Comparação de dois métodos de estimação de gasto com saúde

Jorge Eduardo Ortiz Aguirre

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDAD DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO
DE
MESTRE EM MATEMÁTICA APLICADA

Programa: Matemática Aplicada
Orientador: Profa. Dra. Cláudia Peixoto

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro do CAPES

São Paulo, 4 de Agosto de 2023

Resumo

Esta dissertação simula um plano de saúde do tipo "poupança"dirigido ao consumidor no contexto brasileiro. Esse plano de saúde está ativo durante a vida laboral de um indivíduo, o que corresponde as idades de 25 a 65 anos, e prevê que as despesas com saúde sejam pagas pelo empregador e um seguro catastrófico, mas com um gerenciamento do indivíduo. Este trabalho utilizou dados do plano empresarial, SABESPREV, que contém os gastos anuais efetuados por seus funcionários, para estimar os gastos com saúde individual e anual. Analisamos, a partir de simulações, se este novo produto produz um balanço dos gastos ao final da vida laboral, isto é, verificamos quanto do valor total gasto foi pago pela empresa e pelo seguro catastrófico bem como quanto o indivíduo conseguiu poupar.

Neste trabalho, para a realização da predição de gastos utilizamos as técnicas de Regressão e comparamos os resultados aqui encontrados com os resultados de [MMPP], onde a predição de gastos individuais foi feita a partir de Cadeias de Markov aplicadas às faixas etárias, bem como a distribuição empírica dos gastos em cada faixa.

O processo de estimativa dos gastos por meio de regressão consistiu em aplicar os diferentes modelos (linear, suporte vetorial, árvores, floresta aleatória, Bayesiano Lasso, Bayesiano Ridge) na predição dos gastos e, utilizando técnicas como validação cruzada obtivemos o modelo com melhor desempenho. Este estudo mostrou que há uma grande diferença entre as duas técnicas (Regressão e Cadeias de Markov) nos seguintes aspectos: no balanço das contas de saúde poupança ao término dos 41 anos de vida laboral, frequência e severidade do uso de seguro catastrófico.

Os resutlados foram obtidos por meio da individualação de cada sexo, mostrando que o modelo favorece ao sexo masculino com um gasto pagado pela empresa de 74.5%, um 0,5% pelo seguro catastrófico e o O indivíduo conseguiu poupar 25%, caso contrário para sexo feminimo onde o gasto pagado pela empresa é de 97.7%, um 2.3% pelo seguro catastrófico e o indivíduo conseguiu poupar 0.01%.

Palavras-chave: Regressão, Despesa com Saúde, Plano Poupança, Seguro Catastrófico, Plano de Saúde

Conteúdo

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

1.1 Considerações Preliminares

Os planos de saúde vigentes hoje no Brasil são do tipo "mutualismo", em que as despesas mensais da carteira são pagas com o arrecadado daquele mês por todos ali incluídos. Já o plano de saúde poupança funciona como um sistema de capitalização e ainda não existe no Brasil.

Os planos de saúde do tipo capitalização, ou as contas poupança (CP), como vamos denominar aqui, são dispositivos em que cada pessoa recebe contribuições anuais, em princípio de seu empregador, com a finalidade exclusiva de cobrir despesas de saúde. A idéia é que uma CP tenha um período inicial de acumulação de capital, para posteriormente, em idades mais avançadas, ocorrer a desacumulação.

Os gastos anuais pagos pela CP são limitados por um valor pré-definido enquanto seu saldo for positivo. Caso as despesas ultrapassem o limite ou saldo da conta, um seguro catastrofe deve ser acionado. Embora haja uma limitação para os gastos cobertos pelos fundos da conta poupança, neste dispositivo não há limitação para as despesas anuais e individuais com saúde, já que o seguro catástrofe cobre todas as despesas além dos limites estabelecidos. A dinâmica da CP é a seguinte[1]:

- No início de cada ano, o empregador deposita, por exemplo, R$ $2.500,00$ na conta indiviual de seu empregado.
- Caso as despesas anuais com saúde de um empregado não ultrapassem o saldo da conta ou o limite de, por exemplo, R$ $5.000,00$, estas são integralmente pagas com recursos da conta. Se os gastos anuais ultrapassarem o saldo da conta ou o limite estabelecido, é sacado o valor mínimo entre saldo da conta e o limite, e o valor remanescente é coberto por um seguro catástrofe.

Espera-se que dos 25 aos 65 anos esta CP esteja na fase de acumulação de recursos, os quais serão gastos após a aposentadoria ou em idades mais avançadas.

[1] Os valores utilizados como parâmetros podem ser alterados.

O objetivo desta forma de financiamento é a implementação de um compartilhamento de riscos ao longo do ciclo de vida individual, mobilizando recursos de várias fontes para financiar a fase de desacumulação.

Em 2020, Maia, Marcondes, Peixoto e Pereira [MMPP], fizeram um estudo cuja predição de gastos individuais foi feita por matrizes de transição de Cadeias de Markov e pela distribuição empírica dos dados em cada faixa. Faixas de gastos foram criadas (estados da cadeia) e foram estimadas as probabilidades de transição de uma faixa para outra com base nas variáveis sexo, faixa etária e gastos efetuados nos dois anos anteriores.

Nesta dissertação, utilizando o mesmo conjunto de dados de [MMPP], SABESPREV, foram ajustados modelos de regressão para estimar os gastos individuais bem como técnicas de validação cruzada, para "melhor"modelo. O desempenho do modelo foi medido pela raiz quadrada do erro quadrático médio (REQM). Depois de avaliar os modelos de regressão, escolhemos o mais adequado levando em consideração os critérios de *overfitting* e *underfitting*. Em seguida, realizamos uma análise comparativa entre os resultados das duas CP com técnicas diferentes de estimação de gastos; para tanto também simulamos uma carteira com 10 mil vidas.

1.2 Objetivos

Objetivos Gerais

Reproduzir o trabalho [MMPP] utilizando modelos de regressão para a previsão de gastos com saúde e comparar os resultados do novo produto proposto, as contas poupanças.

Objetivos Específicos

- Ajustar um modelo de regressão que prevê gastos com saúde utilizando o banco de dados SABESPREV.
- Simular uma CP para 10.000 vidas utilizando o modelo de regressão ajustado, durante a vida útil laboral (entre 25 e 65 anos).
- Comparar os resultados das simulações das CP com os dois métodos de estimação de gastos (Cadeias de Markov/ distribuição empírica e regressão).

1.3 Organização do Trabalho

No Capítulo 2, apresentamos a revisão da literatura contemplando algumas definições gerais, técnicas e modelos de regressão.

A metodologia utilizada para ajustar modelos de regressão na estimação de gastos com saúde situa-se no Capítulo 3.

Finalmente no Capítulo 4 encontra-se a análise dos resultados e a comparação dos dois métodos de estimação de gastos e da distribuição dos gastos.

Capítulo 2

Revisão da Literatura

2.1 Regressão

2.1.1 Definições Gerais

A análise preditiva tem como objetivo fazer previsões sobre os resultados futuros baseada em dados históricos e técnicas de análise. A operacionalidade desse processo envolve um conjunto de ferramentas, ou algoritmos, utilizados para compreender os dados existentes e gerar regras de predição.

De modo geral, essas ferramentas podem ser alocadas em uma das seguintes categorias de aprendizagem: as supervisionadas e as não supervisionadas (Hastie; Tibshirani; Friedman, 2008 [4]; Kuhn; Johnson, 2013 [7]). A aprendizagem supervisionada usa dados rotulados, enquanto a aprendizagem não supervisionada usa dados não rotulados, Isso significa que o supervisionado conhece a variável objetivo, por outro lado, a não supervisionada, tem como objetivo encontrar grupos semelhantes no conjunto de dados.

Para nosso estudo vamos usar o aprendizado supervisionado, em que cada observação $i, i = 1, \ldots, n$ do conjunto de dados, dispõe de um vetor de medições para variáveis preditoras (variáveis independentes), $X_{i,j},\ j = 1, \ldots, p$, são medidas correspondentes à resposta de interesse Y_i (variável dependente) do elemento i da amostra. Um modelo que relacione a resposta às preditoras é ajustado com o objetivo de predizer essa resposta em observações futuras, para as quais estão disponíveis apenas dados referentes às preditoras (Hastie; Tibshirani; Friedman, 2008 [4]; James, 2014 [6]). A distinção entre o tipo de variável resposta resulta em dois subgrupos de aprendizagem supervisonada: regressão, para variáveis quantitativas, e Classificação, para as categóricas (qualitativa) (Hastie; Tibshirani; Friedman, 2008 [4]).

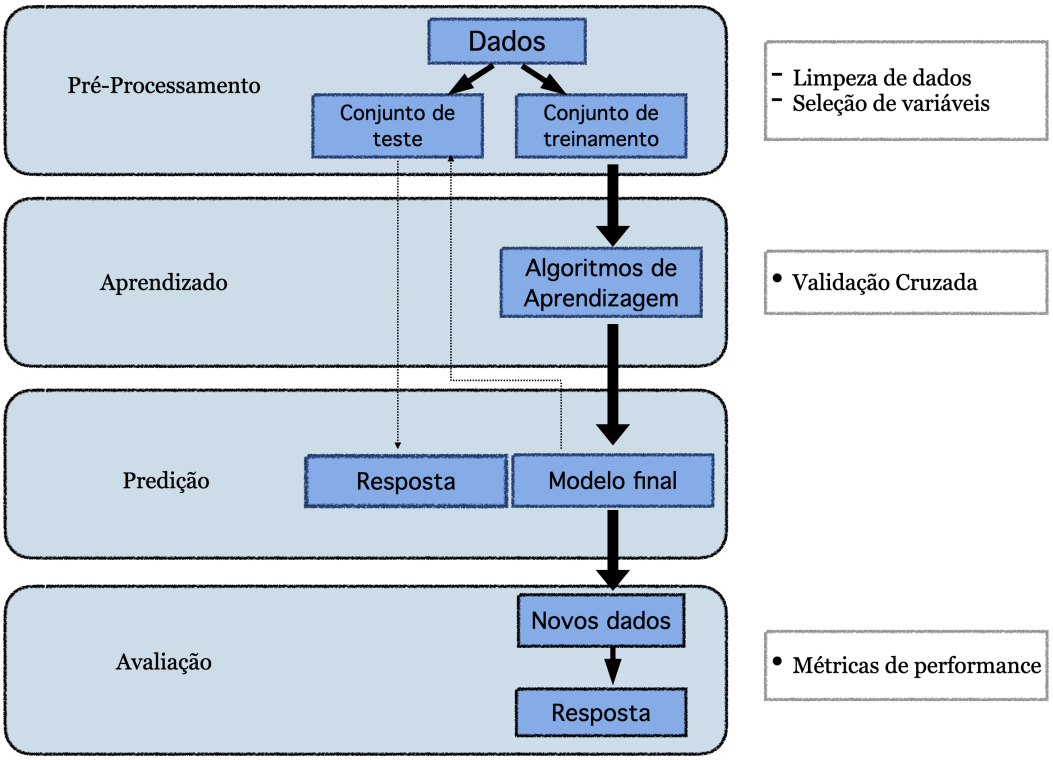

Figura 2.1: Roteiro para aplicação de regressão em análise preditiva.

Na Figura 2.1 o pré-processamento refere-se a divisão do conjunto de dados em treinamento e validação, a qual se realiza com a finalidade de determinar se o modelo tem um bom ajuste. As divisões mais utilizadas são 70:30, 75:25, 80:20 (treinamento:teste), dependendo do tamanho inicial do conjunto de dados.

2.1.2 Transformações de Dados: Preditoras Individuais

Técnicas de pré-processamento estão, de modo geral, relacionadas à organização de dados. A transformação destes apresenta estreita relação com a melhora do desempenho de muitos modelos que requerem preditoras na mesma escala. A normalização e a padronização representam as abordagens mais comuns para modificação da escala de uma variável: a primeira refere-se ao redimensionamento das variáveis para uma escala que varie entre 0 e 1, ou seja, dentro de um intervalo delimitado, e a segunda, à centralização das variáveis com média 0 com desvio padrão 1 (Kuhn; Johnson, 2013 [7]; Raschka, 2017 [5])

2.1.3 Métricas para Avaliação do Desempenho de um Modelo

A avaliação do desempenho de um modelo de regressão em um determinado conjunto de dados é realizado pela mensuração do quão bem as predições decorrentes do modelo ajustado reproduzem o valor observado para a resposta de interesse. Portanto,

é preciso quantificar a diferença entre o valor previsto pelo modelo ($\hat{y}_i$) e seu respectivo valor observado y_i, (James, 2014 [6]). A medida utilizada com mais frequência para avaliar o desempenho de um modelo é o Erro Quadrático Médio (EQM, em inglês: *MSE*), que é dado por:

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.1}$$

A raiz quadrada do EQM é chamada de REQM (em inglês: *RMSE*). Se a REQM for pequena significa que os valores preditos estão próximos dos observados, enquanto que se for grande, indica que os valores preditos diferem substancialmente dos observados. O desempenho do modelo preditivo selecionado deve ser avaliado a partir da mensuaração de REQM em dados de teste (James, 2014 [6]).

Para obter um melhor desempenho nos modelos, a técnica de validação cruzada é uma das mais utilizadas em regressão. Esse processo consiste na divisão aleatória do banco de treinamento original em k partes de tamanhos aproximadamente iguais, em que $k-1$ irão representar dados de treinamento para o ajuste do modelo preditivo e a outra parte ficará reservada para a estimativa de seu desempenho (dados de teste). O processo se repete até que todas as partes tenham participado tanto do treinamento como da validação do modelo, resultando em k estimativas de performance que serão resumidas, usualmente, pelo cálculo da média e do erro padrão (Kuhn; Johnson, 2013 [7])

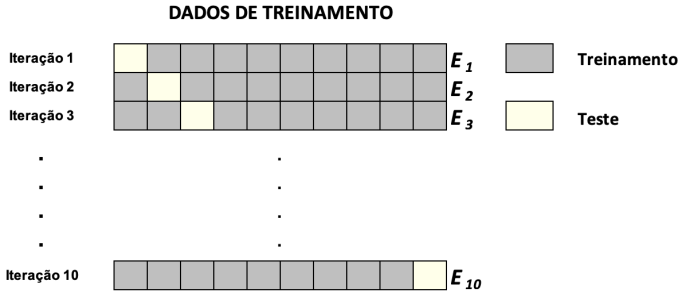

Figura 2.2: Processo de validação cruzada k-fold, k=10.

A Figura 2.2, resume o processo de validação cruzada *k-fold*, com $k = 10$. Os dados de treinamento foram divididos em 10 partes, e em cada iteração, 9 partes foram utilizadas para o treinamento (estimação do modelo) e uma parte para estimar seu desempenho. Ao final do processo, e observando a perfomance de cada modelo, E_k, usando a equação (2.1), escolhemos o modelo com o menor valor em EQM.

2.1.4 Metodos para deteção de dados atípicos

Um *outlier* é uma observação que se encontra a uma distância anormal de outros valores em uma amostra. A presença de valores atípicos em un conjunto de dados pode levar a problemas no momento de estimar os modelos de regressão. Por esta razão, existem maneiras para remover estes dados atípicos.

1. Invervalo Interquartil: Para o método baseado em interquartis, o ponto de corte foi definido como 1,5 vezes a amplitude do intervalo interquartílico (IQR). Observações com valores maiores que o terceiro quartil (Q3) mais 1,5 IQR foram descartadas.
2. Classificação: Neste abordagem, após a ordenação das observaçoes, 2,5% das maiores observações foram removidas.
3. Desvio Padão: Para valores superiores, à média mais 3*STD (Desvio Padrão) foi usado para identificar os *outliers*.
4. Distribução Gama:

$$T_k = \frac{X_{n-k} - X_{(1)}}{\sum_{j=n-k+1}^{n} X_{(j)} - X_{(1)}}$$

em que X representa o valor na posição dada, n representa a ultima observação, k representa a k-ésima observação de acima e j representa a posição k-ésima+1. A saída desta equação é uma constante obtida ao dividir a observação $n-k$-ésima pela soma acumulativa dos valores que se encontram acima da posição k-ésima. A distribuição termina quando a constante T_k for menor que $0,05$. Assim, um valor maior de T_k indicou que a k-ésima observação tem uma influência relativamente maior nos dados, mesmo na presença de valores maiores. Para este análise, observações como $T_k > 0,05$ calculados forma considerados valores atípicos.

Mais detalhes pode ver [16]

2.2 Modelos de Regressão

Diversos modelos de regressão foram desenvolvidos para solucionar diferentes problemas, sendo essencial compará-los para selecionar aquele que obtêm melhor predição em algum sentido, pois nenhum modelo domina todos os outros, ou seja, em um conjunto de dados particular, um determinado modelo pode funcionar melhor, mas o mesmo pode não ser o que apresenta melhor performance em outro conjunto de dados (James, 2014 [6]). Também é bom observar o custo computacional do ajuste, a facilidade de implementar a função de predição estimada, e também é muito importante a interpretabilidade do modelo. A seguir temos uma pequena descrição dos modelos, desde o mais complexo até o mais simples para a escolha do modelo final. Uma boa descrição das etapas de como chegar ao modelo final pode ser vista em Kuhn & Johnson 2013 [7], que, em resumo, descreve os seguintes passos a serem seguidos:

1. Avaliar, inicialmente, modelos menos interpretáveis e mais flexíveis, como por exemplo, *support vector machine* (suporte vetorial), *random forest* (floresta aletória) e *bosting*. Para muitos problemas, estes modelos irão apresentar resultados mais exatos.
2. Posteriormente utilize as regressões penalizadas (*lasso* e *ridge*)
3. Se os modelos forem equivalentes, utilize o modelo mais simples e que tenha a performance próxima a dos modelos mais complexos (regressão linear).

2.2.1 Modelo de Regressão Linear Múltipla

Tendo p variáveis preditivas $X_1, ..., X_p$ e uma variável resposta Y, o **modelo de regressão linear múltipla** é dado por

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + e \tag{2.2}$$

em que $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$ denota os estimadores do modelo, o coeficiente β_0 é chamado **intercepto** e a variável preditora associada a ele é, X_0 e tem valor constante igual a 1.

O modelo (2.2) tem $p+1$ parâmetros desconhecidos, a saber, β_0, β_1, ..., β_p, que são estimados através dos dados observados. O termo do erro (e) é um valor aleatório que pode variar entre diferentes observações dos dados. Em outras palavras, o erro não é um valor constante, mas sim um número que varia e se assume que segue uma distribuição normal com média zero e variância constante.

Definido $X_0 = 1$, podemos escrever para cada elemento da amostra,

$$y_i = \sum_{j=0}^{p} \beta_j x_{ij} + e_i, \qquad i = 1, \ldots, n; j = 1, \ldots, p. \tag{2.3}$$

A (2.3) pode ser escrita também na forma matricial:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & ... & x_{np} \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} \tag{2.4}$$

A forma de obter os parámetros é minimizando a soma dos quadrados dos erros e_i, i.e.,

$$\sum_{j=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - \sum_{j=0}^{p} \hat{\beta}_j x_{ij} \right]^2, \tag{2.5}$$

Assim, obtemos os **estimadores por quadrados mínimos** (EMQ) de $\widehat{\beta}_j, j = 0, ..., p$, de modo que

$$\widehat{y_i} = \sum_{j=0}^{p} \hat{\beta}_j x_{ij}, \qquad i = 1, ..., n \tag{2.6}$$

são os **valores estimados** (sob o modelo). Os termos

$$\widehat{e_i} = y_i - \widehat{y_i}, \qquad i = 1, ..., n \tag{2.7}$$

são os **resíduos**, cuja análise é fundamental para avaliar se o modelo da forma (2.2)

se ajusta bem aos dados.

Para obter os estimadores de quadrados mínimos e facilitar a notação vamos reescrever e_i^2 por $e^{'}e$, em que o apóstrofe significa que a matriz foi transposta.

Substituindo o erro e por $y - Xb$, temos:

$$\begin{aligned} e^{'}e &= (y - Xb)^T(y - Xb) \\ &= y^Ty - y^TXb - (Xb)^Ty - (Xb)^TXb \\ &= y^Ty - (Xb)^Ty - (Xb)^Ty - (Xb)^2 \\ &= y^2 - 2(Xb)^Ty + (Xb)^2. \end{aligned} \tag{2.8}$$

Logo vamos sustitiur $e^{'}e = S(b)$, para assim conseguir a minimização derivando em relação ao b e igualando a zero, a equação (2.8).

$$\begin{aligned} \frac{\partial S}{\partial b} &= -2X^Ty + 2X^TXb = 0 \\ X^Ty &= X^TXb \\ b &= \frac{X^Ty}{X^TX} \\ b &= (X^TX)^{-1}(X^Ty). \end{aligned} \tag{2.9}$$

Para ilustrar o procedimento apresentamos o seguinte exemplo[1]. Um modelo de regressão linear múltipla em que a variável independente é "Valor Total do Serviço de Táxi na Colômbia COP (*Pesos Colombianos*)"e as variáveis preditivas são: distância percorrida em metros, X_1, e tempo do viagem em segundos, X_2.

Tabela 2.1: Serviço de Táxi

i	Y	X_1	X_2
1	11.120	8.000	600
2	6.968	3.260	399
3	9.790	6.340	633

Aplicando a expressão acima, temos:

[1]Dados coletados pelo autor.

$$
\begin{aligned}
b &= \left(\begin{pmatrix} 1 & 1 & 1 \\ 8000 & 3260 & 6340 \\ 600 & 399 & 633 \end{pmatrix} * \begin{pmatrix} 1 & 8000 & 600 \\ 1 & 3260 & 399 \\ 1 & 6340 & 633 \end{pmatrix} \right)^{-1} * \left(\begin{pmatrix} 1 & 1 & 1 \\ 8000 & 3260 & 6340 \\ 600 & 399 & 633 \end{pmatrix} * \begin{pmatrix} 11120 \\ 6968 \\ 9770 \end{pmatrix} \right) \\
b &= \begin{pmatrix} 3 & 17600 & 1632 \\ 17600 & 114823200 & 10113960 \\ 1632 & 10113960 & 919890 \end{pmatrix}^{-1} * \begin{pmatrix} 27858 \\ 173617480 \\ 15636642 \end{pmatrix} \\
b &= \begin{pmatrix} 3799.4809 \\ 0.8250 \\ 1.2000 \end{pmatrix}
\end{aligned}
\tag{2.10}
$$

Assim, o modelo ajustado é dado por

Valor Taxi COP $= 3.799,4809 + 0,8250 \times$ metros $+ 1,200 \times$ segundos

2.2.2 Técnicas de Regularização

Nesta subseção apresentamos técnicas que nos ajudam a selecionar o modelo linear mais adequado para prever valores, de acordo com algum critério.

As técnicas de regularização servem para evitar *overfitting* isto acontece quando um modelo tem pouco vício e se ajusta muito bem aos dados de treinamento, por tanto vai gerar uma alta variância do modelo quando ajustado a outros conjuntos de dados. Particularmente em regressão, o termo regularização é uma compensação entre vício e variância, isso quer dizer, que ao modelo será adicionado vício para diminuir sua alta variância (veja Figura 2.3 a seguir), evitando assim o *overfitting*. Esta estratégia consiste em ajustar diferentes modelos, a um conjunto de treinamento e escolher aquele que gere as melhores estimações com dados de um conjunto teste. Em geral, esse processo é concretizado por meio de validação cruzada 2.1.3.

Na Figura 2.3 começando da esquerda para a direita, ilustramos situações de baixo erro em dados de treinamento e um elevado erro em dados de teste, isto indica uma variância alta. Depois, temos um elevado erro em dados de treinamento e erro parecido em dados de teste indicam presença de um vício alto. Finalmente, baixo erro em dados de treinamento e baixo erro em dados de teste indicando variância e vício baixos.

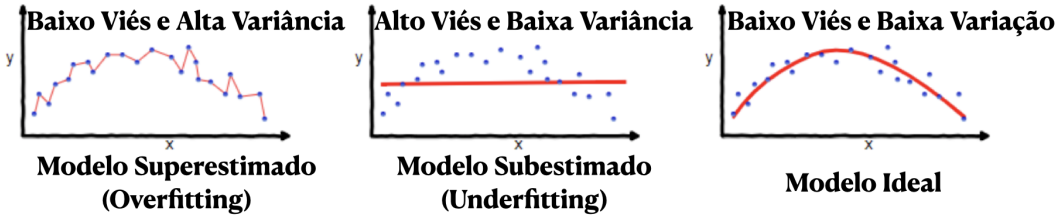

Figura 2.3: Diferentes tipos de viés e variâncias que o modelo pode apresentar

A seguir temos mais um exemplo de *overfitting*. Na Figura 2.4 consideramos um exemplo cujo objetivo é ajustar um modelo (cor azul) de regressão polinomial de

grau, 1, 2, 4, 8, que relacione os pontos de cor vermelha com a curva verde correspondente à função $Y = \frac{1}{1+35X^2}$.

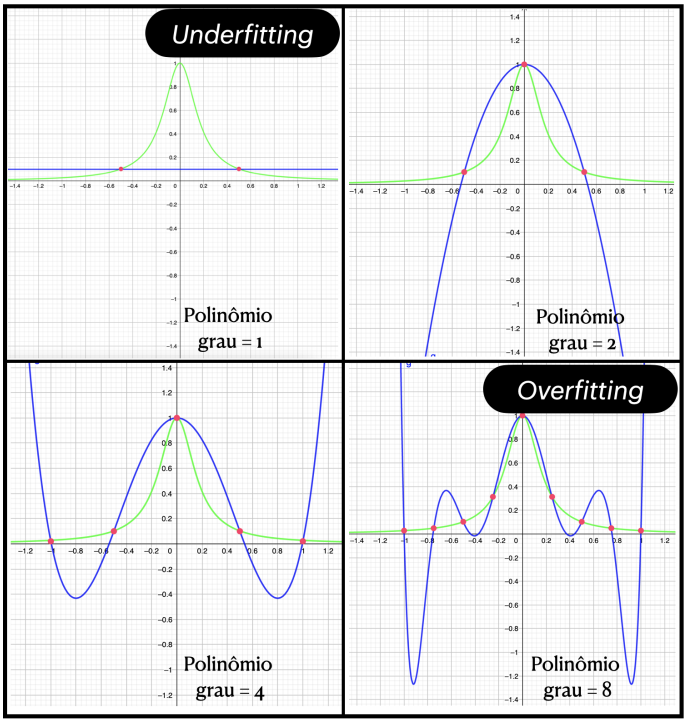

Figura 2.4: Ajuste de modelos polinomias a um conjunto de dados

O modelo polinomial de grau 8 tem um ajuste perfeito, mas não reproduz a curva que gerou os dados. Esse fenômeno é conhecido como *overfitting*.

Há varias técnicas de regularização que evitam o *overfitting*, mas aqui consideramos apenas duas: a regularização L_2, ou ***Ridge*** e a regularização L_1, ou ***Lasso*** que serão descritas a seguir.

Regularização L_2 (Ridge) foi introduzida por Hoerl e Kennard (1970) [11] para tratar o problema de multicolinearidade, mas também pode ser usada para corrigir problemas ligados ao *overfitting*. Considere um modelo linear como descrito pela equação (2.2).No momento de minimizar a soma dos quadrados dos erros, podemos ter o problema de variância alta, isto ocorre quando o modelo ajusta-se demasiadamente bem aos dados de treinamento, mas não aos dados de teste. Neste momento que entra a regularização Ridge (L_2), com os **estimadores de quadrados mínimos penalizados** que correspondem às soluções de

$$\hat{\beta}_{Ridge}(\lambda) = \arg\min_{\beta} \left[\sum_{i=1}^{n} ([y_i - \sum_{j=0}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right]. \tag{2.11}$$

A penalização depende do valor que atribuímos a λ. Se λ cresce implica uma maior penalização dos parâmetros β. Se $\lambda = \infty$, não há variáveis a serem incluídas no modelo e se $\lambda = 0$, obtemos os estimadores de quadrados mínimos que já conhcemos. A escolha de λ dever ser um dos componente da estratégia para a determinação de

estimadores regularizados, pois o objetivo é penalizar parâmetros associados aos preditores. Dizemos que $\hat{\beta}_{Ridge}(\lambda)$ é o estimador Ridge. Pode-se mostrar que

$$\hat{\beta}_{Ridge}(\lambda) = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \tag{2.12}$$

Em que $\mathbf{X} = (x_1^T, ..., x_n^T)^T$ é a matriz de preditores padronizados e $\mathbf{y} = (y_1, ..., y_n)^T$ é o vetor de respostas. Algumas propiedades dessa clase de estimadores são:

1. Em geral, o estimador *Ridge* não é consistente. Sua consistência assintótica vale quando $\lambda = v\lambda_n \to \infty$, $\lambda_n/n \to 0$ e $p < n$.
2. O estimador *Ridge* é viesado para os parâmetros não nulos.
3. A escolha do coeficiente de regularização λ pode ser feita via validação cruzada.

Regularização L_1 (Lasso) foi introduzida por Tibshirani (1996) [12] é um método de análise de regressão que realiza seleção e regularização de variáveis para melhorar a precisão e interpretabilidade do modelo. Consideremos, agora, o estimador Lasso, que é obtido de

$$\hat{\beta}_{Lasso}(\lambda) = \arg\min_{\beta} \left[\frac{1}{n} \sum_{i=1}^{n} ([y_i - \sum_{j=0}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right] \tag{2.13}$$

Algumas propiedades dos estimadores lasso são:

1. Seleciona zero para os parâmetros que correspondem a preditores redundantes.
2. é viesado para os parâmetros nulos.
3. Sob certas condições, o estimador *Lasso* identifica as variáveis não relevantes do modelo atribuindo pesos nulos aos respectivos coeficientes.

é evidente que Lasso tem uma vantagem sobre Ridge, porque não inclui todos os preditores no modelo final (elimina preditores insignificates, desnecessários ou redundantes) e produz modelos mais simples gerando melhor interpretabilidade. Lasso tem melhor desempenho quando o modelo tem menos preditores, enquanto Ridge é melhor quando o modelo tem muitos preditores.

Para outras propiedades, veja Medeiros [13] e Bühlmann and Van de Geer [14]

2.2.3 Árvores

Antes de introducir formalmente o modelo, vamos mostrar um exemplo simples de árvore de decisão para que tenhamos uma ideia de como o modelo funciona.

Em algum momento da sua vida, você gostaria de fazer um passeio de bicicleta. Vamos supor que há dois critérios para sair de bicicleta, o primeiro deles é o clima, ou seja, se está chovendo ou não, e o segundo critério é se a bicicleta está danificada, com um pneu furado ou tem algum defeito mecânico. Com esses dois critérios vamos obter uma árvore para decidir se passeamos de bicicleta ou ficamos em casa.

Na parte esquerda da figura 2.5 há quatro regiões delimitadas por eixos horizontal e vertical. As quatro regiões correspondem a quatro situações: "Bicicleta não dani-

ficada e não está chovendo", "Bicicleta não danificada e está chovendo", "Bicicleta danificada e não está chovendo"e "Bicicleta danificada e está chovendo". Há pontos pretos ou cinzas nas quatro regiões. Devemos escolher a região onde encontram-se os pontos cinzas. Observando o gráfico, dizemos que estes estão no quadro inferior esquerdo, concluindo que a pessoa sai para um passeio de bicicleta se esta não estiver danificada e se não estiver chovendo.

Na parte direita da figura 2.5 temos a árvore de decisão. Começamos com o parâmetro "bicicleta danificada", mas também podemos começar com o parâmetro "chuva"; a partir do primeiro parâmetro temos duas situações, a bicicleta está danificada e ficamos em casa e o algoritmo acabou, ou se a bicicleta não estiver danificada passamos para o próximo parâmetro, isto é, se estiver chovendo ficamos em casa, mas se não estiver chovendo passeamos de bicicleta.

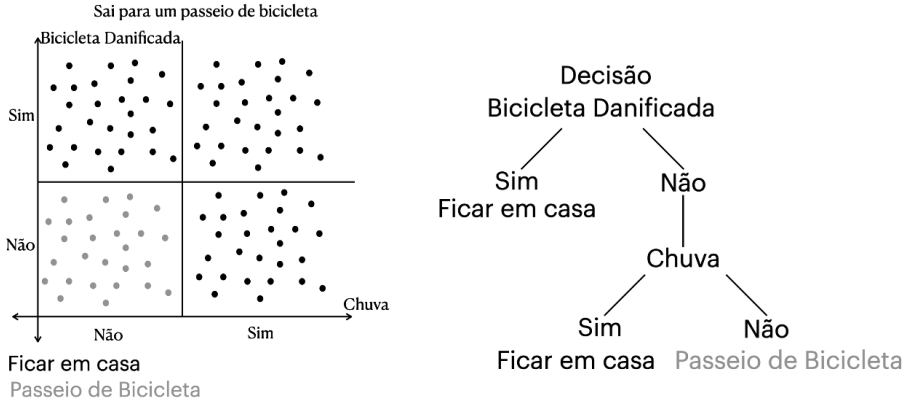

Figura 2.5: Árvore de decisão sobre o passeio de bicicleta

Agora vamos mostrar outro exemplo para explicar matematicamente a construção de uma árvore de decisão.

Abaixo segue uma tabela de quatro colunas e seis linhas composta por três variáveis independentes: preço do ingresso (caro ou econômico); local do show (na mesma cidade ou não); artista (internacional ou nacional) além de uma variável dependente (decisão de ir ou não ao concerto).

Tabela 2.2: Assirtir ao Concerto

Preço	**Lugar**	**Artista**	**Decisão**
Caro	Outra Cidade	Internacional	Sim
Econômico	Mesma Cidade	Nacional	Não
Econômico	Outra Cidade	Internacional	Sim
Caro	Outra Cidade	Nacional	Não
Caro	Mesma Cidade	Internacional	Sim
Econômico	Outra Cidade	Nacional	Não

Primeiramente vamos calcular a entropia da variável Decisão, para isso temos que contar quantas vezes a decisão foi favorável (sim) e desfavorável (não) na Tabela 2.2. Em seguida, vamos calcular a probabilidade de assirtir ou não ao concerto.

$$P(\text{Sim}) = \frac{3}{6} = \frac{1}{2}$$
$$P(\text{Não}) = \frac{3}{6} = \frac{1}{2}$$

$$\text{Entropia}_D = -\left(\frac{1}{2} \cdot log_2\frac{1}{2} + \frac{1}{2} \cdot log_2\frac{1}{2}\right) = 1$$

$$\begin{aligned} \text{Ganho de Informação (GI)} &= Entropia_{Pai} - \sum Peso_{Filho} \cdot Entropia_{Filho} \\ \text{Peso} &= \frac{\text{Número de unidades amostrais do Filho}}{\text{Número de unidades amostrais do Pai}} \\ \text{Entropia} &= -\sum_{i=1} p_i log_2 p_i \end{aligned} \tag{2.14}$$

Um exemplo para explicar o peso é: Na primeira coluna da Tabela 2.2 chamada preço, o dado "caro"é repetido 3 vezes e sabemos que o número total de dados é 6, portanto, o peso quando o **preço é caro** é $3/6 = 1/2$

Para entender o que significa pai e filho, vejamos a variável preço em relação à decisão. Quando o preço é caro temos duas decisões sim e não e quando o preço é econômico temos uma decisão sim e duas decisões não; então pai é a variável objetivo que, para nosso exemplo, é Decisão e tem 2 filhos, a saber, preço caro e preço econômico. Desse modo temos que calcular a entropia e os pesos de cada um dos filhos.

Entropia e peso quando o preço é caro

$$P(\text{Sim}) = \frac{2}{3}$$
$$P(\text{Não}) = \frac{1}{3}$$
$$E = -\left(\frac{2}{3} \cdot log_2\frac{2}{3} + \frac{1}{3} \cdot log_2\frac{1}{3}\right) = 0.92$$
$$Peso = \frac{1}{2}$$

Entropia e peso quando o preço é econômico

$$P(\text{Sim}) = \frac{1}{3}$$
$$P(\text{Não}) = \frac{2}{3}$$
$$E = -\left(\frac{1}{3} \cdot log_2\frac{1}{3} + \frac{2}{3} \cdot log_2\frac{2}{3}\right) = 0.92$$

$$Peso = \frac{1}{2}$$

Ganho de Informação $= 1 - \left(\frac{1}{2} \cdot 0.92 + \frac{1}{2} \cdot 0.92\right)$
GI $= 0.08$

A árvore de decisão começa com a variável que tem o maior valor no ganho de informação (GI), agora vamos calcular o GI para a variável Local e Artista.

Entropia e peso quando o local é em outra cidade

$$P(\text{Sim}) = \frac{2}{4} = \frac{1}{2}$$

$$P(\text{Não}) = \frac{2}{4} = \frac{1}{2}$$

$$E = -\left(\frac{1}{2} \cdot log_2\frac{1}{2} + \frac{1}{2} \cdot log_2\frac{1}{2}\right) = 1$$

$$Peso = \frac{2}{3}$$

Entropia e peso quando o local é na mesma cidade

$$P(\text{Sim}) = \frac{1}{2}$$

$$P(\text{Não}) = \frac{1}{2}$$

$$E = -\left(\frac{1}{2} \cdot log_2\frac{1}{2} + \frac{1}{2} \cdot log_2\frac{1}{2}\right) = 1$$

$$Peso = \frac{1}{3}$$

Ganho de Informação $= 1 - \left(\frac{2}{3} \cdot 1 + \frac{1}{3} \cdot 1\right)$
GI $= 0$

Nesse caso, não houve ganho de informação para a variável Local.

Entropia e peso quando o artista é internacional

$$P(\text{Sim}) = \frac{3}{3} = 1$$

$$P(\text{Não}) = \frac{0}{3} = 0$$

$$E = -\left(1 \cdot log_2 1 + 0 \cdot log_2 0\right) = 0$$

$$Peso = \frac{1}{2}$$

Entropia e peso quando o artista é nacional

$$P(\text{Sim}) = \frac{0}{3} = 0$$

$$P(\text{Não}) = \frac{3}{3} = 1$$

$$E = -\left(0 \cdot log_2 1 + 1 \cdot log_2 1\right) = 0$$

$$Peso = \frac{1}{2}$$

Ganho de Informação $= 1 - \left(\frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 0\right)$
GI $= 1$

Como a variável Artista tem ganho de informação igual a 1, concluimos duas coisas importantes; primeiro, a variável Artista é a raiz da árvore e esta terá apenas dois ramos, pois o maior valor que pode ser obtido no ganho de informação é 1. Sendo assim, a árvore de decisão para os dados da Tabela 2.2 é

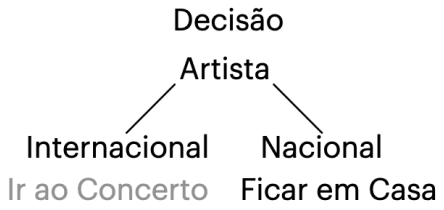

Figura 2.6: Árvore de decisão sobre ida ao concerto

As árvores representam uma alternativa interessante para a construção de modelos preditivos, quando a relação entre os preditores e a resposta de interesse é não linear e complexa. Um algoritmo de árvore de decisão pode ser escrito da seguinte manera:

1. Escolhe-se o conjunto super entrópico (**Entropia**: Medida que nos diz o quanto nosso dados estão desorganizados e misturados. Quanto menor a entropia, maior o ganho de informação e vice-versa).
2. Selecionado o conjunto entrópico, as árvores de decisão são contruídas por nós, levando de novo em consideração a entropia, ou seja, aumentar o ganho de informação. Normalmente são gerados dois nós (divisão binária).
3. Comparando os resultados, nossa ramificação estará finalizada escolhendo a de maior ganho e, este processo será repetido recursivamente para cada lado da ramificação, parando quando o ganho de informação for igual a 0.

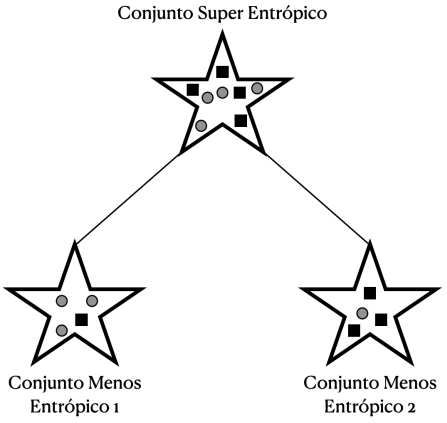

Figura 2.7: Divisão do conjunto entrópico

Em geral, os modelos preditivos resultantes de árvores de decisão são instáveis, ou seja, pequenas alterações nos dados de treinamento podem implicar mudanças na estrutura da árvore ou em suas regras e, portanto, altera a interpretação do modelo ajustado. Nesse cenário, diversos métodos foram desemvolvidos a fim de melhorar o desempenho preditivo de árvores de decisão simples por meio da construção de múltiplas árvores, em vez de uma árvore única, a fim de combinar suas predições em uma predição final mais acurada.(James, 2014 [6]).

2.2.4 Floresta Aleatória

A dinâmica deste modelo é a seguinte: cria várias árvores de decisão e as combina para obter uma predição com maior acurácia e mais estável. Desse modo, floresta aleatória inicia-se com a obtenção de um conjunto de dados de tamanho n por amostragem com reposição do conjunto de treinamento e, posteriormente, para cada conjunto, estima uma árvore de decisão (Hastie; Tibshirani; Friedman,2008[4]; James,2014[6]).

A diferença principal entre floresta aleatória e árvores de decisão é o sobreajuste (*overfitting*), já que o algoritmo de floresta aleatória evita o sobreajuste na maioria dos casos, pois trabalha com subconjuntos aleatórios das características e constrói árvores a partir de tais subconjuntos. Esta abordagem torna o rendimento computacional mais lento, e isto depende de quantas árvores de decisão serão construídas.

Para entender melhor o algoritmo da floresta aleatória, a Figura 2.8 mostra graficamente o proceso para obter o modelo. Uma vez que o conjunto de treinamento foi escolhido, construa uma árvore de decisão associada a esse conjunto de dados. Depois escolha m árvores que deseja construir, Em geral, uma quantidade elevada de árvores aumenta a performance e torna as predições mais estáveis, mas também torna a computação mais lenta.

Para finalizar com a descrição deste algoritmo, podemos dizer que o modelo de regressão floresta aleatória é poderoso e preciso. Ele, geralmente, tem um ótimo desempenho em muitos problemas, incluindo recursos com relacionamentos não li-

neares. As desvantagens, no entanto, incluem o seguinte: não há interpretabilidade, sobreajuste pode ocorrer facilmente, e devemos escolher o número de árvores a incluir no modelo.

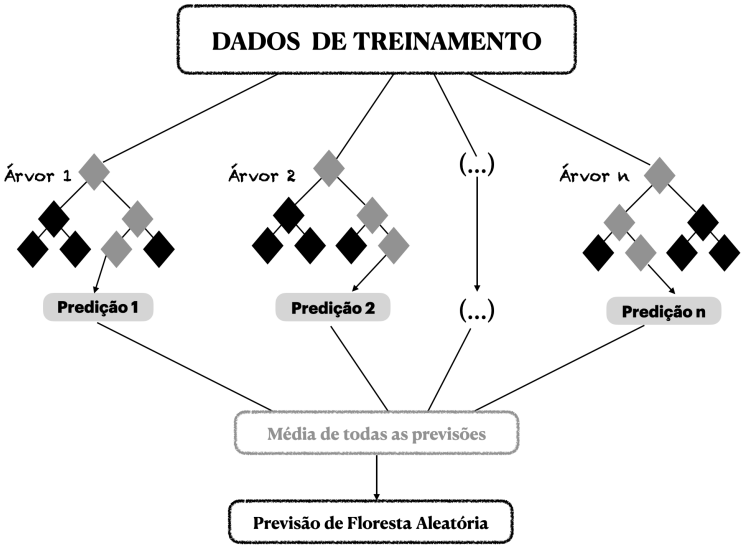

Figura 2.8: Técnica do Modelo Floresta Aletória

2.2.5 Regressão de Vetores de Suporte

A regressão de vetores de suporte (RVS) foi introduzida pela primeira vez por Smola e Schölkopt em 2004 [8]. A ideia desse algoritmo é encontrar um hiperplano ótimo para regressões lineares e não lineares. O RVS funciona como o princípio da máquina de vetores de suporte (MVS), este difere no sentido en que MVS é um classificador usado para prever variáveis categóricas ou qualitativa, dada uma certa margem $\epsilon > 0$. RVS é um regressor usado para prever variáveis ordinárias contínuas através de um hiperplano, tal que represente o número máximo de pontos do conjunto de dados de treinamento, e este submetido às restrições definidas pelo usuário.

Por razões pedagógicas, começamos com funções lineares $f(x) = \beta_1 x + \beta_0$, em que nosso objetivo é garantir a planície, o que significa buscar um pequeno β_1, isso pode ser feito minimizando a norma, ou seja $\|\beta_1\|^2 = < \beta_1, \beta_1.x >$. Podemos escrever isso como um problema de otimização convexa:

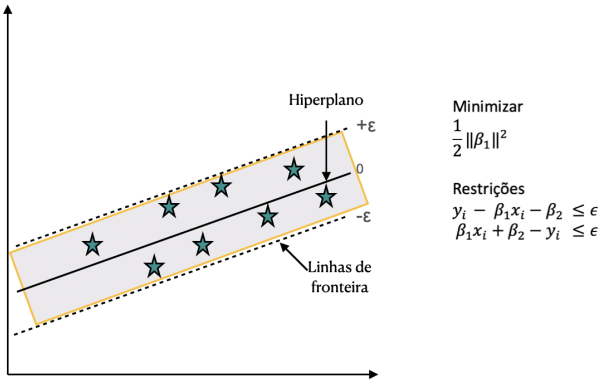

Figura 2.9: Regressão de Vetores de Suporte

Quando temos dados atípicos, como da Figura 2.9 adicionamos variáveis de folga e uma constante $C > 0$ a qual determina o *trade-off* (equilíbrio alcançado) entre a uniformidade de f e o valor até o qual desvios de magnitude maior que ϵ são tolerados.

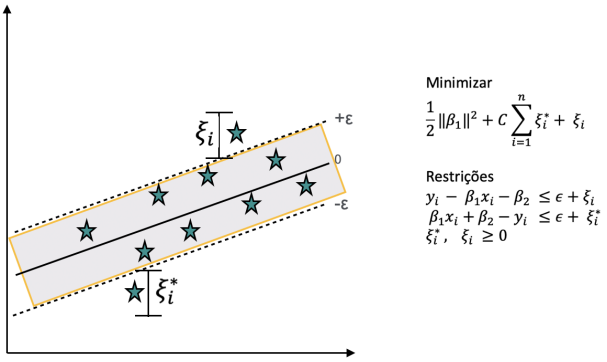

Figura 2.10: Regressão de Vetores de Suporte

As variáveis de folga devem ser mantidas no mínimo, o que implica que deve haver uma função que penaliza valores diferentes de zero.

Temos un exemplo quando o problema não é linear, uma função chamada kernel transforma os dados em um recurso de espaço dimensional mais alto para possibilitar a separação linear:

$$y = \sum_{i=1}^{N}([\beta_{1i} - \beta_{1i}^{*}). < \psi(x_i), \psi(x) > +\beta_0 \tag{2.15}$$

$$y = \sum_{i=1}^{N}[\beta_{1i} - \beta_{1i}^{*}).K(x_i, x) + \beta_0 \tag{2.16}$$

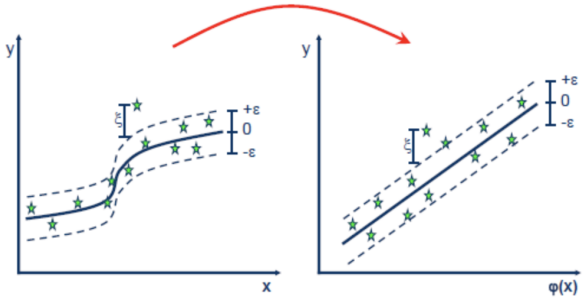

Figura 2.11: http://www.saedsayad.com/support_vector_machine_reg.htm

Capítulo 3

Metodologia e Análise Descritiva

3.1 Metodologia

Procedimentos Iniciais

1. Banco de dados: Os dados são referentes a SABESPREV *(Data set Original)*.
2. Limpeza dos dados: Ajustamos os dados em um *data frame* e definimos como 0 (zero) os dados que estão como Nan (*Not is a number*).
3. Escolha das Variáveis: Consideramos a variável de gasto anual do indivíduo como dependente e as variáveis idade, gasto anual do ano anterior e gasto anual do ano anteanterior como independentes.
4. Selecão dos dados: Selecionamos indivíduos com 25 anos completos em janeiro de 2007 e com no máximo 65 concluídos em dezembro de 2009.
5. Análise descritiva: Foi feita uma anáise descritiva, isto é, calculamos as medidas de posição e dispersão (Percentis - Mínimo - Máximo - Média - Desvio Padrão) dos gastos anuais para os indivíduos selecionados entre 2005 e 2009 em relação à variável expensetot (despesas com saúde). Também foram feitos gráficos de perfis de médias para cada ano, sexo e faixa etária.

Procedimentos de Ajuste do Modelo

1. Divisão dos dados: De acordo com a Seção 2.1.1 a divisão que vamos fazer é de 80% para treinamento e 20% para teste.
2. Ajuste do Modelo: Estimamos os modelos de regressão: Linear Múltipa, Suporte Vetorial, Árvores, Floresta Aleatória, Ridge e Lasso Bayesiano; levando em conta validação cruzada (*Cross Validation* 2.1.3).
3. Desempenho do modelo: Avaliamos a precisão de cada modelo segundo RMSE (*Root Mean Squared Error*).
4. Gráfico dos Modelos: Fizemos um gráfico para observar o comportamento e a diferença de 50 indivíduos dos dados de teste contra a previsão de 50 indivíduos feita pelos modelos de regressão mencionados anteriormente.
5. Escolha do modelo: Escolhemos o melhor modelo considerando a métrica de avaliação RMSE.

Procedimentos para Simulação da CP

1. Escolha das vidas iniciais: Escolhemos as pessoas que tem entre 25 e 30 anos do modelo ganador tanto para as mulherers como para homens.
2. Simulação: As vidas que estão na faixa etária de 25 a 30 anos, serão replicadas aleatoriamente com repetição até alcançar um total de 10.000 pessoas. Esse processo será realizado em 100 vezes, e a simulação final será obtida calculando a média dos resultados obtidos em todas as iterações.
3. Estimação: Com os valorers médios estima-se o gasto em saúde nos 41 anos de vida laboral.
4. Visualização Tabular: Para analisar as estadisticas descritivas do balance da CP, gastos cobertos pela CP e pelo seguro catastrófico.
5. Visualização Gráfica: Histograma: Para observar a severidade e frequência do seguro catastrófico. Dispersão: Entre os gastos totais no período de 41 anos de cada vida e o percentagem dessos gastos que foram cobertos pelo seguro catastrófico.

3.2 Análise Descritiva

O banco de dados (SABESPREV) é formado por, aproximadamente, 70.000 vidas e os registros referem-se ao periodo de 2005 a 2009. Os valores de todas as despesas com saúde foram corrigidos pelo índice de inflação IPCA (índice Nacional de Preços ao Consumidor Amplo) para valores de dezembro de 2009.

Consideraremos neste estudo apenas indivíduos em idade laboral, ou seja, aqueles com 25 anos completos em janeiro de 2005 e com no máximo 65 anos concluídos em dezembro de 2009. Além disso, serão considerados apenas indivíduos que foram acompanhados durante os cinco anos, de 2005 a 2009. Nessas condições foram consideradas na análise 27.780 pessoas.

A Tabela 3.1 apresenta algumas estatísticas descritivas das despesas anuais com saúde dessas 27.780 pessoas. A porcentagem de indivíduos com gastos nulos anuais está entre 5% e 6% em todos os anos. Analisando essa tabela, podemos dizer que em 2005, cerca de 5% das pessoas eram responsáveis por gastos anuais que variaram de R$ 7.000 a R$ 426.772, sendo que a despesa máxima foi 206 vezes a média. Em 2006, metade das pessoas tinham um gasto menor do que R$ $736,26$ e a despesa máxima foi 258 vezes a média. Considerando agora o ano 2007, 10% das vidas tinham um gasto de saúde superior R$ $4.130,22$. Em geral, grandes gastos estão concentrados em uma pequena parcela de pessoas físicas, por exemplo, em 2008 e 2009 menos de 1% dos indivíduos eram responsáveis por despesas que variaram de R$ $31.158,38$ a R$ $1.044.525,30$ e R$ $33.729,95$ a R$ $965.286,66$, respectivamente. Ainda, observamos que 385 vezes a média é equivalente à despesa máxima em 2008, e em 2009, temos que 343 vezes a média é equivalente à despesa máxima.

Tabela 3.1: Estatística descritiva dos gastos

	2005	2006	2007	2008	2009
% Sem despesas	5,97	5,52	5,77	5,55	5,76
Percentil(25)	319,19	303,20	331,85	367,63	359,43
Percentil(50)	763,59	736,26	790,73	903,99	871,60
Percentil(75)	1.746,19	1.665,50	1.789,25	2.090,46	2.008,25
Percentil(90)	3.993,20	3.735,18	4.130,22	4.860,63	4.760,99
Percentil(95)	7.083,06	6.531,47	7.477,29	8.726,47	9.113,72
Percentil(96)	8.496,86	7.960,25	8.863,62	10.246,59	10.779,97
Percentil(97)	10.449,89	9.853,32	10.935,47	12.909,75	14.190,37
Percentil(98)	13.537,60	13.133,72	15.379,85	18.237,42	20.160,33
Percentil(99)	22.910,00	21.682,97	25.850,55	31.158,38	33.729,95
Percentil(99,5)	33.904,68	35.694,56	43.746,81	54.407,36	55.601,09
Percentil(99,9)	92.961,88	105.584,84	92.548,68	139.810,85	181.947,41
Máximo	426.772,21	528.292,65	355.219,72	1.044.525,30	965.286,66
Média	2.063,58	2.041,02	2.219,96	2.709,81	2.813,19
Desvio Padrão	6.833,05	8.127,08	7.536,06	11.611,99	12.672,69

Nas Figuras 3.1, 3.2, 3.3 temos os gráficos de perfis para os gastos médios anuais de 2005, 2006, 2007, 2008 e 2009 com um total de 13.539 mulheres e 14.241 homens em cada gráfico. Podemos verificar que as pessoas do sexo feminimo têm um gasto médio maior do que as do sexo masculino até os 55 anos, mas a diferença diminui a ponto de em 2005 e 2007, 2006 e 2008 nas faixa etária de “56 a 60” e “61 a 64” respectivamente, o gasto médio dos homens é maior; além disso em 2009 o gasto é quase o mesmo tanto paras as mulheres como para os homens.

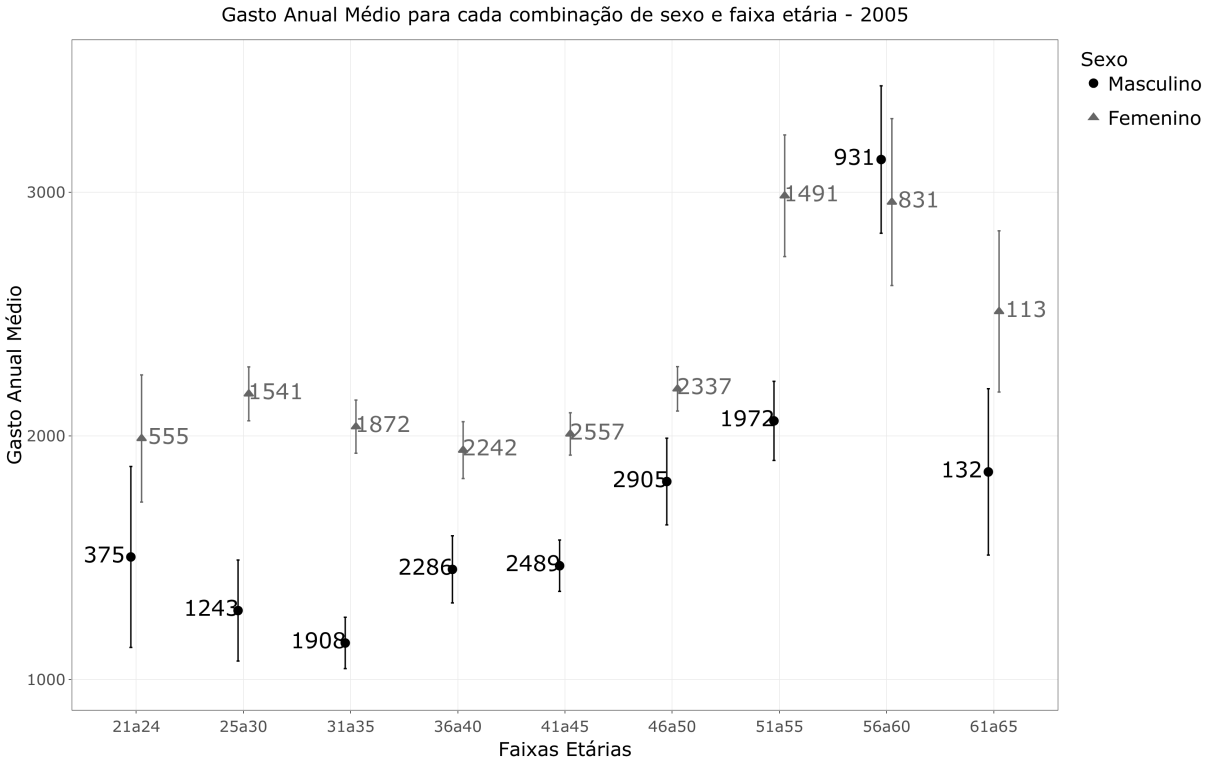

Figura 3.1: Gráfico de Perfis para os gastos de 2005

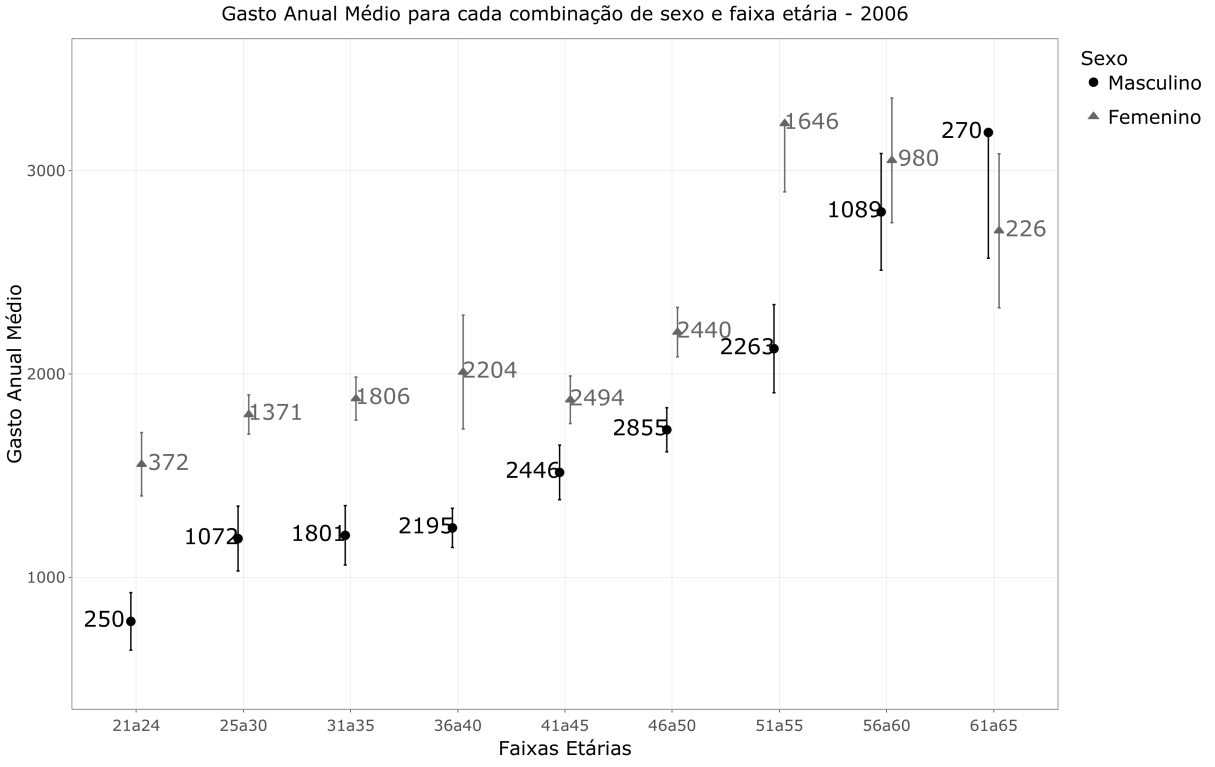

Figura 3.2: Gráfico de Perfis para os gastos de 2006

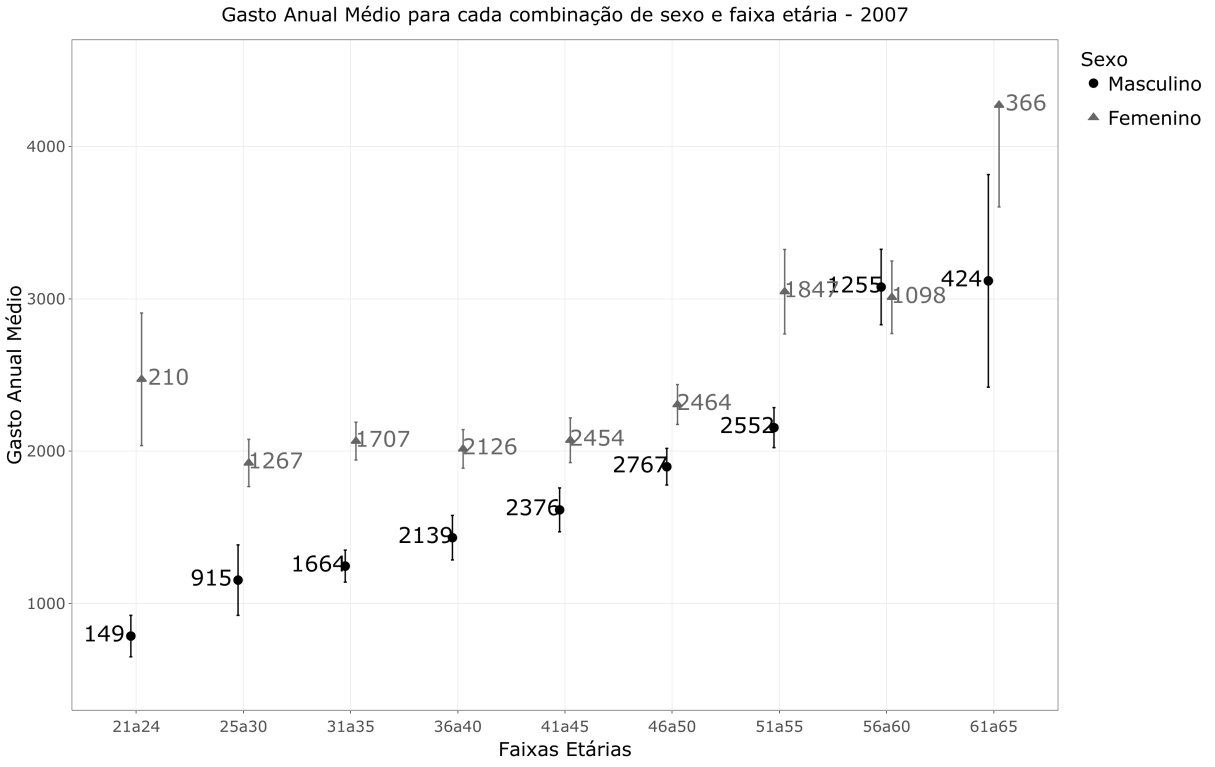

Figura 3.3: Gráfico de Perfis para os gastos de 2007

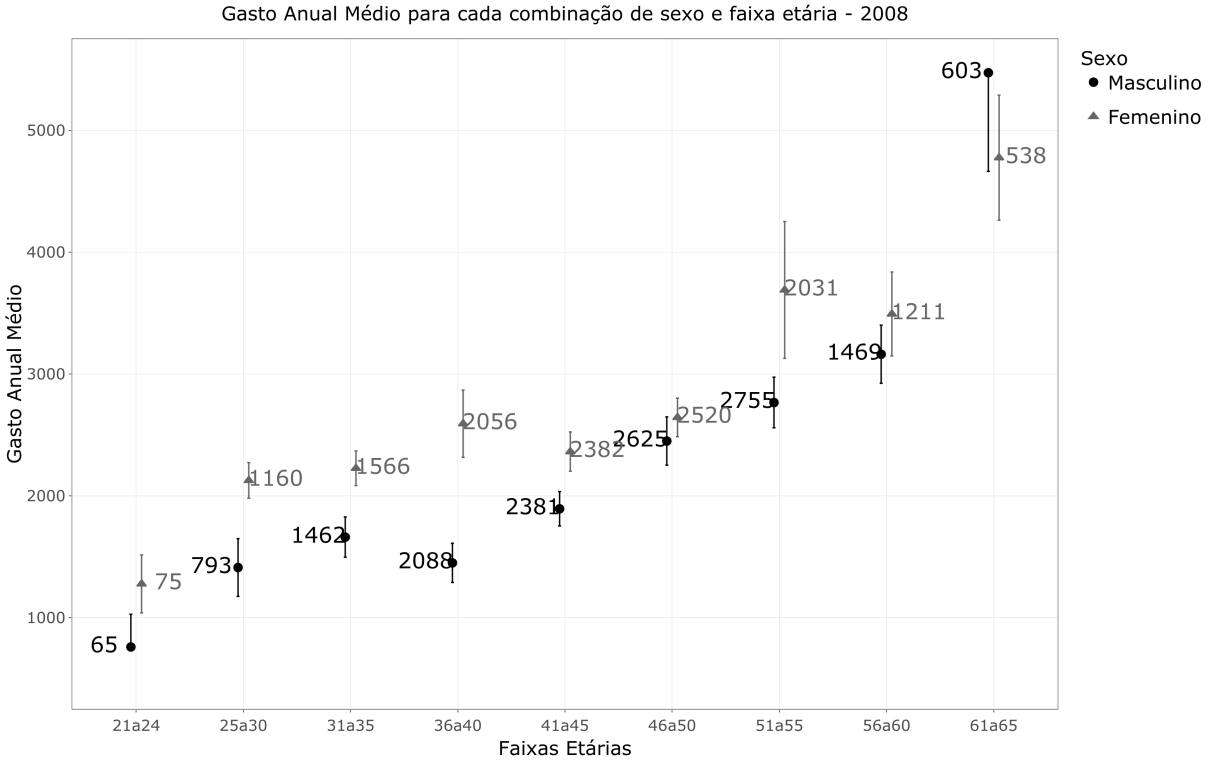

Figura 3.4: Gráfico de Perfis para os gastos de 2008

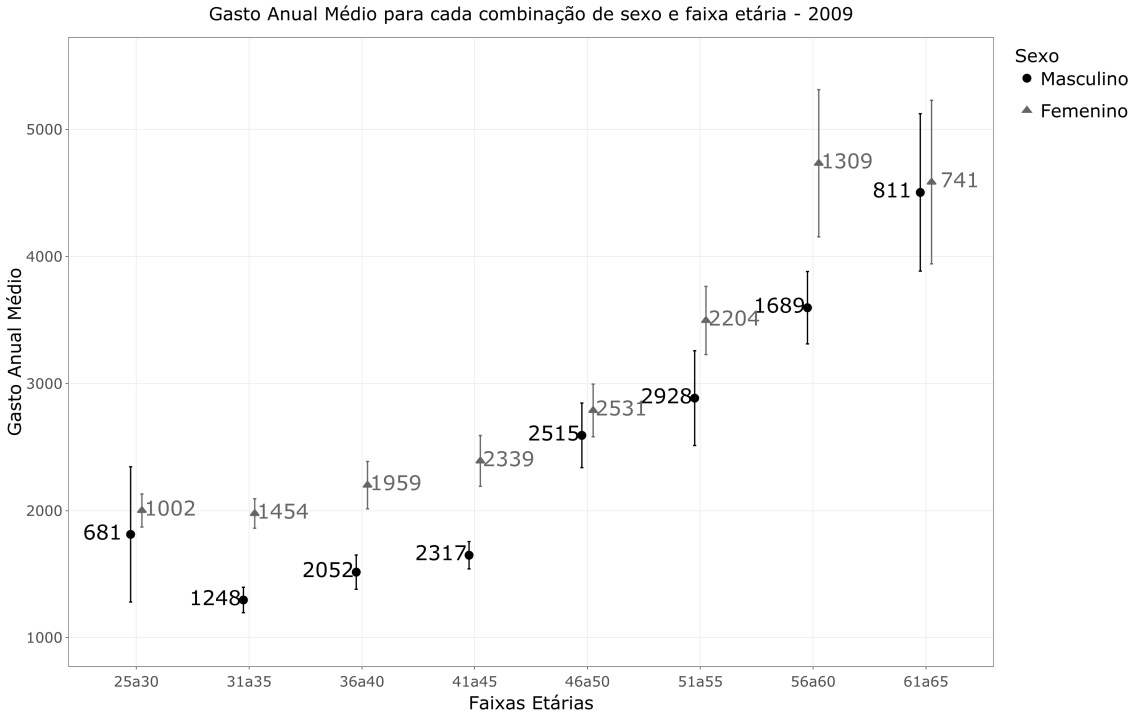

Figura 3.5: Gráfico de Perfis para os gastos de 2009

Capítulo 4

Modelagem e Simulação

Neste capítulo, mostramos os ajustes de diferentes modelos para estimar os gastos e os comparamos a fim de escolher o mais adequado. Posteriormente comparamos os resultados encontrados para a conta poupança que usa como modelo de previsão de gastos Cadeias de Markov e as contas poupanças que utilizam regressão.

4.1 Estimação dos Modelos

Os indivíduos foram divididos, aleatoriamente, em **dados de treinamento** e **dados de teste** nas proporções de 80% e 20%, respectivamente. Uma vez feito isso, passamos a validação cruzada (Cross Validation - CV, mencionado na subseção 2.1.3) usando *k-fold* = 5.

Foram ajustados os modelos de regressão linear múltipla, suporte vectorial, árvores, floresta aleatória, *bayesian ridge* e *bayesian lasso*, com a intenção de avaliar o efeito das variáveis idade, gastoprev1 (gasto com saúde no ano anterior) e gastoprev2 (gasto com saúde há dois anos) no gasto com saúde do presente ano.

Na seção a seguir justificaremos a escolha do modelo e mostraremos seu ajuste.

4.2 Comparação dos Modelos de Regressão

Nesta seção, ajustaremos os modelos para cada ano e sexo. A escolha do modelo será baseada no RMSE associado a cada modelo.

4.2.1 Ajuste com dados atípicos

Ano 2007 - Feminino

Modelo *Bayesian Ridge* ajustado:

$$y_i = 73,61 + 25,87\text{*idade} + 0,28\text{*gasto}_{2006} + 0,25\text{*gasto}_{2005} + e_i,\ i = 1, \ldots, 13539. \quad (4.1)$$

Na Tabela 4.1 temos a comparação, quanto ao RMSE, dos modelos de regressão: linear múltipla, suporte vetorial (SVR), árvores, floresta aleatória, *bayesian ridge* e *bayesian lasso*. Observa-se que o menor valor é dado pelo modelo de Bayesian Ridge.

Tabela 4.1: Comparação dos Modelos - 2007 - Feminino

	B. Ridge	Árvores	B. Lasso	Linear	SVR	F. Aleatória
RMSE	7.230, 10	7.230, 30	7.236, 43	7.238, 22	7.253, 27	7.292, 32

Ano 2007 - Masculino

Modelo de regressão linear múltipla ajustado:

$$y_i = -637,89 + 42,46\text{*idade} + 0,18\text{*gasto}_{2006} + 0,17\text{*gasto}_{2005} + e_i,\ i = 1, \ldots, 14241. \quad (4.2)$$

Na Tabela 4.2 temos a comparação, quanto ao RMSE, dos modelos de regressão: linear múltipla, suporte vetorial (SVR), árvores, floresta aleatória, *bayesian ridge* e *bayesian lasso*. Observa-se que o menor valor é dado pelo modelo de regressão linear múltipla.

Tabela 4.2: Comparação dos Modelos - 2007 - Masculino

	Linear	B. Lasso	B. Ridge	SVR	Árvores	F. Aleatória
RMSE	6.691, 99	6.692, 42	6.693, 77	6.890, 52	7.119, 59	7.123, 21

Ano 2008 - Feminino

Modelo de regressão linear múltipla ajustado:

$$y_i = 26,24 + 40,84\text{*idade} + 0,29\text{*gasto}_{2007} + 0,14\text{*gasto}_{2006} + e_i,\ i = 1, \ldots, 13539. \quad (4.3)$$

Na Tabela 4.3 temos a comparação, quanto ao RMSE, dos modelos de regressão: linear múltipla, suporte vetorial (SVR), árvores, floresta aleatória, *bayesian ridge* e *bayesian lasso*. Observa-se que o menor valor é dado pelo modelo de regressão linear múltipla.

Tabela 4.3: Comparação dos Modelos - 2008 - Feminino

	Linear	B. Ridge	B. Lasso	SVR	F. Aleatória	Árvores
RMSE	8.528, 26	8.539, 57	8.551, 60	9.690, 25	10.118, 41	10.503, 81

Ano 2008 - Masculino

Modelo *Bayesian Ridge* ajustado:

$$y_i = -645,81 + 40,63\text{*idade} + 0,45\text{*gasto}_{2007} + 0,16\text{*gasto}_{2006} + e_i,\ i = 1, \ldots, 14241. \tag{4.4}$$

Na Tabela 4.4 temos a comparação, quanto ao RMSE, dos modelos de regressão: linear múltipla, suporte vetorial (SVR), árvores, floresta aleatória, *bayesian ridge* e *bayesian lasso*. Observa-se que o menor valor é dado pelo modelo de regressão *bayesian ridge*.

Tabela 4.4: Comparação dos Modelos - 2008 - Masculino

	B. Ridge	B. Lasso	Linear	SVR	F. Aleatória	Árvores
RMSE	10.103, 54	10.115, 59	10.118, 18	10.657, 55	10.976, 77	11.139, 28

Ano 2009 - Feminino

Modelo de regressão linear múltipla ajustado:

$$y_i = -668,40 + 52,16\text{*idade} + 0,20\text{*gasto}_{2008} + 0,22\text{*gasto}_{2007} + e_i,\ i = 1, \ldots, 13539. \tag{4.5}$$

Na Tabela 4.5 temos a comparação, quanto ao RMSE, dos modelos de regressão: linear múltipla, suporte vetorial (SVR), árvores, floresta aleatória, *bayesian ridge* e *bayesian lasso*. Observa-se que o menor valor é dado pelo modelo de regressão linear múltipla.

Tabela 4.5: Comparação dos Modelos - 2009 - Feminino

	Linear	B. Ridge	B. Lasso	F. Aleatória	Árvores	SVR
RMSE	14.425, 46	14.429, 24	14.430, 10	14.623, 31	14.888, 43	15.105, 82

Ano 2009 - Masculino

Modelo de regressão linear múltipla ajustado:

$$y_i = -1129,00 + 56,81\text{*idade} + 0,26\text{*gasto}_{2008} + 0,14\text{*gasto}_{2007} + e_i,\ i = 1, \ldots, 14241. \tag{4.6}$$

Na Tabela 4.6 temos a comparação, quanto ao RMSE, dos modelos de regressão: linear múltipla, suporte vetorial (SVR), árvores, floresta aleatória, *bayesian ridge* e *bayesian lasso*. Observa-se que o menor valor é dado pelo modelo de regressão linear múltipla.

Tabela 4.6: Comparação dos Modelos - 2009 - Masculino

	Linear	B. Ridge	B. Lasso	F. Aleatória	Árvores	SVR
RMSE	11.764, 57	11.766, 41	11.768, 41	11.617, 49	11.883, 84	12.261, 19

A tabela a seguir apresenta um resumo do exposto nas seis tabelas acima.

Tabela 4.7: Resumo: melhor modelo estimado

	2007	2008	2009
Feminino	*B. Ridge*	Linear	Linear
Masculino	Linear	*B. Ridge*	Linear

Pela tabela 4.7, temos que o modelo de regressão linear múltipla teve o menor RMSE em 4 dos 6 ajustes feitos e modelo de regressão *bayesian ridge* teve menor RMSE em 2 dos 6 ajustes feitos. Além disso os valores do RMSE não eram muito diferentes. Assim, o modelo regressão linear múltipla é o escolhido para utilizarmos na simulação das 10.000 vidas, sendo um para o sexo feminino e um para o masculino. Utilizando o valor do RMSE, o modelo escolhido para o sexo feminino foi o ajustado em 2007 e para o sexo masculino foi o ajustado em 2007. Vale ressaltar que os modelos foram selecionados sem eliminar dados supostamente atípicos.

Para escolhermos o modelo que será utilizado na simulação, ajustamos a regressão linear múltipla para todos os anos e sexo. Os resultados encontram-se na Tabela 4.8 a seguir.

Tabela 4.8: Resumo: Regressão Linear Múltipla

Ano - Sexo	intercepto	idade	prev1	prev2	**RSME**
2007 - Feminino	73.61	25.87	0.28	0.25	7230.1
2007 - Masculino	-606.60	40.88	0.20	0.14	7116.4
2008 - Feminino	26.24	40.84	0.29	0.14	8528.2
2008 - Masculino	-788.58	42.01	0.49	0.17	10118.1
2009 - Feminino	-668.40	52.16	0.20	0.22	14425.4
2009 - Masculino	-1129.00	56.81	0.26	0.14	11764.5
2007 Sexo Junto	-252.70	35.21	0.18	0.19	8756.4
2008 Sexo Junto	-550.07	41.34	0.46	0.14	7935.5
2009 Sexo Junto	-1348.00	61.83	0.20	0.30	11700.2

Assim, escolhemos para o sexo feminino o modelo ajustado para o ano 2007 e para o sexo masculino o modelo ajustado para o ano 2007

Feminino: Regressão Linear Múltipla

$$y_i = 73,61 + 25,87\text{*idade} + 0,28\text{*gasto}_{prev1} + 0,25\text{*gasto}_{prev2} + e_i,\ i = 1, \ldots, n. \quad (4.7)$$

Masculino: Regressão Linear Múltipla

$$y_i = -606,60 + 40,88\text{*idade} + 0,20\text{*gasto}_{prev1} + 0,14\text{*gasto}_{prev2} + e_i,\ i = 1, \ldots, n. \quad (4.8)$$

4.2.2 Análise dos dados atípicos

Apesar de gastos com saúde apresentarem valores naturalmente atípicos, faremos uma análise utilizando algumas técnicas para detectar e remover dados atípicos.

As Tabelas 4.9, 4.10, 4.11, 4.12, 4.13, 4.14, são compostas por 3 linhas e 4 colunas; a primera linha nos diz a partir de qual valor serão removidos os dados, por exemplo, na posição (1,1) da Tabela 4.9 temos "$>$ *R*$ 2.513", isto quer dizer que todos os registros, que tiverem um gasto maior do que o mencionado, serão removidos. Nas segunda e terceira linhas encontramos a quantidade e porcentagem de dados que foram eliminados. O nome de cada coluna refere-se à técnica utilizada para remover *Outliers*.

Tabela 4.9: Dados atípicos para 2007 e sexo Feminino a partir de diferentes técnicas

	Intervalo Interquartil	**Classificação**	**Desvio-Padrão**	**Distribução Gama**
Gasto	$>$ R$ 2.513	$>$ R$ 12.417	$>$ R$ 25.294	$>$ R$ 104.236
Outliers	2.730	338	124	11
% Outliers	20,12	2,50	2,41	0,08

Tabela 4.10: Dados atípicos para 2007 e sexo Masculino a partir de diferentes técnicas

	Intervalo Interquartil	**Classificação**	**Desvio-Padrão**	**Distribução Gama**
Gasto	$>$ R$ 1.709	$>$ R$ 11.609	$>$ R$ 22.883	$>$ R$ 103.583
Outliers	2.778	356	164	11
% Outliers	19,50	2,50	1,15	0,07

Tabela 4.11: Dados atípicos para 2008 e sexo Feminino a partir de diferentes técnicas

	Intervalo Interquartil	**Classificação**	**Desvio-Padrão**	**Distribuição Gama**
Gasto	> R$ 3.020	> R$ 13.946	> R$ 41.558	> R$ 167.305
Outliers	2.666	338	80	9
% Outliers	19,69	2,50	0,59	0,06

Tabela 4.12: Dados atípicos para 2008 e sexo Masculino a partir de diferentes técnica

	Intervalo Interquartil	**Classificação**	**Desvio-Padrão**	**Distribuição Gama**
Gasto	> R$ 2.021	> R$ 14.835	> R$ 30.856	> R$ 140.614
Outliers	2.742	356	154	13
% Outliers	19,25	2,50	1,08	0,09

Tabela 4.13: Dados atípicos para 2009 e sexo Feminino a partir de diferentes técnica

	Intervalo Interquartil	**Classificação**	**Desvio-Padrão**	**Distribuição Gama**
Gasto	> R$ 2.868	> R$ 15.362	> R$ 37.818	> R$ 202.785
Outliers	2.740	338	100	14
% Outliers	20,23	2,50	0,73	0,10

Tabela 4.14: Dados atípicos para 2009 e sexo Masculino a partir de diferentes técnica

	Intervalo Interquartil	**Classificação**	**Desvio-Padrão**	**Distribuição Gama**
Gasto	> R$ 1.938	> R$ 16.518	> R$ 41.202	> R$ 185.299
Outliers	2.738	356	114	10
% Outliers	19,22	2,50	0,80	0,07

As análises de pontos discrepantes nos dizem que a distribuição Gama é uma boa opção devido à quantidade mínima de dados atípicos removidos, devemos levar em consideração que em gastos com saúde é comum dados atípicos portanto, ao aplicar essa técnica, retiramos entre 9 e 14 observações consideradas *outliers* em cada um dos anos.

4.2.3 Ajuste dos modelos sem dados atípicos

Aplicaremos a técnica da Distribuição Gama para remover *outliers* para cada ajuste. O objetivo é detectar se há ou não mudanças substanciais nos ajustes.

Ano 2007 - Feminino

Modelo de Linear Múltipla ajustado:

$$y_i = 607,90 + 25,17\text{*idade} + 0,10\text{*gasto}_{2006} + 0,12\text{*gasto}_{2005} + e_i,\ i = 1, \ldots, 13528. \quad (4.9)$$

Na Tabela 4.15 temos a comparação, quanto ao RMSE, dos modelos de regressão: linear múltipla, suporte vetorial (SVR), árvores, floresta aleatória, *bayesian ridge* e *bayesian lasso*. Observa-se que o menor valor é dado pelo modelo de regressão linear múltipla.

Tabela 4.15: Comparação dos Modelos - 2007 - Feminino

	Linear	B. Ridge	B. Lasso	árvores	SVR	F. Aleatória
RMSE	6.142, 44	6.144, 02	6.145, 47	6.239, 37	6.260, 93	6.388, 41

Ano 2007 - Masculino

Modelo de Linear Múltipla ajustado:

$$y_i = -646,70 + 44,83\text{*idade} + 0,14\text{*gasto}_{2006} + 0,08\text{*gasto}_{2005} + e_i,\ i = 1, \ldots, 14230. \quad (4.10)$$

Na Tabela 4.16 temos a comparação, quanto ao RMSE, dos modelos de regressão linear múltipla, suporte vetorial (SVR), árvores, floresta aleatória, *bayesian ridge* e *bayesian lasso*. Observa-se que o menor valor é dado pelo modelo de regressão linear múltipla.

Tabela 4.16: Comparação dos Modelos - 2007 - Masculino

	Linear	B. Ridge	B. Lasso	árvores	SVR	F. Aleatória
RMSE	4.474, 67	4.475, 46	4.476, 46	4.482, 68	4.652, 22	4.660, 30

Ano 2008 - Feminino

Modelo de árvores ajustado:

Na Tabela 4.17 temos a comparação, quanto ao RMSE, dos modelos de regressão: linear múltipla, suporte vetorial (SVR), árvores, floresta aleatória, *bayesian ridge* e *bayesian lasso*. Observa-se que o menor valor é dado pelo modelo de regressão árvores.

Tabela 4.17: Comparação dos Modelos - 2008 - Feminino

	Árvores	B. Lasso	B. Ridge	Linear	F. Aleatória	SVR
RMSE	7.388, 66	7.504, 01	7.506, 42	7.508, 33	7.578, 53	7.581, 70

Ano 2008 - Masculino

Modelo de árvores é ajustado

Na Tabela 4.18 temos a comparação, quanto ao RMSE, dos modelos de regressão: linear múltipla, suporte vetorial (SVR), árvores, floresta aleatória, *bayesian ridge* e *bayesian lasso*. Observa-se que o menor valor é dado pelo modelo de regressão árvores.

Tabela 4.18: Comparação dos Modelos - 2008 - Masculino

	Árvores	F. Aleatória	B. Lasso	B. Ridge	Linear	SVR
RMSE	7.451, 05	7.462, 89	7.504, 06	7.504, 73	7.505, 36	7.663, 05

Ano 2009 - Feminino

Modelo de árvores é ajustado

Na Tabela 4.19 temos a comparação, quanto ao RMSE, dos modelos de regressão: linear múltipla, suporte vetorial (SVR), árvores, floresta aleatória, *bayesian ridge* e *bayesian lasso*. Observa-se que o menor valor é dado pelo modelo de regressão árvores.

Tabela 4.19: Comparação dos Modelos - 2009 - Feminino

	Árvores	F. Aleatória	B. Lasso	B. Ridge	Linear	SVR
RMSE	8.165, 48	8.218, 32	8.286, 72	8.286, 81	8.287, 13	8.393, 85

Ano 2009 - Masculino

Modelo de Linear Múltipla ajustado é

$$y_i = -1074,00 + 53,74\text{*idade} + 0,21\text{*gasto}_{2008} + 0,16\text{*gasto}_{2007} + e_i,\ i = 1, \ldots, 14231. \tag{4.11}$$

Na Tabela 4.20 temos a comparação, quanto ao RMSE, dos modelos de regressão: linear múltipla, suporte vetorial (SVR), árvores, floresta aleatória, *bayesian ridge* e *bayesian lasso*. Observa-se que o menor valor é dado pelo modelo de regressão linear múltipla.

Tabela 4.20: Comparação dos Modelos - 2009 - Masculino

	Linear	B. Ridge	B. Lasso	Árvores	SVR	F. Aleatória
RMSE	6.892, 19	6.892, 28	6.892, 92	6.899, 23	7.023, 37	7.028, 66

Segundo o método da distribuição Gama temos poucos dados atípicos e decidimos dar continuidade a simulação de uma CP a partir dos modelos que consideraram todos os dados, isto é, sem retirar dados atípicos. Além disso, sabemos que dados atípicos são comuns em gastos com saúde. Os modelos escolhidos para a simulação, tanto para sexo feminino quanto para o masculino, foram os de regressão linear múltipla de menor RMSE de 2007, da Tabela 4.8.

4.3 Simulação e Resultados

Para avaliar a eficácia das contas poupança de saúde, apresentamos nesta seção os resultados das estimações dos saldos das contas individuais ao longo do tempo, além da frequência e severidade do seguro catastrófico.

Nas simulações, cada indivíduo possui uma conta, a partir dos 25 anos com duração de 41 anos. A dinâmica da CP aqui simulada foi extremamente simplificada, sem atualizações monetárias e supondo que os procedimentos médicos não sofrem acréscimos devido ao avanço tecnológico, inflação, etc., ao longo do tempo. A dinâmica dessa CP é a seguinte:

- O empregador deposita, anualmente, o valor de R$ 2.500,00 (este valor poderá ser alterado).
- Se o empregado gastar, no ano, menos de R$ 5.000,00 ele deve pagar todos os gastos mesmo que não tenha saldo suficiente em suas contas, mas se ele gastar mais de R$ 5.000,00, um seguro catastrófico será acionado e deve cobrir a despesa excedente.

Para dar início à simulação, foi escolhido o modelo de regressão linear estimado em 2007 (4.7) para o sexo feminino a partir de uma amostra de 13.539 mulheres. Foram selecionadas 1.267 pessoas que estavam na faixa etária de 25 a 30 anos. Para obter as 10.000 vidas usamos a função ***sample()*** do pacote R que permite obter uma amostra aleatória de elementos de um conjunto de dados (neste caso das 1.267 mulheres), com ou sem repetição. Aplicamos a função *sample()* com repetição.

Foi feito o mesmo procedimento para o sexo masculino em que o modelo de regressão linear múltipla escolhido também foi o de 2007 (4.8). Este modelo foi estimado a partir de 14.241 homens. Desses dados foram selecionadas 915 homens que estavam na faixa etária de 25 a 30 anos e, em seguida, para obter as 10.000 vidas, usamos a função *sample()* do pacote R, com repetição.

Considerando a média dos gastos da primeira faixa etária das 10.000 vidas para cada sexo de forma independente, a simulação começa supondo que todas as vidas da carteira tenham 25 anos e terminando aos 65 anos (41 anos de vida laboral).

Foram feitas 100 simulações da CP com 10.000 vidas para ambos os sexos.

Observação: Todas as tabelas e gráficos a seguir apresentam os valores médios obtidos nas 100 simulações.

4.3.1 Saldos das Contas Poupança

Para acompanharmos os saldos das CP apresentamos nas Tabelas 32 [1] [MMPP] 4.21 e 4.22, medidas de posição e dispersão das contas poupança individuais referente às médias das 100 simulações das 10.000 vidas, para cada faixa etária e sexo, ao longo do tempo.

Tabela 4.21: Estatística descritivas dos saldos das CP por faixa etária para o sexo feminino

	25 anos	30 anos	35 anos	40 anos	45 anos	50 anos	55 anos	60 anos	65 anos
n_0	1.510	0	0	0	0	0	0	212	9.995
Percentis(5)	155,70	2.552,06	5.382,09	7.112,74	7.519,25	6.568,37	4.255,19	781,23	8,67
Percentis(10)	275,45	2.902,82	5.788,49	7.527,46	7.935,17	6.984,47	4.671,32	1.130,95	12,43
Percentis(15)	382,37	3.205,41	6.135,69	7.880,18	8.288,71	7.338,12	5.024,99	1.468,92	16,27
Percentis(20)	535.25	3.822,34	6.808,15	8.562,55	8.972,38	9.021,99	5.708,89	2.116,35	23,88
Percentis(40)	698,86	4.395,84	7.425,28	9.186,31	9.597,21	8.647,00	6.333,93	2.693,49	27,57
Percentis(50)	791,61	4.650,06	7.706,62	9.470,04	9.881,22	8.931,03	6.617,96	2.968,65	30,04
Percentis(75)	1.003,14	5.190,05	8.292,77	10.063,77	10.476,26	9.526,25	7.213,21	3.547,34	59,74
Percentis(85)	1.096,38	5.436,05	8.558,65	10.332,17	10.745,00	9.795,05	7.482,02	3.813,01	118,57
Percentis(95)	1.234,26	5.764,92	8.920,07	10.697,87	11.111,32	10.161,47	7.848,44	4.179,59	177,39
Percentis(98)	1.318,20	5.980,95	9.154,03	10.933,82	11.347,58	10.397,76	8.084,75	4.416,20	195,04
Máximo	1.586,48	6.733,58	9.977,46	11.769,26	12.184,79	11.235,24	8.922,26	5.246,11	206,81
Média	756,12	4.436,30	7.466,70	9.226,97	9.637,83	8.687,59	6.374,51	2.767,67	65,07
Desvio Padrão	324,47	1.017,11	1.118,86	1.133,88	1.136,10	1.136,43	1.136,43	1.043,63	81,65

n_0 : Total de indivíduos com saldo zero na conta poupança

Ao analisar a Tabela 4.21, observa-se que as mulheres acumulam capital nos primeiros 20 anos de vida laboral, aproximadamente até os 50 anos. No entanto, a partir dessa idade, ocorre uma descapitalização na conta da CP.

Nas 100 simulações realizadas, houve uma variabilidade entre 0 e 4500 pessoas que acumularam dinheiro ao finalizar os 41 anos de vida laboral. Isso significa que em várias simulações não houve mulheres que acumularam dinheiro para sua aposentadoria, enquanto em outras simulações o máximo foi de 4500 pessoas que conseguiram acumular capital.

A média das mulheres que tiveram saldo no momento da aposentadoria foi de 5, com uma média de R$ 65,07. O valor máximo alcançado foi de R$ 206,81, e o desvio padrão foi de R$ 81,65 ao final dos 41 anos de vida laboral. Em outras palavras, 99,95% dos indivíduos não poderão desfrutar de dinheiro quando se aposentarem.

[1] As tabelas 28, 29, 30 encontra-se na seção do apêndice

Tabela 4.22: Estatística descritivas dos saldos das CP por faixa etária para o sexo masculino.

	25 anos	30 anos	35 anos	40 anos	45 anos	50 anos	55 anos	60 anos	65 anos
n_0	1.033	0	0	0	0	0	0	0	0
Percentis(5)	1.106,03	7.004,22	13.979,25	19.431,16	23.311,81	25.619,80	26.355,09	25.517,67	23.107,54
Percentis(10)	1.229,93	7.232,56	14.214,52	19.666,62	23.547,28	25.855,28	26.590,56	25.753,14	23.343,01
Percentis(15)	1.307,54	9.158,43	16.165,79	21.618,64	25.499,32	27.807,32	28.542,60	27.705,18	25.295,05
Percentis(20)	1.419,19	9.493,44	16.506,37	21.959,40	25.840,09	28.148,08	28.883,37	28.045,95	25.635,82
Percentis(40)	1.540,16	9.749,44	16.764,73	22.217,82	26.098,51	28.406,50	29.141,79	28.304,36	25.894,24
Percentis(50)	1.602,71	9.868,74	16.885,95	22.339,09	26.219,78	28.527,77	29.263,06	28.425,64	26.015,51
Percentis(75)	1.743,12	10.124,83	17.145,18	22.598,43	26.479,13	28.787,12	29.522,40	28.684,98	26.274,85
Percentis(85)	1.805,60	10.235,64	17.257,62	22.710,91	26.591,60	28.899,60	29.634,88	28.797,46	26.387,33
Percentis(95)	1.894,66	10.386,56	17.410,27	22.863,60	26.744,30	29.052,29	29.787,57	28.950,15	26.540,02
Percentis(98)	1.942,04	10.471,14	17.496,19	22.949,56	26.830,26	29.138,25	29.873,54	29.036,12	26.625,99
Máximo	2.064,27	10.712,91	17.741,10	23.194,57	27.075,27	29.383,26	30.118,55	29.281,13	26.871,00
Média	1.562,01	9.562,83	16.575,37	22.028,38	25.909,07	28.217,07	28.952,35	28.114,93	25.704,80
Desvio Padrão	248,79	974,40	989,01	989,44	989,45	989,45	989,45	989,45	989,45

n_0 = Total de indivíduos com saldo zero na conta poupança

Observando a Tabela 4.22 os homens acumulam capital nos primeiros 30 anos de vida laboral, aproximadamente até os 60 anos, porém, a partir dos 60 anos começam a descapitalizar a CP. Nas simulações, os 10.000 dos homens acumularam capital, com saldo médio de $R\$$ 25.704, máximo de $R\$$ 26.871 e desvio padrão de $R\$989,45$ ao final de 41 anos de vida laboral, ou seja, 100% (também nas 100 simulações) dos indivíduos poderão desfrutar de dinheiro quando se aposentarem.

4.3.2 Frequência de utilização e severidade do seguro catastrófico

Nesta seção, analisamos o uso do seguro catastrófico referente às médias das 100 simulações das 10.000 vidas. Especificamente analisamos a frequência com que o seguro é acionado e a e severidade. Nas Tabelas 4.23 4.24, apresentamos a frequência e o valor total coberto (severidade) das vidas que utilizaram o seguro catastrófico.

Sexo Feminino

Das 10.000 vidas simuladas para o sexo feminino o seguro foi usado, no máximo, 11 vezes por 2 mulheres e, há 5 mulheres que não utilizaram o seguro. Aproximadamente 47% da mulheres utilizou o seguro catastrófico mais de duas vezes, e aproximadamente 92% utilizou 5 vezes ou menos o seguro. Além disso, observamos que cerca de 60% dos gastos cobertos pelo seguro catastrófico foram com pessoas que o utilizaram no máximo 3 vezes, o que corresponde quase a 78% das vidas, enquanto 20% dos gastos coberto foram com vidas que o utilizaram 5 ou mais, o que corresponde 8,5% das vidas. Também temos que cerca de 6% dos gastos cobertos pelo seguro foram com 212 vidas, que o utilizaram mais de 7 vezes.

Do ponto de vista populacional, o total de gastos com saúde da carteira simulada durante o período de 41 anos foi de R$ 1.049.147.823, sendo R$ 24.148.148 (2,30%) cobertos pelo seguro catastrófico e R$ 1.024.999.675 (97,69%) cobertos pelo seguro HSA individual. Foram depositados na carteira de poupança R$ 1.025.000.000,

dos quais 99,99% foram utilizados para custear despesas com saúde durante a vida profissional, isto quer dizer que só têm R$ 325 para dar às cinco pessoas depois de que se aposentarem.

Tabela 4.23: Frequência e severidade do seguro catastrófico pelas vidas simuladas durante o período de 41 anos, baseadas no Modelo de regressão linear múltipla - feminino

Número de vezes	n	%	% acum	Value (R$)	%	% acum	Média Individual
0	5	0,05%	0,05%	0,00	0,00%	0	0,00
1	1100	11,00%	11,05%	920.450,90	3,81%	3,81%	836,77
2	4193	41,93%	52,98%	6.969.533,02	28,86%	32,67%	1.662,18
3	2496	24,96%	77,94%	6.512.839,00	26,97%	59,64%	2.609,31
4	685	6,85%	84,79%	2.382.032,54	9,86%	69,51%	3.477,42
5	673	6,73%	91,52%	2.714.295,17	11,24%	80,75%	4.033,13
6	636	6,36%	97,88%	3.177.690,97	13,16%	93,91%	4.996,37
7	165	1,65%	99,53%	1.079.431,24	4,47%	98,38%	6.542,01
8	10	0,10%	99,63%	75.373,98	0,31%	98,69%	7.537,40
9	16	0,16%	99,79%	127.478,18	0,53%	99,22%	7.967,39
10	19	0,19%	99,98%	165.705,20	0,69%	99,90%	8.721,33
11	2	0,02%	100,00%	23.318,60	0,10%	100,00%	11.659,30
TOTAL	**10.000**	**100%**		**24.148.148,79**	**100%**		**2.414,81**

Sexo Masculino

Das 10.000 vidas, 8.967 vidas simuladas não utilizaram o seguro no período de 41 anos, aproximadamente 10% utilizou uma vez , e 0.2% utilizou duas vezes o seguro. Além disso, observamos que 98% dos gastos cobertos pelo seguro catastrófico foram com pessoas que o utilizaram no máximo uma vez, o que corresponde a 99% das vidas.

Tabela 4.24: Frequência e severidade do seguro catastrófico pelas vidas simuladas durante o período de 41 anos, baseadas no Modelo de regressão linear múltipla - masculino

Número de vezes	n	%	% acum	Value (R$)	%	% acum	Média Individual
0	8967	89,67%	89,67%	0,00	0,00%	0,00%	0
1	1031	10,31%	99,98%	534.362,93	98,47%	98,47%	518,29577
2	2	0,02%	100,00%	8.310,41	1,53%	100,00%	4155,2027
TOTAL	**10.000**	**100%**		**542.673,34**	**100%**		**54,267334**

Do ponto de vista populacional, o total de gastos com saúde da carteira simulada durante o período de 41 anos foi de R$ 768.494.664, sendo R$ 542.673 (0,07%) cobertos pelo seguro catastrófico e R$ 767.951.991 (99,92%) cobertos pelo seguro HSA individual. Foram depositados na carteira de poupança R$ 1.025.000.000, dos quais 74.8% foram utilizados para custear despesas com saúde durante a vida profissional e 25.06% permaceram para cobrir os gastos com saúde dessa população ápos a aposentadoria.

4.3.3 Balanço da conta poupança, coberturas dos gastos com saúde pelas conta poupança e pelo seguro catastrófico

Há três aspectos que devemos avaliar quando uma CP é implementada: o saldo aos 65 anos, as despesas cobertas pela conta poupança saúde e os gastos cobertos pelo seguro catastrófico dos indivíduos. As duas primerias características são mais suscetíveis ao comportamento de um indivíduo. As Tabelas 34 [MMPP], 4.25 4.26 exibem estatísticas descritivas dessas três características.

Sexo Feminino

Tabela 4.25: Estatísticas descritivas do balanço da conta poupança, gastos com saúde cobertos pela conta poupança e pelo seguro catastrófico durante a vida laboral, baseadas no Modelo de regressão linear múltipla - feminino

	Balanço das contas poupança (BCP) depois dos 65 anos	Gastos cobertos pelo Balanço das contas poupança	Gastos cobertos pelo Seguro Catastrófico	Percentagem do total dos gastos cobertos pelo seguro catastrófico
% de valores nulos	99.95	0,00	0,05	0,05
P(5)	8,67	102.500	867,01	0,83
p(10)	12,47	102.500	1.080,19	1,04
P(15)	16,27	102.500	1.233,43	1,18
P(25)	23,88	102.500	1.502,24	1,44
P(40)	27,58	102.500	1.857,39	1,77
P(50)	30,04	102.500	2.097,51	2,00
P(60)	41,92	102.500	2.381,54	2,27
P(75)	59,74	102.500	3.006,58	2,84
P(85)	118,57	102.500	3.734,67	3,51
P(90)	147,98	102.500	4.253,09	3,98
P(95)	177,40	102.500	2.975,15	4,62
P(97)	189,16	102.500	5.543,53	5,13
P(98)	195,04	102.500	6.049,11	5,57
P(99)	200,93	102.500	6.773,98	6,19
P(995)	203,87	102.500	7.443,38	6,77
Máximo	206,81	102.500	11.938,08	10,43
Média	65,07	102.500	2.414,81	2,28
Desvio Padrão	81,65	2,18	1.318,77	1,20

Se uma mulher não tivesse gasto com saúde no período de 41 anos, seu saldo em CP seria de R$102.500, neste caso temos um saldo médio simulado aos 65 anos de R$65, 07 e o máximo de R$206, 81. O gasto médio coberto durante a vida laboral pelo contas poupança foi em R$102.500 isso quer dizer que quase todas gastam R$2.500 anuais e apenas cinco pessoas vão ter dinheiro quando se aposentarem. Para 25% das vidas femininas, o seguro catastrófico cobriu o valor de R$ 1.502, 24, e para terceira parte das vidas foi coberto um total de R$3.006, 58.

Por outro lado, a metade das vidas teve 2% dos gastos cobertos pelo seguro catastrófico; enquanto que 90% das vidas teve 4% das despesas cobertas por ele, e 1% das vidas teve mais de 3,8% de suas despesas cobertas pelo seguro.

A Figura 4.1 apresenta o histograma da severidade do uso do seguro catastrófico,

evidenciando a cobertura de grandes gastos com saúde pelo seguro.

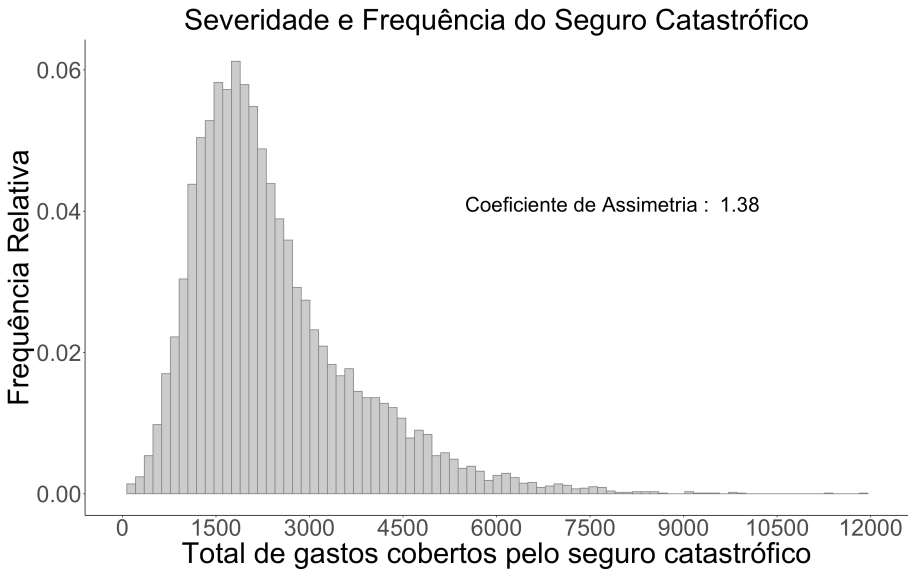

Figura 4.1: Severidade do Seguro Catastrófico

Na Figura 4.2 apresentamos o gráfico de dispersão entre a despesa total no período de 41 anos de cada vida e a porcentagem dessa despesa que foi coberta pelo seguro catastrófico. Os pontos de cor roxa referem-se à quantidade mínima de vezes em que o seguro foi utilizado durante a vida laboral, os pontos de cor amarelo representam as pessoas que usaram o seguro aproximadamente 5 vezes, e os pontos de cor vermelho são os indivíduos que acionaram o seguro mais de 8 vezes.

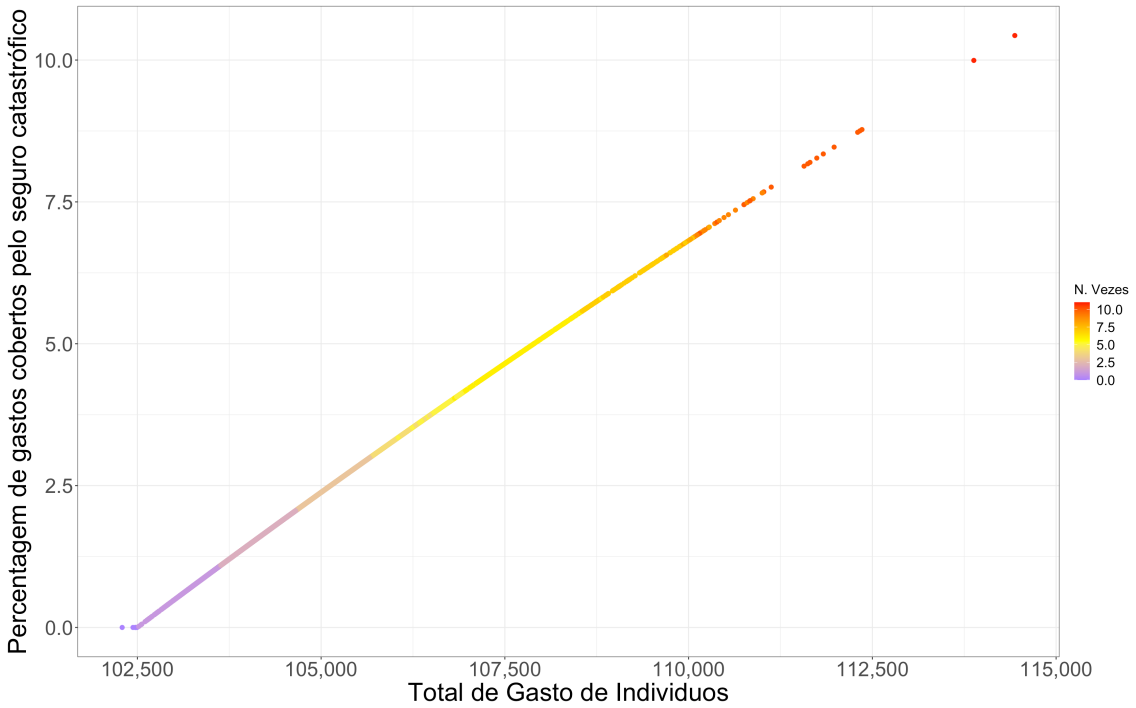

Figura 4.2: Gráfico de dispersão entre os gastos totais no período de 41 anos de cada vida e o percentagem dessos gastos que foram cobertos pelo seguro catastrófico

Sexo Masculino

Tabela 4.26: Estatística descritivas do Balance da conta poupança, cobertura das contas poupança e pelo seguro catastrófico durante a vida laboral, baseadas no Modelo de regressão linear múltipla - masculino

	Balanço das contas poupança (BCP) depois dos 65 anos	Gastos cobertos pelo Balanço das contas poupança	Gastos cobertos pelo Seguro Catastrófico	Percentagem do total dos gastos cobertos pelo seguro catastrófico
% de valores nulos	0,00	0,00	89,67	89,67
P(5)	23.107,54	75.957,97	0,00	0,00
p(10)	23.343,01	76.044,24	0,00	0,00
P(15)	25.295,05	76.112,66	0,00	0,00
P(25)	25.635,82	76.225,14	0,00	0,00
P(40)	25.894,24	76.377,98	0,00	0,00
P(50)	26.015,51	76.484,48	0,00	0,00
P(60)	26.122,01	76.605,75	0,00	0,00
P(75)	26.274,85	76.864,17	0,00	0,00
P(85)	26.387,33	76.204,94	0,00	0,00
P(90)	26.465,75	79.156,98	77,30	0,09
P(95)	26.540,02	79.392,45	411,83	0,51
P(97)	26.587,11	79.478,42	549,53	0,68
P(98)	26.625,99	79.561,48	644,26	0,80
P(99)	26.676,38	79.671,55	881,03	1,09
P(995)	26.715,83	80.553,57	2.131,68	2,57
Máximo	26.871,00	81.937,26	4.208,83	4,88
Média	25.704,80	76.795,19	54,26	0,06
Desvio Padrão	989,45	989,45	232,53	0,28

Se uma vida não tivesse gasto no período de 41 anos, seu saldo em conta seria de R$102.500, neste caso temos um saldo médio simulado aos 65 anos de R$25.704, sendo o máximo de R$26.871. O gasto médio coberto durante a vida laboral pelo

contas poupança foi em R$76.795 e o máximo foi de R$ 81.937. Para 95% das vidas, o seguro catastrófico foi coberto com R$ $411,83$, e para 100 vidas foi coberto com R$$881,03$ o mais.

Por outro lado, 3% das vidas tinhan 0,6% dos gastos cobertos pelo seguro. enquanto 1% das vidas tiveram mais do 1% despesas cobertas por ele, e 0,5% das vidas tiveram de 2,5% de suas despesas cobertas pelo seguro.

As Figura 4.3 apresentam o histograma da severidade do uso do seguro catastrófico, evidenciando a cobertura de grandes gastos com saúde pelo seguro.

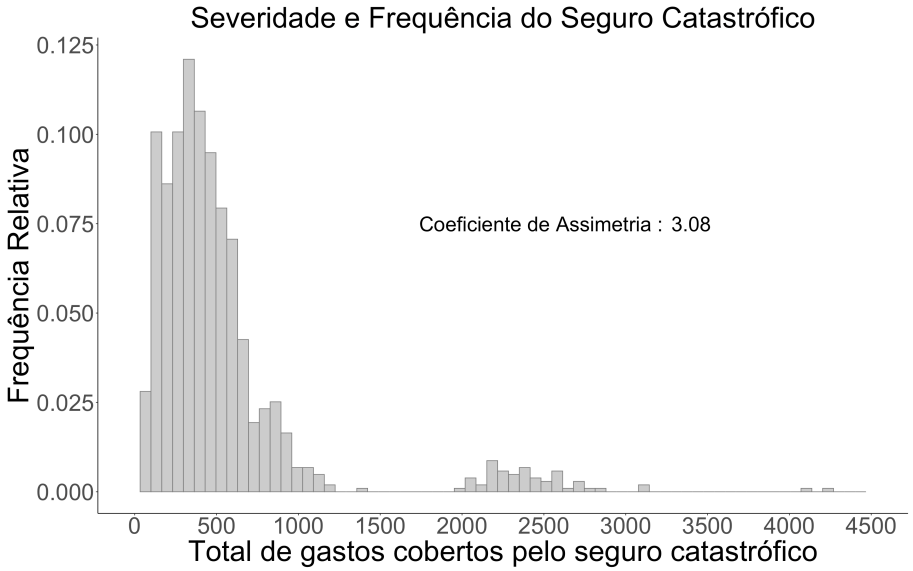

Figura 4.3: Severidade do Seguro Catastrófico

Na Figura 4.4 apresentamos o gráfico de dispersão entre as despesas totais no período de 41 anos de cada vida e a porcentagem dessas despesas que foram cobertas pelo seguro catastrófico. Os pontos de cor amarelo referem-se à quantidade de vezes que o seguro foi utilizado uma vez, e os dois pontos de cor vermelho representam as pessoas que usaram o seguro duas vezes durante a sua vida laboral.

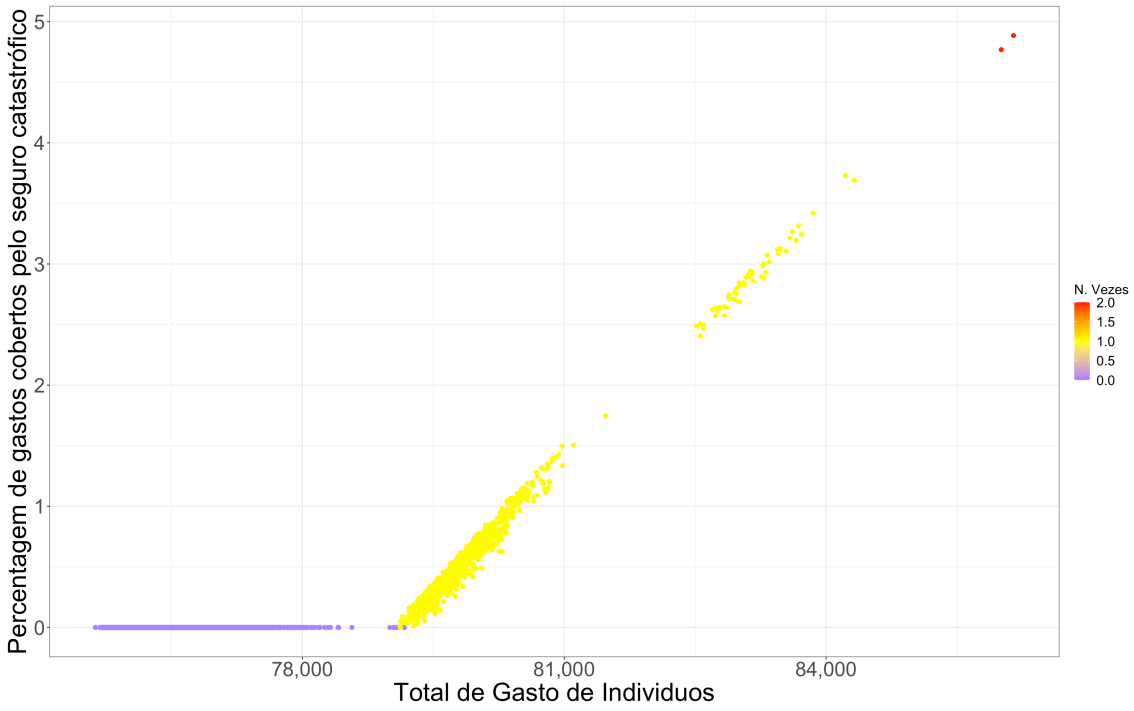

Figura 4.4: Gráfico de dispersão entre os gastos totais no período de 41 anos de cada vida e o percentagem dessos gastos que foram cobertos pelo seguro catastrófico

4.3.4 Seleção e análise de indivíduos representativos

A seguir, apresentamos um estudo adicional no qual selecionamos três indivíduos representativos, tanto do sexo feminino quanto do sexo masculino. O objetivo principal deste estudo é observar seus padrões de gastos com saúde, seu poupança e sua frequência no uso seguro catastrófico. Em seguida, analisaremos se as condições iniciais desses indivíduos têm influência em seu comportamento ao longo dos 41 anos do período estudado.

No caso das mulheres, nosso modelo de trabalho baseou-se em dados do ano de 2007. Para isso, começamos com uma população de 1.267 mulheres na faixa etária de 25 a 30 anos naquele ano. A fim de selecionar três mulheres representativas, realizamos os seguintes passos:

1. Primeiro, calculamos a média de cada variável em nosso banco de dados, ou seja, determinamos a média das diferentes características estudadas.
2. Usando essas médias calculadas, determinamos a distância euclidiana de cada indivíduo em nosso banco de dados em relação a essas médias.
3. Em seguida, classificamos os indivíduos com base nessa distância euclidiana e selecionamos as três primeiras mulheres que mostraram maior semelhança em relação às médias calculadas.

De maneira análoga, selecionamos três homens no ano de 2007, com um ponto de diferença a população masculina na mesma faixa etária era composta por 965 indivíduos.

As mulheres estudadas são:

Sex	*Year*	*Age*	*Agerange*	*Previous1 expensetot*	*Previous2 expensetot*	*Expensetot*
F	2007	29	25a30	255.84	1654.41	1910.54
F	2007	29	25a30	437.84	2521.24	1938.37
F	2007	29	25a30	730.18	379.96	1928.69

Uma vez identificados os indivíduos, realizei o seguinte:

Tabela 4.27: Estatística descritivas dos saldos das CP por faixa etária para o sexo femenino das trés mulheres representativas.

	25 anos	30 anos	35 anos	40 anos	45 anos	50 anos	55 anos	60 anos	65 anos
Mulher(1)	589,46	5.173,43	8.305,14	10.080,54	10.493,64	9.543,74	7.230,71	3.554,54	0,00
Mulher(2)	561,63	5.021,05	8.135,85	9.908,74	10.321,47	9.371,51	7.058,47	3.382,31	0,00
Mulher(3)	571,31	4.888,36	7.985,43	9.755,66	10.168,00	9.217,98	6.904,93	3.228,77	0,00
Máximo	589,46	5.173,43	8.305,14	10.080,54	10,493,64	9.543,74	7.230,71	3.554,54	0,00
Média	574,13	5.027,61	8.142,14	9.914,98	10.327,71	9.377,74	7.064,70	3.388,54	0,00
Desvio Padrão	14,12	142,65	159,94	162,53	162,91	162,96	162,97	162,97	0,00

A seguir, apresenta-se uma tabela que mostra a média de três pessoas representativas e das 10.000 vidas simuladas, todas pertencentes ao sexo feminino.

Tabela 4.28: Média das estatística descritivas das 10.000 vidas simuladas e dos três indíviduos representativos para o sexo femenino

	25 anos	30 anos	35 anos	40 anos	45 anos	50 anos	55 anos	60 anos	65 anos
Média (10.000)	756,12	4.436,30	7.466,70	9.226,97	9.637,83	8.687,59	6.374,51	2.767,67	65,07
Média (3)	574,13	5.027,61	8.142,14	9.914,98	10.327,71	9.377,74	7.064,70	3.388,54	0,00

A Tabela 4.28 mostra que os valores médios das 10.000 vidas em comparação com a média das três mulheres representativas que selecionamos diferem em uma faixa de 590 a 690, favorecendo a média dos três dados representativos. No entanto, ao analisar os dados observamos que as condições iniciais foram as mesmas e as variáveis do modelo, assim como a duração do acompanhamento, também coincidiram, a diferença nos resultados provavelmente se deve à forma como o gasto com saúde está sendo modelado por meio do modelo de regressão linear múltipla. Ou seja, o modelo gerado é um fator mais influente nos resultados do que as condições iniciais, já que as diferenças observadas nas três pessoas representativas em comparação com a médias das 10.000 vidas podem ser atribuídas ao funcionamento interno do modelo. Logo fazemos o mesmo proceso mas com o sexo masculino

Os homens estudadas são:

Sex	*Year*	*Age*	*Agerange*	*Previous1 expensetot*	*Previous2 expensetot*	*Expensetot*
M	2007	29	25a30	629.56	1266.13	1172.42
M	2007	27	25a30	1353.08	1943.214	1137.42
M	2007	29	25a30	1177.15	650.42	1171.88

Uma vez identificados os indivíduos, realizei o seguinte:

Tabela 4.29: Estatística descritivas dos saldos das CP por faixa etária para o sexo masculino das trés mulheres representativas.

	25 anos	30 anos	35 anos	40 anos	45 anos	50 anos	55 anos	60 anos	65 anos
Homem(1)	1.327,58	9.602,56	16.617,67	22.070,76	25.951,45	28.259,44	28.994,73	28.157,30	25.747,17
Homem(2)	1.362,58	9.501,52	16.513,61	21.966,60	25.847,29	28.155,28	28.890,57	28.053,15	25.643,02
Homem(3)	1.328,12	9.486,64	16.498,99	21.951,99	25.832,68	28.140,67	28.875,96	28.038,54	25.628,41
Máximo	1.362,58	9.602,56	16.617,67	22.070,76	25.951,45	28.259,44	28.994,73	28.157,30	25.747,17
Média	1.339,42	9.530,24	16.543,43	21.996,45	25.877,14	28.185,13	28.920,42	28.083,00	25.672,87
Desvio Padrão	20,05	63,07	64,71	64,76	64,76	64,76	64,76	64,76	64,76

A seguir, apresenta-se uma tabela que mostra a média de três pessoas representativas e das 10.000 vidas simuladas, todas pertencentes ao sexo masculino.

Tabela 4.30: Média das estatística descritivas das 10.000 vidas simuladas e dos três indíviduos representativos para o sexo masculino

	25 anos	30 anos	35 anos	40 anos	45 anos	50 anos	55 anos	60 anos	65 anos
Média (10.000)	1.562,01	9.562,83	16.575,37	22.028,38	25.909,07	28.217,07	28.952,35	28.114,93	25.704,80
Média (3)	1.339,42	9.530,24	16.543,43	21.996,45	25.877,14	28.185,13	28.920,42	28.083,00	25.672,87

A Tabela 4.30 mostra que os valores médios das 10.000 vidas em comparação com a média dos 3 homens representativos são bastante similares. Os valores diferem em uma média de R $ 32, favorecendo ligeiramente a média das 10.000 vidas. Portanto, podemos concluir novamente que as condições iniciais não parecem ter tanta influência quanto o modelo de regressão linear múltiple gerado, especialmente no caso masculino, onde os resultados foram muito semelhantes.

Em relação ao seguro catastrófico, podemos dizer que as mulheres acionaram esse serviço aos 64 e 65 anos, com uma média de R$ 1.600 reais por 2 anos, ou seja, R $ 800 de seguro catastrófico anual. Vale ressaltar que durante 39 anos de vida laboral, elas pagaram as despesas de saúde com sua poupança ou dinheiro próprio. Por outro lado, nenhum homem utilizou o seguro, mesmo também que a média de 10.000 homens simulados, onde nenhum deles o utilizou o seguro catastrófico.

Conclusões

Este trabalho tem como objetivo, avaliar a viabilidade de uma plano saúde tipo CP para o empregador, empregado e para órgãos governamentais de regulação. Apresentamos, sob certas quantidades fixadas, a simulação de um produto de plano de saúde direcionado ao consumidor que nunca foi comercializado no Brasil, porém baseado em dados de gastos com saúde de vidas do Brasil. Este estudo foi feito para cada sexo separadamente e obtivemos as Tabelas 4.21 e 4.22 que mostram o balanço das contas poupança ao longo do tempo e a frequência e severidade do seguro catastrófico.

Um subproduto deste trabalho era comparar os resultados obtidos com os do artigo [MMPP], o qual utiliza outro método para estimar os gastos. Vamos começar comparando as tabelas e depois os gráficos.

1. A Tabela 32 [MMPP] apresenta resultados obtidos por meio de cadeias de Markov, mostrando que o saldo na conta poupança aumenta anualmente, resultando em 57 indivíduos sem balanço positivo ao final de sua vida laboral, ou seja, 99,5% dos indivíduos poderão desfrutar de dinheiro quando se aposentarem, com média de R$ 43.937,00, máximo de R$ 92.628,00 e desvio padrão de R$ 18.500,00 ao final de 41 anos de vida laboral.

Por outro lado, nas Tabelas 4.21 e 4.22 são apresentados resultados obtidos através de um modelo de regressão linear múltipla para mulheres e homens, respectivamente. Esses modelos analisam as relações entre variáveis para prever despesas de saúde em função de diferentes fatores, incluindo idade e gastos anuais dos dois anos anteriores.

Na Tabela 4.21, observa-se que as despesas com saúde das mulheres são maiores após os primeiros 15 anos de vida laboral, No entanto, esses maiores gastos em saúde resultam em apenas 5 (com uma variabiliade entre 0 como o mínimo e 4500 o máximo de mulheres), em cada 10.000 mulheres, conseguindo um balanço positivo ao final da vida laboral.

Por sua vez, a Tabela 4.22 mostra que o modelo de regressão para homens permitiu prever o aumento do saldo na conta poupança anualmente durante os primeiros 30 anos de vida laboral, mas depois há um aumento nas despesas de saúde nos últimos 10 anos. É importante destacar que, diferentemente das mulheres, 100% dos homens conseguiram economizar ao final da vida laboral (também nas 100 simulações), com média de R$ 25.704,00, máximo de R$ 26.871,00 e desvio padrão de R$989,00.

2. A Tabela 33 [MMPP] mostra que 607 pessoas não usaram o seguro catastrófico ao

longo de 41 anos, enquanto apenas uma pessoa o utilizou 24 vezes. Além disso, vemos que os gastos totais foram de R$ 1.088.620.494, 00, dos quais 46% foram cobertos pelo seguro catastrófico e o restante foi coberto pela conta HSA individual. Foram depositados R$ 1.025.000.000 nas contas poupança, dos quais 57% foram usados para cobrir os custos de assistência médica durante a vida ativa e 43% foram destinados a cobrir os custos de assistência médica dessa população após a aposentadoria.

Na tabela 4.23 para o sexo feminino, 5 pessoas não usaram o seguro, 9995 pessoas o utilizaram com um máximo de 2 mulheres que usaram o seguro 11 vezes. Além disso, vemos que os gastos totais foram de R$ 1.049.147.823, dos quais 2,3% foram cobertos pelo seguro catastrófico e o restante foi coberto pela conta HSA individual. Foram depositados R$ 1.025.000.000 nas contas de poupança, dos quais 99,99% foram usados para cobrir os custos de assistência médica durante a vida ativa e R$ 325,00 foram destinados a cobrir os custos de assistência médica dessa população após a aposentadoria.

Na tabela 4.24 para o sexo masculino, 8.967 homens não usaram o seguro, 1033 homens usaram o seguro e 2 de eles o utilizaram como máximo 2 vezes. Além disso, vemos que os gastos totais foram de R$ 768.494.664, dos quais 0,7% foram cobertos pelo seguro catastrófico e o restante foi coberto pela conta HSA individual. Foram depositados R$ 1.025.000.000 nas contas de poupança, dos quais 75% foram usados para cobrir os custos de assistência médica durante a vida ativa e 25% foram destinados a cobrir os custos de assistência médica dessa população após a aposentadoria.

Tabela 4.31: Resumo do Seguro Catastrófico

	Linear Femenino	Linear Masculino	Cadeias de Markov
Não utilizaram o Seguro	5	8.967	607
Utilizaram o Seguro	9.995	1.033	9.393
Gastos Totais (R$)	R$ 24.148.148	R$ 542.673	R$ 500.490.020
% Coberto Seguro	2,3 %	0,7%	46 %
Media	R$ 2.414,8	R$ 54,2	R$ 50.049

3. Pode-se dizer que as Tabelas 34 [MMPP], 4.25 e 4.26 são um resumo geral dos resultados do modelo linear. A primeira coluna mostra o saldo das contas de poupança após os 65 anos, a coluna seguinte é a despesa coberta pelas contas de poupança, a terceira coluna é a despesa coberta pelo seguro catastrófico e a última é a porcentagem do total dos gastos cobertos pelo seguro catastrófico. É importante mencionar que se uma pessoa não tiver gasto durante seus 41 anos de vida profissional, seu saldo de conta será de R$ 102.500, 00.

Na Tabela 34 [MMPP], o saldo médio e máximo simulado aos 65 anos é R$ 44.000, 00 e R$ 92.628, 00, respectivamente. Há 57 indivíduos que não terão balanço positivo quando se aposentarem, o que significa que essas pessoas gastaram os R$ 102.500 que o empregador depositou em suas contas poupança durante seus 41 anos de vida laboral. O gasto médio coberto pelo seguro durante a vida laboral foi cerca de R$ 53.000, 00 e o máximo foi de quase R$ 1.200.000, 00, o que equivale a 22 vezes o valor médio. Ao considerar a porcentagem de despesas cobertas pelo seguro, o 5%

das pessoas tiveram mais de 75% de suas despesas cobertas.

A Tabela 4.25 se refere ao sexo feminino e mostra que o saldo médio simulado aos 65 anos é de R$ 65 e o máximo é de R$ 206. A despesa média coberta durante a vida laboral pelas contas de poupança foi de R$ 102.500, o que significa que quase todas as mulheres gastaram R$ 2.500 por ano e apenas cinco pessoas terão poupança quando se aposentarem. A despesa média coberta pelo seguro durante a vida laboral foi de R$ 2.414 e a máxima foi de quase R$ 12.000, o que equivale a 5 vezes o valor médio. Ao considerar a porcentagem de despesas cobertas pelo seguro, o 5% das mulheres tiveram mais de 4,6% de suas despesas cobertas pelo seguro.

A Tabela 4.26 se refere ao sexo masculino e mostra que o saldo médio simulado aos 65 anos é de R$ 25.704, sendo o máximo de R$ 26.871. A despesa média coberta durante a vida laboral pela conta de poupança foi de R$ 76.795 e a máxima foi de R$ 81.937. A despesa média coberta durante a vida laboral pelo seguro foi de R$ 54 e a máxima foi de quase R$ 4.20, o que equivale a 77 vezes o valor médio. Ao considerar a porcentagem de despesas cobertas pelo seguro, o 5% dos homens tiveram mais de 0,6% de suas despesas cobertas pelo seguro.

4. Foram analisados três gráficos 4.1, 4.3 e 5 de distribuições de dados e foi constatado que todos apresentavam um viés à direita, o que se reflete em um coeficiente de assimetria maior que zero. O viés à direita ou viés positivo indica que a cauda da distribuição se estende a valores maiores que a média, enquanto a cauda à esquerda é mais curta. Os valores do coeficiente de assimetria foram obtidos através da função "*skewness*" em **R**, que mede o grau de assimetria em uma distribuição. Em particular, o primeiro gráfico 4.1 teve um coeficiente de assimetria de 1.3, enquanto o segundo 4.3 teve um coeficiente de assimetria de 3.1. Embora o valor exato do coeficiente de assimetria do terceiro gráfico 5[2] seja desconhecido, pode-se observar no gráfico que a distribuição apresenta um viés ainda maior do que os dois gráficos anteriores.

5. Foi realizada uma análise de dispersão com as gráficas 4.2, 4.4 e 6 para examinar a relação entre os gastos totais em um período de 41 anos de vida e a porcentagem desses gastos cobertos por seguros catastróficos. Foram obtidos três gráficos diferentes. O primeiro 4.2 apresenta uma dispersão linear sem grandes espaços, o que sugere que à medida que aumenta a porcentagem de gastos cobertos por seguros catastróficos, também aumenta o total de gastos. Em contraste, o segundo gráfico 4.4 mostra uma dispersão horizontal no início e depois com uma linha reta com inclinação positiva, mas com alguns espaços. Isso indica que, embora exista uma relação positiva entre a porcentagem de gastos cobertos por seguros catastróficos e os gastos totais, essa relação não é tão clara. No terceiro gráfico 6, observa-se uma dispersão logarítmica crescente, o que significa que à medida que aumenta a porcentagem de gastos cobertos por seguros catastróficos, os gastos totais também aumentam, mas em menor proporção e de maneira mais gradual do que nos outros dois gráficos. Os espaços presentes nos segundo e terceiro gráficos indicam que há alguns casos em que a porcentagem de gastos cobertos por seguros catastróficos não se correlaciona com os gastos totais.

[2]Encontra-se no apêdice

Os modelos de regressão podem não atender ao objetivo, já que os dados com saúde apresentam valores muito diferentes quanto a magnitude, isto é, pessoas com baixos gastos e pessoas com gastos elevados e nenhuma outra variável foi observada, tais como doenças hereditárias, qualidade de vida, tais como alimentação e hábitos saudáveis.

As análises podem ser melhoradas com um banco de dados que tenha uma evolução mais longa das vidas além de incluir outras variáveis explicativas.

Apêndice

A. Cadeiras de Markov

Tabela 32: Estatística descritivas baseadas em Cadeias de Markov.

	25 anos	30 anos	35 anos	40 anos	45 anos	50 anos	55 anos	60 anos	65 anos
n_0	1.374	310	104	56	59	37	45	44	57
Percentis(5)	566,58	2.072,38	4.397,13	6.609,37	9.048,97	10.607,68	11.934,60	11.909,12	12.316,18
Percentis(10)	895,15	3.427,07	6.830,20	9.831,50	13.312,45	15.594,78	17.547,70	18.493,80	18.693,20
Percentis(15)	1.136,74	4.449,13	8.557,96	12.365,07	16.319,98	19.048,37	21.628,85	22.768,62	23.409,57
Percentis(20)	1.525,50	6.268,47	11.301,34	16.193,88	20.815,60	24.687,96	27.565,74	28.990,94	30.593,04
Percentis(40)	1.891,27	8.288,77	14.609,43	20.669,76	26.228,27	30.771,80	34.699,46	37.178,03	39.090,09
Percentis(50)	2.056,15	9.469,60	16.637,79	23.338,33	29.376,19	34.595,86	39.003,65	42.001,54	44.468,84
Percentis(75)	2.343,09	12.177,58	20.737,21	29.105,46	36.854,46	43.426,86	49.042,70	53.624,93	57.598,55
Percentis(85)	2.433,67	13.075,04	22.631,53	31.457,58	39.826,89	47.115,36	53.668,58	59.088,06	64.240,68
Percentis(95)	2.500,00	13.989,51	24.586,95	34.674,47	43.650,29	52.366,79	60.237,11	67.081,57	73.867,92
Percentis(98)	2.500,00	14.298,58	25.456,69	46.003,51	45.613,35	55.000,48	63.403,48	71.526,45	78.924,27
Máximo	2.500,00	14.860,01	26.947,64	38.371,59	50.042,01	61.331,32	72.807,91	81.323,54	92.628,25
Média	1.859,62	8.978,01	15.776,12	22.272,80	28.321,27	33.457,82	37.855,17	41.064,89	43.937,40
Desvio Padrão	611,92	3.734,34	6.219,20	8.582,72	10.628,11	12.717,10	14.635,67	16.602,97	18.590,75

n0 = total de indivíduos com saldo zero na conta poupança

Tabela 33: Frequência e severidade do seguro catastrófico pelas vidas simuladas durante o período de 41 anos baseadas em Cadeias de Markov

Número de vezes	n	%	% acum	Value (R$)	%	% acum	Média Individual
0	607	6,07%	6,07%	0	0,00%	0,00%	0
1	1256	12,56%	18,63%	16.562.610	3,31%	3,31%	13.187
2	1462	14,62%	33,25%	38.501.061	7,69%	11,00%	26.335
3	1539	15,39%	48,64%	56.458.891	11,28%	22,28%	36.685
4	1337	13,37%	62,01%	68.418.592	13,67%	35,95%	51.173
5	1059	10,59%	72,60%	66.889.913	13,36%	49,32%	63.163
6	843	8,43%	81,03%	60.368.229	12,06%	61,38%	71.611
7	612	6,12%	87,15%	51.846.249	10,36%	71,74%	84.716
8	429	4,29%	91,44%	42.659.966	8,52%	80,26%	99.440
9	302	3,02%	94,46%	32.243.761	6,44%	86,70%	106.767
10	191	1,91%	96,37%	19.493.188	3,89%	90,60%	102.059
11	100	1,00%	97,37%	11.252.039	2,25%	92,85%	112.520
12	93	0,93%	98,30%	11.905.913	2,38%	95,23%	128.021
13	51	0,51%	98,81%	5.852.635	1,17%	96,40%	114.758
14	43	0,43%	99,24%	5.872.259	1,17%	97,57%	136.564
15	27	0,27%	99,51%	4.161.713	0,83%	98,40%	154.138
16	14	0,14%	99,65%	1.561.301	0,31%	98,71%	111.522
17	10	0,10%	99,75%	1.504.172	0,30%	99,01%	150.417
18	14	0,14%	99,89%	2.053.594	0,41%	99,42%	146.685
19	3	0,03%	99,92%	366.848	0,07%	99,50%	122.283
20	4	0,04%	99,96%	653.094	0,13%	99,63%	163.274
21	2	0,02%	99,98%	1.310.683	0,26%	99,89%	655.342
22	1	0,01%	99,99%	327.002	0,07%	99,95%	327.002
24	1	0,01%	100,00%	226.307	0,05%	100,00%	226.307
Total	**10.000**	**100%**		**500.490.020**	**100%**		**50.049**

Tabela 34: Estatística descritivas do Balance da conta poupança, cobertura das contas poupança e pelo seguro catastrófico durante a vida laboral, baseadas em Cadeias de Markov

	Balanço das contas poupança (BCP) depois dos 65 anos	Gastos cobertos pelo Balanço das contas poupança	Gastos cobertos pelo Seguro Catastrófico	Percentagem do total dos gastos cobertos pelo seguro catastrófico
% de valores nulos	0,57	0,00	6,07	6,07
P(5)	12.316,29	28.687,98	1.728,93	0,04
p(10)	18.693,20	34.254,51	3.950,16	0,08
P(15)	23.409,57	38.301,33	6.336,53	0,11
P(25)	30.593,04	44.998,90	11.727,88	0,18
P(40)	39.090,09	52.926,12	21.768,46	0,27
P(50)	44.468,84	58.205,78	29.761,83	0,33
P(60)	49.702,05	63.605,69	40.478,00	0,39
P(75)	57.598,55	72.177,97	64.916,43	0,51
P(85)	64.240,68	79.442,62	95.775,73	0,60
P(90)	68.286,85	84.257,69	126.886,29	0,66
P(95)	73.867,92	91.133,91	189.300,00	0,75
P(97)	77.077,36	95.467,51	245.579,73	0,79
P(98)	78.924,27	97.786,85	280.285,42	0,82
P(99)	82.345,93	101.433,77	345.623,61	0,85
P(995)	84.575,63	102.500,00	409.668,36	0,88
Máximo	92.628,25	102.500,00	1.198.174,14	0,95
Média	43.937,40	58.813,05	53.283,30	0,35
Desvio Padrão	18.590,75	18.830,50	73.468,62	0,22

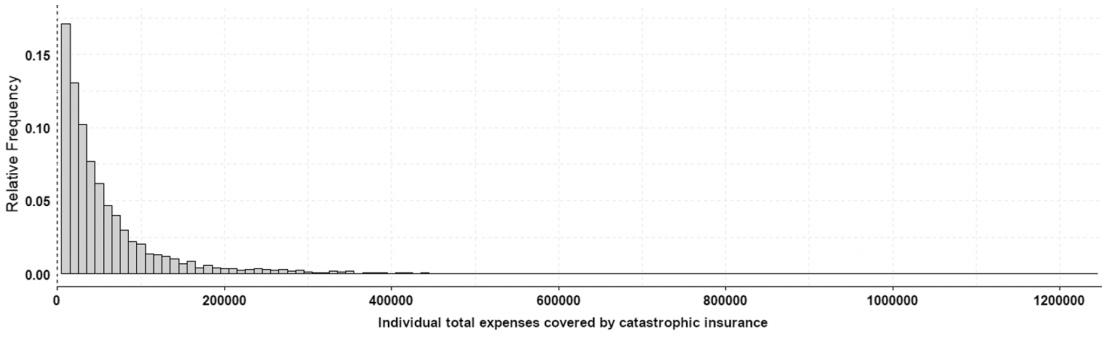

Figura 5: Severidade do Seguro Catastrófico

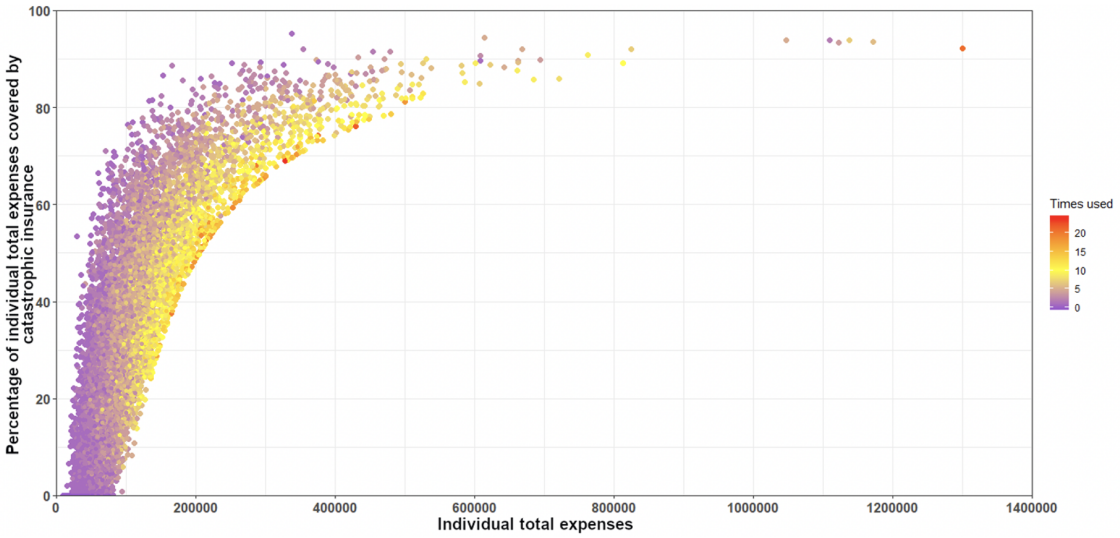

Figura 6: Gráfico de dispersão entre os gastos totais no período de 41 anos de cada vida e o percentagem dessos gastos que foram cobertos por seguro catastrófico

B. Código

1. Para compilar o código, precisamos instalar as seguintes pacotes

   library(tidyverse); library(broom); library(caret); library(glmnet); library(ggthemes); library(magrittr); library(car); library(gdata); library(ggplot2); library(gridExtra); library(ggpubr); library(readxl); library(ISLR); library(leaps); library(psych); library(dplyr); library(sandwich); library(lmtest); library(mltools); library(tseries); library(astsa); library(GGally); library(knitr); library(kableExtra); library(corrplot); library(class); library(texreg); library(xtable); library(MASS); library(data.table); library(e1071); library(kernlab); library(RColorBrewer); library(tree); library(partykit); library(rpart); library(plyr); library(rpart.plot); library(ipred); library(randomForest); library(plotly); library(gvlma); library(datarium); library(DMwR2); library(Metrics); library(FNN); library(rsample);library(outliers); library(scales); library(monomvn); library(yaml); library(AICcmodavg); library(MLmetrics).

2. Chamando os dados do archivo

   dados = read.table(file.path(getwd(),“archivo.csv”),sep=“;”,dec= “,”,header=T)

3. Fazendo um quadro de dados, substituir os valores NAN por zero e observar a quantidade de dados que têm por ano

   dados = as.data.frame(dados)
   dados[is.na(dados)] = 0
   table(dados$year)

4. Selecionando as variáveis dependentes a serem usadas no marco de dados e trocando dados por df para uma melhor manipulação da variável

   dados = dados[,c(2,3,4,5,22,28,16)]
   df = dados

5. Encontrando o número de dados atípcios

```
df = df[order(as.integer(df$expensetot)), ]
sumatoria = 0
df_atipicos = 0
n = 10000 for (i in 1:n) {
   sumatoria[i] = df$expensetot[138900-i+1]
   df_atipicos[i] = df$expensetot[138900-i] / sum(sumatoria)
    if (df_atipicos[i] < 0.05) {
       df_atipicos = df_atipicos[-length(df_atipicos)]
       break
    }
}
print(length(df_atipicos))
df = head(df,-13)
```

6. A função serve para se poder reproduzir os resultados dos geradores de números pseudo-aleatórios, alem disso os dados foram divididos aleatoriamente, com um 80% para treinamentos e um 20% para teste

```
set.seed(12)
training.samples = df$expensetot %>% createDataPartition(p=0.8, list=F)
train.data = df[training.samples, ]
test.data = df[-training.samples, ]
```

7. A seguir temos os modelos de regressão

```
set.seed(12)
train.control = trainControl(method = "cv", number = 5)
```

MODELO LINEAR MúLTIPLA

```
Linear_Model_CV = train(expensetot ~ age + previous1_expensetot +
                                previous2_expensetot,data=train.data,method="lm",
                                trControl = train.control)
predicion_reg_cv_test = predict(Linear_Model_CV$finalModel, newdata = test.data)
reg_rmse_test_1 = sqrt(mean((predicion_reg_cv_test - test.data$expensetot)^2))
```

MODELO SOPORTE VETORIAL

```
Regresor_SVR = train(expensetot ~ age + previous1_expensetot +
                           previous2_expensetot,data=train.data,method="svmLinear",
                           trControl = train.control)
y_svr_test_predict = predict(Regresor_SVR, test.data)
reg_rmse_test_2 = sqrt(mean((y_svr_test_predict - test.data$expensetot)^2))
```

MODELO áRVORES

```
Tree_Model = train(expensetot ~ age + previous1_expensetot +
                   previous2_expensetot, data = train.data, method = "ctree",
                   tuneLength=5,metric = "RMSE", trControl = train.control)
tree.pred = predict(Tree_Model, test.data, type = "raw")
reg_rmse_test_3 = sqrt(mean((tree.pred - test.data$expensetot)^2))
```

MODELO FLORESTA ALETóRIA

```
RF_Model = train(expensetot ~ age + previous1_expensetot + previous2_expensetot,
                 data =train.data,method="rf", tuneLength=5,metric="RMSE",
                 trControl = train.control,n_estimator=500)
rf.pred = predict(RF_Model, test.data)
reg_rmse_test_4 = sqrt(mean((rf.pred - test.data$expensetot)^2))
```

MODELO BAYESIAN RIDGE

```
Bayesian_ridge = train(expensetot ~ age + previous1_expensetot +
                       previous2_expensetot, data = train.data, method="bridge",
                       RJ=TRUE, trControl = train.control)
ridge.pred = predict(Bayesian_ridge, test.data)
reg_rmse_test_5 = sqrt(mean((ridge.pred - test.data$expensetot)^2))
```

MODELO BAYESIAN RIDGE

```
Bayesian_lasso = train(expensetot ~ age + previous1_expensetot +
                       previous2_expensetot, data = train.data, method="blasso",
                       RJ=TRUE, trControl = train.control)
lasso.pred = predict(Bayesian_lasso, test.data)
reg_rmse_test_6 = sqrt(mean((lasso.pred - test.data$expensetot)^2))
```

8. Simulando as 10.000 vidas

```
set.seed(12)
dados_25a30 = filter(df, agerange == '25a30',year >= 2007)
dados_simulados = dados_25a30[sample(nrow(dados_25a30,10000,replace=T),]
```

9. Estimação das 10.000 vidas começando aos 25 e culminando aos 65 anos de vida (ou seja, 41 anos de vida laboral)

```
n = 10000
df_dados = data.frame("Gasto_Anual"= head(dados_simulados$expensetot,n),
                      "Cont_Ind"= rep(2500,n),"Poupança"= rep(0,n),
                      "Seguro_Catas"= rep(0,n),"Individuo"= rep(0,n),
                      "N.Times"=rep(0,n),"Idade"=rep(25,n),"Gasto_Total"=rep(0,n))
new_matrix = matrix(0, ncol = 41, nrow = 14)
new_matrix = as.data.frame(new_matrix)
colnames(new_matrix) = c("25 anos","26 anos","27 anos","28 anos","29 anos",
                         "30 anos", "31 anos","32 anos","33 anos","34 anos",
```

```
                                “35 anos”,“36 anos”,“37 anos”,“38 anos”,“39 anos”,
                                “40 anos”,“41 anos”,“42 anos”,“43 anos”,“44 anos”,
                                “45 anos”,“46 anos”,“47 anos”,“48 anos”,“49 anos”,
                                “50 anos”,“51 anos”,“52 anos”,“53 anos”,“54 anos”,
                                “55 anos”,“56 anos”,“57 anos”,“58 anos”,“59 anos”,
                                “60 anos”,“61 anos”,“62 anos”,“63 anos”,“64 anos”,
                                “65 anos”)
rownames(new_matrix)= c(“no”,“P(5)”,“P(10)”,“P(15)”,“P(25)”,“P(40)”,“P(50)”,“P(75)”
                                “P(85)”,“P(95)”,“P(98)”,“Maximum”,“Mean”,“SD”)
new_matrix_pre = head(dados_simulados,n)
new_matrix_pre$age = rep(25,n)
row.names(new_matrix_pre) = 1 : n
new_matrix_pre = new_matrix_pre[,c(1,3,5,6,7)]
head(new_matrix_pre)
for (years in 1:41){
    df_dados$Idade[df_dados$Idade<=65] = 24 + years
    if (years > 1){
        for (fila_pre in 1:n){
            new_matrix_pre[fila_pre,1] = dados_simulados[fila_pre,1]
            new_matrix_pre[fila_pre,2] = new_matrix_pre[fila_pre,2] + 1
            new_matrix_pre[fila_pre,4] = new_matrix_pre[fila_pre,3]
            new_matrix_pre[fila_pre,3] = new_matrix_pre[fila_pre,5]
            new_matrix_pre[fila_pre,5] = predict(Linear_Model,new_matrix_pre[fila_pre,])
        }
        df_dados[,1] = new_matrix_pre$expensetot
    }
    for (fila in 1:n){
        df_dados[fila,8] = df_dados[fila,8] + df_dados[fila,1]
        if (df_dados[fila,1] <= 5000){
            df_dados[fila,3] = df_dados[fila,3] + (df_dados[fila,2] - df_dados[fila,1])
            if (df_dados[fila,3] <= 0){
                df_dados[fila,4] = df_dados[fila,4] + abs(df_dados[fila,3])
                df_dados[fila,3] = 0
                df_dados[fila,6] = df_dados[fila,6] + 1
            }
        }
        else if (df_dados[fila,1] > 5000){
            df_dados[fila,3] = df_dados[fila,3] + df_dados[fila,2]
            df_dados[fila,3] = df_dados[fila,3] - 5000
            df_dados[fila,4] = df_dados[fila,4] + df_dados[fila,1] - 5000
            df_dados[fila,6] = df_dados[fila,6] + 1
            if (df_dados[fila,3] <= 0){
                df_dados[fila,4] = df_dados[fila,4] + abs(df_dados[fila,3])
                df_dados[fila,3] = 0
            }
        }
```

```
    }
    print(df_dados)
    Tabela = df_dados %>% filter(Poupança > 0)
    new_matrix[1,years] = n - dim(Tabela)[1]
    if (new_matrix[1,years] < n){
       new_matrix[2:11,years] = as.vector(quantile(Tabela$Poupança,c(.05,.10,.15,.25,
                                                                     .40,.50,.75,.85,.95,.98)))
       new_matrix[12,years] = max(Tabela$Poupança)
       new_matrix[13,years] = mean(Tabela$Poupança)
       new_matrix[14,years] = sd(Tabela$Poupança)
    }
    else {
       new_matrix[2:14,years] = 0
    }
}

new_matrix = new_matrix[,c(1,6,11,16,21,26,31,36,41)]
new_matrix = as.data.frame(new_matrix)
new_matrix
table(df_dados$N.Times)
```

Bibliografia

[1] Pedro A. Morettin Julio M. Singer, *Introdução á Ciência de Dados*, São Paulo: Universidad de São Paulo 2020.

[2] Michael W. BerryAzlinah MohamedBee Wah Yap, *Supervised and Unsupervised Learning for Data Science*, Springer 2020.

[3] Christopher Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[4] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference and Prediction, New York: Springer New York, 2008

[5] Raschka S, Python Machine Learning, 2nd ed. Packt Publishing Ltd, Birmingham, 2017.

[6] James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning. Vol 103. New York, NY: Springer New York, 2014.

[7] Kuhn M, Johnson K. Applied Predictive Modeling. (Springer, ed.). New York, NY: Springer New York; 2013.

[8] A. J. Smola and B. Schölkop. A tutorial on support vector regression. Statistics and Computing, 14:199–222, 2004.

[9] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. Springer, 2001.

[10] James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning. Vol 103. New York, NY: Springer New York; Chapter 6: Exercise 7, 2014.

[11] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. Technometrics, 12, 55–67, 1970.

[12] Tibshirani, R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B (methodological), 58, 267–288. 2006.

[13] Medeiros, M. C. Machine Learning Theory and Econometrics. Lecture Notes. 2019.

[14] Bühlmann, P. and van de Geer, S. Statistics for High-Dimensional Data. Berlin: Springer. 2011

[15] S PATRO, K Sahu. Normalization: A Preprocessing Stage. Department of CSE &, VSSUT, Burla, Odisha, India. 2015

[16] G. Bhattarai, Golden Valley. Understanding the Outliers in Healthcare Expenditure Data, PhD, OptumHealth, Golden Valley MN, 2013.

[17] Anderson TW, Goodman LA. Statistical Inference about Markov Chains. Ann. Math. Statist 28(1):89-110. doi:10.1214/aoms/1177707039, 1957.