UNIVERSIDADE DE SÃO PAULO
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
MESTRADO PROFISSIONALIZANTE
"MODELAGEM MATEMÁTICA EM FINANÇAS"

PRECIFICAÇÃO DE SEGUROS DE AUTOMÓVEL

Felipe Villarino Prieto

Orientador: Profº Dr Gilberto Alvarenga Paula

São Paulo
2005

PRECIFICAÇÃO DE SEGUROS DE AUTOMÓVEL

FELIPE VILLARINO PRIETO

Dissertação apresentada à Faculdade de Economia, Administração e Contabilidade e ao Instituto de Matemática e Estatística da Universidade de São Paulo para obtenção do Título de Mestre.

Orientador: Prof. Dr. Gilberto Alvarenga Paula

São Paulo

2005

FICHA CATALOGRÁFICA

Elaborada pela Seção de Processamento Técnico do SBD/FEA/USP

Prieto, Felipe Villarino
Precificação de seguros de automóvel / Felipe Villarino Prieto.
– São Paulo, 2005.
95 p.

Dissertação (Mestrado Profissionalizante) – Universidade de São Paulo, 2005
Bibliografia.

1. Finanças 2. Seguro de automóveis I. Universidade de São Paulo. Faculdade de Economia, Administração e Contabilidade. II. Universidade de São Paulo. Instituto de Matemática e Estatística. III. Título.

CDD – 332

A meus pais Felipe e Lucila pelo carinho, amor e pelos esforços dedicados à educação, saúde e caráter dos seus filhos. À minha noiva Maristela pelo apoio e compreensão durante a realização deste trabalho.

Com destaque especial, faço meus agradecimentos ao Prof. Gilberto Alvarenga Paula, por sua segura orientação e, sobretudo por sua dedicação e apoio na elaboração desta dissertação. Agradeço ainda aos demais mestres do curso, pelos conhecimentos transmitidos, e à FEA e ao IME USP, pelo apoio institucional e pelas facilidades oferecidas.

RESUMO

O mercado segurador brasileiro e, especialmente, o seguro de automóveis é extremamente competitivo, o que obriga as seguradoras a realizarem uma tarifação correta e bem ajustada de acordo com o perfil de risco do segurado. Cada vez mais tem se buscado precificar de uma forma mais granularizada havendo um preço para cada segurado, conforme as características de cada um. Os modelos estatísticos vêm ganhando grande importância em seguros, na medida que eles se adaptam muito bem aos problemas e necessidades apresentadas. Neste trabalho, descrevemos alguns aspectos do seguro de automóveis e a partir de uma base de dados fornecida por uma seguradora brasileira, calculamos o preço individual a ser cobrado para suprir todos os prejuízos ocorridos na carteira. Utilizamos três técnicas combinadas para calcular a freqüência e o custo médio de sinistros: árvores de decisão, regressão logística e modelos lineares generalizados. A árvore de decisão foi utilizada para identificar as possíveis interações entre as variáveis e trouxe um enorme ganho na qualidade de ajuste dos modelos. Os diagnósticos do modelo e análise de resíduos apresentados sugerem um bom ajuste dos modelos. Assim, conseguimos encontrar uma forma bastante eficaz de precificar o seguro de auto mediante o perfil de risco do segurado.

ABSTRACT

The Brazilian insurance market and more specifically car insurance is extremely competitive, which forces insurance companies to practice tight fees, well adjusted to the client's risk-profile. There has been an increasing attempt to price insurance in a more customized way. Statistical models have gained great importance in as much as they fit very well the problems and needs at hand. In the present work, some aspects of car insurance are described, and from a database supplied by a Brazilian insurance company, the individual fee is calculated in order to cover all damage caused to the cast. Three combined techniques have been employed to calculate the frequency and the average cost of accidents: decision trees, logistic regression and generalized linear models. The decision tree was been used to identify possible interactions among the variables, and has brought an enormous gain to the goodness-of-fit. The model's diagnostics and further analyses of the remains presented suggest the model's fine tuning. Thus, highly efficient means of pricing car insurance according to the client's risk profile has been found.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 CAPÍTULO I

1.1 Introdução

A atividade seguradora está difundida em todo o mundo, destacando-se os seguros de automóveis, incêndios, roubo, vida etc. Essa atividade responde à grande incerteza que as pessoas sentem diante de certas situações, que provocam perdas materiais e pessoais. O medo provocado pela possibilidade de ocorrer tais situações é, de certa forma, amenizado mediante a contratação de um seguro, que indenizará o segurado em um eventual prejuízo.

No início, o seguro era uma forma de solidariedade entre os membros de uma comunidade. Consistia basicamente em um fundo em que as pessoas depositavam seu dinheiro. Com o capital acumulado por todos, pagava-se o eventual prejuízo sofrido por alguns membros do grupo. No século XIV, existia em alguns portos o costume dos armadores e comerciantes de uma determinada rota depositarem em um fundo comum uma determinada quantia, em função do número de navios que possuíssem. Aqueles marinheiros que por ventura tivessem seus barcos naufragados, ou fossem abordados por piratas, recebiam uma compensação econômica procedente do fundo comum para adquirir um outro navio, a fim de continuar a exercer suas atividades.

A Ciência Atuarial tal como se conhece hoje tem início no século XVII. Durante esse período, as necessidades comerciais deram lugar a operações que acarretaram um interesse composto. Os seguros marítimos eram algo habitual e a álgebra das rendas vitalícia começava sua caminhada; esse tipo de operação requeria algo mais que o raciocínio intuitivo e comercial das primeiras seguradoras. Um dos pilares da Ciência Atuarial foi a teoria das probabilidades; as bases da análise estatística nos seguros foram estabelecidas por Pascal em 1654, em colaboração com o também matemático Pierre de Fermat. Outro pilar foi o conceito de tábua de vida, baseado em investigações sobre a mortalidade. As primeiras tábuas são devidas a John Graunt (1662). Em 1693, Edmund Halley, matemático inglês, publicou um famoso documento descrevendo a construção de tábuas de vida completas, a partir da hipótese de estacionariedade da população. As tábuas de Halley foram utilizadas pela grande maioria das companhias de seguros inglesas criadas durante o século XVII. Atualmente, a Ciência

Atuarial se enriquece com as aplicações da matemática nos seguros de não-vida, da teoria estatística e da moderna teoria da decisão.

Neste trabalho utilizamos três ferramentas estatísticas na construção de modelos de precificação para uma carteira de automóveis: Árvore de Decisão, Análise de Regressão Logística e Regressão Gama. Para a aplicação das técnicas, utilizamos uma base de dados reais fornecida por uma seguradora brasileira.

Esta dissertação foi organizada do seguinte modo: no Capítulo II, apresentamos a descrição do estudo, a seleção da amostra e o estudo das variáveis; o Capítulo III descreve as técnicas estatísticas utilizadas; no Capítulo IV, ajustamos os modelos, avaliamos e interpretamos os resultados e finalmente, no Capítulo V, apresentamos nossas conclusões e sugestões para trabalhos futuros.

2 CAPÍTULO II

2.1 Descrição do Estudo

No Brasil, um a cada quatro veículos é assegurado principalmente para as garantias de roubo e furto. O seguro de automóveis é composto por grandes portfólios, que sofrem concorrência acirrada das grandes seguradoras.

Nos últimos anos muitas mudanças ocorreram tornando este mercado ainda mais competitivo. Dentre elas vale mencionar a abertura do mercado brasileiro às seguradoras estrangeiras, a associação das seguradoras com grandes bancos comerciais, o crescimento dos canais diretos de comercialização e a utilização do perfil na precificação.

O motivo principal de um estudo de tarifa é identificar segmentos rentáveis em uma carteira. Uma boa tarifa controla melhor a diferença entre o prêmio cobrado e o custo real do seguro. Além disso, pode-se promover descontos no prêmio, de forma a promover um aumento dos segmentos rentáveis em uma carteira ou carregamentos para dissuadir o crescimento de segmentos não rentáveis.

O objetivo deste trabalho é estimar a Freqüência de sinistros e o Custo de sinistro de uma carteira de automóveis, por meio de modelos estatísticos baseados na experiência passada. A multiplicação da "Freqüência de sinistros" pelo "Custo de sinistro" fornece o Prêmio de risco, ou seja, o valor de prêmio necessário para pagar todas as indenizações.

2.2 Terminologia do Seguro

A atividade seguradora, como qualquer outra, se estabelece como uma especialidade e tem sua própria forma de se expressar. Veremos agora uma série de termos de uso freqüente:

- **Seguro:** entendido como um contrato de convênio entre duas partes - a companhia ou entidade seguradora de um lado, e o tomador ou contratante do outro - mediante a qual a primeira se compromete a ressarcir economicamente a perda ou dano que o segurado possa

sofrer durante a vigência do contrato. A obrigação do segurado é pagar o preço do seguro total ou parcial na assinatura do contrato.

• **Risco:** é a possibilidade de perda ou dano. O homem, desde que nasce, vive sob a constante ameaça de doença, acidente, morte etc. Da mesma forma, suas propriedades podem vir a sofrer incêndios, roubos etc.

• **Sinistro:** é a concretização do risco. Por exemplo, um incêndio que destrói uma fábrica, o roubo de lojas, morte em um acidente, colisão de um veículo etc.

• **Seguradora:** é a pessoa jurídica que assume o compromisso de oferecer indenização quando ocorre um sinistro. Para que uma empresa possa operar legalmente como seguradora, esta deve ter uma autorização da Superintendência de Seguros Privados (SUSEP). A SUSEP é o órgão responsável pela regulamentação de toda atividade seguradora no Brasil.

• **Tomador:** é a pessoa física ou jurídica que assina o contrato e paga seu preço.

• **Segurado:** é a pessoa titular do interesse segurado. É quem sofre o prejuízo econômico em seus bens em caso de ocorrência do sinistro ou a pessoa cuja vida ou integridade física é segurada e, portanto, quem receberá a indenização no caso de um sinistro afetar o segurado (exceto no caso de seguros de vida, em que a indenização em caso de morte é recebida pelo beneficiário). O tomador e o segurado podem ser a mesma pessoa ou pessoas distintas.

• **Beneficiário:** quando se segura a vida, a integridade física de uma pessoa pode ser designada a outra pessoa, para que receba as indenizações.

• **Apólice:** é o documento em que se estabelece o contrato de seguro. Tem duas características importantes:

- A prova de que o contrato existe;
- As normas que regulam a relação entre os contratantes.

• **Prêmio:** é o preço do seguro. É a quantidade de dinheiro que o tomador paga para que a seguradora indenize o segurado em um eventual sinistro. O prêmio é em geral por uma vigência anual do seguro.

• **Prêmio de risco:** também chamado de prêmio puro, matemático ou estatístico. É a quantidade necessária e suficiente que a seguradora deve cobrar para cobrir os riscos. Nasce do conceito de esperança matemática como preço justo de uma eventualidade.

• **Prêmio de tarifa:** também chamado prêmio comercial, é o prêmio de risco mais os encargos (gastos administrativos comissões e reservas técnicas).

• **Prêmio de venda:** é o prêmio de tarifa mais os impostos.

• **Sinistralidade:** é um coeficiente (sinistro/prêmio) que mede a relação entre os sinistros pagos em um determinado período e os prêmios auferidos neste mesmo período.

• **Exposição:** é a proporção de dias em que o veículo fica exposto ao risco durante a vigência do contrato. Forma de cálculo: Exposição = (nº de dias vigentes)/365.

• **Freqüência de sinistros:** é a razão entre a quantidade de sinistros e a exposição de um particular segmento ou perfil.

• **Percentual de sinistros:** é a razão entre a quantidade de sinistros e o número de segurados de um determinado perfil ou segmento.

• **Classe de bônus:** é a forma de classificar o segurado quanto ao número de suas renovações na companhia sem a ocorrência de sinistro. A cada classe de bônus é concedido um desconto no prêmio.

2.3 Base de Dados

Para realização deste estudo tem-se à disposição uma base de dados com 1.524.511 apólices de segurados, que contrataram o seguro de automóvel no período de janeiro de 2004 a dezembro de 2004. Trata-se de uma base com informações reais de sinistro e prêmio para a garantia de colisão fornecida por uma seguradora brasileira.

O motivo de utilizar-se uma base tão grande deve-se ao fato do evento sinistro ser relativamente raro, com o percentual de sinistro na carteira girando em torno de 5,2%. Logo, existe a preocupação de termos um número de eventos relativamente grande, de forma que possamos segmentar a base em busca de padrões de comportamento nos segurados. Dessa forma, podemos assegurar que o procedimento de precificação utilizado esteja bem ajustado. Além disso, a freqüência de sinistros de colisão tem um comportamento sazonal e nesse caso optamos por selecionar as apólices durante um ano, para que o modelo ajustado absorvesse essas variações mensais.

Cada linha de nosso banco de dados representa um segurado e em cada coluna são armazenadas informações sobre o mesmo. Existem três tipos de informações: variáveis com as características do veículo, variáveis do comportamento do segurado e variáveis do perfil do

motorista. Quanto ao tipo de variáveis, nossa base de dados contempla as categóricas e as contínuas.

2.4 Descrição das Variáveis

Uma vez descrita nossa base de dados, bem como sua estrutura e tipos de variáveis, agora iremos explorar mais a fundo cada um dos campos da base. A seguir apresentaremos as variáveis utilizadas no estudo, sua descrição e o conteúdo dos campos.

- ANO - Ano / Modelo do veículo;
  - ✓ 1 = 1934
  - ✓ 2 = 1949
  - ✓ 3 = 1956
  - ✓ 4 = 1962
  - ✓ 5 = 1968
  - ✓ 6 = 1970
  - ✓ 7 = 1972
  - ✓ .
  - ✓ .
  - ✓ .
  - ✓ 40 = 1997
  - ✓ 41 = 1998
  - ✓ 42 = 1999
  - ✓ 43 = 2000
  - ✓ 44 = 2001
  - ✓ 45 = 2002
  - ✓ 46 = 2003
  - ✓ 47 = 2004
  - ✓ 48 = 2005

- GAR - Garantia contratada;
  - ✓ 1 = COLISÃO
  - ✓ 2 = ROUBO E FURTO
  - ✓ 3 = DANOS MATERIAIS A TERCEIROS

- DTSIN – Data de ocorrência do sinistro (dd/mm/aa);

- SINISTRO – indicadora de sinistro;
  - ✓ 1 = OCORREU SINISTRO
  - ✓ 0 = NÃO OCORREU SINISTRO

- VLRSIN – Valor dos sinistros (em reais);

- INIVIG – Data de início de vigência do seguro (dd/mm/aa);

- FIMVIG – Data de final de vigência do seguro (dd/mm/aa);

- IDADE – Idade do principal condutor;
  - ✓ 1 = MENOS DE 21 ANOS
  - ✓ 2 = DE 21 A 24 ANOS
  - ✓ 3 = DE 25 A 28 ANOS
  - ✓ 4 = DE 29 A 33 ANOS
  - ✓ 5 = DE 34 A 39 ANOS
  - ✓ 6 = DE 40 A 44 ANOS
  - ✓ 7 = DE 45 A 49 ANOS
  - ✓ 8 = DE 50 A 54 ANOS
  - ✓ 9 = DE 55 A 59 ANOS
  - ✓ 10 = MAIS DE 59 ANOS

- PREMIO – Prêmio pago pelo segurado (em reais);

- IS – Importância segurada (em reais);

- FABR – Fabricante do veículo;
  - ✓ 2 = FABRICANTE 1
  - ✓ 3 = FABRICANTE 2
  - ✓ .
  - ✓ .
  - ✓ .
  - ✓ 123 = FABRICANTE 37
  - ✓ 128 = FABRICANTE 38

- BONUS – Classe de bônus;
  - ✓ 0 = NOVOS CLIENTES
  - ✓ 1 = PRIMEIRA RENOVAÇÂO
  - ✓ .
  - ✓ .
  - ✓ .
  - ✓ 19 = DÉCIMA NONA RENOVAÇÃO
  - ✓ 20 = VIGÉSIMA RENOVAÇÃO

- PROCED – Procedência do veículo;
  - ✓ 1 = NACIONAL
  - ✓ 2 = IMPORTADO

- ORIG – Origem da apólice;
  - ✓ 1 = EMISSÃO NOVA
  - ✓ 2 = RENOVAÇÃO DE APÓLICES DE CONGÊNERES
  - ✓ 4 = RENOVAÇÃO INTERNA 1
  - ✓ 5 = RENOVAÇÃO INTERNA 2
  - ✓ 6 = RENOVAÇÃO INTERNA 3
  - ✓ 7 = RENOVAÇÃO INTERNA 4
  - ✓ 8 = RENOVAÇÃO INTERNA 5

- CIVIL – Estado civil do principal condutor;
  - ✓ 1 = CASADO
  - ✓ 2 = DESQUITADO
  - ✓ 3 = DIVORCIADO
  - ✓ 4 = OUTRO
  - ✓ 5 = SEPARADO
  - ✓ 6 = SOLTEIRO
  - ✓ 7 = UNIÃO ESTÁVEL
  - ✓ 8 = VIÚVO

- FURTO – Sistema antifurto instalado no veículo;
  - ✓ 1 = NÃO
  - ✓ 2 = NÃO DEFINIDO
  - ✓ 3 = SIM

- TIPVEI – Tipo de veículo;
  - ✓ 1 = POPULAR
  - ✓ 2 = ESPORTIVO
  - ✓ 3 = MÉDIO
  - ✓ 4 = LUXO
  - ✓ 5 = UTILITÁRIO LEVE
  - ✓ 6 = UTILITÁRIO PESADO

- SEXO – Sexo do condutor;
  - ✓ 1 = MASCULINO
  - ✓ 2 = FEMININO

- DAF – Rastreador de veículo instalado;
  - 1 = SIM
  - 2 = NÃO

- ❖ PIN – Tipo de pintura do veículo;
  - 1 = COMUM
  - 2 = METÁLICA
  - 3 = PEROLIZADA

- ❖ EXPO – Exposição.

2.5 Variável resposta

Uma vez definida a base de dados e listadas todas as variáveis do estudo, iremos agora definir a variável resposta de acordo com os objetivos do estudo. Para estimar a freqüência de sinistros, utilizaremos um modelo de Regressão Logística (Hosmer e Lemeshow, 1989). Nesse tipo de modelo a variável resposta é binária (1= "sinistro" ou 0= "não sinistro") e as variáveis explicativas podem ser categóricas, contínuas ou de ambos os tipos. Com base nessa informação, nossa variável resposta será a seguinte:

- ❖ SINISTRO (0=se não houve sinistro; 1=se houve sinistro).

Já o Custo de sinistro será estimado utilizando-se um Modelo Linear Generalizado – MLG (McCullagh e Nelder, 1989 e Paula, 2004). Os MGLs são procedimentos estatísticos utilizados para medir o efeito das variáveis explicativas, categóricas ou contínuas, em relação a uma variável resposta contínua ou discreta pertencente à família exponencial de distribuições. Esses procedimentos permitem flexibilidade no desenho do modelo, multiplicativo, aditivo ou planos mistos. Pode-se também utilizar associação entre fatores e interações entre as variáveis. Além disso, diferentes distribuições do erro podem ser utilizadas (Normal, Gama, Binomial, Poisson, etc). Uma outra variável resposta de interesse no nosso estudo será:

- ❖ VLRSIN – Custo de sinistro por apólice (em reais).

2.6 Variáveis explicativas

Definida qual será a variável resposta, nossa atenção se volta agora para as variáveis explicativas. Quando observamos as informações que as seguradoras possuem em sua carteira de automóvel, percebemos que grande parte das variáveis sobre seus segurados são categóricas (ex: estado civil, sexo, tipo de veículo). Raras exceções como a variável valor de sinistros e idade do motorista principal são contínuas.

Em seguros, a interação entre as variáveis é um fator importante para determinar o perfil de um segurado, e esse assunto será mais bem abordado adiante. Apesar das variáveis na sua quase totalidade já estarem representadas em categorias, muitas delas possuem um número excessivo de categorias, ou categorias com um número muito pequeno de eventos.

A alternativa encontrada para solucionar as questões levantadas é categorizar as variáveis quantitativas e promover uma nova categorização das variáveis qualitativas. Por hora, vamos deixar as explicações mais detalhadas para serem discutidas no Capítulo IV.

2.7 Amostra de Modelagem e Validação

Utilizaremos nossa base original, com as 1.524.511 apólices, para realizar a modelagem dos dados. Para realizar a validação dos modelos, iremos utilizar uma base de dados com apólices emitidas no primeiro semestre de 2005. A amostra de validação servirá para verificar se os modelos estimados mantêm o poder de discriminação. A variação do poder de discriminação é um indicativo de que o modelo não é estável ou pode estar ocorrendo uma superestimação.

3 CAPÍTULO III

3.1 Metodologia

O objetivo deste capítulo é descrever as técnicas estatísticas utilizadas no estudo. O motivo da escolha de cada técnica já foi abordado no capítulo anterior e se deve basicamente ao tipo de variável alvo e ao objetivo do estudo. No entanto, vale gastar algumas palavras para discutir a utilização da técnica Árvore de Decisão em conjunto com a Regressão Logística e a Regressão Gama. Ao utilizarmos uma árvore de decisão antes de rodar um dos modelos anteriores, estamos implicitamente capturando interações naturais entre as variáveis e, conseqüentemente, temos um ganho enorme de força bruta e tempo. Além disso, temos uma indicação das variáveis mais importantes para explicar a variação da variável resposta. Logo, a utilização conjunta destas técnicas traz um ganho significativo na qualidade de ajuste do modelo.

3.2 Árvore de Decisão

Árvores de decisão são meios de representar resultados de "Data Mining" na forma de árvore, que lembram um gráfico organizacional horizontal. A árvore de decisão consiste de uma hierarquia de nós internos e externos que são conectados por ramos. O nó interno, também conhecido como nó pai, é a unidade de tomada de decisão que avalia através de teste lógico qual será o próximo nó descendente ou filho. Em contraste, um nó externo (não tem nó descendente), também conhecido como nó filho, está associado a um rótulo ou a um valor.

Em geral, o procedimento de uma árvore de decisão é o seguinte: apresenta-se um conjunto de dados ao nó inicial (ou nó raiz, que também é um nó interno) da árvore; dependendo do resultado do teste lógico usado pelo nó, a árvore ramifica-se para um dos nós filhos e esse procedimento é repetido até que um nó terminal é alcançado. A repetição deste procedimento caracteriza a recursividade da árvore de decisão (Figura 1).

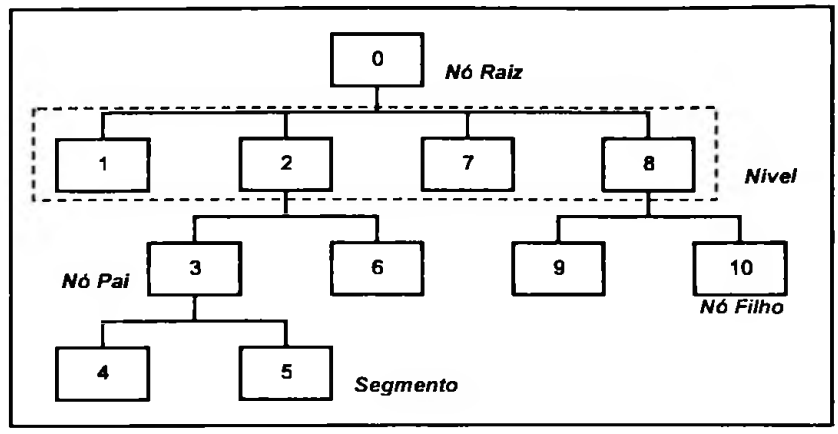

Figura 1 - Estrutura de uma árvore de decisão.

Existem diversos algoritmos para fazer a classificação ou segmentação. No entanto, nós utilizaremos o algoritmo CHAID ("Chi-squared Automatic Interaction Detection"). O CHAID é um método estatístico baseado em árvores que examinam as relações entre muitas variáveis categóricas ou discretas e uma variável alvo categórica. Variáveis preditoras contínuas são discretizadas antes de começar a construção da árvore.

Para determinar a melhor divisão em qualquer nó, Kass (1980) agrupa qualquer par permitido de categorias da variável preditora, se não houver diferença estatística significante dentro do par com respeito à variável alvo. O processo se repete até que nenhum par não significante seja encontrado. O conjunto resultante de categorias do preditor é a melhor divisão com respeito a esse divisor. Esse processo segue para todos os preditores. O preditor que dá a melhor predição é selecionado, e o nó é dividido. O processo continua até que uma regra de parada seja alcançada.

3.2.1 "Chi-squared Automatic Interaction Detection" - CHAID

O CHAID é um dos métodos mais tradicionais de árvores de classificação propostos originalmente por Kass (1980). De acordo com Ripley (1996) o algoritmo CHAID é um descendente do THAID desenvolvido por Morgan e Messenger (1973). Ele foi construído para árvores categóricas, em que um nó pai dá origem a mais de dois nós filho, baseadas em

um algoritmo relativamente simples que seja particularmente rápido em análises de grandes bases de dados. O algoritmo também é muito eficaz quando temos que classificar uma variável com resposta categórica com muitos níveis, baseada em variáveis preditoras categóricas com muitas classes.

O CHAID utiliza o teste qui-quadrado para determinar se uma ramificação deve ou não perseguir e que variáveis independentes devem ser selecionadas. Daí vem o seu nome, Detecção Automática de Interações Aplicando o Teste de Qui-Quadrado ("Chi Squared Automatic Interaction Detection").

O algoritmo foi desenhado para identificar as interações a serem incluídas nos modelos de regressão. Manipula com facilidade as interações, quando outras técnicas de modelagem têm enormes dificuldades. As interações são combinações de variáveis independentes que influenciam o resultado.

3.2.1.1 O Algoritmo

O algoritmo divide em grupos os registros que possuem a mesma probabilidade de resultado, baseando-se nos valores das variáveis independentes. O algoritmo parte de um nó raiz e vai se ramificando em nós descendentes até chegar aos nós terminais, onde terminam as ramificações.

A ramificação ocorre em duas ou mais categorias e está determinada pelo teste χ^2. O teste é aplicado em uma tabela cruzada entre a variável resposta e cada uma das variáveis independentes. O resultado é um p-valor. Esse valor representa a probabilidade de detectar associação entre as variáveis, quando não existe, ou seja, é a probabilidade de rejeitar a hipótese nula (ausência de associação) quando esta hipótese é verdadeira. Assim, um p-valor pequeno é indício de que se deve fazer uma ramificação no nó correspondente.

Os p-valores calculados para cada tabela cruzada, com todas as variáveis independentes, são classificados, e se o menor deles estiver abaixo de um valor crítico determinado então se realiza uma ramificação do nó raiz para aquela variável independente. Se uma variável

explicativa tem mais de duas categorias, o CHAID compara e agrupa, segundo o teste χ^2, aquelas que não diferem significantemente.

O processo continua até a construção da árvore. Quanto maiores forem os ramos, menos variáveis independentes restarão, uma vez que o crescimento acontece justamente com as variáveis. O processo chega ao fim quando o menor p-valor não for inferior ao valor crítico determinado. Os nós filhos serão aqueles que não sofreram ramificações.

3.2.2 Exemplo de árvore de decisão

Suponha que estamos interessados em identificar quais segurados, de uma carteira de automóvel, tiveram sinistro de colisão. Dado que temos as seguintes informações sobre eles, procedência, sexo, idade, e fabricante, estamos interessados em descobrir quais subgrupos são mais propensos a sofrer uma colisão.

Então vejamos, temos uma amostra com uma variável resposta categórica que identifica quem sofreu ou não sinistro e quatro variáveis preditoras, sendo as três primeiras nominais (procedência, sexo e fabricante) e idade que é contínua e será agrupada em categorias ordenadas. Num primeiro instante o CHAID irá examinar as tabelas cruzadas de cada variável explicativa em relação a variável resposta, e testar a significância utilizando o teste de qui-quadrado. Se mais de uma dessas relações são significantes, o CHAID selecionará a variável explicativa que for mais significante (menor p-valor). Caso uma variável possua mais de duas categorias, por exemplo fabricante, o CHAID as compara e une aquelas categorias que não apresentam diferenças em relação a variável resposta. Suponha que a variável fabricante é a mais significante, logo a árvore principia sua divisão com essa variável (Figura 2).

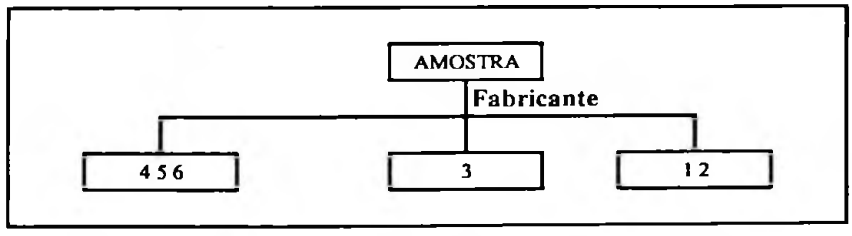

Figura 2 - Primeira divisão da árvore.

Note que algumas categorias foram unidas e a amostra dividiu-se em três subgrupos ou nós. No primeiro, os fabricantes 4, 5 e 6 não diferem significantemente quanto ao percentual de sinistros e consequentemente foram agrupados. O mesmo aconteceu com os fabricantes 1 e 2, que ficaram em um subgrupo a parte.

O CHAID então se volta par o primeiro nó (fabricantes 4, 5 e 6), e para essas observações ele examina as três variáveis restantes para descobrir qual resulta na maior diferença significante. A variável sexo apresentou a maior diferença, e consequentemente esse subgrupo foi dividido pelo sexo (Figura 3).

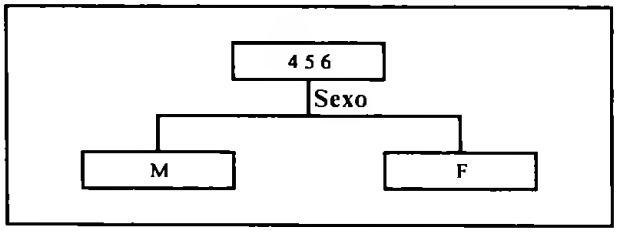

Figura 3 - Divisão do primeiro nó.

Em seguida, o algoritmo examina cada um dos outros subgrupos de fabricantes tentando dividir cada um deles pela variável preditora mais significante.

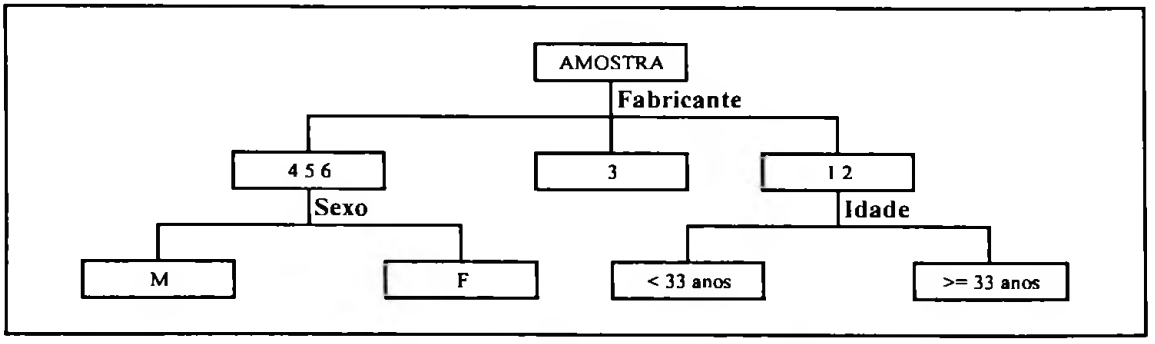

Figura 4 - Divisão dos três subgrupos da variável fabricante.

A Figura 4 mostra que o segundo nó (fabricante 3) não cresceu e tornou-se um nó terminal, devido aos critérios de parada e a variável idade dividiu o terceiro nó. As seguintes condições representam um critério de parada:

- o nível máximo permitido foi alcançado;
- não há uma variável explicativa significante para dividir o nó;

- O número de casos no segmento é menor que o número mínimo de casos para um nó pai;
- Se o nó for dividido, o número de casos em um ou mais nós filhos seria menor que o número mínimo de casos para nós filhos.

O CHAID então desce mais um nível na árvore e toma o primeiro subgrupo de sexo (masculino) dentro do nó 1 e verifica se algum dos dois preditores restantes resultam numa diferença significante no resultado. Se isso não acontecer ou se uma das regras de parada for alcançada, o CHAID declara que esse nó é um segmento e passa a examinar o subgrupo das mulheres da mesma maneira. Assim, nível a nível, o algoritmo, sistematicamente divide os dados em subgrupos (chamados de nós) que exibem diferenças em relação a variável resposta. A Figura 5 mostra como ficou organizada a árvore final.

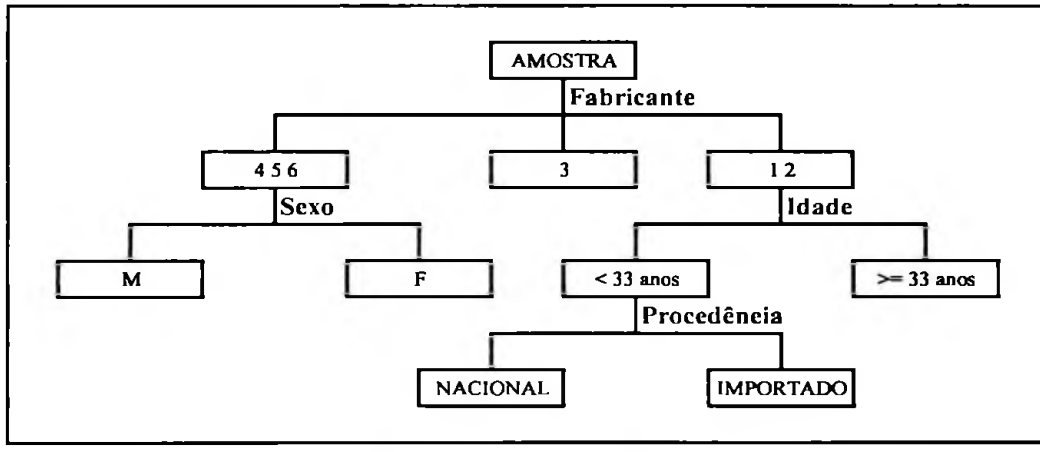

Figura 5 - Árvore final.

3.3 Regressão Logística

Há muito tempo a regressão logística vem sendo largamente utilizada na área médica, e nesses últimos anos o seu uso também se estendeu para o mercado financeiro, na área de crédito e de seguros de grandes instituições. Seu principal objetivo consiste em investigar a relação entre uma variável categórica binária e um grupo de variáveis explicativas, categóricas ou contínuas. Neste capítulo estaremos preocupados em descrever os principais conceitos dessa técnica estatística, sem preocupar-nos com aspectos mais profundos. Para

aqueles que desejarem se aprofundar um pouco mais no tema, nós sugerimos os seguintes textos: Agresti (1990), Cox e Snell (1989) e Hosmer e Lemeshow (1989).

Considere uma variável resposta binária Y com dois resultados possíveis 0 ou 1 (Y=1 se ocorre um determinado evento e Y=0 se não ocorre um determinado evento). E suponha que $\mathbf{x} = (1, x_1, x_2, ..., x_p)^T$ seja um vetor de valores de variáveis explicativas independentes. O interesse se concentra em descrever o efeito das variáveis explicativas sobre a variável resposta que segue uma distribuição binomial (vide, por exemplo, Bussab e Morettin, 2002) com probabilidade $\pi(\mathbf{x})$ dada por:

$$\pi(\mathbf{x}) = P(Y = 1 \mid \mathbf{x}) \text{ e } 1 - \pi(\mathbf{x}) = P(Y = 0 \mid \mathbf{x})$$

com esperança igual a:

$$E(Y \mid \mathbf{x}) = 1\pi(\mathbf{x}) + 0[1 - \pi(\mathbf{x})] = \pi(\mathbf{x})$$

e variância igual a:

$$Var(Y \mid \mathbf{x}) = E(Y^2 \mid \mathbf{x}) - [E(Y \mid \mathbf{x})]^2 = \pi(\mathbf{x})[1 - \pi(\mathbf{x})].$$

3.3.1 Função Logística

Dadas todas as características mencionadas anteriormente uma função sugerida para descrever a probabilidade $\pi(x)$ é a função logística dada por:

$$\pi(x) = \frac{e^{\eta}}{1 + e^{\eta}},$$

em que $\eta = \beta_0 + \beta_1 x$ é um preditor linear e β_0 e β_1 são parâmetros a serem estimados. Temos então a seguinte forma para $\pi(x)$:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}.$$

Como podemos observar na Figura 6 a função logística tem como características um formato em "S", seus valores variam sempre no intervalo (0, 1), é simétrica em torno de $\eta = 0$ $(\pi(x) = 0{,}5)$ e é sempre monótona (crescente ou decrescente) se x é contínua.

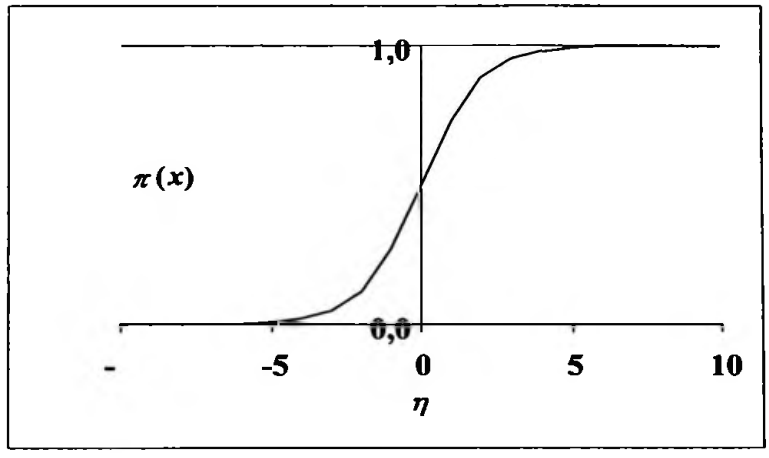

Figura 6 - Representação gráfica da função logística.

3.3.2 Regressão Logística Simples

No modelo logístico simples $\pi(x)$ denota a probabilidade de ocorrer um evento (Y=1) dado o valor x de uma variável explicativa X, sendo definida por

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}.$$

Esse modelo poderia, por exemplo, ser aplicado para analisar a associação entre a ocorrência ou não de sinistro e a ocorrência ou não de um fator particular. Se X=1 denota a ocorrência do fator e X=0 a ausência do fator, então

$$\pi(1) = \frac{e^{\alpha+\beta}}{1+e^{\alpha+\beta}}$$

e

$$\pi(0) = \frac{e^{\alpha}}{1+e^{\alpha}}.$$

Portanto, $\pi(1)$ denota a probabilidade de ocorrer sinistro no grupo em que há presença do fator e $\pi(0)$ denota a probabilidade de sinistro no grupo em que há ausência do fator. Será assumido para cada indivíduo amostrado do primeiro grupo que Y~Bernoulli($\pi(1)$) (vide, por exemplo, Bussab e Morettin, 2002) e para cada indivíduo amostrado do segundo grupo que Y~Bernoulli($\pi(0)$).

3.3.2.1 Razão de chances ("Odds Ratio")

A razão entre uma probabilidade e a probabilidade complementar é denominada chance (" Odds") e indica quantas vezes a ocorrência de um determinado evento é mais provável do que a não ocorrência. No caso da regressão logística simples, a chance de sinistro no grupo com presença do fator (X=1) é definida por

$$\frac{\pi(1)}{1-\pi(1)} = e^{\alpha+\beta}$$

e, similarmente, a chance de sinistro no grupo com ausência do fator (X=0) fica dada por

$$\frac{\pi(0)}{1-\pi(0)} = e^{\alpha}.$$

Assim, a razão de chances entre um indivíduo do grupo com presença do fator e um indivíduo do grupo com ausência do fator, é definida por

$$\varphi = \frac{\dfrac{\pi(1)}{1-\pi(1)}}{\dfrac{\pi(0)}{1-\pi(0)}} = \frac{\pi(1)\{1-\pi(0)\}}{\pi(0)\{1-\pi(1)\}} = \frac{e^{\alpha+\beta}}{e^{\alpha}} = e^{\beta}.$$

Portanto, obtemos $\varphi = e^{\beta}$. Por exemplo se $\pi(1)$=0,75 e $\pi(0)$=0,5 então $\pi(1)\{1-\pi(0)\}$=0,75.0,50 e $\pi(0)\{1-\pi(1)\}$=0,50.0,25. A razão de chances fica dada por

$$\varphi = \frac{0{,}75.0{,}50}{0{,}50.0{,}25} = 3 \quad (\beta = \ln(3)).$$

Logo, um indivíduo do grupo com a presença do fator tem 3 vezes a chance de sinistro do que um indivíduo do grupo com ausência do fator.

Se X for uma variável contínua a interpretação é um pouco diferente. Vamos supor um valor observado x para a variável X, sendo que este valor foi acrescido de uma unidade, então:

$$\pi(x+1) = \frac{e^{\alpha+\beta(x+1)}}{1+e^{\alpha+\beta(x+1)}}.$$

Logo,

$$\frac{\pi(x+1)}{1-\pi(x+1)} = e^{\alpha+\beta(x+1)} = e^{\alpha}e^{\beta}e^{\beta x}.$$

Similarmente, temos que

$$\frac{\pi(x)}{1-\pi(x)} = e^{\alpha+\beta x} = e^{\alpha}e^{\beta x}.$$

Portanto, a razão de chances entre um indivíduo com X=x+1 e um indivíduo com X=x fica dada por

$$\varphi = \frac{\dfrac{\pi(x+1)}{1-\pi(x+1)}}{\dfrac{\pi(x)}{1-\pi(x)}} = \frac{e^{\alpha}e^{\beta}e^{\beta x}}{e^{\alpha}e^{\beta x}} = e^{\beta}.$$

Ou seja, acrescendo em uma unidade o valor da variável X a razão de chances fica novamente dada por e^{β}.

3.3.2.2 Estimação dos parâmetros

O método mais eficiente para estimar os parâmetros do modelo de regressão logística é o da máxima verossimilhança. O primeiro passo a ser dado é construir a função de verossimilhança $l(\boldsymbol{\theta})$ para o modelo estimado. A função de verossimilhança representa a probabilidade de reproduzir os dados da amostra para o modelo, supondo $\boldsymbol{\theta} = (\beta_0, \beta_1)^T$ desconhecido, em que $\boldsymbol{\theta}$ é o vetor de parâmetros.

Vamos supor que Y ~ Bernoulli $(\pi(\mathbf{x}))$ em que $\mathbf{x}$ é o valor observado das variáveis explicativas. Assim,

$$\pi(\mathbf{x}) = P(Y = 1 \mid \mathbf{x}) \text{ e } 1 - \pi(\mathbf{x}) = P(Y = 0 \mid \mathbf{x}),$$

em que $\mathbf{x} = (1, \mathbf{x})^T$. Para uma amostra de n indivíduos independentes, assumimos para o i-ésimo indivíduo que $Y_i \sim$ Bernoulli $(\pi(\mathbf{x}_i))$, então

$$P(Y_i = y_i \mid \mathbf{x}_i) = \pi(\mathbf{x}_i)^{y_i}\left[1 - \pi(\mathbf{x}_i)\right]^{1-y_i}.$$

Assim, a função de verossimilhança conjunta será dada pelo produto das probabilidades

$$l(\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi(\mathbf{x}_i)^{y_i}\left[1 - \pi(\mathbf{x}_i)\right]^{1-y_i}.$$

A forma como a função de verossimilhança está representada torna o seu cálculo muito dispendioso de forma que iremos calcular o logaritmo natural, uma vez que este transforma o produto em uma soma de probabilidades. A estimação dos parâmetros é obtida maximizando-se a função:

$$L(\boldsymbol{\theta}) = \ln(\boldsymbol{\theta}) = \sum_{i=1}^{n} \{y_i \ln[\pi(\mathbf{x}_i)] + (1 - y_i) \ln[1 - \pi(\mathbf{x}_i)]\}.$$

Os valores de $\boldsymbol{\theta}$ que maximizam $L(\boldsymbol{\theta})$ são obtidos derivando-se $L(\boldsymbol{\theta})$ em relação aos parâmetros β_0 e β_1. Logo, será necessário calcular a primeira derivada parcial do logaritmo da função de verossimilhança com relação a cada parâmetro e igualar a zero, ou seja,

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \beta_0} = \sum_{i=1}^{n} [y_i - \pi(\mathbf{x}_i)] = 0$$

e

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \beta_1} = \sum_{i=1}^{n} [y_i - \pi(\mathbf{x}_i)] x_i = 0.$$

A solução deste sistema de equações não-lineares com os parâmetros β_0 e β_1 é obtida através de um processo iterativo baseado no método de Newton-Raphson (McCullagh e Nelder, 1989). As estimativas de máxima verossimilhança dos parâmetros obtidas são denotadas por $\hat{\boldsymbol{\theta}}$.

Uma conseqüência interessante derivada da equação $\sum_{i=1}^{n} [y_i - \pi(\mathbf{x}_i)] = 0$ é que:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(\mathbf{x}_i),$$

ou seja, a soma dos valores observados é igual à soma dos valores preditos.

O valor da função de verossimilhança $l(\hat{\theta})$ varia entre 0 e 1, logo, o logaritmo da função de verossimilhança $L(\hat{\theta})$ será um número negativo e alcançará o valor 0 em um modelo hipotético em que se reproduziram os dados de forma exata, tal como se pode observar na Figura 7.

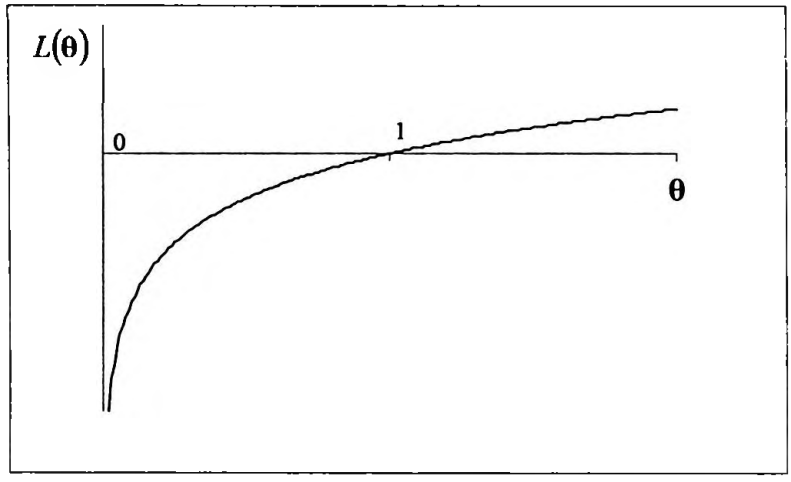

Figura 7 - Gráfico do logaritmo natural da função de verossimilhança.

Alguns softwares soltam o valor $L(\hat{\theta})$, que é negativo. Por exemplo o SAS apresenta na saída do "proc logistic" o valor positivo -2 $L(\hat{\theta})$ que servirá para a comparação de modelos.

3.3.3 Regressão Logística Múltipla

Considere a coleção de valores de variáveis explicativas denotados pelo vetor $\mathbf{x} = (1, x_1, x_2, ..., x_p)^T$. Seja $\pi(\mathbf{x})$ a probabilidade de Y assumir o valor 1 dado **X=x,**

$$\pi(\mathbf{x}) = P(Y = 1 \mid \mathbf{x}).$$

O modelo de regressão logística múltiplo assume que Y ~ Bernoulli ($\pi(\mathbf{x})$), sendo

$$\pi(\mathbf{x}) = \frac{e^{\eta}}{1+e^{\eta}},$$

com $\eta = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$ sendo o preditor linear.

3.3.3.1 Estimação dos parâmetros do modelo

Suponha uma amostra de n observações independentes do par (x_i, y_i), i=1,2, ... , n. Assim, como foi descrito para o modelo de regressão logística simples, o método de estimação dos parâmetros do modelo será o de máxima verossimilhança, com o qual obteremos a estimação do vetor $\boldsymbol{\theta} = (\beta_0, \beta_1, ..., \beta_p)^T$. O número de equações de verossimilhança será igual a p+1, as soluções são obtidas calculando-se as derivadas parciais em relação aos p+1 coeficientes do logaritmo da função de verossimilhança e aplicando-se o método de Newton-Raphson. As equações de verossimilhança são expressas como:

$$\sum_{i=1}^{n}[y_i - \pi(\mathbf{x}_i)] = 0$$

$$\sum_{i=1}^{n}[y_i - \pi(\mathbf{x}_i)]x_{i1} = 0$$

.

.

.

$$\sum_{i=1}^{n}[y_i - \pi(\mathbf{x}_i)]x_{ip} = 0.$$

Em forma matricial podemos expressar essas equações como,

$$\mathbf{X}^T(\mathbf{y} - \mathbf{m}) = \mathbf{0},$$

em que

$$\mathbf{y} = (y_1, ..., y_n)^T,$$

$$\mathbf{m} = \left(\pi(x_1),...,\pi(x_n)\right)^T \text{ e}$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

é a matriz modelo. Aplicando-se o método de Newton-Raphson (McCullagh e Nelder, 1989; Paula, 2004) chega-se ao processo iterativo para estimar $\boldsymbol{\theta} = \left(\beta_0, \beta_1,..., \beta_p\right)^T$:

$$\boldsymbol{\theta}^{(r+1)} = \left(\mathbf{X}^T\mathbf{V}^{(r)}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{(r)}\mathbf{z}^{(r)}, \text{ r=0, 1,2, ...,}$$

em que $\mathbf{z} = \left(z_1,..., z_n\right)^T$ com

$$z_i = \eta_i + \left[y_i - \pi(\mathbf{x}_i)\right]/\pi(\mathbf{x}_i)\{1-\pi(\mathbf{x}_i)\} \text{ e}$$

$$\mathbf{V} = diag\left[V_1,..., V_n\right], \text{ com}$$

$$V_i = \pi(\mathbf{x}_i)\{1-\pi(\mathbf{x}_i)\}.$$

3.3.3.2 Ajuste do modelo

Em um modelo de regressão logística estimado pelo método de máxima verossimilhança o ajuste global se faz com estatísticas derivadas do modelo de verossimilhança. Ou seja, se trata de comparar a verossimilhança do modelo que contém apenas o intercepto, com a verossimilhança do modelo que contém as variáveis preditoras ou explicativas. Este valor poderá ser usado como referência para a qualidade de ajuste do modelo, sendo que quanto mais próximo estiver de 0, melhor será o ajuste do modelo.

3.3.3.2.1 Qui-Quadrado de Pearson

O Qui-Quadrado de Pearson é uma estatística da qualidade de ajuste que compara os valores observados com os preditos pelo modelo, segundo a expressão:

$$\chi^2 = \sum_{j=1}^{n} \frac{\{y_i - \hat{\pi}(\mathbf{x}_i)\}^2}{\hat{\pi}(\mathbf{x}_i)\{1-\hat{\pi}(\mathbf{x}_i)\}}.$$

Existem também dois índices que representam a proporção de incerteza dos dados que é explicada pelo modelo ajustado e que são análogos ao coeficiente de determinação da regressão linear.

O índice de Cox e Snell compara a função de verossimilhança do modelo que contém apenas o intercepto $l(\beta_0)$ com a função de verossimilhança do modelo considerado $l(\boldsymbol{\beta})$:

$$R^2 = 1 - \left[\frac{l(\beta_0)}{l(\boldsymbol{\beta})}\right]^{2/n}.$$

No entanto, o índice de Cox e Snell parece não ser adequado, pois ele não atinge o valor 1 quando o modelo reproduz exatamente os dados, ou seja

$$\text{Se } l(\boldsymbol{\beta}) = 1 \rightarrow R^2_{\max} = 1 - [l(\beta_0)]^{2/n}.$$

Por essa razão Nagelkerke (1991) propôs o índice corrigido que vale 1 no caso em que o modelo explique 100% da incerteza dos dados, definido por

$$R^2_c = \frac{R^2}{R^2_{\max}}.$$

3.3.3.2.2 Teste da qualidade do ajuste de Hosmer e Lemeshow

Como a estatística χ^2 não segue a distribuição qui-quadrado para grandes amostras, dificultando a verificação da qualidade do ajuste, Hosmer e Lemeshow (1989) sugerem uma estatística alternativa que consiste em ordenar os n indivíduos segundo as probabilidades preditas e dividí-los em aproximadamente g=10 grupos com aproximadamente o mesmo tamanho. Essa estatística assume a forma

$$\hat{C} = \sum_{j=1}^{g} \frac{(O_j - n_j^{\bullet}\overline{\pi}_j)^2}{n_j^{\bullet}\overline{\pi}_j(1-\overline{\pi}_j)},$$

em que O_j denota o número de sucessos no j-ésimo grupo formado enquanto $n_j^{\bullet}$ e $\overline{\pi}_j$ representam o tamanho e a probabilidade predita média desse grupo, respectivamente. A estatística $\hat{C}$, supondo que o modelo é correto, segue distribuição qui-quadrado com g-2 graus de liberdade para amostras grandes.

3.3.4 Significância dos parâmetros do modelo

Uma vez estimados os parâmetros teremos que verificar quais deles são significativamente diferentes de zero. Utilizamos o teste de Wald para verificar se cada um deles é igual a zero. Se o objetivo é comprovar que o conjunto de covariáveis elegidas explica o fenômeno em estudo, ou seja, verificar se todos os coeficientes estimados são iguais a zero ou existe pelo menos algum deles com valor distinto, podemos utilizar a razão de verossimilhanças ou outro teste ou procedimento estatístico.

3.3.4.1 Teste da razão de verossimilhanças

O teste da razão de verossimilhanças consiste em construir um teste baseado na função de verossimilhança, para o qual se calcula qual é a distância que existe entre o modelo ajustado

com as variáveis preditoras incluídas no modelo $l(\boldsymbol{\beta})$ e o modelo que contém parte dos parâmetros $l(\boldsymbol{\beta}_1)$, segundo a seguinte expressão:

$$D = -2\ln\left[\frac{l(\boldsymbol{\beta}_1)}{l(\boldsymbol{\beta})}\right] = -2\ln l(\boldsymbol{\beta}_1) - [-2\ln l(\boldsymbol{\beta})].$$

A estatística D segue uma distribuição qui-quadrado para amostras grandes com graus de liberdade igual à diferença de parâmetros entre os dois modelos. Por exemplo, suponha que o preditor linear seja expresso na forma matricial

$$\boldsymbol{\eta} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2,$$

em que $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_2$ são vetores paramétricos de dimensões q_1 e q_2, respectivamente, enquanto $\mathbf{X}_1$ e $\mathbf{X}_2$ são as matrizes modelo. Se desejarmos testar $\mathrm{H}: \boldsymbol{\beta}_2 = \mathbf{0}$ contra $\mathrm{A}: \boldsymbol{\beta}_2 \neq \mathbf{0}$ então a estatística $D = -2\ln[l(\boldsymbol{\beta}_1)/l(\boldsymbol{\beta}_2)]$ segue sob H distribuição qui-quadrado com q_1 graus de liberdade. Para ilustrar, suponha o modelo logístico

$$\pi(\mathbf{x}) = \frac{e^{\beta_0+\beta_1 x_1+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_p x_p}}.$$

Se queremos testar $\mathrm{H}: \beta_2 = 0$ contra $\mathrm{A}: \beta_2 \neq 0$ então a estatística D deverá seguir sob H e para amostras grandes uma distribuição qui-quadrado com 1 grau de liberdade.

3.3.4.2 O teste de Wald

Quando estamos trabalhando com grandes amostras os estimadores de máxima verossimilhança dos parâmetros têm distribuição aproximadamente normal. Por tanto, a significância de um coeficiente particular pode ser verificada através da estatística $z = \hat{\beta}/dp(\hat{\beta})$ que tem distribuição aproximadamente normal padrão. Uma variação dessa estatística é obtida elevando-se esse quociente ao quadrado e se denomina estatística de Wald, a qual segue a lei de uma qui-quadrado com 1 grau de liberdade

$$W = \left[\frac{\hat{\beta}}{dp(\hat{\beta})}\right]^2,$$

em que $dp(\hat{\beta})$ denota o desvio padrão assintótico de $\hat{\beta}$. Expressões para a estatística de Wald para testar mais de um parâmetro podem ser encontradas, por exemplo, em Paula (2004). A estatística de Wald no caso do interesse em testar $H: \beta_2 = 0$ contra $A: \beta_2 \neq 0$ segue também distribuição qui-quadrado com q_1 graus de liberdade. Não existem diferenças importantes de poder entre as estatísticas da razão de verossimilhanças e Wald.

3.3.5 Interpretação dos parâmetros

A interpretação dos resultados obtidos se realiza a partir da interpretação dos coeficientes do modelo. Assim, como foi dito para o modelo logístico simples, na regressão logística múltipla a medida de associação mais utilizada também é a razão de chances.

Para ilustrar vamos supor o modelo logístico múltiplo

$$\pi(\mathbf{x}) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}.$$

Supor inicialmente que a variável explicativa X_j é binária (X_j=1 indica a presença de um fator e X_j=0 a ausência do fator) e que a resposta seja a ocorrência ou não de sinistro. Assim, a razão de chances entre um indivíduo que tem o fator em relação a um indivíduo que não tem o fator continua sendo $\varphi = e^{\beta_j}$, desde que os valores das demais variáveis explicativas sejam os mesmos para ambos os indivíduos. Se X_j for uma variável contínua, a razão de chances entre dois indivíduos, com valor x_j e um valor $x_j + 1$, continua sendo $\varphi = e^{\beta_j}$ desde que os valores das demais variáveis explicativas permaneçam os mesmos para ambos os indivíduos. Assim, os coeficientes de uma regressão logística sempre têm uma fácil interpretação tornando o método bastante atrativo.

3.3.6 Análise de diagnóstico

Uma vez ajustados os modelos temos que verificar possíveis desvios nas suposições feitas para os mesmos, por exemplo, através de técnicas de diagnóstico. Em particular, utilizaremos técnicas gráficas que buscarão identificar pontos suspeitos de serem aberrantes ou influentes. Neste trabalho o assunto será abordado de forma bem superficial, de modo que maiores explicações poderão ser encontradas em Paula(2004).

Para identificar pontos aberrantes analisaremos os resíduos, ou seja, o desvio entre o valor observado e o valor ajustado. Particularmente, para os modelos binomiais o componente do desvio padronizado é definido como:

$$t_{D_i} = \pm\sqrt{\frac{2}{1-\hat{h}_{ii}}}\left\{y_i \ln\left(\frac{y_i}{n_i\hat{\pi}_i}\right)+(n_i-y_i)\ln\left(\frac{n_i-y_i}{n_i-n_i\hat{\pi}_i}\right)\right\}^{1/2},$$

em que h_{ii} é o i-ésimo elemento da diagonal principal da matriz $\mathbf{H} = \mathbf{V}^{1/2}\mathbf{X}\left(\mathbf{X}^T\mathbf{V}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{1/2}$, $\mathbf{V} = diag\{n_1\pi_1,...,n_n\pi_n\}$, n_i : número de repetições de $\mathbf{x}_i^T$ e $0< y_i < n_i$.

Para $y_i = 0$ e $y_i = n_i$ o componente do desvio padronizado assume as seguintes formas:

$$t_{D_i} = -\frac{\{2n_i|\ln(1-\hat{\pi}_i)|\}^{1/2}}{\sqrt{1-\hat{h}_{ii}}} \quad \text{e}$$

$$t_{D_i} = \frac{\{2n_i|\ln\hat{\pi}_i|\}^{1/2}}{\sqrt{1-\hat{h}_{ii}}},$$

respectivamente. O resíduo studentizado também pode ser usado na detecção de pontos aberrantes, sendo definido por

$$t_{S_i} = \frac{1}{\sqrt{1-\hat{h}_{ii}}}\frac{(y_i-n_i\hat{\pi}_i)}{\{n_i\hat{\pi}_i(1-\hat{\pi}_i)\}^{1/2}}.$$

As observações influentes, ou seja, observações que exercem um grande peso nas estimativas dos parâmetros do modelo serão medidas pela distância de Cook aproximada:

$$LD_i \cong \frac{\hat{h}_{ii}}{\left(1-\hat{h}_{ii}\right)^2}\frac{\left(y_i - n_i\hat{\pi}_i\right)^2}{n_i\hat{\pi}_i\left(1-\hat{\pi}_i\right)}.$$

3.4 Modelos Lineares Generalizados

A partir de agora apresentaremos uma extensão dos modelos lineares para uma família mais geral, proposta por Nelder e Wedderburn (1972), denominada Modelos Lineares Generalizados (MLGs). Esta família possibilita modelar variável resposta contínua ou categórica com a distribuição da variável original não necessariamente homocedástica e que considera outras distribuições tais como a Gama, Binomial, Binomial Negativa, Poisson, além da Normal. Mostraremos uma introdução breve à teoria de modelos lineares generalizados e aqueles que desejarem obter uma fonte de informação mais detalhada poderão consultar Paula (2004) e McCullagh e Nelder (1989). Utilizaremos o PROC GENMOD do SAS para ajustar um modelo linear generalizado e por isso muitas vezes iremos mencionar esse procedimento ao longo do texto.

3.4.1 Distribuição da variável resposta

Em modelos lineares generalizados a variável resposta pode possuir uma distribuição de probabilidade pertencente à família exponencial. Isto é, a densidade de probabilidade da variável resposta Y assume uma forma contínua ou discreta. Suponha $Y_1, Y_2, ..., Y_n$, variáveis aleatórias independentes, cada uma com densidade dada por:

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

que é expressa por algumas funções $a(.)$, $b(.)$, e $c(.)$ que determinam a distribuição específica, enquanto $\phi>0$ é o parâmetro de dispersão, que pode ser conhecido ou desconhecido. As funções $a(.)$ e $c(.)$ são determinadas de tal forma que $a(\phi) = \phi w$ e

$c = c(y, \phi/w)$, em que w é um peso conhecido para cada observação. A média e a variância da distribuição apresentada acima podem ser expressas nas formas:

$$E(Y) = b'(\theta) \text{ e}$$

$$Var(Y) = \frac{b''(\theta)\phi}{w},$$

em que a primeira denota a derivada em relação a θ. Se μ representa a média de Y então a variância expressa em função da média é:

$$Var(Y) = \frac{V(\mu)\phi}{w},$$

em que $V(\mu) = b''(\theta)$ é denominada função de variância.

As distribuições de probabilidade da variável resposta Y nos modelos lineares generalizados são parametrizadas em função da média μ e do parâmetro ϕ em vez do parâmetro θ. As distribuições que estão disponíveis no procedimento de GENMOD são mostradas na lista dada abaixo. Além disso, também apresentaremos o parâmetro de escala e a variância de Y.

- Normal:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right] para -\infty < y < +\infty,$$

$$\phi = \sigma^2,$$

$$escala = \sigma,$$

$$Var(Y) = \sigma^2.$$

- Gama:

$$f(y) = \frac{1}{\Gamma(\nu) y}\left(\frac{y\nu}{\mu}\right)^{\nu} \exp\left(-\frac{y\nu}{\mu}\right) para \quad 0 < y < +\infty,$$
$$\phi = \nu^{-1},$$
$$escala = \nu,$$
$$Var(Y) = \frac{\mu^2}{\nu}.$$

- Poisson:

$$f(y) = \frac{\mu^y e^{-\mu}}{y!} para \quad y = 0,1,2,...,$$
$$\phi = 1,$$
$$Var(y) = \mu.$$

- Inversa da Gaussiana:

$$f(y) = \frac{1}{\sqrt{2\pi y^3 \sigma}} \exp\left[-\frac{1}{2y}\left(\frac{y-\mu}{\mu\sigma}\right)^2\right] para \quad 0 < y < +\infty,$$
$$\phi = \sigma^2,$$
$$escala = \sigma,$$
$$Var(Y) = \sigma^2 \mu^3.$$

- Binomial:

$$f(y) = \binom{n}{\tau} \mu^{\tau} (1-\mu)^{n-\tau} para \frac{\tau}{n}, \tau = 0,1,2,...,n,$$
$$\phi = 1,$$
$$Var(y) = \frac{\mu(1-\mu)}{n}.$$

- Binomial Negativa:

$$f(y) = \frac{\Gamma(y+1/k)}{\Gamma(y+1)\Gamma(1/k)} \frac{(k\mu)^k}{(1+k\mu)^{y+1/k}} \quad para \quad y = 0,1,2,...,$$
$$dispersão = k,$$
$$Var(y) = \mu + k\mu^2.$$

A distribuição binomial negativa contém um parâmetro k, chamado parâmetro da dispersão binomial negativa. Este não é o mesmo parâmetro ϕ do modelo linear generalizado, mas é um parâmetro adicional da distribuição que deve ser estimado.

Para a distribuição binomial, a variável resposta é a proporção binomial Y = nº de eventos / nº de ensaios. A função de variância é $V(\mu) = \mu(1-\mu)$, e o parâmetro binomial n dos experimentos é considerado como um peso w.

O PROC GENMOD trabalha com um parâmetro de escala que está relacionado ao parâmetro de dispersão ϕ da família exponencial. Os parâmetros de escala são relacionados aos parâmetros de dispersão como foi visto em cada uma das distribuições de probabilidades anteriormente.

3.4.2 Função de ligação

A média μ_i da variável resposta da i-ésima observação está relacionada a um preditor linear através de uma função de ligação $g(.)$ diferenciável e monótona:

$$g(\mu_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta},$$

em que $\mathbf{x}_i = (1, x_{i1}, ..., x_{ip})^T$ é um vetor fixo de valores de variáveis explanatórias, e $\boldsymbol{\beta}$ é um vetor de parâmetros desconhecidos.

3.4.3 Logaritmo da função de verossimilhança

O logaritmo da função de verossimilhança para as distribuições apresentadas são parametrizados em função da média μ_i e do parâmetro de dispersão ϕ. O termo y_i representa variável resposta para a i-ésima observação, e w_i representa o peso conhecido da dispersão. O logaritmo da função de verossimilhança é dado por:

$$L(\mathbf{y},\boldsymbol{\mu},\phi) = \sum_i \ln\left(f(y_i,\mu_i,\phi)\right),$$

em que a soma se dá sobre todas as observações. As contribuições individuais são $l_i = \ln\left(f(y_i,\mu_i,\phi)\right)$, i=1, ..., n, são apresentadas abaixo em função da média e dos parâmetros de dispersão.

- Normal:

$$l_i = -\frac{1}{2}\left[\frac{w_i(y_i-\mu_i)^2}{\phi}\right] + \ln\left(\frac{\phi}{w_i}\right) + \ln(2\pi).$$

- Gama:

$$l_i = \frac{w_i}{\phi}\ln\left(\frac{w_i y_i}{\phi\mu_i}\right) - \frac{w_i y_i}{\phi\mu_i} - \ln(y_i) - \ln\left(\Gamma\left(\frac{w_i}{\phi}\right)\right).$$

- Poisson:

$$l_i = y_i \ln(\mu_i) - \mu_i.$$

- Inversa Gaussiana:

$$l_i = -\frac{1}{2}\left[\frac{w_i(y_i-\mu_i)^2}{y_i\mu^2\phi}\right] + \ln\left(\frac{\phi y_i^3}{w_i}\right) + \ln(2\pi).$$

- Binomial:

$$l_i = r_i \ln(\mu_i) + (n_i - r_i)\ln(1-\mu_i).$$

- Binomial Negativa:

$$l_i = y\ln(k\mu) - (y+1/k)\ln(1+k\mu) + \ln\left(\frac{\Gamma(y+1/k)}{\Gamma(y+1)\Gamma(1/k)}\right).$$

Para a binomial e a Poisson, os termos que envolvem coeficientes binomiais ou fatoriais das contagens observadas são retirados do cálculo do logaritmo da função de verossimilhança desde que não comprometam a estimação dos parâmetros.

3.4.4 Estimativa de máxima verossimilhança

O procedimento GENMOD utiliza o processo iterativo de Newton-Raphson para maximizar o logaritmo da função de verossimilhança $L(\mathbf{y},\boldsymbol{\mu},\phi)$ em relação aos parâmetros da regressão. Via de regra, o procedimento produz também estimativas de máxima verossimilhança do parâmetro de escala.

Na r-ésima iteração o algoritmo atualiza o vetor de parâmetros $\boldsymbol{\beta}$ da seguinte forma:

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} - \left[\mathbf{J}^{(r)}\right]^{-1}\mathbf{S}^{(r)},$$

em que $\mathbf{J}$ é a matriz Hessiana (segunda derivada) e $\mathbf{S}$ é o vetor gradiente (primeira derivada) da $L(\mathbf{y},\boldsymbol{\mu},\phi)$, ambos calculados do valor atual do vetor de parâmetros. Ou seja:

$$\mathbf{S} = [S_j] = \left[\frac{\partial L}{\partial \beta_j}\right] \text{ e}$$

$$\mathbf{J} = [J_{ij}] = \left[\frac{\partial^2 L}{\partial \beta_i \partial \beta_j}\right].$$

Em alguns casos, o parâmetro de escala é estimado pelo método de máxima verossimilhança. Nestes casos, os elementos que correspondem ao parâmetro de escala são computados e incluídos em **S** e **J**.

O vetor gradiente e a matriz hessiana para os parâmetros de regressão são dados por:

$$\mathbf{S} = \sum_i \frac{w_i (y_i - \mu_i)\mathbf{x}_i}{V(\mu_i) g'(\mu_i) \phi} \text{ e}$$

$$\mathbf{J} = -\mathbf{X}^T \mathbf{W}_0 \mathbf{X}$$

em que $\mathbf{X}$ é a matriz modelo e $\mathbf{W}_0$ é uma matriz diagonal com elementos

$$w_{oi} = w_{ei} + w_i (y_i - \mu_i) \frac{V(\mu_i) g''(\mu_i) + V'(\mu_i) g'(\mu_i)}{(V'(\mu_i))^2 (g'(\mu_i))^3 \phi},$$

em que

$$w_{ei} = \frac{w_i}{\phi V(\mu_i)(g'(\mu_i))^2}, \text{ i=1, ..., n.}$$

A matriz $-\mathbf{J}$ é chamada de matriz de informação observada. O valor esperado de $\mathbf{W}_0$ é a matriz diagonal $\mathbf{W}_e$ com os valores da diagonal igual a w_{ei}. Se substituirmos $\mathbf{W}_0$ por $\mathbf{W}_e$ a matriz $-\mathbf{J}$ é chamada de matriz de informação esperada. A matriz $\mathbf{W}_e$ é chamada de matriz de pesos para o método de ajuste de Fisher.

3.4.5 Matriz de Covariância e Correlação

A matriz estimada de covariância para os parâmetros estimados é dada por $\boldsymbol{\Sigma} = -\mathbf{J}^{-1}$, em que $\mathbf{J}$ é a matriz hessiana calculada usando as estimativas do parâmetro na última iteração. Note que o parâmetro de dispersão, seja estimado ou especificado, será incorporado em $\mathbf{J}$.
A matriz de correlação é a matriz de covariância normalizada. Isto é, σ_{ij} é um elemento de Σ então o elemento correspondente da matriz de correlação é $\sigma_{ij} / \sigma_i \sigma_j$, em que $\sigma_i = \sqrt{\sigma_{ii}}$.

3.4.6 Qualidade do ajuste

Duas estatísticas bastante úteis para avaliar a qualidade do ajuste de um modelo linear generalizado são o Desvio e a estatística Qui-Quadrado de Pearson.

Podemos expressar o logaritmo da função de verossimilhança para μ na forma

$$l(\boldsymbol{\mu}, \mathbf{y}) = \sum_{i=1}^{n} L(\mu_i, y_i),$$

em que $L(\mu_i, y_i)$ denota a contribuição da i-ésima observação.

O Desvio é definido por

$$D^*(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2\{l(\mathbf{y}, \mathbf{y}) - l(\mathbf{y}, \hat{\boldsymbol{\mu}})\}$$

sendo $\hat{\boldsymbol{\mu}}$ a estimativa de máxima verossimilhança de $\boldsymbol{\mu}$.

Para uma distribuição específica também pode ser definido como

$$D^*(\mathbf{y}, \hat{\boldsymbol{\mu}}) = \frac{D(\mathbf{y}, \hat{\boldsymbol{\mu}})}{\phi}$$

em que $D(\mathbf{y},\hat{\boldsymbol{\mu}})$ é sem escala. A Tabela 1 mostra o desvio sem escala para cada uma das distribuições de probabilidade disponíveis no PROC GENMOD.

Tabela 1 - Tabela de desvios sem escala.

Distribuição	**Desvio**
Normal	$\sum_i w_i(y_i-\mu_i)^2$
Gama	$2\sum_i w_i\left[-\ln\left(\frac{y_i}{\mu_i}\right)-(y_i-\mu_i)\right]$
Poisson	$2\sum_i w_i\left[y_i\ln\left(\frac{y_i}{\mu_i}\right)-(y_i-\mu_i)\right]$
Inversa Gaussiana	$\sum_i \frac{w_i(y_i-\mu_i)^2}{\mu_i^2 y_i}$
Binomial	$2\sum_i w_i m_i\left[y_i\ln\left(\frac{y_i}{\mu_i}\right)+(1-y_i)\ln\left(\frac{1-y_i}{1-\mu_i}\right)\right]$
Binomial Negativa	$2\sum_i w_i\left[y\ln\left(\frac{y}{\mu}\right)-(y+\frac{1}{k})\ln\left(\frac{y+1/k}{\mu+1/k}\right)\right]$

No caso binomial, $y_i = r_i/m_i$, em que r_i é a contagem binomial e m_i é o número de réplicas para o i-ésimo grupo.

A estatística Qui-Quadrado de Pearson é definida como

$$\chi^2 = \sum_i \frac{w_i(y_i-\mu_i)^2}{V(\mu_i)},$$

e o Qui-Quadrado de Pearson escalado é dado por χ^2/ϕ.

A versão escalada de ambas as estatísticas, sob determinadas condições de regularidade, segue uma distribuição qui-quadrado com os graus de liberdade iguais ao número de observações menos o número de parâmetros estimados. A versão escalada pode ser usada

como orientação da qualidade de ajuste de um dado modelo. Antes de sair aplicando estas estatísticas temos que assegurar que todas as condições assintóticas estejam satisfeitas.

3.4.7 Parâmetro de Dispersão

Existem diversas opções disponíveis no PROC GENMOD para lidar com o parâmetro de dispersão. As opções NOSCALE e ESCALE na indicação do MODELO afetam a maneira com que o parâmetro da dispersão é tratado. Se você especificar a opção SCALE=DEVIANCE, o parâmetro de dispersão está estimado pelo desvio dividido pelos graus de liberdade. Se você especificar a opção SCALE=PEARSON, o parâmetro de dispersão será estimado pela estatística qui-quadrado de Pearson dividida pelos graus de liberdade. Se a opção ESCALE e NOSCALE não forem especificadas o parâmetro de escala é estimado por máxima verossimilhança.

3.4.8 Análise de diagnóstico

A metodologia de diagnóstico dos pontos aberrantes e influentes utilizada para a regressão logística também será estendida para os MLGs. Para quem quiser se aprofundar no assunto uma boa referencia é Paula (2004).

Assim, definimos o resíduo padronizado como:

$$t_{D_i} = \frac{\phi^{1/2} d(y_i; \hat{\mu}_i)}{\sqrt{1-\hat{h}_{ii}}},$$

em que $d(y_i; \hat{\mu}_i)$ denota a raiz quadrada sinalizada do i-ésimo componente de $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ e h_{ii} denota o i-ésimo elemento de uma matriz $\mathbf{H}$ especifica.

A distância de Cook aproximada fica definida como:

$$LD_i \cong \left\{ \frac{\hat{h}_{ii}}{(1-\hat{h}_{ii})} \right\} t^2_{S_i},$$

em que

$$t_{S_i} = \frac{\phi^{1/2}(y_i - \hat{\mu}_i)}{\sqrt{\hat{V}_i(1-\hat{h}_{ii})}},$$

com $\hat{V}_i = V(\hat{\mu}_i)$.

4 CAPÍTULO IV

4.1 Desenvolvimento

Mostraremos neste capítulo a aplicação das três técnicas estatísticas descritas no capítulo anterior. Inicialmente mostraremos a análise descritiva das variáveis, que foi muito útil para nos orientar no processo de categorização das variáveis e também verificar o comportamento de cada variável explicativa em função de sua correspondente variável resposta. Em seguida apresentamos um processo de categorização (WOE), com o objetivo de obter um alto poder de discriminação das variáveis independentes em função da variável resposta. Tendo as variáveis categorizadas, faltava apenas identificar as interações entre as variáveis para rodar os modelos, e isso foi realizado com a construção das duas árvores de decisão. As interações encontradas foram incorporadas nos dois modelos, de freqüência e de custo de sinistro. E por fim apresentaremos e discutiremos os resultados obtidos em cada modelo.

As análises foram feitas nos softwares "SAS System for windows V8" e "Answer Tree for windows v. 2.0".

4.2 Análise descritiva

Para estudar o comportamento das variáveis explicativas, construímos tabelas de freqüência de sinistro para cada variável qualitativa, um histograma e um gráfico de densidade para a variável contínua, valor de sinistro.

Como podemos observar na Figura 8 e Figura 9 a distribuição da variável valor do sinistro mostra-se bastante assimétrica sugerindo que uma distribuição gama possa se ajustar bem aos dados.

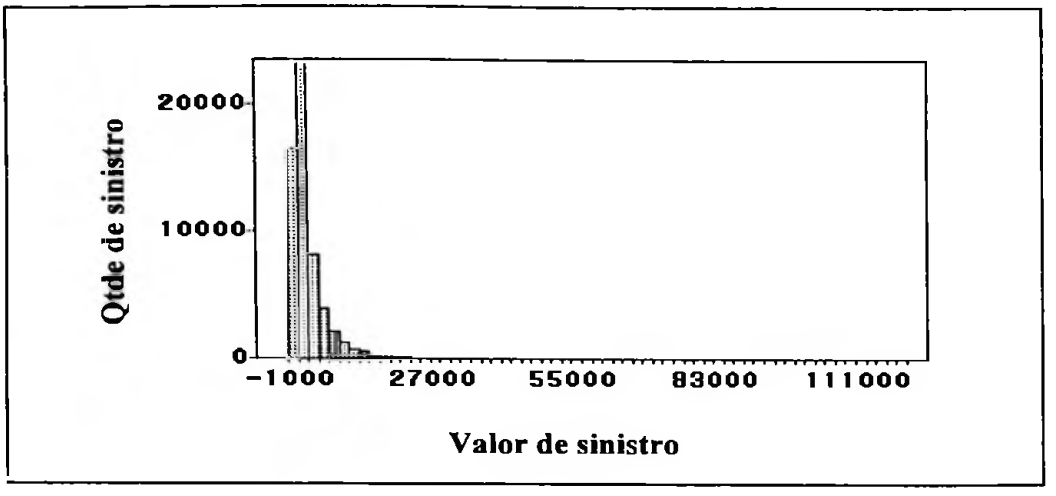

Figura 8 - Histograma do valor de sinistros.

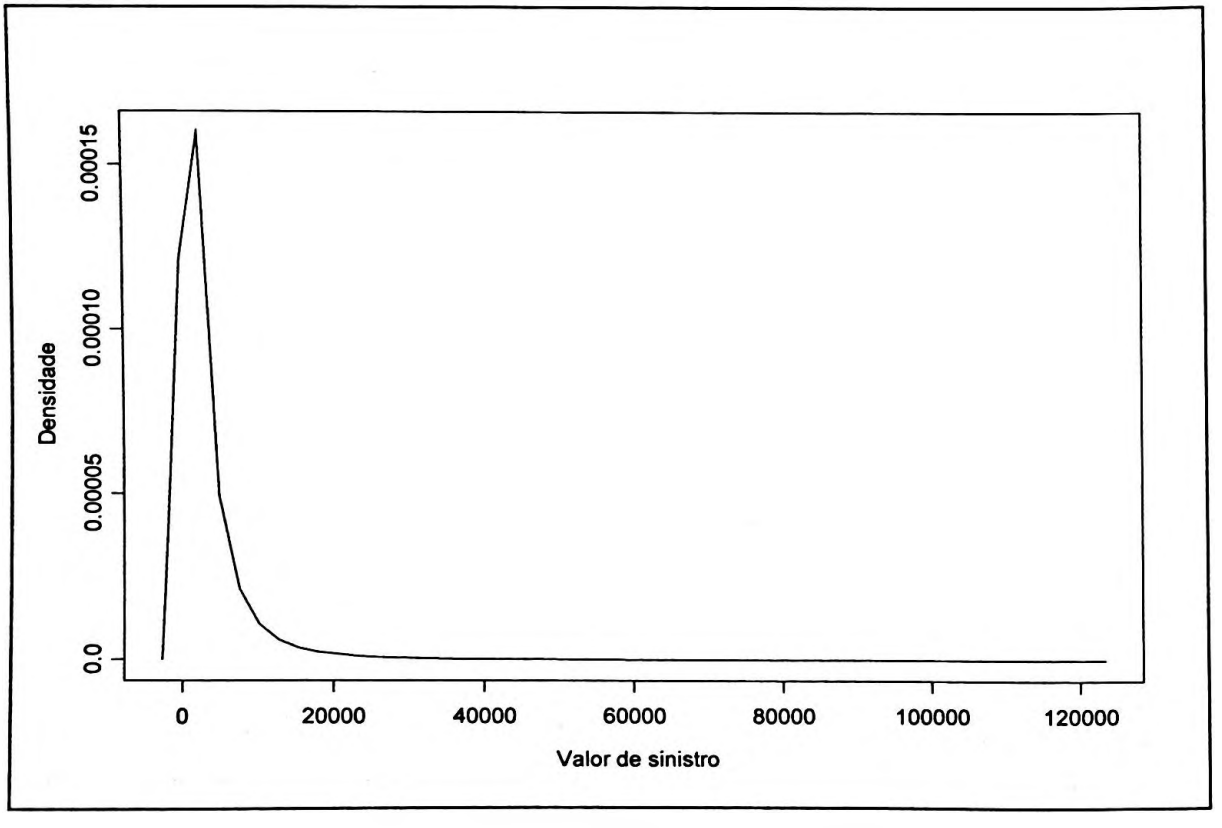

Figura 9 - Gráfico de densidade do valor de sinistros.

No Apêndice 1, podemos observar que o percentual de apólices que tiveram sinistro é de 5,2%, e que muitas variáveis apresentam ausência de respostas ou seu percentual é muito

pequeno nas categorias. Esse comportamento já nos indica que uma categorização das variáveis se faz necessária.

Vamos explorar um pouco mais cada uma das variáveis, começando pela variável ANO, que apresenta maior concentração de veículos novos. Aproximadamente 50% da carteira tem ano de fabricação igual ou superior a 2001 e uma pequena parte (17%) foi fabricada antes de 1996. A carteira tem veículos com ano de fabricação variando de 1936 até 2005, e podemos perceber que os veículos mais velhos apresentam categorias sem ocorrência de sinistro ou com percentual de sinistro muito baixo, quando comparado com o total da carteira.

A variável FABR apresenta 38 categorias diferentes, porém as montadoras 1, 2, 3 e 4 dominam o mercado com quase 90% de participação. As demais categorias, além de ter pouca participação, algumas delas não tiveram veículos sinistrados.

Comportamento semelhante ocorre com BONUS: temos muitas classes (0 a 20), porém as oito primeiras classes representam mais de 97% da carteira. Aqui, novamente observamos categorias sem a ocorrência de sinistro ou com apenas um caso.

Já PROCED apresenta apenas duas categorias (1 e 10) e a categoria 1 representa 95% dos veículos, porém aqui não observamos categorias com baixa ocorrência de sinistros.

A variável ORIG difere das variáveis apresentadas até aqui, apresentando uma distribuição mais uniforme.

A próxima variável, TIPVEI, tem seis classes, mas como acontece com a maioria das variáveis analisadas até aqui apenas duas delas concentram 82% da carteira.

Quando olhamos para o estado civil dos segurados (CIVIL), percebemos que os casados aparecem como a categoria dominante (78%), seguida pelos solteiros com 11%, e as demais categorias se distribuem de forma mais ou menos uniforme apresentando um comportamento mais ou menos igual ao da carteira.

Com exceção das três primeiras categorias, motoristas com menos de trinta anos, as demais categorias da variável IDADE distribuem-se de forma mais ou menos homogênea. No

entanto, o percentual de sinistro, principalmente nas duas primeiras categorias (jovens até 25 anos), é muito alto.

Aproximadamente 70% dos veículos possuem dispositivo antifurto instalado em seus veículos. A variável FURTO apresenta a categoria "não definida" com apenas 3 automóveis sem sinistro.

Quando olhamos para o tipo de pintura (PINTURA), a classe com mais representatividade é a "metálica", com 68%, seguida pela "comum" e "perolizada" com 24% e 8%, respectivamente.

Quanto ao sexo dos motoristas (SEXO), os homens são maioria, com quase 57%.

E finalmente observamos que o sistema de rastreamento do veículo (DAF) está instalado em apenas 0,7% dos veículos, mas que reflete um percentual de sinistro muito menor que a carteira (3%).

4.3 Categorização das variáveis

Uma vez definidas as variáveis resposta, nosso objetivo é encontrar as variáveis explicativas mais relevantes para explicar a variável ocorrência e valor do sinistro. Além disso, estamos interessados em identificar agrupamentos com comportamento interno homogêneo em relação às variáveis resposta e heterogêneo em relação às demais categorias da variável.

Categorizar uma variável consiste em agrupar suas respostas em categorias com comportamento interno igual e diferindo em relação às outras categorias da variável. Em nosso trabalho, as variáveis foram agrupadas em categorias semelhantes quanto ao risco de ocorrência de sinistro. Para cada variável, categorias com o mesmo comportamento receberam um valor de 1 a n (número de categorias). Desta forma, criamos uma nova variável explicativa (variável categorizada), que será utilizada no lugar da variável original. A variável ANOMOD, por exemplo, que originalmente apresentava 40 categorias com o ano de fabricação variando de 1936 a 2005, após a categorização passou a ter apenas três categorias (1: veículos <= 1995; 2: veículos entre 1996 e 2000; 3: veículos >=2001).

A utilização de variáveis categorizadas tem algumas vantagens sobre as variáveis originais. O modelo resultante é mais simples, pois teremos que estimar um número menor de coeficientes. Além disso, os segurados com mesmo comportamento em relação à ocorrência de sinistro terão pesos iguais no modelo. Outro aspecto relevante está no fato das interações entre as variáveis serem determinantes para a obtenção de uma tarifa equilibrada e competitiva. Logo, variáveis contínuas como a ANOMOD devem ser categorizadas, uma vez que as interações entre cada ano e modelo e as categorias de uma outra variável se tornariam inviáveis.

O processo de categorização utilizado está baseado na variável resposta, ou seja, as categorias das variáveis independentes foram agrupadas de acordo com a relação existente entre cada uma delas e a variável dependente SINISTRO. Todo o processo está fundamentado em uma medida descritiva "Weights of Evidence" - WOE (Good, 1950).

4.3.1 Weights of Evidence – WOE

O WOE é uma medida descritiva que auxilia a identificação de categorias com alto ou baixo poder de discriminação em relação à variável alvo, além de identificar aquelas categorias que melhor discriminam os segurados sinistrados dos que não apresentaram sinistros. Outra vantagem é que podemos agrupar categorias com valores de WOE próximos, desde que apresentem uma explicação plausível quanto à lógica de seguros. Desta forma, podemos reduzir o número de categorias da variável, tornando-a mais estável.

A seguir mostraremos todos os passos necessários para o cálculo da medida WOE.

- Distribuir a variável em percentis. Se já estivermos trabalhando com uma variável categórica neste primeiro instante, assumiremos a distribuição de categorias existente. Se a priori estivermos trabalhando com uma variável contínua, sem nenhum conhecimento técnico, sugerimos que inicialmente a variável seja dividida em percentis;

- Em seguida, calculamos para cada categoria o número e o percentual de segurados com e sem sinistro;

- Após as duas primeiras etapas, calculamos a razão entre "segurados sem sinistro" / "segurados com sinistro" (% s/sin / % c/sin);

- Finalmente, teremos WOE, que nada mais é que o logaritmo natural da razão entre segurados sem e com sinistro.

Tabela 2 - Construção da medida WOE.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro Na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
A	s_1	c_1	C1/(c1+s1)	s1/s.	c1/c.	(s1/s.) / (c1/c.)	ln[(s1/s.) / (c1/c.)]
B	s_2	c_2	C2/(c2+s2)	s2/s.	c2/c.	(s2/s.) / (c2/c.)	ln[(s2/s.) / (c2/c.)]
C	s_3	c_3	C3/(c3+s3)	s3/s.	c3/c.	(s3/s.) / (c3/c.)	ln[(s3/s.) / (c3/c.)]
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
J	S_{10}	c_{10}	c10/(c10+s10)	s10/s.	c10/c.	(s10/s.) / (c10/c.)	ln[(s10/s.) / (c10/c.)]
Total	s.	c.	c./(c.+s.)	1	1	1	0

Note na Tabela 2 que o WOE pode assumir os seguintes valores:

- < 0: valores negativos indicam que a categoria tem poder para discriminar segurados com maior propensão a ter sinistros. Além disso, quanto mais distante de zero, maior será o poder de discriminação.

- = 0: valores nulos indicam que a categoria é neutra, ou seja, a razão entre segurados sem e com sinistro é igual a um.

- > 0: valores positivos indicam que a categoria tem poder para discriminar segurados com menor propensão a ter sinistros. Além disso, quanto mais distante de zero, maior será o poder de discriminação.

Como exemplo, mostraremos como foi categorizada a variável idade do motorista principal. A variável IDADE foi dividida em classes, variando de cinco em cinco anos, de acordo com a experiência obtida pelos analistas de precificação. Calculamos o valor de WOE seguindo o critério descrito anteriormente.

Tabela 3 - Categorização da variável idade do motorista principal.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
< DE 21 ANOS	3.791	438	10,36	0,26	0,55	0,47	-0,75
DE 21 A 25 ANOS	23.574	1.770	6,98	1,61	2,23	0,72	-0,32
DE 26 A 30 ANOS	52.375	3.044	5,49	3,58	3,83	0,93	-0,07
DE 31 A 35 ANOS	173.844	9.435	5,15	11,88	11,87	1,00	0,00
DE 36 A 40 ANOS	279.190	14.459	4,92	19,08	18,19	1,05	0,05
DE 41 A 45 ANOS	249.737	13.324	5,06	17,07	16,77	1,02	0,02
DE 46 A 50 ANOS	189.997	10.508	5,24	12,99	13,22	0,98	-0,02
DE 51 A 55 ANOS	166.146	9.678	5,50	11,36	12,18	0,93	-0,07
DE 56 A 60 ANOS	123.582	6.697	5,14	8,45	8,43	1,00	0,00
MAIS DE 60 ANOS	200.952	10118	4,79	13,73	12,73	1,08	0,08
Total	1.463.188	79.471	5,15	1,00	1,00	1,00	0,00

Analisando o valor de WOE, apresentados na Tabela 3, podemos tirar algumas conclusões. Note que os jovens de até 25 anos apresentam valores negativos e distantes de zero, mostrando que esses motoristas apresentam maiores chances de se envolverem em acidentes. Os motoristas que estão na faixa dos 46 a 55 anos também apresentam um risco maior, com valores de WOE negativo. As demais faixas de idade se constituem em um risco bom, com valores de WOE POSITIVO. Estes valores de WOE confirmam o que na prática já se conhece.

Com base nos valores do WOE e na experiência em seguros, a variável IDADE, que era contínua, foi categorizada e deu origem à variável IDMOT, com quatro categorias (Tabela 4).

Tabela 4 - Categorização da variável idade do motorista principal.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	Woe
<= DE 30 ANOS	79.740	5.252	6,18	0,05	0,07	0,82	-0,19
DE 31 A 45 ANOS	702.771	37.218	5,03	0,48	0,47	1,03	0,03
DE 46 A 55 ANOS	356.143	20.186	5,36	0,24	0,25	0,96	-0,04
MAIS DE 55 ANOS	324.534	16.815	4,93	0,22	0,21	1,05	0,05
Total	1.463.188	79.471	5,15	1,00	1,00	1,00	0,00

As tabelas contendo as categorias e WOE para cada variável independente, antes e após o procedimento de categorização se encontram no Apêndice 1 e 2.

4.4 Interação entre as variáveis – Árvore de Decisão

Como já foi dito anteriormente, a experiência obtida em seguros mostra que a interação entre as variáveis da carteira é um fator muito importante para segmentar os clientes mais ou menos rentáveis. Um exemplo clássico em seguros é a interação entre sexo, estado civil e idade. Em geral, os homens tem maior probabilidade de colidir seus veículos do que as mulheres, porém quando observamos as mulheres casadas na faixa de 40 a 50 anos essa situação não ocorre. Uma hipótese plausível para esse caso talvez esteja no fato dessas mulheres possuírem filhos jovens que tenham acesso aos seus automóveis.

Com as variáveis categorizadas, faremos uso da árvore de decisão para detectar as possíveis interações entre as mesmas e também para confirmar as interações já conhecidas como sexo, estado civil e idade. Construímos duas árvores, uma para as variáveis do modelo de freqüência e a outra para o modelo de custo de sinistros. Isto porque cada variável resposta é explicada por variáveis e interações distintas. Por exemplo, um determinado perfil pode ter um alto índice de colisão, porém essas colisões são leves, de tal forma que o valor total de sinistro é baixo.

4.4.1 Árvore de Decisão para freqüência de sinistros

Para construir a árvore de freqüência de sinistro (Apêndice 6), selecionamos todas as variáveis descritivas e utilizamos um nível de significância de 0,01. Desta forma, a árvore cresceu bastante e se ramificou, atingindo sete níveis e selecionou nove variáveis significativas. A variável mais importante para explicar a ocorrência de sinistro foi o ano-modelo do veículo (ANOMOD). Esta divide a árvore em três grandes ramos, a saber: veículos inferiores a 1996, veículos entre 1996 e 2000 e finalmente os automóveis superiores a 2000. A partir de agora iremos analisar estes grandes segmentos separadamente.

O segmento de carros inferiores a 1996 é o menor dos três e subdivide-se em mais dois níveis, segmentando os segurados em duas categorias de classe de bônus (BONUS) e duas categorias de dispositivo antifurto (FURTO). Esse ramo produziu três perfis distintos quanto à freqüência de sinistros e duas interações, ANOMOD x BONUS e BONUS x FURTO.

O segundo grande ramo da árvore cresceu mais quatro níveis, dando origem a 14 nós. A classe de bônus também surgiu como a variável mais significativa, segmentando o ramo em dois, classe de bônus inferior e superior a 5. Daí para frente a árvore continuou crescendo e segmentando o ramo em tipo de veículo (TIVEI), fabricante (FABR), dispositivo antifurto (FURTO) e idade do motorista (IDMOT). Além das interações já observadas no primeiro ramo, também foi possível detectar as seguintes interações: BONUS x TIVEI, TIVEI x FABR, TIVEI x FURTO e FURTO x IDMOT.

E por fim vamos analisar o último e maior ramo da árvore, com 18 nós e cinco níveis. Assim como ocorreu com os dois ramos anteriores, a classe de bônus divide o grupo em dois, mostrando que essa variável é a mais discriminante depois do ano e modelo. O ramo continua subdividindo-se e as variáveis responsáveis por isso são: idade do motorista (IDMOT), tipo de pintura (PIN), sistema antifurto (FURTO), origem (ORIGEM) e tipo de veículo (TIVEI). Neste ramo as interações que merecem destaque foram BONUS x PINTURA, PINTURA x IDMOT, BONUS x IDMOT, IDMOT x ORIGEM, SEXO x TIVEI, SEXO x IDMOT e FURTO x SEXO.

As variáveis ANOMOD, BONUS, TIVEI e IDMOT parecem ser as mais significativas. A variável estado civil (CIVIL) não apareceu nesta árvore, porém durante a construção da mesma ela estava presente em níveis inferiores, e optamos por não apresentá-la apenas para facilitar a visualização e análise da árvore. Logo, além das treze interações descritas acima também foram identificadas as interações CIVIL x SEXO e CIVIL x IDMOT, totalizando 15 interações.

Com as interações obtidas esperamos obter um melhor ajuste do modelo de freqüência, visto que apenas a experiência em seguros não é capaz de apontar quais são as variáveis mais relevantes e como elas interagem.

4.4.2 Árvore de Decisão para valor de sinistros

Assim como fizemos com a árvore anterior, para segmentar o valor de sinistro nós também selecionamos todas as variáveis explicativas e permitimos o seu crescimento com nível de significância de 0,01. A árvore construída (Apêndice 7) apresenta no seu primeiro nível a variável procedência do veículo (PROCED) como a mais significante. Esta divide a carteira em dois segmentos, os veículos nacionais e os importados. A partir do nó dos veículos nacionais ocorre uma nova divisão entre tipo de veículo (TIPVEI), que dá origem aos dois grandes segmentos, os "populares" e as "demais categorias".

O segmento dos veículos "populares" segue subdividindo-se em fabricante (FABR) , sexo (SEXO) e origem (ORIGEM), enquanto que o segmento das "demais categorias" subdivide-se em origem, fabricante, ano (ANO) e sexo. A análise conjunta das variáveis nos permite concluir que existem pelo menos seis interações entre as mesmas. São elas: PROCED x TIPVEI, TIPVEI x FABR, FABR x SEXO, FABR x ORIGEM, FABR x ANO e ANO x SEXO.

Diferente do que ocorreu com a primeira árvore, as variáveis que melhor explicam a variação do valor de sinistro foram PROCED, TIVEI, ORIGEM e FABR. Isso mostra que a decisão de construir árvores distintas para cada uma das variáveis resposta foi acertada. Essas interações serão testadas no modelo e esperamos que o mesmo confirme a importância das mesmas.

4.5 Estimação da freqüência de sinistros – Regressão Logística

Por definição, a regressão logística é utilizada quando nossa variável dependente assume apenas os valores 0 ou 1, mas as variáveis independentes podem ser categóricas ou contínuas. O modelo de regressão logística estima a probabilidade de ocorrência de um evento específico (1=ocorrência de sinistro; 0=não ocorrência de sinistro); em outras palavras, descrever a probabilidade associada aos valores da variável resposta.

Esse tipo de modelo vem sendo largamente utilizado em diversas áreas do conhecimento para modelagem estatística de dados. Em parte, isso se deve à facilidade de interpretação dos

parâmetros do modelo e à facilidade de uma revisão periódica. Enquanto o comportamento dos segurados da carteira de seguros se mantiver estável, o modelo será robusto. No entanto, em algum instante no tempo fatores agirão sobre essa carteira, provocando mudanças, e então será preciso rodar um novo modelo.

A freqüência de sinistro é calculada dividindo-se a quantidade de sinistro pela exposição, o que conceitualmente pode ser visto como a probabilidade de ocorrência de sinistro durante a vigência do seguro. Essa forma de cálculo nos permite dimensionar o risco de cada segurado ao longo da vigência da apólice.

O modelo logístico parece se aplicar perfeitamente ao nosso problema. Então vejamos: o modelo de regressão logística estima a probabilidade de ocorrência de sinistro, que é uma variável binária, mas como descrevemos anteriormente isso é justamente nosso objetivo. O único cuidado que devemos tomar é com a forma de entrada dos dados no modelo, para incorporar a exposição no modelo.

Existem duas formas de entrarmos com os dados para rodar um modelo logístico no proc logistic do SAS. A primeira é considerar que cada linha representa uma observação, e então quando escrevemos nosso modelo a variável resposta está representada pela variável binária SINISTRO. Na segunda agrupamos a quantidade de sinistros e de segurados segundo os diversos perfis existentes, de forma que nosso modelo passe a considerar como variável resposta a razão entre o somatório de sinistros e o somatório de segurados, que nada mais é que o percentual de sinistrados na carteira. Como estamos interessados em modelar a freqüência de sinistros, substituiremos a quantidade de segurados pela exposição e, desse modo, nossa variável resposta será a razão entre a soma de sinistros e a soma da exposição.

Antes de rodar o modelo logístico, seguimos a recomendação de Hosmer e Lemeshow (1989) e criamos variáveis "dummy" para representar as categorias das variáveis explicativas. Assim a variável ANOMOD, com três categorias, passou a ser representada por apenas duas variáveis "dummy" (ANO1 e ANO2), pois a última categoria é representada pela combinação das demais "dummy". Com isso o modelo se torna mais parcimonioso, uma vez que necessitamos estimar menos parâmetros. A Tabela 5 mostra a relação criada entre a variável preditora e as "dummy".

Tabela 5 - Exemplo de criação de variável "dummy".

ANOMOD	ANO1	ANO2
Ano <= 1995	1	0
1996 <= Ano <= 2000	0	1
2001 <= Ano	0	0

Inicialmente selecionaremos as 12 variáveis explicativas, cada uma representada por k-1 variáveis "dummy" (k é o número de categorias da variável explicativa), o que totaliza 19 variáveis. Além disso, incluímos no modelo as 40 interações que foram diagnosticadas pela árvore de decisão. No Apêndice 5 podemos verificar todas as variáveis "dummy" que serão utilizadas no modelo assim como as interações entre elas.

Ajustamos um modelo logístico múltiplo através do procedimento proc logistic do SAS (Apêndice 3). As variáveis explicativas mais relevantes para determinação da ocorrência de sinistro foram selecionadas através do método automático de seleção stepwise, adotando-se o nível de significância 15% como critério de entrada e saída das variáveis no modelo.

A Tabela 6 resume todos os passos realizados até o modelo final, mostrando a ordem de entrada das variáveis e indicando qual variável foi incluída ou removida em cada passo com os seus respectivos p-valores para inclusão e remoção no modelo.

Tabela 6 - Regressão Logística – Stepwise.

Passo	Incluída	P-valor	Removida	P-valor
1	BON1	0,0000	-	< 0,15
2	ANO1	0,0000	-	< 0,15
3	ANO2	0,0000	-	< 0,15
4	TVEI1	0,0000	-	< 0,15
5	IDMOT1	0,0000	-	< 0,15
6	IDMOT3	0,0000	-	< 0,15
7	FAB1	0,0000	-	< 0,15
8	FURTO1	0,0000	-	< 0,15
9	PIN1	0,0000	-	< 0,15
10	RASTR1	0,0000	-	< 0,15
11	FAB5	0,0000	-	< 0,15
12	FAB2	0,0000	-	< 0,15
13	BON1*IDMOT1	0,0000	-	< 0,15
14	CPCD1	0,0000	-	< 0,15
15	CIVIL1	0,0000	-	< 0,15
16	CIVIL1*IDMOT1	0,0000	-	< 0,15
17	CIVIL1*IDMOT3	0,0000	-	< 0,15
18	BON1*PIN1	0,0001	-	< 0,15
19	FAB5*TVEI1	0,0003	-	< 0,15
20	BON1*TVEI1	0,0014	-	< 0,15
21	SEX1	0,0023	-	< 0,15
22	CIVIL1*SEX1	0,0000	-	< 0,15
23	IDMOT3*SEX1	0,0000	-	< 0,15
24	IDMOT1*SEX1	0,0000	-	< 0,15
25	FAB1*TVEI1	0,0023	-	< 0,15
26	ANO1*IDMOT1	0,0432	-	< 0,15
27	IDMOT2	0,0869	-	< 0,15
28	ANO1*IDMOT2	0,0001	-	< 0,15
29		0,1827	ANO1*IDMOT1	> 0,15
30	BON1*IDMOT2	0,0003	-	< 0,15
31	ANO2*IDMOT2	0,0015	-	< 0,15
32	ANO1*IDMOT3	0,0079	-	< 0,15
33	CIVIL1*IDMOT2	0,0550	-	< 0,15
34	IDMOT2*SEX1	0,0661	-	< 0,15
35	ORIG1	0,1363	-	< 0,15
36	BON1*CIVIL1	0,1352	-	< 0,15
37	FURTO1*IDMOT2	0,1380	-	< 0,15

Podemos verificar que apenas a interação ANO1*IDMOT1 foi removida após a sua entrada, e das 59 variáveis iniciais (variáveis explicativas e interações) o modelo final selecionou apenas 35 variáveis explicativas.

No passo-0 o intercepto entra no modelo com o p-valor de inclusão (< 0,0001) sendo menor do que o valor crítico de 0,15. No próximo passo a variável mais relevante é ANO1 com p-valor de inclusão sendo menor do que 0,15. E o processo seguiu até o último passo com a inclusão da interação FURTO1*IDMOT2 no modelo.

As variáveis selecionadas pelo modelo final, com as respectivas estimativas dos parâmetros e a sua significância são mostradas na Tabela 7. O modelo parece estar bem ajustado, pois o valor observado do desvio é menor do que o nº de graus de liberdade.

Tabela 7 - Variáveis e coeficientes estimados pelo modelo.

Parâmetro	Estimativas	Erro Padrão	P-valor
Intercepto	-2,271	0,040	0,0000
ANO1	-0,519	0,025	0,0000
ANO2	-0,180	0,011	0,0000
BON1	0,371	0,026	0,0000
CIVIL1	-0,101	0,032	0,0017
CPCD1	-0,112	0,021	0,0000
FAB1	0,059	0,015	0,0001
FAB2	-0,286	0,036	0,0000
FAB5	-0,098	0,012	0,0000
FURTO1	-0,082	0,011	0,0000
IDMOT1	0,674	0,050	0,0000
IDMOT2	0,077	0,034	0,0223
IDMOT3	-0,085	0,037	0,0204
ORIG1	0,014	0,008	0,0863
PIN1	-0,144	0,017	0,0000
RASTR1	-0,492	0,069	0,0000
SEX1	-0,171	0,027	0,0000
TVEI1	-0,092	0,016	0,0000
BON1*TVEI1	-0,060	0,017	0,0006
BON1*PIN1	0,087	0,020	0,0000
BON1*IDMOT1	-0,232	0,043	0,0000
BON1*IDMOT2	-0,073	0,017	0,0000
BON1*CIVIL1	-0,040	0,023	0,0876
CIVIL1*SEX1	0,207	0,022	0,0000
CIVIL1*IDMOT1	-0,319	0,041	0,0000
CIVIL1*IDMOT2	-0,077	0,030	0,0098
CIVIL1*IDMOT3	0,160	0,036	0,0000
FURTO1*IDMOT2	-0,028	0,017	0,0898
IDMOT1*SEX1	-0,225	0,038	0,0000
IDMOT2*SEX1	-0,048	0,023	0,0339
IDMOT3*SEX1	0,074	0,025	0,0037
ANO1*IDMOT2	0,199	0,033	0,0000
ANO1*IDMOT3	0,107	0,036	0,0030
ANO2*IDMOT2	0,058	0,016	0,0003
FAB1*TVEI1	0,098	0,027	0,0003
FAB5*TVEI1	0,088	0,018	0,0000

Ainda, o valor observado da estatística Wald foi igual a 3.069,0286. Comparando-o com o valor crítico de uma distribuição Qui-quadrado com 35 graus de liberdade, temos indícios para rejeitar a hipótese nula. Ou seja, não há motivos para acreditarmos que todos os coeficientes estimados sejam iguais a zero.

Note que os coeficientes negativos diminuem a probabilidade de ocorrência de sinistro e coeficientes positivos aumentam a probabilidade ou a freqüência de sinistro. Essa possibilidade de interpretação dos coeficientes é uma das grandes vantagens da regressão logística, e isso pode ser melhor explorado com a interpretação das razões de chances apresentadas no Apêndice 8.

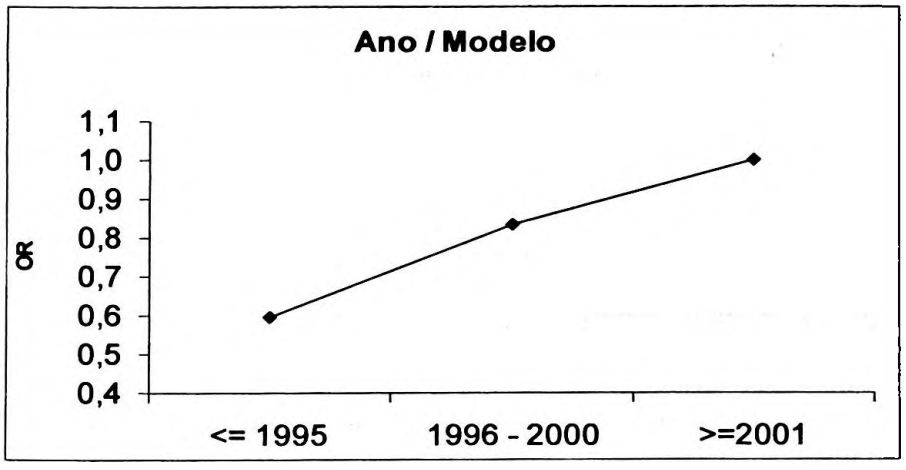

Gráfico 1: Razão de chances - Ano / Modelo.

Analisando o Gráfico 1 podemos verificar, por exemplo, que a chance de uma colisão entre os segurados que possuem veículos mais novos é consideravelmente maior que a dos segurados que possuem veículos mais velhos, dado que as demais variáveis permaneçam inalteradas.

Uma vez considerado estatisticamente significante, precisamos testar o poder explicativo desse modelo, isto é, necessitamos verificar com que eficiência o modelo consegue prever a ocorrência de sinistro. Verificar o poder explicativo nada mais é que olhar para a diferença entre os valores estimados e observados, e se essa diferença for pequena, dizemos que o modelo tem um alto poder explicativo. O teste utilizado para verificar o poder explicativo do modelo foi o de Hosmer-Lemeshow.

Para estimar se as diferenças entre os valores estimados e observados são consideradas estatisticamente grandes, vamos calcular o valor da estatística $\hat{C}$ de Hosmer-Lemeshow, a partir da Tabela 8.

Tabela 8 - Partição para o teste de Hosmer-Lemeshow.

Grupo	Total	Evento		Não Evento	
		Observado	Estimado	Observado	Estimado
1	73.511	3.732	3.734,08	69.779	69.776,84
2	72.908	4.507	4.489,33	68.401	68.418,96
3	74.349	5.021	5.022,78	69.328	69.326,31
4	73.197	5.285	5.272,76	67.912	67.924,73
5	74.082	5.601	5.649,78	68.481	68.431,82
6	72.556	5.831	5.804,84	66.725	66.751,03
7	72.997	6.176	6.164,24	66.821	66.832,61
8	73.421	6.620	6.553,32	66.801	66.867,61
9	73.356	7.034	7.091,48	66.322	66.264,25
10	72.867	8.238	8.230,16	64.629	64.637,14

O valor da estatística $\hat{C}$ é 2,7424 que é menor do que o valor crítico da distribuição Qui-quadrado com 8 graus de liberdade. Logo, não temos indícios suficientes para rejeitar a hipótese nula de modelo correto.

Finalmente, depois de concluirmos que o modelo é eficiente e está bem ajustado, e a partir dos coeficientes estimados e mostrados anteriormente, podemos determinar a equação do modelo para estimar a probabilidade de ocorrer um sinistro, dados os valores das variáveis explicativas:

$$\log\left\{\frac{\hat{\mu}}{1-\hat{\mu}}\right\} = -2{,}2713 - ANO1 * 0{,}5189 + ... + FAB5 * TIPVEI1 * 0{,}0879$$

em que,

$\hat{\mu}$:probabilidade estimada de ocorrer um sinistro para um determinado perfil de segurado.

$$\hat{\mu} = \frac{e^{-2{,}2713 - ANO1*0{,}5189 + ... + FAB5*TIPVEI1*0{,}0879}}{1 + e^{-2{,}2713 - ANO1*0{,}5189 + ... + FAB5*TIPVEI1*0{,}0879}}$$

4.5.1 Taxa de acerto do modelo

Para calcular a taxa de acerto do modelo rodamos o modelo de freqüência de sinistros para a base de validação. Dado que a freqüência real está próxima de 7,9% utilizamos um ponto de corte igual a 0,08 para classificar um determinado perfil de segurado como um provável sinistrado.

Tabela 9 - Taxa de acerto do modelo.

		REAL		
		1	0	
PREVISTO	1	4.368	24.655	29.023
	0	34.573	303.026	337.599
		38.941	327.681	366.622

Porcentagem de acerto	**83,8**
Sensibilidade	**15,1**
Especificidade	**89,8**
Falso +	**88,8**
Falso -	**7,5**

A Tabela 9, mostra que a taxa de acerto é alta (83,8%) indicando que o modelo parece estar bem ajustado. O valor de falso positivo é alto, provavelmente por se tratar de um evento raro e termos utilizado uma amostra prospectiva.

4.5.2 Análise de diagnóstico

No diagnóstico do modelo procuramos identificar a existência de pontos aberrantes (Figura 10) e influentes (Figura 11) que pudessem comprometer a qualidade do ajuste do modelo. De um modo geral não encontramos desvios sérios. Os desvios se distribuem homogeneamente ao redor de zero e parece não existir pontos influentes de forma desproporcional nas estimativas dos parâmetros do modelo ajustado.

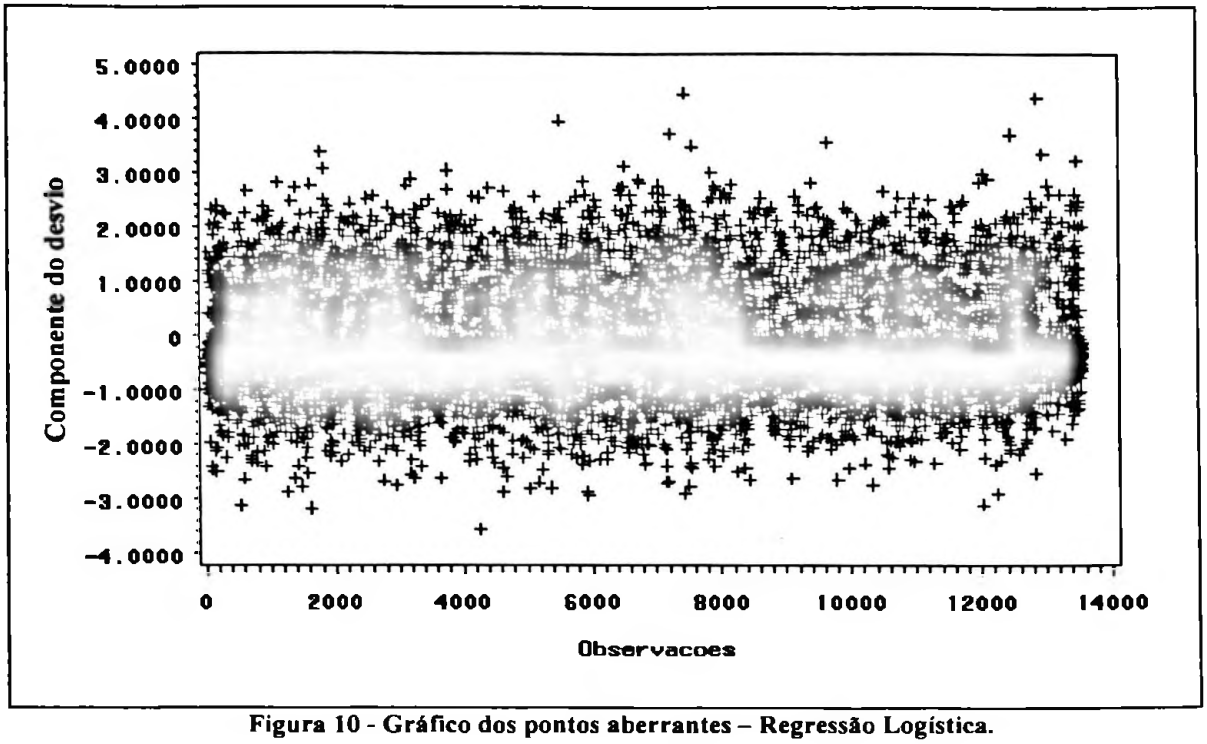

Figura 10 - Gráfico dos pontos aberrantes – Regressão Logística.

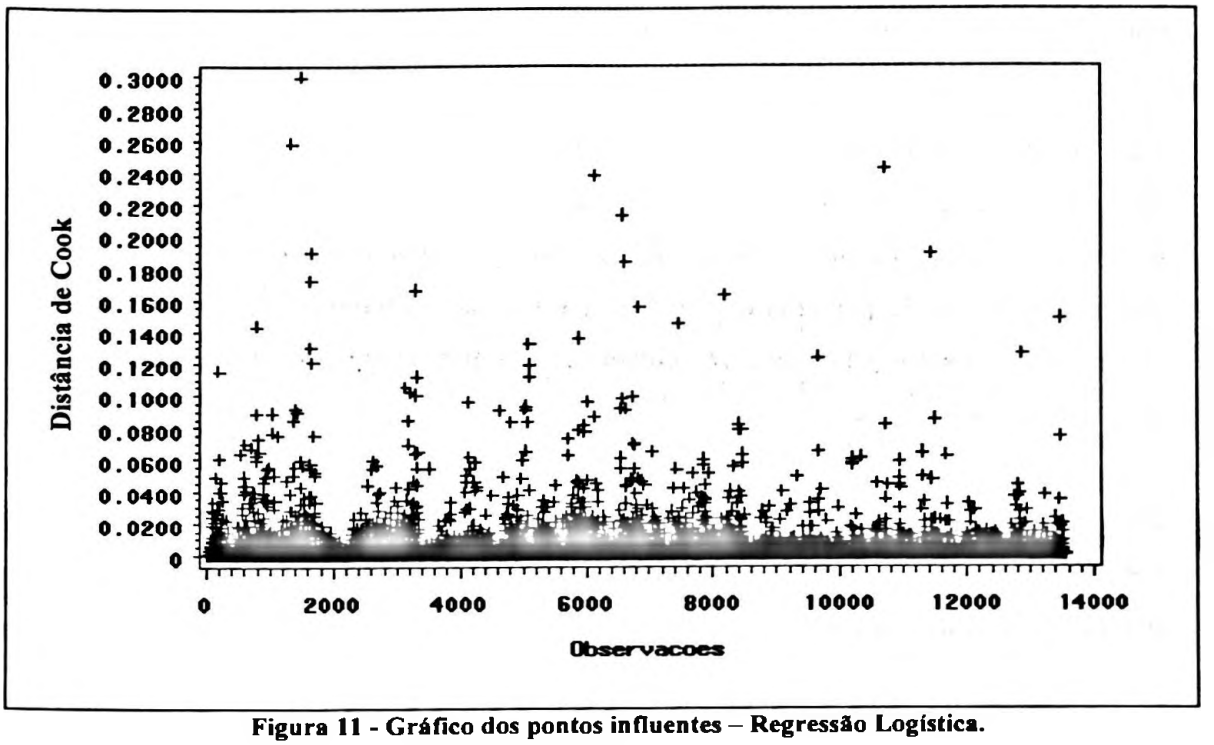

Figura 11 - Gráfico dos pontos influentes – Regressão Logística.

4.6 Estimação do valor de sinistros – Regressão Gama

Uma vez estimada a freqüência de sinistros, nosso objetivo agora é encontrar um modelo que possa explicar a relação existente entre o valor de sinistro de colisão e as variáveis explicativas mencionadas anteriormente. Como já foi visto, o valor de sinistro (VLRSIN) pode ser aproximado por uma distribuição gama e assumiremos que eles são independentes. Assim sendo, os Modelos Lineares Generalizados parece ser a técnica mais adequada para estimar o valor dos sinistros.

Utilizamos o proc genmod do SAS e, inicialmente, tentaremos ajustar um modelo com as 19 variáveis explicativas utilizadas para estimar a freqüência de sinistros, mais as 29 interações detectadas com a árvore de decisão.

Analisando a Tabela 10 e a saída do programa no Apêndice 4, podemos observar que apesar do modelo estar bem ajustado nem todos os coeficientes estimados são significantes, ou seja, o p-valor é maior do que 0,05. Logo, retiramos a interação ANO3*FAB5 que apresenta o maior p-valor não significativo e rodamos o modelo novamente. Mais uma vez observamos a significância dos coeficientes e repetimos o processo até obtermos um modelo final com todas os coeficientes significantes.

Tabela 10 - Coeficientes estimados para o primeiro modelo gama.

Parâmetro	Estimativa	Erro Padrão	LI	LS	W	P-valor
Intercepto	7,3467	0,7046	5,9657	8,7276	108,72	<,0001
ANO2	0,1276	0,1158	-0,0994	0,3547	1,21	0,2705
ANO2*FAB1	-0,0744	0,1286	-0,3265	0,1777	0,33	0,5629
ANO2*FAB3	-0,0294	0,1184	-0,2615	0,2027	0,06	0,8038
ANO2*FAB4	0,0662	0,1229	-0,1746	0,3070	0,29	0,5901
ANO2*FAB5	0,0271	0,1174	-0,2030	0,2571	0,05	0,8177
ANO2*FAB6	0,1676	0,1172	-0,0622	0,3974	2,04	0,1528
ANO2*SEX2	0,0152	0,0327	-0,0488	0,0792	0,22	0,6420
ANO3	0,2764	0,1206	0,0400	0,5128	5,25	0,0219
ANO3*FAB1	-0,1272	0,1306	-0,3832	0,1288	0,95	0,3301
ANO3*FAB3	-0,1917	0,1231	-0,4329	0,0495	2,43	0,1193
ANO3*FAB4	0,0540	0,1272	-0,1953	0,3034	0,18	0,6711
ANO3*FAB5	-0,0024	0,1223	-0,2420	0,2372	0,00	0,9841
ANO3*FAB6	0,1119	0,1220	-0,1273	0,3510	0,84	0,3593
ANO3*SEX2	0,0243	0,0319	-0,0381	0,0868	0,58	0,4450
BON1	0,0958	0,0093	0,0776	0,1140	106,60	<,0001
CIVIL1	0,0530	0,0114	0,0308	0,0753	21,83	<,0001
CPCD1	-0,3528	0,0876	-0,5245	-0,1810	16,20	<,0001
CPCD1*TVEI2	-0,0956	0,0908	-0,2735	0,0824	1,11	0,2926
FAB1	0,7039	0,7100	-0,6878	2,0955	0,98	0,3215
FAB1*ORIG2	0,0119	0,0798	-0,1445	0,1683	0,02	0,8815
FAB1*SEX2	-0,1946	0,0764	-0,3444	-0,0448	6,48	0,0109
FAB1*TVEI2	-0,8933	0,6999	-2,2650	0,4785	1,63	0,2019
FAB3	0,5997	0,7084	-0,7887	1,9881	0,72	0,3972
FAB3*ORIG2	0,0509	0,0790	-0,1039	0,2057	0,41	0,5197
FAB3*SEX2	-0,3350	0,0748	-0,4816	-0,1884	20,06	<,0001
FAB3*TVEI2	-1,0684	0,7000	-2,4404	0,3036	2,33	0,1269
FAB4	0,4661	0,7094	-0,9244	1,8566	0,43	0,5112
FAB4*ORIG2	-0,0214	0,0814	-0,1810	0,1382	0,07	0,7929
FAB4*SEX2	-0,2120	0,0773	-0,3636	-0,0604	7,51	0,0061
FAB4*TVEI2	-0,8885	0,7000	-2,2604	0,4835	1,61	0,2043
FAB5	0,2864	0,7083	-1,1019	1,6746	0,16	0,6860
FAB5*ORIG2	0,0596	0,0782	-0,0937	0,2129	0,58	0,4461
FAB5*SEX2	-0,2319	0,0746	-0,3781	-0,0857	9,66	0,0019
FAB5*TVEI2	-0,8030	0,7000	-2,1749	0,5690	1,32	0,2513
FAB6	0,4210	0,7083	-0,9672	1,8092	0,35	0,5522
FAB6*ORIG2	0,0665	0,0788	-0,0879	0,2209	0,71	0,3988
FAB6*SEX2	-0,3169	0,0749	-0,4637	-0,1701	17,90	<,0001
FAB6*TVEI2	-0,9869	0,7000	-2,3588	0,3851	1,99	0,1586
FURTO2	0,0400	0,0091	0,0221	0,0579	19,16	<,0001
IDMOT2	-0,0290	0,0176	-0,0635	0,0054	2,73	0,0988
IDMOT3	-0,0063	0,0187	-0,0428	0,0303	0,11	0,7370
IDMOT4	-0,0775	0,0191	-0,1148	-0,0401	16,50	<,0001
ORIG2	-0,1183	0,0782	-0,2716	0,0351	2,28	0,1307
ORIG2*TVEI2	-0,0595	0,0185	-0,0958	-0,0231	10,30	0,0013
PIN2	-0,0807	0,0096	-0,0994	-0,0619	71,19	<,0001
RASTR2	0,0680	0,0649	-0,0591	0,1952	1,10	0,2942
SEX2	0,3740	0,0779	0,2214	0,5267	23,06	<,0001
TVEI2	1,4768	0,6943	0,1160	2,8377	4,52	0,0334

Os coeficientes estimados para este novo modelo são todos significantes, com p-valor inferior a 0,05. Pelo valor da estimativa de ϕ que foi 1,05 podemos confirmar a forte assimetria da variável valor de sinistro, ou seja, uma forte concentração de sinistros com valores baixos e uma freqüência decrescente de sinistros com valores maiores.

E finalmente apresentamos na Tabela 11 todos os coeficientes ajustados pelo modelo final.

Tabela 11 - Variáveis e coeficientes estimados para o modelo final.

Parâmetro	Estimativa	Erro Padrão	LI	LS	W	P-valor
Intercept	7,811	0,036	7,740	7,883	46161,600	< 0,0001
ANO2	0,143	0,019	0,107	0,180	59,390	< 0,0001
ANO3	0,270	0,020	0,231	0,310	181,960	< 0,0001
BON1	0,096	0,009	0,078	0,114	108,070	< 0,0001
CIVIL1	0,052	0,011	0,031	0,073	23,130	< 0,0001
CPCD1	-0,459	0,022	-0,503	-0,415	423,700	< 0,0001
FAB1	0,299	0,034	0,233	0,365	79,790	< 0,0001
FAB3	0,264	0,030	0,206	0,323	78,920	< 0,0001
FAB4	0,215	0,031	0,155	0,276	48,900	< 0,0001
FAB6	0,113	0,043	0,030	0,197	7,090	0,0078
FURTO2	0,040	0,009	0,022	0,058	19,400	< 0,0001
IDMOT2	-0,024	0,009	-0,042	-0,006	6,600	0,0102
IDMOT4	-0,072	0,012	-0,095	-0,050	39,810	< 0,0001
ORIG2	-0,125	0,022	-0,167	-0,083	33,890	< 0,0001
PIN2	-0,081	0,010	-0,100	-0,062	71,710	< 0,0001
SEX2	0,382	0,071	0,242	0,521	28,790	< 0,0001
TVEI2	1,081	0,059	0,966	1,196	340,090	< 0,0001
ANO2*FAB6	0,163	0,034	0,097	0,230	23,110	< 0,0001
ANO3*FAB3	-0,142	0,021	-0,182	-0,101	46,440	< 0,0001
ANO3*FAB6	0,136	0,035	0,068	0,203	15,450	< 0,0001
FAB1*SEX2	-0,179	0,075	-0,325	-0,033	5,780	0,0162
FAB1*TVEI2	-0,583	0,063	-0,706	-0,460	86,430	< 0,0001
FAB3*ORIG2	0,056	0,026	0,006	0,107	4,730	0,0297
FAB3*SEX2	-0,324	0,073	-0,467	-0,181	19,620	< 0,0001
FAB3*TVEI2	-0,769	0,060	-0,887	-0,652	165,380	< 0,0001
FAB4*SEX2	-0,205	0,076	-0,353	-0,057	7,330	0,0068
FAB4*TVEI2	-0,599	0,063	-0,722	-0,477	91,610	< 0,0001
FAB6*ORIG2	0,072	0,026	0,022	0,122	7,840	0,0051
FAB6*SEX2	-0,306	0,073	-0,450	-0,163	17,500	< 0,0001
FAB6*TVEI2	-0,689	0,060	-0,806	-0,572	133,430	< 0,0001
ORIG2*FAB5	0,064	0,025	0,016	0,112	6,840	0,0089
ORIG2*TVEI2	-0,057	0,018	-0,093	-0,021	9,640	0,0019
SEX2*FAB5	-0,221	0,073	-0,364	-0,079	9,220	0,0024
TVEI2*FAB5	-0,507	0,057	-0,619	-0,394	78,040	< 0,0001

E assim o modelo final ajustado fica determinado da seguinte forma:

$$\hat{\mu} = e^{7,811 - ANO2*0,143 + ... + TVEI2*FAB5*(-0,507)}$$

em que $\hat{\mu}$ é a estimativa do valor esperado do custo de sinistro dado um particular perfil de risco do segurado.

4.6.1 Análise de diagnóstico

Assim como fizemos na regressão logística, também utilizaremos métodos gráficos de diagnóstico para os MLGs. De um modo geral não encontramos desvios sérios que pudessem comprometer a qualidade do ajuste (Figura 12).

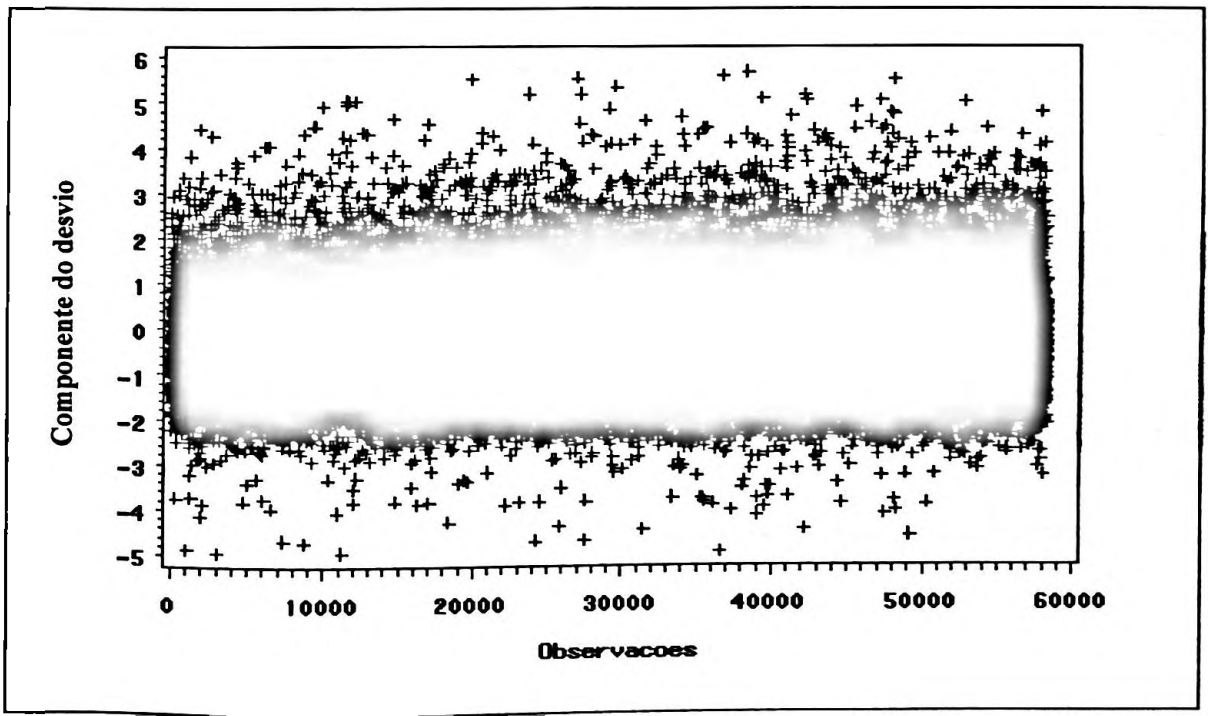

Figura 12 - Gráfico dos pontos aberrantes – MLGs.

Observando a Figura 13, concluímos que apesar de alguns pontos descolarem um pouco da massa de observações, ainda assim não classificaremos como pontos influentes.

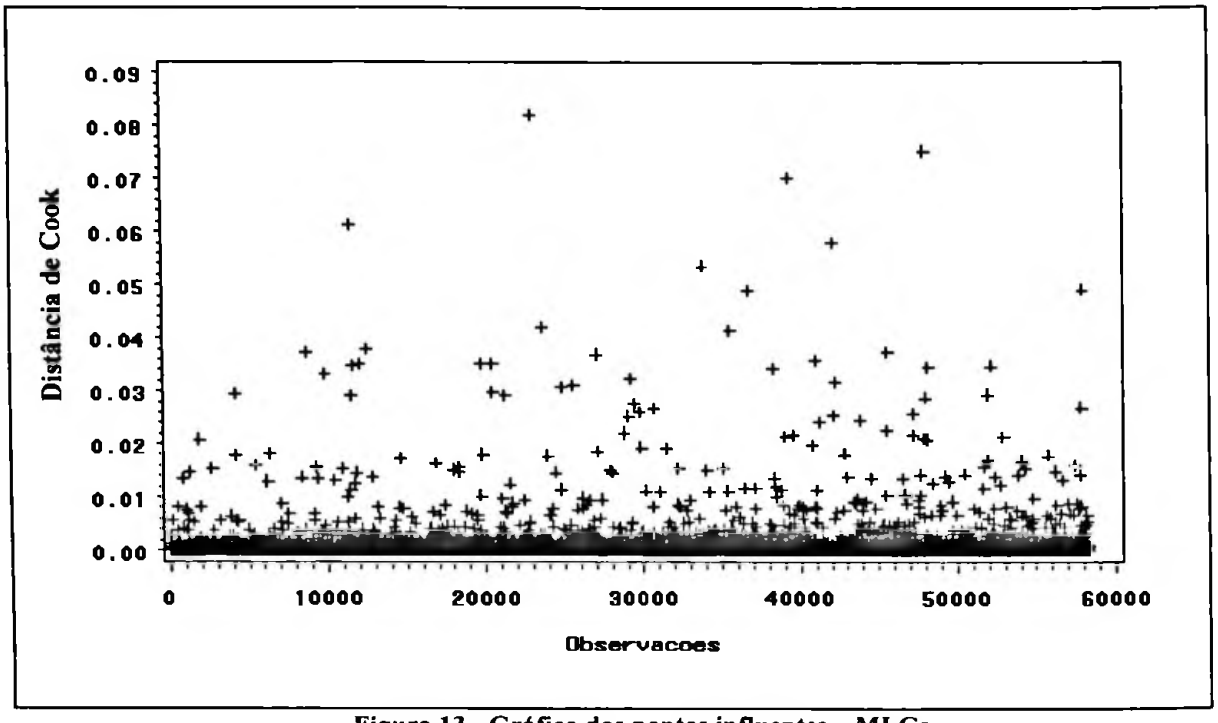

Figura 13 - Gráfico dos pontos influentes – MLGs.

5 CAPÍTULO V

5.1 Conclusões

Neste trabalho foi descrita uma alternativa para precificação de sinistros de colisão. Foram apresentadas as características do seguro e as variáveis disponíveis para o estudo, assim como os métodos de tratamento dessas variáveis.

Uma técnica de categorização de variáveis foi apresentada (WOE) e outra de segmentação de carteiras (Árvore de Decisão - CHAID), com a intenção de verificar as interações existentes entre as variáveis explicativas.

O WOE permitiu categorizar cada uma das variáveis explicativas em função da variável alvo, unindo categorias com mesmo comportamento em relação à variável dependente e mantendo separadas categorias distintas. Através do CHAID foi possível identificar diversas interações entre as variáveis explicativas, que foram utilizadas com grande sucesso no ajuste dos modelos. Como ferramenta de árvore de decisão, o CHAID destacou-se por apresentar diversos mecanismos de controle do crescimento e fácil interpretação do método de classificação do algoritmo.

Apresentamos ainda um modelo de regressão logística para estimar a freqüência de sinistros de colisão e regressão gama para obter o custo de sinistros. Dentre as diversas vantagens da regressão logística, destacamos a facilidade de interpretação dos parâmetros e o resultado da análise ser uma probabilidade, o que permitiu adequá-la perfeitamente para estimar a freqüência de sinistros. A regressão gama conseguiu um bom ajuste e parte deve-se à utilização conjunta com o CHAID, uma vez que o mesmo não possui um método de seleção de variáveis.

5.2 Sugestões

Quando estudamos eventos raros, como por exemplo doenças raras, pode-se optar por uma amostragem retrospectiva em que amostras independentes são extraídas dos grupos em que

ocorreu o evento (casos) e em que não ocorreu o evento (controle). Pode-se, por exemplo extrair para cada caso N controles (em geral N não deve ser superior a 4). Assim, consegue-se representatividade de ambos os grupos (casos e controles) na amostra sem que seja preciso trabalhar com banco de dados muito grandes. Conhecendo-se os tamanhos dos grupos pode se fazer as correções necessárias de modo que as freqüências estimadas pelo modelo de regressão logística tenham as mesmas interpretações inferenciais de um modelo ajustado a dados prospectivos (para maiores detalhes, ver Paula, 2004, pp. 105-106). Como sinistro é um evento raro este tipo de procedimento poderia também ser utilizado e teria a vantagem de podermos trabalhar com um banco de dados bem menor do que aquele utilizado nas nossas análises, porém exigiria a escolha de uma amostra de apólices sem ocorrência de sinistro comparável ao grupo de apólices com sinistro.

REFRÊNCIAS

AGRESTI, A. ***Categorical Data Analysis.*** New York. Wiley. 1990.

BRASIL, Gilberto. ***O ABC da Matemática Atuarial e Princípios Gerais de Seguros.*** Porto Alegre. Sulina. 1985.

BREIMAN, L. et al. ***Classification and Regression Trees. California.*** Wadsworth. 1984.

BUSSAB; MORETTIN, P. A.. ***Estatística Básica.*** São Paulo. Atlas. 2002.

COX, D. R.; SNELL, E. J.. ***The Analysis of Binary Data.*** 2nd ed. London. Chapman and Hall. 1989.

GOOD, I. J. ***Probability and the Weighing of Evidence.*** London. Charles Griffin. 1950.

HOSMER, David W.; LEMESHOW, Stanley. ***Applied Logistic Regression.*** New York. Wiley. 1989.

KASS, G.. ***An exploratory technique for investigating large quantities of categorical data.*** Applied Statistics. 1980. 29:2, 119-127.

Loh,W. Y.; Shih, Y. S. ***Split Selection Methods for Classification Trees.*** Statistica Sinica. 1997.

MCCULLAGH, P.; NELDER, J. A. ***Generalized Linear Models.*** 2nd ed. New York. Chapman and Hall. 1989.

MENDES, João José de Souza. ***Bases Técnicas do Seguro.*** São Paulo. EMTS. 1977.

MORGAN, J. N.; MESSENGER, R. C. (1973). ***THAID: A sequential analysis program for the analysis of nominal scale dependent variables.*** *Technical report,* Institute of Social Research, University of Michigan, Ann Arbor. 1973.

NAGELKERKE, N. ***A Note on a General Definition of the Coefficient of Determination.*** Biometrika. 1991.

NELDER, J. A.; WEDDERBURN, R. W. N. ***Generalized Linear Models.*** Journal of the Royal Statistical Society A, 135, 370-384.

PAULA, G. A. ***Modelos de Regressão com Apoio Computacional.*** São Paulo. IME-USP. 2004.

RIPLEY, B. ***Pattern Recognition and Neural Networks.*** Cambridge: Cambridge University Press. 1996.

SAS Institute Inc. ***SAS Language: Reference.*** 1nd ed. Cary, NC: SAS Institute Inc. 1990.

SAS INSTITUTE INC. ***SAS OnlineDoc®.*** Version 8, Cary, NC: SAS Institute Inc. 1999.
SAS Institute Inc. ***SAS Procedures Guide.*** Version 6, 3nd ed. Cary, NC: SAS Institute Inc 1990.

SAS Institute Inc. ***SAS/STAT User's Guide.*** Version 6, 4nd ed, Volume 1. Cary, NC: SAS Institute Inc. 1989.

SAS Institute Inc. ***SAS/STAT User's Guide.*** Version 6, 4nd ed, Volume 2. Cary, NC: SAS Institute Inc. 1989.

APÊNDICES

APÊNDICE 01 – TABELA DE FREQÜÊNCIA DAS VARIÁVEIS ORIGINAIS

Tabela 12 - Ano e modelo do veículo.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
1934	2	0	0,00	0,00	0,00	-	-
1949	2	0	0,00	0,00	0,00	-	-
1956	2	0	0,00	0,00	0,00	-	-
1962	4	0	0,00	0,00	0,00	-	-
1968	3	0	0,00	0,00	0,00	-	-
1974	11	0	0,00	0,00	0,00	-	-
1975	11	1	8,33	0,00	0,00	0,61	-0,50
1976	18	0	0,00	0,00	0,00	-	-
1977	9	0	0,00	0,00	0,00	-	-
1978	11	0	0,00	0,00	0,00	-	-
1979	24	2	7,69	0,00	0,00	0,66	-0,41
1980	29	0	0,00	0,00	0,00	-	-
1981	49	3	5,77	0,00	0,00	0,90	-0,10
1982	75	1	1,32	0,00	0,00	4,14	1,42
1983	147	2	1,34	0,00	0,00	4,05	1,40
1984	248	7	2,75	0,00	0,00	1,95	0,67
1985	453	17	3,62	0,00	0,00	1,47	0,39
1986	1.646	78	4,52	0,00	0,00	1,16	0,15
1987	1.666	71	4,09	0,00	0,00	1,29	0,26
1988	3.466	161	4,44	0,00	0,00	1,19	0,17
1989	7.119	331	4,44	0,00	0,00	1,19	0,17
1990	7.968	337	4,06	0,01	0,00	1,30	0,27
1991	11.502	550	4,56	0,01	0,01	1,15	0,14
1992	13.383	622	4,44	0,01	0,01	1,19	0,17
1993	24.746	1.198	4,62	0,02	0,01	1,14	0,13
1994	40.833	2.009	4,69	0,03	0,02	1,12	0,11
1995	69.612	3.599	4,92	0,05	0,04	1,07	0,06
1996	90.852	4.464	4,68	0,06	0,05	1,12	0,12
1997	114.876	5.913	4,90	0,08	0,07	1,07	0,07
1998	107.675	5.515	4,87	0,07	0,07	1,08	0,07
1999	137.106	7.331	5,08	0,09	0,09	1,03	0,03
2000	155.087	8.557	5,23	0,10	0,10	1,00	0,00
2001	212.264	11.826	5,28	0,14	0,14	0,99	-0,01
2002	167.437	9.788	5,52	0,11	0,12	0,94	-0,06
2003	183.676	11.089	5,69	0,12	0,13	0,91	-0,09
2004	130.508	8.417	6,06	0,09	0,10	0,86	-0,16
2005	41.471	2.172	4,98	0,03	0,03	1,05	0,05
Total	1.523.991	84.061	5,23	1,00	1,00	1,00	0,00

Tabela 13 - Fabricante do veículo.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
1	2	0	0,00	0,00	0,00	-	-
2	370.543	20.535	5,25	0,24	0,24	1,00	0,00
3	157.884	8.496	5,11	0,10	0,10	1,03	0,02
4	418.675	22.183	5,03	0,27	0,26	1,04	0,04
5	385.797	21.237	5,22	0,25	0,25	1,00	0,00
6	41	2	4,65	0,00	0,00	1,13	0,12
7	8.908	660	6,90	0,01	0,01	0,74	-0,30
8	1.959	107	5,18	0,00	0,00	1,01	0,01
9	11.884	714	5,67	0,01	0,01	0,92	-0,09
10	30.750	1.850	5,67	0,02	0,02	0,92	-0,09
11	1.330	80	5,67	0,00	0,00	0,92	-0,09
12	3.995	182	4,36	0,00	0,00	1,21	0,19
13	10.105	544	5,11	0,01	0,01	1,02	0,02
14	602	35	5,49	0,00	0,00	0,95	-0,05
15	8.336	485	5,50	0,01	0,01	0,95	-0,05
16	1.668	90	5,12	0,00	0,00	1,02	0,02
17	26.099	1.864	6,67	0,02	0,02	0,77	-0,26
18	50.796	2.946	5,48	0,03	0,04	0,95	-0,05
19	655	35	5,07	0,00	0,00	1,03	0,03
20	1.252	55	4,21	0,00	0,00	1,26	0,23
21	24.396	1.500	5,79	0,02	0,02	0,90	-0,11
22	771	43	5,28	0,00	0,00	0,99	-0,01
23	42		0,00	0,00	0,00	-	-
24	1		0,00	0,00	0,00	-	-
25	87	2	2,25	0,00	0,00	2,40	0,88
26	1		0,00	0,00	0,00	-	-
27	11		0,00	0,00	0,00	-	-
28	2		0,00	0,00	0,00	-	-
29	3.639	182	4,76	0,00	0,00	1,10	0,10
30	953	52	5,17	0,00	0,00	1,01	0,01
31	2.230	160	6,69	0,00	0,00	0,77	-0,26
32	58	2	3,33	0,00	0,00	1,60	0,47
33	1	1	50,00	0,00	0,00	0,06	-2,90
34	10		0,00	0,00	0,00	-	-
35	1		0,00	0,00	0,00	-	-
36	514	18	3,38	0,00	0,00	1,58	0,45
37	7		0,00	0,00	0,00	-	-
38	8	1	11,11	0,00	0,00	0,44	-0,82
Total	1.524.013	84.061	5,23	1,00	1,00	1,00	0,00

Tabela 14 - Classe de bônus.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
0	225.415	15.000	6,24	0,15	0,18	0,83	-0,19
1	185.132	10.978	5,60	0,12	0,13	0,93	-0,07
2	166.114	9.495	5,41	0,11	0,11	0,96	-0,04
3	142.998	8.110	5,37	0,09	0,10	0,97	-0,03
4	139.249	7.864	5,35	0,09	0,09	0,98	-0,02
5	156.449	8.770	5,31	0,10	0,10	0,98	-0,02
6	274.730	12.625	4,39	0,18	0,15	1,20	0,18
7	192.323	10.316	5,09	0,13	0,12	1,03	0,03
8	32.809	623	1,86	0,02	0,01	2,90	1,07
9	6.235	207	3,21	0,00	0,00	1,66	0,51
10	2.007	51	2,48	0,00	0,00	2,17	0,78
11	256	8	3,03	0,00	0,00	1,77	0,57
12	130	7	5,11	0,00	0,00	1,02	0,02
13	78	1	1,27	0,00	0,00	4,30	1,46
14	19	1	5,00	0,00	0,00	1,05	0,05
15	6		0,00	0,00	0,00	-	-
16	4		0,00	0,00	0,00	-	-
17	1		0,00	0,00	0,00	-	-
18	1		0,00	0,00	0,00	-	-
19	1		0,00	0,00	0,00	-	-
20	56	5	8,20	0,00	0,00	0,62	-0,48
Total	1.524.013	84.061	5,23	1,00	1,00	1,00	0,00

Tabela 15 - Procedência do veículo.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
Nacional	1.455.279	80.033	5,21	0,95	0,95	1,00	0,00
Importado	68.734	4.028	5,54	0,05	0,05	0,94	-0,06
Total	1.524.013	84.061	5,23	1,00	1,00	1,00	0,00

Tabela 16 - Origem do cliente.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
Em	197.660	12.241	5,83	0,13	0,15	0,89	-0,12
R1	317.378	17896	5,34	0,21	0,21	0,98	-0,02
R2	241.527	13707	5,37	0,16	0,16	0,97	-0,03
R3	167.364	8925	5,06	0,11	0,11	1,03	0,03
R4	97.608	4836	4,72	0,06	0,06	1,11	0,11
R5	261.757	13443	4,88	0,17	0,16	1,07	0,07
Rc	240.719	13013	5,13	0,16	0,15	1,02	0,02
Total	1.524.013	84.061	5,23	1,00	1,00	1,00	0,00

Tabela 17 - Tipo de veículo.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
1	710.508	37.556	5,02	0,47	0,45	1,04	0,04
2	553.445	30783	5,27	0,36	0,37	0,99	-0,01
3	118.495	7468	5,93	0,08	0,09	0,88	-0,13
4	3.933	253	6,04	0,00	0,00	0,86	-0,15
5	60.843	3783	5,85	0,04	0,05	0,89	-0,12
6	76.789	4218	5,21	0,05	0,05	1,00	0,00
Total	1.524.013	84.061	5,23	1,00	1,00	1,00	0,00

Tabela 18 - Estado civil do segurado.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
Casado	1.150.339	61.890	5,11	0,79	0,78	1,01	0,01
Desquitado	4.109	219	5,06	0,00	0,00	1,02	0,02
Divorciado	26.873	1.516	5,34	0,02	0,02	0,96	-0,04
Outro	3.879	195	4,79	0,00	0,00	1,08	0,08
Separado	23.842	1.449	5,73	0,02	0,02	0,89	-0,11
Solteiro	164.453	9.028	5,20	0,11	0,11	0,99	-0,01
União Estável	55.074	3.405	5,82	0,04	0,04	0,88	-0,13
Viúvo	30.043	1625	5,13	0,02	0,02	1,01	0,01
Total	1.458.612	79.327	5,16	1,00	1,00	1,00	0,00

Tabela 19 - Idade do motorista principal.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
< De 21 Anos	3.791	438	10,36	0,26	0,55	0,47	-0,75
De 21 A 25 Anos	23.574	1.770	6,98	1,61	2,23	0,72	-0,32
De 26 A 30 Anos	52.375	3.044	5,49	3,58	3,83	0,93	-0,07
De 31 A 35 Anos	173.844	9.435	5,15	11,88	11,87	1,00	0,00
De 36 A 40 Anos	279.190	14.459	4,92	19,08	18,19	1,05	0,05
De 41 A 45 Anos	249.737	13.324	5,06	17,07	16,77	1,02	0,02
De 46 A 50 Anos	189.997	10.508	5,24	12,99	13,22	0,98	-0,02
De 51 A 55 Anos	166.146	9.678	5,50	11,36	12,18	0,93	-0,07
De 56 A 60 Anos	123.582	6.697	5,14	8,45	8,43	1,00	0,00
Mais De 60 Anos	200.952	10118	4,79	13,73	12,73	1,08	0,08
Total	1.463.188	79.471	5,15	1,00	1,00	1,00	0,00

Tabela 20 - Sistema antifurto.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
Não	502.214	26.603	5,03	0,33	0,32	1,04	0,04
Não Definido	3		0,00	0,00	0,00	-	-
Sim	1.017.127	57.310	5,33	0,67	0,68	0,98	-0,02
Total	1.519.344	83.913	5,23	1,00	1,00	1,00	0,00

Tabela 21 - Tipo de pintura.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
Comum	348.075	19.628	5,34	0,24	0,24	0,99	-0,01
Metálica	1.002.924	55.808	5,27	0,68	0,68	1,00	0,00
Perolizada	118.265	6.515	5,22	0,08	0,08	1,01	0,01
Total	1.469.264	81.951	5,28	1,00	1,00	1,00	0,00

Tabela 22 - Sexo do motorista principal.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
F	638.479	34.405	5,11	0,44	0,43	1,01	0,01
M	824.733	45.069	5,18	0,56	0,57	0,99	-0,0
Total	1.463.212	79.474	5,15	1,00	1,00	1,00	0,0(

Tabela 23 - Sistema de rastreamento do veículo.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
Instalado	10.286	318	3,00	0,01	0,00	1,78	0,58
Não Instalado	1.513.727	83.743	5,24	0,99	1,00	1,00	0,00
Total	1.524.013	84.061	5,23	1,00	1,00	1,00	0,00

APÊNDICE 02 – CATEGORIZAÇÃO DAS VARIÁVEIS

Tabela 24 - Ano e modelo do veículo.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
<= 1995	273.891	13.453	4,68	0,18	0,16	1,12	0,12
1996 - 2000	514.744	27.316	5,04	0,34	0,32	1,04	0,04
>= 2001	735.356	43.292	5,56	0,48	0,52	0,94	-0,0
Total	1.523.991	84.061	5,23	1,00	1,00	1,00	0,00

Tabela 25 - Fabricante do veículo.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
Fab1	166.111	10.339	5,86	0,11	0,12	0,89	-0,1
Fab2	25.003	1.271	4,84	0,02	0,02	1,09	0,08
Fab3	370.543	20.535	5,25	0,24	0,24	1,00	0,00
Fab4	157.884	8.496	5,11	0,10	0,10	1,03	0,02
Fab5	418.675	22.183	5,03	0,27	0,26	1,04	0,04
Fab6	385.797	21.237	5,22	0,25	0,25	1,00	0,00
Total	1.524.013	84.061	5,23	1,00	1,00	1,00	0,00

Tabela 26 - Classe de bônus.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
<= 5	1.015.357	60.217	5,60	0,67	0,72	0,93	-0,0
>5	508.656	23.844	4,48	0,33	0,28	1,18	0,1
Total	1.524.013	84.061	5,23	1,00	1,00	1,00	0,0

Tabela 27 - Procedência do veículo.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	wo
Nacional	1.455.279	80.033	5,21	0,95	0,95	1,00	0,0
Importado	68.734	4.028	5,54	0,05	0,05	0,94	-0,0
Total	1.524.013	84.061	5,23	1,00	1,00	1,00	0,0

Tabela 28 - Origem do cliente.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
Em/R1/R2	756.565	43.844	5,48	0,50	0,52	0,95	-0,05
R3/R4/R5/Rc	767.448	40.217	4,98	0,50	0,48	1,05	0,05
Total	1.524.013	84.061	5,23	1,00	1,00	1,00	0,00

Tabela 29 - Tipo de veículo.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
1	710.508	37.556	5,02	0,47	0,45	1,04	0,04
2/3/4/5/6	813.505	46.505	5,41	0,53	0,55	0,96	-0,04
Total	1.524.013	84.061	5,23	1,00	1,00	1,00	0,00

Tabela 30 - Estado civil do segurado.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
Casado/União Estável	1.209.292	65.490	5,14	0,83	0,83	1,00	0,00
Solteiro/Sep/Div/Viu	249.320	13.837	5,26	0,17	0,17	0,98	-0,02
Total	1.458.612	79.327	5,16	1,00	1,00	1,00	0,00

Tabela 31 - Idade do motorista principal.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
<= De 30 Anos	79.740	5.252	6,18	0,05	0,07	0,82	-0,19
De 31 A 45 Anos	702.771	37.218	5,03	0,48	0,47	1,03	0,03
De 46 A 55 Anos	356.143	20.186	5,36	0,24	0,25	0,96	-0,04
Mais De 55 Anos	324.534	16.815	4,93	0,22	0,21	1,05	0,05
Total	1.463.188	79.471	5,15	1,00	1,00	1,00	0,00

Tabela 32 - Sistema antifurto.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
Não / Ndf	502.217	26.603	5,03	0,33	0,32	1,04	0,04
Sim	1.017.127	57.310	5,33	0,67	0,68	0,98	-0,02
Total	1.519.344	83.913	5,23	1,00	1,00	1,00	0,00

Tabela 33 - Tipo de pintura.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
Comum	348.075	19.628	5,34	0,24	0,24	0,99	-0,01
Metálica / Perol	1.121.189	62.323	5,27	0,76	0,76	1,00	0,00
Total	1.469.264	81.951	5,28	1,00	1,00	1,00	0,00

Tabela 34 - Sexo do motorista principal.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
F	638.479	34.405	5,11	0,44	0,43	1,01	0,01
M	824.733	45.069	5,18	0,56	0,57	0,99	-0,01
Total	1.463.212	79.474	5,15	1,00	1,00	1,00	0,00

Tabela 35 - Sistema de rastreamento do veículo.

Categoria	Apol s/ sin	Apol c/ sin	% sinistro na categoria	% s/ sin	% c/ sin	(% s/ sin) / (% c/ sin)	woe
Instalado	10.286	318	3,00	0,01	0,00	1,78	0,58
Não instalado	1.513.727	83.743	5,24	0,99	1,00	1,00	0,00
Total	1.524.013	84.061	5,23	1,00	1,00	1,00	0,00

APÊNDICE 03 – SAÍDA DO PROC LOGISTIC

The LOGISTIC Procedure

Model Information

Data Set	MODELA.BASE
Response Variable (Events)	SINISTRO
Response Variable (Trials)	EXPO
Number of Observations	13326
Model	binary logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	Binary Outcome	Total Frequency
1	Event	58045
2	Nonevent	675199.1

The LOGISTIC Procedure

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square
1	BON1		1	1	892.7459	.
2	ANO1		1	2	470.7110	.
3	ANO2		1	3	319.4365	.
4	TVEI1		1	4	190.0736	.
5	IDMOT1		1	5	151.3398	.
6	IDMOT3		1	6	129.1892	.
7	FAB1		1	7	123.6283	.
8	FURTO1		1	8	97.6137	.
9	PIN1		1	9	60.1745	.
10	RASTR1		1	10	41.2335	.
11	FAB5		1	11	29.1757	.
12	FAB2		1	12	30.3171	.
13	BON1*IDMOT1		1	13	26.0846	.
14	CPCD1		1	14	21.6091	.
15	CIVIL1		1	15	21.6623	.
16	CIVIL1*IDMOT1		1	16	91.4389	.
17	CIVIL1*IDMOT3		1	17	37.6215	.
18	BON1*PIN1		1	18	14.5522	.
19	FAB5*TVEI1		1	19	13.1656	.
20	BON1*TVEI1		1	20	10.2010	.
21	SEX1		1	21	9.2610	.
22	CIVIL1*SEX1		1	22	116.8759	.
23	IDMOT3*SEX1		1	23	39.4528	.
24	IDMOT1*SEX1		1	24	20.7856	.
25	FAB1*TVEI1		1	25	9.2587	.
26	ANO1*IDMOT1		1	26	4.0877	.
27	IDMOT2		1	27	2.9313	.
28	ANO1*IDMOT2		1	28	14.8113	.
29		ANO1*IDMOT1	1	27	.	1.7755
30	BON1*IDMOT2		1	28	13.2325	.
31	ANO2*IDMOT2		1	29	10.0580	.
32	ANO1*IDMOT3		1	30	7.0637	.
33	CIVIL1*IDMOT2		1	31	3.6818	.
34	IDMOT2*SEX1		1	32	3.3772	.
35	ORIG1		1	33	2.2194	.
36	BON1*CIVIL1		1	34	2.2315	.
37	FURTO1*IDMOT2		1	35	2.1999	.

Summary of Stepwise Selection

Step	Pr > ChiSq
1	<.0001
2	<.0001
3	<.0001
4	<.0001
5	<.0001
6	<.0001
7	<.0001
8	<.0001
9	<.0001
10	<.0001
11	<.0001
12	<.0001
13	<.0001
14	<.0001
15	<.0001
16	<.0001
17	<.0001
18	0.0001
19	0.0003
20	0.0014
21	0.0023
22	<.0001
23	<.0001
24	<.0001
25	0.0023
26	0.0432
27	0.0869
28	0.0001
29	0.1827
30	0.0003
31	0.0015
32	0.0079
33	0.0550
34	0.0661
35	0.1363
36	0.1352
37	0.1380

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	DF	Value	Value/DF	Pr > ChiSq
Deviance	13290	10156.0034	0.7642	1.0000
Pearson	13290	10956.9116	0.8244	1.0000

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.2713	0.0401	3206.2049	<.0001
ANO1	1	-0.5169	0.0254	417.8095	<.0001
ANO2	1	-0.1799	0.0113	251.5436	<.0001
BON1	1	0.3714	0.0256	210.8484	<.0001
CIVIL1	1	-0.1012	0.0322	9.8780	0.0017
CPCD1	1	-0.1118	0.0209	28.5483	<.0001
FAB1	1	0.0593	0.0151	15.3814	<.0001
FAB2	1	-0.2860	0.0358	63.9756	<.0001
FAB5	1	-0.0978	0.0120	65.9556	<.0001
FURTO1	1	-0.0824	0.0114	52.0548	<.0001
IDMOT1	1	0.6740	0.0499	182.2441	<.0001
IDMOT2	1	0.0766	0.0335	5.2214	0.0223
IDMOT3	1	-0.0853	0.0368	5.3749	0.0204
ORIG1	1	0.0137	0.00799	2.9415	0.0863
PIN1	1	-0.1435	0.0172	69.5773	<.0001
RASTR1	1	-0.4919	0.0688	51.1440	<.0001
SEX1	1	-0.1710	0.0269	40.4814	<.0001
TVEI1	1	-0.0915	0.0164	31.2677	<.0001
BON1*TVEI1	1	-0.0598	0.0173	11.9063	0.0006
BON1*PIN1	1	0.0874	0.0201	18.9360	<.0001
BON1*IDMOT1	1	-0.2316	0.0432	28.6861	<.0001
BON1*IDMOT2	1	-0.0731	0.0174	17.7176	<.0001
BON1*CIVIL1	1	-0.0397	0.0232	2.9173	0.0876
CIVIL1*SEX1	1	0.2068	0.0219	89.3066	<.0001
CIVIL1*IDMOT1	1	-0.3190	0.0406	61.8321	<.0001
CIVIL1*IDMOT2	1	-0.0766	0.0297	6.6694	0.0098
CIVIL1*IDMOT3	1	0.1595	0.0360	19.6626	<.0001
FURTO1*IDMOT2	1	-0.0283	0.0167	2.8787	0.0898
IDMOT1*SEX1	1	-0.2253	0.0381	34.9362	<.0001
IDMOT2*SEX1	1	-0.0483	0.0228	4.4989	0.0339
IDMOT3*SEX1	1	0.0737	0.0254	8.4458	0.0037
ANO1*IDMOT2	1	0.1989	0.0327	36.9570	<.0001
ANO1*IDMOT3	1	0.1065	0.0359	8.7882	0.0030
ANO2*IDMOT2	1	0.0582	0.0162	12.9219	0.0003
FAB1*TVEI1	1	0.0976	0.0272	12.8649	0.0003
FAB5*TVEI1	1	0.0879	0.0181	23.5383	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
CPCD1	0.894	0.858	0.932
FAB2	0.751	0.700	0.806
ORIG1	1.014	0.998	1.030
RASTR1	0.611	0.534	0.700

Association of Predicted Probabilities and Observed Responses

Percent Concordant	54.9	Somers' D	0.133
Percent Discordant	41.6	Gamma	0.138
Percent Tied	3.6	Tau-a	0.019
Pairs	39191929136	c	0.566

Partition for the Hosmer and Lemeshow Test

Group	Total	Event		Nonevent	
		Observed	Expected	Observed	Expected
1	73511	3732	3734.08	69779	69776.84
2	72908	4507	4489.33	68401	68418.96
3	74349	5021	5022.78	69328	69326.31
4	73197	5285	5272.76	67912	67924.73
5	74082	5601	5649.78	68481	68431.82
6	72556	5831	5804.84	66725	66751.03
7	72997	6176	6164.24	66821	66832.61
8	73421	6620	6553.32	66801	66867.61
9	73356	7034	7091.48	66322	66264.25
10	72867	8238	8230.16	64629	64637.14

APÊNDICE 04 – SAÍDA DO PROC GENMOD

The GENMOD Procedure

Model Information

Data Set	MODELA.BSE_VLR2
Distribution	Gamma
Link Function	Log
Dependent Variable	VLR100
Observations Used	58269

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	58E3	63714.2289	1.0941
Scaled Deviance	58E3	66890.0806	1.1486
Pearson Chi-Square	58E3	90521.9351	1.5544
Scaled Pearson X2	58E3	95034.0236	1.6319
Log Likelihood		-526913.4007	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	7.8114	0.0364	7.7402	7.8827	46161.6	<.0001
ANO2	1	0.1432	0.0186	0.1068	0.1797	59.39	<.0001
ANO3	1	0.2703	0.0200	0.2310	0.3095	181.96	<.0001
BON1	1	0.0963	0.0093	0.0781	0.1144	108.07	<.0001
CIVIL1	1	0.0520	0.0108	0.0308	0.0732	23.13	<.0001
CPCD1	1	-0.4590	0.0223	-0.5027	-0.4153	423.70	<.0001
FAB1	1	0.2990	0.0335	0.2334	0.3646	79.79	<.0001
FAB3	1	0.2642	0.0297	0.2059	0.3225	78.92	<.0001
FAB4	1	0.2153	0.0308	0.1549	0.2756	48.90	<.0001
FAB6	1	0.1132	0.0425	0.0299	0.1966	7.09	0.0078
FURTO2	1	0.0401	0.0091	0.0223	0.0580	19.40	<.0001
IDMOT2	1	-0.0239	0.0093	-0.0421	-0.0057	6.60	0.0102
IDMOT4	1	-0.0723	0.0115	-0.0948	-0.0499	39.81	<.0001
ORIG2	1	-0.1249	0.0215	-0.1670	-0.0828	33.89	<.0001
PIN2	1	-0.0809	0.0096	-0.0996	-0.0622	71.71	<.0001
SEX2	1	0.3818	0.0712	0.2424	0.5213	28.79	<.0001
TVEI2	1	1.0810	0.0586	0.9661	1.1959	340.09	<.0001
ANO2*FAB6	1	0.1634	0.0340	0.0968	0.2301	23.11	<.0001
ANO3*FAB3	1	-0.1416	0.0208	-0.1823	-0.1009	46.44	<.0001
ANO3*FAB6	1	0.1355	0.0345	0.0680	0.2031	15.45	<.0001
FAB1*SEX2	1	-0.1791	0.0745	-0.3252	-0.0331	5.78	0.0162
FAB1*TVEI2	1	-0.5829	0.0627	-0.7058	-0.4600	86.43	<.0001
FAB3*ORIG2	1	0.0561	0.0258	0.0055	0.1067	4.73	0.0297
FAB3*SEX2	1	-0.3238	0.0731	-0.4670	-0.1805	19.62	<.0001
FAB3*TVEI2	1	-0.7692	0.0598	-0.8865	-0.6520	165.38	<.0001
FAB4*SEX2	1	-0.2049	0.0757	-0.3532	-0.0566	7.33	0.0068
FAB4*TVEI2	1	-0.5992	0.0626	-0.7219	-0.4765	91.61	<.0001
ORIG2*FAB5	1	0.0640	0.0245	0.0160	0.1120	6.84	0.0089
SEX2*FAB5	1	-0.2214	0.0729	-0.3642	-0.0785	9.22	0.0024
TVEI2*FAB5	1	-0.5065	0.0573	-0.6189	-0.3941	78.04	<.0001
FAB6*ORIG2	1	0.0718	0.0256	0.0215	0.1220	7.84	0.0051
FAB6*SEX2	1	-0.3063	0.0732	-0.4498	-0.1628	17.50	<.0001
FAB6*TVEI2	1	-0.6892	0.0597	-0.8062	-0.5723	133.43	<.0001
ORIG2*TVEI2	1	-0.0567	0.0183	-0.0925	-0.0209	9.64	0.0019
Scale	1	1.0498	0.0054	1.0392	1.0606		

NOTE: The scale parameter was estimated by maximum likelihood.

APÊNDICE 05 – VARIÁVEIS DUMMY

Tabela 36 - Dummy Ano e modelo.

Variável	Categoria	Dummy	
		Ano1	Ano2
Ano	<= 1995	1	0
	1996 – 2000	0	1
	>= 2001	0	0

Tabela 37 - Dummy Classe de bônus.

Variável	Categoria	Dummy
		Bon1
Bonus	<= 5	1
	>5	0

Tabela 38 - Dummy Estado civil.

Variável	Categoria	Dummy
		Civil1
Civil	Casado/União Estável	1
	Solteiro/Sep/Div/Viu	0

Tabela 39 - Dummy Procedência.

Variável	Categoria	Dummy
		Cpcd1
Proced	Nacional	1
	Importado	0

Tabela 40 - Dummy Fabricante.

Variável	Categoria	Dummy				
		Fab1	Fab2	Fab3	Fab4	Fab5
Fabr	Fabricante 1	1	0	0	0	0
	Fabricante 2	0	1	0	0	0
	Fabricante 3	0	0	1	0	0
	Fabricante 4	0	0	0	1	0
	Fabricante 5	0	0	0	0	1
	Fabricante 6	0	0	0	0	0

Tabela 41 - Dummy Sistema antifurto.

Variável	Categoria	Dummy Furto1
Furto	Não / Ndf	1
	Sim	0

Tabela 42 - Dummy Idade do motorista.

Variável	Categoria	Dummy Idmot1	Idmot2	Idmot3
Idmot	<= De 30 Anos	1	0	0
	De 31 A 45 Anos	0	1	0
	De 46 A 55 Anos	0	0	1
	Mais De 55 Anos	0	0	0

Tabela 43 - Dummy Origem do cliente.

Variável	Categoria	Dummy Orig1
Origem	Em/R1/R2	1
	R3/R4/R5/Rc	0

Tabela 44 - Dummy Tipo de pintura.

Variável	Categoria	Dummy Pin1
Pin	Comum	1
	Metálica / Perol	0

Tabela 45 - Dummy Sistema de rastreamento.

Variável	Categoria	Dummy Rastr1
Daf	Instalado	1
	Não instalado	0

Tabela 46 - Dummy Sexo do motorista.

Variável	Categoria	Dummy Sex1
Sexo	Feminino	1
	Masculino	0

Tabela 47 - Dummy Tipo de veículo.

Variável	Categoria	Dummy Tveil
Tipvei	1	1
	2/3/4/5/6	0

APÊNDICE 06 – ÁRVORE DE DECISÃO PARA FREQÜÊNCIA DE SINISTRO

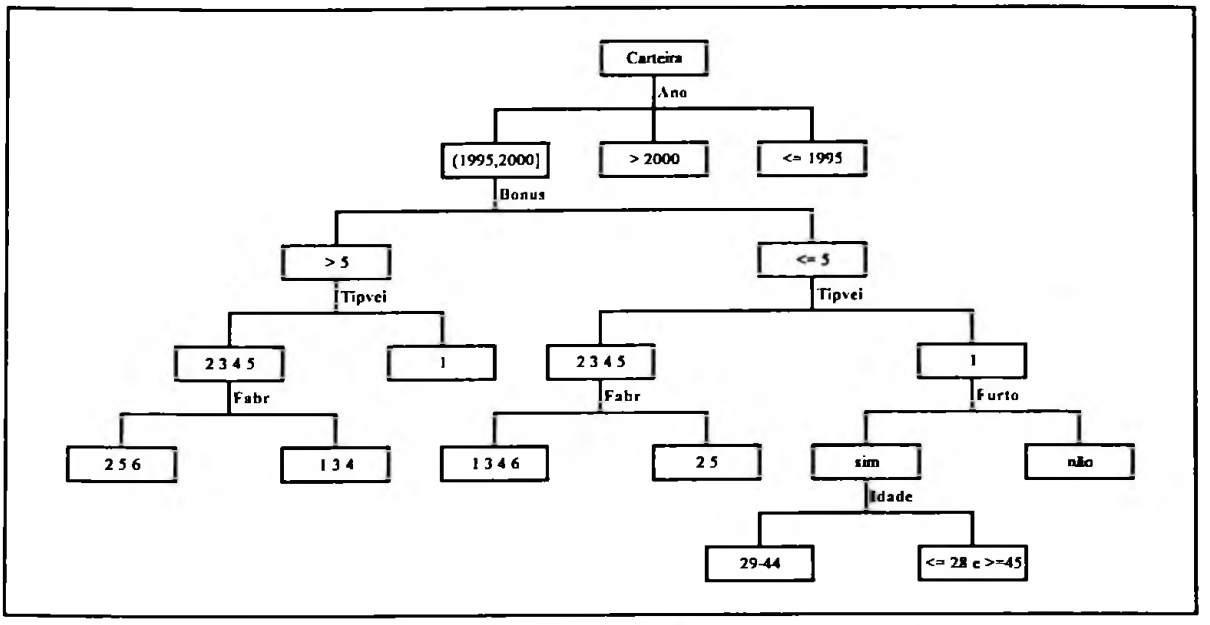

Gráfico 2 - Árvore de decisão para freqüência de sinistro (ramo 1).

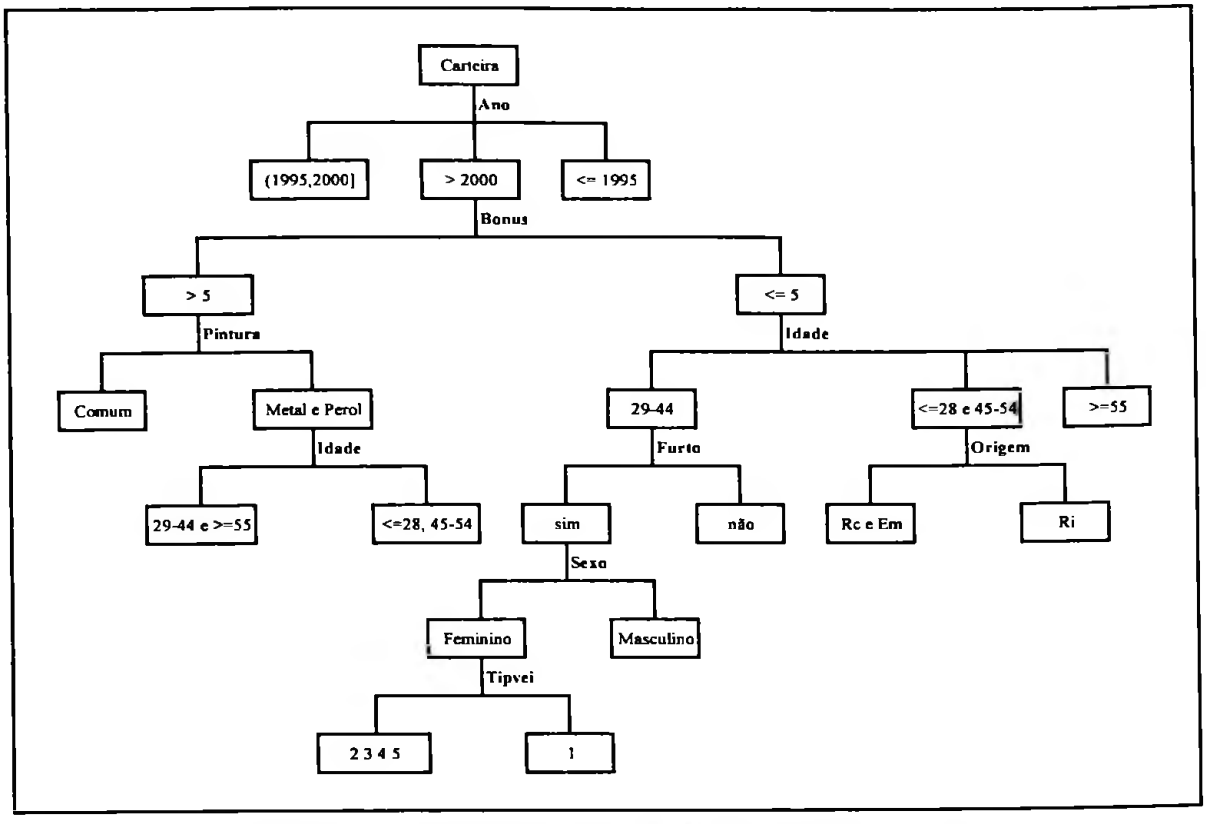

Gráfico 3 - Árvore de decisão para freqüência de sinistro (ramo 2).

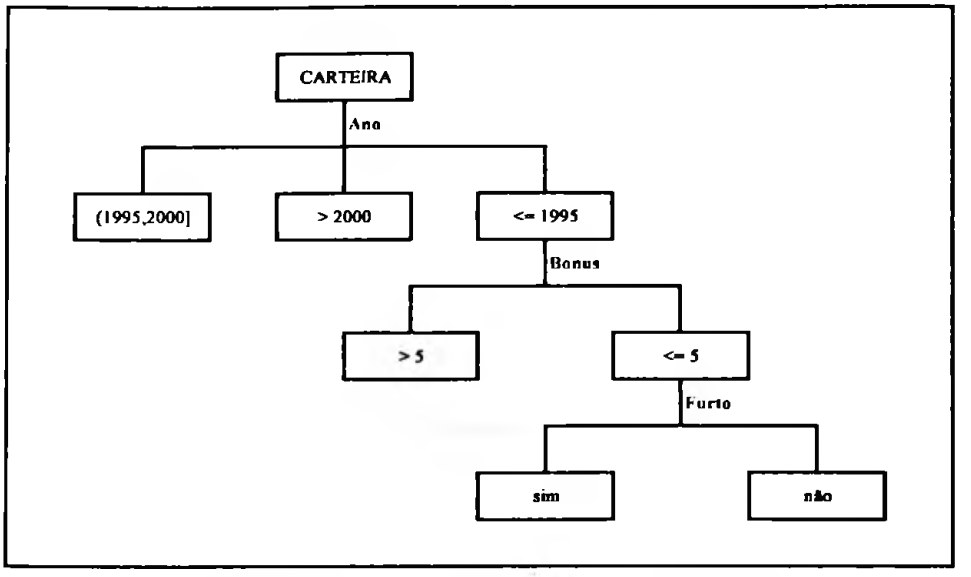

Gráfico 4 - Árvore de decisão para freqüência de sinistro (ramo 3).

APÊNDICE 07 – ÁRVORE DE DECISÃO PARA SEVERIDADE

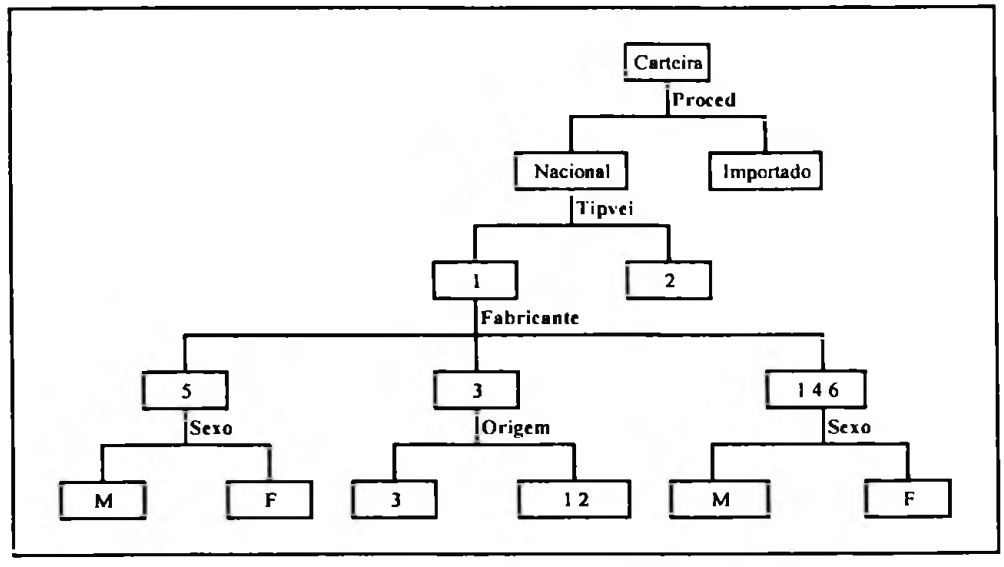

Gráfico 5 - Árvore de decisão para severidade (ramo 1).

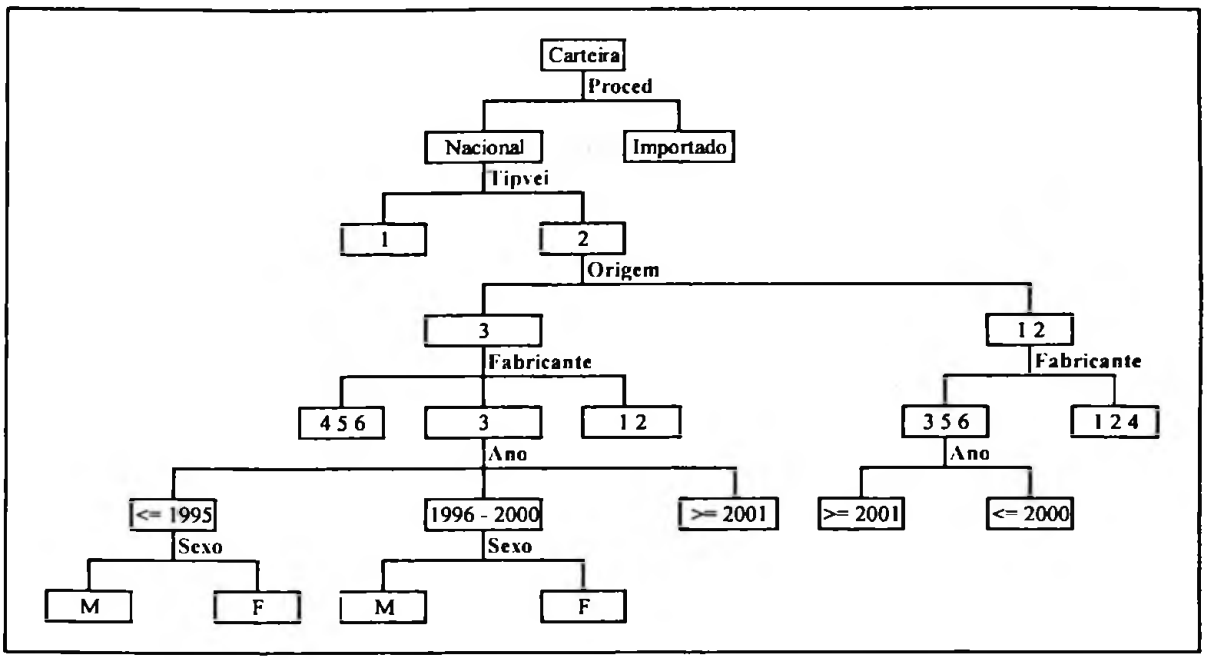

Gráfico 6 - Árvore de decisão para severidade (ramo 2).

APÊNDICE 08 – RAZÃO DE CHANCES DO MODELO LOGÍSTICO

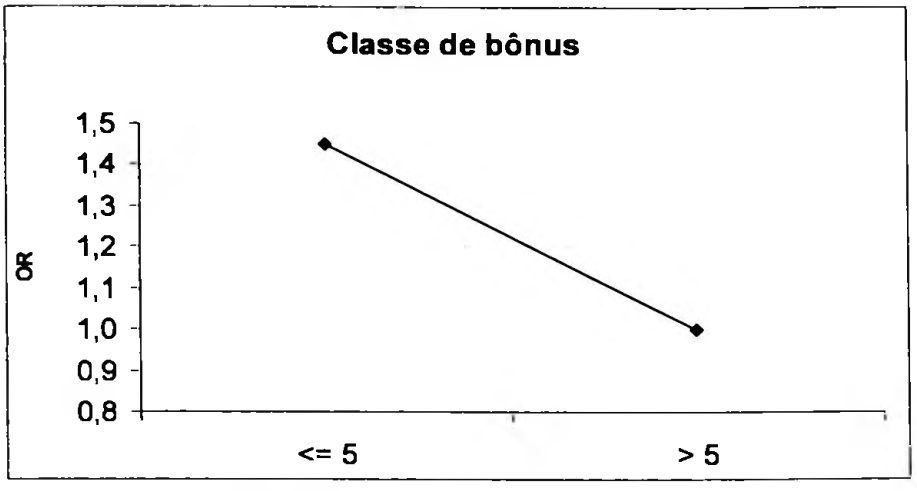

Gráfico 7: Razão de chances - Classe de bônus.

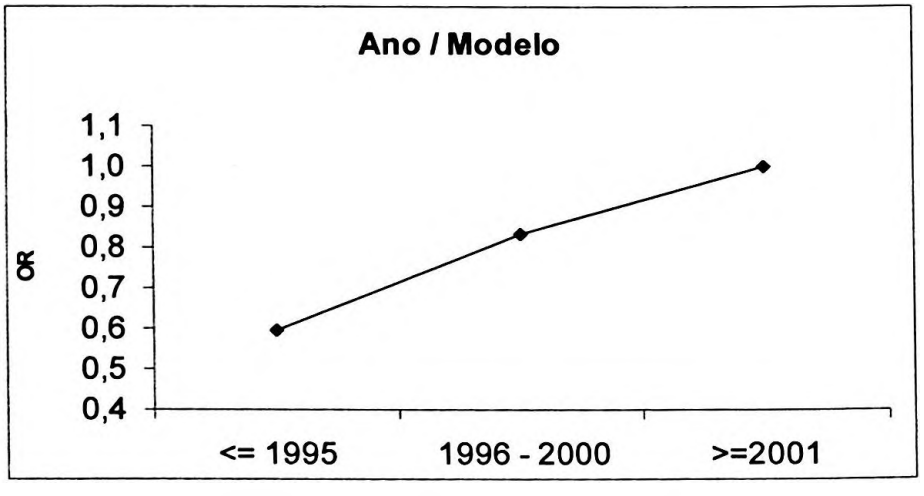

Gráfico 8: Razão de chances – Ano / Modelo.

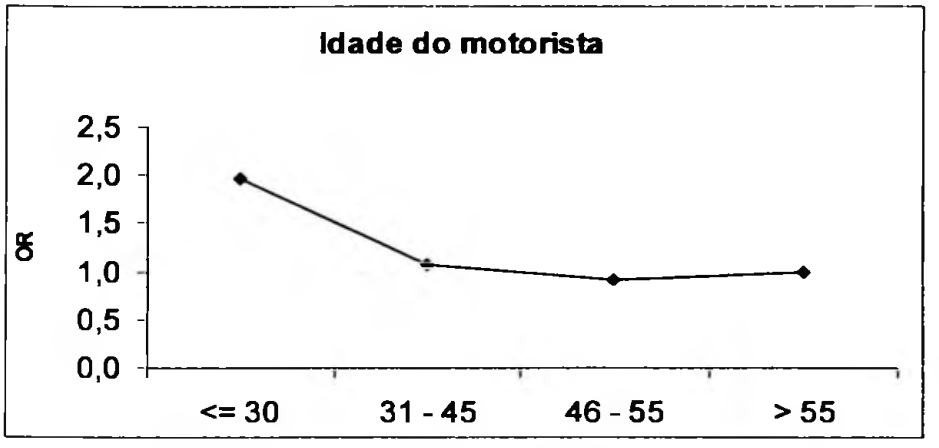

Gráfico 9: Razão de chances – Idade do motorista.

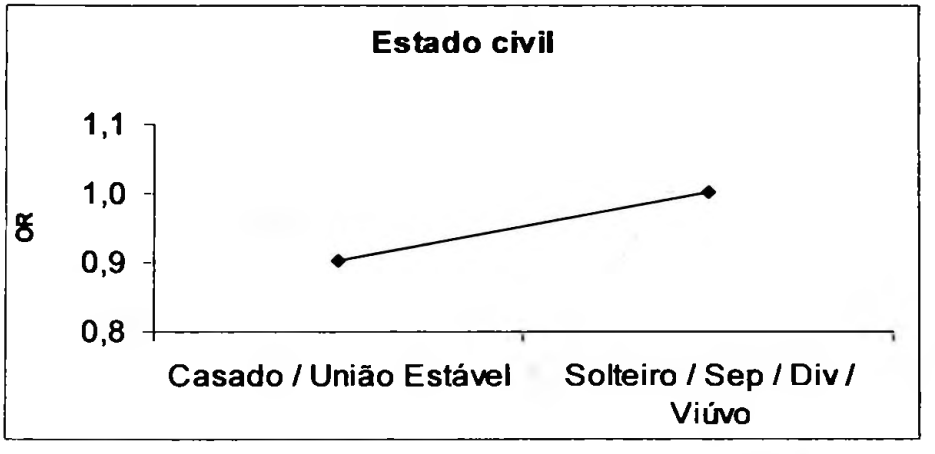

Gráfico 10: Razão de chances – Estado civil.

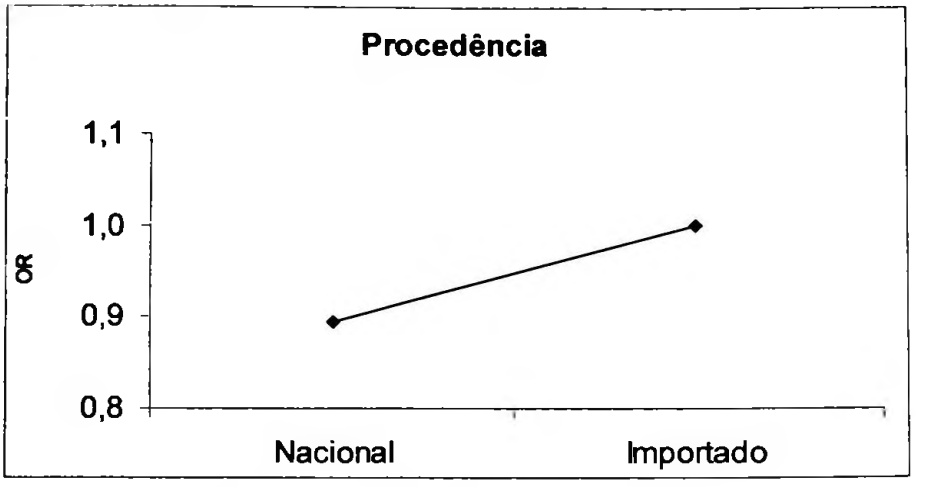

Gráfico 11: Razão de chances – Procedência.

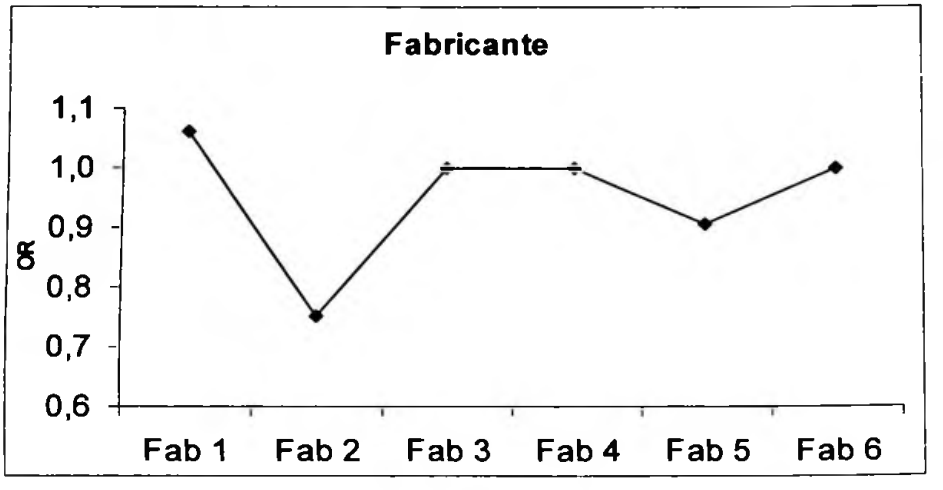

Gráfico 12: Razão de chances – Fabricante.

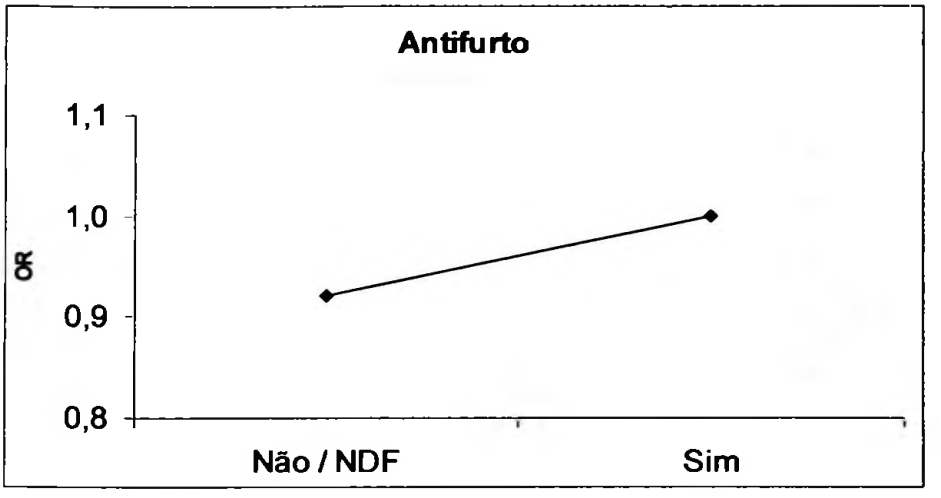

Gráfico 13: Razão de chances – Dispositivo antifurto.

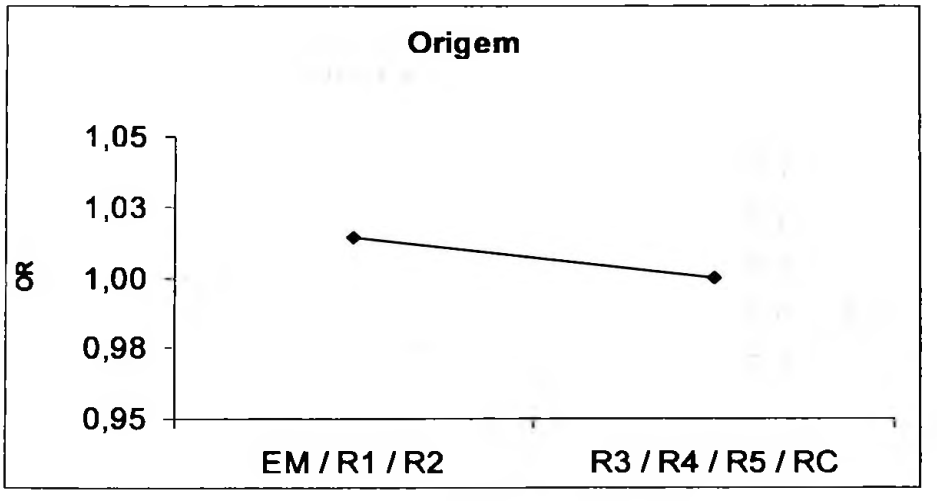

Gráfico 14: Razão de chances – Origem.

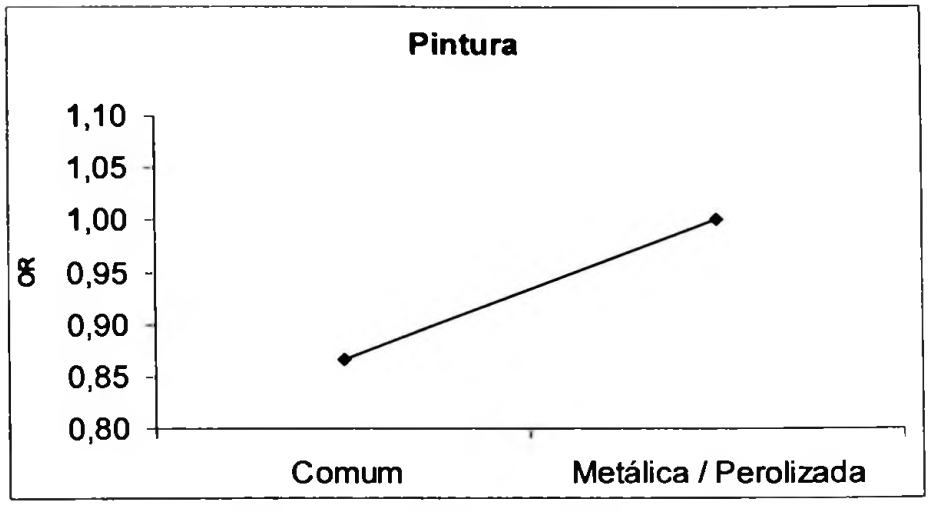

Gráfico 15: Razão de chances – Tipo de pintura.

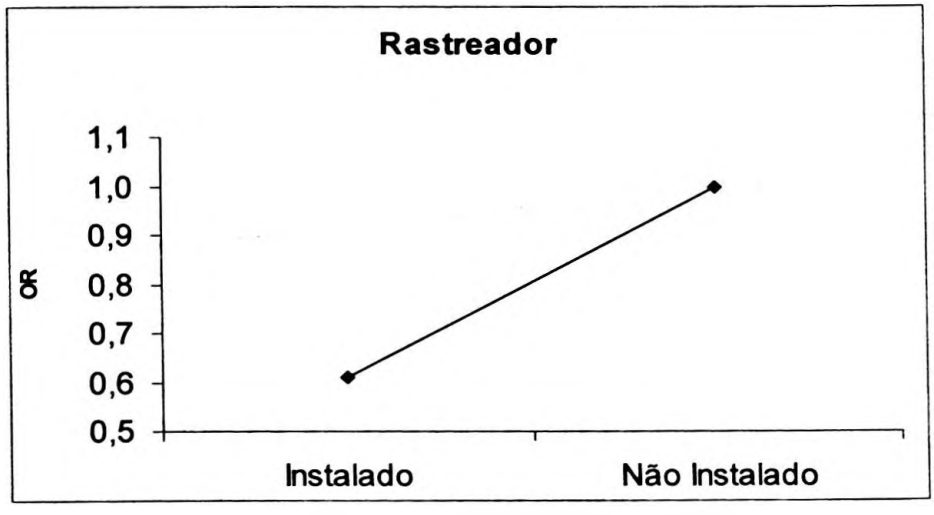

Gráfico 16: Razão de chances – Rastreador.

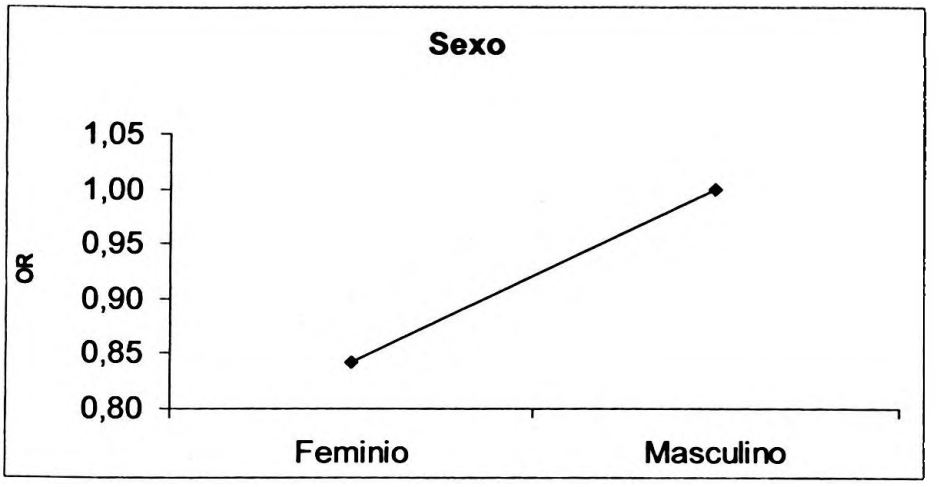

Gráfico 17: Razão de chances – Sexo do motorista.

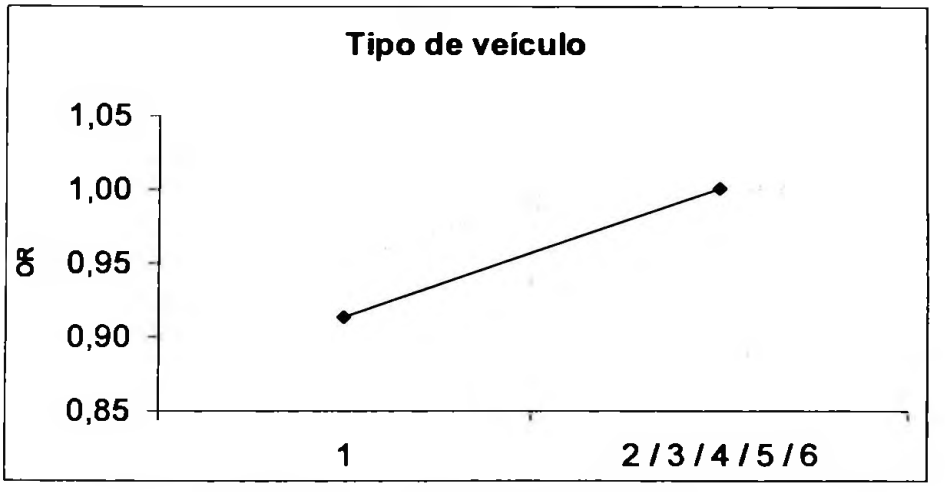

Gráfico 18: Razão de chances – Tipo de veículo.

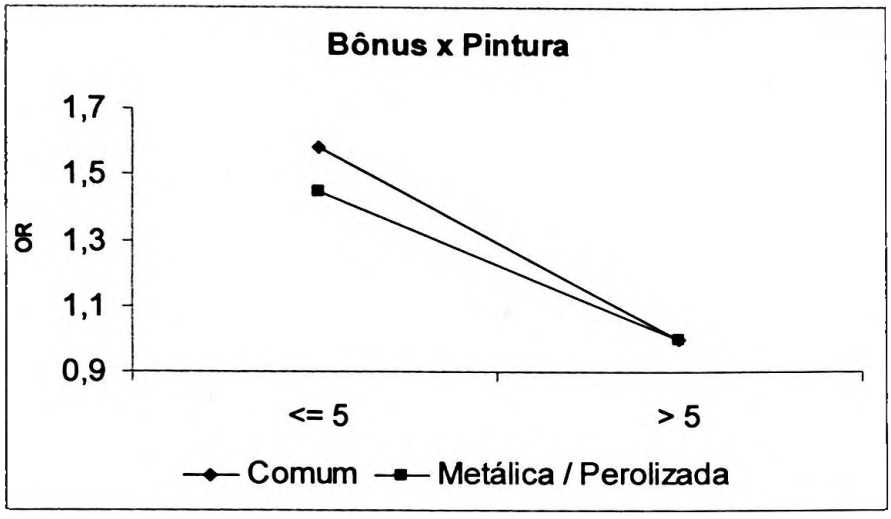

Gráfico 19: Razão de chances – Interação entre Bônus e tipo de Pintura.

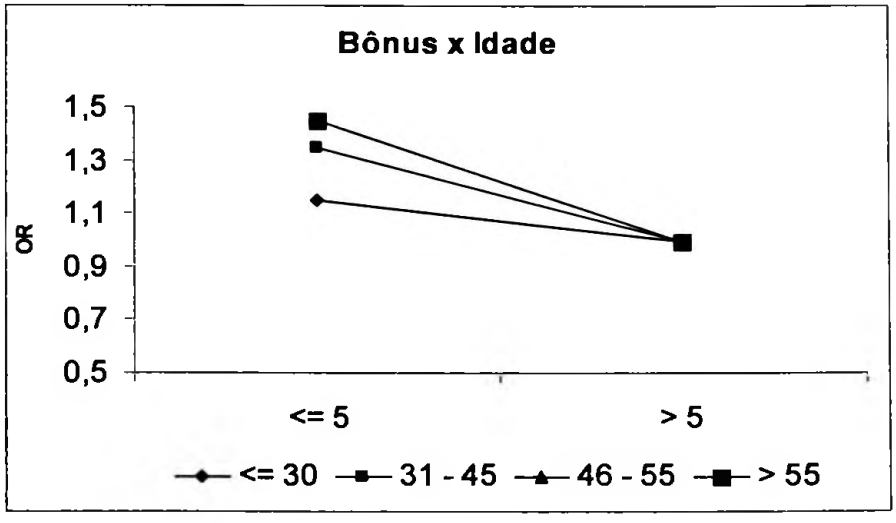

Gráfico 20: Razão de chances – Interação entre Bônus e Idade.

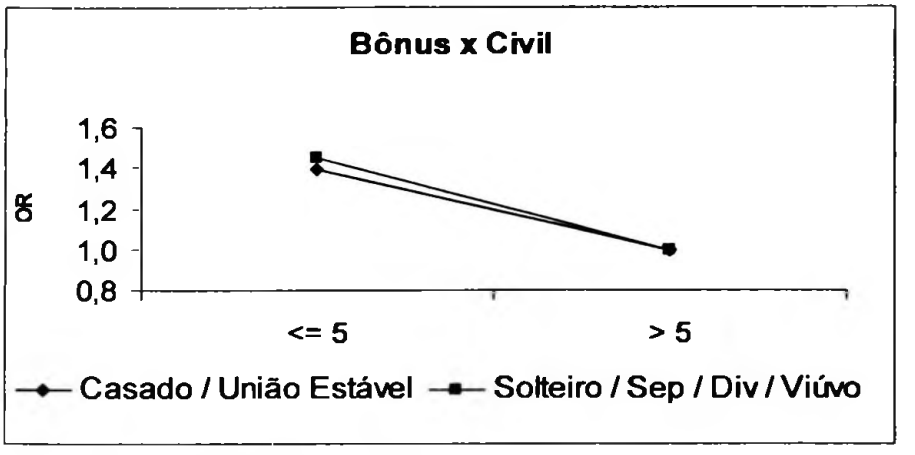

Gráfico 21: Razão de chances – Interação entre Bônus e Estado civil.

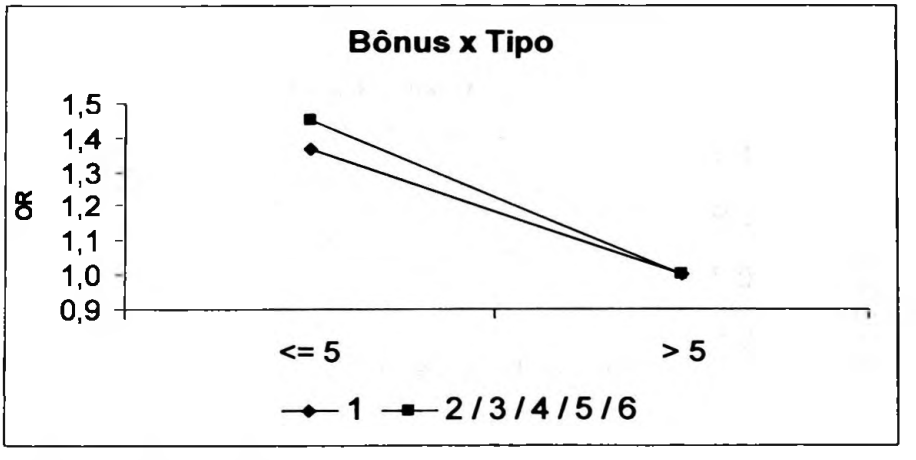

Gráfico 22: Razão de chances – Interação entre Bônus e Tipo de veículo.

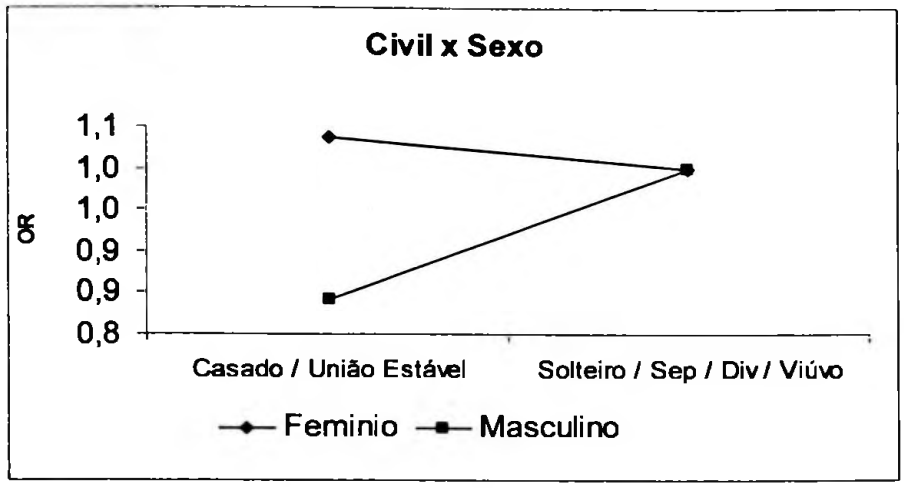

Gráfico 23: Razão de chances – Interação entre Estado civil e Sexo.

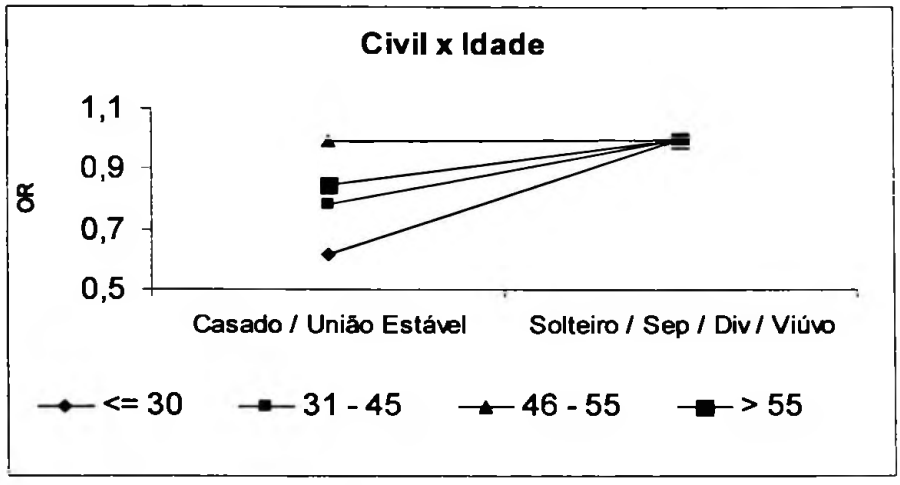

Gráfico 24: Razão de chances – Interação entre Estado civil e Idade.

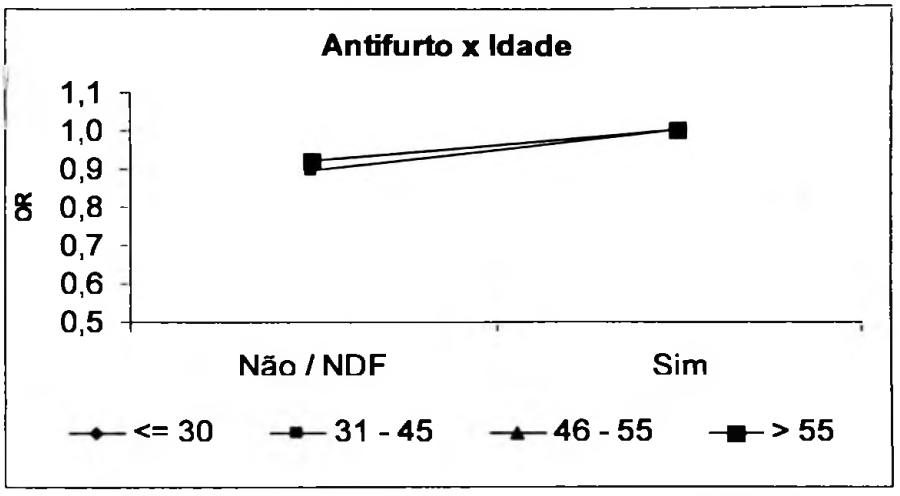

Gráfico 25: Razão de chances – Interação entre Dispositivo antifurto e Idade.

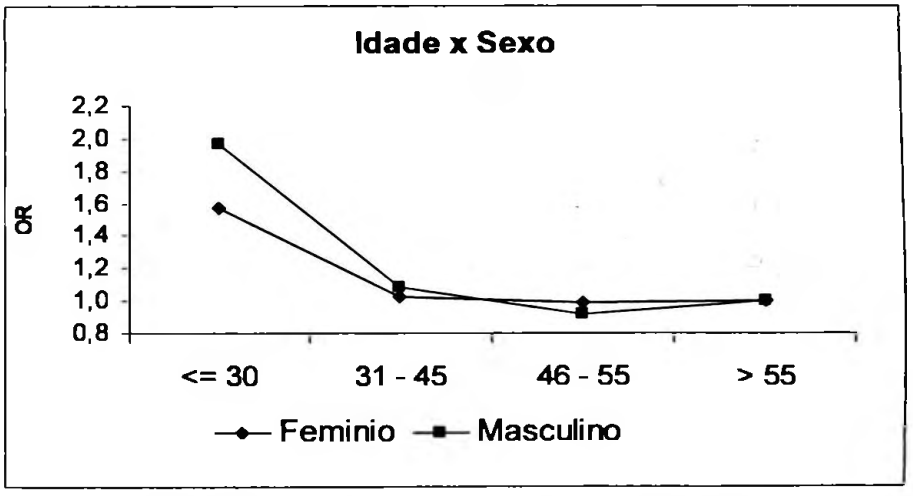

Gráfico 26: Razão de chances – Interação entre Idade e Sexo.

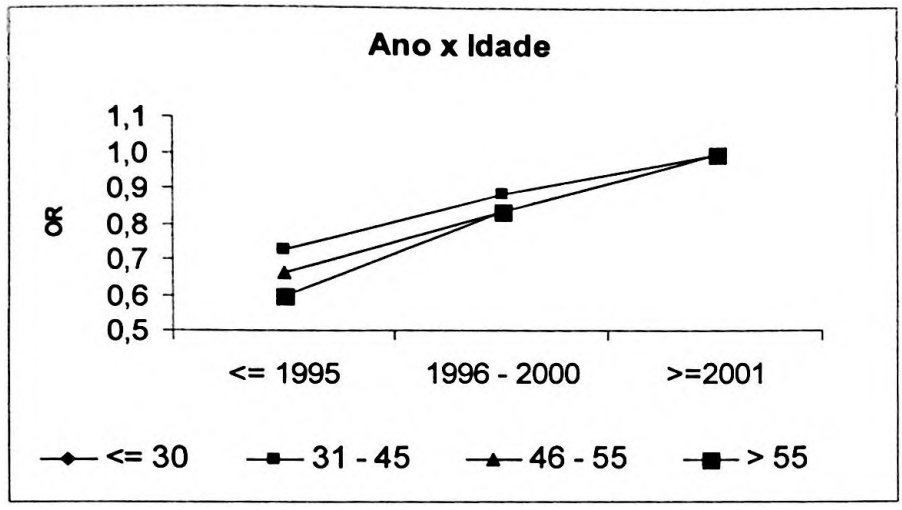

Gráfico 27: Razão de chances – Interação entre Ano / Modelo e Idade.

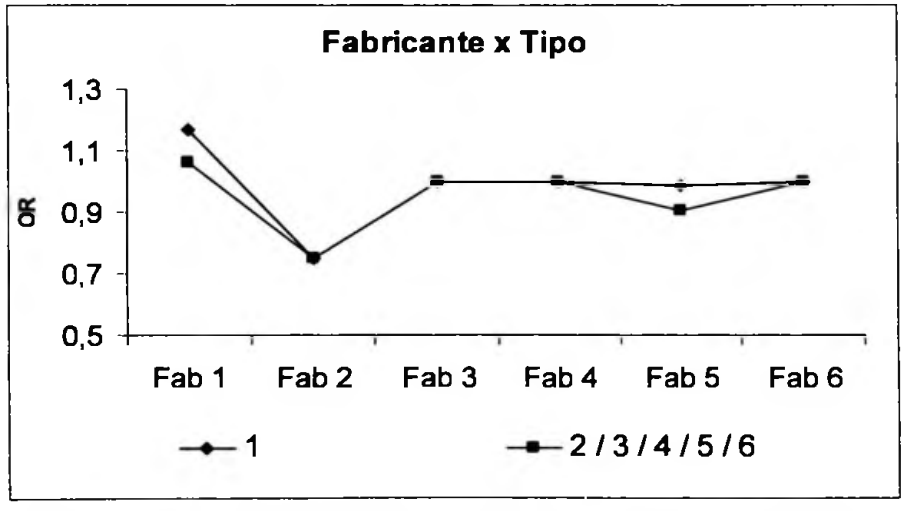

Gráfico 28: Razão de chances – Interação entre Fabricante e Tipo de veículo.