525374

Winkelmann, R. & Zimmermann, K. (1995). Recent developments in count data modelling: theory and application. *Journal of Economic Surveys,* 9, 1-24. DOI: 0950-0804/95/01 0001-24

Winkelmann, R. & Zimmermann, K. (1998). Is job stability declining in Germany? Evidence from count data models. *Applied Economics,* 30, 1413-1420. DOI: 10.1080/000368498324760

Wooldridge, J. (2002). *Introductory econometrics – a modern approach*. (2ª ed.) Ohio: South-Western. ISBN: 0324113641

Xie, Y., Lord, D. & Zhang, Y. (2007). Predicting motor vehicle collisions using Bayesian neural networks: an empirical analysis. *Accident Analysis & Prevention*, 39, 922-933. Recuperado a 2 de Setembro, 2011, de http://www.sciencedirect.com/science/article/pii/S0001457507000073

Zeileis A., Kleiber C., Jackman S. (2008). Regression Models for Count Data in R. *Journal of Statistical Software*. 27. 1-27. Recuperado a 16 de Setembro, 2011, de http://cran.r-project.org/web/packages/pscl/vignettes/countreg.pdf

Zorn, C. (1998). An analytic and empirical examination of zero-inflated and hurdle Poisson specifications. *Sociological Methods & Research*, 26, 368-400. Recuperado a 23 de Março, 2011, de http://www.polmeth.wustl.edu/media/Paper/zorn96.pdf