UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS MATEMÁTICAS E DA NATUREZA
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE MÉTODOS ESTATÍSTICOS

DAVID GABRIEL PEÇANHA PINHEIRO

Modelos para seleção de variáveis aplicados à subscrição de seguros

RIO DE JANEIRO
2024

DAVID GABRIEL PEÇANHA PINHEIRO

Modelos para seleção de variáveis aplicados à subscrição de seguros

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Métodos Estatísticos da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção dos graus de Bacharel em Ciências Atuariais e Bacharel em Estatística.

Orientador: Carlos Tadeu Pagani Zanini (UFRJ)

RIO DE JANEIRO
2024

CIP - Catalogação na Publicação

P654m

Pinheiro, David
Modelos para seleção de variáveis aplicados à subscrição de seguros / David Pinheiro. -- Rio de Janeiro, 2024.
46 f.

Orientador: Carlos Zanini.
Trabalho de conclusão de curso (graduação) - Universidade Federal do Rio de Janeiro, Instituto de Matemática, Bacharel em Ciências atuariais, 2024.

1. risco de subscrição. 2. modelos de regressão linear. 3. seguro patrimonial catastrófico. 4. seleção de variáveis. 5. spike-and-slab. I. Zanini, Carlos, orient. II. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

DAVID GABRIEL PEÇANHA PINHEIRO

Modelos para seleção de variáveis aplicados à subscrição de seguros

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Métodos Estatísticos da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção dos graus de Bacharel em Ciências Atuariais e Bacharel em Estatística.

Aprovado em ___ de _____ de _____

BANCA EXAMINADORA:

Carlos Tadeu Pagani Zanini
Doutorado (UFRJ)

Hugo Tremonte de Carvalho
Doutorado (UFRJ)

João Batista de Morais Pereira
Doutorado (UFRJ)

Àqueles que, na busca pelo autodesenvolvimento,
harmonizam as vozes internas com as externas.

AGRADECIMENTOS

A graduação foi por muito a experiência mais avassaladora que vivi até então. Tê-la feito em tenra idade foi paradoxalmente uma dádiva e o maior dos desafios. A quantidade de desenvolvimento pessoal, acadêmico e profissional realizado durante esse período expandiu as noções prévias do que entendia como evolução. Muitas das vezes, senti como se meu cérebro estivesse fisicamente dilatando e eu que sempre fui ávido por conhecimento, desliguei tudo que interpretei como ruído e me agarrei a essa sensação. Antes de agradecer, queria pedir desculpas a todos que se sentiram colocados em segundo plano por mim em algum momento enquanto eu aprendia a lidar com minhas multitudes.

De maneira geral, gostaria de agradecer a cada pessoa que de alguma maneira me atravessou em toda minha vida. Umas mais e outras menos, todas foram primordiais e tiveram algum papel no efeito borboleta que me dirigiu à vida que levo hoje. Nominalmente, agradeço:

À minha mãe Zaira Peçanha, por ser a maior apoiadora da minha vida. Ao seu jeito, você me deu tudo que eu nunca soube que precisava. Sua bondade sem limites permeia todos os passos que dou.

Ao meu irmão João Daniel, por ter me ensinado o amor em sua forma mais intrínseca. Consigo ver o melhor da minha mãe e de mim em você e ainda assim você é tão único que eu não poderia ter mais orgulho.

À minha madrinha Valéria Sampaio e tia Renata Peçanha, por serem meus referenciais de sentimento e por sempre me fazerem sentir pertencido. As emoções que aprendi com vocês para sempre guiarão minha vida.

Ao meu amigo Edson Lucas, pelas confidências diárias, pela companhia e troca infinita. Você me ensinou que família é sobre quem nos escolhe.

Ao meu namorado André Rodrigues, pela revolução que inspirou em minha vida. Com você, redescobri o afeto, a paixão e tudo que faz a vida valer a pena.

Ao meu pai Jorge Pinheiro, pelo apoio, amor e segurança a mim oferecidas; ao meu padrinho Iuri Sampaio, por todas as vezes que você compareceu a uma formatura, a um aniversário e sempre fez questão de me prestigiar; às minhas avós Creuzeli Peçanha e Marly Pinheiro, bem como ao meu finado avô Laercio Santos, por tantas vezes que me senti cuidado, amado e entendido; às minhas amigas Isabelle Marques e Laura Frantelmo pelos dias de diversão, conversas edificantes e acolhimento; à minha terapeuta Adriane Pereira que segurou minha mão enquanto eu reaprendia a enxergar o mundo.

A todos amigos que fiz na faculdade, especialmente Diene Cordeiro, que me acompanhou em viagens de ônibus, em todas salas de aulas e em todos os trabalhos em grupo. Na frustração e no sucesso, você é a pessoa que mais entende o que foi minha experiência com a graduação.

Ao corpo docente do Instituto de Matemática da UFRJ, em especial ao meu orientador Carlos Zanini, que relevou várias sextas-feiras que não pude comparecer às nossas reuniões devido aos infinitos compromissos do trabalho e foi crucial para produção deste estudo, para além do aspecto acadêmico.

À minha sede por mudança, ao meu inconformismo com a infelicidade e à minha ética, por serem valores fundamentais que me mantêm no chão quando as vozes decidem me levitar.

"I sometimes fear that I am misunderstood.
It is simply because what I want to say,
what I need to say, won't be heard.
Heard in a way I so rightfully deserve.
What I choose to say is of so much substance
That people just won't understand the depth of my message.
So my voice is not my weakness,
It is the opposite of what others are afraid of.
My voice is my suit and armor,
My shield, and all that I am.
I will comfortably breathe in it, until I find the moment to be silent.
I live loudly in my mind, so many hours of the day.
The world is pin drop sound compared to the boom
That thumps and bumps against the walls of my cranium.
I live it and love it and despise it and I am entrapped in it.
So being misunderstood, I am not offended by the gesture, but honored.
If they let us..."

Chloë Mitchell

RESUMO

O seguro patrimonial catastrófico é um ramo de seguros com crescimento célere em todo mundo frente às incertezas oriundas das mudanças climáticas. Um dos riscos mais proeminentes para seguradoras que assumem risco patrimonial catastrófico ocorre durante o processo de avaliação para determinar a aceitação de segurados, chamado de risco de subscrição. Uma seguradora com carteira pouco diversificada tem maiores chances de sofrer perdas econômicas irreversíveis e, portanto, urge a necessidade de uma subscrição de seguros adequada. Neste trabalho, a partir de uma base de dados de apólices patrimonais sob riscos catastróficos de uma seguradora chilena, utilizaremos modelos de regressão linear múltipla, com penalização Lasso, Ridge, *Elastic Net* ingênuo e o modelo Bayesiano *Spike-and-Slab* com o objetivo de selecionar e quantificar o efeito de covariáveis descritivas, em sua maioria qualitativas nominais, sobre o valor líquido de sinistro a fim de auxiliar a subscrição de seguros da seguradora.

Palavras-chave: risco de subscrição. modelos de regressão linear. seguro patrimonial catastrófico. seleção de variáveis. *spike-and-slab*.

ABSTRACT

Catastrophic property insurance is a branch of insurance experiencing rapid growth worldwide in face of uncertainties stemming from climate change. One of the most prominent risks for insurers assuming catastrophic property risk happens during the process of evaluation to determine acceptance of insureds, called underwriting risk. An insurer with a poorly diversified portfolio is more likely to suffer irreversible economic losses, thus emphasizing the need for adequate insurance underwriting. In this paper, based on a database of property policies under catastrophic risks from a Chilean insurer, we will use penalized multiple linear regression models, Lasso, Ridge, naive Elastic Net, and the Bayesian Spike-and-Slab model with the aim of discussing the effect of descriptive covariates, mostly nominal qualitative, on the net loss value in order to support the underwriting sector of the insurer.

Keywords: underwriting risk. linear regression models. catastrophic property insurance. variable selection. spike-and-slab.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O seguro é um mecanismo de proteção financeira que visa transferir o risco de um evento incerto de uma pessoa ou empresa para uma seguradora, mediante o pagamento de um prêmio. A precificação de seguros é um processo fundamental para as seguradoras, pois é por meio dele que se determina o valor do prêmio a ser cobrado do segurado para que a seguradora possa assumir o risco. Nesse sentido, a precificação de seguros envolve uma série de variáveis que são importantes para se chegar a um preço justo e adequado: há de se estimar a perda esperada para uma apólice com determinadas características e assim calcular o preço do seguro.

Nesse sentido, surge a subscrição de seguros, que é o processo de avaliação de riscos para determinar se a seguradora irá segurar uma pessoa física ou jurídica, o preço do seguro e as condições da apólice. A subscrição é uma das etapas primordiais da atividade seguradora e é responsável por grande parte do risco admitido por uma seguradora. Urge-se, pois, a necessidade de construir ferramentas que auxiliem os subscritores, trabalhadores da seguradora responsáveis pela aceitação do risco, a identificar as variáveis que mais são impactantes para subscrição.

De acordo com a Swiss Re (2021), um dos ramos de seguro mais importantes mundialmente é o ramo patrimonial, cujo prêmio de seguro espera-se crescer de US$450 bilhões em 2020 para US$1,3 trilhão até 2040. O seguro patrimonial é um tipo de cobertura de seguro que protege indivíduos ou empresas contra perdas financeiras relacionadas à sua propriedade e é o foco da análise deste trabalho. Ele fornece cobertura para danos ou perdas em edifícios, estruturas e pertences pessoais causados por riscos cobertos, como incêndio, roubo, vandalismo, desastres naturais e outros eventos especificados. Consoante dados do SES SUSEP (2023), Sistema de Estatísticas da Superintendência Nacional de Seguros Privados, as seguradoras emitiram no Brasil, no ano de 2022, R$24 bilhões em prêmios de seguro associados ao ramo patrimonial, representando cerca de 14,9% do mercado brasileiro em termos de prêmio emitido.

O seguro de propriedade geralmente consiste em dois tipos principais de cobertura:

1. Seguro Patrimonial Residencial: Esse tipo de seguro é projetado para indivíduos que possuem uma casa. Ele fornece cobertura para a própria residência, outras estruturas na propriedade (como garagens ou galpões), pertences pessoais e proteção de responsabilidade por lesões ou danos que ocorram na propriedade.

2. Seguro Patrimonial Comercial: Esse seguro é destinado a empresas que possuem ou alugam propriedades comerciais. Ele oferece cobertura para o prédio, equipamentos, estoque e outros ativos de propriedade do negócio. Também fornece cobertura de responsabilidade por lesões ou danos que ocorram nas instalações.

As apólices de seguro patrimonial geralmente possuem diferentes níveis de cobertura, franquias e prêmios, dependendo do tipo e valor da propriedade segurada, bem como da localização e riscos específicos envolvidos.

Desta maneira, podemos perceber que o seguro patrimonial possui diversas variáveis que podem impactar a perda esperada para a seguradora, e do ponto de vista estatístico, a grande maioria é de natureza qualitativa ao invés de quantitativa. Isso adiciona uma camada a mais de complexidade à seleção do modelo para identificação de variáveis que melhor discriminam sinistros, pois a maioria das técnicas de seleção de variáveis conhecidas são ajustadas para variáveis quantitativas.

No escopo deste trabalho, serão utilizados dados reais fornecidos por uma empresa chilena, estando anonimizada ou excluída toda e qualquer informação privada. Serão aplicados a esses dados os modelos de regressão linear múltipla, Lasso, Ridge, *Elastic Net* ingênuo e *Spike-and-Slab* Bayesiano com foco na identificação das variáveis mais relevantes em termos de perda financeira para a seguradora. O processo envolverá diversas etapas, como limpeza e transformação de dados, anáise exploratória, ajuste dos modelos e as devidas comparações.

O texto do trabalho está estruturado da seguinte maneira: o Capítulo 2 apresenta a descrição da base de dados utilizada, incluindo contexto mais detalhado acerca da origem dos dados, bem como a análise exploratória dos dados. No Capítulo 3 são apresentados a inferência clássica e Bayesiana e descritos todos os modelos de regressão a serem comparados em relação à acurácia preditiva. Já no Capítulo 4, os resultados da aplicação dos modelos são apresentados. Por fim, o Capítulo 5 amarra todo o conteúdo visto e conclui acerca das técnicas discutidas de regressão linear e sugere implementações para trabalhos futuros.

2 BASE DE DADOS

2.1 DESCRIÇÃO

A base de dados utilizada neste trabalho consiste de seguros patrimoniais de cobertura de catástrofes do mercado chileno de seguros.

Uma das coberturas mais impactantes do seguro patrimonial é a cobertura contra catastrófes. A cobertura de catástrofe no seguro patrimonial refere-se à proteção oferecida aos bens e propriedades contra danos causados por eventos catastróficos, como terremotos, furacões, incêndios em larga escala, inundações, entre outros. Esses eventos são caracterizados por sua natureza imprevisível e potencialmente devastadora, causando grandes prejuízos financeiros.

As apólices de seguro patrimonial normalmente incluem coberturas básicas para riscos comuns, como incêndios e roubos. No entanto, essas apólices geralmente não oferecem proteção adequada contra eventos catastróficos, pois os danos e prejuízos associados a essas situações excepcionais são muito maiores e podem ultrapassar os limites de cobertura das apólices tradicionais.

Para abordar essa falta de proteção, as seguradoras oferecem coberturas de catástrofe como uma extensão do seguro patrimonial. Essas coberturas são projetadas especificamente para proteger os segurados contra os riscos e danos resultantes de eventos catastróficos. Os termos e condições dessas coberturas variam de acordo com cada apólice e seguradora, e podem incluir franquias mais altas e limites de cobertura específicos para eventos catastróficos.

O Chile está localizado no chamado anel de fogo do Pacífico, e por isso enfrenta perigos naturais como terremotos, tsunamis, aluviões, deslizamentos de terra e erupções vulcânicas e, mais recentemente, incêndios florestais. Dessa maneira, há uma grande necessidade de compra de seguro contra catástrofe, o que pode ser percebido por meio da participação do mercado do principal risco catastrófico, o terremoto. De acordo com o CMF (2021) (Comissão para o Mercado Financeiro, em tradução livre do espanhol), esse risco sozinho representava US$666,9 milhões em prêmio emitido, 65,1% de todo o mercado segurador não vida do Chile em dados datados em 31 de março de 2021.

A base de dados foi fornecida por uma seguradora chilena com informações relacionadas a apólices de seguros patrimonais para catástrofes com vigência de janeiro de 2017 a junho de 2021. Cada linha contém a informação agregada de uma única apólice, ou seja, se um segurado avisou à seguradora mais de um sinistro, os valores das perdas serão somados. Na Tabela 1, constam o nome original, a interpretação e o tipo de cada variável presente na base de dados referente a cada imóvel segurado.

A base contém 12.118 linhas únicas, representando cada uma das apólices da segura-

dora. A variável resposta é o valor do sinistro líquido da apólice, representada pela coluna *LOSSES*, enquanto as demais 19 variáveis serão todas consideradas covariáveis para os modelos aplicados neste trabalho. A variável de interesse encontra-se originalmente na moeda de código CLF, chamada de unidade de fomento, em tradução livre do espanhol, e é uma unidade de conta não circulante utilizada no Chile. A taxa de câmbio entre essa moeda e o peso chileno, moeda circulante do Chile, é ajustada diariamente para que o poder de compra da unidade de fomento permaneça constante diariamente quando há baixa inflação. Para referência, em 14 de março de 2024, uma unidade de fomenta valia 39,18 dólares americanos.

Tabela 1 – Variáveis presentes nos dados originais

Nome original	Interpretação	Tipo
CITY	Cidade no Chile	Qualitativa nominal
STATE	Região no Chile	Qualitativa nominal
LATITUDE	Latitude	Quantitativa contínua
LONGITUDE	Longitude	Quantitativa contínua
ZONA_CRESTA	Zona CRESTA	Qualitativa nominal
OCCTYPE	Tipo de ocupação	Qualitativa nominal
SIC_CODE	Código SIC	Qualitativa nominal
NUMEDIFICIO	N° de prédios	Qualitativa nominal
NUMSTORIES	N° de andares	Qualitativa nominal
SHAPECONF	Configuração da estrutura	Qualitativa nominal
STORYPROF	Presença ou tipo de piso frágil	Qualitativa nominal
OVERPROF	Presença ou tipo de sacada	Qualitativa nominal
TORSION	Presença ou tipo de resistência	Qualitativa nominal
CLOUDING	Presença ou tipo de revestimento	Qualitativa nominal
SHORTCOL	Presença ou tipo de coluna curta	Qualitativa nominal
STRUCTUP	Presença ou tipo de estruturas	Qualitativa nominal
ENGFOUND	Presença ou tipo de fundações	Qualitativa nominal
POUNDING	Presença ou tipo de edifícios laterais	Qualitativa nominal
BASEISOL	Presença ou tipo de isolamento	Qualitativa nominal
LOSSES	Valor do sinistro líquido da apólice	Quantitativa contínua

Algumas variáveis desta base são específicas para a análise de modelos de catástrofe complexos, como Zona CRESTA (Avaliação de Risco de Catástrofes e Padronização de Acúmulos de Riscos, em tradução livre do inglês) e Código SIC (Classificação Industrial Padrão, em tradução livre do inglês).

Segundo CRESTA (2023), a classificação foi criada com objetivo de estabelecer um sistema global uniforme para o controle de riscos de acumulação de desastres naturais - especialmente terremotos, tempestades e inundações. Essas zonas de risco são essencialmente baseadas na atividade sísmica observada e esperada, bem como em outros desastres naturais, como secas, inundações e tempestades. As zonas CRESTA consideram a distribuição de valores segurados dentro de uma região ou país para facilitar a avaliação de riscos.

Consoante o Governo do Reino Unido (2024), o código SIC diz respeito a um sistema de classificação de indústrias por meio de um código de quatro dígitos, como um método de padronização da classificação industrial para fins estatísticos em diversas agências. Estabelecido nos Estados Unidos em 1937, é utilizado por agências governamentais para classificar áreas industriais.

Ademais, a base de dados foi analisada utilizando a linguagem de programação *Python*, que segundo KUHLMAN (2011), é uma linguagem de programação de alto nível lançada por Guido van Rossum em 1991, interpretada de *script*, imperativa, orientada a objetos, funcional, de tipagem dinâmica e forte. Atualmente, a linguagem *Python* possui um modelo de desenvolvimento comunitário, aberto e gerenciado pela organização sem fins lucrativos *Python Software Foundation*.

O pacote utilizado para a leitura e limpeza dos dados foi o *pandas*, que de acordo com Pandas (2023) é uma biblioteca para manipulação e análise de dados. Em particular, a biblioteca *pandas* oferece estruturas e operações para manipular tabelas numéricas e séries temporais.

Todas as variáveis qualitativas nominais foram transformadas em *dummies* para permitir a identificação dos efeitos de cada uma de suas categorias pelos modelos. As covariáveis quantitativas contínuas, latitude e longitude, foram padronizadas pois a magnitude distinguia das demais variáveis *dummies* a fim de evitar problemas numéricos nos procedimentos de inferência.

2.2 ANÁLISE EXPLORATÓRIA

O primeiro passo para entender a influência das covariáveis na variável resposta é performar uma análise exploratória nos dados. Todas covariáveis, à exceção da latitude e longitude de cada apólice, têm caratér qualitativo nominal, ou seja, medidas clássicas de análise exploratória como média, variância, moda e quantis não podem ser calculadas. Assim, podemos explorar como medida de resumo outros recursos, como *boxplots* das covariáveis frente à resposta, frequência e severidade das observações de cada covariável, etc.

Primeiro, estudemos a quantidade de apólices da nossa carteira de estudo e a população de cada uma das 16 regiões do Chile ao verificar a tabela 2, que contém o número de apólices e habitantes por região.

Tabela 2 – Quantidade de apólices e habitantes por região

Região	**Número de Apólices**	**Número de habitantes**
Antofagasta	361	607.534
Araucanía	511	957.224
Arica y Parinacota	103	226.068
Atacama	187	286.168
Aysén	68	103.158
Bío-bío	791	1.538.194
Coquimbo	450	757.586
Lagos	769	828.708
Magallanes	97	166.533
Maule	534	1.044.950
Metropolitana de Santiago	5.677	7.112.808
Ñuble	222	480.609
O'Higgins	476	914.555
Ríos	667	384.837
Tarapacá	189	330.558
Valparaíso	1.016	1.815.902
Total	**12.118**	**17.555.392**

Observando os dados de quantidade de apólices por região bem como os dados de população oriundos do CENSO (2017) chileno, podemos observar que existe uma aparente relação positiva entre o número de habitantes e a quantidade de apólices por região. Isso faz sentido pensando no fato de que temos uma carteira massificada de seguro patrimonial, onde a maioria das apólices são de cunho residencial e comercial, assim onde mais há pessoas, haverá mais apólices de seguros subscritas, se a seguradora tiver presença de mercado em todo o país.

Outrossim, podemos iniciar o estudo da relação entre a covariável de localização e a variável de interesse. Primeiramente, vejamos o mapa do Chile e sua divisão por região na Figura 1.

Figura 1 – Mapa do Chile dividido por região

Em sequência, vejamos os gráficos de mapa de cada apólice versus o log do seu respectivo valor líquido de sinistro, na Figura 2, e também agrupamento por município do log do valor líquido de sinistro, na Figura 3.

Existe maior concentração de sinistros nas regiões com mais apólices, como a região metropolitana de Santiago e Valparaíso. Regiões menos habitadas consequentemente têm menos apólices subscritas e sinistros menos vultosos, como nas regiões no extremo sul, Magallanes e Aysén.

Os gráficos acima foram produzidos utilizando a biblioteca *geopandas*. De acordo com GeoPandas (2024), essa biblioteca é um projeto para adicionar suporte a dados geográficos a objetos do *pandas*.

Além da localidade, uma variável tida como de suma importância para avaliação do potencial sinistral é o tipo de ocupação referente à apólice,. A variável de tipo de ocupação presente na base de dados, chamada originalmente de OCCTYPE, que fomenta todo trabalho é uma classificação feita pela Moody's RMS, empresa líder no mundo em riscos catastróficos e discorre sobre a natureza/finalidade da propriedade segurada. Dessa maneira, estudemos a quantidade de cada tipo de ocupação e seu respectivo impacto na

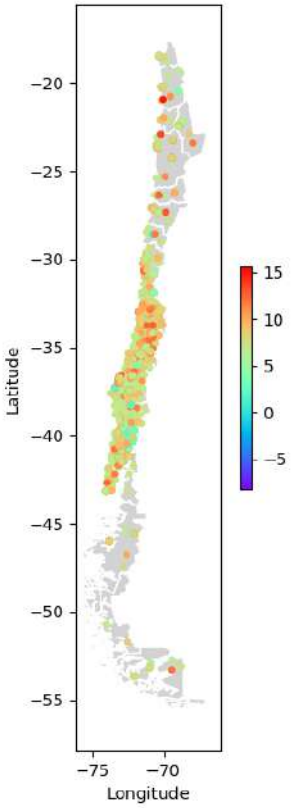

Figura 2 – Sinistro líquido por localidade

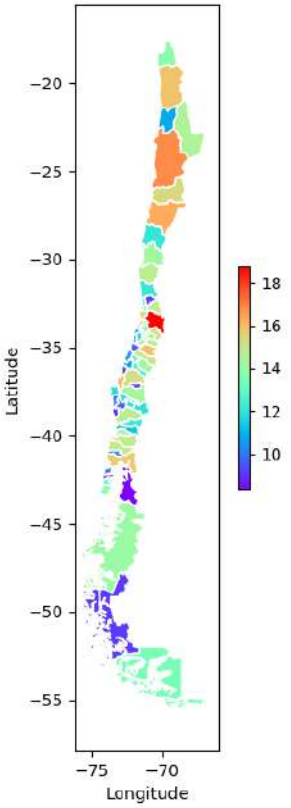

Figura 3 – Sinistro líquido por município

variável de interesse a partir da Tabela 3.

O tipo de ocupação mais frequente, com 5387 apólices subscritas, é "Telefone e Telégrafo", que são seguros patrimoniais feitos para empresas de telefonia, geralmente para torres de telefonia celular. Essas apólices na maioria das vezes têm importância segurada baixa, mas são precificados como um risco volátil, devido à alta sensibilidade a raios, chuvas fortes, enchentes, etc. Nos dados deste trabalho, embora esse tipo de ocupação seja o mais frequente, é apenas o 13° mais severo. Isso pode significar que a empresa seguradora notou sua baixa severidade e aumentou a subscrição desse risco.

O resto do top 5 (varejistas, serviços profissionais técnicos-empresariais, processamento de alimentos e medicamentos e atacadistas) é composto de seguros frequentes e também severos, pois possuem importância segurada de alta magnitude.

Apólices com tipo de ocupação "Mar/Água", por exemplo, têm apenas a 24ª maior frequência, mas são as 7ª maiores contribuidoras para a soma de sinistros. Embora possivelmente enviesado pelo pequeno número de dados, podemos desconfiar que esse tipo de apólice tem grande impacto nos sinistros e deve ser um potencial tipo de ocupação para a seguradora revisar durante o período de subscrição.

A próxima variável a ser escrutinada é o número de andares. Como fizemos com as

Tabela 3 – Quantidade de apólices e log de soma de sinistros por tipo de ocupação

Tipo de Ocupação	**Quantidade**	**Ordem**	**Sinistros**	**Ordem**
Telefone e Telégrafo	5387	1	15,66	13
Varejistas	1654	2	17,05	3
Profissionais Técnicos e Empresariais	998	3	17,46	2
Alimentos e Medicamentos	876	4	17,68	1
Atacadistas	721	5	16,79	5
Habitação Permanente (multifamiliar)	705	6	16,97	4
Educação	421	7	16,38	8
Fabricação Pesada e Montagem	221	8	15,36	16
Habitação Permanente (unifamiliar)	220	9	14,98	17
Agricultura	150	10	16,29	10
Serviços de Saúde	125	11	15,70	12
Processamento de Metais e Minerais	109	12	15,61	15
Religião e ONGs	94	13	14,64	20
Entretenimento e Recreação	86	14	14,92	18
Comércio Geral	76	15	14,64	19
Fabricação Leve e Montagem	62	16	16,36	9
Elétrico	49	17	14,05	22
Mineração	34	18	16,71	6
Alojamento Temporário	34	18	15,65	14
Processamento de Produtos Químicos	24	20	14,10	21
Serviços Pessoais e de Reparação	22	21	12,55	28
Habitação Institucional em Grupo	12	22	13,55	24
Petróleo	10	23	12,34	30
Mar/Água	9	24	16,59	7
Serviços Gerais	6	25	13,24	25
Água	5	26	15,95	11
Aéreo	2	27	12,68	26
Comunicação (Rádio e TV)	2	27	13,64	23
Diversos	2	27	11,35	31
Rodovia	1	30	12,49	29
Controle de Inundações	1	30	12,65	27

Ordem indica a posição na ordenação de cada variável classificados do maior para o menor

demais variáveis, podemos estudar a quantidade de apólices com diferentes números de andares. Nesse sentido, verifiquemos a Tabela 4.

Como podemos notar, existe uma concentração de 82% de apólices com construções seguradas que possuem apenas o térreo, 4% de ocupações com 1 andar, 3% com 2 andares, e as construções com outras configurações são menos ocorrentes. No que tange ao comportamento dessa variável frente à variável resposta, podemos construir um *boxplot* do log da soma de sinistros por número de andares, representado pela Figura 4.

Embora um possível comportamento para esse *boxplot* era o log da soma dos sinistros

Tabela 4 – Quantidade de apólices por número de andares

Número de andares	Quantidade
0	9950
1	473
2	365
3	121
4	129
5	163
6	27
7	43
8	89
9	112
10	64
11	28
12	173
13	21
14	38
15	22
16	29
17	27

Número de andares	Quantidade
18	27
19	13
20	124
21	15
22	12
23	11
24	11
25	9
26	2
27	6
28	2
29	1
30	2
31	1
32	4
33	2
55	1
60	2

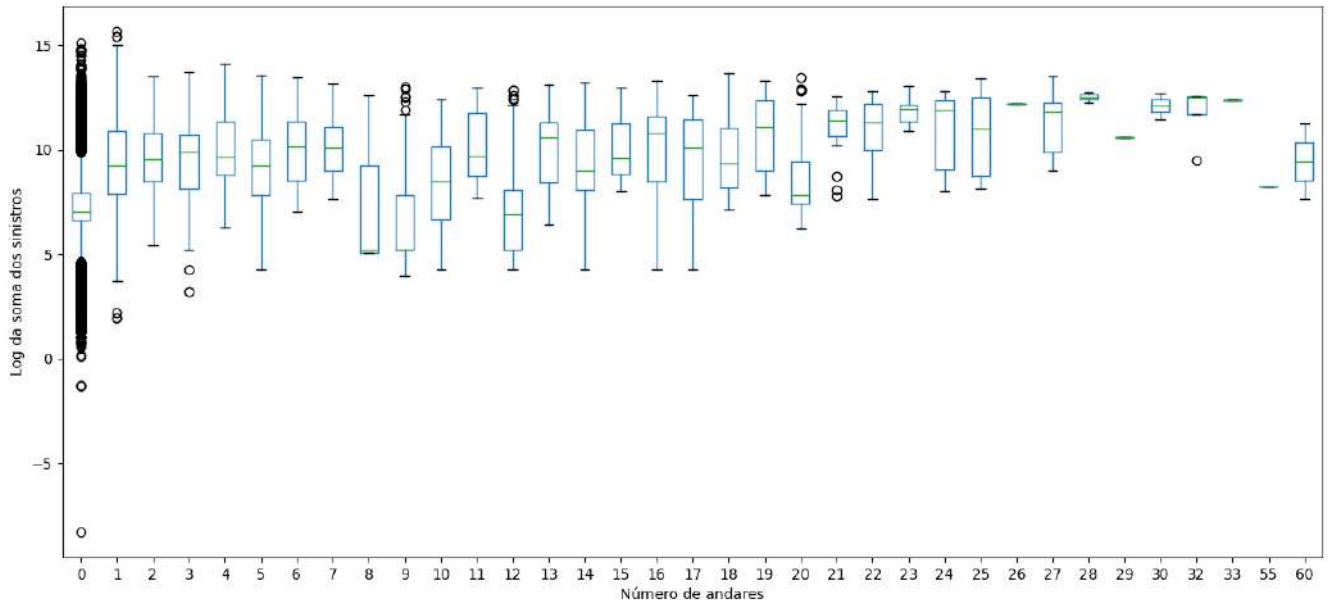

Figura 4 – *Boxplot* do log da soma de sinistros por número de andares

tivesse comportamento crescente à medida que o número de andares crescesse, isso não ocorre. Não existe comportamento qualquer aparente, mesmo que geralmente prédios com mais andares tenham importância segurada maior. O motivo dessa falta de padrão característico claro pode ser devido à incerteza inerente aos eventos, ao baixo número

de apólices seguradas com andares mais altos ou mesmo até ao perfil de subscrição da seguradora, que pode simplesmente subscrever apenas prédios com mais andares que sejam mais antigos ou em localidades mais baratas a fim de reduzir sua exposição.

A mesma análise pode ser replicada para a variável de presença ou tipo de piso frágil (tradução livre do espanhol, *piso débil*). Segundo LANDERO (2003), piso frágil diz respeito a construções cujo térreo é menos rígido que andares superiores como podemos observar na Figura 5, onde o segundo andar, de alvenaria, está em grande parte apoiado em uma estrutura de vidro, no primeiro andar.

Figura 5 – Exemplo de construção com piso frágil

Essa variável está inclusa na base de dados, com a seguinte codificação, "0" significa que não há presença de piso frágil, "1" indica nível leve ou parcial de piso frágil, "2" indica nível grave ou total de piso frágil. Dessa forma, vejamos a Tabela 5 com a quantidade de apólices por tipo de piso frágil.

Tabela 5 – Quantidade de apólices por presença ou tipo de piso frágil

Piso frágil	**Quantidade**
0	11989
1	91
2	10

Como vimos anteriormente, 9950 são construções térreas, então elas já automaticamente não poderiam ser classificadas como piso frágil. Ainda assim, 2039 construções com andares não tem piso frágil, compondo absoluta maioria. Em sequência, podemos construir o *boxplot* do log da soma de sinistros por presença ou tipo de piso frágil, representado pela Figura 6.

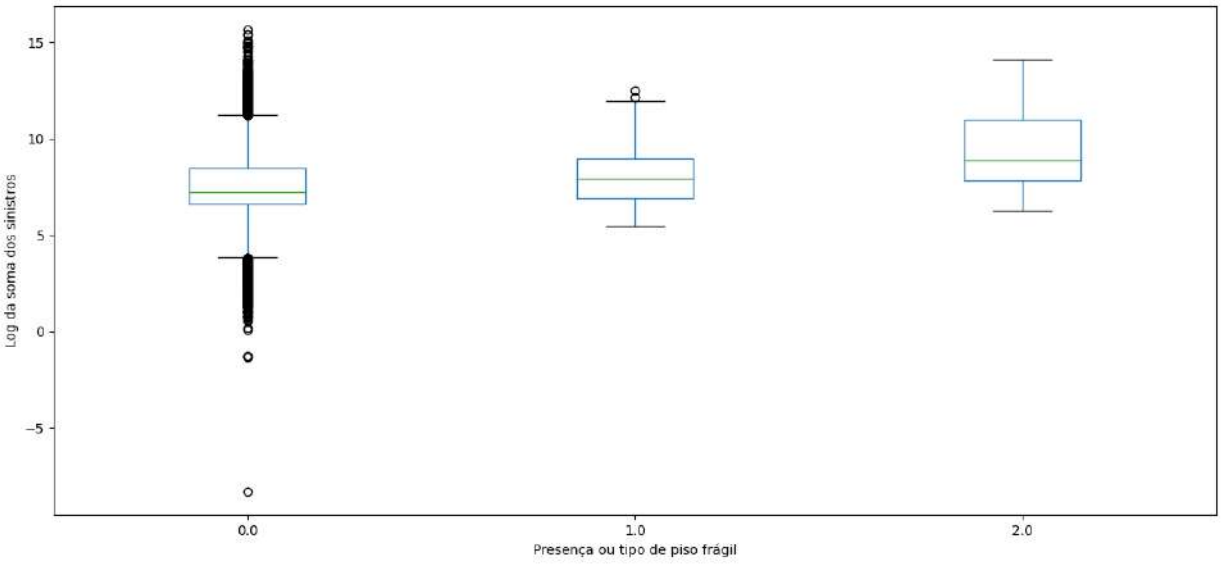

Figura 6 – *Boxplot* do log da soma de sinistros por presença ou tipo de piso frágil

A Figura 6 parece indicar um comportamento crescente no log da soma dos sinistros frente à presença ou tipo de piso frágil. Consoante LANDERO (2003), que performou um estudo comparativo do comportamento sísmico de edifícios com piso frágil, existe evidência contundente para concluir que estruturas com piso frágil são bem menos resistentes a abalos sísmicos e em nosso contexto mais propensas a apresentarem um sinistro grave para a seguradora.

Em suma, observamos o comportamento da distribuição de algumas variáveis a partir das tabelas de quantidade de apólices e o log da soma do sinistro (variável de interesse deste trabalho). Para as variáveis de localidade, vimos que locais mais populosos têm mais apólices subscritas e a soma acumulada é maior do que em regiões mais remotas e menos popularmente densas. Já para a variável do tipo de ocupação, concluimos que a natureza do edifício segurado parece ter relação com a magnitude dos sinistros. No que tange à variável do número de andares, a relação com a variável de interesse foi inconclusiva. Finalmente, notamos que a presença ou gravidade de piso frágil em uma construção diz respeito à resistência sísmica, e consequentemente, gravidade do sinistro na nossa carteira de seguros patrimoniais catastróficos.

Outrossim, existem outras variáveis presentes na base de dados que são importantes para a magnitude do valor do sinistro líquido da apólice, o que faz sentido pensando que elas foram coletadas pela seguradora em primeiro lugar. Entretanto, a análise exploratória de tais variáveis tem interpretação similar às já apresentadas e foi optado não incluí-las a fim de evitar redundância.

3 INFERÊNCIA E MODELOS DE REGRESSÃO

Neste capítulo, discorreremos sobre os procedimentos de inferência na seção 3.1 e os modelos de regressão adotados na aplicação deste trabalho na seção 3.2.

Inicialmente, podemos listar considerações e variáveis de interesse que serão relevantes para este capítulo:

- Tomaremos a abordagem probabilística com os modelos de regressão;
- A variável resposta y_i representa o desfecho a ser estudado neste trabalho, ou seja, o log dos sinistros associados a i-ésima apólice;
- As regressoras $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ik})^T$ representam as características de cada apólice, por exemplo, x_{ij} indica a presença leve de piso frágil na construção, para algum j;
- O parâmetro $\boldsymbol{\theta}$ representa o vetor aleatório que relaciona as regressoras à variável resposta.
- A função $p(y_i|\boldsymbol{\theta}, \mathbf{x}_i)$ representa a distribuição de probabilidade dos sinistros dadas as características do imóvel, por exemplo no caso da regressão linear múltipla sem penalidade, $y_i|\boldsymbol{\theta}, \mathbf{x}_i \sim N(\boldsymbol{\theta}^T\mathbf{x}_i, \sigma^2)$.

3.1 INFERÊNCIA

Nesta seção, apresentaremos brevemente inferência clássica e inferência Bayesiana via MCMC. Inferência clássica e Bayesiana serão o fundamento para discutir os modelos na seção 3.2 e aplicá-los no capítulo 4.

3.1.1 Inferência clássica

Conforme DEGROOT; SCHERVISH (2012), um modelo estatístico consiste na identificação de variáveis aleatórias de interesse (tanto observáveis quanto apenas hipoteticamente observáveis) na especificação de distribuições conjuntas possíveis para as variáveis aleatórias observáveis e na identificação de quaisquer parâmetros dessas distribuições que são assumidos como desconhecidos. Quando tratamos o parâmetro desconhecido $\boldsymbol{\theta}$ como aleatório, a distribuição conjunta das variáveis aleatórias observáveis indexadas por $\boldsymbol{\theta}$ é a distribuição condicional das variáveis aleatórias observáveis dado $\boldsymbol{\theta}$.

Tomemos o seguinte exemplo relativo à operação de seguros: uma seguradora vende proteção patrimonial e está interessada em prever o tempo até o aviso de um sinistro de imóveis de 5 andares na região de Magallanes. A empresa pode coletar dados sobre as apólices de imovéis de 5 andares na região de Magallanes, optando por usar uma

distribuição da família exponencial para modelar o tempo desde quando uma apólice é emitida até o aviso de um sinistro. Eles gostariam de modelar as apólices como tendo todos a mesma taxa de falha θ, mas o valor de θ embora fixo é desconhecido. Essa hipótese é razoável visto que estamos interessados em imóveis sob risco similar, de mesma quantidade de andares e na mesma região.

Mais precisamente, denotemos $X_1, X_2, ...$ como uma sequência de tempos até o aviso de um sinistro, em meses. A empresa acredita que se conhecessem a taxa de falha θ, então $X_1, X_2, ...$ seriam variáveis aleatórias independentes e identicamente distribuídas, seguindo a distribuição exponencial com parâmetro θ. Essa descrição é exatamente o modelo estatístico para o problema apresentado.

Dessa forma, DEGROOT; SCHERVISH (2012) define inferência estatística como um procedimento que gera uma afirmação probabilística sobre alguma ou todas as partes de um modelo estatístico. Aqui "afirmação probabilística" significa uma declaração que faz uso de quaisquer dos conceitos da teoria da probabilidade. Alguns exemplos incluem média, média condicional, quantil, variância, distribuição condicional para uma variável aleatória dada outra, a probabilidade de um evento, a probabilidade condicional de um evento dado algo, e assim por diante.

De volta ao exemplo proposto acima, suponha-se que tenhamos os dados observados pela seguradora, denotados por $X_1, ..., X_m$ e existe interesse pelo o que acontecerá com o tempo até aviso de sinistro das próximas apólices, ou seja $X_{m+1}, X_{m+2}, ...$ Alguns exemplos de inferência estatística para o caso proposto incluiria obtenção da estimador de máxima verossimilhança para a média global dado os valores observados, construir intervalos de confiança para o valor da taxa de falha θ, calcular probabilidade acerca da média da vida útil dos componentes, etc.

3.1.2 Inferência Bayesiana via MCMC

Semelhante ao método clássico, a abordagem Bayesiana também considera os dados observados, aqui denotado $\mathbf{y}$. Entretanto, a abordagem Bayesiana trata $\boldsymbol{\theta}$ como um vetor aleatório usando modelos probabilísticos que incorporam sua incerteza.

Segundo DEGROOT; SCHERVISH (2012), a estrutura Bayesiana é composta por elementos específicos. O primeiro é a distribuição de probabilidade atribuída ao parâmetro $\boldsymbol{\theta}$ antes da observação das demais variáveis de interesse, chamada distribuição a priori. Dada a natureza subjetiva dessa distribuição, é comum que opiniões de especialistas auxiliem na modelagem dos dados, determinando se distribuições a priori mais ou menos informativas serão utilizadas, assim como se parâmetros secundários, chamados hiperparâmetros (pensados como parâmetros do parâmetros de $\boldsymbol{\theta}$), serão incorporados. A notação usada para descrever a função de distribuição a priori de $\boldsymbol{\theta}$ neste artigo é $p(\boldsymbol{\theta})$.

O segundo elemento é a função de verossimilhança $\boldsymbol{\theta} \mapsto f(\mathbf{y}|\boldsymbol{\theta})$, que representa a probabilidade conjunta dos dados (variáveis aleatórias a serem observadas), condicionada

a e em função de $\boldsymbol{\theta}$.

O terceiro e último elemento é a distribuição de probabilidade condicional do parâmetro $\boldsymbol{\theta}$, dado as variáveis aleatórias a serem observadas, $\mathbf{y}$, chamada distribuição a posteriori. A notação usada para descrever a função de distribuição a posteriori de $\boldsymbol{\theta}$ neste estudo é $p(\boldsymbol{\theta}|\mathbf{y})$, que pode ser obtida a partir do Teorema de Bayes:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}) d\boldsymbol{\theta}} \propto p(\mathbf{y}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}), \tag{3.1}$$

onde a proporcionalidade segue uma vez que o denominador não depende de $\boldsymbol{\theta}$.

A equação (3.1) pode levar a modelos muito complexos, mesmo quando a priori e verossimilhança são relativamente simples, e tratá-los analiticamente pode ser impossível, de modo que métodos numéricos são necessários para avaliação. A simulação permite estudar a distribuição a posteriori com mais detalhes.

Para simular valores dessas distribuições numericamente, usaremos o Método de Monte Carlo via Cadeias de Markov (MCMC). Esse método consiste em gerar valores para $\boldsymbol{\theta}$ oriundos de distribuições aproximadas, amostrando sequencialmente e utilizando os últimos valores gerados para estimar a distribuição a posteriori $p(\boldsymbol{\theta}|\mathbf{y})$, objetivo do modelo. As amostras em sequência formam uma Cadeia de Markov.

3.1.2.1 Amostrador de Gibbs

Segundo GELMAN et al. (2013) e GAMERMAN; LOPES (2006), o Amostrador de Gibbs consiste em tomar $\boldsymbol{\theta}$ como um vetor de K subvetores ou componentes tal que $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K)^T$. Cada iteração do amostrador de Gibbs gera valores para os K componentes de $\boldsymbol{\theta}$, onde cada cada $\boldsymbol{\theta}_k^{(t)}$ (valor do k-ésimo componente de $\boldsymbol{\theta}$ na t-ésima iteração) é gerado condicionado aos demais.

A distribuição de um determinado $\boldsymbol{\theta}_k$ condicionado aos demais $K-1$ componentes de $\boldsymbol{\theta}$ e em $\mathbf{y}$ é chamada de distribuição condicional completa de $\boldsymbol{\theta}_k$. A notação utilizada neste trabalho para essa distribuição é $p(\boldsymbol{\theta}_k|\mathbf{y}, \boldsymbol{\theta}_{-k})$, onde $\boldsymbol{\theta}_{-k} = (\boldsymbol{\theta}_i;\ i = 1, ..., K,\ i \neq k.)$.

O algoritmo pode ser resumido a seguir:

1. Inicie o contador de iterações em $t = 1$ e defina os valores iniciais $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\theta}_1^{(0)}, ..., \boldsymbol{\theta}_k^{(0)})^T$
2. Obtenha um novo valor de $\boldsymbol{\theta}^{(t)} = (\boldsymbol{\theta}_1^{(t)}, ..., \boldsymbol{\theta}_k^{(t)})^T$ através de uma geração sucessiva de valores:

$$\begin{aligned}
&\boldsymbol{\theta}_1^{(t)} \sim p(\boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_2^{(t-1)}, \boldsymbol{\theta}_3^{(t-1)}, ..., \boldsymbol{\theta}_k^{(t-1)});\\
&\boldsymbol{\theta}_2^{(t)} \sim p(\boldsymbol{\theta}_2|\mathbf{y}, \boldsymbol{\theta}_1^{(t)}, \boldsymbol{\theta}_3^{(t-1)}, ..., \boldsymbol{\theta}_k^{(t-1)});\\
&\quad\vdots\\
&\boldsymbol{\theta}_k^{(t)} \sim p(\boldsymbol{\theta}_k|\mathbf{y}, \boldsymbol{\theta}_1^{(t)}, \boldsymbol{\theta}_2^{(t)}, ..., \boldsymbol{\theta}_{k-1}^{(t)}).
\end{aligned}$$

3. Passe o contador de t para $t+1$ e volte ao passo 2 até que a convergência ocorra.

3.1.2.2 Algoritmo de Metropolis-Hastings

O algoritmo de Metropolis-Hastings, tal como o amostrador de Gibbs, também é um método MCMC bastante utilizado em contextos Bayesianos para gerar amostras aproximadas da distribuição a posteriori. Nesta seção descreveremos iterativamente o funcionamento do algoritmo.

Segundo PETRIS; PETRONE; CAMPAGNOLI (2009), o algoritmo de Metropolis-Hastings com passeio aleatório como proposta é um método flexível e geral o bastante para gerar uma cadeia de Markov utilizando uma distribuição invariante.

Tomemos $\boldsymbol{\theta}$ como o parâmetro de interesse e $\boldsymbol{\theta}^{(t)}$ como o valor do parâmetro na t-ésima iteração do algoritmo de Metropolis. Inicializando o algorito em um valor inicial $\boldsymbol{\theta}^{(0)}$, o algoritmo gera um valor proposto $\boldsymbol{\theta}^*$ para o próximo estado da cadeia através de uma distribuição proposta $q(\cdot|\boldsymbol{\theta}^{(t)})$, isto é, $\boldsymbol{\theta}^* \sim q(\cdot|\boldsymbol{\theta}^{(t)})$. Usando como objetivo a distribuição a posteriori $p(\boldsymbol{\theta}|\mathbf{y})$ do parâmetro a ser estimado, o valor proposto com probabilidade $\alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})$ (o valor α será exibido a frente) ou rejeitado com probabilidade $1-\alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})$. Caso o valor proposto seja aceito, o próximo estado da cadeia será o valor proposto, ou seja, $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^*$, e caso seja rejeitado, será o valor do estado anterior da cadeia, ou seja, $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$, e assim se segue com as iterações subsequentes.

No algoritmo de Metropolis-Hastings com proposta em passeio aleatório, $q(\cdot|\boldsymbol{\theta}^{(t)})$ representa a densidade de transição para $\boldsymbol{\theta}^*$ condicionalmente ao estado atual $\boldsymbol{\theta}^{(t)}$ da cadeia e modela, por exemplo, uma distribuição Normal $\boldsymbol{\theta}^*|\boldsymbol{\theta} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, se o suporte de $\boldsymbol{\theta}$ for $\mathbb{R}^k$.

O algoritmo é dado por:

1. Inicialize $t=0$ e defina o valor inicial $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\theta}_1^{(0)}, ..., \boldsymbol{\theta}_k^{(0)})^T$.
2. Obtenha um novo valor proposto $\boldsymbol{\theta}^*$ a partir da distribuição $q(\cdot|\boldsymbol{\theta}^{(t)})$.
3. Compute $\alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}) = \min\left\{1, \frac{p(\boldsymbol{\theta}^*|\mathbf{y})q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{(t-1)}|\mathbf{y})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})}\right\}$.
4. Gere uma variável independente aleatória $U \sim \text{Ber}(\alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)}))$.
5. Se $U=1$, faça $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$, caso contrário faça $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.
6. Passe o contador de t para $t+1$ e volte aos passos de 2 a 6 até que evidências de convergência sejam observadas.

3.1.2.3 Amostrador de Gibbs com passo de Metropolis

O amostrador de Gibbs é o mais simples dos algoritmos de simulação de cadeias de Markov e costuma ser a primeira escolha para modelos condicionalmente conjugados, onde

podemos amostrar diretamente de cada distribuição condicional completa a posteriori, mas requer o uso de aproximações em caso de modelos que não são condicionalmente conjugados.

Segundo GELMAN et al. (2013), tanto o amostrador de Gibbs quanto o algoritmo de Metropolis podem ser usados em conjunto para amostrar de distribuições complicadas. Quando parâmetros a serem estimados pelo amostrador de Gibbs não possuem forma fechada para suas respectivas distribuições condicionais completas, o algoritmo de Metropolis é uma opção viável para amostras dessas distribuições de forma aproximada.

Se algumas das distribuições condicionais completas em um modelo não puderem ser amostradas diretamente em um dos passos da etapa 2 do algoritmo de Gibbs, os parâmetros podem ser amostrados de forma aproximada usando o algoritmo de Metropolis-Hastings com passeio aleatório.

De forma mais geral, os parâmetros podem ser atualizados em blocos, onde cada bloco é amostrado em seu respectivo passo no amostrador de Gibbs, podendo-se, inclusive, utilizar um passo de Metropolis para amostrar dentro do bloco, caso a respectiva condicional completa do bloco seja desconhecida. Esse método é frequentemente referido como *Metropolis-Hastings within Gibbs* ou amostrador de Gibbs com passo de Metropolis.

3.2 MODELOS DE REGRESSÃO

Nesta seção, introduziremos os modelos de regressão que serão utilizados para seleção de variáveis e quantificação do efeitos de regressoras no capítulo 4. Discorreremos sobre os modelos de regressão linear múltipla, com penalização Ridge, Lasso, *Elastic Net* ingênuo e o modelo Bayesiano *Spike-and-Slab*.

3.2.1 Regressão linear múltipla

De acordo com MONTGOMERY; PECK; VINING (2013), para avaliar a relação da variável resposta y em relação a k covariáveis X_j, $j = 1, ..., k$, podemos tomar o modelo dado por:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i, \ i = 1, ..., n, \tag{3.2}$$

em que, n é o número de observações e $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)'$ é um vetor de coeficientes de regressão (parâmetros), ϵ_i é um componente de erro aleatório e x_{ij} representa o valor númerico para a j-ésima covariável, X_j, à i-ésima observação da base de dados.

Assume-se que os erros são independentes e seguem distribuição normal com média zero e variância desconhecida σ^2, isto é, $\epsilon_i \sim N(0, \sigma^2)$.

O modelo é chamado de regressão linear múltipla, pois envolve mais de um coeficiente de regressão. O adjetivo linear indica que o preditor estipulado no modelo é uma função linear dos parâmetros $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)^T$.

Um dos métodos tradicionais de estimação dos parâmetros do modelo de regressão linear múltipla é obtida através solução analítica já conhecida dos estimadores de mínimos quadrados, que, sob a hipótese de erros independentes e identicamente distribuídos com média zero, é equivalente ao estimador de máxima verossimilhança, e é apresentada em MONTGOMERY; PECK; VINING (2013) como:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \text{ se a matriz } (\mathbf{X}'\mathbf{X})^{-1} \text{ existir,} \tag{3.3}$$

em que a matriz $\mathbf{X}$ é tal que a i-ésima linha é composta por $\mathbf{x}_i = (1, x_{i1}, ..., x_{ik})^T$ e a variável resposta $\mathbf{y} = (y_1, y_2, ..., y_n)^T$.

Esse estimador segue distribuição normal com média $\boldsymbol{\beta}$ e tem variância $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$, isto é, $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2)$.

A implementação do modelo em *Python* pode ser obtida através da biblioteca *scikit-learn*, na classe *linear_model*, método *LinearRegression*. Esse biblioteca é uma pacote de aprendizado de máquina de software livre para a linguagem de programação *Python* e possui vários algoritmos de classificação, regressão e agrupamento, incluindo máquinas de vetor de suporte, florestas aleatórias, aumento de gradiente, k-médias e DBSCAN.

3.2.2 Ridge

Segundo HOERL; KENNARD (1970), para obtermos o modelo Ridge, suponhamos y_i e $\mathbf{x}_i = (x_{i1}, ..., x_{ik})^T$ com a mesma interpretação que no modelo de regressão linear múltipla. Assim, ao tomar $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)^T$, o estimador de Ridge para o vetor de efeito das regressoras sob o método de mínimos quadrados com penalização Ridge em sua formulação como problema de otimização com restrição é definido por:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \right\}, \text{ sujeito a } \sum_{j=1}^{k} \beta_j^2 \leq t, \tag{3.4}$$

onde $t \geq 0$ é um parâmetro de ajuste. Uma forma alternativa de descrever esse estimador é a partir da penalização quadrática:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{k} \beta_j^2 \right\}, \tag{3.5}$$

onde $\lambda \in \mathbb{R}^+$ é um parâmetro de penalização. Note que β_0 não sofre penalização por ser o intercepto e não interagir com nenhuma covariável diretamente. Podemos pensar também

em Ridge como um modelo Bayesiano ao elicitar β_j para a distribuição a priori:

$$\beta_j \sim N\left(0, \frac{1}{2\lambda}\right) \implies p(\beta_j) \propto \exp\{-\lambda\beta_j^2\}, j = 1, ..., k,$$

$$\text{com modelo: } (y_i|\beta_0, \beta_j) \sim N\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}, \sigma^2\right)$$

$$\implies p(y_i|\beta_0, \beta_j) \propto \exp\left\{ -\left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij}\right)^2 \right\}$$

Agora a estimativa dos parâmetros do modelo Ridge via máxima verossimilhança é equivalente à estimativa obtida pela moda a posteriori do modelo Ridge Bayesiano:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \arg\max_{\boldsymbol{\beta}} \log(p(\boldsymbol{\beta}|\mathbf{y})) \\
&= \arg\max_{\boldsymbol{\beta}} \{\log(\mathbf{y}|\boldsymbol{\beta})p(\boldsymbol{\beta})) - \log(p(\mathbf{y}))\} \\
&= \arg\max_{\boldsymbol{\beta}} \{\log(p(\mathbf{y}|\boldsymbol{\beta})p(\boldsymbol{\beta}))\} \\
&= \arg\max_{\boldsymbol{\beta}} \left\{ \log\left(\prod_{i=1}^{n} \exp\left\{ -\left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij}\right)^2 \right\} \right) + \log\left(\prod_{j=1}^{k} \exp\{-\lambda\beta_j^2\} \right) \right\} \\
&= \arg\max_{\boldsymbol{\beta}} \left\{ -\left(\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{k} \beta_j^2 \right) \right\} \\
&= \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{k} \beta_j^2 \right\}
\end{aligned}$$

Existe solução analítica para $\hat{\boldsymbol{\beta}}$, dada por:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_0)^{-1}\mathbf{X}'\mathbf{Y}, \tag{3.6}$$

em que $\mathbf{I}_0$ é uma matriz identidade $(k+1) \times (k+1)$ tal que o elemento da primeira linha e primeira coluna é igual a zero.

A regressão Ridge reduz a magnitude das estimativas para os coeficientes β_j da regressão linear, em comparação com as estimativas de minímos quadrados ordinários. Sua aplicação não está relacionada a seleção de variáveis, pois embora o Ridge encolha os coeficientes $\hat{\beta}_j$, não há garantias de forçar $\hat{\beta}_j = 0$, assim não consegue excluir a influência de uma covariável sobre a resposta.

Além disso, a regressão Ridge também está relacionada a uma maior estabilidade numérica em comparação com mínimos quadrados ordinário devido ao acréscimo de $\lambda > 0$ na diagonal de $\mathbf{X}^T\mathbf{X}$, o que muitas vezes evita instabilidades com o procedimento de inversão em (3.6) em comparação com (3.3).

A implementação em *Python* neste trabalho foi realizada pelo método *Ridge* da biblioteca *sckit-learn* e classe *linear_model*.

3.2.3 Lasso

Segundo TIBSHIRANI (1996), nos mesmos moldes que o Ridge, podemos obter o estimador de Lasso enquanto:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\left\{ \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2\right\}, \text{ sujeito a } \sum_{j=1}^{k}|\beta_j| \leq t, \tag{3.7}$$

onde $t \geq 0$ é um parâmetro de ajuste. Uma forma alternativa de descrever esse estimador é como:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\left\{ \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{k}|\beta_j|\right\}, \tag{3.8}$$

onde $\lambda \in \mathbb{R}^+$ é um parâmetro de penalização. Semelhante ao Ridge, podemos tomar o modelo de regressão Lasso a partir de um modelo Bayesiano ao incluir uma priori para β_j:

$$\beta_j \sim \text{Laplace}\left(0, \frac{1}{\lambda}\right) \implies p(\beta_j) \propto \exp\{-\lambda|\beta_j|\},$$

$$\text{com modelo: } (y_i|\boldsymbol{\beta}) \sim N\left(\beta_0 + \sum_{j=1}^{k}\beta_j x_{ij}, \sigma^2\right)$$

$$\implies p(y_i|\boldsymbol{\beta}) \propto \exp\left\{ -\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2\right\}$$

Agora a estimativa dos parâmetros do modelo Lasso via máxima verossimilhança é equivalente à estimativa obtida pela moda a posteriori do modelo Lasso Bayesiano:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \arg\max_{\boldsymbol{\beta}} \log(p(\boldsymbol{\beta}|\mathbf{y})) \\
&= \arg\max_{\boldsymbol{\beta}}\{\log(\mathbf{y}|\boldsymbol{\beta})p(\boldsymbol{\beta})) - \log(p(\mathbf{y}))\} \\
&= \arg\max_{\boldsymbol{\beta}}\{\log(p(\mathbf{y}|\boldsymbol{\beta})p(\boldsymbol{\beta}))\} \\
&= \arg\max_{\boldsymbol{\beta}}\left\{ \log\left(\prod_{i=1}^{n}\exp\left\{ -\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2\right\}\right) + \log\left(\prod_{j=1}^{k}\exp\{-\lambda|\beta_j|\}\right)\right\} \\
&= \arg\max_{\boldsymbol{\beta}}\left\{ -\left(\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{k}|\beta_j|\right)\right\} \\
&= \arg\min_{\boldsymbol{\beta}}\left\{ \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{k}|\beta_j|\right\}
\end{aligned}$$

Repare que o estimador encontrado é igual ao proposto em (3.8). Já o modelo proposto por PARK; CASELLA (2008) chamado *Bayesian Lasso* inclui β_j como uma priori condicionada em σ^2. Isso deve-se ao fato de que o modelo acima com priori não condicionada gera bimodalidade na densidade da posteriori. Na Figura 7, supondo uma priori independente para σ^2, podemos observar o fenômeno.

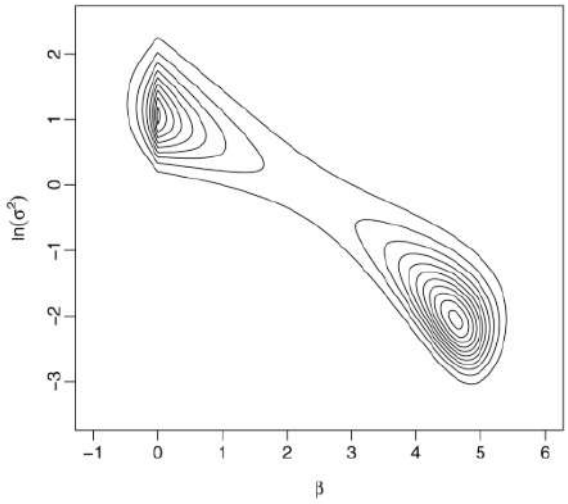

Figura 7 – Gráfico de contorno da densidade da distribuição a posteriori de $(\beta, \ln(\sigma^2))$ da forma apresentada acima

Entretanto, para os efeitos de compreender a regressão Lasso enquanto um modelo Bayesiano, o modelo sem condicionamento serve seu propósito. Já em uma aplicação da abordagem Bayesiana para o Lasso, é importante utilizar o modelo com a priori condicionada a σ^2 a fim de evitar uma posteriori com duas modas.

Note que não existe solução analítica para $\hat{\boldsymbol{\beta}}$, por isso deve ser computada como um problema númerico com restrição a partir de uma desigualdade.

Diferentemente do Ridge, o Lasso pode ser utilizado para seleção de variáveis, pois é possível ter λ tal que $\hat{\beta}_j = 0$, assim identificando uma covariável que não exerce influência para a resposta.

A implementação no *Python* é obtida pelo método *Lasso* da biblioteca *sckit-learn* e classe *linear_ model*, as mesmas que os demais modelos. O algoritmo usado pelo *sckit-learn* para obtenção de $\hat{\beta}_j$ é por coordenadas decrescentes.

3.2.4 *Elastic Net* ingênuo

Consoante ZOU; HASTIE (2005), o modelo *Elastic Net* ingênuo, nos mesmos moldes que o Ridge e Lasso, tem seu estimador definido como:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \right\}, \text{ sujeito a } (1-\gamma) \sum_{j=1}^{k} |\beta_j| + \gamma \sum_{j=1}^{k} \beta_j^2 \leq t, \tag{3.9}$$

onde $t \geq 0$ é um parâmetro de ajuste.

Chama-se a função $(1-\gamma)\sum_{j=1}^{k} |\beta_j| + \gamma \sum_{j=1}^{k} \beta_j^2$ de penalidade *Elastic Net*. Quando $\gamma = 1$, o modelo se torna uma regressão de Ridge simples. Já quando $\gamma = 0$, o modelo é uma regressão de Lasso simples. Uma forma alternativa de descrever esse estimador é

como:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 + \lambda_1 \sum_{j=1}^{k} |\beta_j| + \lambda_2 \sum_{j=1}^{k} \beta_j^2 \right\}, \tag{3.10}$$

onde $\lambda_1, \lambda_2 \in \mathbb{R}^+$ são parâmetros de penalização.

Ainda de acordo com ZOU; HASTIE (2005), esse modelo é chamado de *Elastic Net* ingênuo, pois existe evidência empírica de que este modelo não performa bem ao menos que esteja bem próximo ao Ridge ou ao Lasso. O estimador de *Elastic Net* procura corrigir o duplo encolhimento (*double shrinkage*) que introduz viés extra ao estimador ingênuo, e é dado por:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \boldsymbol{\beta}^T \left(\frac{\mathbf{X}^T\mathbf{X} + \lambda_2 \mathbf{I}}{1 + \lambda_2} \right) \boldsymbol{\beta} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \lambda_1 \sum_{j=1}^{k} |\beta_j| \right\} \tag{3.11}$$

O modelo implementado no *Python*, disponível pelo método *ElasticNet* da biblioteca *scikit-learn* e classe *linear_ model*, é similar ao *Elastic Net* ingênuo, e consiste na minimização do seguinte estimador:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \left(\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 + \gamma L_1 \sum_{j=1}^{k} |\beta_j| + 0.5\gamma(1 - L_1) \sum_{j=1}^{k} \beta_j^2 \right) \right\}, \tag{3.12}$$

onde $\gamma = \lambda_1 + \lambda_2$ e $L_1 = \frac{\lambda_1}{\gamma}$. Essa parametrização é equivalente ao *Elastic Net* ingênuo com $\lambda_2^* = 0.5 \times \lambda_2$, ou ainda:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 + \lambda_1 \sum_{j=1}^{k} |\beta_j| + 0.5 \cdot \lambda_2 \sum_{j=1}^{k} \beta_j^2 \right\} \tag{3.13}$$

Exceto quando $\lambda_1 = 0$, não existe solução analítica para $\hat{\boldsymbol{\beta}}$, assim a obtenção da estimativa é feita numericamente pelo algoritmo de coordenadas decrescentes, tal qual no caso Lasso.

3.2.5 *Spike-and-Slab* Bayesiano

O modelo de *Spike-and-Slab* Bayesiano é proposto a partir de um modelo de regressão com uma priori para os efeitos das regressoras, aqui denotado β_k, cuja distribuição de probabilidade é uma mistura entre uma distribuição contínua normal e uma distribuição discreta com massa de probabilidade em 0, ou seja, o modelo fornece $P(\beta_k = 0|\mathbf{y}) > 0$ e dá a probabilidade que o efeito de uma regressora seja irrelevante. Essa característica é importante para seleção de variáveis e para a finalidade deste trabalho, que é discorrer sobre a importância de cada regressora sobre a variável resposta.

Consoante TADESSE; VANNUCCI (2021), tomando as notações já apresentadas para os modelos anteriores, um modelo *Spike-and-Slab* Bayesiano pode ser descrito como:

- Modelo: $y_i \sim N(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2)$
- Priori mista para os efeitos das regressoras: $\beta_k \sim (1-\pi_k)\delta_0 + \pi_k N(0, \sigma^2\tau^2)$, onde δ_x denota a medida delta de Dirac no ponto x
- Priori para os demais parâmetros, considerando independência a priori:
  - $\pi_k \sim \text{Bernoulli}(p)$
  - $p \sim \text{Beta}(a, b)$
  - $\tau^{-2} \sim \text{Gama}(\alpha_1, \alpha_2)$
  - $\sigma^{-2} \sim \text{Gama}(\frac{1}{2}, \frac{s^2}{2})$

em que a, b, α_1, α_2 e s^2 são conhecidos.

Note que se $\pi_k = 0$, então $\beta_k = 0$ quase certamente, e portanto $\mathbf{X}_k$ não tem efeito sobre $\mathbf{y}$.

Seja o vetor paramétrico $\boldsymbol{\theta} = (\boldsymbol{\beta}, p, \boldsymbol{\pi}, \tau^{-2}, \sigma^{-2})^T$. Seguem as condicionais completas para descrição do modelo:

$$(p|\mathbf{y}, \boldsymbol{\theta}_{-p}) \sim \text{Beta}\left(a + \sum_{k=1}^{d}\pi_k, n - \sum_{k=1}^{d}\pi_k + b\right) \tag{3.14}$$

$$(\tau^{-2}|\mathbf{y}, \boldsymbol{\theta}_{-\tau^{-2}}) \sim \text{Gama}\left(\alpha_1 + \frac{|S_1|}{2}, \frac{\boldsymbol{\beta}^T\boldsymbol{\beta}}{2\sigma^2} + \alpha_2\right) \tag{3.15}$$

$$(\sigma^{-2}|\mathbf{y}, \boldsymbol{\theta}_{-\sigma^{-2}}) \sim \text{Gama}\left(\frac{n + |S_1| + 1}{2}, \frac{1}{2}\left[s^2 + \tau^{-2}\boldsymbol{\beta}^T\boldsymbol{\beta} + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right]\right) \tag{3.16}$$

$$(\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\theta}_{-\boldsymbol{\beta}}) \sim N(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*) \tag{3.17}$$

$$P(\pi_k = 1|\mathbf{y}, \boldsymbol{\theta}_{-(\beta_k, \pi_k)}) = 1 - \frac{(1-p)}{p\tau^{-1}(\mathbf{x}_{\cdot k}^T\mathbf{x}_{\cdot k} + \tau^{-2})^{-\frac{1}{2}} \exp\left\{\frac{(\mathbf{x}_{\cdot k}^T\mathbf{x}_{\cdot k})^2}{\mathbf{x}_{\cdot k}^T\mathbf{x}_{\cdot k} + \tau^{-2}}\right\} + (1-p)}, \tag{3.18}$$

onde $S_x = \{x : \pi_k = x\}, d_x = |S_x|$, $x = \{0, 1\}$. Ainda, $\boldsymbol{\Sigma}_* = \sigma^2(\mathbf{X}_{-S_0}^T\mathbf{X}_{-S_0} + \tau^{-2}I_{d-d_0})$, $\boldsymbol{\mu}_* = \sigma^{-2}\boldsymbol{\Sigma}_*\mathbf{X}_{-S_0}^T\mathbf{y}$.

A aplicação no *Python* foi realizada através da biblioteca *pymc*, que de acordo com seu site, é uma biblioteca de programação probabilística que permite aos usuários construir modelos Bayesianos de maneira simples e ajustá-los usando método de Monte Carlo via Cadeias de Markov (MCMC). O método que o *pymc* utiliza para amostragem é o *Metropolis-Hastings within Gibbs*, embora neste caso o amostrador de *Gibbs* conseguirá executar a amostragem vide a forma fechada das condicionais completas.

4 RESULTADOS

Após termos elucidado parte do arcabouço teórico presente neste trabalho, podemos utilizar os modelos apresentados para entender o efeito de cada variável presente na base de dados fornecida na severidade dos sinistros para auxiliar a subscrição de seguros. A seção 4.1 diz respeito ao pré processamento da base de dados e a seção 4.2 a seguir discorre sobre a aplicação dos modelos e interpretação dos resultados obtidos.

4.1 PRÉ-PROCESSAMENTO DA BASE DE DADOS

O primeiro passo para ajustar os modelos apresentados é transformar a base de dados contendo 20 colunas, sendo 17 covariáveis qualitativas nominais e 2 covariáveis mais a variável resposta sendo quantitativas contínuas, em uma base totalmente númerica para que possa ser lida pelos modelos. Para isso, iremos manter as variáveis quantitativas contínuas e transformar as variáveis qualitativas nominais em variáveis *dummies* binárias, onde "0" indica a ausência daquela variável na apólice e "1" indica a presença da variável na ápolice, como o seguinte exemplo da Tabela 6:

Tabela 6 – Transformação de variáveis qualitativas nominais em variáveis *dummies*

Apólice	**Cidade**
A	Santiago
B	Valparaíso
C	Valdivia

$\longrightarrow$

Apólice	**Santiago**	**Valparaíso**	**Valdivia**
A	1	0	0
B	0	1	0
C	0	0	1

O resultado dessa transformação é uma base de dados com 372 covariáveis e 1 resposta. As colunas de latitude e longitude, as únicas covariáveis originalmente quantitativas contínuas, foram padronizadas para que todas covariáveis tenham magnitude similar a fim de evitar problemas númericos.

Adicionalmente, os dados foram separados aleatoriamente em conjuntos de treino e teste, com proporção de 70% treino e 30% teste. Essa separação foi feita utilizado a função *train_test_split*, da biblioteca *sckit-learn*.

Ultimamente, foram eliminadas colunas com variância zero no conjunto de teste, ou seja, variáveis que não tiveram ocorrência nenhuma no conjunto de teste e a permanência delas implicariam em problemas númericos na aplicação. No fim, de 372 covariáveis, foram eliminadas 24 covariáveis, como por exemplo, o código SIC 8072, que caracteriza consultórios dentais, em que havia uma única ocorrência na base original. Ao fim da exclusão, 348 covariáveis permaneceram na base de dados.

4.2 APLICAÇÃO

Os primeiros modelos a serem aplicados serão estimados via máxima verossimilhança penalizada sob a abordagem clássica, ou equivalentemente, via máximo a posteriori do ponto de vista Bayesiano: regressão linear múltipla sem penalidade, Ridge, Lasso e *Elastic Net* ingênuo. Os procedimentos de inferência para os modelos são realizados a partir das funções do *scikit-learn*: *LinearRegression*, *Lasso*, *Ridge* e *ElasticNet*, respectivamente.

Para os modelos que incluem hiperparâmetros a serem ajustados, como por exemplo o parâmetro λ para a regressão Lasso, Ridge e *Elastic Net* e a razão entre as penalidades Lasso e Ridge denotada como γ para o *Elastic Net*, iremos performar o método de validação cruzada.

Segundo REFAEILZADEH; TANG; LIU (2016), na validação cruzada, os dados são inicialmente divididos em k segmentos ou blocos de tamanho igual. Posteriormente, são realizadas k iterações de treinamento e validação, de modo que em cada iteração um bloco diferente dos dados é reservado para validação, enquanto os $k-1$ blocos restantes são utilizadas para treinamento. A Figura 8 demonstra um exemplo do esquema de particionamento e execução do método com $k = 3$.

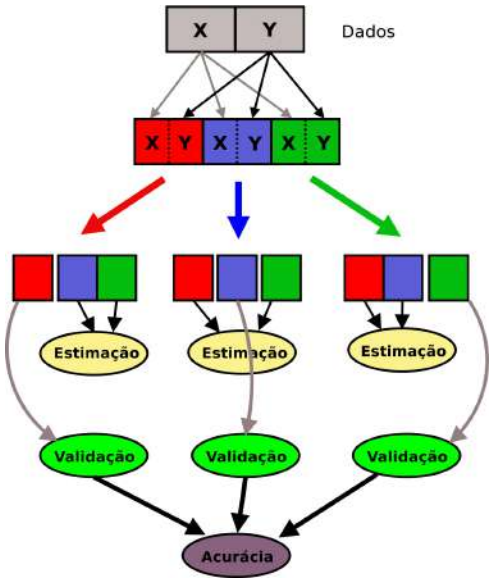

Figura 8 – Validação cruzada com número de blocos $k = 3$

Desta maneira, faremos a busca em grade com validação cruzada para obter o valor ótimo para o hiperparâmetro em cada um dos os modelos. Para modelos que incluem apenas um único hiperparâmetro, propomos uma grade de valores para teste, e então o modelo é aplicado para cada valor proposto para o hiperparâmetro. Em sequência, performa-se o método de validação cruzada e calcula-se uma medida de qualidade preditiva para o valor proposto. O hiperparâmetro que resultar na melhor medida preditiva é selecionado como o valor ótimo dado a lista de valores introduzida.

No caso de modelos com dois ou mais hiperparâmetros, a locução adverbial "em grade" entra em jogo: dado listas de valores para cada hiperparâmetro a ser testado, o domínio dos hiperparâmetros é dividido em uma grade discreta, ver Figura 9 para exemplo de grade discreta de dois hiperparâemtros. Dessa forma, todas as combinações possíveis de valores da grade são testados e a validação cruzada é feita para obter um valor de medida preditiva para cada combinação. De maneira similar, o conjunto de hiperparâmetros que apresentar a melhor medida é tomado como a combinação ótima dado as listas de valores propostas.

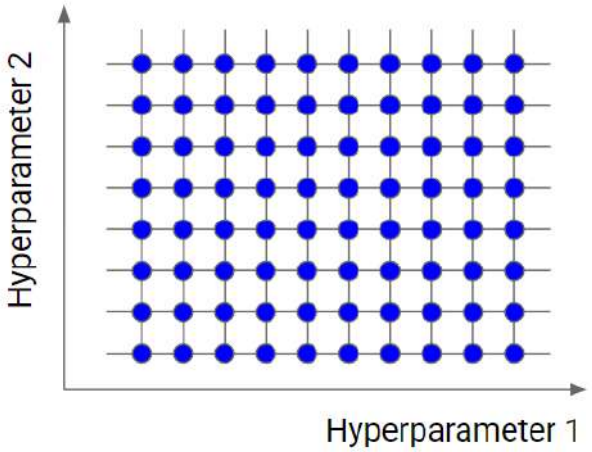

Figura 9 – Exemplo de busca em grade para dois hiperparâmetros

A busca em grade foi tal que λ_{Lasso} e $\lambda_{\text{Elastic Net}}$ variaram de 0 a 1, com incrementos de tamanho $0,001$, já λ_{Ridge} variou de 1 a 100, com incrementos de tamanho $0,1$ e por fim γ variou de 0 a 1, com incrementos de tamanho $0,01$. As estimativas para os hiperparâmetros ótimos na busca em grade foram:

$$\hat{\lambda}_{\text{Lasso}} = 0,001 \tag{4.1}$$

$$\hat{\lambda}_{\text{Ridge}} = 1,2. \tag{4.2}$$

$$\hat{\lambda}_{\text{Elastic Net}} = 0,001 \text{ e } \hat{\gamma} = 0,05. \tag{4.3}$$

Selecionado os hiperparâmetros, os modelos foram ajustados. Para avaliação da acurácia de cada modelo, a medida de qualidade preditiva adotada foi o erro quadrático médio entre as respostas verdadeiras do conjunto de teste e a estimativa obtida a partir do ajuste do modelo nas covariáveis do conjunto de teste. Os erros resultaram:

$$\text{EQM}_{\text{Sem Penal.}} = 5,36. \tag{4.4}$$

$$\text{EQM}_{\text{Lasso}} = 2,12. \tag{4.5}$$

$$\text{EQM}_{\text{Ridge}} = 2,07. \tag{4.6}$$

$$\text{EQM}_{\text{Elastic Net}} = 2,12. \tag{4.7}$$

Observe que o menor erro quadrático médio foi oriundo da regressão Ridge. Ademais, podemos analisar o comportamento gráfico do ajuste do modelo versus a resposta esperada ao redor da reta $y = x$ na Figura 10:

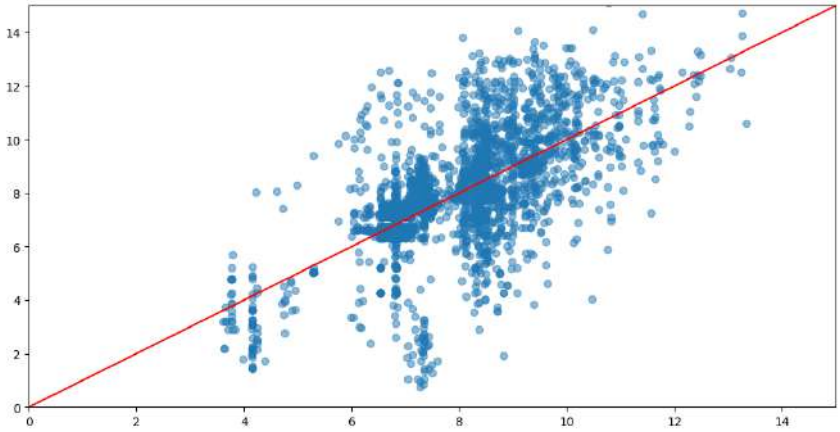

Figura 10 – Gráfico de dispersão do ajuste da regressão Ridge versus a resposta esperada

Repare que o comportamento embora não perfeitamente alinhado com a reta $y = x$, não apresenta grandes desvios ao redor da reta, isso o erro quadrático médio foi considerado satisfatório.

Em sequência, realizou-se uma seleção de variáveis na regressão Ridge utilizando a função *SelectFromModel* da biblioteca *sckit-learn*. Essa função define a média dos coeficientes ajustados da regressão escolhida como um limite, e então variáveis cuja importância, dada pelo coeficiente estimado, ultrapassem esse limite são selecionadas.

O limite determinado pela função foi $0,79$. De 348 variáveis presentes no modelo, 134 foram selecionadas. Os modelos foram ajustados novamente a fim de verificar se a seleção de variáveis conseguiu diminuir a variabilidade do modelo. Entretanto, no que tange ao erro quadrático médio, essa seleção de variável não foi capaz de alcançar este objetivo, conforme ilustrado nas equações 4.8 a 4.11 a seguir em que EQM^* representa o erro após a remoção das variáveis:

$$\text{EQM}^*_{\text{Sem Penal.}} = 2,18 < \text{EQM}_{\text{Sem Penal.}} \tag{4.8}$$

$$\text{EQM}^*_{\text{Lasso}} = 2,19 > \text{EQM}_{\text{Lasso}} \tag{4.9}$$

$$\text{EQM}^*_{\text{Ridge}} = 2,15 > \text{EQM}_{\text{Ridge}} \tag{4.10}$$

$$\text{EQM}^*_{\text{Elastic Net}} = 2,20 > \text{EQM}_{\text{Elastic Net}} \tag{4.11}$$

O ganho foi apenas positivo para regressão linear múltipla sem penalização. Para termos uma mudança mais significativa, poderíamos tomar uma seleção de variáveis mais elaborada, como nas abordagens mais comuns que testam todas combinações de variáveis possíveis de covariáveis para verificar qual explica melhor a variabilidade dos dados.

Depois de aplicar os modelos analisados sob inferência clássica, resta aplicar o modelo Spike-and-Slab cuja abordagem é Bayesiana.

As prioris para os parâmetros foram selecionadas como:

- $\pi_k|p \sim \text{Bernoulli}(p)$;
- $p \sim \text{Beta}(1,1)$;
- $\tau^{-2} \sim \text{Gama}(1,1)$;
- $\sigma^{-2} \sim \text{Gama}(\frac{1}{2}, \frac{3}{8})$.

A inferência via MCMC foi realizada considerando *burn-in* de 5.000 iterações e espaçamento de tamanho 1.

O EQM foi calculado para este modelo estimando as respostas usando o modelo Bayesiano ajustado aplicado no conjunto de teste. Em detalhe, seja ℓ a ℓ-ésima iteração do MCMC realizado no conjunto de treino, gerou-se $y_i|\boldsymbol{\theta}^{(\ell)} \sim N(\mathbf{x}_i^T\boldsymbol{\beta}^{(\ell)}, \sigma^{2(\ell)})$ para cada ℓ, onde $\mathbf{x}_i^T$ são as amostras das covariáveis presente no conjunto de teste. Desta maneira, a resposta estimada $\hat{y}_i$ para cada amostra i foi obtida através da média das ℓ gerações. As estimativas foram então comparadas com os valores verdadeiros e o erro quadrático médio resultou em:

$$\text{EQM}_{\text{Spike-and-Slab}} = 1,98 \tag{4.12}$$

Repare que o EQM resultante do ajuste do modelo *Spike-and-Slab* foi o menor de todos os erros vistos até agora, ou seja, é o que menos erra em média em previsão.

Ademais, as estatísticas mais importantes para este trabalho incluem a probabilidade de inclusão da covariável, que é dada como a média a posteriori do parâmetro π_k, e as estimativas para os efeitos das regressoras β_k. Na Tabela 7, pode-se encontrar as 10 covariáveis com maiores valores absolutos de efeitos de regressora estimados pelo *Spike-and-Slab*.

Tabela 7 – 10 covariáveis com maiores valores absolutos de efeitos de regressora estimados pelo *Spike-and-Slab*

Coluna Original	**Variável**	**Probabilidade de Inclusão** (π_k)	**Estimativa de Efeitos de Regressora** ($\hat{\beta}_k$)
Código SIC	7521	1	-4,71
Código SIC	2026	1	-4,04
Código SIC	8062	1	3,35
Número de Prédios	4	1	3,20
Tipo de ocupação	Agricultura	1	3,15
Zona CRESTA	5	1	3,12
Código SIC	5311	1	2,86
Número de Andares	23	1	2,82
Número de Prédios	12	1	2,56
Código SIC	3663	1	-2,46

Em primeira instância, podemos observar que todas as probabilidades de inclusão apresentadas na tabela são igual a 1. Em mais detalhe, 71 das 348 covariáveis presentes, cerca de 20% das covariáveis, tiveram probabilidade de inclusão exatamente igual a 1; 142 de 348 covariáveis, cerca de 41%, tiveram probabilidade de inclusão maior que $0,99$; 338 das 347 covariáveis, cerca de 97%, tiveram probabilidade de inclusão maior que $0,95$; finalmente, 347 de 348 covariáveis, quase 100%, tiveram probabilidade de inclusão maior que $0,9$.

A única covariável que teve probabilidade de inclusão menor que $0,9$ foi a covariável de longitude, que junto à covariável de latitude, são as únicas covariáveis que não são binárias. Se observarmos a geografia do Chile, o país é bem estreito e faz sentido que a variável longitude não seja tão significativa.

Como as probabilidades de inclusão foram em grande maioria consideráveis, significa que todas as variáveis são importantes para o poder preditivo do modelo e assim podemos utilizar as estimativas de efeitos de regressora para poder interpretar a influência no valor de sinistro.

Ao consultar no site da agência americana de administração da saúde e segurança ocupacional, a OSHA, na ferramenta de pesquisa de código SIC, podemos ver que o código SIC, número 7521, variável que tem a estimativa de efeito de regressora de maior magnitude, diz respeito a estacionamento de automóveis, como garagens, parques de estacionamento, etc. Voltando na base original, podemos conferir que há 46 registros de sinistros desse código SIC, todos entre CLF 156 e CLF 180. Existem bastante registros, pouca variabilidade e baixa magnitude no valor dos sinistros, então faz sentido que a estimativa de efeito de regressora tenha um valor negativo proeminente.

Da mesma maneira, o código SIC 2026 diz respeito a estabelecimentos que se dedicam principalmente ao processamento de leite fluido, creme e produtos relacionados. Aqui temos 223 sinistros registrados, tendo 5 com magnitude mais relevante, entre CLF 1.000 e CLF 100.000, e os demais de valor baixo, de até CLF 100. Existem ainda mais registros, um pouco mais de variabilidade e em geral baixa magnitude, o que implica na estimativa de efeito de regressora ajustada ter um valor negativo também proeminente mas menos que o código SIC 7521.

Outrossim, o código SIC 8062 é relacionado a estabelecimentos que se dedicam principalmente à prestação de serviços médicos e cirúrgicos gerais e outros serviços hospitalares. Repare que a estimativa de efeito de regressora para essa variável é proeminente e positiva. Observando a base de dados original, temos 74 sinistros registrados com esse código SIC, em grande maioria acima de CLF 100.000, podendo chegar a CLF 600.000. Contextualizando, hospitais costumam conter maquinário, equipamentos e estruturas caras, caracterizando apólices de grande risco financeiro, e o modelo comprova isso. É relevante para seguradora evitar subscrever apólices de hospitais a fim de diminuir seu risco de subscrição.

Já para as variáveis número de prédios 4 e 12 que aparecem na tabela 7, a relevância da interpretação do coeficiente de alta magnitude pode não ser verossímil. Ao retornar à base original, vemos que cada uma das variáveis tem apenas 4 registros de sinistros, todos com magnitude bem expressiva, o que explica a estimativa de efeito de regressora altamente positiva, mas devido ao baixo número de ocorrências, o viés é consideravelmente alto. Em contexto, faz sentido um condomínio com 4 ou 12 edifícios apresentar alto risco de subscrição para seguradora, mas a conclusão apenas pelo ajuste do modelo pode ser errôneo.

No que tange à variável zona CRESTA 5, a aparição dela no top 10 em si já é relevante, pois ela descreve uma grande região do Chile, como é possível ver na Figura 11 contendo um mapa da zona topográfica do Chile, em que a região é codificada como *Austral Zone*.

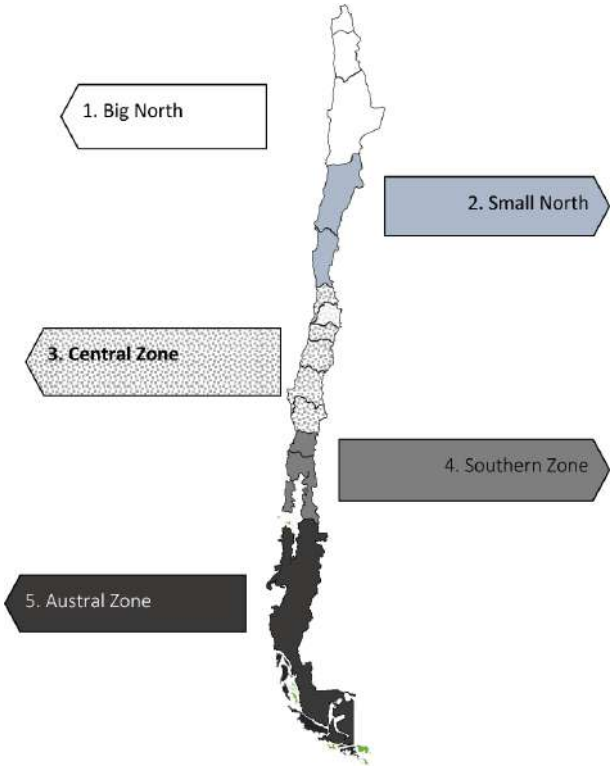

Figura 11 – Mapa das zonas topográficas do Chile

Note que a zona CRESTA 5 compreende a região do extremo sul chileno, a zona austral, como as regiões de Magallanes, Los Rios, Aysen, Los Lagos, etc. E a estimativa de efeito de regressora para essa região é proeminente e positiva. Retomando à base original, vemos que há 2243 sinistros registrados nessa zona e existe uma porcentagem significativa, em torno de 10%, de sinistros acima de CLF 10.000, valor que já as seguradoras já consideram como sinistro vultoso. Essa região embora não seja densamente populosa é uma região onde se encontram penitenciárias, redes de hotéis de ski, indústrias, fazendas de agropecuária, e outros riscos grandes em maior porcentual do que as demais regiões.

Para a seguradora, isso significa que ela deve diminuir o seu risco de subscrição nessa zona e diversificar a carteira, seja ao subscrever mais apólices pequenas ou assumir menos

riscos vultosos nessa zona. Esse tipo de análise é muito produtivo para seguradora e não foi possível capturá-lo imediatamente na análise exploratória de dados, onde a análise de localidade foi bastante contaminada pela magnitude da chamada *key zone*, a zona CRESTA 3.

Para o código SIC 5311, que diz respeito a hipermercados e shoppings, a interpretação é similar às já vistas. Existem 40 sinistros registrados, todos com magnitude alta devido à natureza do risco. Shoppings e hipermercado costumam ter valor segurado imponente atrelado às suas apólices e representam um risco de subscrição muito alto.

De maneira similar, podemos prosseguir tecendo análises sobre cada uma das 348 covariáveis e suas estimativas de efeito de regressora pelo modelo *Spike-and-Slab*. Como o modelo Bayesiano teve o menor EQM dentre os modelos analisados, as inferências feitas a partir dos seus resultados são as mais acuradas. Ainda tentamos ajustar o *Spike-and-Slab* usando apenas as covariáveis selecionadas a partir do Ridge, mas o EQM desse ajuste resultou em 17,56, muito acima de qualquer nível aceitável de acurácia.

As análises apresentadas têm como objetivo nortear a subscrição de seguros. A partir de uma base de dados contendo cerca de 12 mil amostras com 19 covariáveis, sendo 2 quantitativas contínuas e 17 qualitativa nominal, e cada uma contendo vários valores distintos, a análise exploratória desses dados em relação ao impacto com a variável resposta é complexo e visões mais pormenorizadas do risco podem ser facilmente ignoradas.

A aplicação de um modelo de regressão linear com boa acurácia, como o modelo Bayesiano de regressão *Spike-and-Slab*, identificam muito mais facilmente regressoras importantes para a análise, que têm efeito significativo frente à variável resposta, como foi possível constatar ao longo deste capítulo. Dessa forma, essas análises terão valor muito significativo para seguradora no auxílio ao controle de risco de subscrição por seleção adversa de maneira mais simples, uma vez que a aplicação de modelos de regressão é menos custoso que a compra de modelos de riscos catastróficos, tradicionalmente utilizados pelas seguradoras.

5 CONCLUSÕES

Neste estudo, apresentamos uma aplicação de modelos de regressão para seleção de variáveis no contexto ao risco de subscrição que é responsável por grande parte do resultado financeiro de uma seguradora. Uma seguradora com carteira saudável é aquela que subscreve riscos saudáveis com bom histórico de sinistros. Nesse sentido, surge a problemática que permeia este trabalho: a partir de uma base de dados de sinistros históricos contendo diversas variáveis em sua maioria de natureza qualitativa nominal, como fazer inferência acerca de quais variáveis são mais relevantes para o valor de sinistro esperado.

O caminho tomado foi aplicar modelos de regressão, como regressão linear múltipla, com penalização Lasso, Ridge e *Elastic Net* ingênuo e regressão *Spike-and-Slab*. Os resultados da aplicação revelaram que a regressão Ridge e a regressão *Spike-and-Slab* obtiveram melhor poder preditivo devido a apresentarem os menores erros quadráticos médios, com o modelo Bayesiano sendo o superior entre as duas abordagens.

Uma vez determinado o melhor modelo em termos de acurácia de predição, foi possível fazer interpretações sobre as regressoras com maior efeito na resposta, como estacionamentos e estabelecimentos de processamento de leite, que representam um risco positivo para seguradora, e estabelecimentos que prestam serviços médicos e apólices na zona CRESTA 5, que por sua vez indicam um risco negativo para seguradora.

Como pontos fortes deste estudo, podemos destacar que ele introduz uma proposta mais parcimoniosa para a avaliação de covariáveis mais relevantes para a variável de interesse no contexto de seguros patrimoniais de catástrofe, sendo usual que seguradoras arquem financeiramente com a compra de *softwares* que implementem modelos tradicionais para esses contextos, além de *expertise* a custo extremamente elevados para, dentre outras coisas, realizar o trabalho a que este estudo se propõe. Além disso, este estudo performa essa avaliação em covariáveis qualitativas nominais, quando a maioria dos estudos o fazem a partir de covariáveis quantitativas contínuas.

Embora este estudo proponha uma análise da importância de cada covariável e possa auxiliar inicialmente a subscrição de seguros de uma seguradora, realisticamente as análises feitas neste trabalho nunca irão substituir completamente a informação obtida pela compra dos serviços de empresas como a Moody's RMS e a Verisk, que detêm informação de todo mercado global, possuem modelos ajustados e confiáveis, além de uma estrutura gigante para fornecer informação mais precisa que modelos de regressão linear ajustados usando dados de uma única seguradora, que datam de 2017 a 2021.

Como trabalho futuro a ser desenvolvido a partir deste estudo, a seleção de variáveis que performamos usando a média dos coeficientes de regressão ajustados no Ridge poderia ser realizado de maneira diferente. Uma alternativa a ser explorada é seleção de variáveis tradicionais, como os métodos *forward* e *backward*, que exploram combinações

de covariáveis, dentre as quais escolhe-se aquela que resulta no maior poder preditivo observado.

Outrossim, tendo em vista que a variável de localidade é entendida por todo o mercado, incluindo órgãos reguladores da operação de seguros por volta do globo, como sendo pivotal para o sinistro esperado é impreterível que seja incluído uma modelagem espacial e geoprocessamento nos dados apresentados. Aliás, esse é o tipo de modelagem que as empresas de avaliação de riscos catastróficos fazem, assim urge a necessidade de futuramente introduzirmos uma abordagem mais parcimoniosa para esse tipo de modelo, a fim de que as próprias seguradoras tenham estrutura para implementar.

Em geral, este trabalho propõe uma ferramenta simples para o auxílio de subscrição de seguros, a partir da qual as seguradoras podem fazer um julgamento inicial no que tange à relevância das informações coletadas dos segurados a fim de reduzir a seleção adversa e consequentemente seu risco de subscrição.

REFERÊNCIAS

CENSO. 2017. Disponível em: https://www.ine.gob.cl/estadisticas/sociales/censos-de-poblacion-y-vivienda/censo-de-poblacion-y-vivienda.

CMF. **Informe Financiero del Mercado Asegurador**. 2021. 11 p. Disponível em: https://www.cmfchile.cl/portal/estadisticas/617/articles-47810_recurso_1.pdf.

CRESTA. **CRESTA About**. 2023. Disponível em: https://about.cresta.org/.

DEGROOT, M.; SCHERVISH, M. **Probability and Statistics**. Addison-Wesley, 2012. ISBN 9780321500465. Disponível em: https://books.google.com.br/books?id=4TlEPgAACAAJ.

GAMERMAN, D.; LOPES, H. **Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition**. Taylor & Francis, 2006. (Chapman & Hall/CRC Texts in Statistical Science). ISBN 9781584885870. Disponível em: https://books.google.com.br/books?id=yPvECi_L3bwC.

GELMAN, A. et al. **Bayesian Data Analysis, Third Edition**. Taylor & Francis, 2013. (Chapman & Hall/CRC Texts in Statistical Science). ISBN 9781439840955. Disponível em: https://books.google.com.br/books?id=ZXL6AQAAQBAJ.

GeoPandas. **Geopandas**. 2024. Disponível em: https://github.com/geopandas/geopandas.

Governo do Reino Unido. **Standard industrial classification of economic activities (SIC)**. 2024. Disponível em: https://www.gov.uk/government/publications/standard-industrial-classification-of-economic-activities-sic.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. **Technometrics**, Taylor & Francis, v. 12, n. 1, p. 55–67, 1970.

KUHLMAN, D. **A Python Book: Beginning Python, Advanced Python, and Python Exercises**. Platypus Global Media, 2011. ISBN 9780984221233. Disponível em: https://books.google.com.br/books?id=1FL-ygAACAAJ.

LANDERO, H. H. Análisis comparativo del comportamiento sísmico de edificios reticulares rigidizados, no rigidizados y con piso débil. 2003. Disponível em: https://www.researchgate.net/publication/37613129_Analisis_comparativo_del_comportamiento_sismico_de_edificios_reticulares_rigidizados_no_rigidizados_y_con_piso_debil.

MONTGOMERY, D.; PECK, E.; VINING, G. **Introduction to Linear Regression Analysis**. Wiley, 2013. (Wiley Series in Probability and Statistics). ISBN 9781118627365. Disponível em: https://books.google.com.br/books?id=lSyiRZh09oEC.

Pandas. **Package overview**. 2023. Disponível em: https://pandas.pydata.org/pandas-docs/stable/getting_started/overview.html.

PARK, T.; CASELLA, G. The bayesian lasso. **Journal of the American Statistical Association**, Taylor Francis, v. 103, n. 482, p. 681–686, 2008. Disponível em: https://doi.org/10.1198/016214508000000337.

PETRIS, G.; PETRONE, S.; CAMPAGNOLI, P. **Dynamic Linear Models with R**. Springer New York, 2009. (Use R!). ISBN 9780387772370. Disponível em: https://books.google.com.br/books?id=5hGCNQEACAAJ.

REFAEILZADEH, P.; TANG, L.; LIU, H. Cross-validation. In: _____. [S.l.: s.n.], 2016. p. 1–7. ISBN 978-1-4899-7993-3.

SES SUSEP. **SES-SUSEP**. 2023. Disponível em: https://www2.susep.gov.br/menuestatistica/ses/principal.aspx.

Swiss Re. **Global Property casualty insurance premiums expected to more than double to USD 4.3 trillion by 2040, Swiss Re Institute forecasts: Swiss re**. Swiss Re Group, 2021. Disponível em: https://shorturl.at/uJLPT.

TADESSE, M.; VANNUCCI, M. **Handbook of Bayesian Variable Selection**. CRC Press, 2021. (Chapman & Hall/CRC Handbooks of Modern Statistical Methods). ISBN 9781000510256. Disponível em: https://books.google.com.br/books?id=Cn1TEAAAQBAJ.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society. Series B (Methodological)**, [Royal Statistical Society, Wiley], v. 58, n. 1, p. 267–288, 1996. ISSN 00359246. Disponível em: http://www.jstor.org/stable/2346178.

ZOU, H.; HASTIE, T. Regularization and Variable Selection Via the Elastic Net. **Journal of the Royal Statistical Society Series B: Statistical Methodology**, v. 67, n. 2, p. 301–320, 03 2005. ISSN 1369-7412. Disponível em: https://doi.org/10.1111/j.1467-9868.2005.00503.x.