UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO
PROGRAMA DE PÓS-GRADUAÇÃO EM SAÚDE PÚBLICA

DAIANE PRISCILA SAMPAIO BUSSOLA

Distribuições estáveis sob um enfoque Bayesiano: aplicações a dados de saúde pública

Ribeirão Preto

2021

DAIANE PRISCILA SAMPAIO BUSSOLA

Distribuições estáveis sob um enfoque Bayesiano: aplicações a dados de saúde pública

Versão original

Dissertação apresentada ao Programa de Pós-Graduação em Saúde Pública da Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, para obtenção do título de mestre, como parte dos requisitos para obtenção do título de Mestre em Ciências.

Área de concentração: Saúde Pública

Orientador: Prof. Dr. Jorge Alberto Achcar

Ribeirão Preto

2021

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Bussola, Daiane Priscila Sampaio

Distribuições estáveis sob um enfoque Bayesiano: aplicações a dados de saúde pública. Ribeirão Preto, 2021. 103 p.

Dissertação de Mestrado, apresentada à Faculdade de Medicina de Ribeirão Preto/USP. Área de concentração: Saúde Pública.

Orientador: Achcar, Jorge Alberto.

1. Distribuição estável. 2. Dados de saúde pública. 3. Modelos de regressão. 4. Métodos Bayesianos.

Dissertação de autoria de Daiane Priscila Sampaio Bussola, sob o título “**Distribuições estáveis sob um enfoque Bayesiano: aplicações a dados de saúde pública**”, apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Saúde Pública, na área de concentração Epidemiologia e Bioestatística, aprovada em ____ de _____ de _____ pela comissão julgadora constituída pelos doutores:

Prof. Dr.
Instituição
Presidente

Prof. Dr.
Instituição

Prof. Dr.
Instituição

Dedico esta dissertação a Deus por me conceder sabedoria e força. Aos meus pais, Madalena e Sidnei, aos meus irmãos, Matheus e Isabelle, por todo apoio, colaboração e incentivo.

AGRADECIMENTOS

Agradeço primeiramente a Deus, por sua infinita misericórdia! Agradeço a Ele por me conceder saúde, sabedoria, força e luz, para sempre seguir em frente sem desistir. A ti, Senhor, toda honra e toda a glória!

Aos meus pais, Sidnei e Madalena, por não medirem esforços em investir em minha educação desde criança, me ensinaram o caminho do bem, sempre me apoiaram em todas as etapas da minha vida, me guiando para o correto. Eu não chegaria até aqui, se não fosse por vocês. Agradeço por tudo, meu amor por vocês é incondicional.

Aos meus irmãos, Matheus e Isabelle, são exemplos de amor fraternal. Me apoiaram e ajudaram quando mais precisei, vocês são essenciais em minha vida. Sou extremamente grata por tudo, amo vocês demais.

À minha prima, Ana Livia, que com sua inocência de criança, me fez sorrir em momentos difíceis. Te amo, Livinha!

Ao meu orientador, professor Jorge Alberto Achcar, pela oportunidade de realizar este trabalho. Agradeço por todos ensinamentos, pela paciência e por me guiar em todos os passos. Muito obrigada por tudo!

A todos que de alguma forma contribuíram para a realização deste trabalho, o meu muito obrigada e que Deus os abençoe!

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

“É notável uma ciência que começou com jogos de azar tenha se tornado o mais importante objeto do conhecimento humano”
(Pierre Simon Laplace)

RESUMO

BUSSOLA, Daiane Priscila Sampaio. **Distribuições estáveis sob um enfoque Bayesiano**: aplicações a dados de saúde pública. 2021. 103 f. Dissertação (Mestrado em Ciências) – Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, São Paulo, 2021.

Nesta dissertação é apresentada uma abordagem Bayesiana para distribuições estáveis na presença de covariáveis. Esta classe de distribuição, apesar da grande flexibilidade de ajuste para os dados, dado que a distribuição generaliza as distribuições gaussianas para diferentes situações de dados assimétricos e de cauda pesada, não é muito popular nas aplicações, pois não existe uma forma analítica para sua função densidade de probabilidade, que implica em grandes dificuldades para obter estimadores de máxima verossimilhança padrão para os parâmetros do modelo. Alternativamente, o uso de métodos Bayesianos pode ser uma boa alternativa para obter as inferências de interesse, especialmente usando métodos MCMC (*Markov Chain Monte Carlo*), mas em geral há grandes dificuldades para obter convergência dos algoritmos de simulação como o Gibbs padrão ou algoritmos de Metropolis-Hastings em aplicações. Também é apresentado alguma discussão na escolha de distribuições a priori e, a introdução de modelos de regressão para os parâmetros de localização e escala da distribuição estável que pode ser muito útil em aplicações. Aplicações com conjuntos de dados reais são apresentadas para ilustrar a aplicabilidade da abordagem da modelagem proposta.

Palavras-chaves: Distribuição Estável. Dados de Saúde Pública. Modelos de regressão. Métodos Bayesianos.

ABSTRACT

BUSSOLA, Daiane Priscila Sampaio. **Stable distributions under a Bayesian approach**: applications to public health data. 2021. 103 p. Dissertation (Master of Science) – Ribeirão Preto School of Medicine, University of São Paulo, Ribeirão Preto, 2020.

This dissertation presents a Bayesian approach for stable distributions in the presence of covariates. This class of distribution, despite the great adjustment flexibility for the data, since the distribution generalizes the Gaussian distributions for different situations of asymmetric and heavy tail data, is not very popular in applications, as there is no analytical form to its density probability function, which implies great difficulties in obtaining maximum likelihood estimators for the model parameters. Alternatively, the use of Bayesian methods can be a good alternative to obtain inferences of interest, especially using MCMC (Markov Chain Monte Carlo) methods, but in general there are great difficulties to obtain convergence simulation algorithms like the standard Gibbs or Metropolis-Hastings algorithms in applications. Some discussion is also presented in the choice of distributions a priori and the introduction of regression models for the parameters of location and scale of the stable distribution that can be very useful in applications. Applications with real data sets are presented to illustrate the applicability of the proposed modeling approach.

Keywords: Stable Distributions. Public health data. Regression Models. Bayesian methods.

LISTA DE FIGURAS

Figura 1 – Gráficos com as taxas de mortalidade em mulheres, segundo Regionais de Saúde, por 100 mil mulheres, nos anos de 2009 a 2015. Fonte: https://mortalidade.inca.gov.br/MortalidadeWeb	20
Figura 2 – Representação espacial das taxas brutas de mortalidade por câncer de MAMA, por 100.000 mulheres, São Paulo, entre 2009 e 2015. Fonte: https://mortalidade.inca.gov.br/MortalidadeWeb	21
Figura 3 – Densidades da distribuição α -estável com parâmetros $\mu = 0, \sigma = 1$. Imagem à direita: forma fechada para as densidades conhecidas Gauss, Cauchy e Lévy. Imagem à esquerda: funções densidade de probabilidade estáveis para $\alpha = 1.2$ e $\beta = 0, 0.5, 0.8$ e 1	23
Figura 4 – Histogramas: original e escala transformada.	37
Figura 5 – Gráfico dos resíduos (modelo de regressão linear normal - dados FEV)	38
Figura 6 – Traceplots - distribuição estável - sem covariáveis	39
Figura 7 – Traceplots (distribuição estável - presença de covariáveis)	41
Figura 8 – Gráfico de dispersão da resposta FEV versus cada covariável	41
Figura 9 – Gráfico da probabilidade normal para a resposta HbA1c	42
Figura 10 – Gráfico da probabilidade normal para a resposta HbA1c	43
Figura 11 – Gráfico de resíduos (regressão linear normal para resposta HbA1c)	44
Figura 12 – Gráfico de resíduos (regressão linear normal para resposta HbA1c)	45
Figura 13 – Gráfico de resíduos (regressão linear normal para resposta HbA1c)	47
Figura 14 – Boxplot de y_1 para as regiões	49
Figura 15 – Boxplot de y_1 para os anos	50
Figura 16 – Boxplot de y_1 para os meses	50
Figura 17 – Gráfico intervalo 95% $C_r I$ para a média: y_1 vs Região	51
Figura 18 – Gráfico de resíduos fator região para y_1	52
Figura 19 – Gráfico intervalo 95% $C_r I$ para a média: y_1 vs Ano	53
Figura 20 – Gráfico de resíduos fator Ano para y_1	54
Figura 21 – Gráfico intervalo 95% $C_r I$ para a média: y_1 vs Mês	55
Figura 22 – Gráfico de resíduos fator Mês para y_1	56
Figura 23 – Gráfico de dispersão para cada covariável vs y_1	56

Figura 24 – Gráfico de resíduos (regressão linear normal para resposta y_1)	57
Figura 25 – Boxplot de y_2 para as regiões	60
Figura 26 – Boxplot de y_2 para os anos	61
Figura 27 – Boxplot de y_2 para os meses	61
Figura 28 – Gráfico intervalo 95% C_rI para a média: y_2 vs Região	62
Figura 29 – Gráfico de resíduos fator região para y_2	63
Figura 30 – Gráfico intervalo 95% C_rI para a média: y_2 vs Ano	64
Figura 31 – Gráfico de resíduos fator Ano para y_2	64
Figura 32 – Gráfico intervalo 95% C_rI para a média: y_2 vs Mês	65
Figura 33 – Gráfico de resíduos fator Mês para y_2	66
Figura 34 – Gráfico de dispersão para cada covariável vs y_2	66
Figura 35 – Gráfico de resíduos (regressão linear normal para resposta y_2)	67
Figura 36 – Boxplot de y_1 para as regiões - 10/2013 a 06/2015	70
Figura 37 – Boxplot de y_1 para os anos - 10/2013 a 06/2015	71
Figura 38 – Boxplot de y_2 para os meses - 10/2013 a 06/2015	71
Figura 39 – Gráfico intervalo 95% C_rI para a média: y_1 vs Região	72
Figura 40 – Gráfico de resíduos fator região para y_1	73
Figura 41 – Gráfico intervalo 95% C_rI para a média: y_1 vs Ano – 10/2013 a 06/2015	74
Figura 42 – Gráfico de resíduos fator Ano para y_1 – 10/2013 a 06/2015	74
Figura 43 – Gráfico intervalo 95% C_rI para a média: y_1 vs Mês – 10/2013 a 06/2015	75
Figura 44 – Gráfico de resíduos fator Mês para y_1 – 10/2013 a 06/2015	76
Figura 45 – Gráfico de dispersão para cada covariável vs y_1	76
Figura 46 – Gráfico de resíduos (regressão linear normal para resposta y_1 ~ 10/2013 a 06/2015)	77
Figura 47 – Boxplot de y_2 para as regiões - 10/2013 a 06/2015	80
Figura 48 – Boxplot de y_2 para os anos - 10/2013 a 06/2015	81
Figura 49 – Boxplot de y_2 para os meses - 10/2013 a 06/2015	81
Figura 50 – Gráfico intervalo 95% C_rI para a média: y_2 vs Região – 10/2013 a 06/2015	82
Figura 51 – Gráfico de resíduos fator região para y_2 – 10/2013 a 06/2015	83
Figura 52 – Gráfico intervalo 95% C_rI para a média: y_2 vs Ano – 10/2013 a 06/2015	84
Figura 53 – Gráfico de resíduos fator Ano para y_2 – 10/2013 a 06/2015	84
Figura 54 – Gráfico intervalo 95% C_rI para a média: y_2 vs Mês – 10/2013 a 06/2015	85
Figura 55 – Gráfico de resíduos fator Mês para y_2 – 10/2013 a 06/2015	86

Figura 56 – Gráfico de dispersão para cada covariável vs y_2 – 10/2013 a 06/2015 . .	86
Figura 57 – Gráfico de resíduos (regressão linear normal para resposta y_2) - 10/2013 a 06/2015	87

LISTA DE TABELAS

Tabela 1 – Modelo de regressão linear normal (dados FEV)	37
Tabela 2 – Análise Bayesiana - Distribuição α –Estável sem presença de covariáveis.	39
Tabela 3 – Análise Bayesiana - distribuição estável	40
Tabela 4 – Modelo de regressão linear normal (dados HbA1c)	44
Tabela 5 – Análise Bayesiana – distribuição estável sem considerar a presença de covariáveis (dados HbA1c)	45
Tabela 6 – Análise Bayesiana - distribuição estável (dados HbA1c)	46
Tabela 7 – Análise descritiva - 06/2009 a 06/2013 - para y_1 fator região, ($n = 49$ para cada região)	48
Tabela 8 – Análise descritiva - 06/2009 a 06/2013 - para y_1 fator ano	48
Tabela 9 – Análise descritiva - 06/2009 a 06/2013 - para y_1 fator mês	49
Tabela 10 – ANOVA: fator Região – 06/2009 a 06/2013 - para y_1	51
Tabela 11 – ANOVA: fator Ano – 06/2009 a 06/2013 - para y_1	53
Tabela 12 – ANOVA: fator Mês – 06/2009 a 06/2013 - para y_1	55
Tabela 13 – Modelo de regressão linear normal (y_1 - 06/2009 a 06/2013)	57
Tabela 14 – Análise Bayesiana - distribuição estável para y_1	58
Tabela 15 – Análise descritiva - 06/2009 a 06/2013 - para y_2 fator região, ($n = 49$ para cada região)	59
Tabela 16 – Análise descritiva - 06/2009 a 06/2013 - para y_2 fator ano	59
Tabela 17 – Análise descritiva - 06/2009 a 06/2013 - para y_2 fator mês	60
Tabela 18 – ANOVA: fator Região – 06/2009 a 06/2013 - para y_2	62
Tabela 19 – ANOVA: fator Ano – 06/2009 a 06/2013 - para y_2	63
Tabela 20 – ANOVA: fator Mês – 06/2009 a 06/2013 - para y_2	65
Tabela 21 – Modelo de regressão linear normal (y_2 - 06/2009 a 06/2013)	67
Tabela 22 – Análise Bayesiana - distribuição estável para y_2	68
Tabela 23 – Análise descritiva - 10/2013 a 06/2015 - para y_1 fator região, ($n = 21$ para cada região)	69
Tabela 24 – Análise descritiva - 10/2013 a 06/2015 - para y_1 fator ano	69
Tabela 25 – Análise descritiva - 10/2013 a 06/2015 - para y_1 fator mês	70
Tabela 26 – ANOVA: fator Região – 10/2013 a 06/2015 - para y_1 ($n = 21$)	72

Tabela 27 – ANOVA: fator Ano – 10/2013 a 06/2015 - para y_1	73
Tabela 28 – ANOVA: fator Mês – 10/2013 a 06/2015 - para y_1	75
Tabela 29 – Modelo de regressão linear normal (y_1 ~ 10/2013 a 06/2015)	77
Tabela 30 – Análise Bayesiana - distribuição estável para y_1 ~ 10/2013 a 06/2015 . .	78
Tabela 31 – Análise descritiva - 10/2013 a 06/2015 - para y_2 fator região	79
Tabela 32 – Análise descritiva - 10/2013 a 06/2015 - para y_2 fator ano	79
Tabela 33 – Análise descritiva - 10/2013 a 06/2015 - para y_2 fator mês	80
Tabela 34 – ANOVA: fator Região – 10/2010 a 06/2015 - para y_2	82
Tabela 35 – ANOVA: fator Ano – 10/2013 a 06/2015 - para y_2	83
Tabela 36 – ANOVA: fator Mês – 10/2013 a 06/2015 - para y_2	85
Tabela 37 – Modelo de regressão linear normal (y_2 - 10/2013 a 06/2015)	87
Tabela 38 – Análise Bayesiana - distribuição estável para y_2 - 10/2013 a 06/2015 .	88

SUMÁRIO

1	Introdução	16
1.1	<i>Objetivos</i>	18
1.2	<i>Um conjunto de dados especial: câncer de mama no estado de São Paulo</i>	18
2	Distribuição Alfa-Estável	22
2.1	<i>Conceitos básicos</i>	22
2.2	<i>Propriedades da distribuição Alfa-Estável</i>	24
2.2.1	<i>Método de amostragem da Função Característica</i>	25
2.2.2	<i>Método da Máxima Verossimilhança</i>	26
2.3	<i>Algoritmo para simulação de variáveis aleatórias com uma distribuição estável</i>	27
3	Introdução aos métodos Bayesianos	28
3.1	<i>Fórmula de Bayes</i>	28
3.2	<i>Distribuição a Priori</i>	29
3.3	<i>Métodos de simulação para a posteriori</i>	30
3.3.1	<i>Amostrador de Gibbs</i>	31
3.3.2	<i>Algoritmo de Metropolis-Hastings</i>	31
3.4	<i>Inferência Bayesiana para os parâmetros da distribuição estável</i>	32
3.5	<i>Análise Bayesiana assumindo modelos de regressão para os parâmetros de locação e escala de uma distribuição estável</i>	34
4	Aplicações	36
4.1	<i>Efeitos do tabagismo na saúde</i>	36
4.1.1	<i>Abordagem Bayesiana usando uma distribuição estável não considerando a presença de covariáveis</i>	38
4.1.2	<i>Abordagem Bayesiana usando uma distribuição estável na presença de covariáveis</i>	40
4.2	<i>Associação entre níveis de hemoglobina (HbA1c) e algumas covariáveis para diabetes mellitus tipo 2 (T2DM)</i>	42

4.2.1	Abordagem Bayesiana usando uma distribuição estável sem considerar a presença de covariáveis	44
4.2.2	Uma abordagem Bayesiana assumindo uma distribuição estável na presença de covariáveis	45
4.3	<i>Aplicação a dados de câncer de mama</i>	47
4.3.1	Câncer de mama - junho/2009 a junho/2013 - para y_1	48
4.3.2	Câncer de mama - junho/2009 a junho/2013 - para y_2	59
4.3.3	Câncer de mama - outubro/2013 a junho/2015 - para y_1	69
4.3.4	Câncer de mama - outubro/2013 a junho/2015 - para y_2	79
4.3.5	Interpretação dos resultados obtidos	89
5	Conclusão	92
	REFERÊNCIAS	94
	Apêndice A – Códigos OpenBugs - dados FEV	96
A.1	<i>Código sem presença de covariáveis</i>	96
A.2	<i>Código com presença de covariáveis</i>	96
	Apêndice B – Código OpenBugs - dados câncer de mama	99
B.1	<i>Período 06/2009 a 06/2013 - para y_1</i>	99
B.2	<i>Período 06/2009 a 06/2013 - para y_2</i>	100
B.3	<i>Período 10/2013 a 06/2015 - para y_1</i>	101
B.4	<i>Período 10/2013 a 06/2015 - para y_2</i>	102

1 INTRODUÇÃO

Modelos estatísticos de regressão foram introduzidos em 1885 por Francis Galton que estava interessado na área de Antropometria, ou seja, estudo das medidas e relações existentes com os corpos humanos. A análise de regressão é uma técnica estatística que permite estimar o comportamento de uma variável resposta em relação a uma ou mais variáveis explicativas. Por exemplo, estimar a altura média dos filhos a partir da altura de seus pais; estimar a produção média de uma lavoura a partir da quantidade de chuva, quantidade de adubo, etc. Há modelos de regressão chamados de modelos simples, quando existe apenas uma variável resposta e uma variável explicativa, e modelos de regressão múltipla quando existe uma variável resposta e mais de uma explicativa. Casos em que existe mais de uma variável resposta são analisados pela regressão Multivariada, é uma das técnicas estatísticas mais usadas na análise de dados.

Modelos de regressão lineares são frequentemente usados para explorar a relação entre resultados sustentáveis e variáveis independentes. Esses modelos usualmente assumem normalidade para os termos de erro. Com o tamanho amostral muito grande, no entanto, violações da suposição de normalidade podem não afetar visivelmente os resultados, e as transformações das respostas podem, na verdade, influenciar as estimativas de interesse.

Portanto, a metodologia pode ser usada quando a normalidade é satisfeita e isso pode ser observado a partir de um teste de normalidade. Nesse caso, o teste de normalidade deve ser realizado com os resíduos que são estimativas dos erros (DRAPER; SMITH, 1981). Se o requisito de normalidade não for atendido, o modelo de regressão deve ser modificado e usando técnicas de verificação do modelo e possível transformação dos dados, a análise de regressão deve ser realizada atendendo ao requisito de normalidade. Ainda, além da suposição normal com relação aos termos de erro, o modelo de regressão linear assume homoscedasticidade, independência e linearidade. Portanto, a análise de regressão deve ser feita de forma que satisfaça normalidade, homoscedasticidade, independência e linearidade, modificando o modelo de regressão usando técnicas de verificação do modelo, em especial, análises dos resíduos. Essa análise de sensibilidade é feita para garantir a robustez do modelo, como mostram Schmidt e Finan (2018).

Achcar, Achcar e Martinez (2013) mostram que a presença de *outliers* podem afetar significativamente as inferências para modelos de regressão sob suposições de normalidade

para os erros. Como alternativa, nesta dissertação mostramos que é possível o uso de modelos de regressão assumindo distribuições estáveis para obter resultados de inferência robustos, pois tal distribuição apresenta grande flexibilidade para ajuste aos dados. Com esse modelo, o uso de métodos Bayesianos e algoritmos de simulação MCMC (*Markov Chain Monte Carlo*), possibilitam obter inferências para os parâmetros do modelo mesmo com a inexistência de uma forma analítica para sua função densidade. Além disso, o uso de técnicas de *data augmentation* (aumento de dados) nos algoritmos de simulação MCMC é a chave para se obter uma boa performance para o método de simulação MCMC para aplicações usando distribuição estável (DAMIEN; WAKEFIELD; WALKER, 1999). É importante salientar que o uso de técnicas usuais de inferência estatística como o método de máxima verossimilhança apresenta grandes dificuldades para este modelo como será observado na Seção 2.

Dessa forma, propomos neste trabalho para analisar os dados, modelagens estatísticas utilizando a distribuição α -Estável. Esta classe de distribuição foi caracterizada por Paul Lévy em seus estudos relacionados às somas de variáveis aleatórias identicamente distribuídas e independentes, por volta de 1920 (ver, Nolan (2009)).

A distribuição α -Estável (também chamada apenas de distribuição estável), é utilizada quando queremos supor distribuições mais robustas para os dados, como em casos em que a suposição de normalidade para erros em regressão múltipla pode não ser apropriada.

Essa distribuição apresenta grande flexibilidade em termos de regulação dos pesos das caudas, fator que permite bom ajuste em diversas situações. A distribuição normal faz parte dessa classe, porém, outras distribuições estáveis quando comparadas com a normal, podem oferecer modelos mais realísticos.

A distribuição α -estável estudada por Nolan (2009) em geral, não possui forma fechada para sua função densidade de probabilidade, sendo definida por meio de sua função característica. Existem apenas três casos especiais em que é possível escrever sua função densidade: distribuições normal, de Cauchy e Lévy.

As inferências de interesse, serão obtidas empregando métodos Bayesianos. Os sumários a posteriori de interesse serão obtidos via métodos de simulação de Monte Carlo via Cadeias de Markov (MCMC), como o amostrador de Gibbs e/ou o algoritmo de Metropolis-Hastings (ver, por exemplo, Gelfand e Smith (1990); Chib e Greenberg (1995)).

O delineamento deste estudo é considerado epidemiológico observacional analítico ecológico. É epidemiológico, pois estuda a distribuição e os determinantes das doenças ou condições relacionadas à saúde em populações especificadas; observacional analítico são os delineamentos para examinar a existência de associação entre uma exposição e uma doença ou condição relacionada à saúde. Um dos delineamentos de estudo analítico é o ecológico, em que possibilita examinar associações entre exposição e doença/condição relacionada na coletividade (LIMA-COSTA; BARRETO, 2003).

Este documento está organizado em quatro capítulos. No Capítulo 2, apresentamos uma descrição da distribuição α -estável. No Capítulo 3 apresentamos uma introdução aos métodos Bayesianos. No Capítulo 4 consideramos algumas aplicações do modelo proposto a alguns conjuntos de dados de saúde pública. Finalmente, no Capítulo 5, apresentamos as conclusões.

1.1 OBJETIVOS

Objetivo Geral:

Propor modelos estatísticos, com distribuição α -Estável, para aplicações a dados de saúde pública.

Objetivos Específicos:

Aplicar a distribuição estável na presença de covariáveis para dados médicos. Em especial, vamos analisar dados de câncer da mama em períodos distintos no estado de São Paulo.

1.2 UM CONJUNTO DE DADOS ESPECIAL: CÂNCER DE MAMA NO ESTADO DE SÃO PAULO

Para motivação do presente estudo, entre algumas aplicações da metodologia proposta neste estudo que serão apresentadas na seção de aplicações, apresentamos aqui uma aplicação especial considerando dados de câncer de mama no estado de São Paulo, em períodos diferentes na presença de algumas covariáveis. Por serem dados secundários, para este trabalho foi realizada a varredura no DATASUS (Departamento de Informática do Sistema Único de Saúde) e no site da Fundação SEADE (Fundação Sistema Estadual

de Análise de Dados¹). Para a coleta de dados referente às mamografias, novos casos notificados no estado de São Paulo, mensalmente no períodos de junho/2009 a junho/2013 e de outubro/2013 a junho/2015, foram acessadas as informações do banco de dados originário do DATASUS.

Os dados referentes às covariáveis: grau de urbanização, PIB real per capita, renda média, grau de urbanização e número de médicos; que indicam o nível socioeconômico das regiões do estado de São Paulo, foram obtidos do site da Fundação SEADE, que é vinculada à Secretaria de Governo.

As fontes utilizadas para a obtenção dos dados são as seguintes:

- SISMAMA - Sistema de informação do câncer de Mama

Banco de dados integrado ao DATASUS, o SISMAMA é específico para buscas relacionadas ao câncer de mama em todo o Brasil. Nele, encontram-se as seguintes informações: quantidade de exames realizados, exames sem achados, exames contendo nódulo na mama, exames realizados em mulheres sem cirurgia anterior, exames com microcalcificação na mama direita ou esquerda e exames que diagnosticaram linfonodos nas axilas direita ou esquerda. Para este estudo serão considerados apenas a quantidade de exames realizados e os exames em que foram diagnosticados nódulos.

- Fundação SEADE

A Fundação SEADE disponibiliza um menu “Produtos > Informações dos Municípios Paulistas (IMP)”, onde é possível selecionar as regiões administrativas e encontrar informações que possibilitam obter um perfil das mesmas, como indicadores relacionados à: economia, educação, demografia, saúde, trabalho, saneamento, entre tantos outros, que são provenientes de levantamentos feitos pelo estado.

Foram considerados para este estudo as regiões administrativas do estado de São Paulo e a Região Metropolitana de São Paulo (RMSPL), definidas, atualmente, pelo Decreto 60.135 de 10 de fevereiro de 2014, em que foi criada a RA de Itapeva e alteradas as composições da RA de Sorocaba. As 15 Regiões Administrativas são: Araçatuba (ARCTB), Barretos (BRRS), Bauru (BAURU), Campinas (CMPNS), Central (CNTRL), Franca (FRNCA), Itapeva (ITPVA), Marília (MRLIA), Presidente Prudente (PSTPD),

¹ Os dados podem ser acessados em: [http : //www.seade.gov.br/lista – produtos/](http://www.seade.gov.br/lista-produtos/) Último acesso: 13/06/2019.

Registro (RGSTR), Ribeirão Preto (RBPRT), Santos (SNTOS), São José do Rio Preto (SJDRP), São José dos Campos (SJCPS) e Sorocaba (SRCBA). .

No estado de São Paulo, onde a taxa de mortalidade para câncer de mama ainda é considerada elevada, diversos fatores podem estar relacionados, inclusive melhora nas condições de vida e no acesso aos recursos de saúde. Na Figura 1 é apresentado o gráfico com as taxas de mortalidade por câncer de mama nas regionais de saúde do estado. Note que mesmo com alguns decaimentos temporais essas taxas ainda são relativamente altas.

Figura 1 – Gráficos com as taxas de mortalidade em mulheres, segundo Regionais de Saúde, por 100 mil mulheres, nos anos de 2009 a 2015. Fonte: <https://mortalidade.inca.gov.br/MortalidadeWeb>

Na Figura 2 é mostrado a representação espacial² das taxas brutas de mortalidade por câncer de mama, por 100 mil mulheres, São Paulo, entre 2009 e 2015.

² Mapa pode ser acessado em: <https://mortalidade.inca.gov.br/MortalidadeWeb/pages/Modelo06/consultar.xhtml/panelResultado>

Figura 2 – Representação espacial das taxas brutas de mortalidade por câncer de MAMA, por 100.000 mulheres, São Paulo, entre 2009 e 2015. Fonte: <https://mortalidade.inca.gov.br/MortalidadeWeb>

Portanto, nota-se que há necessidade de estudos do câncer de mama no estado de São Paulo para que possam ser tomadas medidas de divulgação e alerta para o risco que pode ser evitado através da prevenção via exames de rastreamento.

2 DISTRIBUIÇÃO ALFA-ESTÁVEL

2.1 CONCEITOS BÁSICOS

Uma ampla classe de distribuições que engloba a distribuição gaussiana é dada pela classe de distribuições estáveis. Essa grande classe define famílias locação-escala que são fechadas sob convolução. A condição *estável* refere-se ao fato de que se somarmos variáveis aleatórias estáveis independentes, identicamente distribuídas (i.i.d.), com distribuição α -estável, a variável aleatória que representar essa soma, também terá distribuição α -estável.

Essa classe de distribuições é descrita por quatro parâmetros α, β, μ e σ . O parâmetro $\alpha \in (0, 2]$ refere-se à curtose, ou seja, ele define o grau de concentração das observações em torno da média da distribuição, e quando $\alpha = 2$ essa classe reduz-se à distribuição Gaussiana. O parâmetro $\beta \in [-1, 1]$ refere-se à assimetria da distribuição, se $\beta = 0$ a distribuição é simétrica, se $\beta < 0$ a distribuição é simétrica à direita, e se $\beta > 0$ a distribuição é simétrica à esquerda. Esses dois parâmetros α e β determinam a forma da distribuição. Os parâmetros de locação da distribuição e dispersão (escala da distribuição) são, respectivamente, $\mu \in (-\infty, \infty)$ e $\sigma \in (0, \infty)$ (ver (NOLAN, 2009)).

Distribuições estáveis são denotadas por $S_\alpha(\beta, \mu, \sigma)$. Se uma variável aleatória $X \sim S_\alpha(\beta, \mu, \sigma)$, então $Z = (X - \mu)/\sigma \sim S_\alpha(\beta, 0, 1)$ (ver Nolan (2009), Gnedenko e Kolmogorov (1968), Skorohod (1961)).

Definição 1. *Sejam as variáveis aleatórias independentes e identicamente distribuídas $X_1, X_2, \dots, X_n$ e X , então X é dita que segue uma distribuição α -Estável se existir uma constante positiva C_n e um número real D_n , tal que a seguinte relação é observada:*

$$X_1 + X_2 + \dots + X_n \stackrel{d}{=} C_n X + D_n \quad (1)$$

em que, $(\stackrel{d}{=})$ representa a igualdade em distribuição. A equação 1 é chamada de estritamente estável se $D_n = 0 \forall n$, (NOLAN, 2009).

Em geral, a distribuição α -estável não possui forma fechada para as expressões das funções densidades de probabilidade e acumulada, mas há três casos especiais dessa distribuição em que é possível escrever uma forma fechada da função densidade de probabilidade, (veja, por exemplo, Buckle (1995)):

1. Caso distribuição de Gauss ($\alpha = 2$)

$$f(x) = \frac{1}{2\sigma\sqrt{\pi}} e^{-\frac{(x-\mu)^2}{4\sigma^2}}$$

2. Caso distribuição de Cauchy ($\alpha = 1, \beta = 0$)

$$f(x) = \frac{\sigma}{\pi((x-\mu)^2 + \sigma^2)}$$

3. Caso distribuição de Lévy ($\alpha = \frac{1}{2}, \beta \neq 1$)

$$f(x) = \frac{\sqrt{\sigma}}{\sqrt{2\pi}(x-\mu)^{\frac{3}{2}}} e^{-\frac{\sigma}{2(x-\mu)}}$$

Uma grande dificuldade associada às distribuições estáveis, é que em geral não existe forma fechada simples para funções densidade probabilidade. Contudo, sabe-se que as funções densidade de probabilidade de distribuições estáveis são contínuas e unimodais (ver [Skorohod \(1961\)](#), [Ibragimov e Cernin \(1959\)](#) e [Kanter \(1976\)](#)). Também, o suporte de todas distribuições estáveis é dado em $(-\infty, \infty)$, exceto para $\alpha < 1$ e $|\beta| = 1$, quando o suporte é $(-\infty, 0)$ para $\beta = 1$ e $(0, \infty)$ para $\beta = -1$ (ver [Nolan \(2009\)](#)). É importante salientar que se $\alpha < 1$, a variância é infinita e a média da distribuição estável não existe.

Figura 3 – Densidades da distribuição α -estável com parâmetros $\mu = 0, \sigma = 1$. Imagem à direita: forma fechada para as densidades conhecidas Gauss, Cauchy e Lévy. Imagem à esquerda: funções densidade de probabilidade estáveis para $\alpha = 1.2$ e $\beta = 0, 0.5, 0.8$ e 1 .

Pelos gráficos da Figura 3, é observado que o parâmetro α controla a intensidade local e o parâmetro β controla a assimetria da distribuição. A Figura 3 também traz a densidade de uma distribuição normal (linha azul), e a densidade de uma distribuição de Cauchy (linha vermelha).

Quando $\alpha > 1$, a média da distribuição existe e é igual a μ . Em geral, o n -ésimo momento de uma variável aleatória estável é infinito se, e somente se, $p < \alpha$. Quando o parâmetro de simetria β é positivo, a distribuição é simétrica à direita, i.e., a cauda à direita é mais pesada, veja a o gráfico à direita na Figura 3. Quando β é negativo, a simetria é à esquerda. Quando $\beta = 0$, a distribuição é simétrica em μ . Conforme α se aproxima de 2, β perde efeito e a distribuição se aproxima da distribuição Gaussiana, independentemente de β . Os dois últimos parâmetros, σ e μ , são os parâmetros usuais de dispersão e locação, i.e. σ determina a amplitude e μ o pico da densidade. Para $\sigma = 1$ e $\mu = 0$ a distribuição é chamada de estável padrão.

Embora essa classe de distribuições seja uma boa alternativa para modelagem de dados em diferentes áreas, geralmente tem-se dificuldades em obter estimativas sob a abordagem da inferência clássica, devido a falta de uma expressão de forma fechada para as funções densidade de probabilidade. Uma possibilidade em aplicações, é obter a f.d.p da fórmula da função característica inversa, onde a função característica é dada por,

$$\begin{aligned} \phi_{stable}(\mathbf{t}; \alpha, \sigma, \beta, \mu) &= \mathbf{E}[e^{itX}] = \\ &\begin{cases} \exp(i\mu t - |\sigma t|^\alpha (1 - i\beta(\text{sign}t)\tan\frac{\pi\alpha}{2})) & \alpha \neq 1 \\ \exp(i\mu t - \sigma|t|(1 + i\beta\frac{2}{\pi}(\text{sign}t)\ln|t|)) & \alpha = 1 \end{cases} \end{aligned} \quad (2)$$

onde

$$\text{sign}t = \begin{cases} 1, & t > 0 \\ 0, & t = 0 \\ -1, & t < 0 \end{cases}$$

Assim, a densidade é obtida da relação,

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \Phi(t) dt$$

onde $\Phi(t)$ é a função característica. Em aplicações, em geral, precisamos usar métodos numéricos para resolver essa integral, muitas vezes exigindo grande tempo computacional.

2.2 PROPRIEDADES DA DISTRIBUIÇÃO ALFA-ESTÁVEL

Existem várias propriedades básicas da distribuição α -Estável:

- A cauda da função densidade decai como uma função de potência,

$$P(|X| > x) \propto C.x^{-\alpha}, x \rightarrow \infty$$

para uma constante C .

- Momentos de ordem p satisfazem a propriedade

$$E|X|^p < \infty, 0 < p < \alpha, \quad E|X|^p = \infty, p \geq \alpha$$

- A esperança é dada por

$$E[X] = \mu, \alpha > 1, \quad E[X] = \infty, \alpha \leq 1$$

- A propriedade de estabilidade é preservada sob transformação linear. Supor que uma variável aleatória X_i que seja i.i.d., tal que $X_i \sim S_\alpha(\sigma_i, \beta_i, \mu_i)$, então

- $Y = \sum_{i=1}^n X_i$ tem uma distribuição α -Estável com índice de estabilidade α e parâmetros

$$\beta = \frac{\sum_{i=1}^n \beta_i \sigma_i^\alpha}{\sum_{i=1}^n \sigma_i^\alpha}, \sigma = \left(\sum_{i=1}^n \sigma_i^\alpha \right)^{\frac{1}{\alpha}}, \mu = \sum_{i=1}^n \mu_i$$

- $Y = X_1 + a$ tem uma distribuição α -Estável com índice de estabilidade α e parâmetros

$$\beta = \beta_1, \sigma = \sigma_1, \mu = \mu_1 + a$$

- $Y = aX_1, a \neq 0$ tem uma distribuição α -Estável com índice de estabilidade α e parâmetros

$$\beta = (\text{sign } a)\beta_1, \sigma = |a|\sigma_1, \mu = \begin{cases} a\mu_1 & \alpha \neq 1 \\ a\mu_1 - \frac{2}{\pi}a(\ln(a))\sigma_1\beta_1 & \alpha = 1 \end{cases}$$

2.2.1 MÉTODO DE AMOSTRAGEM DA FUNÇÃO CARACTERÍSTICA

Definição 2. Assumir uma amostra aleatória de tamanho N , $x = x_1, x_2, \dots, x_N$, com uma distribuição α -Estável. Então, define-se a função característica amostral por,

$$\hat{\phi}_x(u) = \frac{1}{N} \sum_{j=1}^N e^{iux_j}$$

Pela lei dos grandes números, $\hat{\phi}x(u)$ é um estimador consistente da função característica $\phi x(u)$. Realizando uma transformação, temos para todos os α

$$|\phi x(u)| = \exp(-\sigma^\alpha |u|^\alpha)$$

Portanto,

$$-\log|\phi x(u)| = \sigma^\alpha |u|^\alpha$$

Assumindo $\alpha \neq 1$, escolhem-se dois valores diferentes de zero, u_k , $k = 1, 2$, então $-\log|\hat{\phi}x(u_k)| = \sigma^\alpha |u_k|^\alpha$. Resolvendo essas duas equações e tendo $\hat{\alpha}, \hat{\sigma}$, segue,

$$\hat{\sigma} = \frac{\log \frac{\log|\hat{\phi}(u_1)|}{\log|\hat{\phi}(u_2)|}}{\log \frac{|u_1|}{|u_2|}}$$

$$\log \hat{\sigma} = \frac{\log|u_1| \log(-\log|\hat{\phi}(u_2)|) - \log|u_2| \log(-\log|\hat{\phi}(u_1)|)}{\log \frac{|u_1|}{|u_2|}}$$

A estimação de $\hat{\beta}$ e $\hat{\mu}$ baseado nas partes imaginária e real da função característica,

$$\operatorname{Re}(\phi x(u)) = \exp(-|\sigma u|^\alpha) \cos(\mu u + |\sigma u|^\alpha \beta(\operatorname{sign} u) \tan \frac{\pi \alpha}{2}),$$

$$\operatorname{Im}(\phi x(u)) = \exp(-|\sigma u|^\alpha) \sin(\mu u + |\sigma u|^\alpha \beta(\operatorname{sign} u) \tan \frac{\pi \alpha}{2})$$

Então, temos

$$\left(\arctan \frac{\operatorname{Im}(\phi x(u))}{\operatorname{Re}(\phi x(u))} \right) = \mu u + |\sigma u|^\alpha (\operatorname{sign} u) \tan \frac{\pi \alpha}{2}$$

Baseados em $\hat{\alpha}$, $\hat{\sigma}$ e em dois valores diferentes de zero u_k , $k = 3, 4$, podemos resolver o sistema de equações para obter as estimativas de $\hat{\beta}$ e $\hat{\mu}$,

$$\hat{\mu} = \frac{u_4 \hat{\alpha} \arctan \frac{\operatorname{Im}(\phi x(u_3))}{\operatorname{Re}(\phi x(u_3))} - u_3 \hat{\alpha} \arctan \frac{\operatorname{Im}(\phi x(u_4))}{\operatorname{Re}(\phi x(u_4))}}{u_3 u_4^{\hat{\alpha}} - u_4 u_3^{\hat{\alpha}}}$$

$$\hat{\beta} = \frac{u_4 \arctan \frac{\operatorname{Im}(\phi x(u_3))}{\operatorname{Re}(\phi x(u_3))} - u_3 \arctan \frac{\operatorname{Im}(\phi x(u_4))}{\operatorname{Re}(\phi x(u_4))}}{\hat{\sigma}^{\hat{\alpha}} \tan \frac{\pi \hat{\alpha}}{2} (u_4 u_3^{\hat{\alpha}} - u_3 u_4^{\hat{\alpha}})}$$

2.2.2 MÉTODO DA MÁXIMA VEROSSIMILHANÇA

Definição 3. Usando estimação por Máxima Verossimilhança (MV) para obter os estimadores dos parâmetros da distribuição α -estável baseada em um vetor de observações

$x = (x_1, x_2, \dots, x_n)$, as estimativas de MV dos componentes do vetor de parâmetros $\theta = (\alpha, \sigma, \beta, \mu)$ são obtidas maximizando-se a função log-verossimilhança

$$L_\theta(x) = \sum_{i=1}^n \log \tilde{f}(x_i; \theta)$$

onde $\tilde{f}(\cdot; \theta)$ é a função densidade estável.

O til denota o fato de não existir uma forma fechada explícita da função densidade de probabilidade que deve ser aproximada numericamente. A aproximação da função densidade pode ser obtida pela transformação inversa da função característica pela transformada rápida de *Fourier*. As estimativas pela Máxima Verossimilhança são quase sempre as mais precisas, seguida de perto pela estimação por métodos de regressão, método quantil e pelo método dos momentos. Apesar disso, essa técnica é quase que inviável nas aplicações com dados reais dada a dificuldade na obtenção da aproximação da função densidade a partir de expressões matemáticas nem sempre simples.

2.3 ALGORITMO PARA SIMULAÇÃO DE VARIÁVEIS ALEATÓRIAS COM UMA DISTRIBUIÇÃO ESTÁVEL

Gerar uma variável aleatória U uniformemente distribuída em $(-\frac{\pi}{2}, \frac{\pi}{2})$ e uma variável aleatória exponencial E independente com média 1. Para $\alpha \neq 1$, calcule

$$X = S_{\alpha, \beta} \frac{\sin(\alpha(U + B_{\alpha, \beta}))}{(\cos(U))^{\frac{1}{\alpha}}} \left(\frac{\cos(U - \alpha(U + B_{\alpha, \beta}))}{E} \right)^{\frac{1-\alpha}{\alpha}}$$

onde, $B_{\alpha, \beta} = \frac{\arctan(\beta \tan(\frac{\pi\alpha}{2}))}{\alpha}$, e $S_{\alpha, \beta} = (1 + \beta^2 \tan^2(\frac{\pi\alpha}{2}))^{\frac{1}{2\alpha}}$.

Para $\alpha = 1$, calcule

$$X = \frac{2}{\pi} \left[\left(\frac{\pi}{2} + \beta U \right) \tan(U) - \beta \log \left(\frac{\frac{\pi}{2} E \cos(U)}{\frac{\pi}{2} + \beta U} \right) \right]$$

Generalização da escala e locação

$$Y = \begin{cases} \sigma X + \mu, & \alpha \neq 1, \\ \sigma X + \frac{2}{\pi} \beta \sigma \log(\sigma) + \mu, & \alpha = 1 \end{cases}$$

3 INTRODUÇÃO AOS MÉTODOS BAYESIANOS

Métodos Bayesianos são considerados alternativas eficientes e poderosas para análise de dados. Diferente do método clássico, em que os parâmetros do modelo são considerados constantes desconhecidas, o método Bayesiano considera todos os parâmetros quantidades aleatórias, além do fato de que permite a inclusão da informação de um especialista aos dados.

A base da inferência Bayesiana é a Fórmula de Bayes, que associa os dados com a informação a priori, para então obter a distribuição a posteriori (priori combinada à informação dos dados) onde é realizada a estimação dos parâmetros (inferência) (ver, por exemplo, [Box e Tiao \(1973\)](#); [Paulino, Turkman e Murteira \(2003\)](#)).

3.1 FÓRMULA DE BAYES

Definição 4. *Considere uma partição do espaço amostral Ω , que contém uma sequência de eventos $A_1, A_2, \dots, A_k$, mutuamente exclusivos e exaustivos. Ou seja, $\bigcup_{j=1}^k A_j = \Omega$ e $A_i \cap A_j = \phi$ (conjunto vazio), para $i \neq j$ tal que $P\left(\bigcup_{j=1}^k A_j\right) = \sum_{j=1}^k P(A_j) = 1$. Assim, para qualquer evento $B (B \subset \Omega)$, temos*

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^k P(B|A_j)P(A_j)} \quad (3)$$

para i variando de 1 até k .

Suponha, um vetor de dados $y = (y_1, \dots, y_n)'$ e θ (quantidades desconhecidas) os parâmetros de uma distribuição de probabilidade associada com a variável aleatória Y_i com valores observados $y_i, i = 1, \dots, n$.

Assumindo uma amostra aleatória $y = (y_1, \dots, y_n)$ independentes e identicamente distribuídos, supor que o vetor aleatório Y possui uma distribuição conjunta dada pela função densidade de probabilidade $f(y|\theta)$, que também é conhecida como função de verossimilhança para θ . Uma vez que os dados foram observados e uma distribuição a

priori foi atribuída para θ , dada por $\pi(\theta)$, de (3) tem-se a distribuição a posteriori para θ_i dado y ,

$$\pi(\theta_i|y) = \frac{f(y|\theta_i)\pi(\theta_i)}{\sum_{j=i}^k f(y|\theta_j)\pi(\theta_j)} \quad (4)$$

onde o parâmetro θ também é considerado como uma quantidade aleatória, no enfoque Bayesiano.

Suponha que o parâmetro θ assume valores contínuos num dado intervalo; considerando uma amostra aleatória $y = (y_1, \dots, y_n)$, podemos escrever (4) como,

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta} \quad (5)$$

em que a integral no denominador é definida no intervalo de variação de θ .

Em geral não é necessário calcular a integral no denominador, de modo que a distribuição a posteriori é proporcional à distribuição a priori multiplicada pela verossimilhança, ou seja:

$$\pi(\theta | y) \propto f(y | \theta) \pi(\theta) \quad (6)$$

3.2 DISTRIBUIÇÃO A PRIORI

No enfoque Bayesiano, a distribuição a priori representa o que se conhece sobre os parâmetros a serem estimados, antes da avaliação dos dados. É necessário ter cautela ao definir a distribuição a priori, pois se a mesma não for definida corretamente, pode-se chegar a deduções errôneas. Uma distribuição a priori para um parâmetro θ pode ser descrita de várias formas, podendo ocorrer a partir de procedimentos subjetivos ou objetivos.

A distribuição priori conjugada é uma priori informativa, em que a distribuição a priori e a posteriori pertencem à mesma classe de distribuições. A passagem da priori para a posteriori envolve apenas uma mudança nos parâmetros, não necessitando de cálculos adicionais (ver, por exemplo, [Paulino, Turkman e Murteira \(2003\)](#)).

Na prática, quando não é possível obter o conhecimento do especialista (não existe ou não é fidedigno), são utilizadas as prioris não informativas, isto é, prioris de referência “neutras” (ver, exemplo, [Box e Tiao \(1973\)](#)). Com o uso dessas distribuições a priori não-informativas encontramos resultados similares aos resultados obtidos usando inferência

clássica, pois utilizando uma distribuição a priori não informativa, a inferência baseia-se apenas na informação dos dados amostrais.

3.3 MÉTODOS DE SIMULAÇÃO PARA A POSTERIORI

Nesta seção será apresentado um resumo de técnicas de simulação usadas para se obter sumários a posteriori de interesse. A palavra simulação refere ao tratamento de um problema real através da reprodução em um ambiente, geralmente computacional, controlado pelo pesquisador (ver por exemplo, [Gamerman e Lopes \(2006\)](#)).

Nas aplicações, em geral, é necessário resolver integrais Bayesianas que não apresentam solução analítica na obtenção de sumários a posteriori de interesse. Na prática, nem sempre os modelos utilizados são simples para se obter esses sumários. Ainda que se tenha uma priori e uma verossimilhança simples, a junção das mesmas pode produzir uma distribuição a posteriori complexa.

Existem métodos que são baseados em amostragem, por exemplo, o método de Monte Carlo via Cadeias de Markov (*Markov Chain Monte Carlo (MCMC)*), que, com o avanço computacional, passaram a ser muito utilizados. Quando os métodos de simulação direta são inviáveis e existe complexidade na forma matemática das distribuições a posteriori, é de interesse utilizar os métodos MCMC (ver por exemplo, [West e Harrison \(1997\)](#)).

O método MCMC consiste em simular uma variável aleatória através de uma cadeia de Markov, na qual sua distribuição assintoticamente se aproxima da distribuição a posteriori de interesse (ver por exemplo, [Bernardo e Smith \(2000\)](#)). A cadeia de Markov é um processo estocástico em que o próximo estado da cadeia depende apenas do estado atual e dos dados. Porém, o estado inicial é descartado após um período de aquecimento, chamado de *burn-in*.

Entre os métodos de simulação que usam o MCMC estão os amostradores de Gibbs e o algoritmo de Metropolis-Hastings, ambos simulam amostras da distribuição a posteriori conjunta a partir das distribuições condicionais (ver por exemplo, [Gelfand e Smith \(1990\)](#); [Chib e Greenberg \(1995\)](#), [Bernardo e Smith \(2000\)](#)).

O amostrador de Gibbs, permite gerar amostras da distribuição a posteriori conjunta desde que seja especificado o conjunto completo de condicionais, geralmente assumindo que uma distribuição condicional depende apenas de algum subconjunto da “vizinhança” das

variáveis, ou seja, as distribuições condicionais completas devem possuir formas fechadas ou conhecidas (ver por exemplo, [Gelfand e Smith \(1990\)](#), [Chib e Greenberg \(1955\)](#)), [Bernardo e Smith \(2000\)](#). Em contrapartida, o algoritmo de Metropolis-Hastings permite gerar amostras da distribuição a posteriori conjunta com distribuições condicionais completas possuindo ou não uma forma conhecida ou fechada.

3.3.1 AMOSTRADOR DE GIBBS

Supor que interessa-se em obter inferências da distribuição a posteriori conjunta, $\pi(\theta|y)$, sendo $\theta = (\theta_1, \dots, \theta_k)$. Para isso, simulam-se quantidades aleatórias de distribuições condicionais completas $\pi(\theta_i|y, \theta_{(i)})$ que produzem uma cadeia de Markov, isto é, a cadeia irá sempre se mover para um novo valor, não existindo um mecanismo de aceitação-rejeição.

Se as distribuições condicionais completas forem completamente conhecidas, podemos ter que dado um conjunto inicial arbitrário de valores para o vetor de parâmetros $\theta = \theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_k^{(0)}$. Sendo assim, o algoritmo é dado por:

$$\begin{aligned}
 (i) \quad & \text{Gerar } \theta_1^{(1)} \text{ de } \pi(\theta_1 | y, \theta_2^{(0)}, \dots, \theta_k^{(0)}); \\
 (ii) \quad & \text{Gerar } \theta_2^{(1)} \text{ de } \pi(\theta_2 | y, \theta_1^{(1)}, \theta_3^{(0)}, \dots, \theta_k^{(0)}); \\
 (iii) \quad & \text{Gerar } \theta_3^{(1)} \text{ de } \pi(\theta_3 | y, \theta_1^{(1)}, \theta_2^{(1)}, \theta_4^{(0)}, \dots, \theta_k^{(0)}); \\
 & \vdots \\
 (k) \quad & \text{Gerar } \theta_k^{(1)} \text{ de } \pi(\theta_k | y, \theta_1^{(1)}, \theta_2^{(1)}, \dots, \theta_{k-1}^{(1)});
 \end{aligned} \tag{7}$$

Substitua os valores iniciais por $\theta^{(1)} = \theta_1^{(1)}, \theta_2^{(1)}, \dots, \theta_k^{(1)}$ para uma nova realização. Assim, cada iteração será completa após k movimentos ao longo dos eixos coordenados das componentes de θ . Após a obtenção da convergência, as amostras resultantes formam uma amostra de $\pi(\theta|y)$.

3.3.2 ALGORITMO DE METROPOLIS-HASTINGS

Quando as distribuições condicionais não são facilmente identificadas, devemos usar o algoritmo de Metropolis-Hastings ou métodos de amostragem por importância.

Supor que pretendemos simular amostras da densidade não-regular $\pi(\theta_{(i)}|\theta_{(i)})$, em que $\theta_{(i)} = (\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_k)$, definir o núcleo de transição $q(\theta, \beta)$ da distribuição $p(\theta)$ que representa $\pi(\theta_{(i)}|\theta_{(i)})$ e que transforma θ em β . Se θ é uma variável real com amplitude em toda reta $\mathbb{R}$, podemos construir q tal que $\beta \leftarrow \theta + \sigma z$, com $Z \sim N(0, \sigma^2)$, em que σ^2 reflete a variância condicional de θ em $p(\theta)$.

Se θ é limitado com amplitude (a, b) usar uma transformação que leva (a, b) em $(-\infty, \infty)$ e a partir daí usar o núcleo de transição q e aplicar o algoritmo de Metropolis-Hastings para a densidade da variável transformada. Desta forma o algoritmo é dado por:

(i) Iniciar com um valor $\theta^{(0)}$ e indicador de estágio, $j = 0$;

(ii) Gerar um ponto β do núcleo de transição $q(\theta^{(j)}, \beta)$;

(iii) Atualizar $\theta^{(j)}$ por $\theta^{(j+1)} = \beta$ com probabilidade,

$$p = \min \left\{ 1, \frac{p(\beta)q[\theta^{(j)}, \beta]}{p[\theta^{(j)}]q[\beta, \theta^{(j)}]} \right\}$$

Ficar com $\theta^{(j)}$ com probabilidade $1 - p$;

(iv) Repetir os estágios (ii) e (iii) até conseguir uma distribuição estacionária.

É importante observar que: o algoritmo de Metropolis-Hastings é especificado pela densidade candidata para geração $q(x, y)$; se um valor candidato é rejeitado, então o valor atual é considerado na próxima etapa; o cálculo de p no passo (iii) não depende da constante normalizadora.

Se a densidade candidata para a geração das amostras é simétrica, isto é, $q(x, y) = q(y, x)$, a probabilidade de movimento se reduz à $\frac{p(\beta)}{p[\theta^{(j)}]}$. Assim, se $p(\beta) > p[\theta^{(j)}]$, a cadeia se move para β ; caso contrário, ela se move para β com probabilidade $\frac{p(\beta)}{p[\theta^{(j)}]}$, ou seja, um salto na direção “ascendente” é sempre aceito, porém, um salto na direção “descendente” é aceito com uma dada probabilidade (ver, por exemplo, [Bernardo e Smith \(2000\)](#), Chib).

3.4 INFERÊNCIA BAYESIANA PARA OS PARÂMETROS DA DISTRIBUIÇÃO ESTÁVEL

Uma alternativa ao uso da inferência clássica, é utilizar métodos Bayesianos. Contudo, o custo computacional pode ser mais alto para obter sumários a posteriori de interesse.

Uma alternativa é usar variáveis latentes artificiais (*data augmentation*) que podem melhorar a simulação computacional de amostras da distribuição conjunta a posteriori de interesse (ver, por exemplo, [Damien, Wakefield e Walker \(1999\)](#), [Tanner e Wong \(1987\)](#)).

Desse modo, uma análise Bayesiana de distribuições estáveis foi introduzida por [Buckle \(1995\)](#) usando métodos MCMC com introdução de variáveis latentes. O uso de métodos Bayesianos com simulação MCMC pode trazer grande flexibilidade ao considerar variáveis latentes onde amostras dessas variáveis são simuladas em cada passo dos algoritmos de Gibbs ou Metropolis-Hastings.

Considerando uma variável latente ou uma variável auxiliar, [Buckle \(1995\)](#) provou um teorema que é útil para simular amostras da distribuição conjunta a posteriori para os parâmetros α, β, γ e δ . Esse teorema estabelece que a distribuição estável para uma variável aleatória Z definida em $(-\infty, \infty)$ é obtida como a distribuição marginal de uma distribuição bivariada para a própria variável aleatória Z e uma variável aleatória auxiliar Y . Essa variável Y é definida no intervalo $(-0.5, a_{\alpha, \beta})$, quando $Z \in (-\infty, 0)$ e em $(a_{\alpha, \beta}, 0.5)$ quando $Z \in (0, \infty)$. A quantidade $a_{\alpha, \beta}$ é dada por,

$$a_{\alpha, \beta} = -\frac{b_{\alpha, \beta}}{\alpha\pi}$$

onde, $b_{\alpha, \beta}(\frac{\beta\pi}{2}) \min(\alpha, 2 - \alpha)$.

A f.d.p. conjunta para as variáveis aleatórias Z e Y é dada por

$$f(z, y|\alpha, \beta) = \frac{\alpha}{|\alpha - 1|} \frac{1}{|z|} \left| \frac{z}{t_{\alpha, \beta}(y)} \right|^\theta \exp \left[- \left| \frac{z}{t_{\alpha, \beta}(y)} \right|^\theta \right] \sigma^{-1} \quad (8)$$

onde $\theta = \frac{\alpha}{\alpha - 1}$,

$$t_{\alpha, \beta}(y) = \frac{\sin(\pi\alpha y + b_{\alpha, \beta})}{\cos(\pi y)} \left\{ \frac{\cos(\pi y)}{\cos[\pi(\alpha - 1)y + b_{\alpha, \beta}]} \right\}^{\frac{1}{\theta}}$$

e $Z = \frac{(X - \mu)}{\alpha}$ para $\alpha \neq 0$.

Da densidade bivariada (8), [Buckle \(1995\)](#) mostra que a distribuição marginal para a variável aleatória Z é uma distribuição estável $S_\alpha(\beta, 0, 1)$. Como mencionado, os custos computacionais para obtenção dos sumários a posteriori de interesse usando os métodos MCMC são altos para essa classe de modelos, o que pode gerar algumas limitações para aplicações práticas. Um problema pode ser a convergência do algoritmo de simulação. Propomos então, o uso de um *software* popular gratuito, o *software OpenBugs* (ver [Spiegelhalter et al. \(2003\)](#)), para obter as distribuições a posteriori de interesse que

requer apenas a definição da função de verossimilhança e as distribuições a priori para os parâmetros do modelo.

Para a análise Bayesiana do modelo, assume-se distribuições uniforme a priori no domínio de cada parâmetro, isto é, $\alpha \sim U(0, 2)$, $\beta \sim U(-1, 1)$, $\sigma \sim U(0, a)$ e $\mu \sim (-b, b)$, onde a e b são hiperparâmetros conhecidos, usualmente com valores muito grandes para ter distribuições a priori aproximadamente não informativas. Outra possibilidade na análise de dados, é assumir uma distribuição gamma a priori para o parâmetro de escala σ e uma distribuição a priori normal para o parâmetro de localização μ , mas observou-se nos exemplos considerados no nesse estudo que assumindo priori uniforme para todos os parâmetros, as inferências posteriores foram melhores, especialmente em termos de convergência dos algoritmos de simulação.

3.5 ANÁLISE BAYESIANA ASSUMINDO MODELOS DE REGRESSÃO PARA OS PARÂMETROS DE LOCAÇÃO E ESCALA DE UMA DISTRIBUIÇÃO ESTÁVEL

Nesta seção, assumiremos que a resposta x_i para $i = 1, \dots, n$, na presença de um vetor de covariáveis $\mathbf{v} = (v_1, v_2, \dots, v_k)$ tem uma distribuição estável tem uma distribuição estável $X_i \sim S_\alpha(\beta, \mu_i, \sigma_i)$, isto é, $Z_i = \frac{X_i - \mu_i}{\sigma_i} \sim S_\alpha(\beta, \mu_i, \sigma_i)$, onde os parâmetros de localização μ e escala σ da distribuição estável são relacionados conjuntamente às variáveis explicativas pelos modelos de regressão dados por,

$$\begin{aligned}\mu_i &= d_0 + d_1 v_{1i} + d_2 v_{2i} + \dots + d_k v_{ki} \\ \sigma_i &= \exp(e_0 + e_1 v_{1i} + e_2 v_{2i} + \dots + e_k v_{ki})\end{aligned}\tag{9}$$

Assumindo a distribuição conjunta a posteriori para α, β, d e e , onde $d = (d_0, d_1, d_2, \dots, d_k)$ e $e = (e_0, e_1, e_2, \dots, e_k)$ são os parâmetros de regressão associados aos modelos de regressão, dado por $\pi_0(\alpha, \beta, d, \sigma)$, [Buckle \(1995\)](#) mostra que a distribuição a posteriori conjunta para os parâmetros α, β, d e e , é dada por,

$$\begin{aligned}\pi(\alpha, \beta, \mathbf{d}, \mathbf{e} | \mathbf{x}) &\propto \int - \left(\frac{\alpha}{|\alpha - 1|} \right)^n \exp \left\{ - \sum_{i=1}^n \left| \frac{z_i}{t_{\alpha, \beta}(y_i)} \right|^\theta \right\} \times \\ &\quad \prod_{i=1}^n \left| \frac{z_i}{t_{\alpha, \beta}(y_i)} \right|^\theta \sigma_i^{-1} \frac{1}{|z_i|} \pi_0(\alpha, \beta, \mathbf{d}, \mathbf{e}) dy\end{aligned}\tag{10}$$

onde $\theta = \frac{\alpha}{\alpha - 1}$, para $i = 1, \dots, n$, $\alpha \in (0, 2]$, $\beta \in [-1, 1]$, $\mathbf{x} = (x_1, x_2, \dots, x_n)$ e $\mathbf{y} = (y_1, y_2, \dots, y_n)$ são respectivamente, os vetores de dados observados e não observados. Note que, a expressão da distribuição bivariada (8) é dada em termos de x_i

$\sigma \in (0, \infty)$; $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ e $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n)$ são respectivamente, o observado e não-observado em vetores de dados. Observe que a distribuição bivariada na expressão (8) é dada em termos de x_i e as variáveis latentes y_i , e não em termos de z_i e y_i (há o Jacobiano σ_i^{-1} multiplicando o lado direito da expressão acima). Perceba também que, quando $\alpha = 2$, temos $\theta = 2$ e $b_{\alpha, \beta} = 0$. Nesse caso temos uma distribuição Gaussiana com média igual a μ e variância igual a $2\sigma^2$.

Para a análise Bayesiana do modelo proposto, assumimos prioris uniformes $U(a, b)$ para os parâmetros $\alpha, \beta, d_0, d_1, d_2, \dots, d_k, e_0, e_1, e_2, \dots, e_k$ onde os hiperparâmetros a e b são assumidos serem conhecidos em cada aplicação seguindo as restrições $\alpha \in (0, 2]$ e $\beta \in [-1, 1]$. Além disso, assumimos independência entre todos os parâmetros.

No algoritmo de simulação para obter amostras de Gibbs para as quantidades aleatórias α, β, d e e , tendo a distribuição a posteriori conjunta (10), também assumimos distribuições uniformes para a variável latente Y definida no intervalo $(-0.5, 0.5)$. Com essa escolha de prioris, usamos o padrão disponível nos pacotes do *software OpenBugs* (Spiegelhalter *et al.* (2003)), que oferece grande simplificação para obter as amostras de Gibbs simuladas para a distribuição a posteriori conjunta de interesse.

Da expressão (10), a distribuição de probabilidade a posteriori conjunta para $\alpha, \beta, \mathbf{d}, \mathbf{e}$ e $\mathbf{y} = (y_1, y_2, \dots, y_n)$ é dada por,

$$\pi(\alpha, \beta, \mathbf{d}, \mathbf{e} | \mathbf{x}) \propto \left(\frac{\alpha}{|\alpha - 1|} \right)^n \exp \left\{ - \sum_{i=1}^n \left| \frac{z_i}{t_{\alpha, \beta}(y_i)} \right|^\theta \right\} \times \\ \prod_{i=1}^n \left| \frac{z_i}{t_{\alpha, \beta}(y_i)} \right| \sigma_i^{-1} \frac{1}{|z_i|} \prod_{i=1}^n h(y_i) \pi_0(\alpha, \beta, \mathbf{d}, \mathbf{e})$$

onde θ e $t_{\alpha, \beta}(\cdot)$ são respectivamente definidos em (8) e (3.4), e a variável Y é definida no intervalo $(-0.5, 0.5)$.

Como utilizamos o *software OpenBugs* para simular amostras da distribuição a posteriori conjunta, não apresentamos todas as distribuições condicionais necessárias para o algoritmo de Gibbs. Esse *software* apenas requer a distribuição dos dados e as distribuições a priori de interesse. Isso traz grande simplificação computacional para determinar os sumários posteriores de interesse como será mostrado nas aplicações a seguir.

4 APLICAÇÕES

4.1 EFEITOS DO TABAGISMO NA SAÚDE

Existem muitos estudos sobre os efeitos do tabagismo na saúde. Nesse sentido, [Tager et al. \(1983\)](#) relatou análises de um estudo em que objetivou avaliar a respiração de crianças na ausência ou presença de pais fumantes, bem como a exposição à fumaça passiva de pelo menos um dos pais ($n = 654$ indivíduos). Esse estudo representa uma das primeiras tentativas de documentar os óbvios sinais de redução da função pulmonar pelo ato de fumar e pela exposição secundária ao fumo. Os dados desta investigação são provenientes de um estudo observacional onde os sujeitos se auto-selecionaram em qual grupo acreditavam pertencer (fumantes ou não fumantes), também informaram sua condição de fumante.

Utilizando-se um espirômetro (aparelho para medir o volume de ar inspirado e expirado pelos pulmões), registrou-se os valores de *Forced Expiratory Volume*(FEV) para cada sujeito, no Brasil, esse teste é chamado de espirometria, nesse exame a pessoa enche os pulmões com o máximo de ar que puder e depois força rapidamente rapidamente em um espirômetro, esse aparelho mede todo o ar que a pessoa consegue soprar e também a velocidade de saída do ar dos pulmões¹.

As variáveis descritas associadas a cada indivíduo são: idade (em anos), FEV (em litros), altura (em centímetros), gênero (feminino (0), masculino (1)), fumante (não fumante (0), fumante (1)).

Na Figura 4 é apresentado o histograma normal da variável de resposta FEV, nas escalas originais e logarítmicas. A partir desses gráficos, observa-se que a normalidade é otimizada na escala logarítmica. Não considerando a transformação logarítmica, a distribuição é aproximadamente enviesada (veja Figura 4). Desse modo, é possível assumir uma distribuição estável para obter melhor ajuste para os dados na escala original.

¹ Detalhes sobre o exame podem ser encontrados no site da Sociedade Paulista de Pneumologia e Tisiologia: <https://sppt.org.br/serie-1-metodos-diagnosticos-em-pneumologia-espirometria/f>

Figura 4 – Histogramas: original e escala transformada.

Uma observação importante é que neste estudo será considerado o ponto (.) como separador decimal.

Em uma análise inicial, assumimos um modelo de regressão linear com erros normais (ver, por exemplo, [Draper e Smith \(1981\)](#) ou [Seber e Lee \(2003\)](#)) sob uma abordagem clássica (Estimadores de Mínimos Quadrados) relacionando a resposta FEV em escala logarítmica com as covariáveis idade, altura, gênero e fumante. O ajuste do modelo de regressão é dado por,

$$\begin{aligned} \log(FEV) = & -1.9414 + 0.0168 \times (altura) + 0.0236 \times (idade) + \\ & + 0.0288 \times (genero) - 0.0471 \times (fumante) \end{aligned} \quad (11)$$

Na Tabela 1, é apresentado os sumários dos resultados obtidos com uma análise com os estimadores de Mínimos Quadrados-EMQ ($S = 0.1459$ (erro padrão residual); $R - sq = 80.96\%$ (múltiplo R-quadrado (RQ))). Dos resultados apresentados na Tabela 1, é observado que assumindo um modelo de regressão linear múltipla com erros normais (resposta em escala logarítmica), todas as covariáveis (altura, idade, gênero e fumante) mostram efeitos significativos na resposta (FEV), uma vez que os valores- p são menores do que 0,05 (nível de significância assumido).

Tabela 1 – Modelo de regressão linear normal (dados FEV)

Termo	<i>EMQ</i>	<i>ErroPd</i>	<i>T – value</i>	<i>P – value</i>
Constante	-1.9414	0.0790	-24.58	<0.001
Altura	0.0168	0.0007	25.34	<0.001
Idade	0.0236	0.0034	7.04	<0.001
Gênero	0.0288	0.0118	2.45	0.015
Fumante	-0.0471	0.0210	-2.25	0.025

Segundo os dados, nota-se que: pessoas mais velhas tem maior FEV (em litros); homens também apresento maior FEV quando comparado com mulheres; fumantes tem

FEV menor quando comparado aos não fumantes; e, maior altura implica em maior FEV. A figura 5 apresenta a análise residual a partir da qual, observa-se que os pressupostos necessários (normalidade e variância constante dos resíduos) assumindo a escala logarítmica para as respostas (uma distribuição log-normal para FEV) são razoavelmente aceitos (não totalmente). Isso nos motiva a buscar melhores modelos estatísticos.

Figura 5 – Gráfico dos resíduos (modelo de regressão linear normal - dados FEV)

4.1.1 ABORDAGEM BAYESIANA USANDO UMA DISTRIBUIÇÃO ESTÁVEL NÃO CONSIDERANDO A PRESENÇA DE COVARIÁVEIS

Agora, assumiremos uma distribuição estável para o banco de dados FEV, inicialmente não considerando a presença de covariáveis. Assumimos as seguintes distribuições a priori para os parâmetros α, β, μ e σ : $\alpha \sim U(1, 2)$, $\beta \sim U(-1, 1)$, $\mu \sim U(-2, 2)$ e $\sigma \sim U(0, 2)$, onde $U(a, b)$ denota uma distribuição uniforme no intervalo (a, b) .

Na elicitação das prioris para os parâmetros do modelo, algumas informações sobre os parâmetros da distribuição estável puderam ser obtidas a partir dos histogramas apresentados na Figura 4, com simetria próxima levando à distribuições a priori que deveriam ser concentradas em $\alpha = 2$. Dessa forma, considerou-se uma priori uniforme para o parâmetro α concentrado no intervalo $(1, 2)$. A Tabela 2 apresenta um sumário das posteriores de interesse ($burn-in = 100000$ e outras 400000 amostras de Gibbs a cada

centésima amostra) obtido usando o *software* OpenBugs (SPIEGELHALTER *et al.*, 2003). A convergência do algoritmo MCMC foi verificada a partir dos *traceplots* padrão das amostras simuladas. O código usado no OpenBugs é apresentado no Apêndice A no final desta dissertação..

Tabela 2 – Análise Bayesiana - Distribuição α –Estável sem presença de covariáveis.

	<i>Média</i>	<i>DesP</i>	$95\%C_rI$
α	1.912	0.0468	(1.83, 1.99)
β	-0.076	0.4961	(-0.90, 0.92)
μ	1.997	0.0025	(1.99, 2.00)
σ	0.739	0.0261	(0.69, 0.79)

Dos resultados apresentados na Tabela 2, é possível notar que a média posteriori para o parâmetro α possui um estimado de Monte Carlo baseado nas amostras simuladas de Gibbs próximas ao valor 2 (estimador de Monte Carlo da média a posteriori é igual a 1,9252) indicando que o ajuste da distribuição estável é próximo a distribuição normal usual. A Figura 6 apresenta os *traceplots* das amostras simuladas de Gibbs para cada parâmetro.

Figura 6 – Traceplots - distribuição estável - sem covariáveis

4.1.2 ABORDAGEM BAYESIANA USANDO UMA DISTRIBUIÇÃO ESTÁVEL NA PRESENÇA DE COVARIÁVEIS

Para uma segunda análise estatística agora considerando a presença de um vetor de covariáveis, assume-se que o modelo de regressão para os parâmetros de locação e escala da distribuição estável definido em (9) sob uma abordagem Bayesiana (uso de métodos MCMC) relacionando a resposta FEV com as covariáveis altura, idade, gênero e fumante, isto é,

$$\begin{aligned}\mu_i &= d_0 + d_1 * altura_i + d_2 * idade_i + d_4 * genero_i + d_4 * fumante_i \\ \sigma_i &= exp(e_0 + e_1 * altura_i + e_2 * idade_i + e_2 * genero_i + e_4 * fumante_i)\end{aligned}\tag{12}$$

Para a análise Bayesiana, assume-se uma distribuição uniforme $U(1.5, 2)$ para α , uma uniforme $U(-0.5, 0.5)$ para β e uma uniforme $U(-1, 1)$ para os parâmetros de regressão $d_0, d_1, d_2, d_3, d_4, e_0, e_1, e_2, e_3, e_4$. A Tabela 3 apresenta os sumários de interesse a posteriori (usando o *software OpenBugs* - códigos apresentados no Apêndice A).

Tabela 3 – Análise Bayesiana - distribuição estável

	<i>Média</i>	<i>DesvP</i>	<i>95%CrI</i>
α	-1.975	0.0215	(1.923, 1.999)
β	-0.014	0.2831	(-0.478, 0.471)
d_0	2.015	0.0158	(2.000, 2.061)
d_1	-0.012	0.3110	(-0.012, -0.011)
d_2	0.239	0.0049	(0.229, 0.248)
d_3	0.236	0.0426	(0.149, 0.318)
d_4	0.050	0.1060	(-0.162, 0.243)
e_0	-0.898	0.1053	(-0.998, -0.613)
e_1	-0.007	0.0013	(-0.010, -0.004)
e_2	0.085	0.0156	(0.055, 0.117)
e_3	0.353	0.0582	(0.243, 0.469)
e_4	0.168	0.1060	(-0.028, 0.394)

Dos resultados apresentados na Tabela 3, observa-se que assumindo a distribuição estável na presença de covariáveis, a covariável fumante não mostra efeito significativo nos parâmetros de locação e escala da distribuição estável assumida para a resposta FEV, uma vez que zero não está dentro do intervalo de credibilidade 95% para os parâmetros de regressão correspondentes d_4 e e_4 . A figura 7 apresenta os *traceplots* para as amostras de Gibbs simuladas para cada parâmetro, onde pode-se observar uma boa convergência do algoritmo de simulação MCMC.

Figura 7 – Traceplots (distribuição estável - presença de covariáveis)

A Figura 8 apresenta os gráficos de dispersão para a resposta FEV versus cada covariável, onde pode-se notar uma pequena dependência linear entre a resposta FEV e fumante, que também está intimamente relacionado à covariável gênero que mostra efeitos significativos na resposta (homens são mais fumantes quando comparados com as mulheres).

Figura 8 – Gráfico de dispersão da resposta FEV versus cada covariável

4.2 ASSOCIAÇÃO ENTRE NÍVEIS DE HEMOGLOBINA (HbA1c) E ALGUMAS COVARIÁVEIS PARA DIABETES MELLITUS TIPO 2 (T2DM)

Apresentamos agora outro exemplo ilustrativo, a associação entre algumas covariáveis para pacientes com diabetes mellitus tipo 2 (em inglês, *type 2 diabetes mellitus* - *T2DM*, sigla que será usada) com níveis de hemoglobina glicada (em inglês, *glycated hemoglobin* - *HbA1c*, sigla que será usada no estudo), que é considerado neste exemplo (conjunto de dados introduzido por [Shu, Chan e L. \(2017\)](#)).

O tamanho da amostra consiste em $n = 154$ pacientes. Algumas covariáveis são relacionadas à resposta hemoglobina glicada (HbA1c) em cada paciente. Neste estudo, considerou-se as seguintes covariáveis: idade, educação (em anos), tempo do diagnóstico (em meses), cumprimento da medicação, escore de barreira (*score barrier*), pontuação de habilidade de conhecimento dietético, índice de massa corporal - BMI (em inglês, *body mass score*), gênero e índice total de alimentação saudável - HEI (*healthy eating index*).

A Figura 9 apresenta o gráfico da probabilidade normal para a média de resposta HbA1c, de onde é possível notar que a distribuição normal não é apropriada para os dados.

Figura 9 – Gráfico da probabilidade normal para a resposta HbA1c

A Figura 10 apresenta os gráficos de dispersão para a resposta HbA1c versus cada covariável, de onde é possível notar a relação linear entre cada covariável com a resposta.

Figura 10 – Gráfico da probabilidade normal para a resposta HbA1c

A Tabela 4 apresenta os sumários a posteriori para a análise de estimadores de mínimos quadrados (EMQ). O modelo de regressão ajustado (uso do *software R* (2019)), é dado por,

$$\begin{aligned}
 HbA1c = & 12.37 + 0.0188 \times idade + 0.0010 \times educação + 0.0022 \times tempo.diagnóstico - \\
 & - 0.0263 \times cumprimento.medicação - 0.0100 \times escore.barreira - \\
 & - 0.1267 \times conhecimento.dieta + 0.0581 \times BMI + 0.057 \times genero - 0.0221 \times HEI
 \end{aligned}
 \tag{13}$$

Dos resultados da Tabela 4, observou-se que assumindo um modelo de regressão linear múltiplo com erros normais, as covariáveis cumprimento da medicação, conhecimento da dietético e BMI, apresentam efeitos significativos na resposta HbA1c, uma vez que os p-valores são menores que 0.05 (nível de significância assumido). A Figura 11 apresenta os resultados da análise residual, de onde observou-se que as suposições necessárias (normalidade e variância constante dos resíduos) assumindo a escala logarítmica para as resposta são razoavelmente aceitáveis (não totalmente).

Tabela 4 – Modelo de regressão linear normal (dados HbA1c)

Termo	<i>EMQ</i>	<i>ErroPd</i>	<i>T – value</i>	<i>P – value</i>
Constante	12.37	2.39	5.18	<0.001
Idade	0.0188	0.0207	0.91	0.37
Educação	0.0010	0.0459	0.02	0.98
T. diagnóstico	0.0022	0.0014	1.52	0.13
Cump. Med.	-0.0263	0.0127	-2.07	0.040
Escore barreira	-0.0100	0.0166	-0.60	0.549
C. de dieta	-0.1267	0.0446	-2.84	0.005
BMI	0.0581	0.0289	2.01	0.046
Gênero	0.0570	0.3870	0.15	0.88
Total HEI	-0.0221	0.0179	-1.23	0.22

Figura 11 – Gráfico de resíduos (regressão linear normal para resposta HbA1c)

4.2.1 ABORDAGEM BAYESIANA USANDO UMA DISTRIBUIÇÃO ESTÁVEL SEM CONSIDERAR A PRESENÇA DE COVARIÁVEIS

Assumiremos agora, uma distribuição estável para o conjunto de dados, não considerando a presença de covariáveis, assumindo as mesmas distribuições a priori para os parâmetros α, β, μ e σ assumidas na seção 4.1.1. A Tabela 5 apresenta os sumários a posteriori de interesse (*burn-in* = 300.000 e outras 511.000 amostras de Gibbs a cada 100 amostra) obtidos usando o *software* OpenBugs. A convergência do algoritmo MCMC foi

verificada a partir de *traceplots* padrão das amostras simuladas. A Figura 12 apresenta os *traceplots* das amostras de Gibbs simuladas para cada parâmetro.

Figura 12 – Gráfico de resíduos (regressão linear normal para resposta HbA1c)

Tabela 5 – Análise Bayesiana – distribuição estável sem considerar a presença de covariáveis (dados HbA1c)

	<i>Média</i>	<i>DesvP</i>	<i>95%CrI</i>
α	1.650	0.1424	(1.389, 1.909)
β	-0.225	0.7033	(-0.988, 0.975)
μ	8.791	0.1734	(8.484, 9.153)
σ	1.393	0.103	(1.208, 1.616)

4.2.2 UMA ABORDAGEM BAYESIANA ASSUMINDO UMA DISTRIBUIÇÃO ESTÁVEL NA PRESENÇA DE COVARIÁVEIS

Considerando a presença de um vetor de covariáveis (idade, educação, cumprimento da medicação, escore de barreira, conhecimento de dieta, gênero e pontuação HEI total), assume-se os modelos de regressão para os parâmetros de locação e escala da distribuição estável, definidos em 9 sob uma abordagem Bayesiana (usando métodos MCMC), dados por,

$$\begin{aligned}
 \mu_i &= d_0 + d_1 idade_i + d_2 educação_i + d_3 tempo.diagnóstico_i + d_4 cumprimento.medicação_i + \\
 &\quad + d_5 escore.barreira_i + d_6 conhecimento.dieta_i + d_7 BMI_i + d_8 HEI.total_i + d_9 genero_i \\
 \sigma_i &= \exp(e_0 + e_1 idade_i + e_2 educação_i + e_3 tempo.diagnóstico_i + e_4 cumprimento.medicação_i + \\
 &\quad + e_5 escore.barreira_i + e_6 conhecimento.dieta_i + e_7 BMI_i + e_8 HEI.total_i + e_9 genero_i)
 \end{aligned}
 \tag{14}$$

Para uma análise Bayesiana, assume-se uma priori uniforme $U(0.5, 2)$ para α , uma uniforme $U(-1, 1)$ para β , uma uniforme $U(-10, 10)$ para o parâmetro de regressão d_0 , uma priori uniforme $U(-1, 1)$ para os parâmetros d_9 e e_0 , e distribuições a priori uniforme $U(-0.5, 0.5)$ para os outros parâmetros de regressão. Na Tabela (6) apresentam-se os sumários a posteriori de interesse (usando o *software* OpenBugs).

Tabela 6 – Análise Bayesiana - distribuição estável (dados HbA1c)

	<i>Média</i>	<i>DesvP</i>	<i>95%C_rI</i>
α	1.8200	0.1479	(1.4700, 1.9940)
β	0.0197	0.6218	(-0.9489, 0.9678)
d_0	8.9200	0.2138	(8.5060, 9.3380)
d_1	0.0267	0.0186	(-0.0099, -0.0629)
d_2	-0.4750	0.0414	(-0.0804, 0.0879)
d_3	0.0022	0.0014	(0.0004, 0.0059)
d_4	-0.0165	0.130	(-0.0405, 0.0122)
d_5	-0.0428	0.0228	(-0.0897, 0.0014)
d_6	-0.1468	0.0404	(-0.2229, -0.0646)
d_7	0.0734	0.0310	(0.0100, 0.1313)
d_8	-0.0251	0.0180	(-0.0604, 0.0093)
d_9	0.2887	0.3430	(-0.4444, 0.9100)
e_0	0.1920	0.0996	(0.0026, 0.3804)
e_1	0.0157	0.0080	(0.3140, 0.0316)
e_2	-0.0214	0.0194	(-0.0592, 0.0163)
e_3	-0.6850	0.0010	(-0.0028, 0.0012)
e_4	0.0014	0.0058	(-0.0113, 0.0118)
e_5	-0.0117	0.0067	(-0.0246, 0.0014)
e_6	-0.0385	0.0206	(-0.0814, 0.0021)
e_7	0.0156	0.0127	(-0.0106, 0.0394)
e_8	-0.0079	0.0082	(-0.0234, 0.0099)
e_9	0.2315	0.1399	(-0.0732, 0.4706)

Dos resultados apresentados na Tabela 6, observa-se que assumindo a distribuição estável na presença de covariáveis, a covariável idade, tempo de diagnóstico, conhecimento de dieta e BMI, mostram efeito significativo no parâmetro de locação da distribuição estável assumido para a resposta HbA1c, uma vez que o zero não está nos intervalos 95% de credibilidade para os parâmetros de regressão correspondentes: d_1 , d_3 , d_6 e d_7 . Além disso, a covariável idade, mostra efeito significativo no parâmetro escala, uma vez que o zero não está no intervalo 95% de credibilidade para o parâmetro de regressão correspondente e_1 .

A Figura 13 apresenta os *traceplots* das amostras de Gibbs simuladas para cada parâmetro, de onde observa-se boa convergência da simulação do algoritmo MCMC.

Figura 13 – Gráfico de resíduos (regressão linear normal para resposta HbA1c)

4.3 APLICAÇÃO A DADOS DE CÂNCER DE MAMA

Apresentamos agora uma aplicação com dados relacionados ao câncer de mama no estado de São Paulo (ver Seção 1.2) mensalmente no períodos de junho/2009 a junho/2013 (dados 1) e de outubro/2013 a junho/2015 (dados 2). Os dados foram analisados da seguinte maneira: primeiramente no período de junho/2009 a junho/2013, consideramos duas opções para a resposta: razão entre o número de exames para cada 1000 mulheres (denotado por y_1) e razão nódulos a cada 100 exames (denotado por y_2). Na segunda parte, utilizamos as mesmas respostas para o período de outubro/2013 a junho/2015.

O tamanho da amostra para os dados 1 consiste em $n = 784$ e para os dados 2 é $n = 336$. Existem diversas covariáveis que são relacionadas às respostas y_1 e y_2 , neste estudo considerou-se as seguintes: grau de urbanização (x_1), PIB (Produto Interno Bruto) (x_2), renda (x_3) e número de médicos pela população feminina (x_4).

Os dados também estão classificados por regiões administrativas e região metropolitana de São Paulo: Araçatuba (ARCTB), Barretos (BRRTS), Bauru (BAURU), Campinas (CMPNS), Central (CNTRL), Franca(FRNCA), Itapeva (ITPVA), Marília (MRLIA),

Presidente Prudente (PSTPD), Registro (RGSTR), Ribeirão Preto (RBPRT), Santos (SNTOS), São José do Rio Preto (SJDRP), São José dos Campos (SJCPS), Sorocaba (SRCBA) e Região Metropolitana de São Paulo (RMSPL).

4.3.1 CÂNCER DE MAMA - JUNHO/2009 A JUNHO/2013 - PARA y_1

Inicialmente realizou-se a análise descritiva dos dados para a resposta y_1 : razão entre o número de exames e a população feminina (a cada 1000 mulheres).

A Tabela 7 apresenta os dados com o fator região administrativa, dessa tabela observa-se a maior média para a região de Barretos e a menor média para a região de Bauru. A Tabela 8 apresenta os dados com fator ano, sendo 2012 o ano com maior média e 2009 o ano para a menor média. Na tabela 9 apresenta-se os dados com fator o mês, sendo o mês 5 com a maior média e o mês 2 com a menor média.

Tabela 7 – Análise descritiva - 06/2009 a 06/2013 - para y_1 fator região, ($n = 49$ para cada região)

Região	Média	<i>DesvP</i>	<i>Variância</i>	<i>Mín.</i>	<i>Máx.</i>
ARCTB	67.74	31.84	1013.81	1.81	124.47
BAURU	27.298	6.042	36.510	10.710	42.250
BRRTS	71.14	16.61	275.78	36.05	105.31
CMPNS	35.31	14.42	207.99	1.27	57.63
CNTRL	43.86	10.07	101.43	19.42	69.38
FRANCA	40.21	14.89	221.68	1.19	66.57
ITPVA	66.42	29.24	855.22	18.25	113.37
MRLIA	52.62	15.75	248.09	4.69	74.33
PSTPD	56.41	15.24	232.32	10.61	82.67
RBPRT	27.03	11.24	126.31	0.32	42.68
RGSTR	51.69	20.38	415.48	1.34	96.42
RMSPL	32.86	12.05	145.24	0.75	49.92
SJCPS	46.58	14.91	222.45	2.92	68.21
SJDRP	69.66	14.28	204.05	21.33	93.28
SNTOS	35.17	10.74	115.40	0.56	51.70
SRCBA	34.38	9.19	84.40	5.17	54.43

Tabela 8 – Análise descritiva - 06/2009 a 06/2013 - para y_1 fator ano

Ano	N	<i>Média</i>	<i>DesvP</i>	<i>Variância</i>	<i>Mín.</i>	<i>Máx.</i>
2009	112	27.84	19.45	378.27	0.32	85.18
2010	192	38.80	17.55	307.89	0.69	110.01
2011	192	54.85	20.92	437.84	18.57	124.47
2012	192	57.04	20.41	416.60	23.42	115.88
2013	96	53.23	19.50	380.10	5.71	101.73

Tabela 9 – Análise descritiva - 06/2009 a 06/2013 - para y_1 fator mês

Mês	N	Média	DesvP	Variância	Mín.	Máx.
1	64	44.21	19.05	363.09	9.78	115.88
2	64	43.77	18.80	353.43	1.34	82.84
3	64	46.37	21.14	447.11	1.42	103.35
4	64	51.39	21.67	469.76	2.09	101.73
5	64	55.02	22.56	508.96	19.08	113.37
6	80	46.09	26.54	704.47	0.32	105.31
7	64	44.80	27.19	739.56	1.81	124.47
8	64	46.71	24.28	589.70	1.19	96.57
9	64	48.16	21.56	465.00	3.36	97.70
10	64	49.24	20.49	419.64	19.48	106.91
11	64	46.75	18.87	356.18	13.13	90.63
12	64	46.62	21.50	462.37	0.69	110.01

Nas Figuras 14, 15 e 16 são apresentados os *boxplots* para região, ano e mês respectivamente, em todos pode-se notar a presença de diversos *outliers*.

Figura 14 – Boxplot de y_1 para as regiões

Figura 15 – Boxplot de y_1 para os anos Figura 16 – Boxplot de y_1 para os meses

A Tabela 10 apresenta o resultado da Análise de Variância (ANOVA) para o fator Região, por ela nota-se que as médias são estatisticamente diferentes para as regiões (com valor- $p < 0.05$). Note que, a região BAURU (27.298) e RBPRT (27.03) apresentam as menores médias, enquanto as regiões ARCTB (67.74), BRRTS (71.14) e SJDRP (69.66) apresentam as maiores médias. Nos resultados do sumário do modelo, o fator Região explica 44.38% da variação na resposta y_1 . Indica também que o desvio padrão (DesvP) entre os pontos de dados e os valores ajustados é de aproximadamente 16.7620. A Figura 17 apresenta o gráfico de intervalos de confiança 95% C_rI para a média, da resposta y_1 (razão entre o número de exames e a população feminina (a cada 1000 mulheres)) versus cada região. A Figura 18 apresenta a análise residual, onde nota-se que há observações que não se ajustam bem com a distribuição normal.

Tabela 10 – ANOVA: fator Região – 06/2009 a 06/2013 - para y_1

Análise de Variância					
Fonte	GL	$SQ(Aj\cdot)$	$QM(Aj\cdot)$	$Valor F$	$Valor - p$
Amostra	15	172584	11505.6	40.85	<0.001
Erro	768	216296	281.6		
Total	783	388880			
Sumário do Modelo					
S	R^2	$R^2(Aj\cdot)$	$R^2(pred)$		
16.7820	44.38%	43.29%	42.04%		
Médias					
Região	$Média$	$DesvP$	$95\%C_{-}\{r\}I$		
ARCTB	67.74	31.84	(63.03, 72.44)		
BAURU	27.298	6.042	(22.592, 32.004)		
BRRTS	71.14	16.61	(66.43, 75.85)		
CMPNS	35.31	14.42	(30.61, 40.02)		
CNTRL	43.86	10.07	(39.15, 48.57)		
FRNCA	40.21	14.89	(39.15, 48.57)		
ITPVA	66.42	29.24	(61.72, 71.13)		
MRLIA	52.62	15.75	(47.91, 71.13)		
PSTPD	56.41	15.24	(51.70, 61.11)		
RBPRT	27.03	11.24	(22.32, 31.74)		
RGSTR	51.69	20.38	(46.99, 56.40)		
RMSPLS	32.86	12.05	(28.16, 37.57)		
SJCPS	46.58	14.91	(41.87, 51.28)		
SJDRP	69.66	14.28	(64.96, 74.37)		
SNTOS	35.17	10.74	(30.46, 39.88)		
SRCBA	34.38	9.19	(29.67, 39.08)

Figura 17 – Gráfico intervalo 95% $C_r I$ para a média: y_1 vs Região

Figura 18 – Gráfico de resíduos fator região para y_1

A Tabela 11 apresenta o resultado da ANOVA para o fator Ano, por ela nota-se que as médias são estatisticamente diferentes para os anos (com valor- $p < 0.05$). Note que, o ano 2009 (27.84) apresenta a menor média, enquanto o ano 2012 (57.04) apresenta a maior média. Nos resultados do sumário do modelo, o fator Ano explica 22.83% da variação na resposta y_1 . Indica também que o desvio padrão entre os pontos de dados e os valores ajustados é de aproximadamente 19.6275. A Figura 19 apresenta o gráfico de intervalos de confiança 95% $C_r I$ para a média, da resposta y_1 (razão entre o número de exames e a população feminina (a cada 1000 mulheres)) versus cada ano. A Figura 20 mostra a análise residual, de onde é possível notar que há observações que não se ajustam bem com a distribuição normal, ou seja, indicando a presença de *outliers*.

Tabela 11 – ANOVA: fator Ano – 06/2009 a 06/2013 - para y_1

Análise de Variância					
Fonte	GL	$SQ(Aj\cdot)$	$QM(Aj\cdot)$	$ValorF$	$Valor - p$
Amostra	4	88778	22194.5	57.61	<0.001
Erro	779	300102	385.2		
Total	783	388880			
Sumário do Modelo					
S	R^2	$R^2(Aj\cdot)$	$R^2(pred)$		
19.6275	22.83%	22.43%	21.84%		
Médias					
Ano	N	$Média$	$DesvP$	$95\%C_{\{r\}}I$	
2009	112	27.84	19.45	(24.20, 31.48)	
2010	192	38.80	17.55	(36.02, 41.58)	
2011	192	54.85	20.92	(52.07, 57.63)	
2012	192	57.04	20.41	(54.26, 59.82)	
2013	96	53.23	19.50	(49.29, 57.16)

Figura 19 – Gráfico intervalo 95% C_rI para a média: y_1 vs Ano

Figura 20 – Gráfico de resíduos fator Ano para y_1

Na Tabela 12 apresenta-se o resultado da ANOVA para o fator Mês, por ela nota-se que as médias não são estatisticamente diferentes para os meses (com valor- $p > 0.05$). Note que, o mês 2 (43.77) apresenta a menor média, enquanto o mês 5 (55.02) apresenta a maior média. Nos resultados do sumário do modelo, o fator Mês explica 1.86% da variação na resposta y_1 . Mostra também que o desvio padrão entre os pontos de dados e os valores ajustados é de aproximadamente 22.2348, o que indica que o fator Mês não afeta a resposta de forma sistemática. A Figura 21 apresenta o gráfico de intervalos de confiança 95% $C_r I$ para a média, da resposta y_1 (razão entre o número de exames e a população feminina (a cada 1000 mulheres)) versus cada mês. A Figura 22 mostra a análise residual, de onde é possível notar que há observações que não se ajustam bem com a distribuição normal, ou seja, indicando a presença de *outliers*.

Tabela 12 – ANOVA: fator Mês – 06/2009 a 06/2013 - para y_1

Análise de Variância					
Fonte	<i>GL</i>	<i>SQ(Aj·)</i>	<i>QM(Aj·)</i>	<i>ValorF</i>	<i>Valor – p</i>
Amostra	11	7214	655.8	1.33	0.205
Erro	772	381666	494.4		
Total	783	388880			
Sumário do Modelo					
S	<i>R</i> ²	<i>R</i> ² (<i>Aj·</i>)	<i>R</i> ² (<i>pred</i>)		
22.2348	1.86%	0.46%	0.0%		
Médias					
Mês	<i>N</i>	<i>Média</i>	<i>DesvP</i>	<i>95%C_{\{r\}}I</i>	
1	64	44.21	19.05	(38.75, 49.66)	
2	64	43.77	18.80	(38.31, 49.22)	
3	64	46.37	21.14	(40.91, 51.82)	
4	64	51.39	21.67	(45.93, 56.85)	
5	64	55.02	22.56	(49.56, 60.47)	
6	80	46.09	26.54	(41.21, 50.97)	
7	64	44.80	27.19	(39.35, 50.26)	
8	64	46.71	24.28	(41.26, 52.17)	
9	64	48.16	21.56	(42.70, 53.61)	
10	64	49.24	20.49	(43.78, 54.69)	
11	64	46.75	18.87	(41.30, 52.21)	
12	64	46.62	21.50	(41.16, 52.07)

Figura 21 – Gráfico intervalo 95% C_rI para a média: y_1 vs Mês

Figura 22 – Gráfico de resíduos fator Mês para y_1

REGRESSÃO MÚLTIPLA CLÁSSICA (ERROS NORMAIS)

A Figura 23 apresenta os gráficos de dispersão da resposta y_1 para cada uma das covariáveis consideradas para o estudo: x_1 (grau de urbanização), x_2 (PIB), x_3 (renda), x_4 (número de médicos para população feminina), Ano e Mês, de onde pode-se observar como se comportam as relações de cada covariável com a resposta y_1 .

Figura 23 – Gráfico de dispersão para cada covariável vs y_1

Agora, a Tabela 13 mostra os sumários inferenciais para a análise de mínimos quadrados (EMQ). O modelo de regressão ajustado por mínimos quadrados é dado por,

$$y_1 = -24109 + 12.049 \times Ano + 1.151 \times Mes - 77.8 \times x_1 - 13.93 \times x_2 + 21.4 \times x_3 - 0.502 \times x_4 \quad (15)$$

A partir dos resultados da Tabela 13 observa-se que assumindo um modelo de regressão linear múltiplo com erros normais, as covariáveis Ano, Mês, x_1 e x_2 apresentam efeitos significativos na resposta y_1 , tendo em vista que o valor-p de cada covariável é menor que 0.05 (nível de significância assumido). Pela Figura 24 obtêm-se os resultados da análise residual, de onde observa-se que as suposições necessárias (normalidade e variância constante dos resíduos) são razoavelmente verificadas.

Tabela 13 – Modelo de regressão linear normal (y_1 - 06/2009 a 06/2013)

Termo	<i>EMQ</i>	<i>ErroPd</i>	<i>T - Value</i>	<i>P - Value</i>
Constante	-24109	1573	-15.33	0.000
Ano	12.049	0.786	15.33	0.000
Mês	1.151	0.198	5.82	0.000
x_1	-77.8	31.5	-2.47	0.014
x_2	-13.93	2.10	-6.63	0.000
x_3	21.4	15.1	1.42	0.156
x_4	-0.502	0.861	-0.58	0.560

Figura 24 – Gráfico de resíduos (regressão linear normal para resposta y_1)

ABORDAGEM BAYESIANA USANDO UMA DISTRIBUIÇÃO ESTÁVEL

Assume-se agora, uma distribuição estável para o conjunto de dados, considerando a presença de um vetor de covariáveis (ano, mês, x_1 , x_2 , x_3 e x_4), com o modelo de regressão para o parâmetro de locação μ da distribuição estável, definido em (9) sob uma abordagem Bayesiana (usando métodos MCMC), dado por,

$$\mu_i = d_0 + d_1 * ano_i + d_2 * mês_i + d_3 * x_{1i} + d_4 * x_{2i} + d_5 * x_{3i} + d_6 * x_{4i} \quad (16)$$

Para esta análise Bayesiana, assume-se uma priori uniforme $U(0, 2)$ para α , uma uniforme $U(-1, 0)$ para β , uma uniforme $U(0, 100)$ para δ , uma normal $N(-24000, 0.001)$ para o parâmetro de regressão d_0 , uma normal $N(12, 0.1)$ para o parâmetro de regressão d_1 , uma normal $N(1, 0.1)$ para o parâmetro de regressão d_2 , uma normal $N(-78, 0.1)$ para o parâmetro de regressão d_3 , uma normal $N(-14, 0.1)$ para o parâmetro de regressão d_4 , uma normal $N(21, 0.1)$ para o parâmetro de regressão d_5 e uma normal $N(-0.5, 0.1)$ para o parâmetro de regressão d_6 . Na Tabela 14, são apresentados os sumários a posteriori de interesse (usando o *software OpenBugs*).

Tabela 14 – Análise Bayesiana - distribuição estável para y_1

	<i>Média</i>	<i>DP</i>	<i>95%$C_r I$</i>
α	1.165	0.002	(1.16, 1.169)
β	-0.288	0.005	(-0.294, -0.275)
d_0	-23960.0	18.35	(-24020.0, -23930.0)
d_1	12.01	0.010	(12.0, 12.04)
d_2	-0.108	0.002	(-0.113, -0.103)
d_3	-74.06	0.056	(-74.23, -73.95)
d_4	-8.585	0.006	(-8.598, -8.568)
d_5	17.77	0.012	(17.74, 17.81)
d_6	-3.099	0.003	(-3.105, -3.093)
δ	92.52	0.070	(92.38, 92.72)

Dos resultados apresentados na Tabela 14, é possível observar que assumindo a distribuição estável para os dados, todas covariáveis mostram efeito significativos no parâmetro de locação da distribuição estável assumido para a resposta y_1 (razão entre o número de exames para cada 1000 mulheres), uma vez que os intervalos de credibilidade 95% para os parâmetros de regressão (d_1 , d_2 , d_3 , d_4 , d_5 e d_6) não contém o valor zero.

4.3.2 CÂNCER DE MAMA - JUNHO/2009 A JUNHO/2013 - PARA y_2

Apresentamos agora, a análise descritiva dos dados para a resposta y_2 : razão nódulos a cada 100 exames.

A Tabela 15 apresenta os dados com o fator região administrativa, dessa tabela observa-se a maior média para a região de BAURU e a menor média para a região de FRNCA. A Tabela 16 apresenta os dados com fator ano, sendo 2009 o ano com maior média e 2011 o ano para a menor média. Na tabela 17 apresentam-se os dados com fator mês, sendo o mês 8 com a maior média e o mês 12 com a menor média.

Tabela 15 – Análise descritiva - 06/2009 a 06/2013 - para y_2 fator região, ($n = 49$ para cada região)

Região	Média	<i>DesvP</i>	<i>Variância</i>	<i>Mín.</i>	<i>Máx.</i>
ARCTB	8.615	3.340	11.158	5.310	25.760
BAURU	12.551	1.847	3.410	8.070	18.050
BRRTS	12.112	2.435	5.929	6.530	18.660
CMPNS	8.436	0.850	0.723	6.420	10.630
CNTRL	8.623	3.116	9.711	4.530	15.010
FRNCA	6.950	3.959	15.673	3.770	30.950
ITPVA	7.218	2.471	6.105	1.660	15.410
MRLIA	8.298	2.858	8.170	4.830	22.420
PSTPD	7.882	2.293	5.258	0.680	13.120
RBPRT	8.652	2.772	7.686	6.050	22.730
RGSTR	11.507	3.603	12.984	3.570	20.980
RMSPL	9.574	0.991	0.981	7.320	12.530
SJCPS	11.264	2.222	4.936	7.370	17.180
SJDRP	10.867	1.100	1.210	8.270	13.300
SNTOS	10.133	1.908	3.641	7.350	18.750
SRCBA	10.766	1.654	2.735	5.530	13.930

Tabela 16 – Análise descritiva - 06/2009 a 06/2013 - para y_2 fator ano

Ano	N	<i>Média</i>	<i>DesvP</i>	<i>Variância</i>	<i>Mín.</i>	<i>Máx.</i>
2009	112	10.391	4.096	16.779	0.680	30.950
2010	192	9.662	2.771	7.680	3.570	22.730
2011	192	9.177	2.590	6.708	3.370	17.170
2012	192	9.429	2.997	8.980	4.400	17.180
2013	96	9.664	2.598	6.750	5.220	18.050

Tabela 17 – Análise descritiva - 06/2009 a 06/2013 - para y_2 fator mês

Mês	N	Média	DesvP	Variância	Mín.	Máx.
1	64	9.436	2.626	6.898	3.370	17.180
2	64	9.380	2.524	6.368	5.130	15.630
3	64	9.549	2.924	8.552	4.790	17.170
4	64	9.458	2.734	7.474	3.570	15.680
5	64	9.870	2.902	8.422	4.920	18.660
6	80	9.813	3.377	11.405	0.680	22.420
7	64	10.106	3.767	14.189	1.660	25.760
8	64	10.169	4.019	16.150	3.770	30.950
9	64	9.456	2.609	6.808	3.602	14.180
10	64	9.226	2.670	7.127	4.240	17.800
11	64	9.455	2.632	6.927	4.400	15.980
12.	64	9.114	2.809	7.892	4.420	22.730

Nas Figuras 25, 26 e 27 são apresentados os *boxplots* para região, ano e mês respectivamente, em todos nota-se a presença de diversos *outliers*.

Figura 25 – Boxplot de y_2 para as regiões

Figura 26 – Boxplot de y_2 para os anos Figura 27 – Boxplot de y_2 para os meses

A Tabela 18 apresenta o resultado da ANOVA para o fator Região, dela nota-se que as médias são estatisticamente diferentes para as regiões (com valor- $p < 0.05$). Note que, a região FRNCA (6.905) apresenta a menor média, enquanto as regiões BAURU (12.551) e BRRTS (12.112) apresentam as maiores médias. Nos resultados do sumário do modelo, o fator Região explica 31.98% da variação na resposta y_2 . Indica também que o desvio padrão entre os pontos de dados e os valores ajustados é de aproximadamente 2.50386. A Figura 28 apresenta o gráfico de intervalos de confiança 95% C_rI para a média da resposta y_2 (razão nódulos por exames(a cada 100 exames)) versus cada região. A Figura 29 apresenta a análise residual, onde nota-se que há observações que não se ajustam bem com a distribuição normal, indicando presença de *outliers*.

Tabela 18 – ANOVA: fator Região – 06/2009 a 06/2013 - para y_2

Análise de Variância					
Fonte	GL	$SQ(Aj\cdot)$	$QM(Aj\cdot)$	$ValorF$	$Valor - p$
Amostra	15	2263	150.895	24.07	<0.001
Erro	768	4815	6.269		
Total	783	388880			
Sumário do Modelo					
S	R^2	$R^2(Aj\cdot)$	$R^2(pred)$		
2.50386	31.98%	30.65%	29.11%		
Médias					
Região	$Média$	$DesvP$	$95\%C_{-}\{r\}I$		
ARCTB	8.615	3.340	(7.913, 9.317)		
BAURU	12.551	1.847	(11.849, 9.317)		
BRRTS	12.112	2.435	(11.410, 12.814)		
CMPNS	8.436	0.850	(7.734, 9.139)		
CNTRL	8.623	3.116	(7.921, 9.325)		
FRNCA	6.950	3.959	(6.247, 7.652)		
ITPVA	7.218	2.471	(6.516, 7.920)		
MRLIA	8.298	2.858	(7.596, 9.000)		
PSTPD	7.882	2.293	(7.180, 8.585)		
RBPRT	8.652	2.772	(7.950, 9.354)		
RGSTR	11.507	3.603	(10.805, 12.209)		
RMSPL	9.574	0.991	(8.872, 10.277)		
SJCPS	11.264	2.222	(10.562, 11.966)		
SJDRP	10.867	1.100	(10.165, 11.569)		
SNTOS	10.133	1.908	(9.431, 10.835)		
SRCBA	10.766	1.654	(10.064, 11.468)

Figura 28 – Gráfico intervalo 95% C_rI para a média: y_2 vs Região

Figura 29 – Gráfico de resíduos fator região para y_2

A Tabela 19 apresenta o resultado da ANOVA para o fator Ano, por ela nota-se que as médias são estatisticamente diferentes para os anos (com valor- $p < 0.05$). Note que, o ano 2011 (9.662) apresenta a menor média, enquanto o ano 2009 (10.391) apresenta a maior média. Nos resultados do sumário do modelo, o fator Ano explica 1.57% da variação na resposta y_2 . Indica também que o desvio padrão entre os pontos de dados e os valores ajustados é de aproximadamente 2.99063. A Figura 30 apresenta o gráfico de intervalos de confiança 95% C_rI para a média, da resposta y_2 versus cada ano. A Figura 31 mostra a análise residual, de onde é possível notar a presença de *outliers*, o que faz com que os resíduos não se ajustem bem com a distribuição normal.

Tabela 19 – ANOVA: fator Ano – 06/2009 a 06/2013 - para y_2

Análise de Variância					
Fonte	GL	$SQ(Aj\cdot)$	$QM(Aj\cdot)$	$ValorF$	$Valor - p$
Amostra	4	111.0	27.749	3.10	0.015
Erro	779	6979.3	8.944		
Total	783	7078.3			
Sumário do Modelo					
S	R^2	$R^2(Aj\cdot)$	$R^2(pred)$		
2.99063	1.57%	1.06%	0.24%		
Médias					
Ano	N	$Média$	$DesvP$	$95\%C_{\{r\}}I$	
2009	112	10.391	4.096	(9.836, 10.945)	
2010	192	9.662	2.771	(9.239, 10.086)	
2011	192	9.177	2.590	(8.754, 9.601)	
2012	191	9.429	2.997	(9.005, 9.852)	
2013	96	9.664	2.598	(9.065, 10.263)

Figura 30 – Gráfico intervalo 95% C_rI para a média: y_2 vs Ano

Figura 31 – Gráfico de resíduos fator Ano para y_2

Na Tabela 20 apresenta-se o resultado da ANOVA para o fator Mês, por ela nota-se que as médias não são estatisticamente diferentes para os meses (com valor- $p > 0.05$). Note que, o mês 12 (9.114) apresenta a menor média, enquanto o mês 8 (10.169) apresenta a maior média. Nos resultados do sumário do modelo, o fator Mês explica 1.11% da variação na resposta y_2 . Mostra também que o desvio padrão entre os pontos de dados e os valores ajustados é de aproximadamente 3.01118, o que indica que o fator Mês também não afeta a resposta y_2 de forma sistemática. A Figura 32 apresenta o gráfico de intervalos de confiança 95% C_rI para a média da resposta y_2 versus cada mês. A Figura 33 mostra a análise residual, de onde é possível notar que há observações que não seguem um ajuste normal satisfatório, indicando a presença de *outliers*.

Tabela 20 – ANOVA: fator Mês – 06/2009 a 06/2013 - para y_2

Análise de Variância					
Fonte	GL	$SQ(Aj\cdot)$	$QM(Aj\cdot)$	$ValorF$	$Valor - p$
Amostra	11	78.36	7.123	0.79	0.655
Erro	772	6999.90	9.067		
Total	783	7078.26			
Sumário do Modelo					
	S	R^2	$R^2(Aj\cdot)$	$R^2(pred)$	
3.01118	1.11%		0.00%	0.00%	
Médias					
Mês	N	$Média$	$DesvP$	$95\%C_{-}\{r\}I$	
1	64	9.436	2.626	(8.697, 10.175)	
2	64	9.380	2.524	(8.641, 10.119)	
3	64	9.549	2.924	(8.810, 10.288)	
4	64	9.458	2.734	(8.719, 10.197)	
5	64	9.870	2.902	(9.131, 10.608)	
6	80	9.813	3.377	(9.152, 10.474)	
7	64	10.106	3.767	(9.368, 10.845)	
8	64	10.169	4.019	(9.430, 10.908)	
9	64	9.465	2.609	(9.430, 10.908)	
10	64	9.226	2.670	(8.487, 9.965)	
11	64	9.455	2.632	(8.716, 10.194)	
12	64	9.114	2.809	(8.375, 9.853)

Figura 32 – Gráfico intervalo 95% C_rI para a média: y_2 vs Mês

Figura 33 – Gráfico de resíduos fator Mês para y_2

REGRESSÃO MÚLTIPLA CLÁSSICA (ERROS NORMAIS)

A Figura 34 apresenta os gráficos de dispersão da resposta y_2 (razão nódulos a cada 100 exames) para cada uma das covariáveis consideradas para o estudo: x_1 (grau de urbanização), x_2 (PIB), x_3 (renda), x_4 (número de médicos pela população feminina), Ano e Mês, de onde pode-se observar as relações de cada covariável com a resposta y_2 .

Figura 34 – Gráfico de dispersão para cada covariável vs y_2

A Tabela 21 apresenta os sumários a posteriori para a análise de mínimos quadrados (EMQ). O modelo de regressão ajustado por mínimos quadrados, é dado por,

$$y_1 = 15.46 - 0.271 \times Ano - 0.0404 \times Mes - 10.72 \times x_1 + 0.315 \times x_2 + 3.71 \times x_3 - 0.189 \times x_4 \quad (17)$$

A partir dos resultados da Tabela 13 observa-se que assumindo um modelo de regressão linear múltiplo com erros normais, as covariáveis Ano e x_1 apresentam efeitos significativos na resposta y_2 , pois o valor-p de cada covariável é menor que 0.05, nível de significância assumido. Pela Figura 35 obtêm-se os resultados da análise residual, de onde observa-se que as suposições necessárias (normalidade e variância constante dos resíduos) são razoavelmente aceitáveis.

Tabela 21 – Modelo de regressão linear normal (y_2 - 06/2009 a 06/2013)

Termo	<i>EMQ</i>	<i>ErroPd</i>	<i>T - Value</i>	<i>P - Value</i>
Constante	15.46	2.75	5.63	0.000
Anos	-0.271	0.131	-2.06	0.039
Mês	-0.0404	0.0331	-1.22	0.223
x_1	-10.72	5.27	-2.04	0.042
x_2	0.315	0.351	0.90	0.370
x_3	3.71	2.52	1.47	0.141
x_4	-0.189	0.144	-1.31	0.190

Figura 35 – Gráfico de resíduos (regressão linear normal para resposta y_2)

ABORDAGEM BAYESIANA USANDO UMA DISTRIBUIÇÃO ESTÁVEL

Assume-se agora, uma distribuição estável para o conjunto de dados da resposta y_2 , considerando a presença de um vetor de covariáveis (ano, mês, x_1 , x_2 , x_3 e x_4), assume-se o modelo de regressão para o parâmetro de locação μ da distribuição estável, definido em (9) sob uma abordagem Bayesiana (usando métodos MCMC), dado por,

$$\mu_i = d_0 + d_1 * (ano_i - 2009) + d_2 * mês_i + d_3 * x_{1i} + d_4 * x_{2i} + d_5 * x_{3i} + d_6 * x_{4i} \quad (18)$$

Para esta análise Bayesiana, assume-se uma priori uniforme $U(0, 2)$ para α , uma uniforme $U(-1, 0)$ para β , uma uniforme $U(0, 100)$ para δ , uma uniforme $U(10, 20)$ para o parâmetro de regressão d_0 , uma uniforme $U(-0.30, -0.20)$ para o parâmetro de regressão d_1 , uma uniforme $U(-15, -8)$ para o parâmetro de regressão d_3 , uma uniforme $U(1, 5)$ para o parâmetro de regressão d_5 e uma uniforme $U(-0.5, 0.5)$ para os parâmetros de regressão d_2 , d_4 e d_6 . Na Tabela 22, são apresentados os sumários a posteriori de interesse.

Tabela 22 – Análise Bayesiana - distribuição estável para y_2

	<i>Média</i>	<i>DP</i>	<i>95%$C_r I$</i>
α	1.2	0.003	(1.192, 1.206)
β	-0.075	6.709	(-0.076, -0.074)
d_0	12.91	0.059	(12.81, 13.0)
d_1	-0.272	0.002	(-0.275, -0.268)
d_2	0.166	9.616	(0.164, 0.168)
d_3	-14.15	0.057	(-14.24, -14.05)
d_4	-0.477	0.004	(-0.485, -0.470)
d_5	4.321	0.018	(4.289, 4.349)
d_6	-0.080	0.018	(-0.087, -0.075)
δ	97.1	0.408	(96.42, 97.73)

Pelos resultados da Tabela 22, é possível observar que assumindo a distribuição estável para os dados, todas covariáveis mostram efeito significativo no parâmetro de locação da distribuição estável assumido para a resposta y_2 (razão nódulos a cada 100 exames), uma vez que os intervalos de credibilidade 95% para os parâmetros de regressão (d_1 , d_2 , d_3 , d_4 , d_5 e d_6) não contém o valor zero.

4.3.3 CÂNCER DE MAMA - OUTUBRO/2013 A JUNHO/2015 - PARA y_1

Nesta seção, serão analisados os dados para o período entre outubro/2013 a junho/2015. Começando com os dados para a resposta y_1 : razão entre o número de exames e a população feminina (a cada 1000 mulheres).

Realizando uma análise inicial dos dados, a Tabela 23 apresenta os dados com o fator região, dessa tabela observa-se a maior média para a região de BRRTS e a menor média para a região de SRCBA. A Tabela 24 apresenta os dados com fator ano, sendo 2013 o ano com maior média e 2015 o ano para a menor média. Na tabela 25 apresentam-se os dados com fator mês, sendo o mês 11 com a maior média e o mês 7 com a menor média.

Tabela 23 – Análise descritiva - 10/2013 a 06/2015 - para y_1 fator região, ($n = 21$ para cada região)

Região	Média	<i>DesvP</i>	<i>Variância</i>	<i>Mín.</i>	<i>Máx.</i>
ARCTB	78.60	18.36	337.20	56.10	129.51
BAURU	15.03	14.66	214.98	0.24	43.92
BRRTS	93.78	26.57	706.02	57.72	158.62
CMPNS	26.82	7.35	54.08	15.96	48.28
CNTRL	39.72	7.85	61.69	28.41	55.79
FRNCA	38.31	15.66	245.27	2.53	66.49
ITPVA	49.64	36.80	1353.91	12.00	157.02
MRLIA	18.94	22.33	498.59	1.37	70.41
PSTPD	27.85	24.89	619.55	0.47	76.30
RBPRT	18.39	8.98	80.68	9.72	39.77
RGSTR	26.08	19.49	380.03	1.85	65.18
RMSPL	36.72	3.18	10.13	31.80	44.49
SJCPS	58.12	9.79	95.94	40.66	84.88
SJDRP	44.30	24.73	611.61	10.67	105.39
SNTOS	48.24	7.18	51.55	34.70	62.01
SRCBA	12.45	9.76	95.20	2.95	41.74

Tabela 24 – Análise descritiva - 10/2013 a 06/2015 - para y_1 fator ano

Ano	N	<i>Média</i>	<i>DesvP</i>	<i>Variância</i>	<i>Mín.</i>	<i>Máx.</i>
2013	48	62.23	32.23	1038.61	16.99	158.62
2014	192	36.84	26.49	701.52	0.47	137.05
2015	96	33.68	24.71	610.44	0.24	105.84

Tabela 25 – Análise descritiva - 10/2013 a 06/2015 - para y_1 fator mês

Mês	N	Média	DesvP	Variância	Mín.	Máx.
1	32	38.62	21.39	457.37	1.37	93.81
2	32	42.12	26.52	703.11	2.61	129.51
3	32	33.71	25.10	629.84	3.18	98.89
4	32	32.61	27.11	735.05	2.53	105.84
5	32	35.49	24.91	620.45	0.47	88.50
6	32	32.08	25.61	655.99	0.24	99.36
7	16	29.30	25.62	656.50	2.19	90.01
8	16	29.31	25.19	634.52	2.37	76.65
9	16	32.15	27.85	775.85	1.85	89.06
10	32	53.56	37.30	1391.49	2.07	157.02
11	32	56.60	31.50	992.13	13.33	158.62
12	32	45.25	26.29	691.22	1.81	113.50

Nas Figuras 36, 37 e 38 são apresentados os *boxplots* para região, ano e mês respectivamente, para o período 10/2013 a 06/2015, em todos nota-se a presença de *outliers*.

Figura 36 – Boxplot de y_1 para as regiões - 10/2013 a 06/2015

Figura 37 – Boxplot de y_1 para os anos - 10/2013 a 06/2015

Figura 38 – Boxplot de y_2 para os meses - 10/2013 a 06/2015

A Tabela 26 apresenta o resultado da ANOVA para o fator Região, dela nota-se que as médias são estatisticamente diferentes para as regiões (com valor- $p < 0.05$). Note que, a região FRNCA (6.905) apresenta a menor média, enquanto as regiões BAURU (12.551) e BRRTS (12.112) apresentam as maiores médias. Nos resultados do sumário do modelo, o fator Região explica 31.98% da variação na resposta y_1 . Indica também que o desvio padrão entre os pontos de dados e os valores ajustados é de aproximadamente 2.50386. A Figura 39 apresenta o gráfico de intervalos de confiança 95% $C_r I$ para a média da resposta y_1 (razão entre o número de exames e a população feminina (a cada 1000) mulheres) versus cada região. A Figura 40 apresenta a análise residual, onde nota-se que há observações que não se ajustam bem com a distribuição normal, note a presença de *outliers*.

Tabela 26 – ANOVA: fator Região – 10/2013 a 06/2015 - para y_1 ($n = 21$)

Análise de Variância					
Fonte	GL	$SQ(Aj\cdot)$	$QM(Aj\cdot)$	$ValorF$	$Valor - p$
Amostra	15	161892	10792.8	31.88	<0.001
Erro	320	108329	338.5		
Total	335	270220			
Sumário do Modelo					
S	R^2	$R^2(Aj\cdot)$	$R^2(pred)$		
18.3991	59.91%	58.03%	55.80%		
Médias					
Região	$Média$	$DesvP$	$95\%C_{\{r\}}I$		
ARCTB	78.60	18.36	(70.70, 86.50)		
BAURU	15.03	14.66	(7.13, 22.93)		
BRRTS	93.78	26.57	(85.88, 101.68)		
CMPNS	26.82	7.35	(18.92, 34.72)		
CNTRL	39.72	7.85	(31.82, 47.62)		
FRNCA	38.31	15.66	(30.41, 46.21)		
ITPVA	49.64	36.80	(41.74, 57.54)		
MRLIA	18.94	22.33	(11.04, 26.84)		
PSTPD	27.85	24.89	(19.95, 35.75)		
RBPRT	18.39	8.98	(10.49, 26.29)		
RGSTR	26.08	19.49	(18.18, 33.98)		
RMSPL	36.72	3.18	(28.82, 44.61)		
SJCPS	58.12	9.79	(50.22, 66.02)		
SJDRP	44.30	24.73	(36.40, 52.20)		
SNTOS	48.24	7.18	(40.34, 56.14)		
SRCBA	12.45	9.76	(4.55, 20.35)

Figura 39 – Gráfico intervalo 95% $C_r I$ para a média: y_1 vs Região

Figura 40 – Gráfico de resíduos fator região para y_1

A Tabela 27 apresenta o resultado da ANOVA para o fator Ano, por ela nota-se que as médias são estatisticamente diferentes para os anos (com valor- $p < 0.05$). Note que, o ano 2015 (33.68) apresenta a menor média, enquanto o ano 2013 (62.23) apresenta a maior média. Nos resultados do sumário do modelo, o fator Ano explica 10.89% da variação na resposta y_1 . Indica também que o desvio padrão entre os pontos de dados e os valores ajustados é de aproximadamente 26.8908. A Figura 41 apresenta o gráfico de intervalos de confiança 95% C_rI para a média, da resposta y_1 versus cada ano. A Figura 42 mostra a análise residual, de onde é possível notar a presença de *outliers*, note também que os dados não se ajustem bem com a distribuição normal.

Tabela 27 – ANOVA: fator Ano – 10/2013 a 06/2015 - para y_1

Análise de Variância					
Fonte	GL	$SQ(Aj\cdot)$	$QM(Aj\cdot)$	$ValorF$	$Valor - p$
Amostra	2	29423	14711.5	20.34	<0.001
Erro	333	240797	723,1		
Total	335	270220			
Sumário do Modelo					
	S	R^2	$R^2(Aj\cdot)$	$R^2(pred)$	
26.8908	10.89%	10.35%	9.14%		
Médias					
Ano	N	$Média$	$DesvP$	$95\%C_{\{r\}}I$	
2013	48	62.23	32.23	(54.60, 69.87)	
201	192	36.84	26.49	(33.02, 40.65)	
2015	96	33.68	24.71	(28.28, 39.08)

Figura 41 – Gráfico intervalo 95% C_rI para a média: y_1 vs Ano – 10/2013 a 06/2015

Figura 42 – Gráfico de resíduos fator Ano para y_1 – 10/2013 a 06/2015

Na Tabela 28 apresenta-se o resultado da ANOVA para o fator Mês, nela nota-se que as médias são estatisticamente diferentes para os meses (com valor- $p < 0.05$). Note que, o mês 7 (29.30) apresenta a menor média, enquanto o mês 11 (56.60) apresenta a maior média. Nos resultados do sumário do modelo, o fator Mês explica 9.64% da variação na resposta y_1 . É possível ver também que o desvio padrão entre os pontos de dados e os valores ajustados é de aproximadamente 27.4525. A Figura 43 apresenta o gráfico de intervalos de confiança 95% C_rI para a média da resposta y_1 versus cada mês. A Figura 44 mostra a análise residual, de onde é possível notar que não há um ajuste normal satisfatório, indicando a presença de *outliers*.

Tabela 28 – ANOVA: fator Mês – 10/2013 a 06/2015 - para y_1

Análise de Variância					
Fonte	<i>GL</i>	<i>SQ(Aj·)</i>	<i>QM(Aj·)</i>	<i>ValorF</i>	<i>Valor – p</i>
Amostra	11	26041	2367.4	3.14	<0.001
Erro	324	244179	753.6		
Total	335	270220			
Sumário do Modelo					
S	<i>R</i> ²	<i>R</i> ² (<i>Aj·</i>)	<i>R</i> ² (<i>pred</i>)		
27.4525	9.64%	6.57%	2.88%		
Médias					
Mês	<i>N</i>	<i>Média</i>	<i>DesvP</i>	95% <i>C_{r}I</i>	
1	32	38.62	21.39	(29.07, 48.16)	
2	32	42.12	26.52	(32.57, 51.66)	
3	32	33.71	25.10	(24.16, 43.26)	
4	32	32.61	27.11	(23.06, 42.16)	
5	32	35.49	24.91	(25.94, 45.04)	
6	32	32.08	25.61	(22.53, 41.62)	
7	16	29.30	25.62	(15.80, 42.80)	
8	16	29.31	25.19	(15.81, 42.81)	
9	16	32.15	27.85	(18.65, 45.65)	
10	32	53.56	37.30	(44.01, 63.11)	
11	32	56.60	31.50	(47.05, 66.14)	
12	32	45.25	26.29	(35.71, 54.80)

Figura 43 – Gráfico intervalo 95% $C_r I$ para a média: y_1 vs Mês – 10/2013 a 06/2015

Figura 44 – Gráfico de resíduos fator Mês para y_1 – 10/2013 a 06/2015

REGRESSÃO MÚLTIPLA CLÁSSICA (ERROS NORMAIS)– 10/2013 A 06/2015

Figura 45 apresenta os gráficos de dispersão da resposta y_1 (razão entre o número de exames e a população feminina (a cada 1000 mulheres)) para cada uma das covariáveis consideradas para o estudo: x_1 (grau de urbanização), x_2 (PIB), x_3 (renda), x_4 (número de médicos pela população feminina), Ano e Mês, de onde observar-se as relações de cada covariável com a resposta y_1 .

Figura 45 – Gráfico de dispersão para cada covariável vs y_1

A Tabela 29 apresenta os sumários a posteriori para a análise de mínimos quadrados (EMQ). O modelo de regressão ajustado por mínimos quadrados, é dado por,

$$y_1 = -41.4 - 11.04 \times Ano - 0.093 \times Mes + 158.3 \times x_1 - 3.05 \times x_2 - 19.6 \times x_3 - 5.50 \times x_4 \quad (19)$$

A partir dos resultados da Tabela 29 observa-se que assumindo um modelo de regressão linear múltiplo com erros normais, as covariáveis Ano, x_1 e x_4 apresentam efeitos significativos na resposta y_1 , pois o valor-p de cada covariável é menor que 0.05 (nível de significância assumido). Pela Figura 46 obtêm-se os resultados da análise residual, de onde observa-se que as suposições necessárias (normalidade e variância constante dos resíduos) são razoavelmente aceitáveis, mas não totalmente.

Tabela 29 – Modelo de regressão linear normal (y_1 ~ 10/2013 a 06/2015)

Termo	EMQ	ErroPd	T – Value	P – Value
Constante	.41.4	43.0	-0.96	0.336
Anos	-11.04	3.03	-3.64	0.000
Mês	-0.093	0.513	-0.18	0.857
x_1	158.3	74.3	2.13	0.034
x_2	-3.05	3.44	-0.89	0.377
x_3	-19.6	42.9	-0.46	0.647
x_4	-5.50	2.10	-2.62	0.009

Figura 46 – Gráfico de resíduos (regressão linear normal para resposta y_1 ~ 10/2013 a 06/2015)

ABORDAGEM BAYESIANA USANDO UMA DISTRIBUIÇÃO ESTÁVEL – 10/2013 A 06/2015

Assume-se agora, uma distribuição estável para o conjunto de dados da resposta y_1 , considerando a presença de um vetor de covariáveis (ano, mês, x_1 , x_2 , x_3 e x_4), assume-se o modelo de regressão para o parâmetro de locação μ da distribuição estável, definido em 9 sob uma abordagem Bayesiana (usando métodos MCMC), dado por,

$$\mu_i = d_0 + d_1 * (ano_i - 2013) + d_2 * mês_i + d_3 * x_{1i} + d_4 * x_{2i} + d_5 * x_{3i} + d_6 * x_{4i} \quad (20)$$

Para esta análise Bayesiana, assume-se uma priori uniforme $U(0, 2)$ para α , uma uniforme $U(-1, 0)$ para β , uma uniforme $U(0, 100)$ para δ , uma uniforme $U(0, 100)$ para o parâmetro de regressão d_0 , uma uniforme $U(-50, -30)$ para o parâmetro de regressão d_1 , uma uniforme $U(-0.5, 0.5)$ para o parâmetro de regressão d_2 , uma uniforme $U(140, 180)$ para o parâmetro de regressão d_3 , uma uniforme $U(-5, 0)$ para o parâmetro de regressão d_4 , uma uniforme $U(-30, -10)$ para o parâmetro de regressão d_5 e uma uniforme $U(-10, 0)$ para o parâmetro de regressão d_6 . Na Tabela 30, são apresentados os sumários a posteriori de interesse.

Tabela 30 – Análise Bayesiana - distribuição estável para $y_1 \sim 10/2013$ a 06/2015

	<i>Média</i>	<i>DP</i>	<i>95%C_rI</i>
α	0.915	0.0123	(0.0884, 0.930)
β	-0.390	0.0026	(-0.396, -0.388)
d_0	-40.03	0.1022	(-40.18, -39.85)
d_1	-13.16	0.0353	(-13.21, -13.09)
d_2	0.038	0.0082	(0.025, 0.053)
d_3	156.6	0.418	(155.9, 157.2)
d_4	-0.633	0.0106	(-0.654, 0.621)
d_5	-26.23	0.0734	(-26.33, -26.1)
d_6	-8.317	0.0269	(-8.35, -8.271)
δ	33.74	0.0992	(33.57, 33.88)

Pelos resultados da Tabela 30, é possível observar que assumindo a distribuição estável para os dados, todas covariáveis mostram efeito significativo no parâmetro de locação da distribuição estável assumido para a resposta y_1 (razão entre o número de exames pela população feminina (a cada 1000 mulheres)), uma vez que os intervalos de credibilidade 95% para os parâmetros de regressão (d_1 , d_2 , d_3 , d_4 , d_5 e d_6) não contém o valor zero.

4.3.4 CÂNCER DE MAMA - OUTUBRO/2013 A JUNHO/2015 - PARA y_2

Finalmente, a última a análise descritiva dos dados, é do período outubro/2013 a junho/2015 para a resposta y_2 : razão nódulos a cada 100 exames. A Tabela 31 apresenta os dados com o fator região administrativa, dessa tabela observa-se a maior média para a região de MRLIA e a menor média para a região de ITPVA. A Tabela 32 apresenta os dados com fator ano, sendo 2015 o ano com maior média e 2013 o ano para a menor média. Na tabela 33 apresentam-se os dados com fator mês, sendo o mês 3 com a maior média e o mês 8 com a menor média.

Tabela 31 – Análise descritiva - 10/2013 a 06/2015 - para y_2 fator região

Região	N	<i>Média</i>	<i>DP</i>	<i>Variância</i>	<i>Mín.</i>	<i>Máx.</i>
ARCTB	21	9.63	2.35	5.56	5.56	16.24
BAURU	20	8.48	4.85	23.57	2.88	16.35
BRRTS	21	8.90	1.40	1.95	6.66	12.02
CMPNS	21	7.54	0.85	0.72	5.78	8.74
CNTRL	21	4.45	1.56	2.44	1.99	7.97
FRNCA	21	9.00	1.69	2.88	5.74	12.46
ITPVA	21	3.94	2.04	4.16	0.32	7.79
MRLIA	20	17.49	7.70	59.27	5.05	29.25
PSTPD	21	6.98	5.55	30.83	0.31	25.00
RBPRT	21	6.93	1.53	2.35	3.50	10.47
RGSTR	21	11.42	4.08	16.70	3.12	24.00
RMSPL	21	8.62	0.60	0.36	7.61	9.75
SJCPS	21	9.24	1.22	1.50	6.92	11.82
SJDRP	21	11.67	3.44	11.86	6.02	20.97
SNTOS	21	8.56	1.26	1.59	5.61	11.30
SRCBA	21	10.34	2.35	5.53	5.92	14.74

Tabela 32 – Análise descritiva - 10/2013 a 06/2015 - para y_2 fator ano

Ano	N	<i>Média</i>	<i>DesvP</i>	<i>Variância</i>	<i>Mín.</i>	<i>Máx.</i>
2013	48	8.52	2.45	6.00	4.01	16.35
2014	192	8.86	4.58	21.05	0.31	28.20
2015	94	9.26	4.61	21.33	1.80	29.25

Tabela 33 – Análise descritiva - 10/2013 a 06/2015 - para y_2 fator mês

Mês	N	Média	DesvP	Variância	Mín.	Máx.
1	31	9.10	2.76	7.66	1.99	16.28
2	32	9.54	4.46	19.93	2.87	25.00
3	32	9.82	5.08	25.86	1.80	24.85
4	32	9.20	5.60	31.45	2.13	29.25
5	32	9.80	5.59	31.29	2.35	28.20
6	31	8.76	4.23	17.95	2.05	25.75
7	16	8.33	4.84	23.42	0.32	20.97
8	16	7.41	3.35	11.22	1.67	14.52
9	16	9.21	6.50	42.25	2.36	24.00
10	32	9.20	3.52	12.42	0.31	16.35
11	32	7.77	2.62	6.89	2.09	11.92
12	32	8.05	2.78	7.77	1.75	14.85

Nas Figuras 47, 48 e 49 são apresentados os *boxplots* para região, ano e mês respectivamente, em todos nota-se a presença de diversos *outliers*.

Figura 47 – Boxplot de y_2 para as regiões - 10/2013 a 06/2015

Figura 48 – Boxplot de y_2 para os anos - 10/2013 a 06/2015

Figura 49 – Boxplot de y_2 para os meses - 10/2013 a 06/2015

A Tabela 34 apresenta o resultado da ANOVA para o fator Região, dela nota-se que as médias são estatisticamente diferentes para as regiões (com valor- $p < 0.05$). Note que, a região ITPVA (3.944) apresenta a menor média, enquanto as regiões RGSTR (11.421) e SJDRP (11.671) apresentam as maiores médias. Nos resultados mostrados no sumário do modelo, o fator Região explica 46.98% da variação na resposta y_2 . Indica também que o desvio padrão entre os pontos de dados e os valores ajustados é de aproximadamente 3.24250. A Figura 50 apresenta o gráfico de intervalos de confiança 95% C_rI para a média da resposta y_2 (razão nódulos por exames(a cada 100 exames)) versus cada região. A Figura 51 apresenta a análise residual, onde nota-se que as observações não se ajustam bem com a distribuição normal.

Tabela 34 – ANOVA: fator Região – 10/2010 a 06/2015 - para y_2

Análise de Variância					
Fonte	GL	$SQ(Aj\cdot)$	$QM(Aj\cdot)$	$Valor F$	$Valor - p$
Amostra	15	2963	197.54	18.79	<0.001
Erro	318	3343	10.51		
Total	333	6306			
Sumário do Modelo					
S	R^2	$R^2(Aj\cdot)$	$R^2(pred)$		
3.24250	46.98%	44.48%	41.41%		
Médias					
Região	N	$Média$	$DesvP$	$95\%C_{\{r\}}I$	
ARCTB	21	9.63	2.358	(8.241, 11.025)	
BAURU	20	8.48	4.85	(7.06, 9.91)	
BRRTS	21	8.909	1.400	(7.516, 10.301)	
CMPNS	21	7.547	0.851	(6.155, 8.939)	
CNTRL	21	4.453	1.563	(3.061, 5.845)	
FRNCA	21	9.006	1.698	(7.614, 10.398)	
ITPVA	21	3.944	2.040	(2.552, 5.336)	
MRLIA	20	17.49	7.70	(16.06, 18.91)	
PSTPD	21	6.98	5.55	(5.58, 8.37)	
RBPRT	21	6.937	1.535	(5.545, 8.329)	
RGSTR	21	11.421	4.088	(10.029, 12.813)	
RGSTR	21	8.622	0.600	(7.230, 10.014)	
SJCPS	21	9.248	1.226	(7.856, 10.640)	
SJDRP	21	11.671	3.444	(10.279, 13.063)	
SNTOS	21	8.562	1.262	(7.170, 9.954)	
SRCBA	21	10.346	2.352	(8.954, 11.738)

Figura 50 – Gráfico intervalo 95% $C_r I$ para a média: y_2 vs Região – 10/2013 a 06/2015

Figura 51 – Gráfico de resíduos fator região para y_2 – 10/2013 a 06/2015

A Tabela 35 apresenta o resultado da ANOVA para o fator Ano, por ela nota-se que as médias não são estatisticamente diferentes para os anos (com valor- $p > 0.05$). Note que, o ano 2013 (8.529) apresenta a menor média, enquanto o ano 2015 (9.267) apresenta a maior média. Nos resultados do sumário do modelo, o fator Ano explica 0.31% da variação na resposta y_2 . Indica também que o desvio padrão entre os pontos de dados e os valores ajustados é de aproximadamente 4.35827. A Figura 52 apresenta o gráfico de intervalos de confiança 95% $C_r I$ para a média, da resposta y_2 versus cada ano. A Figura 53 mostra a análise residual, de onde é possível notar a presença de *outliers*, o que faz com que os dados não se ajustem bem com a distribuição normal.

Tabela 35 – ANOVA: fator Ano – 10/2013 a 06/2015 - para y_2

Análise de Variância					
Fonte	GL	$SQ(Aj\cdot)$	$QM(Aj\cdot)$	$Valor F$	$Valor - p$
Amostra	2	19.25	9.625	0.51	0.603
Erro	331	6287.18	18.994		
Total	333	6306.43			
Sumário do Modelo					
	S	R^2	$R^2(Aj\cdot)$	$R^2(pred)$	
4.35827		0.31%	0.00%	0.00%	
Médias					
Ano	N	$Média$	$DesvP$	$95\%C_{-}\{r\}I$	
2013	48	8.529	2.450	(7.291, 9.766)	
2014	192	8.863	4.588	(8.244, 9.482)	
2015	94	9.267	4.619	(8.382, 10.151)

Figura 52 – Gráfico intervalo 95% C_rI para a média: y_2 vs Ano – 10/2013 a 06/2015

Figura 53 – Gráfico de resíduos fator Ano para y_2 – 10/2013 a 06/2015

Na Tabela 36 apresenta-se o resultado da ANOVA para o fator Mês, por ela nota-se que as médias não são estatisticamente diferentes para os meses (com valor- $p > 0.05$). Note que, o mês 8 (7.418) apresenta a menor média, enquanto o mês 3 (9.824) apresenta a maior média. Nos resultados do sumário do modelo, o fator Mês explica 2.85% da variação na resposta y_2 . Mostra também que o desvio padrão entre os pontos de dados e os valores ajustados é de aproximadamente 4.36201, o que indica que o fator Mês também não afeta a resposta y_2 de forma sistemática. A Figura 54 apresenta o gráfico de intervalos de confiança 95% C_rI para a média da resposta y_2 versus cada mês. A Figura 55 mostra a análise residual, de onde é possível notar que há observações que não seguem um ajuste normal satisfatório, mostrando a presença de *outliers*.

Tabela 36 – ANOVA: fator Mês – 10/2013 a 06/2015 - para y_2

Análise de Variância					
Fonte	GL	$SQ(Aj\cdot)$	$QM(Aj\cdot)$	$Valor F$	$Valor - p$
Amostra	11	179.7	16.34	0.86	0.582
Erro	322	6126.7	19.03		
Total	333	6306.4			
Sumário do Modelo					
	S	R^2	$R^2(Aj\cdot)$	$R^2(pred)$	
4.36201	2.85%	0.00%	0.00%		
Médias					
Mês	N	$Média$	$DesvP$	$95\%C_{\{r\}}I$	
1	31	9.109	2.769	(7.568, 10.651)	
2	32	9.544	4.464	(8.027, 11.061)	
3	32	9.824	5.086	(8.307, 11.341)	
4	32	9.207	5.609	(7.690, 10.724)	
5	32	9.802	5.594	(8.285, 11.319)	
6	31	8.762	4.237	(7.220, 10.303)	
7	16	8.33	4.84	(6.18, 10.47)	
8	16	7.418	3.350	(5.272, 9.563)	
9	16	9.21	6.50	(7.06, 11.35)	
10	32	9.200	3.525	(7.683, 10.717)	
11	32	7.775	2.627	6.258, 9.292	
12	32	8.050	2.788	6.533, 9.567

Figura 54 – Gráfico intervalo 95% C_rI para a média: y_2 vs Mês – 10/2013 a 06/2015

Figura 55 – Gráfico de resíduos fator Mês para y_2 – 10/2013 a 06/2015

REGRESSÃO MÚLTIPLA CLÁSSICA (ERROS NORMAIS) – 10/2013 A 06/2015

Figura 56 apresenta os gráficos de dispersão da resposta y_2 (razão nódulos a cada 100 exames) para cada uma das covariáveis consideradas para o estudo: x_1 (grau de urbanização), x_2 (PIB), x_3 (renda), x_4 (número de médicos pela população feminina), Ano e Mês, de onde pode-se observar as relações de cada covariável com a resposta y_2 .

Figura 56 – Gráfico de dispersão para cada covariável vs y_2 – 10/2013 a 06/2015

A Tabela 37 apresenta os sumários a posteriori para a análise de mínimos quadrados (EMQ). O modelo de regressão ajustado por mínimos quadrados é dado por,

$$y_1 = 5.64 - 0.438 \times Ano - 0.1588 \times Mes + 11.9 \times x_1 + 0.579 \times x_2 - 12.17 \times x_3 + 0.934 \times x_4 \quad (21)$$

A partir dos resultados da Tabela 37 observa-se que assumindo um modelo de regressão linear múltiplo com erros normais, as covariáveis Mês, x_3 e x_4 apresentam efeitos significativos na resposta y_2 , pois o valor-p é menor que 0.10 (nível de significância assumido). Pela Figura 57 obtêm-se os resultados da análise residual, de onde observa-se que as suposições necessárias (normalidade e variância constante dos resíduos) são razoavelmente aceitáveis (não totalmente).

Tabela 37 – Modelo de regressão linear normal (y_2 - 10/2013 a 06/2015)

Termo	<i>EMQ</i>	<i>ErroPd</i>	<i>Valor - T</i>	<i>Valor - p</i>
Constant	5.64	7.04	0.80	0.42
Anos	-0.43	0.49	-0.88	0.38
Mês	-0.15	0.08	-1.89	0.06
x_1	11.9	12.2	0.98	0.33
x_2	0.57	0.56	1.03	0.30
x_3	-12.17	7.02	-1.73	0.084
x_4	0.93	0.34	2.71	0.007

Figura 57 – Gráfico de resíduos (regressão linear normal para resposta y_2) - 10/2013 a 06/2015

ABORDAGEM BAYESIANA USANDO UMA DISTRIBUIÇÃO ESTÁVEL - 10/2013 A 06/2015

Assume-se agora, uma distribuição estável para o conjunto de dados da resposta y_2 , considerando a presença de um vetor de covariáveis (ano, mês, x_1 , x_2 , x_3 e x_4), assume-se o modelo de regressão para o parâmetro de locação μ da distribuição estável, definido em 9 sob uma abordagem Bayesiana (usando métodos MCMC), dado por,

$$\mu_i = d_0 + d_1 * (ano_i - 2013) + d_2 * mês_i + d_3 * x_{1i} + d_4 * x_{2i} + d_5 * x_{3i} + d_6 * x_{4i} \quad (22)$$

Para esta análise Bayesiana, assume-se uma priori uniforme $U(0, 2)$ para α , uma uniforme $U(-1, 0)$ para β , uma uniforme $U(0, 100)$ para δ , uma uniforme $U(0, 10)$ para o parâmetro de regressão d_0 , uma uniforme $U(-1, 1)$ para o parâmetro de regressão d_1 , uma uniforme $U(-0.5, 0.5)$ para o parâmetro de regressão d_2 , uma uniforme $U(8, 15)$ para o parâmetro de regressão d_3 , uma uniforme $U(0, 1)$ para o parâmetro de regressão d_4 , uma uniforme $U(-20, 0)$ para o parâmetro de regressão d_5 e uma uniforme $U(0, 3)$ para o parâmetro de regressão d_6 . Na Tabela 22, são apresentados os sumários a posteriori de interesse.

Tabela 38 – Análise Bayesiana - distribuição estável para y_2 - 10/2013 a 06/2015

	<i>Média</i>	<i>DP</i>	<i>95%CrI</i>
α	0.707	0.0100	(0.688, 0.724)
β	-0.656	0.0142	(-0.677, -0.627)
d_0	1.217	0.0465	(1.157, 1.307)
d_1	-0.636	0.0463	(-0.735, -0.583)
d_2	-0.092	0.0164	(-0.120, -0.048)
d_3	8.037	0.0526	(8.0, 8.23)
d_4	0.883	0.0317	(0.818, 0.935)
d_5	-10.6	0.0872	(-10.91, -10.52)
d_6	0.277	0.0308	(0.225, 0.321)
δ	19.89	0.1801	(19.69, 20.52)

Pelos resultados da Tabela 38, é possível observar que assumindo a distribuição estável para os dados, todas covariáveis mostram efeito significativo no parâmetro de locação da distribuição estável assumido para a resposta y_2 (razão nódulos a cada 100 exames), uma vez que os intervalos de credibilidade 95% para os parâmetros de regressão (d_1 , d_2 , d_3 , d_4 , d_5 e d_6) não contém o valor zero.

4.3.5 INTERPRETAÇÃO DOS RESULTADOS OBTIDOS

Apresentamos a seguir um sumário dos resultados obtidos da análise estatística.

Para a **resposta** Y_1 : razão número de exames/1000 mulheres no período 2009/2013. Os fatores significativos (análise clássica e análise Bayesiana) foram:

- **Região:** melhores razões exames/1000 mulheres foram observados para as regiões ARCTB, BRRTS, ITPVA e SJDRP apesar de mostrarem em alguns casos grande variabilidade (ver, boxplots da Figura 14). Essas regiões são de cidades ricas do interior do estado de São Paulo que possivelmente tem melhores estruturas de saúde pública.
- **Ano:** os últimos anos do período observado (2011, 2012, 2013) apresentam razões maiores, o que indica melhoria do serviço de saúde no estado de São Paulo no decorrer dos anos do período observado.
- Os fatores x_1 (grau de urbanização), x_2 (PIB), x_3 (renda) e x_4 (número de médicos para a população feminina) que são fatores relacionados a estrutura da região apresentam algum efeito significativo usando a metodologia Bayesiana assumindo uma distribuição estável (maior sensibilidade para detectar fatores significativos).

Para a **resposta** Y_2 : razão nódulos a cada 100 exames no período 2009/2013. Os fatores significativos (análise clássica e análise Bayesiana) foram:

- **Região:** melhor (valor menor de razão nódulos/100 exames) foi observado para a região FRNCA (ver, boxplots da Figura 25). As piores regiões (valores maiores de razão nódulos/100 exames) são BAURU, BRRTS, RGSTR e SJCPS (ver, boxplots da Figura 25). Possivelmente existe algum outro fator não considerado afetando a resposta.
- **Ano:** os últimos anos do período observado (2011, 2012, 2013) apresentam razões em geral menores, o que indica melhoria do serviço de saúde no estado de São Paulo no decorrer dos anos do período observado. O melhor ano (valor menor de razão nódulos /100 exames) observado foi o ano 2011.
- Os fatores x_1 (grau de urbanização), x_2 (PIB), x_3 (renda) e x_4 (número de médicos para a população feminina) que são fatores relacionados a estrutura da região

apresentam algum efeito significativo usando a metodologia Bayesiana assumindo uma distribuição estável (maior sensibilidade para detectar fatores significativos).

Para a **resposta** Y_1 : razão número de exames/1000 mulheres no período 2013/2015. Os fatores significativos (análise clássica e análise Bayesiana) foram:

- **Região:** melhores razões exames/1000 mulheres foram observados para as regiões ARCTB e BRRTS apesar de mostrarem em alguns casos grande variabilidade (ver, boxplots da Figura 36).
- **Ano:** os dois últimos anos do período observado (2014 e 2015) apresentam razões menores o que indica piora do serviço de saúde no estado de São Paulo no decorrer dos anos do período observado (ver Figuras 37 e 41).
- Os fatores x_1 (grau de urbanização), x_2 (PIB), x_3 (renda) e x_4 (número de médicos para a população feminina) que são fatores relacionados a estrutura da região apresentam algum efeito significativo usando a metodologia Bayesiana assumindo uma distribuição estável (maior sensibilidade para detectar fatores significativos).

Para a **resposta** Y_2 : razão nódulos a cada 100 exames no período 2013/2015. Os fatores significativos (análise clássica e análise Bayesiana) foram:

- **Região:** melhores (valor menor de razão nódulos/100 exames) foi observado para as regiões CNTRL e ITPVA (ver, boxplots da Figura 47). A pior região (valor maior de razão nódulos/100 exames) é MRLIA (ver, boxplots da Figura 47). Possivelmente existe algum outro fator não considerado afetando a resposta.
- Os fatores x_1 (grau de urbanização), x_2 (PIB), x_3 (renda) e x_4 (número de médicos para a população feminina) que são fatores relacionados a estrutura da região apresentam algum efeito significativo usando a metodologia Bayesiana assumindo uma distribuição estável (maior sensibilidade para detectar fatores significativos).

Em geral observa-se que para a resposta Y_1 (razão número exames/1000 mulheres) a média amostral observada para o período (2009/2013) é igual à 47.3986, enquanto que para o período (2013/2015) a média amostral observada é igual à 39.5627, o que indica uma piora no serviço de saúde do estado de São Paulo.

Considerando-se a resposta Y_2 (razão nódulos a cada 100 exames) a média amostral observada para o período (2009/2013) é igual à 9.59055, enquanto que para o período

(2013/2015) a média amostral observada é igual à 8.92853, o que indica uma pequena melhoria no estado de São Paulo, apesar da piora do serviço de saúde.

5 CONCLUSÃO

O uso de distribuições que generalizam a habitual distribuição Gaussiana é de grande interesse na análise de dados, uma vez que em muitas aplicações é possível haver desvios da suposição normal.

Em muitas aplicações, é comum ter transformações da resposta de interesse como logaritmo ou transformações [Box e Cox \(1964\)](#), mas em aplicações práticas, [Schmidt e Finan \(2018\)](#) aponta que as transformações de resultado mudam a estimativa alvo e, portanto, o viés. Portanto, não recomendam transformações, especialmente para amostras grandes.

O uso de distribuições estáveis pode ser uma boa alternativa nas aplicações, devido à sua flexibilidade para ser ajustada por diferentes conjuntos de dados e resultados robustos na presença de *outliers* (ver, por exemplo, ([ACHCAR et al., 2013](#))). Essa distribuição não é muito popular na análise de dados, uma vez que há grandes dificuldades para obter as inferências de interesse sob a abordagem clássica usando métodos de estimativa de máxima verossimilhança ou usando abordagem Bayesiana na presença de um vetor de covariáveis, uma situação comum especialmente em estudos médicos.

Neste estudo foi exposto que usando modelos de regressão conjunto para os parâmetros de locação e escala da distribuição estável sob uma abordagem Bayesiana e, usando métodos MCMC existentes com prioris uniformes para todos os parâmetros, foram obtidas boas inferências para os parâmetros do modelo. É importante apontar que usando um único modelo de regressão linear para o parâmetro de locação da distribuição estável assumindo distribuições normais como prioris para os parâmetros de regressão não foi possível obter bons resultados de convergência do algoritmo de simulação MCMC usando *software* OpenBugs mesmo após um grande número de iterações, possivelmente devido a problemas de identificabilidade do modelo.

Com o uso da distribuição estável observamos que para os quatro conjuntos de dados do câncer de mama, todas as covariáveis apresentaram efeitos significativos nas respostas. Em todos os casos, tivemos estimadores Bayesianos diferentes de 2 (próximos dos valores de 1), indicando que a distribuição normal não se ajusta muito bem aos dados, por esse motivo, a maior sensibilidade em detectar efeitos significativos de todas as covariáveis.

As inferências para o modelo com distribuição estável são muito sensíveis a escolha das distribuições a priori para os parâmetros. No presente estudo, foram escolhidas prioris usando a informação dos resultados clássicos (métodos Bayesianos empíricos). Possivelmente outras estruturas computacionais devam ser usadas para essa classe de modelos (objeto de pesquisas futuras). Outro problema que pode existir é a multicolinearidade.

É importante ressaltar a respeito dos resultados obtidos neste estudo: a utilização do *software* gratuito OpenBugs que requer apenas a função de verossimilhança obtida do método proposto por [Buckle \(1995\)](#) e a especificação das distribuições a priori que podem ser muito úteis em aplicações.

REFERÊNCIAS

- ACHCAR, J. A.; ACHCAR, A.; MARTINEZ, E. Z. Robust linear regression models: use of a stable distribution for the response data. *Open Journal of Statistics*, v. 3, p. 409–416, 2013. Citado na página 16.
- ACHCAR, J. A.; LOPES, S.; MAZUCHELI, J.; LINHARES, R. A bayesian approach for stable distributions: Some computational aspects. *Open Journal of Statistics*, v. 3, p. 268–277, 2013. Citado na página 92.
- BERNARDO, J.; SMITH, A. F. M. *Bayesian Theory*. [S.l.]: Statistics in Medicine - STAT MED, 2000. v. 15. Citado 3 vezes nas páginas 30, 31 e 32.
- BOX, G. E. P.; COX, D. R. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, Blackwell Publishing for the Royal Statistical Society, v. 26, n. 2, p. 211–252, 1964. Citado na página 92.
- BOX, G. E. P.; TIAO, G. C. Bayesian inference in statistical analysis. Addison-Wesley, 1973. Citado 2 vezes nas páginas 28 e 29.
- BUCKLE, D. J. Bayesian inference for stable distributions. *Journal of the American Statistical Association*, v. 90, p. 605–613, 1995. Citado 4 vezes nas páginas 22, 33, 34 e 93.
- CHIB, S.; GREENBERG, E. Understanding the metropolis-hastings algorithm. *The American Statistician*, v. 49, n. 4, p. 327–335, 1995. Citado 3 vezes nas páginas 17, 30 e 31.
- DAMIEN, P.; WAKEFIELD, J.; WALKER, S. Gibbs sampling for bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society, Series B*, v. 61, p. 331–344, 1999. Citado 2 vezes nas páginas 17 e 33.
- DRAPER, N.; SMITH, H. Applied regression analysis. Wiley series in probability and mathematical statistics, 1981. Citado 2 vezes nas páginas 16 e 37.
- GAMERMAN, D.; LOPES, H. F. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. London, GB: Chapman and Hall CRC, 2006. Citado na página 30.
- GELFAND, A. E.; SMITH, A. F. M. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, v. 85, n. 410, p. 398–409, 1990. Citado 3 vezes nas páginas 17, 30 e 31.
- GNEDENKO, B. V.; KOLMOGOROV, A. N. Limit distributions for sums of independent random variables. Addison-Wesley, Massachussetts, 1968. Citado na página 22.
- IBRAGIMOV, A.; CERNIN, K. E. On the unimodality of stable laws. *Teor. Veroyatnost. i Primenen.*, v. 4, p. 453–456, 1959. Citado na página 23.
- KANTER, M. On the unimodality of stable densities. *Annals of Probability*, n. 4, p. 1006–1008, 1976. Citado na página 23.
- LIMA-COSTA, M. F.; BARRETO, S. M. Tipos de estudos epidemiológicos: conceitos básicos e aplicações na área do envelhecimento. *Epidemiologia e Serviços de Saúde*, v. 12, n. 4, p. 189–201, 2003. Citado na página 18.

- NOLAN, J. P. Models for heavy tailed data. *Birkhauser, Boston*, In progress, Chapter 1 online at academic2.american.edu/~jpnolan., 2009. Citado 3 vezes nas páginas 17, 22 e 23.
- PAULINO, C.; TURKMAN, M. A. A.; MURTEIRA, B. Estatística bayesiana. Fundação Calouste Gulbenkian, Lisboa, 2003. Citado 2 vezes nas páginas 28 e 29.
- SCHMIDT, A.; FINAN, C. Linear regression and the normality assumption. *Journal of Clinical Epidemiology*, ELSEVIER SCIENCE INC, v. 98, p. 146–151, 2018. Citado 2 vezes nas páginas 16 e 92.
- SEBER, G.; LEE, A. Linear regression analysis. 2nd ed., Wiley series in probability and mathematical statistics, 2003. Citado na página 37.
- SHU, P. S.; CHAN, Y. M.; L., S. Higher body mass index and lower intake of dairy products predict poor glycaemic control among type 2 diabetes patients in malaysia. *PLoS One*, Public Library of Science, v. 12, n. 2, p. 1–15, 2017. Citado na página 42.
- SKOROHOD, A. V. On a theorem concerning stable distributions. *Select. Transl. Math. Statist. and Probability*, Inst. Math. Statist. and Amer. Math. Soc., Providence, R.I., v. 1, p. 169–170, 1961. Citado 2 vezes nas páginas 22 e 23.
- SPIEGELHALTER, D. J.; THOMAS, A.; BEST, N. G.; LUNN, D. Winbugs user's manual. MRC Biostatistics Unit, Cambridge, 2003. Citado 3 vezes nas páginas 33, 35 e 39.
- TAGER, I.; WEISS, S.; A., M.; ROSNER, B.; SPEIZER, F. Longitudinal study of the effects of maternal smoking on pulmonary function. *New England Journal of Medicine*, n. 309(12), p. 699–703, 1983. Citado na página 36.
- TANNER, M. A.; WONG, W. H. The calculation of posterior distributions by data augmentation. *Journal of American Statistical Association*, v. 82, p. 528–550, 1987. Citado na página 33.
- WEST, M.; HARRISON, J. *Bayesian Forecasting and Dynamic Models*. New York, NY, USA: 2nd, Springer-Verlag New York, Inc., 1997. Citado na página 30.

Apêndice A – CÓDIGOS OPENBUGS - DADOS FEV

A.1 CÓDIGO SEM PRESENÇA DE COVARIÁVEIS

Listing A.1 – Distribuição estável para os dados da resposta FEV (sem presença de covariáveis)

```

1  model{
2    for (in in 1:n)  {
3      dummy[i] <- 0
4      x[i] <- FEV[i]
5      dummy[i] ~ dloglik(logLike[i])
6      z[i] <- (x[i] - mu)/sigma
7      w[i] <- abs(z[i]/t[i])
8      a1[i] <- (sin(3.14159*alpha*y[i] + b1))/cos(3.14159*y[i])
9      a2[i] <- (cos(3.14159*y[i]))/cos(3.14159*(alpha - 1) *
10     * y[i] + b1)
11     a3[i] <- pow(a2[i], theta2)
12     t[i] <- a1[i]*a3[i]
13     logLike[i] <- log(alpha) - log(abs(alpha - 1)) - log(sigma) -
14     - pow(w[i], theta1) + theta1*log(w[i]) - log(abs(z[i]))
15     y[i] ~ dunif(-0.5, 0.5)
16   }
17   theta1 <- alpha/(alpha - 1)
18   theta2 <- (alpha - 1)/alpha
19   b1 <- (beta*min(alpha, 2-alpha)*3.14159)/2
20   alpha ~ dunif(1, 2)
21   beta ~ dunif(-1, 1)
22   sigma ~ dunif(0, 2)
23   mu ~ dunif (-2, 2)
24 }

```

A.2 CÓDIGO COM PRESENÇA DE COVARIÁVEIS

Listing A.2 – Distribuição estável para os dados (regressão nos parâmetros de locação e escala)

```

1  model{
2    for (in in 1:n)  {
3      dummy[i] <- 0
4      x[i] <- FEV[i]
5      dummy[i] ~ dloglik(logLike[i])
6      z[i] <- (x[i] - mu[i])/sigma [i]
7      mu[i] <- d0 + d1*altura[i] + d2*idade[i] + d3*genero[i] +
8      + d4*fumante[i]
9      sigma[i]<- exp(e0 + e1*altura[i] + e2*idade[i] + e3*genero[i] +
10     + e4*fumante[i])
11     w[i] <- abs(z[i]/t[i])
12     a1[i] <- (sin(3.14159*alpha*y[i] + b1))/cos(3.14159*y[i])
13     a2[i] <- (cos(3.14159*y[i]))/cos(3.14159*(alpha - 1) *
14     * y[i] + b1)
15     a3[i] <- pow(a2[i], theta2)
16     t[i] <- a1[i]*a3[i]
17     logLike[i] <- log(alpha) - log(abs(alpha - 1)) - log(sigma) -
18     - pow(w[i], theta1) + theta1*log(w[i]) - log(abs(z[i]))
19     y[i] ~ dunif(-0.5, 0.5)
20   }
21   theta1 <- alpha/(alpha - 1)
22   theta2 <- (alpha - 1)/alpha
23   b1 <- (beta*min(alpha, 2-alpha)*3.14159)/2
24   alpha ~ dunif(1.5, 2)
25   beta ~ dunif(-0.5, 0.5)
26   d0 ~ dunif(2, 5)
27   d1 ~ dunif(-0.5, 0.5)
28   d2 ~ dunif(-1, 1)
29   d3 ~ dunif(-1, 1)
30   d4 ~ dunif(-1, 1)
31   e0 ~ dunif(-1, 1)
32   e1 ~ dunif(-1, 1)
33   e2 ~ dunif(-1, 1)

```

```
34 e3 ~ dunif(-1, 1)
35 e4 ~ dunif(-1, 1)
36 }
```

Apêndice B – CÓDIGO OPENBUGS - DADOS CÂNCER DE MAMA

B.1 PERÍODO 06/2009 A 06/2013 - PARA Y1

Listing B.1 – Distribuição estável para os dados 06/2009-06/2013 (y_1)

```

1  model {
2    for (i in 1:n) {
3      dummy[i] <- 0
4      dummy[i] ~ dloglik(logLike[i])
5      z[i] <- (response[i] - mu[i])/delta
6      mu[i] <- d0 + d1 * ano[i] + d2 * mes[i] + d3*x1[i] + d4*x2[i] +
7      + d5*x3[i] + d6*x4[i]
8      w[i] <- abs(z[i]/t[i])
9      a1[i] <- (sin(3.14159*alpha*y[i] + b1))/cos(3.14159*y[i])
10     a2[i] <- (cos(3.14159 * y[i]))/cos(3.14159*(alpha-1)*y[i] + b1)
11     a3[i] <- pow(a2[i], theta2)
12     t[i] <- a1[i]*a3[i]
13     logLike[i] <- log(alpha) - log(abs(alpha-1)) - log(delta) -
14     - pow(w[i], theta1) + theta1*log(w[i]) - log(abs(z[i]))
15     y[i] ~ dunif(-0.5, 0.5)
16   }
17   theta1 <- alpha/(alpha-1)
18   theta2 <- (alpha-1)/alpha
19   b1 <- (beta*min(alpha, 2-alpha)*3.14159)/2
20   alpha ~ dunif(0, 2)
21   beta ~ dunif(-1, 0)
22   delta ~ dunif(0, 100)
23   d0 ~ dnorm(-24000, 0.001)
24   d1 ~ dnorm(12, 0.1)
25   d2 ~ dnorm(1, 0.1)
26   d3 ~ dnorm(-78, 0.1)
27   d4 ~ dnorm(-14, 0.1)
28   d5 ~ dnorm(21, 0.1)

```

```

29 d6~ dnorm( -0.5,0.1)
30 }

```

B.2 PERÍODO 06/2009 A 06/2013 - PARA Y_2

Listing B.2 – Distribuição estável para os dados 06/2009-06/2013 (y_2)

```

1  model {
2    for (i in 1:n) {
3      dummy[i] <- 0
4      dummy[i] ~ dloglik(logLike[i])
5      z[i] <- (response[i]- mu[i])/delta
6      mu[i] <- d0+d1*(ano[i]-2009)+d2*mes[i]+d3*x1[i]+d4*x2[i] +
7      + d5*x3[i]+d6*x4[i]
8      w[i] <- abs(z[i]/t[i])
9      a1[i] <- (sin(3.14159*alpha*y[i]+b1))/cos(3.14159*y[i])
10     a2[i] <- (cos(3.14159 *y[i]))/cos(3.14159*(alpha-1)*y[i]+b1)
11     a3[i] <- pow(a2[i],theta2)
12     t[i] <- a1[i]*a3[i]
13     logLike[i]<-log(alpha)-log(abs(alpha-1))-log(delta)-
14     - pow(w[i],theta1)+theta1*log(w[i])-log(abs(z[i]))
15     y[i] ~ dunif(-0.5,0.5)
16   }
17   theta1<-alpha/(alpha-1)
18   theta2<-(alpha-1)/alpha
19   b1<-(beta*min(alpha,2-alpha)*3.14159)/2
20   alpha ~ dunif(0, 2)
21   beta ~ dunif(-1, 0)
22   delta~ dunif(0,100)
23   d0~ dunif(10,20)
24   d1~ dunif(-0.30,-0.20)
25   d2~ dunif(-0.5,0.5)

```

```

26 d3~ dunif(-15,-8)
27 d4~ dunif(-0.5,0.5)
28 d5~ dunif(1,5)
29 d6~ dunif(-0.5,0.5)
30 }

```

B.3 PERÍODO 10/2013 A 06/2015 - PARA Y_1

Listing B.3 – Distribuição estável para os dados 10/2013-06/2015 (y_1)

```

1  model {
2    for (i in 1:n) {
3      dummy[i] <- 0
4      dummy[i] ~ dloglik(logLike[i])
5      z[i] <- (response[i] - mu[i]) / delta
6      mu[i] <- d0 + d1 * (ano[i] - 2013) + d2 * mes[i] + d3 * x1[i] +
7      + d4 * x2[i] + d5 * x3[i] + d6 * x4[i]
8      w[i] <- abs(z[i] / t[i])
9      a1[i] <- (sin(3.14159 * alpha * y[i] + b1)) / cos(3.14159 * y[i])
10     a2[i] <- (cos(3.14159 * y[i])) / cos(3.14159 * (alpha - 1) * y[i] + b1)
11     a3[i] <- pow(a2[i], theta2)
12     t[i] <- a1[i] * a3[i]
13     logLike[i] <- log(alpha) - log(abs(alpha - 1)) - log(delta) -
14     - pow(w[i], theta1) + theta1 * log(w[i]) - log(abs(z[i]))
15     y[i] ~ dunif(-0.5, 0.5)
16   }
17   theta1 <- alpha / (alpha - 1)
18   theta2 <- (alpha - 1) / alpha
19   b1 <- -(beta * min(alpha, 2 - alpha) * 3.14159) / 2
20   alpha ~ dunif(0, 2)
21   beta ~ dunif(-1, 0)
22   delta ~ dunif(0, 100)

```

```

23 d0~ dunif(-50,-30)
24 d1~ dunif(-20,0)
25 d2~ dunif(-0.5,0.5)
26 d3~ dunif(140,180)
27 d4~ dunif(-5,0)
28 d5~ dunif(-30,-10)
29 d6~ dunif(-10,0)
30 }

```

B.4 PERÍODO 10/2013 A 06/2015 - PARA Y_2

Listing B.4 – Distribuição estável para os dados 10/2013-06/2015 (y_2)

```

1  model {
2    for (i in 1:n) {
3      dummy[i] <- 0
4      dummy[i] ~ dloglik(logLike[i])
5      z[i] <- (response[i]-mu[i])/delta
6      mu[i] <- d0+d1*(ano[i]-2013)+d2*mes[i]+d3*x1[i]+
7      + d4*x2[i] +d5*x3[i]+d6*x4[i]
8      w[i] <- abs(z[i]/t[i])
9      a1[i] <- (sin(3.14159*alpha*y[i]+b1))/cos(3.14159*y[i])
10     a2[i] <- (cos(3.14159 *y[i]))/cos(3.14159*(alpha-1)*y[i]+b1)
11     a3[i] <- pow(a2[i],theta2)
12     t[i] <- a1[i]*a3[i]
13     logLike[i] <- log(alpha)-log(abs(alpha-1))-log(delta)-
14     - pow(w[i],theta1)+theta1*log(w[i])-log(abs(z[i]))
15     y[i] ~ dunif(-0.5,0.5)
16   }
17   theta1<-alpha/(alpha-1)
18   theta2<-(alpha-1)/alpha
19   b1<-(beta*min(alpha,2-alpha)*3.14159)/2

```

```
20 alpha ~ dunif(0, 2)
21 beta ~ dunif(-1, 0)
22 delta~ dunif(0,100)
23 d0~ dunif(0,10)
24 d1~ dunif(-1,1)
25 d2~ dunif(-0.5,0.5)
26 d3~ dunif(8,15)
27 d4~ dunif(0,1)
28 d5~ dunif(-20,0)
29 d6~ dunif(0,3)
30 }
```