

FUNDAÇÃO GETULIO VARGAS
ESCOLA de PÓS-GRADUAÇÃO em
ECONOMIA

Cristiana Caldeira Garcia de Freitas

Demanda por Seguro de Automóvel no
Rio de Janeiro

Rio de Janeiro
2018

Cristiana Caldeira Garcia de Freitas

Demanda por Seguro de Automóvel no Rio de Janeiro

Dissertação para obtenção do grau
de mestre apresentada à Escola de
Pós-Graduação em Economia

Orientador: Luis Henrique Bertolino
Braido

Rio de Janeiro
2018

Freitas, Cristiana Caldeira Garcia de
Demanda por seguro de automóvel no Rio de Janeiro / Cristiana
Caldeira Garcia de Freitas. – 2018.
63 f.

Dissertação (mestrado) - Fundação Getulio Vargas, Escola de Pós-
Graduação em Economia.

Orientador: Luis Henrique Bertolino Braido.

Inclui bibliografia.

1. Seguro de automóveis. 2. Seguros - Modelos econômicos.
3. Oferta e procura. I. Braido, Luís H. B. II. Fundação Getulio
Vargas. Escola de Pós-Graduação em Economia. III. Título.

CDD – 330

CRISTIANA CALDEIRA GARCIA DE FREITAS

“DEMANDA POR SEGURO DE AUTOMÓVEL NO RIO DE JANEIRO”.

Dissertação apresentado(a) ao Curso de Mestrado em Economia do(a) Escola de Pós-Graduação em Economia para obtenção do grau de Mestre(a) em Economia.

Data da defesa: 27/03/2018

ASSINATURA DOS MEMBROS DA BANCA EXAMINADORA



Luis Henrique Bertolino Braido
Orientador(a)



Marcelo Castello Branco Sant'Anna



Rodrigo dos Santos Targino

Resumo

Esta dissertação tem como objetivo estimar a demanda por seguro de automóvel na cidade do Rio de Janeiro. Em razão da disponibilidade de dados a nível do consumidor, foi feita uma breve revisão de modelos de escolha discreta, com destaque para o Logit com coeficientes aleatórios. Para ser possível a efetiva implementação dos modelos, foi preciso lidar com uma peculiaridade do setor de seguros: os preços variam de acordo com o indivíduo. Como só temos os preços efetivos da apólice contratada, foi testado um modelo *random forest* para gerar valores esperados de prêmio das seguradoras.

Palavras-chave: demanda, escolha discreta, seguro de automóvel

Abstract

This thesis aims to estimate the demand for auto insurance in the city of Rio de Janeiro. Due to the availability of consumer level data, a brief review of discrete choice models was made, focusing on Logit with random coefficients. In order to effectively implement the models, we had to deal with a peculiarity of the insurance industry: prices vary according to each individual. As we only have the effective prices of the contracted policy, a random forest model was tested to generate expected premium values for the insurers.

Keywords: demand, discrete choice, auto insurance

Lista de Figuras

1	Distribuição das Operações entre os Principais Ramos de Seguro . . .	1
2	Floresta Aleatória (Classificação)	12
3	Exemplo de Árvore Aleatória	13
4	Sequencialidade de Sinistros	22
5	Importância das Variáveis Usadas no Random Forest Para Explicar Prêmio	28

Lista de Tabelas

1	Número de Ocorrências por Tipo de Sinistro	20
2	Gastos Médios por Tipo de Sinistro	20
3	Comparação dos Erros Quadráticos Médios do Custo Estimado	24
4	Resultados da Regressão Prêmio X Custo Esperado	25
5	Resultados da Regressão de Prêmio	27
6	Prêmio Médio e <i>Market Share</i> por Seguradora	29
7	Resultados do Modelo Logit Misto	30
8	Resultados do Modelo Logit com Coeficientes Aleatórios	32
9	Elasticidades-Preço de Substituição: Logit Misto	33
10	Elasticidades-Preço de Substituição: Logit com Coeficientes Aleatórios	33
11	R_Auto: Descrição das Variáveis	36
12	S_Auto: Descrição das Variáveis	38

Sumário

1	Introdução	1
2	Literatura	3
2.1	Modelos de Escolha Discreta	3
2.1.1	Logits	3
2.1.2	Independência das Alternativas Irrelevantes	7
2.1.3	Logit com Coeficientes Aleatórios	8
2.2	Custo Esperado e Prêmio	10
2.2.1	Random Forests	10
3	Dados	15
4	Modelo para Prêmio	17
4.1	Custo Esperado	18
4.1.1	Especificação dos Tipos de Sinistro	19
4.1.2	Modelos Testados	21
4.1.3	Comparação do Modelos	24
4.2	Prêmio Esperado	26
5	Estimação da Demanda	29
5.1	Logit Misto	30
5.2	Logit com Efeitos Aleatórios	30
5.3	Elasticidades	32

6	Considerações Finais	34
A	Anexos	36
B	Código R	40

1 Introdução

Esta dissertação tem como objetivo estimar a demanda por seguro de automóvel na cidade do Rio de Janeiro, como um primeiro passo para a compreensão deste mercado. O setor de seguros no Brasil ainda é pouco estudado, dado seu tamanho e constante crescimento.¹ Em 2003, ano estudado neste trabalho, o segmento de automóveis era o mais expressivo em volume de prêmio dentre os demais ramos de seguro, conforme ilustrado na Figura 1.

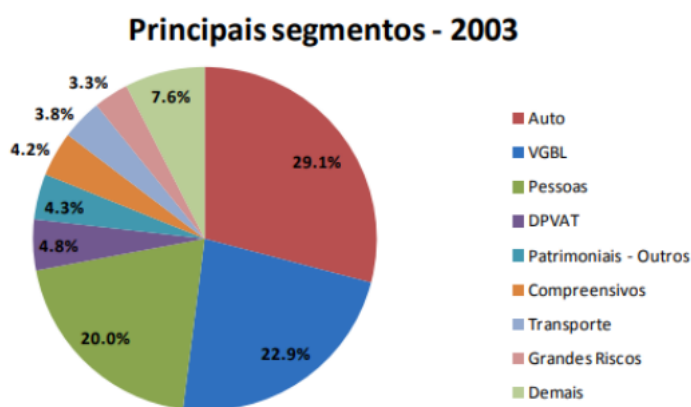


Figura 1: Distribuição das Operações entre os Principais Ramos de Seguro

Fonte: Superintendência de Seguros Privados (SUSEP)

O mercado de seguros de automóvel no Brasil apresenta alta complexidade. Existe uma estrutura vertical, em que as seguradoras vendem seus produtos por meio de corretores licenciados pela SUSEP. Os dados revelam algumas peculiaridades, como a existência de uma fração considerável de apólices vendida com taxa de corretagem nula coexistindo com uma grande dispersão entre as taxas positivas. Merece destaque o artigo de Braido e Ledo (2018), que constroem um modelo dinâmico de competição em preços com custos de busca e troca, visando racionalizar esse fato estilizado. Do lado da demanda, no entanto, não há nenhum estudo até então conhecido.

Busquei fazer uma revisão de modelos de escolha discreta derivados a partir de

¹Dados e tabelas descritivas referentes aos diversos setores de seguro no Brasil estão disponíveis em <http://www.susep.gov.br>.

uma hipótese de maximização de utilidade por parte do agente tomador de decisão. O foco foi dado para os modelos logit e nos estudos realizados por McFadden (1974), McFadden e Train (2000), Train (2009) e Greene (2011). O logit com coeficientes aleatórios assume um papel importante por relaxar a hipótese de independência das alternativas irrelevantes (IIA), gerando padrões de substituição mais realistas.

Para viabilizar a efetiva implementação do modelo, foi preciso lidar com uma peculiaridade do setor de seguros: os preços variam de acordo com o indivíduo. Em geral, quando temos dados de escolha a nível do consumidor, esse não é um problema, já que os preços do mesmo bem costumam ser constantes em um mesmo mercado. Não é possível rodar um modelo logit condicional ou com efeitos aleatórios se não conhecermos as características de todos os produtos disponíveis na cesta de escolhas de cada indivíduo. Como só temos os preços efetivos da apólice contratada, foi testado um modelo *random forest* para gerar valores esperados de prêmio de todas as seguradoras para cada consumidor. A principal vantagem do método é captar relações não lineares entre as variáveis, o que será importante para evitar posteriores problemas de multicolinearidade no modelo de demanda.

Os valores de prêmio estimados pelo *random forest* são então utilizados para o exercício de estimação de demanda via logit com coeficientes aleatórios. Computou-se, por fim, as elasticidades da demanda em relação aos preços para obter uma melhor interpretação econômica dos padrões de substituição vigentes no mercado em questão.

A próxima seção faz uma breve revisão da literatura, descrevendo os modelos relevantes no contexto dessa dissertação. A seção 3 apresenta as bases de dados utilizadas e relata alguns tratamentos e filtros realizados na mesma. A seção 4 detalha o modelo de prêmio, incluindo a construção da variável de custo esperado. A seção 5 apresenta os resultados obtidos para a estimação de demanda e elasticidades. A seção 6 conclui.

2 Literatura

2.1 Modelos de Escolha Discreta

Esta primeira subseção da revisão de literatura tem o propósito de discutir brevemente os métodos de estimação dos parâmetros da demanda em mercados de produtos diferenciados, com uma abordagem de escolha discreta.

Para que um modelo se encaixe na estrutura de escolha discreta, o conjunto de alternativas precisa exibir três características, enumeradas por Train (2009). Em primeiro lugar, as opções devem ser mutuamente exclusivas da perspectiva do tomador de decisão, que escolhe apenas uma alternativa do conjunto de opções. Em segundo lugar, esse conjunto deve ser exaustivo, na medida em que todas as alternativas possíveis estão incluídas e o decisor escolhe necessariamente uma delas. Em terceiro lugar, o número de alternativas deve ser finito.

Vamos construir uma visão geral desses modelos sob a ótica do comportamento individual maximizador de utilidade. O foco vai ser o modelo logit com coeficientes aleatórios, passando, primeiramente, por uma revisão dos modelos que o antecederam e são úteis para a compreensão de sua formulação, vantagens e deficiências.

2.1.1 Logits

Descreverei o modelo logit binário e depois algumas de suas extensões para tratar variáveis dependentes multicategóricas. Várias dessas versões envolvendo múltiplas alternativas de escolha foram desenvolvidas por McFadden nos anos 70, com destaque para o trabalho de 1974, em que ele completa a análise da relação entre a formulação logit e a distribuição da utilidade não observada, mostrando que a probabilidade de escolha na forma logística implica que a utilidade subjacente tem distribuição de valor extremo.

Em geral, modelos de escolha discreta são derivados de um modelo de utilidade aleatória em que os agentes seguem um comportamento de maximização da própria utilidade. Considera-se aqui que a utilidade do agente i , ao escolher a alternativa j

pode ser decomposta em uma parte observável (utilidade representativa) $V_{i,j}$ e outra, conhecida pelo agente mas não observável pelo econometrista $\epsilon_{i,j}$, tratada como sendo aleatória:

$$U_{i,j} = V_{i,j} + \epsilon_{i,j}. \quad (1)$$

A utilidade representativa $V_{i,j}(X_{i,j}, Z_i)$ será especificada como uma função linear nos parâmetros. O vetor Z_i representa as características dos consumidores e $X_{i,j}$ se refere às alternativas. No caso multinomial, por exemplo, temos:

$$V_{i,j} = Z_i' \beta_j. \quad (2)$$

O modelo logit é obtido assumindo que cada $\epsilon_{i,j}$ é independente e identicamente distribuído (IID) segundo uma distribuição de valor extremo. A distribuição acumulada de cada componente não observado é definida da seguinte forma:

$$F(\epsilon_{i,j}) = \exp(-\exp(-\epsilon_{i,j})). \quad (3)$$

No caso mais simples, o modelo binário, a variável de escolha Y assume valores dicotômicos, um para a ocorrência do evento e outro para a não-ocorrência do evento. Se trata de um modelo linear generalizado onde a variável resposta tem distribuição de Bernoulli (ou Binomial) e a função de ligação é a função logística. A probabilidade de ocorrência do evento é dada pela seguinte fórmula:

$$Pr_i = E(Y_i = 1|X_i) = \frac{1}{1 + \exp(-V_{i,j})} = \frac{\exp(V_{i,j})}{1 + \exp(V_{i,j})}. \quad (4)$$

Conforme vemos na equação acima, a probabilidade em questão é não linear nas variáveis X e nos parâmetros β . Para podermos realizar uma estimação via método dos mínimos quadrados ordinários, são necessárias transformações de forma a remover as restrições de intervalo e criar uma função linear das variáveis preditoras. Por isso, em vez de trabalhar com Pr_i diretamente, usamos os seguintes artifícios algébricos:

$$\frac{Pr_i}{1 - Pr_i} = \exp(V_{i,j}) \quad (5)$$

e

$$\text{logit}(Pr_i) = \log\left(\frac{Pr_i}{1 - Pr_i}\right) = V_{i,j}. \quad (6)$$

Dessa forma, no modelo de regressão logística, assume-se que o logit (log dos *odds*) da probabilidade subjacente é uma função linear das variáveis explicativas.

Os coeficientes $\beta_0, \beta_1, \dots, \beta_p$ são estimados com o método da máxima verossimilhança e determinam o efeito que as variáveis associadas exercem sobre as chances de um evento ocorrer.

Logit Multinomial

Consideraremos aqui $J > 2$ possíveis escolhas e que variáveis dependentes Z_i são características específicas do indivíduo, constantes ao longo das alternativas. Escolhemos uma classe como nível de referência e regredimos separadamente as outras $J-1$ contra a referência.

$$\begin{aligned} \log\left(\frac{Pr(Y_i = 1|Z_i)}{Pr(Y_i = J|Z_i)}\right) &= Z_i' \beta_1 \\ &\vdots \\ \log\left(\frac{Pr(Y_i = J-1|Z_i)}{Pr(Y_i = J|Z_i)}\right) &= Z_i' \beta_{J-1}. \end{aligned} \quad (7)$$

Aplicando a função exponencial e usando o fato de que as J probabilidades devem somar 1, chega-se à fórmula:

$$Pr(Y_i = j|Z_i) = \frac{\exp(Z_i' \beta_j)}{1 + \sum_{l=1}^J \exp(Z_i' \beta_l)}. \quad (8)$$

Para escolhas $j = 1, \dots, J - 1$. E, para a escolha dada como nível de referência:

$$Pr(Y_i = 0|Z_i) = \frac{1}{1 + \sum_{l=1}^J \exp(Z_i' \beta_l)}, \quad (9)$$

onde β_l são parâmetros específicos da escolha.

Esse modelo leva a uma função de verossimilhança bem comportada e fácil de estimar. O logit multinomial pode ser visto como um caso especial do logit condicional, como veremos a seguir.

Logit Condicional

É uma extensão do logit multinomial, onde as variáveis dependentes divergem entre as possíveis escolhas de cada indivíduo. É útil quando as variáveis explicativas podem incluir atributos do produto, e não apenas características do agente tomador de decisão. A probabilidade de que o indivíduo i escolha a alternativa j é dada por:

$$Pr(Y_i = j|X_{i,1}, \dots, X_{i,J}) = \frac{\exp(X_{i,j}'\beta)}{\sum_{l=1}^J \exp(X_{i,l}'\beta)}, \quad (10)$$

para $j = 1, \dots, J$.

Aqui o vetor de parâmetros β é comum a todas as escolhas, enquanto as variáveis $X_{i,j}$ são específicas para cada escolha j . Os *odds* de um indivíduo escolher a opção j e não a opção k são dados em função da diferença entre $X_{i,j}$ e $X_{i,k}$:

$$\log \left(\frac{Pr(Y_i = j|X_{i,1}, \dots, X_{i,K})}{Pr(Y_i = k|X_{i,1}, \dots, X_{i,K})} \right) = [X_{i,j} - X_{i,k}]' \beta. \quad (11)$$

Logit Misto

O logit misto é uma combinação dos modelos multinomial e condicional, nos permitindo examinar como as características específicas do indivíduo i e do produto j afetam a probabilidade de escolha do bem j :

$$Pr(Y_i = j | X_{i,1}, \dots, X_{i,J}, Z_i) = \frac{\exp(X'_{i,j}\beta + Z'_i\beta_j)}{\sum_{l=1}^J \exp(X'_{i,l}\beta + Z'_i\beta_l)}. \quad (12)$$

Os logits multinomial, condicional e misto são diferentes formas do mesmo modelo, com variações na indexação das variáveis e parâmetros para se adequar aos diferentes tipos de questionamentos e dados. Uma das vantagens deste tipo de modelo é que a função log-máxima verossimilhança com as probabilidades de escolha na forma exibida abaixo é globalmente côncava nos parâmetros β , conforme demonstrado por McFadden (1974). É importante destacar que o que estou chamando de logit misto não é o que a literatura por vezes denomina de *mixed logit* e se refere a modelos com coeficientes aleatórios, os quais ainda descreverei nesta seção.

2.1.2 Independência das Alternativas Irrelevantes

O principal problema dos modelos logits até então analisados é a propriedade conhecida como independência das alternativas irrelevantes (IIA). Para quaisquer duas alternativas j e k , a razão das probabilidades logit será dada por $\exp(V_{i,j} - V_{i,k})$, conforme demonstrado na equação 11 anteriormente descrita, que não depende de nenhuma das outras possibilidades contidas no conjunto de escolha. Essa propriedade os torna pouco atrantes do ponto de vista estrutural, pois geram padrões de substituição irrealistas.

IIA vai ser violada se os agente tomadores de decisão considerarem alternativas como substitutos ou se houverem variáveis omitidas que são comuns à algumas das possíveis escolhas. McFadden (1974), usa um exemplo que ficou famoso como *red bus/blue bus* para ilustrar a limitação dos modelos que dependem da IIA. Ele supõe um cenário em que uma população dispõe inicialmente de duas alternativas de transporte: ônibus e carro, sendo que $\frac{2}{3}$ da população opta por utilizar o carro. Em um dado momento, um novo tipo de ônibus é introduzido, com todas as características essenciais idênticas às do primeiro ônibus (um exemplo de variação de pouca importância seria a cor do veículo). Intuitivamente, $\frac{2}{3}$ da população continuaria utilizando carro, enquanto o restante se dividiria entre as alternativas de ônibus. No entanto, o que ocorre sob a validade da IIA é que apenas metade da população vai permanecer usando carro quando o segundo ônibus surgir no conjunto de escolhas.

O autor sugere, então, que os modelos logit multinomial e condicional só sejam empregados caso possamos assumir de forma plausível que as categorias são distintas e ponderadas de maneira independente aos olhos daquele que realiza a escolha.

Parece interessante migrar para modelos de escolha discreta que geram padrões de substituição mais flexíveis e relaxem a hipótese de IIA.

2.1.3 Logit com Coeficientes Aleatórios

McFadden e Train (2000) demonstraram que qualquer modelo de escolha discreta derivado de um problema de maximização de utilidade aleatória pode ser aproximado por um modelo logit misto com coeficientes aleatórios (MMNL)².

É intuitivo que os indivíduos que demonstrem preferência por um dado bem também o demonstrarão por um substituto próximo em termos de características observáveis e que pessoas diferentes valorizem atributos diferentes dos produtos.

Aqui, as utilidades marginais associadas às características das alternativas variam entre os indivíduos, gerando uma correlação positiva entre os componentes das escolhas que são similares quanto a seus atributos observáveis.

Podemos modelar isso da seguinte forma:

$$U_{i,j} = X'_{i,j}\beta_i + \epsilon_{i,j}, \quad (13)$$

que também pode ser escrito como:

$$U_{i,j} = X'_{i,j}\bar{\beta} + v_{i,j}. \quad (14)$$

Note que o componente estocástico da função utilidade, $v_{i,j}$, não é mais independente entre as escolhas, apenas a parcela $\epsilon_{i,j}$ é i.i.d com distribuição de valor extremo tipo I:

$$v_{i,j} = X'_{i,j}(\beta_i - \bar{\beta}) + \epsilon_{i,j}. \quad (15)$$

²No artigo, o modelo com coeficientes aleatórios é denotado *mixed multinomial logit*. Não estou seguindo a nomenclatura, pois aqui considere o logit misto como um logit que incorpora a lógica multinomial e condicional.

Essa versão do modelo especifica:

$$\beta_{i,k} = \tilde{\beta}_k + \sigma_k \eta_{i,k}, \quad (16)$$

onde $\tilde{\beta}_k$ é a média populacional, $\eta_{k,i}$ é a heterogeneidade específica do indivíduo i com relação à sua preferência pela característica k , com média 0 e desvio padrão 1, e σ_j é o desvio padrão da distribuição dos $\beta_{i,k}$ s ao redor de $\tilde{\beta}_k$. Os vetores β_i s são distribuídos aleatoriamente entre os indivíduos com médias fixas.

Heterogeneidade observada sob a forma de variações determinísticas nos gostos também pode ser incorporada neste modelo, através da inclusão de variáveis que representam características próprias do indivíduo. Greene (2011) apresenta esse tipo de refinamento do modelo ao permitir que as médias das distribuições de parâmetros sejam heterogêneas de acordo com características observadas do indivíduo, Z_i , através do termo γ_k . Este seria um conjunto de características invariantes da escolha que produzem heterogeneidade individual nas médias dos coeficientes distribuídos aleatoriamente, de modo que:

$$\beta_{i,k} = \tilde{\beta}_k + Z_i' \gamma_k + \sigma_k \eta_{i,k}. \quad (17)$$

As variações mais comuns do modelo e aqui consideradas especificam que a distribuição de β_i , dada por $f(\beta_i)$, é contínua e na maioria dos casos normal ou lognormal. Essa distribuição é uma função dos parâmetros que representam a média e a variância do β_i , condicional à Z_i , para o indivíduo i . A probabilidade de escolha, neste caso, é uma média ponderada das formulações logísticas avaliadas em diferentes valores de β_i , com os pesos dados pela densidade $f(\beta_i)$:

$$Pr(Y_i = j | X_{i,j}, Z_i) = \int \left(\frac{\exp(X_{i,j}' \beta_i)}{\sum_{l=1}^K \exp(X_{i,l}' \beta_i)} \right) f(\beta_i) d\beta_i. \quad (18)$$

O MMNL pode ser visto como uma forma de evitar a propriedade IIA, relaxando a hipótese de independência entre o componente não observado da função utilidade. Note que os denominadores da equação 18 estão dentro de uma integral, de forma que não vão se cancelar no momento do cálculo da razão de probabilidades $\frac{Pr_{i,j}}{Pr_{i,k}}$. Assim, é possível gerar previsões mais ricos para novas escolhas, dependendo semelhança entre as características observáveis das alternativas.

A estimação dos parâmetros dessa família de modelos pode ser realizada pelo método da máxima verossimilhança simulada (VS), que consiste em encontrar um estimador $\theta = (\tilde{\beta}, \gamma, \sigma)$ que maximiza a log VS, dada a forma funcional $f(\cdot)$ escolhida pelo pesquisador para a densidade $f(\beta|\theta)$. As probabilidades de escolha são desenhadas através de simulação para dado valor de θ :

1. Define-se β^r a partir de $f(\beta|\theta^r)$, sendo $r=1, \dots, R$ referente ao número da simulação da vez;
2. Calcula-se a formulação logística de acordo com 10.

Repete-se os passos 1 e 2 R vezes para obter a média dos resultados $\check{P}_{i,j}$. As probabilidades simuladas geram a log VS a ser maximizada.

McFadden e Train (2000) sugerem ainda o método dos momentos simulados como alternativa para a estimação, mas como o pacote estatístico utilizado³ implementa MVS, decidi destacá-lo.

2.2 Custo Esperado e Prêmio

Esta subseção da revisão de literatura tem como objetivo apresentar o modelo *random forest*, introduzido por Breiman (2001), o qual será utilizado em alguns modelos de custos testados e no modelo que gera o valor de prêmio esperado.

2.2.1 Random Forests

As floresta aleatória é um método de aprendizagem de máquinas popular capaz de realizar tarefas de regressão e de classificação. A idéia geral de qualquer método baseado em árvores é separar as variáveis em grupos avaliando a homogeneidade dentro de cada grupo. A técnica de *Random Forests* nos permite capturar relações não lineares entre as variáveis explicativas e a variável dependente. Suas principais etapas consistem em:

³O pacote do R **gmnl** lançado em 2017 se mostrou eficiente na estimação de modelos logit multinomiais com heterogeneidade não observada via MVS.

- Gerar, através de *bootstrapping*, um conjunto de n amostras retiradas da amostra original através de seleção aleatória com reposição. Cerca de $\frac{1}{3}$ das observações são deixadas de fora e usadas para calcular os erros *out-of-bag* (OOB).
- Desenvolver uma árvore por amostra utilizando, em cada nó, um subconjunto de apenas m^4 variáveis selecionadas aleatoriamente dentre as k variáveis explicativas existentes, para executar a divisão. Os subconjuntos de cada nó são diferentes em uma mesma árvore. O critério utilizado para realizar as partições está ligado à utilidade dos m atributos em questão para fins de classificação. Será selecionado aquele que, dentre os m possíveis atributos da vez, possibilita o maior ganho de informação, medido através tradicionalmente pelo índice Gini⁵.
- Cada árvore cresce até que o nó terminal atinja um número mínimo de observações pré-definido. Os erros OOB representam a proporção de vezes em que o voto majoritário para os casos deixados de fora da amostra resultou em classificação incorreta. No caso de florestas de regressão, calcula-se o erro quadrático médio das previsões fora da amostra.
- Quando contruímos a j -ésima árvore, sua amostra OOB é testada nela, e a precisão da sua previsão é observada. Então, permuta-se de forma aleatória os valores da k -ésima variável explicativa. O aumento do erro de classificação provocado nos exemplos que sofreram permuta em relação aos originais fornece a medida de importância do atributo.

A figura 2 ilustra como o algoritmo cria múltiplas árvores de decisão e as combina para obter uma previsão mais precisa e estável. Isso vem da idéia que via votação majoritária de um painel de juízes independentes, obtemos uma previsão final melhor do que o melhor juiz. No exemplo, as árvores são classificatórias e a previsão final é uma variável categórica.

⁴Parâmetro definido na função `randomForest` do R através do argumento "`mtry=`". O ajuste padrão considera $m=\sqrt{k}$ para problemas de classificação e $m = \frac{k}{3}$ para regressão.

⁵A divisão em cada nó é feita visando produzir nós filhos mais puros do que o nó pai original, ou seja, com maiores concentrações de casos em dadas classes.

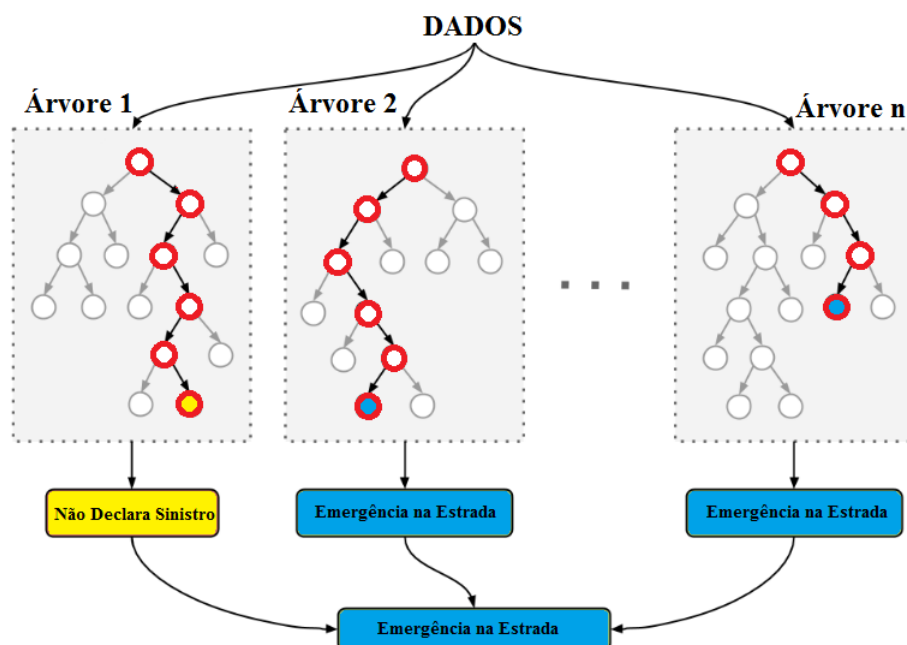


Figura 2: Floresta Aleatória (Classificação)

No caso de uma floresta de regressão, as previsões de cada árvore são uma média da variável de interesse, e a previsão da floresta é uma média das previsões dadas pelas árvores que a compõe. Exemplificamos uma árvore simples na figura 3, levando em consideração apenas as variáveis sexo e idade para explicar os custos associados à apólice. Comparando os nós terminais 4 e 15, fica claro que, conforme esperado, homens jovens proporcionam em média maiores gastos para a seguradora do que mulheres mais velhas.

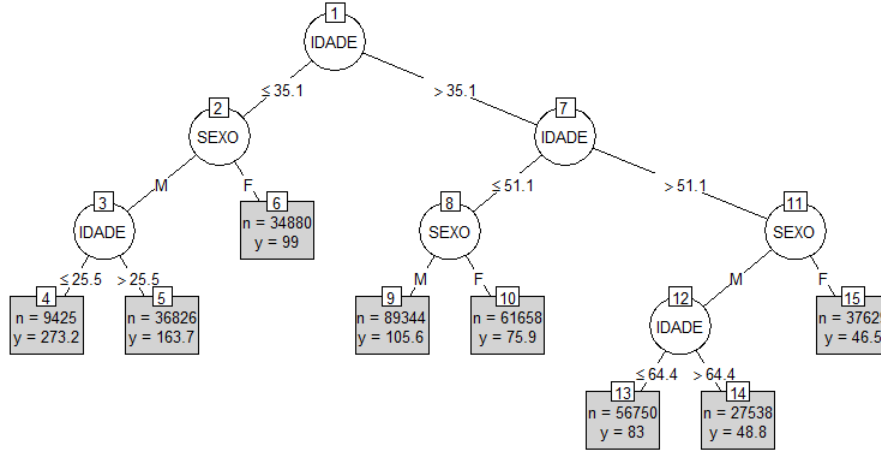


Figura 3: Exemplo de Árvore Aleatória

Uma árvore de regressão pode ser vista como um modelo de regressão constante por partes (*step function*), da forma:

$$y(x) = \sum_{i=1}^N k_i I_{\{x \in D_i\}} \quad (19)$$

Onde:

k_i são constantes

$I_{\{ \cdot \}}$ é função indicadora retornando 1 se o argumento é verdadeiro e 0 caso contrário

D_i são partições disjuntas do conjunto de dados

No artigo original de Breiman (2001), foi mostrado que a taxa de erro da floresta depende essencialmente de dois pontos:

1. A correlação entre duas árvores na floresta. Aumentar a correlação aumenta a taxa de erro da floresta.
2. A força de cada árvore individual na floresta. Uma árvore com uma baixa taxa de erro é um classificador forte. Aumentar a força das árvores individuais diminui a taxa de erro da floresta.

Reduzir "m" diminui tanto a correlação quanto a força, enquanto aumentá-lo gera acréscimos em ambos. Este é o parâmetro ajustável para o qual as florestas aleatórias são mais sensíveis.

Apesar da fama que a floresta aleatória tem de ser uma caixa preta que pega as variáveis e retorna previsões, sem se preocupar com os cálculos que estão acontecendo, existem alguns parâmetros que podemos alterar para melhorar o desempenho do modelo ou seu tempo de computação. Testei neste trabalho algumas alterações no número de árvores, do tamanho mínimo de observações dos nós terminais, dos tamanhos das amostras referentes a cada classe⁶, e do número de variáveis sorteadas a cada ramificação. Os resultados apresentados levam em conta o modelo com os ajustes que proporcionaram as melhores previsões.

⁶Mais precisamente, foi testada a técnica de *undersampling* da categoria dominante com o objetivo de melhorar a previsibilidade das classes pouco representadas na amostra original.

3 Dados

As bases de dados utilizadas foram cedidas pela Superintendência Brasileira de Seguros Privados (SUSEP)⁷. O órgão realiza coleta, semestralmente, dados estatísticos referentes às carteiras de automóveis de todas as companhias seguradoras que operam legalmente no Brasil. O período estudado foi o primeiro semestre de 2003, sendo as bases em questão R_Auto2003B (referente às apólices vigentes no período de 1º de janeiro a 30 de junho do mesmo ano) e S_Auto2003B (referente aos sinistros referentes aos sinistros avisados entre 1º de janeiro a 31 de agosto, ocorridos até 30 de junho). Esses dados reúnem informações sobre todos os contratos não-mandatários de seguro automotivo, não englobando veículos segurados somente por contratos mandatários, qual seja, o Seguro de Danos Pessoais Causados por Veículos Automotores de Via Terrestre (DPVAT). Há uma descrição detalhada de todas as variáveis contidas nas bases aqui descritas nas tabelas 11 e 12 do Anexo.

A estrutura é de um banco de dados relacional, na medida em que temos uma chave (número da apólice, código do endosso e item) para fundir e trabalhar conjuntamente com os dados presentes nas duas tabelas. Criamos assim, uma tabela conjunta denominada M_Auto2003B, necessária para a etapa de cálculo dos custos esperados das seguradoras.

Alguns filtros foram implementados, seja para remover informações incoerentes (possíveis erros de reportagem) ou para ignorar categorias que não são de interesse no momento para este estudo. Em particular, não foram analisadas apólices: (1) que sofreram endosso de qualquer tipo; (2) contratadas por pessoa jurídica; (3) coletivas; (4) com cobertura não-compreensiva; (5) com cobertura para pick-ups, ônibus, caminhões e outros veículos automotivos que não sejam classificados como de “passeio nacional” ou “passeio importado”; (6) contratadas por pessoas com idade superior a 95 anos ou inferior a 18; (7) com vigência não-anual; (8) com valores de prêmio total inferior a R\$300 ou superior a R\$4000.

Utilizou-se ainda uma base auxiliar "Modelos_Auto" que associa os códigos de modelos da SUSEP, presentes nas demais bases aqui citadas, a suas características

⁷De acordo com o “Manual de Orientações para Envio de Dados” da Susep, contido em diversas circulares emitidas.

principais. Em geral, a coluna de descrição na tabela em questão inclui a marca do veículo, modelo, número de válvulas e cilindradas. Foi possível, então, unir as bases através da chave `COD_MODELO` e, no R, através do pacote **stringr**, extrair as informações de interesse dentre as características, gerando novas variáveis potencialmente importantes na nossa análise. Por exemplo, criou-se uma dummy de motor 1.0 e uma variável categórica de marca. Esse exercício é importante uma vez que a classificação original da SUSEP envolve mais de 2 mil veículos, tornando inviável sua utilização de forma direta.

Como havia originalmente um grande número de registros, após as observações citadas serem descartadas da análise, ainda resta um volume tal que o tempo de processamento computacional é demasiadamente longo⁸ para algumas funções. Assim, utilizo somente as informações referentes à cidade do Rio de Janeiro⁹ (totalizando mais de 340 mil observações).

Em casos de haver mais de um sinistro em uma mesma apólice, o arquivo `S_AUTO`, segundo orientações da SUSEP, contém registros separados para cada sinistro avisado no período em estudo. Dessa forma, os registros representando a mesma apólice e acidente (identificados pelas variáveis `COD_SEG`, `APOLICE`, `D_OCORR`) foram agregados em uma única observação e suas indenizações, somadas. Havia ainda sinistros referentes ao mesmo acidente com indenizações fracionadas em mais de um registro (por exemplo, um acidente de colisão no qual foi necessário o serviço de reboque). Nesses casos, considerou-se como causa do acidente aquela que representasse maior gravidade.

Por fim, foi feita a escolha de focar a análise em segurados cujos percentuais de bônus associados fossem positivos, o que representa cerca de 70% da subamostra do Rio de Janeiro. A classe de bônus é determinada pelo sistema de bônus/malus, que estabelece um desconto progressivo para os segurados que renovam o seguro sem ter sofrido nenhum sinistro na vigência anterior. O vínculo é com segurado e não com seu carro, seguradora ou corretora. Na primeira contratação de seguro, a classe de bônus é 0 e aumenta progressivamente até um máximo do 10, podendo representar até 50% de

⁸Além de um grande número de apólices, há diversas variáveis qualitativas com várias categorias, o que aumenta o tempo computacional em questão.

⁹O filtro para cidade do Rio de Janeiro foi feito de forma a capturar apólices cujos 3 primeiros dígitos do CEP estivessem entre 200 e 237, além de estarem na região 18 (que corresponde à região metropolitana do Rio de Janeiro)

desconto. Se há declaração de um sinistro de colisão, roubo/furto ou incêndio, perde-se uma classe. Conforme constatado por Braido e Ledo (2018), as apólices com desconto de bônus não nulo são particularmente interessantes, pois o cliente obrigatoriamente já comprou uma apólice antes e se torna mais propenso à já estar vinculado a um corretor no seu processo de decisão. Essa hipótese vai ser útil para a construção do modelo de prêmios que veremos a seguir.

4 Modelo para Prêmio

Para estimar a demanda através dos modelos logit condicional, misto ou com coeficientes aleatórios, é preciso que os dados estejam em um formato conhecido como *long*. Neste formato, existe uma linha para cada produto disponível para a escolha por indivíduo. Ou seja, se temos 5 tipos de seguro e 240 mil indivíduos, precisamos de uma base de dados com 1 milhão e 700 mil linhas. Cada uma dessas linhas contém as informações referentes ao indivíduo (constante entre as escolhas) e as informações referentes ao produto (que podem ser distintas para o mesmo indivíduo para diferentes escolhas).

A base de dados original, no entanto, se encontra no formato *wide*, onde cada linha representa uma escolha. Dado que cada indivíduo escolhe apenas um bem, ele é representado apenas em uma linha, associado ao produto escolhido. A rigor, não sabemos ao certo qual o conjunto de escolhas considerado por cada indivíduo. No entanto, estamos lidando apenas com as 4 seguradoras com maior poder de mercado (além de uma categoria única que engloba todas as outras)¹⁰ e clientes que estão renovando sua apólice de seguro. É razoável supor que eles estejam cientes das suas principais opções de escolha e formem suas crenças, baseadas no seu próprio perfil de risco e em cotações anteriores sobre o valor de prêmio que lhe seria cobrado por cada uma dessas companhias de seguro.

Usualmente, o processo de transformação dos dados do formato *wide* para *long* é extremamente simples, conforme descrito por Sarrias e Daziano (2017), pois na maior parte dos casos as características (incluindo preço) dos bens não variam a nível do

¹⁰O modelo logit seria capaz de lidar com um número maior de opções de escolha, ainda que representem *market shares* baixos. A decisão de agregar as seguradoras pouco representativas teve o objetivo de reduzir o tempo computacional de algumas funções que estavam demasiadamente demoradas.

indivíduo, apenas do mercado. No caso de qualquer tipo de seguro, no entanto, o preço é totalmente correlacionado com as características do indivíduo e/ou bem segurado, por isso o processo se torna mais delicado e dependente de alguma hipótese sobre sua formação.

4.1 Custo Esperado

Não é possível saber qual será o custo decorrente de uma apólice para a seguradora antes do fim de período de vigência. Este custo vai depender da ocorrência de sinistros, da gravidade dos mesmos e do custo de cada um deles. Quanto maior o risco associado ao segurado, maior a probabilidade do mesmo declarar sinistro. Além disso, em todos os tipos de seguro, pode haver um limite máximo no total reembolsado e também um valor de franquia, que atua como limite mínimo de perdas que serão reembolsadas. Quanto maiores as importâncias seguradas contratualmente, potencialmente mais custoso poderá ser o acidente.

De maneira geral, as seguradoras atendem uma grande quantidade de indivíduos os quais reportam uma quantidade relativamente pequena de sinistros. É preciso discriminar adequadamente entre os segurados que tem maior propensão a proporcionar perdas. Assim, as companhias de seguro buscam sempre melhores maneiras de prevenir sinistros para poder quantificar suas perdas potenciais e atribuir valores adequados para os prêmios a serem cobrados. Utiliza-se toda a informação disponível que possa ser relevante na estimação dos riscos associados ao segurado, como características do indivíduo, do veículo e da região em questão. As seguradoras possuem amplas bases de dados e procuram sempre aperfeiçoar suas metodologias de previsão de forma a extrair o máximo possível dos dados disponíveis.

Não conhecemos a metodologia exata aplicada por cada companhia, mas sabemos que a tendência do mercado envolve métodos de *machine learning*. A Porto Seguro (empresa com a maior participação no mercado de seguros auto brasileiro) lançou no fim de 2017 uma competição no Kaggle¹¹ visando o desenvolvimento de um algoritmo para previsão de declaração de sinistros que seja mais preciso que os seus próprios. A

¹¹ Maior plataforma de competição de análise de dados e modelagem preditiva no mundo, onde as empresas colocam os problemas que gostariam de solucionar (juntamente com um prêmio) e analistas de todo o mundo competem para produzir os melhores modelos de *machine learning*.

seguradora declara que vem usando métodos de machine learning há 20 anos.

Tentei aplicar alguns métodos simples para prever o custo esperado da seguradora com o fim de gerar uma nova variável potencialmente com um bom poder explicativo do prêmio, e que permitisse o aproveitamento das informações contidas na base *S_Auto*.

4.1.1 Especificação dos Tipos de Sinistro

Em todos os modelos testados, utilizei como variáveis explicativas o ano do modelo do automóvel, o código tarifário, o tipo de franquia do contrato, as importâncias seguradas e o sexo e a idade do condutor principal. Foi incluída uma variável categórica com os 3 primeiros dígitos do CEP¹² de residência para os modelos lineares, enquanto nos modelos de floresta aleatória foi utilizada a variável contínua CEP.

Incluí ainda variáveis geradas para representar a expectativa de declaração de sinistro para prever tipos específicos de acidentes.

Para os modelos descritos abaixo, foram geradas 4 novas variáveis binárias e uma nova variável contínua:

1. SINISTRO: Indicando a declaração ou não de sinistro durante o período de vigência da apólice.
2. ACIDENTEGRAVE: Indicando a declaração ou não de um acidente muito custoso para a seguradora (roubo, furto, incêndio ou colisão de perda total) durante o período de vigência da apólice.
3. COLISAOPARCIAL: Indicando a declaração ou não de sinistro de colisão parcial durante o período de vigência da apólice.
4. EMERGENCIA: Indicando a declaração ou não de uma emergência na estrada durante o período de vigência da apólice.

¹²Os 3 primeiros dígitos do CEP indicam, respectivamente, região, sub-região e setor. No caso da cidade do Rio de Janeiro, essa variável varia entre 200 e 237. Dessa forma, podemos captar diferenciação de risco entre 37 setores da cidade. Uma classificação mais detalhada geraria um número de categorias propenso a problemas computacionais.

5. CUSTO: Valor total do custo decorrente da apólice, calculado como a soma de todas as indenizações pagas subtraindo o valor do salvado correspondente. A não ser nos casos de emergência na estrada (em que não se cobra a taxa de franquia), subtraiu-se ainda desse número o valor da franquia de cada apólice.

Tabela 1: Número de Ocorrências por Tipo de Sinistro

Causa do Sinistro	Número de Ocorrências
Sem Sinistro	224524
Emergência na Estrada	9973
Colisão	4339
Roubo/Furto	1094
Perda Total	505
Incêndio	52

A divisão representada pelas novas variáveis binárias (que indicam ocorrência ou não de acidentes graves, colisão parcial ou emergência 24 horas) foi feita de acordo com o patamar associado ao montante médio de custos decorrentes de cada tipo de acidente, conforme apresentados na Tabela 2. Caso não agregássemos os sinistros que foram denominados como acidentes graves, teríamos problemas para a previsão dos acidente com pequeno número pouco representativo de casos, como é o caso de incêndio (vide Tabela 1).

Tabela 2: Gastos Médios por Tipo de Sinistro

Causa do Sinistro	Custo Médio	Indenização Média
Emergência na Estrada	285.05	303.12
Colisão	2039.64	2938.19
Roubo/Furto	10431.67	11596.68
Perda Total	12753.14	16123.27
Incêndio	6477.89	7932.96

Utilizamos uma parcela de 80% da base tratada conforme já descrito, selecionada de forma aleatória, para “treinar” os modelos em questão. O “teste” foi então realizado levando em conta os 20% remanescentes dos dados.

4.1.2 Modelos Testados

Regressão Linear

Primeiramente, calculei via regressão linear os custos esperados C_i para cada apólice i como função das variáveis explicativas detalhadas no início da seção anterior, aqui representadas por β_0, \dots, β_J :

$$C_i = \beta_0 + X_{i,1}\beta_1 + \dots + X_{i,J}\beta_J + \epsilon_i. \quad (20)$$

Visando capturar as não linearidades contidas no termo de erro da regressão, rodamos uma *random forest* resíduo observado ($\hat{\epsilon}_i = C_i - \hat{\beta}_0 + X_{i,1}\hat{\beta}_1 + \dots + X_{i,J}\hat{\beta}_J$) com as mesmas variáveis explicativas da regressão original e o incluímos em uma nova regressão.

Logit Binário

Estimei regressões logísticas, como candidatos naturais para um modelo de escolha binária. Assume-se que a probabilidade cumulativa de declaração de cada tipo de sinistro seja independente da ocorrência de outro tipo de acidente e que o mesmo indivíduo pode declarar mais de um sinistro no período.

No R, basta ajustar um modelo linear generalizado¹³, informando, além da fórmula, a família de distribuições da resposta. Computei a probabilidade de ocorrência de um sinistro qualquer e também a probabilidade de cada tipo de sinistro¹⁴ como função de todas as variáveis anteriormente descritas.

Logit de Razão Contínua

Em uma segunda etapa, considerei uma possível sequencialidade para as probabilidades ocorrência dos sinistros mais leves para os mais graves. O modelo utilizado é conhecido como *continuation ratio* e pressupõe outcomes ordenados, em que os indivíduos

¹³A função em questão é chamada através do comando `glm()`

¹⁴Aqui considerando os tipos de sinistro de acordo com a classificação de gravidade anteriormente gerada (acidente grave, colisão parcial ou assistência 24 horas)

passariam por estágios levando a um determinado nível. A sequencialidade no nosso problema é dada de acordo com a figura abaixo:

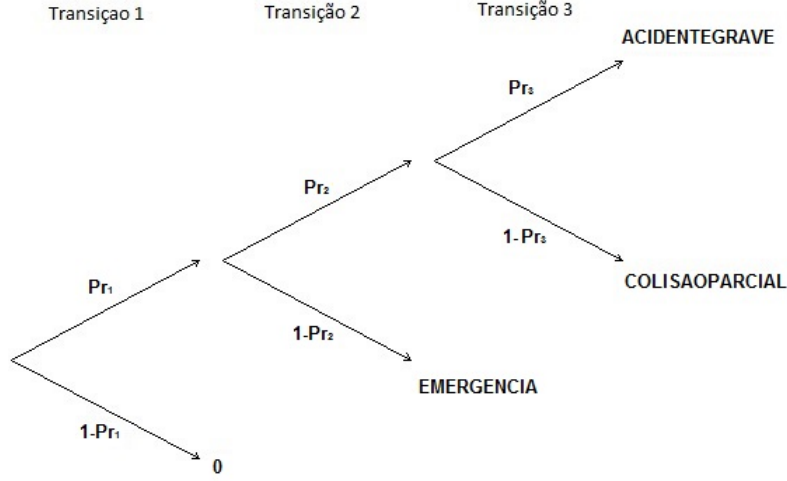


Figura 4: Sequencialidade de Sinistros

Dessa forma, a probabilidade de declaração de uma colisão parcial vai ser dada pelo produto das probabilidades de transição explicadas na figura acima, $Pr1 * Pr2 * (1 - Pr3)$.

O pacote do R utilizado para computar a função foi o **glmnetcr**¹⁵. Ele provê uma função para que se realize um ajuste personalizado de modelos de *continuation ratio*, dado interesse de previsão de uma resposta ordinal (aqui sendo a ocorrência ou não de um determinado tipo de acidente, ordenado de acordo com seu nível de gravidade). A formulação utilizada foi a *backward*:

$$\text{logit}(P_i[Y_i = m | Y_i \leq m, X_i = x]) = \alpha_m + \beta_m x_i \quad (21)$$

Essa formulação é comum quando o interesse reside em estimar as probabilidades do caso mais severo em relação ao mais moderado. De forma semelhante, a seguradora visa diferenciar os clientes baseados na sua propensão de gerar perdas, especialmente a probabilidade de ocorrer acidentes graves em relação a não ocorrência de sinistros.

¹⁵Detalhes em: <https://cran.r-project.org/web/packages/glmnet/glmnet.pdf>

Random Forest

Foi utilizado o modelo disponível no pacote *randomForest* do R, que é baseado no código original de Breiman e Cutler¹⁶. Mantive as configurações padronizadas do pacote para a execução dos meus códigos, com exceção dos parâmetros anteriormente citados.

Primeiramente, testei o modelo de regressão para os custos e, de forma similar aos casos logit descritos anteriormente, executei um algoritmo que visa classificar, dadas as variáveis explicativas, a possível ocorrência de um acidente grave, colisão parcial, assistência na estrada ou não ocorrência de sinistro.

O *random forest* de classificação foi o único modelo aqui testado que não incluiu todas as 224 mil observações referentes à população estudada de apólices. Foi utilizado o método de *undersampling*, que consiste em eliminar observações da classe mais frequente (no nosso caso, apólices em que não há declaração de sinistros). O objetivo é tratar o desbalanceamento dos dados, fazendo com que a classe menos frequente se torne mais observada. Decidi por adotar esta técnica nesse modelo pois, dada a raridade da ocorrência do evento de declaração de sinistros mais graves, inicialmente a floresta estava prevendo sempre que não ocorreria um acidente desse tipo.

Custos de Cada Tipo de Acidente

Dadas as probabilidades de declaração de cada tipo de sinistro, precisamos estimar um valor de dispêndio previsto para obter o custo esperado. Para cada grau de gravidade do acidente, rodei uma regressão linear do custo condicional à ocorrência daquele tipo de sinistro em função dos valores de importâncias seguradas da apólice, ano do modelo, idade e sexo do motorista principal, tipo de franquia, características do veículo segurado e CEP truncado nos primeiros 3 dígitos. Na regressão cuja variável dependente foi o custo de assistência 24 horas, as variáveis dependentes foram importância segurada do casco, sexo e idade, pois, neste caso, não existe relação entre o serviço e as coberturas relacionadas à terceiros.

Considerarei ainda a mediana dos custos de emergência na estrada e a fração média de importância segurada do casco paga pela seguradora em caso de colisão parcial e

¹⁶Código e explicações disponíveis em: <https://www.stat.berkeley.edu/~breiman/RandomForests>

de acidentes graves. Testei então, as combinações dos custos condicionais gerados das diversas maneiras descritas com as probabilidades de declaração de sinistros e reporteio os melhores resultados associados a cada modelo de previsão.

4.1.3 Comparação do Modelos

A principal métrica usada para comparar os resultados dos modelos foi o erro quadrático médio (EQM) do custo estimado na base de teste em relação ao custo real observado na mesma. Sua formulação é dada por:

$$EQM = \sum_{i=1}^n \frac{(CUSTOREAL - CUSTOESTIMADO)^2}{n}. \quad (22)$$

Os modelos testados para prever os custos associados às apólices de seguro obtiveram performances muito similares. Os resultados encontram-se na tabela 3 abaixo.

Tabela 3: Comparação dos Erros Quadráticos Médios do Custo Estimado

Modelo	EQM
(1) Random Forest Regressão	1.765.250
(2) Regressão Linear	1.767.318
(3) Random Forest/Regressão*	1.767.065
(4) Logit	1.763.712
(5) Continuation Ratio	1.769.675
(6) Random Forest Classificação	1.770.380

* : Indica o modelo de regressão com a variável de erro esperada gerada via Random Forest

Além do EQM, analizei as matrizes de confusão dos modelos logit binários e do random forest. Dado desbalanceamento do conjunto de dados - no sentido de que apenas cerca de 7% das observações possuem algum tipo de sinistro declarado - a acurácia¹⁷ de todos os modelos foi elevada, enquanto a sensibilidade¹⁸ foi baixa.

¹⁷Acurácia = Total de Acertos / Total de Dados no conjunto

¹⁸Sensibilidade = Total de Acertos Positivos / Total de Positivos no conjunto

Apesar desta aparente baixa capacidade de previsão, os modelos foram capazes de gerar variáveis de custo esperado com alto poder explicativo para o prêmio, e por isso se tornaram relevantes no escopo mais geral do modelo de preços. Conforme ilustrado na tabela 4, apenas o custo esperado (gerado pela regressão random forest¹⁹) é capaz de explicar aproximadamente 17% da variação do preço das apólices.

Tabela 4: Resultados da Regressão Prêmio X Custo Esperado

<i>Variável Dependente:</i>	
	PRÊMIO
CUSTO ESPERADO	2.649*** (0.012)
CONSTANTE	617.182*** (1.736)
Observations	240,487
R ²	0.168
Adjusted R ²	0.168
F Statistic	48,602.260*** (df = 1; 240485)
<i>Nota:</i>	*p<0.1; **p<0.05; ***p<0.01

¹⁹Dados os resultados semelhantes dos diversos modelos testados, escolhi arbitrariamente um dentre aqueles não lineares, para evitar posteriores potenciais problemas de colinearidade quando a variável gerada fosse inserida em um modelo linear.

4.2 Prêmio Esperado

Para estimar o valor de prêmio cobrado por cada um dos 5 grupos de seguradoras considerados, testamos o poder de explicação de diversas variáveis, via modelo de regressão linear e random forest. Ambos apresentaram bom poder explicativo e variáveis de custo esperado significativas.

A tabela 5 descreve os coeficientes associados a algumas das variáveis usadas no modelo de regressão. *Dummies* de marca, modelo, cilindradas, alguns CEPs, idade, percentual de bônus, importâncias seguradas ex-casco e constante são significativas neste modelos mas foram omitidas nesta tabela. A ideia aqui é demonstrar a significância do custo e a diferença de preços cobrados por seguradoras distintas mantidas as demais características constantes. Também é interessante notar o comportamento do CEP3: 224²⁰ representa um setor da zona sul da cidade, considerada uma região mais segura, com menor taxa relativa de roubos de veículos que o setor 213²¹, na zona norte da cidade. Os coeficientes relacionados com cada um destes CEPs vai de acordo com o que poderíamos intuir: mantidas constantes as demais variáveis, os condutores cuja residência se localiza no setor 213 tem um acréscimo de cerca de R\$41 no valor do seu prêmio, enquanto para os moradores do setor 224 há uma redução de R\$42.

Dadas as intuições obtidas a partir do modelo de regressão, foi rodado um random forest com as mesmas variáveis explicativas. A performance do modelo foi superior, com a capacidade de explicar 70% da variância²², enquanto a regressão apresentou um R^2 de 61%. Na figura 5, vemos uma representação de importância das variáveis no random forest, com destaque para o custo esperado.

²⁰Essa região inclui bairros nobres como Leblon, Ipanema e Jardim Botânico.

²¹Este setor representa o bairro de Madureira e seus entornos, onde a criminalidade é relativamente elevada.

²²A medida é dada pelo “pseudo R^2 ”: $1 - \text{EQM} / \text{Var}(y)$, onde EQM é o erro quadrático médio das previsões OOB e $\text{Var}(y)$ é a variância da variável alvo.

Tabela 5: Resultados da Regressão de Prêmio

<i>Variável Dependente:</i>	
	PRÊMIO
CEP3: 213	41.215*** (3.084)
CEP3: 224	-42.052*** (2.644)
SEXO MASCULINO	60.301*** (1.171)
IS_CASCO	0.023*** (0.0001)
SEGURADORA A	99.429*** (2.233)
SEGURADORA C	68.773*** (1.647)
SEGURADORA D	62.704*** (2.241)
SEGURADORA E	95.998*** (2.218)
CUSTO ESPERADO	0.432*** (0.011)
Observações	240,487
R ²	0.610
Estatística F	27 3,682.611*** (df = 102; 240384)

Nota1: *p<0.1; **p<0.05; ***p<0.01

Nota2: Apenas uma parte das variáveis utilizadas na regressão foram ilustradas

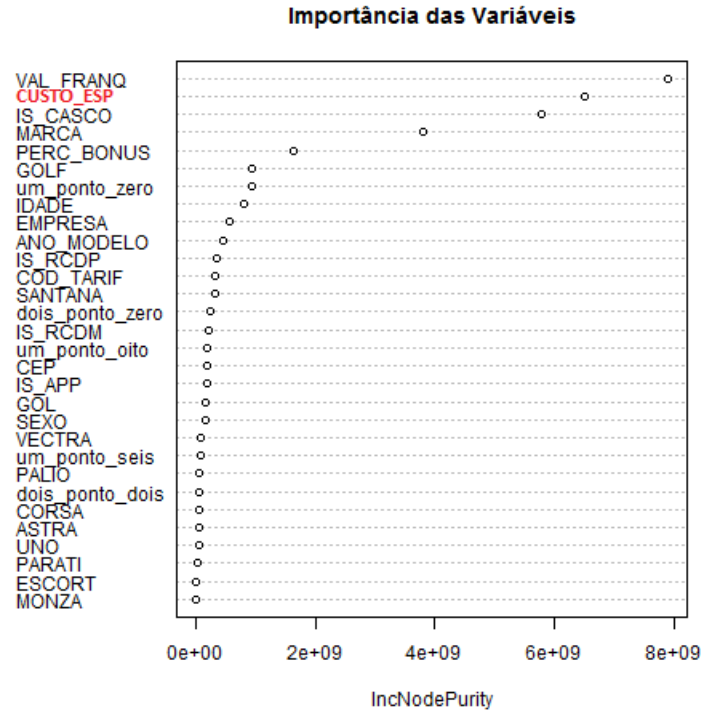


Figura 5: Importância das Variáveis Usadas no Random Forest Para Explicar Prêmio

Dado o excelente desempenho do random forest na determinação de prêmios cobrados dadas as variáveis disponíveis e com o objetivo de evitar uma posterior colinearidade no modelo de demanda, ele foi escolhido para estimar os preços esperados para cada segurado para suas 5 opções de escolha disponíveis, aqui denominadas seguradora A, B, C, D e E.

5 Estimação da Demanda

Supomos que o segurado decide primeiramente o valor de suas coberturas contra terceiros, passageiros e se vai pagar franquia normal, reduzida ou majorada. Dada essa escolha, as características do motorista e o preço esperado para sua apólice em cada seguradora, ele decide a empresa que deseja contratar baseado na confiança atribuída à companhia e o preço que ele espera pagar pela apólice anual. Dessa forma, as alternativas consideradas por cada indivíduo são constantes em todas suas variáveis exceto preço e empresa seguradora.

Cabe reforçar a suposição de que a variável de preço considerado na análise é o prêmio esperado. Como só analisamos clientes cuja classe de bônus é positiva, eles necessariamente obtiveram, em períodos anteriores, informações no seu processo de busca para contratação da primeira apólice. Assim, estão cientes das suas principais opções de escolha e aptos a formação de crenças, baseadas no seu próprio perfil de risco, sobre o valor de prêmio que lhe seria cobrado por cada uma dessas companhias de seguro. A hipótese aqui é que ele toma sua decisão em função desse valor esperado.

A Tabela 6 apresenta o *market share* de cada seguradora considerada na amostra em que estimamos a demanda. Os nomes reais foram substituídos por letras para preservar a confidencialidade dos dados que foram cedidos pela SUSEP.

Tabela 6: Prêmio Médio e *Market Share* por Seguradora

Seguradora	<i>Market Share</i>
A	33,4%
B	25,4%
C	19,8%
D	11,4%
E	10,0%

5.1 Logit Misto

Foi testado o modelo logit misto, para posterior comparação das elasticidades obtidas pelo modelo com coeficientes aleatórios. A formulação considerada foi a seguinte:

$$U_{i,j} = \alpha_j + \beta_0 \text{PremioPrevisto}_{i,j} + \beta_{1,j} \text{Idade}_i + \beta_{2,j} \text{IS_Casco}_i + \epsilon_{i,j} \quad (23)$$

Os resultados encontram-se na tabela 7. Conforme esperado, um aumento do prêmio previsto, por parte de qualquer seguradora, reduz os *odds* da escolha da mesma frente à todas as empresas concorrentes. Enquanto isso, aumento de idade parece reduzir a probabilidade de escolha de qualquer seguradora em relação à B, e importância seguradora do casco apresenta sinais diversos dependendo da companhia que se está analisando comparativamente ao nível de referência B.

Tabela 7: Resultados do Modelo Logit Misto

B é a seguradora de referência				
	A	C	D	E
Constante	-1,026e+00*** (2,406e-02)	4.743e-01*** (2,620e-02)	3,525e-01*** (3,134e-02)	-1,042e-01** (3,297e-02)
Idade	-1,762e-03*** (2,1481e-04)	-1,813e-02*** (2,578e-04)	-1,841e-02*** (3,143e-04)	-1.473e-02*** (3.267e-04)
IS Casco	3,673e-05*** (6,685e-07)	2,490e-05*** (7,591e-07)	1.282e-06 (9.678e-07)	-1.085e-05*** (1.055e-06)
Prêmio Previsto	-2,898e-03*** (6,776e-05)			

*** p<0.01;** p<0.05;*** p<0.1

Observações: 240.487

Log Likelihood: -357450

R² de McFadden: 2,5%

5.2 Logit com Efeitos Aleatórios

Aqui, o coeficiente associado à variável de prêmio previsto foi tratado como aleatório. Variáveis específicas do indivíduo foram incluídas para explicar a média dos parâmetros

aleatórios. A formulação do modelo foi a seguinte:

$$U_{i,j} = \alpha_j + \beta_{i,PremioPrevisto} PremioPrevisto_{i,j} \quad (24)$$

com:

$$\beta_{i,PremioPrevisto} = \bar{\beta} + \pi_{11} Idade_i + \pi_{12} Is_Casco_i + \sigma_1 \eta_{1,i} \quad (25)$$

onde $\eta_{1,i} \sim N(0,1)$

De acordo com os resultados apresentados na tabela 8, indivíduos mais velhos e com carros mais caros seriam mais sensíveis à variações no valor esperado do prêmio. As constantes específicas de cada alternativa indicam as propensões individuais à escolher uma seguradora, em relação à B, por razões que não estão capturadas nas variáveis incluídas no modelo. A empresa A foi a única com uma constante positiva, e também é a única que apresenta um *market share* superior ao de B. Considerando o R^2 de McFadden para fins comparativos, podemos notar que o mesmo dobrou em relação ao modelo anterior.

Tabela 8: Resultados do Modelo Logit com Coeficientes Aleatórios

Prêmio Previsto Constante	-2,088e-03*** (3,425e-04)			
Prêmio Previsto Idade	-2,200e-05*** (6,271e-06)			
Prêmio Previsto IS Casco	-8,663e-08*** (8,479e-09)			
Prêmio Previsto Desvio-Padrão	1,123e-02*** (2,711e-04)			
Constante	B é a seguradora de referência			
	A	C	D	E
	2,775e-01*** (5,597e-03)	-3,967e-01*** (6,828e-03)	-7,560e-01*** (7,669e-03)	-9,144e-01*** (7,841e-03)

*** p<0.01,** p<0.05,*** p<0.1

Observações: 240.487

Log Likelihood: -362320

 R^2 de McFadden: 5,1%

5.3 Elasticidades

As tabela 9 apresenta as elasticidades-preço de substituição geradas pelo modelo logit misto. A diagonal principal representa as elasticidades-preço próprias, que indicam o impacto percentual esperado na probabilidade de escolha do bem quando há o aumento de 1% em seu preço. Conforme poderíamos esperar, em todos os casos os sinais foram negativos. Os valores absolutos não diferem muito, mas os clientes de A e B, que são as firmas com maiores fatias do mercado, parecem mais sensíveis à mudanças de preço. As elasticidades-preço cruzadas representam o efeito relativo sobre as probabilidades de escolha das marcas em cada coluna decorrente de variações de 1% nos preços das marcas em cada linha. Devido à propriedade de IIA, as elasticidades cruzadas são constantes em todas as alternativas, ou seja, este modelo calcula padrões de substituição proporcionais.

Tabela 9: Elasticidades-Preço de Substituição: Logit Misto

		$\Delta \% Q$				
		A	B	C	D	E
$\Delta \% P$	A	-2,183	0,144	0,144	0,144	0,144
	B	0,310	-2,059	0,310	0,310	0,310
	C	0,752	0,752	-1,500	0,752	0,752
	D	0,546	0,546	0,546	-1,833	0,546
	E	0,575	0,575	0,575	0,575	-1,766

Por fim, é feita a análise das elasticidades-preço do modelo logit com coeficientes aleatórios, para obter uma melhor compreensão dos padrões de substituição entre seguradoras. Devido à formulação do modelo, as elasticidades foram calculadas por simulação. Podemos observar na tabela 10 que as elasticidades geradas exibem padrões de substituição que não são proporcionais. O resultado já era esperado dado o relaxamento da hipótese de IIA. Além disso, as demandas de todas as seguradoras, exceto C, mostraram-se mais inelásticas à variações no próprio preço.

Tabela 10: Elasticidades-Preço de Substituição: Logit com Coeficientes Aleatórios

		$\Delta \% Q$				
		A	B	C	D	E
$\Delta \% P$	A	-1,090	0,436	1,176	0,095	0,328
	B	0,331	-1,114	0,871	0,013	0,215
	C	0,669	0,650	-2,513	0,346	0,615
	D	0,546	0,008	0,201	-0,300	0,042
	E	0,103	0,087	0,326	0,008	-1,032

6 Considerações Finais

A demanda por seguros na cidade do Rio de Janeiro foi estimada via Logit com coeficientes aleatórios. Testou-se ainda o modelo Logit misto e suas elasticidades-preço de substituição para fins de comparação com os padrões mais flexíveis gerados pelo modelo principal. Como esperado, todas as elasticidades próprias apresentaram sinais negativos, enquanto as elasticidades cruzadas obtiveram sinais positivos. Idade e valor da importância segurada do casco apareceram com alto nível de significância na determinação do coeficiente de sensibilidade ao valor do prêmio esperado, sendo que um aumento em quaisquer um destes fatores torna este coeficiente ainda mais negativo.

O preço considerado nos modelos de demanda foi o valor de prêmio previsto gerado pelo modelo de *random forest*, que apresentou um bom ajuste, sendo capaz de explicar 70% de sua variância. A variável de custo esperado se mostrou importante neste contexto, apesar do mau desempenho preditivo dos modelos de custo testados. A capacidade do *random forest* de capturar relações não lineares entre as variáveis permitiu o uso concomitante do prêmio esperado gerado pelo modelo com outras variáveis usadas como explicativas para sua formação, como idade e importância segurada do casco sem resultar em problemas de colinearidade.

Referências

- [1] Braido, L. e Ledo, B. (2018). *Dynamic Price Competition in Auto-Insurance Brokerage*. Rand Journal of Economics, accepted for publication, Feb. 2018.
- [2] Breiman, L. (2001). *Random Forests*. Machine Learning 45, 5-32.
- [3] Elshiewy, Ossama, Daniel Guhl, e Yasemin Boztug (2017). *Multinomial Logit Models in Marketing-From Fundamentals to State-of-the-Art*. Marketing ZFP 39.3, 32-49.
- [4] Greene, W. (2011). *Discrete choice modeling*. in Handbook of Econometrics: Vol. 2, Applied Econometrics, Part 4.2: Chapter 11, edited by T. Mills and K. Patterson. London: Palgrave, pp. 473–556.
- [5] McFadden, D (1974). *Conditional Logit Analysis of Qualitative Choice Behavior*. in P. Zarembka, ed., 'Frontiers in Econometrics', Academic Press, New York, pp. 105–42.
- [6] McFadden, D. e Train, K. (2000). *Mixed mnl models of discrete response*. Journal of Applied Econometrics 15, pp. 447–470.
- [7] Sarrias, M. e Daziano, R.A. (2017). *Multinomial logit models with continuous and discrete individual heterogeneity in R: the gmnL package*. Journal of Statistical Software 79(2): 1-46.
- [8] Superintendência de Seguros Privados. (2002). *Circular SUSEP n. 184, de 28 de Março de 2002*. Ministério da Fazenda.
- [9] Train, K (2009). *Discrete Choice Methods with Simulation*. 2nd edition, Cambridge University Press.

A Anexos

Tabela 11: R_Auto: Descrição das Variáveis

Variável	Descrição
COD_SEG	Variável numérica que indica o código FIP da seguradora junto à SUSEP. Exemplo: 08001
APOLICE	Variável numérica que caracteriza o número da apólice do contrato.
ENDOSSO	Variável numérica que indica se houve alguma alteração da apólice original. O código “0000000000” representa o registro da apólice original.
COD_END	Variável qualitativa que caracteriza o tipo de endosso. O valor 0 indica registro de apólice; 1 indica alteração genérica (a substituição do veículo ou alterações das coberturas, importâncias seguradas, categoria tarifária, do tipo de franquia contratada, ou outros fatores de risco do segurado; 2 e 3 indicam cancelamento da apólice; 4 indica o cancelamento do endosso.
ITEM	Variável numérica que indica o item de identificação do veículo em caso de apólice coletiva. Apólices individuais estão associadas ao código “000000”.
COBERTURA	Variável qualitativa que descreve o tipo de cobertura contratada. 1 representa cobertura compreensiva; 2, cobertura contra furto, roubo e incêndio; 3, cobertura somente contra incêndio; 4, cobertura contra perda total, colisão, furto e roubo; e 9 indica outros tipos de cobertura.
COD_MODELO	Variável alfanumérica que caracteriza o código do modelo do veículo conforme a tabela IX da circular 135 da SUSEP.
ANO_MODELO	Ano do modelo do veículo segurado.
COD_TARIF	Variável numérica que indica a categoria tarifária em que o veículo segurado se enquadra. O código 1 representa passeio nacional; 2, passeio importado; 3, pick-ups; 4, veículos de carga; 5, motocicletas; 6, ônibus; 7, utilitários; e 9 todos os demais.

Continua na próxima página

Tabela 11 – Continuação

Variável	Descrição
REGIAO	Variável numérica que descreve a região associada à residência do indivíduo cujo veículo foi segurado. Varia entre 1 a 41, conforme estabelecido na tabela VIII da circular 135 da SUSEP.
TIPO_FRANQ	Variável qualitativa que identifica o tipo de franquia contratada. Os códigos 1, 2, 3 e 4 indicam, respectivamente, tipos de franquia reduzida, normal, majorada e sem franquia.
IS_CASCO	Valor máximo, em R\$, da cobertura contratada de casco para próprio veículo segurado;
IS_RCDM	Valor máximo, em R\$, para cobertura de responsabilidade civil facultativa de veículos – danos materiais (propriedade de terceiros).
IS_RCDP	Valor máximo, em R\$, para cobertura de responsabilidade civil facultativa de veículos – danos pessoais (pedestres e passageiros de veículos terceiros).
IS_APP	Importância segurada, em R\$, contratada para cobertura de acidentes pessoais de passageiros do veículo segurado.
PRE_CASCO	Valor total, em R\$, do prêmio emitido para cobertura de casco.
PRE_RCDM	Valor total, em R\$, do prêmio emitido para cobertura de responsabilidade civil facultativa de veículos – danos materiais.
PRE_RCDP	Valor total, em R\$, do prêmio emitido para cobertura de responsabilidade civil facultativa de veículos – danos pessoais.
PRE_APP	Valor total, em R\$, do prêmio emitido para cobertura de acidentes pessoais de passageiros.
PRE_OUTROS	Valor total, em R\$, do prêmio emitido para as coberturas de acessórios, equipamentos, carrocerias e outras coberturas, as quais são contabilizadas no ramo 31 do FIP, tais como assistência 24 horas, carro reserva, etc.
INICIO_VIG	Data de início da vigência da apólice, no formato AAAAMMDD.
FIM_VIG	Data de término da vigência da apólice, no formato AAAAMMDD.
PERC_BONUS	Percentual de desconto fornecido de acordo com histórico de não ocorrência de sinistro de casco, incidente sobre o prêmio total.
Continua na próxima página	

Tabela 11 – Continuação

Variável	Descrição
PERC_CORR	Valor percentual da comissão de corretagem.
SEXO	Variável qualitativa que indica o sexo do principal condutor do veículo, a saber: M, para sexo masculino; F, feminino; e J para pessoa jurídica.
DATA_NASC	Data de nascimento do principal condutor do veículo, no formato AA-AAMMDD. O código “00000000” está associado à pessoa jurídica.
CEP	CEP de residência do principal condutor do veículo.

Tabela 12: S_Auto: Descrição das Variáveis

Variável	Descrição
COD_SEG, APOLICE, EN- DOSSO, ITEM, COBERTURA, COD_MODELO, ANO_MODELO, COD_TARIF, REGIAO, SEXO, DATA_NASC	Vide Tabela 1.
IND_CASCO	Valor, em R\$, da indenização paga para cobertura de casco, referente à danos físicos do veículo segurado.
IND_RCDM	Valor, em R\$, da indenização paga ao segurado para a cobertura de danos à propriedade de terceiros.
IND_RCDP	Valor, em R\$, da indenização paga ao segurado para cobertura de danos físicos causados a pedestres e passageiros de veículos de terceiros.
IND_APP	Valor, em R\$, da indenização paga ao segurado para cobertura de danos físicos causados aos passageiros do veículo segurado.
Continua na próxima página	

Tabela 12 – Continuação

Variável	Descrição
IND_OUTROS	Valor, em R\$, da indenização paga referente à despesas adicionais, como serviço de emergência na estrada, serviço de aluguel de carro, reposição de acessórios do veículo e de vidros quebrados.
VAL_SALVAD	Valor, em R\$, recuperado pela seguradora em função de sub-rogação.
D_SALVAD	Data de recuperação do salvado e/ou ressarcimento, no formato AA-AAMMDD.
D_AVI_CASC, D_AVI_RCDM, D_AVI_RCDP, D_AVI_APP	Data de acionamento do seguro, consoante, de acordo com os danos decorrentes do sinistro, no formato AAAAMMDD.
D_LIQ_CASC, D_LIQ_RCDM, D_LIQ_RCDP, D_LIQ_APP	Data de liquidação do sinistro pela seguradora de acordo com cada um dos tipos de acionamento, no formato AAAAMMDD.
D_OCORR	Data reportada para a ocorrência do acidente, no formato AAAAMMDD.
CAUSA	Variável qualitativa que indica o código da causa geradora do sinistro. O código 1 indica furto ou roubo; 2, colisão; 3, perda total; 4, incêndios; 5, serviço de emergência na estrada; e 9 indica outras causas.
CEP	CEP da localidade de ocorrência do sinistro.

B Código R

```
#-----#
#-----BASE DE DADOS-----#
#-----#

library("dplyr")
library(lubridate)

#####
##### R AUTO #####
#####

#Importando os dados

R_AUTO= read.csv2("R_Auto2003B.csv")

#####
# Selecionando as informacoes relevantes:#
#####

#Mantemos apenas as apolices sem endosso
R_AUTO2003B <- R_AUTO%>%filter(ENDOSSO==0)

#Mantemos apenas as apolices com cobertura abrangente
R_AUTO2003B <- R_AUTO2003B%>%filter(COBERTURA==1)

#Excluindo apolices coletivas:
R_AUTO2003B <- R_AUTO2003B%>% filter(SEX0=="F"|SEX0=="M")

#Deleta observacoes duplicadas da mesma apolice
R_AUTO2003B<-R_AUTO2003B%>%mutate(dup=duplicated(APOLICE))%>%
  filter(dup=="FALSE")

#Mantendo contratos assinados entre 01/01/2002 e 06/30/2003
```

```

R_AUTO2003B <- R_AUTO2003B%>%
  filter(ymd(INICIO_VIG) %within% interval(ymd("2002-01-01"),ymd("2003-06-30")))

#Mantendo contratos de 1 ano
R_AUTO2003B$CONTRATO_DUR<-
  as.numeric(ymd(R_AUTO2003B$FIM_VIG)- ymd(R_AUTO2003B$INICIO_VIG))
quantile(R_AUTO2003B$CONTRATO_DUR,na.rm=TRUE)
R_AUTO2003B<-filter(R_AUTO2003B,CONTRATO_DUR>360,CONTRATO_DUR<370)

#Excluindo idades sem sentido
R_AUTO2003B$IDADE<-
  as.numeric(ymd(R_AUTO2003B$INICIO_VIG)-ymd(R_AUTO2003B$DATA_NASC))/365
table(trunc(R_AUTO2003B$IDADE))
R_AUTO2003B <- R_AUTO2003B%>% filter(IDADE>=18&IDADE<99)

#Excluindo contratos com valores outliers de premio
R_AUTO2003B$somaPremios=
  as.numeric(R_AUTO2003B$PRE_APP)+as.numeric(R_AUTO2003B$PRE_CASCO)
  +as.numeric(R_AUTO2003B$PRE_RCDM)+as.numeric(R_AUTO2003B$PRE_RCDP)
  +as.numeric(R_AUTO2003B$PRE_OUTROS)
quantile(R_AUTO2003B$somaPremios,probs = c(1,2,10,25,50,75,90,98,99,100)/100)
R_AUTO2003B<-R_AUTO2003B%>%filter(somaPremios>300&somaPremios<4000)

#Acrescentando o nome das seguradoras:
cod_seg = read.csv2("CodxSeg.csv",stringsAsFactors=FALSE,
  fileEncoding="latin1")
R_AUTO2003B<-left_join(R_AUTO2003B,cod_seg,by="COD_SEG")

#Mantendo apenas passeio nacional e importado
table(R_AUTO2003B$COD_TARIF)
R_AUTO2003B<-filter(R_AUTO2003B,COD_TARIF=="1"|COD_TARIF=="2")

#Exclui regioao sem sentido
table(R_AUTO2003B$REGIAO)
R_AUTO2003B< filter(R_AUTO2003B,
  REGIAO%in%c("01","02","03","04","05","06","07","08","09",10:41))

```

```
write.csv2(R_AUTO2003B,"R_AUTO2003B_filtrada.csv")

#####
##### S AUTO #####
#####

#Importando os dados da S_AUTO
S_AUTO= read.csv2("S_Auto2003B.csv",stringsAsFactors = FALSE)
library(plyr)

#Consertando tipos das variaveis
S_AUTO2003B<-S_AUTO%>%mutate(APOLICE=as.character(APOLICE),REGIAO=
      as.factor(REGIAO),ITEM=as.integer(ITEM),ENDOSSO=as.integer(ENDOSSO))

#Mantemos apenas as apolices sem endosso
S_AUTO2003B <- S_AUTO2003B%>%filter(ENDOSSO==0)

#Mantemos apenas as apolices com cobertura comprehensiva
S_AUTO2003B <- S_AUTO2003B%>%filter(COBERTURA==1)

#Excluindo apolices coletivas:
S_AUTO2003B <- S_AUTO2003B%>% filter(SEX0=="F"|SEX0=="M")

#Selecionando somente acidentes que ocorreram entre 01/01/2003 e 31/06/2003
S_AUTO2003B <- S_AUTO2003B%>% filter(ymd(as.numeric(D_OCORR)) %within%
      interval(ymd("2003-01-01"),ymd("2003-06-30")))

#Mantendo apenas passeio nacional e importado
S_AUTO2003B<-filter(S_AUTO2003B,COD_TARIF=="1"|COD_TARIF=="2")

# Agregando indenizacoes em uma mesma apolice/causa/data e
# mantendo apenas o ID e indenizacoes:
S_AUTO2003B_aux <- ddply(S_AUTO2003B, .(COD_SEG,APOLICE,D_OCORR,CAUSA),
      summarize, IND_CASCO = sum(IND_CASCO), IND_RCDP = sum(IND_RCDP),
      IND_RCDM = sum(IND_RCDM), IND_APP = sum(IND_APP),
```

```

IND_OUTROS = sum(IND_OUTROS), VAL_SALVAD=sum(VAL_SALVAD))

# Agregando indenizacoes em uma mesma apolice:
S_AUTO2003B <- ddply(S_AUTO2003B, .(COD_SEG, APOLICE),
  summarize, IND_CASCO = sum(IND_CASCO), IND_RCDP = sum(IND_RCDP),
  IND_RCDM = sum(IND_RCDM), IND_APP = sum(IND_APP),
  IND_OUTROS = sum(IND_OUTROS), VAL_SALVAD=sum(VAL_SALVAD))

#Considerando como causa a causa mais grave do acidente
for(i in 1:nrow(S_AUTO2003B)){
  aux = filter(S_AUTO2003B_aux, APOLICE==S_AUTO2003B$APOLICE[i])
  causa = as.data.frame(aux$CAUSA)
  aux_causa=ifelse(any("3" == causa), "3", ifelse(any("1"==causa), "1",
    ifelse(any("4"==causa), "4", ifelse(any("2"== causa), "2", "9"))))
  S_AUTO2003B$CAUSA[i] = aux_causa
}

write.csv2(S_AUTO2003B, "S_AUTO2003B_filtrada.csv")

#####
# MERGE base S e base R tratadas:#
#####

M_AUTO=merge(R_AUTO2003B, S_AUTO2003B, by="APOLICE", all.x = TRUE)
M_AUTO$CEP3 = trunc(M_AUTO$CEP/100000) #VARIÁVEL DE SETOR REGIONAL
M_AUTO2003B=M_AUTO%>%select(COD_SEG=COD_SEG.x, APOLICE, COD_MODELO, ANO_MODELO,
  REGIAO, COD_TARIF, IS_CASCO, IS_RCDM, IS_RCDP, IS_APP, PRE_CASCO, PRE_RCDM,
  PRE_RCDP, PRE_APP, PRE_OUTROS, PERC_BONUS, PERC_CORR, SEXO, CEP, CEP3, IDADE,
  somaPremios, IND_CASCO, IND_RCDM, IND_RCDP, IND_APP, IND_OUTROS, VAL_SALVAD,
  CAUSA, TIPO_FRANQ, VAL_FRANQ, EMPRESA)

write.csv2(M_AUTO2003B, "M_AUTO2003B.csv")

#####
##### BASES RIO #####

```

```
#####

# Preenchendo valores de Causa e Custo de onde nao houve acidente com zeros:
M_AUTO$CUSTO[is.na(M_AUTO$CAUSA)]<-0
M_AUTO$CAUSA[is.na(M_AUTO$CAUSA)]<-"0"

levels(M_AUTO$CAUSA) <- c("Sem_Sinistro","Roubo/Furto","Colisao",
    "Perda_Total","Incendio","Emergencia_na_Estrada")

#Custo da seguradora:
M_AUTO = M_AUTO %>% mutate(IND_TOTAL=IND_CASCO+IND_RCDM+IND_RCDP+IND_APP+
    IND_OUTROS) %>% mutate(CUSTO = ifelse(CAUSA=="Emergencia_na_Estrada",
    IND_TOTAL-VAL_SALVAD, IND_TOTAL-VAL_SALVAD-VAL_FRANQ))

#Definindo variaveis
M_AUTO$ACIDENTEGRAVE<-ifelse(M_AUTO$CAUSA=="Roubo/Furto"|M_AUTO$CAUSA=="
    "Perda_Total"|M_AUTO$CAUSA=="Incendio",1,0)
M_AUTO$COLISAOPARCIAL<-ifelse(M_AUTO$CAUSA=="Colisao",1,0)
M_AUTO$EMERGENCIA<-ifelse(M_AUTO$CAUSA=="Emergencia_na_Estrada",1,0)
M_AUTO$SINISTRO<-ifelse(M_AUTO$CAUSA=="Sem_Sinistro",0,1)

#Mantendo apenas as colunas que iremos utilizar
M_AUTO<-M_AUTO %>% select(APOLICE,CAUSA,EMPRESA,ANO_MODELO,COD_MODELO,REGIAO,
    COD_TARIF,IS_CASCO,IS_RCDM,IS_RCDP,IS_APP,PERC_BONUS,PERC_CORR,SEXO,
    CEP,CEP3,IDADE,somaPremios,VAL_SALVAD,CAUSA,TIPO_FRANQ,COD_SEG,VAL_FRANQ,
    IND_TOTAL,CUSTO,ACIDENTEGRAVE,COLISAOPARCIAL,EMERGENCIA,SINISTRO)

#Selecionando apenas regiao metropolitana do RJ e CEPs da capital
RIO<-filter(M_AUTO,as.character(REGIAO)==18)%>%
    filter(CEP3%in%c(200:237)) %>% droplevels()
rm(M_AUTO)
glimpse(RIO)

#Selecionando apenas franquias Normal, Reduzida e Majorada
table(RIO$TIPO_FRANQ)
RIO<-filter(RIO,TIPO_FRANQ%in%c(1,2,3))
```

```

#Excluindo os outliers de importancia segura do Casco
quantile(RIO$IS_CASCO, probs = c(0.001, 0.002, 0.1, 0.5, 0.9, 0.99, 0.998, 0.999))
RIO <- RIO%>%filter(IS_CASCO>3500&IS_CASCO<100000)

#Acrescentando características do Automovel
library("stringr")
MODELOS <- read.csv2("Modelos_Auto.csv")
MODELOS$MARCA<-as.factor(sub('.*\\-', '', MODELOS$DESCRICAO))
MODELOS$um_ponto_zero<-str_detect(MODELOS$DESCRICAO, pattern= fixed("1.0"))
MODELOS$um_ponto_dois<-str_detect(MODELOS$DESCRICAO, pattern= fixed("1.2"))
MODELOS$um_ponto_tres<-str_detect(MODELOS$DESCRICAO, pattern= fixed("1.3"))
MODELOS$um_ponto_quatro<-str_detect(MODELOS$DESCRICAO, pattern= fixed("1.4"))
MODELOS$um_ponto_cinco<-str_detect(MODELOS$DESCRICAO, pattern= fixed("1.5"))
MODELOS$um_ponto_seis<-str_detect(MODELOS$DESCRICAO, pattern= fixed("1.6"))
MODELOS$um_ponto_sete<-str_detect(MODELOS$DESCRICAO, pattern= fixed("1.7"))
MODELOS$um_ponto_oito<-str_detect(MODELOS$DESCRICAO, pattern= fixed("1.8"))
MODELOS$dois_ponto_zero<-str_detect(MODELOS$DESCRICAO, pattern= fixed("2.0"))
MODELOS$dois_ponto_dois<-str_detect(MODELOS$DESCRICAO, pattern= fixed("2.2"))
MODELOS$dois_ponto_quatro<-str_detect(MODELOS$DESCRICAO, pattern= fixed("2.4"))
MODELOS$dois_ponto_seis<-str_detect(MODELOS$DESCRICAO, pattern= fixed("2.6"))
MODELOS$dois_ponto_oito<-str_detect(MODELOS$DESCRICAO, pattern= fixed("2.8"))
MODELOS$tres_ponto_zero<-str_detect(MODELOS$DESCRICAO, pattern= fixed("3.0"))
MODELOS$tres_ponto_dois<-str_detect(MODELOS$DESCRICAO, pattern= fixed("3.2"))
MODELOS$tres_ponto_quatro<-str_detect(MODELOS$DESCRICAO, pattern= fixed("3.4"))
MODELOS$tres_ponto_seis<-str_detect(MODELOS$DESCRICAO, pattern= fixed("3.6"))
MODELOS$tres_ponto_oito<-str_detect(MODELOS$DESCRICAO, pattern= fixed("3.8"))
MODELOS$quatro_ponto_zero<-str_detect(MODELOS$DESCRICAO, pattern= fixed("4.0"))
MODELOS$PALIO <- str_detect(MODELOS$DESCRICAO, pattern = fixed("PALIO"))
MODELOS$GOL <- str_detect(MODELOS$DESCRICAO, pattern = fixed("GOL"))
MODELOS$PARATI <- str_detect(MODELOS$DESCRICAO, pattern = fixed("PARATI"))
MODELOS$SANTANA <- str_detect(MODELOS$DESCRICAO, pattern = fixed("SANTANA"))
MODELOS$ESCORT <- str_detect(MODELOS$DESCRICAO, pattern = fixed("ESCORT"))
MODELOS$KA <- str_detect(MODELOS$DESCRICAO, pattern = fixed("KA"))
MODELOS$CORSA <- str_detect(MODELOS$DESCRICAO, pattern = fixed("CORSA"))
MODELOS$ASTRA <- str_detect(MODELOS$DESCRICAO, pattern = fixed("ASTRA"))

```

```

MODELOS$FOX <- str_detect(MODELOS$DESCRICA0,pattern = fixed("FOX"))
MODELOS$UNO <- str_detect(MODELOS$DESCRICA0,pattern = fixed("UNO"))
MODELOS$MONZA <- str_detect(MODELOS$DESCRICA0,pattern = fixed("MONZA"))
MODELOS$VECTRA <- str_detect(MODELOS$DESCRICA0,pattern = fixed("VECTRA"))
MODELOS$CLIO <- str_detect(MODELOS$DESCRICA0,pattern = fixed("CLIO"))
MODELOS$HILUX <- str_detect(MODELOS$DESCRICA0,pattern = fixed("HILUX"))
MODELOS$GOLF <- str_detect(MODELOS$DESCRICA0,pattern = fixed("GOLF"))
MODELOS$STILO <- str_detect(MODELOS$DESCRICA0,pattern = fixed("STILO"))
MODELOS$IPANEMA <- str_detect(MODELOS$DESCRICA0,pattern = fixed("IPANEMA"))

summary(MODELOS)

RIO<-left_join(RIO,MODELOS,by=c("COD_MODELO"="CODIGO")) %>%
  mutate(COD_TARIF=COD_TARIF.x)%>%
  select(-c(DESCRICA0,COD_TARIF.y,COD_TARIF.x,COD_MODELO))

#Removendo os modelos sem sentido
RIO<-filter(RIO,is.na(RIO$MARCA)==FALSE)%>%droplevels()

#Estatísticas descritivas
custo_medio_rio <- RIO%>%filter(CAUSA!="Sem_Sinistro")%>%group_by(CAUSA) %>%
  summarise(custo=mean(CUSTO),SD_custo=sd(CUSTO),indenizacao=
    mean(IND_TOTAL),franquia=mean(VAL_FRANQ),salvado=mean(VAL_SALVAD))
summary(RIO)
table(RIO$MARCA)

levels(RIO$MARCA)

#Agregando marcas pouco frequentes
levels(RIO$MARCA) <- c("Kya_Motors","Audi","BMW","Chrysler","Citroen","Daewoo",
  "Toyota","Outros","Fiat","Ford","General_Motors","General_Motors",
  "Honda","Honda_Nacional", "Hyundai","Rover","Outros","Kya_Motors",
  "General_Motors", "Mazda", "Mercedes_Benz","Ford","Mitsubishi",
  "Japonesa","Outros", "Peugeot","Renault","General_Motors","Seat",
  "Japonesa","Suzuki","Toyota","Outros","Volkswagen","Volvo","Rover")

# Renomeando os Niveis

```

```

levels(RIO$EMPRESA) <- c("ALFA","ALLIANZ","BMG","BRASILVEICULOS","CAIXA",
    "COSESP","GRALHA_AZUL","CONFIANCA","HDI","INDIANA","ITAU","MARITIMA",
    "MITSUI","PORTO_SEGURO","TOKIO_MARINE","TOKIO_MARINE",
    "UNIBANCO_AIG","ZURICH","ZURICH")
table(RIO$EMPRESA)

# Tokio Marine: 10% de desconto para categoria de bonus 1, 15% para 2,...
# ate 30% para categorias >=5

table(RIO$PERC_BONUS,RIO$EMPRESA)
for(i in 1:nrow(RIO)){
  if(RIO$EMPRESA[i]=="TOKIO_MARINE" && RIO$PERC_BONUS[i]>=5)
    {RIO$PERC_BONUS[i]=30}
  if(RIO$EMPRESA[i]=="TOKIO_MARINE" && RIO$PERC_BONUS[i]==1)
    {RIO$PERC_BONUS[i]=10}
  if(RIO$EMPRESA[i]=="TOKIO_MARINE" && RIO$PERC_BONUS[i]==2)
    {RIO$PERC_BONUS[i]=15}
  if(RIO$EMPRESA[i]=="TOKIO_MARINE" && RIO$PERC_BONUS[i]==3)
    {RIO$PERC_BONUS[i]=20}
  if(RIO$EMPRESA[i]=="TOKIO_MARINE" && RIO$PERC_BONUS[i]==4)
    {RIO$PERC_BONUS[i]=25}
}

summary(RIO)

####FILTRO APENAS PARA BONUS>0 (clientes antigos)####
RIO_B=filter(RIO,PERC_BONUS>0)
nrow(RIO_B)/nrow(RIO)
#proporcao de apolices com bonus positivo na base (69%)

market_share_bonus_positivo <- prop.table(table(RIO_B$EMPRESA)) %>%
  as.data.frame() %>% arrange(Freq)
market_share_bonus_positivo

# Considero as 4 maiores (market share do mercado com bonus positivo > 10%)
# e agrupo o resto como "outros"

```

```

# Porto, Itau, Tokio Marine e Unibanco representam mais de 80% deste mercado.
levels(RIO_B$EMPRESA)
levels(RIO_B$EMPRESA)<-c("OUTRAS","OUTRAS","OUTRAS","OUTRAS",
      "OUTRAS","OUTRAS","OUTRAS","ITAU","OUTRAS","OUTRAS","PORTO_SEGURO",
      "TOKIO_MARINE","UNIBANCO_AIG","OUTRAS")
levels(RIO_B$EMPRESA)<-c("C","B","A","D","E")

#-----#
#-----CUSTO ESPERADO-----#
#-----#

# Criando Variavel de Gravidade
RIO_B$GRAVIDADE<-as.factor(ifelse(RIO_B$CAUSA=="Sem_Sinistro",
      "0",ifelse(RIO_B$CAUSA=="Emergencia_na_Estrada","EMERGENCIA",
      ifelse(RIO_B$CAUSA=="Colisao","COLISAOPARCIAL","ACIDENTEGRAVE"))))

#Para dividir a base em treino e teste
library('caTools')
set.seed(100)
split<-sample.split(RIO_B,SplitRatio = 0.8)
treino<-subset(RIO_B,split=="TRUE")
teste<-subset(RIO_B,split=="FALSE")

#####
##### REGRESSAO SIMPLES #####
#####

model1<-lm(CUSTO~PERC_BONUS+CEP+MARCA+IS_CASCO+IS_RCDM+IDADE+ANO_MODELO
      ,data=treino)
summary(model1)
#calculando o MSE
teste$ajustado<-predict(model1,teste)
real<-as.numeric(teste[, "CUSTO"])
MSE_LM_CUSTO<- mean((real-teste$ajustado)^2,na.rm=TRUE)
MSE_LM_CUSTO

```

```
#####
##### RANDOM FOREST #####
#####

library("randomForest")

set.seed(200)
ajuste.rf<-randomForest(CUSTO ~ PERC_BONUS+CEP+MARCA+IS_CASCO+IS_RCDM+IDADE+
                        ANO_MODELO,data=treino,ntree=100,maxnodes=40)
pred.rf<-predict(ajuste.rf,newdata=teste)
MSE_RF_CUSTO<- mean((real-pred.rf)^2)

#####
##### RANDOM FOREST NOS RESIDUOS DO MQO #####
#####

treino$residuo <- treino$CUSTO - predict(model1,treino)

set.seed(100)
residuo.rf1<-randomForest(residuo ~PERC_BONUS+IDADE+ CEP ,
                        data=treino,ntree=100,maxnodes=50)
residuo.rf1
pred.residuo.rf1<-predict(residuo.rf1,newdata=teste)
MSE_MQORF_CUSTO1<- mean((real-teste$ajustado-pred.residuo.rf1)^2)

#####
# CUSTO ESPERADO DE CADA TIPO DE SINISTRO#
#####

#####COM REGRESSOES:

#*****
#Custo de colisao parcial :
```

```

col_pc_treino<-treino%>%select(MARCA,IS_CASCO,
    ANO_MODELO,IDADE,SEXO,COD_TARIF,TIPO_FRANQ,COD_SEG,CUSTO,
    COLISAOPARCIAL,CEP3,um_ponto_zero,um_ponto_seis,um_ponto_oito,
    dois_ponto_dois,dois_ponto_zero)%>%filter(COLISAOPARCIAL==1)
reg_custo_pc<-lm(CUSTO~IS_CASCO+IDADE+
    COD_TARIF+SEXO+TIPO_FRANQ+um_ponto_oito+dois_ponto_dois+
    dois_ponto_zero+CEP3,data=col_pc_treino)
summary(reg_custo_pc)

#CEPS Significativos (5% de significancia):205,215,218,224
col_pc_treino$CEP205=ifelse(col_pc_treino$CEP3=="205","1","0")
col_pc_treino$CEP215=ifelse(col_pc_treino$CEP3=="215","1","0")
col_pc_treino$CEP218=ifelse(col_pc_treino$CEP3=="218","1","0")
col_pc_treino$CEP224=ifelse(col_pc_treino$CEP3=="224","1","0")
table(col_pc_treino$CEP3)
teste$CEP205=ifelse(teste$CEP3=="205","1","0")
teste$CEP215=ifelse(teste$CEP3=="215","1","0")
teste$CEP218=ifelse(teste$CEP3=="218","1","0")
teste$CEP224=ifelse(teste$CEP3=="224","1","0")

#Refazendo a regressao considerando apenas as dummies dos 215 3 224 CEPs:
reg_custo_pc<-lm(CUSTO~IS_CASCO+CEP205+CEP215+CEP218+CEP224+COD_TARIF+
    IDADE+TIPO_FRANQ+um_ponto_oito+dois_ponto_dois+dois_ponto_zero,data=col_pc_treino)
summary(reg_custo_pc)
#Criando Variavel com o custo previsto de colisao parcial
CP_CUSTO_PREVISTO<-predict(reg_custo_pc,teste)

#*****
#Assistencia 24 horas:
ass_24_treino<-treino%>%select(SEXO,IDADE,CEP3,IS_CASCO,CUSTO,CAUSA,
    IND_TOTAL,VAL_SALVAD,COD_TARIF,COD_SEG)%>%
    filter(CAUSA=="Emergencia_na_Estrada")
reg_custo_24<-lm(CUSTO~IS_CASCO+CEP3+SEXO+IDADE,data=ass_24_treino)
summary(reg_custo_24)
#CEPS Significativos (5% de significancia na regressao com CEP3:212
ass_24_treino$CEP212=ifelse(ass_24_treino$CEP3=="212","1","0")

```

```

table(ass_24_treino$CEP3)
teste$CEP212=ifelse(teste$CEP3=="212","1","0")
#Refazendo a regressao considerando apenas as dummies do CEP acima
reg_custo_24<-lm(CUSTO~IS_CASCO+CEP212+SEXO+IDADE,data=ass_24_treino)
summary(reg_custo_24)

#Criando Variavel com o custo previsto de Acidente na Estrada:
ASS24_CUSTO_PREVISTO<-predict(reg_custo_24,teste)

#*****
#Acidente grave:
ac_gv_treino<-treino%>%select(IS_CASCO,IS_APP,
                             IS_RCDM,IS_RCDP,CUSTO,MARCA,um_ponto_oito,um_ponto_zero,
                             dois_ponto_zero,dois_ponto_dois,ACIDENTEGRAVE,SEXO,CEP3,IDADE,
                             COD_SEG,COD_TARIF)%>%filter(ACIDENTEGRAVE==1)
reg_custo_ag<-lm(CUSTO~IS_CASCO+CEP3+IS_RCDM+IS_RCDP+
                 COD_TARIF+IDADE+SEXO+um_ponto_oito+dois_ponto_dois+
                 dois_ponto_zero,data=ac_gv_treino)
summary(reg_custo_ag)
#Criando dummies para os niveis de CEP significativos
ac_gv_treino$CEP205=ifelse(ac_gv_treino$CEP3=="205","1","0")
ac_gv_treino$CEP207=ifelse(ac_gv_treino$CEP3=="207","1","0")
ac_gv_treino$CEP219=ifelse(ac_gv_treino$CEP3=="219","1","0")
table(ac_gv_treino$CEP3)
teste$CEP205=ifelse(teste$CEP3=="205","1","0")
teste$CEP207=ifelse(teste$CEP3=="207","1","0")
teste$CEP219=ifelse(teste$CEP3=="219","1","0")
#Refazendo a regressao considerando apenas as dummies dos CEPs relevantes:
reg_custo_ag<-lm(CUSTO~IS_CASCO+CEP205+CEP207+CEP219+IS_RCDM+COD_TARIF+
                 IDADE+um_ponto_oito+dois_ponto_dois+dois_ponto_zero,data=ac_gv_treino)
summary(reg_custo_ag)
#Criando Variavel com o custo previsto de Acidente na Estrada:
AG_CUSTO_PREVISTO<-predict(reg_custo_ag,teste)

#####COMO PORCENTAGEM DO IS_CASCO

```

```

#####
#Acidente Grave

cor(ac_gv_treino$CUSTO,ac_gv_treino$IS_CASCO)
AG_PERC_ISCASCO<-mean(ac_gv_treino$CUSTO/ac_gv_treino$IS_CASCO)
ac_gv_teste<-teste[,1:30]>%filter(ACIDENTEGRAVE=="1")>%
  select(CUSTO,IS_CASCO)
mean(ac_gv_teste$CUSTO)
mean(ac_gv_teste$IS_CASCO*AG_PERC_ISCASCO)
#erro medio
mean(abs((ac_gv_teste$IS_CASCO*AG_PERC_ISCASCO)-ac_gv_teste$CUSTO))

#####
#Colisao Parcial:
cor(col_pc_treino$CUSTO,col_pc_treino$IS_CASCO)
CP_PERC_ISCASCO<-mean(col_pc_treino$CUSTO/col_pc_treino$IS_CASCO)
col_pc_teste<-teste%>%filter(COLISAOPARCIAL=="1")>%select(CUSTO,IS_CASCO)
#erro medio
mean(abs((col_pc_teste$IS_CASCO*CP_PERC_ISCASCO)-col_pc_teste$CUSTO))

#####
#Assistencia 24hrs
summary(ass_24_treino$CUSTO)
cor(ass_24_treino$CUSTO,ass_24_treino$IS_CASCO)
ASS24_PERC_ISCASCO<-mean(ass_24_treino$CUSTO/ass_24_treino$IS_CASCO)
ass_24_teste<-teste%>%filter(EMERGENCIA=="1")>%select(CUSTO,IS_CASCO)
mean(abs((ass_24_teste$IS_CASCO*ASS24_PERC_ISCASCO)-ass_24_teste$CUSTO))
mean(abs(median(ass_24_teste$CUSTO)-ass_24_teste$CUSTO))
MEDIANA_ASS24<-median(ass_24_teste$CUSTO)

#####
# PROBABILIDE DE SINISTRO #
#####

```

```
#####
#Logits#
#####

#### Para prever a ocorrencia ou nao de um Sinistro ####

model_SINISTRO<-glm(SINISTRO~ANO_MODELO+MARCA+um_ponto_zero+
  um_ponto_oito+TIPO_FRANQ+IS_CASCO+IDADE+PERC_BONUS+
  IS_APP+IS_RCDM+IS_RCDP+CEP3+SEXO,treino,family="binomial")
summary(model_SINISTRO)
res_sin<-predict.glm(model_SINISTRO,teste,type="response")
teste$Logit_SINISTRO<-predict.glm(model_SINISTRO,teste,type="response")
quantile(teste$Logit_SINISTRO,probs = c(0,1,2,3,4,5,6,7,8,9,10)/10)

##Matriz de confusao
table(teste$SINISTRO)[2]/sum(table(teste$SINISTRO))
CF1 = teste$Logit_SINISTRO > 0.1 # faz a limiarizacao
CF_1 = table(teste$SINISTRO, CF1, deparse.level = 2) # matriz de confusao
show(CF_1) # mostra os resultados
ACC1 = sum(diag(CF_1))/sum(CF_1) # calcula a acuracia
show(ACC1) # mostra a acuracia
precisao1= CF_1[2,2]/(CF_1[2,2] + CF_1[1,2])
precisao1 # quantidade de sinistros previstos que efetivamente ocorrem
sensitividade1= CF_1[2,2]/(CF_1[2,2] + CF_1[2,1])
sensitividade1 # quantidade de sinistros que foram previstos

#### Para prever a ocorrencia ou nao de um Acidente Grave ####

model_PT<-glm(ACIDENTEGRAVE~ANO_MODELO+MARCA+um_ponto_zero+um_ponto_oito+
  TIPO_FRANQ+IS_CASCO+IDADE+PERC_BONUS+IS_APP+IS_RCDM+IS_RCDP+CEP3+SEXO,
  treino,family="binomial")
library(stats)
res_logit_ag<-predict.glm(model_PT,teste,type="response")
real_logit_ag<-as.numeric(teste[, "ACIDENTEGRAVE"])
quantile(res_logit_ag,probs = c(0,1,2,3,4,5,6,7,8,9,9.92,10)/10)
```

```

RIO_B$res_logit_ag<-predict.glm(model_PT,RIO_B,type="response")

#Matriz de confusao
CF2 = res_logit_ag > 0.012 # faz a limiarizacao
CF_2 = table(teste$ACIDENTEGRAVE, CF2, deparse.level = 2) # matriz de confusao
show(CF_2) # mostra os resultados
ACC2 = sum(diag(CF_2))/sum(CF_2) # calcula a acuracia
show(ACC2) # mostra a acuracia
precisao2= CF_2[2,2]/(CF_2[2,2] + CF_2[1,2])
precisao2 # quantidade de sinistros previstos que efetivamente ocorrem
sensitividade2= CF_2[2,2]/(CF_2[2,2] + CF_2[2,1])
sensitividade2 # quantidade de sinistros que foram previstos

#### Para prever a ocorrencia ou nao de uma Colisao Parcial ####

model_CP<-glm(COLISAOPARCIAL~ANO_MODELO+um_ponto_zero+um_ponto_oito+
  dois_ponto_zero+TIPO_FRANQ+IS_CASCO+IDADE+PERC_BONUS+IS_APP+IS_RCDM+
  IS_RCDP+CEP3+SEXO,treino,family="binomial")
res_logit_CP<-predict.glm(model_CP,teste,type="response")
real_logit_CP<-as.numeric(teste[, "COLISAOPARCIAL"])
quantile(res_logit_CP,probs = c(0,1,2,3,4,5,6,7,8,9,9.8,10)/10)
nrow(col_pc_treino)/nrow(treino)

RIO_B$res_logit_CP<-predict.glm(model_CP,RIO_B,type="response")

#Matriz de confusao
CF4 = res_logit_CP > 0.034 # faz a limiarizacao
CF_4 = table(teste$COLISAOPARCIAL, CF4, deparse.level=2) # matriz de confusao
show(CF_4) # mostra os resultados
ACC4 = sum(diag(CF_4))/sum(CF_4) # calcula a acuracia
show(ACC4) # mostra a acuracia
precisao4= CF_4[2,2]/(CF_4[2,2] + CF_4[1,2])
precisao4 # quantidade de sinistros previstos que efetivamente ocorrem

```

```

#### Para prever a ocorrencia ou nao de uma Emergencia na Estrada ####

model_24<-glm(EMERGENCIA~ANO_MODELO+TIPO_FRANQ+IS_CASCO+IDADE+PERC_BONUS+
             IS_APP+IS_RCDM+IS_RCDP+CEP3+SEXO,treino,family="binomial")
res_logit_24<-predict.glm(model_24,teste,type="response")
real_logit_24<-as.numeric(teste[, "EMERGENCIA"])
quantile(res_logit_24,probs = c(0,1,2,3,4,5,6,7,8,9,9.4,9.5,10)/10)
nrow(ass_24_teste)/nrow(teste)

RIO_B$res_logit_24<-predict.glm(model_24,RIO_B,type="response")

#Matriz de confusao
CF6 = res_logit_24 > 0.10 # faz a limiarizacao
CF_6 = table(teste$EMERGENCIA, CF6, deparse.level = 2) # matriz de confusao
show(CF_6) # mostra os resultados
ACC6 = sum(diag(CF_6))/sum(CF_6) # calcula a acuracia
show(ACC6) # mostra a acuracia
precisao6= CF_6[2,2]/(CF_6[2,2] + CF_6[1,2])
precisao6 # quantidade de sinistros previstos que efetivamente ocorrem

#####
#####LOGIT SEQUENCIAL (CONTINUATION RATIO MODEL)#####
#####

library("glmnetcr")

#ordenando os dados com a severidade do acidente:
ordem<-c("0","EMERGENCIA","COLISAOPARCIAL","ACIDENTEGRAVE")
library(gdata)
treino$GRAVIDADE <- as.ordered(reorder.factor(treino$GRAVIDADE,

```

```

        new.order=ordem))
treino<-treino %>%arrange(GRAVIDADE)

#Preparando os dados para a funcao,que so aceita dado numerico
library(dummies)
y <- treino$GRAVIDADE
x <- dummy.data.frame(treino[,c("ANO_MODELO","TIPO_FRANQ","COD_TARIF",
    "IS_CASCO","IS_RCDM","IDADE","PERC_BONUS","IS_APP","IS_RCDM","IS_RCDP",
    "CEP3","SEXO","MARCA","um_ponto_zero","um_ponto_oito","dois_ponto_zero")])
x_teste <- dummy.data.frame(teste[,c("ANO_MODELO","TIPO_FRANQ","COD_TARIF",
    "IS_CASCO","IS_RCDM","IDADE","PERC_BONUS","IS_APP","IS_RCDM","IS_RCDP",
    "CEP3","SEXO","MARCA","um_ponto_zero","um_ponto_oito","dois_ponto_zero")])

#Testando
cratio <- glmnetcr(x, y)
BIC.step <- select.glmnetcr(cratio) #extraí o modelo com melhor ajuste
BIC.step # Step 44 (s44) corresponde ao modelos que minimiza o BIC
c_ratio<-fitted(cratio,x_teste,s = BIC.step)
probs_cratio<-as.data.frame(c_ratio$probs)
summary(probs_cratio)

#####
# Custo esperado como soma das probs*customedio #
#####

#LOGITS BINARIOS

teste$CUSTO_LOGIT1<-(res_logit_ag*AG_CUSTO_PREVISTO)+
    (res_logit_CP*CP_CUSTO_PREVISTO)+(res_logit_ag*ASS24_CUSTO_PREVISTO)
MSE_LOGIT1<- mean(((teste$CUSTO_LOGIT1-teste$CUSTO)^2))

teste$CUSTO_LOGIT2<-(res_logit_ag*AG_PERC_ISCASCO*teste$IS_CASCO)+
    (res_logit_ag*CP_PERC_ISCASCO*teste$IS_CASCO)+
    (res_logit_ag*median(ass_24_treino$CUSTO))
MSE_LOGIT2<- mean(((teste$CUSTO_LOGIT2-teste$CUSTO)^2))

```

```

teste$CUSTO_LOGIT3<-(res_logit_ag*AG_PERC_ISCASCO*teste$IS_CASCO)+
  (res_logit_ag*CP_CUSTO_PREVISTO)+
  (res_logit_ag*median(ass_24_treino$CUSTO))
MSE_LOGIT3<- mean((teste$CUSTO_LOGIT3-teste$CUSTO)^2)

#CRATIO

teste$CUSTO_CRATIO<-(probs_cratio$ACIDENTEGRAVE*AG_CUSTO_PREVISTO)+
  (probs_cratio$COLISAOPARCIAL*CP_CUSTO_PREVISTO)+
  (probs_cratio$EMERGENCIA*ASS24_CUSTO_PREVISTO)
MSE_CRATIO<- mean((teste$CUSTO_CRATIO-teste$CUSTO)^2)

teste$CUSTO_CRATIO2<-
  (probs_cratio$ACIDENTEGRAVE*AG_PERC_ISCASCO*teste$IS_CASCO)+
  (probs_cratio$COLISAOPARCIAL*CP_PERC_ISCASCO*teste$IS_CASCO)+
  (probs_cratio$EMERGENCIA*median(ass_24_treino$CUSTO))
MSE_CRATIO2<- mean((teste$CUSTO_CRATIO2-teste$CUSTO)^2)

teste$CUSTO_CRATIO3<-
  (probs_cratio$ACIDENTEGRAVE*AG_PERC_ISCASCO*teste$IS_CASCO)+
  (probs_cratio$COLISAOPARCIAL*CP_CUSTO_PREVISTO)+
  (probs_cratio$EMERGENCIA*median(ass_24_treino$CUSTO))
MSE_CRATIO3<- mean((teste$CUSTO_CRATIO3-teste$CUSTO)^2)

#####
# PROBABILIDE DE SINISTROS COM RF #
#####

#### Para prever o tipo de Acidente ####
table(treino$GRAVIDADE)

# Rebalanceamos o conjunto de dados a ser testado com "sampszie"
set.seed(100)
rf<-randomForest(GRAVIDADE~IS_CASCO+um_ponto_oito+SEXO+IS_RCDM+IDADE+

```

```

    PERC_BONUS+CEP,treino,type="classification",maxnodes=50,ntree=100,
    sampsize=c("0"=15000,"EMERGENCIA"=5000,"COLISAOPARCIAL"=3420,
               "ACIDENTEGRAVE"=1317))
teste_resrf<-as.data.frame(predict(rf,newdata=teste,type="prob"))
summary(teste_resrf) #frequencia media de acidentes proxima a base RIO_B

#####
#Custo esperado como soma das probs*custo#
#####

teste$CUSTO_RF<-(teste_resrf$ACIDENTEGRAVE*AG_CUSTO_PREVISTO)+
  (teste_resrf$COLISAOPARCIAL*CP_CUSTO_PREVISTO)+
  (teste_resrf$EMERGENCIA*ASS24_CUSTO_PREVISTO)
MSE_RF1<- mean((teste$CUSTO_RF-teste$CUSTO)^2)

teste$CUSTO_RF2<-(teste_resrf$ACIDENTEGRAVE*AG_PERC_ISCASCO*teste$IS_CASCO)+
  (teste_resrf$COLISAOPARCIAL*CP_PERC_ISCASCO*teste$IS_CASCO)+
  (teste_resrf$EMERGENCIA*median(ass_24_treino$CUSTO))
MSE_RF2<- mean((teste$CUSTO_RF2-teste$CUSTO)^2)

teste$CUSTO_RF3<-(teste_resrf$ACIDENTEGRAVE*AG_PERC_ISCASCO*teste$IS_CASCO)+
  (teste_resrf$COLISAOPARCIAL*CP_CUSTO_PREVISTO)+
  (teste_resrf$EMERGENCIA*median(ass_24_treino$CUSTO))
MSE_RF3<- mean((teste$CUSTO_RF3-teste$CUSTO)^2)

#-----#
#-----PREMIO ESPERADO-----#
#-----#

#Novas variaveis
RIO_B$IDADE2<-RIO_B$IDADE*RIO_B$IDADE
RIO_B$pred.rf<-predict(ajuste.rf,newdata=RIO_B)

#Regressao de Premio
reg_custo1<-lm(somaPremios~IDADE+um_ponto_zero+um_ponto_tres+um_ponto_quatro+

```

```

        um_ponto_cinco+um_ponto_seis+um_ponto_sete+um_ponto_oito+
        dois_ponto_zero+dois_ponto_dois+MARCA+IDADE2+PERC_BONUS+CEP3+SEX0+
        IS_CASCO+IS_RCDM+IS_RCDP+IS_APP+VAL_FRANQ+ANO_MODELO+COD_TARIF+EMPRESA+
        PALIO+GOL+PARATI+SANTANA+ESCORT+KA+CORSA+ASTRA+UNO+MONZA+VECTRA+CLIO+
        HILUX+GOLF+STILO+IPANEMA+pred.rf,RIO_B)
summary(reg_custo1) #R2 = 61,82%

reg_custo2<-lm(somaPremios~pred.rf,RIO_B)
summary(reg_custo2) #R2: 16,8 %

library(randomForest)
set.seed(100)
rf1 <- randomForest(somaPremios~IDADE+um_ponto_zero+um_ponto_tres+
        um_ponto_quatro+um_ponto_cinco+um_ponto_seis+um_ponto_sete+
        um_ponto_oito+dois_ponto_zero+dois_ponto_dois+MARCA+PERC_BONUS+CEP+SEX0+
        IS_CASCO+IS_RCDM+IS_RCDP+IS_APP+VAL_FRANQ+ANO_MODELO+COD_TARIF+EMPRESA+
        PALIO+GOL+PARATI+SANTANA+ESCORT+KA+CORSA+ASTRA+UNO+MONZA+VECTRA+CLIO+
        HILUX+GOLF+STILO+IPANEMA+pred.rf,data = RIO_B,ntree=100,nodesize=100)
rf1 #70,71% da variancia explicada.

png(filename="imp_var_rf_premio.png")
varImpPlot(rf1,main="Importancia das Variaveis")
dev.off()

#-----#
#-----DEMANDA-----#
#-----#

library(mlogit)
library(psych)
library(mvtnorm)
library(bayesm)
library(gmnl)

# Base de Dados

RIO_B_long <- read.csv("Base_Long_Rio_Bonus_Positivo.csv")%>%

```

```

mutate(CEP3=as.factor(CEP3))
levels(RIO_B_long$EMPRESA)<-c("B","C","A","D","E")
dados <- mlogit.data(RIO_B_long,choice="EMP",shape="long",
  alt.levels = c("B","C","A","D","E"))

# Funcoes =====
# As seguintes funcoes a necessarias para calcular as elasticidades:
# Essas funcoes sao inspiradas no codigo de Elshiewy, Ossama,
# Daniel Guhl, e Yasemin Boztug (2017)

# Esta funcao computa as probabilidades e market shares para um modelo
# (heterogeneo) MNL dados draws da distribuicao populacional estimada.

mnlProb <- function(x, betai) {
  u <- x %*% t(betai)
  num <- exp(u)
  tprobs <- t(num) / colSums(num)
  probs <- t(tprobs)
  shares <- rowMeans(probs)
  res <- list(shares = shares,
    probs = probs)
  return(res)
}

# Esta funcao computa a matriz de primeiras derivadas para um modelo MNL
# dadas preferencias (heterogeneas) e probabilidades correspondentes.

mnlDeriv <- function(betaxi, probi) {
  nJ <- dim(probi)[1]
  nD <- dim(probi)[2]
  betaxim <- matrix(betaxi, nJ, nD, byrow = TRUE)
  x <- rep(1, nJ)
  eta <- -(1 / x) %*% t(rep(1, nJ)) * tcrossprod(betaxim * probi, probi) / nD
  diag(eta) <- (1 / x) * rowSums(betaxim * probi * (1 - probi)) / nD
  return(eta)
}

```

```
# Esta funcao computa a matriz de elasticidades. "coef" especifica a posicao
# da variavel de interesse em x e beta.
```

```
mn1Elast <- function(x, betai, coef = 5) {
  nJ <- nrow(x)
  probs <- mn1Prob(x, betai)
  shares <- probs$shares
  shares <- matrix(shares, nJ, nJ, byrow = TRUE)
  eta <- x[, coef] / shares * mn1Deriv(betai[, coef], probs$probs)
  return(eta)
}
```

```
# Esta e uma funcao conveniente para computar elasticidades para varios
# conjuntos de escolha "coef" especifica a posicao da variavel de interesse
# em "data" e "betai".
```

```
computeElast <- function(data, betai, coef = 5) {
  E_chid <- sapply(data, FUN = mn1Elast,
                  betai = betai, coef = coef,
                  simplify = "array")

  E <- apply(E_chid, c(1, 2), mean)
}
```

```
# MNL -----
# benchmark model: apenas as constantes de marca
mn10 <- gmn1(EMP ~ 1, data = dados)
(ll_0 <- logLik(mn10))
#MNL model (mn10 + PREMIO_PREVISTO, IDADE e ANO_MODELO como covariaveis)
+ ANO_MODELO +TIPO_FRANQ
mn1 <- gmn1(EMP ~ PREMIO_PREVISTO| IS_CASCO + IDADE +ANO_MODELO |
          0 , data = dados)
#parametros estimados
round(summary(mn1)$CoefTable, 3)
```

```

#LL
logLik(mnl)
#McFadden R^2
round(c(1 - logLik(mnl) / ll_0), 3)

#####
# Preparando dados para calculo das elasticidades
M <- model.matrix(mFormula(EMP ~ PREMIO_PREVISTO ),
                  data = dados)
coef_names <- colnames(M)
seg_hb <- NULL
for (i in unique(dados$APOLICE)) {
  seg_hb[[i]] <-
    list(y = match(subset(dados, APOLICE == i & (EMP))[, "EMPRESA"],
                  dados[1:5, "EMPRESA"]), X = M[which(dados$APOLICE == i),])
}

seg_elast <- NULL
j=0
for (i in unique(dados$chid)) {
  M_i <- M[dados$chid == i,]
  j=1+j
  #rownames(M_i) <- c("A", "B", "C", "D", "E")
  seg_elast[[j]] <- M_i
}

#####
#Elasticidade-Preco (media de todas as observacoes)
coefs<-t(coef(mnl)[1:5])
round(computeElast(seg_elast, coefs, coef = 5), 3)

# M-MNL -----
mmnl <- gmm1(EMP ~ PREMIO_PREVISTO | 1 | 0 | IDADE + IS_CASCO -1 ,
            data=dados, model="mixl", ranp = c(PREMIO_PREVISTO= "n" ),
            mvar = list(PREMIO_PREVISTO = c("IDADE", "IS_CASCO")))

summary(mmnl)

```

```
summary(mmn1)
# Medias dos parametros aleatorios
round(summary(mmn1)$CoefTable, 3)
# LL
logLik(mmn1)
# McFadden R^2
round(c(1 - logLik(mmn1) / ll_0), 3)
# Coeficientes individuais
mmn1_beta_i <- effect.gmn1(mmn1)$mean
# Elasticidades Preco
set.seed(1234)
mmn1_beta_r<-rmvnorm(100, mmn1_beta_i[1,], cov.gmn1(mmn1))
for (i in 2:nrow(mmn1_beta_i)){
  aux_mmn1_beta_r <- rmvnorm(100, mmn1_beta_i[i,], cov.gmn1(mmn1))
  mmn1_beta_r <-rbind(mmn1_beta_r,aux_mmn1_beta_r)
}
print(computeElast(seg_elast, mmn1_beta_r,coef = 4))
```