284ª **REUNIÃO DO DEPARTAMENTO DE ESTATÍSTICA**

DATA: 09/07/18

☒ Ordinária ☐ Extraordinária Início: 14:30 Término: 15:00

Lista de Presença

1. Adrian Heringer Pizzinga
2. Ana Beatriz Monteiro Fonseca
3. Ana Maria Lima de Farias
4. Douglas Rodrigues Pinto
5. Eduardo Ferioli Gomes
6. Fábio Nogueira Demarqui
7. Hugo Henrique Kegler dos Santos
8. Jessica Quintanilha Kubrusly
9. Jony Arrais Pinto Junior
10. José Murilo Ferraz Saraiva
11. José Rodrigo de Moraes
12. Keila Mara Cassiano
13. Karina YurikoYaginuma
14. Licinio Esmeraldo da Silva
15. Luciane Ferreira Alcoforado
16. Ludmilla S. Viana Jacobson
17. Luis Guillermo Coca Velarde
18. Luz Amanda Melgar Santander
19. Márcia Marques de Carvalho
20. Marco Aurélio dos Santos Sanfins
21. Mariana Albi de Oliveira Souza
22. Maria Cristina Bessa Moreira
23. Moisés Lima de Menezes
24. Núbia Karla de Oliveira Almeida
25. Patrícia LusiéVelozo da Costa
26. Valentin Sisko
27. Wilson Calmon Almeida dos Santos

Ata da 284ª Reunião Ordinária do Departamento de Estatística

Aos nove dias do mês de julho de dois mil e dezoito (09/07/2018) foi realizada, na sala de reuniões do Instituto de Matemática e Estatística, a 284ª (ducentésima octogésima quarta) reunião ordinária do Departamento de Estatística (GET), que se iniciou às 14h30m horas sob a presidência do professor Jony Arrais Pinto Junior, chefe do GET, para deliberação sobre os seguintes itens de pauta: **1)**Aprovação da ata da reunião anterior; **2)** Aprovação do relatório do projeto de iniciação à pesquisa "Modelos Lineares Hierárquicos Bayesianos" da profa.Patrícia; **3)**Aprovação do relatório do projeto de iniciação à pesquisa "Analisando o impacto socioeconômico e ambiental da hanseníase através de modelos espaciais" da profa. Patrícia; **4)**Aprovação do quadro de horários 2018/2; **5))**Informes. Estavam presentes os seguintes professores: Ana Beatriz Monteiro Fonseca, Ana Maria Lima de Farias, Douglas Rodrigues Pinto, Eduardo Ferioli Gomes, Hugo Henrique Kegler dos Santos, Jessica Quintanilha Kubrusly, Jony Arrais Pinto Junior, José Rodrigo de Moraes, Karina YurikoYaginuma, Licínio Esmeraldo da Silva, Luciane Ferreira Alcoforado, Ludmilla S. Viana Jacobson, Luis Guillermo Coca Velarde, Luz Amanda Melgar Santander, Márcia Marques de Carvalho, Marco Aurélio dos Santos Sanfins, Mariana Albi de Oliveira Souza, Maria Cristina Bessa Moreira, Núbia Karla de Oliveira Almeida, Patrícia Lusié Velozo da Costa, Valentin Sisko e Wilson Calmon Almeida dos Santos. **Item 1)** O prof.Jony submeteu à votação a aprovação da ata da 283ª reunião ordinária, que foi aprovada por maioria. **Item 2)** O prof. Jony apresentou o parecer da comissão de pesquisa favorável à aprovação do relatório projeto de iniciação à pesquisa da profa. Patrícia intitulado "Modelos Lineares Hierárquicos Bayesianos". O prof. Jony colocou em votação a aprovação do relatório do projeto, que foi aprovado por unanimidade. **Item 3)** O prof. Jony apresentou o parecer da comissão de pesquisa favorável à aprovação do relatório projeto de iniciação à pesquisa da profa. Patrícia intitulado "Analisando o impacto socioeconômico e ambiental da hanseníase através de modelos espaciais". O prof. Jony colocou em votação a aprovação do relatório do projeto, que foi aprovado por unanimidade. **Item 4)** O prof. Jony apresentou a proposta de quadro de horários para 2018/2 após a apresentação das preferências dos professores enviadas à chefia. O prof. Jony informou que neste período o GET está contando com 3 professores substitutos que já se encontram no quadro (Rebecca de Oliveira Souza, Rodrigo Ferrari Lucas Lassance e Yasmin Ferreira Cavaliere) e com mais dois que serão contratados em breve (Victor Eduardo Leite de Almeida Duca - aprovado pelo concurso do edital 10/2018 e Deyvid Toledo Santiago de Almeida do edital 170/2018). Por este motivo, o prof. Jony frisou que eventuais necessidades de alterações no quadro deveriam ser tratadas com a próxima chefia de departamento. Em seguida, o prof. Jony colocou em votação o quadro de horários que foi aprovado por unanimidade. Nada mais havendo a tratar e ninguém mais desejando fazer uso da palavra, foi encerrada a reunião às 15h, cuja ata vai datada e assinada por mim, Jony Arrais Pinto Junior, chefe do Departamento de Estatística. Niterói, 09 de julho de 2018.

Disciplinas			Professor	Dias				
Código	Nome	Turma		Seg	Ter	Qua	Qui	Sex
GET00040	Estatística V	A1	Substituto 1	16-18		16-18		
GET00041	Bioestatística	A1	Luciane		08-11			
GET00053	EstatísticaBásicaAplicada as CiênciasHumanas	A1	Hugo		18-20		18-20	

GET00059	Bioestatística I	A1	Rodrigo	16-18	16-18			
GET00116	Fundamentos de EstatísticaAplicada	A1	Marco	16-18		16-18		
GET00116	Fundamentos de EstatísticaAplicada	B1	Eduardo	18-20		18-20		
GET00116	Fundamentos de EstatísticaAplicada	C1	Moisés	18-20		18-20		
GET00117	MétodosEstatísticosAplicados à Economia I	A1	Substituto 2			11-13		11-13
GET00117	MétodosEstatísticosAplicados à Economia I	B1	Substituto 2			20-22		20-22
GET00118	MétodosEstatísticosAplicados à Economia II	A1	Mariana			11-13		11-13
GET00118	MétodosEstatísticosAplicados à Economia II	B1	Rebecca			20-22		20-22
GET00119	EstatísticaBásicaparaEngenharia II	A1	Rodrigo	09-11		09-11		
GET00121	Introdução à Probabilidade e à Estatística	A1	Ana Beatriz		11-13		11-13	
GET00121	Introdução à Probabilidade e à Estatística	B1	Ana Maria	18-20		18-20		
GET00196	Instrumentação no Ensino de CPE	A1	Jony		16-18		16-18	
GET00122	Probabilidade e Estatística	A1	Luz Amanda	14-16		14-16		
GET00122	Probabilidade e Estatística	B1	Valentin	20-22		20-22		
GET00158	EstatísticaBásicaAplicada as Ciências da Vida	A1	Guillermo		11-13		11-13	
GET00169	EstatísticaBásicaparaCiências Humanas I	A1	Rodrigo	07-09		07-09		
GET00170*	EstatísticaGeral*	A1	Nubia	16-18		16-18		
GET00170	EstatísticaGeral*	B1	Ana Beatriz	14-16		14-16		
GET00170	EstatísticaGeral*	C1	José Murilo		16-18		16-18	
GET00170	EstatísticaGeral*	D1	Substituto 1	18-20		18-20		
GET00170	EstatísticaGeral*	E1	Substituto 1	07-09		07-09		
GET00176	EstatísticaAplicadaàsCiências da Vida	A1	Nubia		09-11		09-11	
GET00177	EstatísticaBásicaparaEngenharia	A1	Jéssica	14-16		14-16		
GET00177	EstatísticaBásicaparaEngenharia	B1	Luz Amanda	16-18		16-18		
GET00177	EstatísticaBásicaparaEngenharia	C1	Wilson		09-11		09-11	
GET00177	EstatísticaBásicaparaEngenharia	D1	Wilson		11-13		11-13	
GET00177	EstatísticaBásicaparaEngenharia	E1	Keila		14-16		14-16	
GET00177	EstatísticaBásicaparaEngenh	F1	Keila		16-18		16-18	

	aria							
GET00177	EstatísticaBásicaparaEngenharia	G1	Substituto 2			09-11		09-11
GET00178	EstatísticaAplicadaparaEngenharia	A1	Rebecca			11-13		11-13
GET00180	EstatísticaBásicaparaEngenharia I	A1	José Murilo		11-13		11-13	
GET00181	ModelosProbabilísticos	A1	José Murilo		09-11		09-11	
GET00100	Estatística I	A1/B1	Yasmin	09-13		09-13		
GET00125	Amostragem I	A1	Ludmilla		11-13		11-13	
GET00126	AnáliseMultivariada I	A1	Valentin		09-11		09-11	11-13
GET00127	Análise de Sériestemporais I	A1	Moisés	07-09		07-09		07-09
GET00128	EstatísticaAplicada	A1	Ludmilla		09-11		09-11	
GET00130	MétodosCoputacionaisparaEstatística II	A1	Jony		11-13		11-13	
GET00133	Estatísticas e Indicadores	A1	Marcia		09-11		09-11	
GET00135	Inferência	A1	Jéssica	11-13		11-13		11-13
GET00188	Fundamentos de MatemáticaparaEstatística	A1	Douglas	11-13		11-13		11-13
GET00188	Fundamentos de MatemáticaparaEstatística	B1	Hugo	11-13		11-13		11-13
GET00189	Probabilidade I	A1	Ana Maria	11-13		11-13		11-13
GET00190	Probabilidade II	A1	Karina	11-13		11-13		11-13
GET00136	InferênciaBayesiana I	A1	Maria Cristina		07-09		07-09	
GET00137	Metodologia da PesquisaCientífica	A1	Eduardo	07-09		07-09		
GET00138	ModelosLineares I	A1	José Rodrigo	09-11		09-11		09-11
GET00139	ProgramaçãoEstatística	A1	Maria Cristina		14-16		14-16	09-11
GET00182	Estatística II	A1	Mariana	09-11		09-11		09-11
GET00197	IntroduçãoaoPasseioAleatório	A1	Douglas	14-16		14-16		
GET00155	EstatísticaNãoParamétrica	A1	Marco	09-11		09-11		
GET00187	AnáliseMultivariada II	A1	Luciane		11-13		11-13	
GET00162	ModelosLineares II	A1	José Rodrigo	07-09		07-09		
GET00165	Simulação de EventosDiscretos	A1	Karina		07-09		07-09	
GET00147	Análise de Sobrevivência e Confiabilidade	A1	Guillermo		09-11		09-11	

JONY ARRAIS PINTO JUNIOR
Chefe Depto de Estatística
SIAPE 2722748

Universidade Federal Fluminense
Instituto de Matemática e Estatística
Departamento de Estatística

Assunto: Avaliação do Relatório Final associado ao Projeto de Iniciação à Pesquisa da ProfessoraPatrícia LusiéVelozo da Costa (SIAPE 1805333).

A Comissão de Pesquisa, no uso de sua atribuição, avaliou o relatório final do projeto de iniciação à pesquisa entitulado "Analisando o impacto socioeconômico e ambiental da hanseníase através de modelos espaciais", submetido pela docentes Patrícia LusiéVelozo da Costa (SIAPE 1805333), bem como o relato do aluno envolvido Matheus Camelo dos Santos Araújo (matrícula UFF 114054024). Considerando a Instrução de Serviço GET – e entendendo que o projeto cumpriu seu propósito, a Comissão de Pesquisa do GET considera validada a sua execução e indica o aproveitamento do trabalho realizado como atividade complementar para os alunos envolvidos.

Niterói, 03 de julho de 2018

Ana Beatriz Monteiro Fonseca

Eduardo Ferioli Gomes

Luciane Alcoforado

Wilson Calmon

Universidade Federal Fluminense
Instituto de Matemática e Estatística
Departamento de Estatística

Analisando o impacto socioeconômico e ambiental da hanseníase através de modelos espaciais

Relatório Final

Identificação do Projeto: Projeto de Iniciação à Pesquisa
Docente responsável: Patrícia Lusié Velozo da Costa (SIAPE: 1805333)
Aluno vinculado: Matheus Camelo dos Santos Araujo (matrícula: 114054024)
(aluno do curso de Estatística)
carga horária semanal: 20horas

Resumo do plano inicial

A docente responsável estava co-orientando o aluno Paulo Henrique Leal de Sousa na sua dissertação de mestrado profissional em Epidemiologia. O aluno em questão não tem formação estatística e essa foi a motivação da docente criar esse projeto e incluir o aluno Matheus para auxiliar o Paulo. Desse auxílio, surgiram propostas de modelagem diferentes da usada no trabalho do aluno Paulo. Dessa forma, o objetivo deste trabalho consistiu em recorrer a modelagem de dados espaciais a fim de caracterizar a influência de fatores socioeconômicos e ambientais na ocorrência da hanseníase no estado do Maranhão, em escala municipal. Pretendeu-se utilizar um modelo condicional autoregressivo (CAR) para acomodar a dependência espacial e inferir sobre os parâmetros desconhecidos desses modelos através da abordagem Bayesiana. Pretendeu-se trabalhar com dados simulados

a fim de analisar a sensibilidade da distribuição a priori atribuída e verificar se era possível recuperar os valores verdadeiros dos parâmetros desconhecidos. Posteriormente, pretendeu-se aplicar a modelagem proposta nos dados de hanseníase no Maranhão.

Resultados previamente esperados

Quanto a formação do aluno, a docente espera ter contribuído na formação acadêmica do aluno uma vez que esse precisou aprender sobre modelagem espacial em dados de área, reforçou os conhecimentos sobre inferência bayesiana e sobre modelagem estatística. O aluno se interessou posteriormente em cursar a disciplina de Estatística Espacial para aprender mais sobre essa área. Além disso, o aluno precisou entrar em contato com profissionais de outras áreas, conforme mencionado anteriormente.

Quanto ao trabalho realizado, esperava-se que os dados estivessem variando espacialmente e que o modelo implementado ajustasse bem os dados mostrando ser relevante incluir uma dependência espacial.

Resumo do projeto executado e resultados efetivamente obtidos

O projeto foi previsto para ser desenvolvido entre setembro de 2017 a março de 2018, totalizando 7 meses. Porém, houveram alguns atrasos na obtenção dos dados e o aluno precisou de um pouco mais de tempo para o estudo de modelos espaciais. O aluno produziu um texto contendo a revisão bibliográfica necessária nesse trabalho, descrição e resultados do modelo aplicado nos dados reais e simulados. Esse texto está sendo enviado em conjunto com o Relatório Final. O aluno foi avaliado ao longo do período de execução, citado anteriomente, através de apresentações realizadas sobre o assunto estudado e de implementações de modelos pertinentes ao assunto.

Propos-se um modelo condicional autoregressivo para ajustar a taxa de pessoas menores de 15 anos infectadas pela Hanseníase no Maranhão. Taxas altas nesse grupo de pessoas indica condições ambientais e socioeconômicas alarmantes. Antes de aplicar

nos dados reais, aplicou-se a modelagem proposta em dados artificiais. As estimativas intervalares dos parâmetros contiveram os valores verdadeiros desses e a análise de sensibilidade mostrou que o modelo não é tão sensível a escolha dos hiperparâmetros da distribuição a priori. Aplicou-se o modelo então nos dados reais e comparou-se o modelo espacial com um modelo sem dependência espacial. Os intervalos de credibilidade tiveram um comprimento razoavelmente grande, o que indica uma incerteza alta. Assim como os resíduos não tiveram um comportamento tão adequado. Sendo assim, atualmente estamos trabalhando com extensões desse trabalho e apresentaremos esse estudo em Projeto Final. Maiores detalhes, encontram-se no texto produzido pelo aluno e anexado a esse relatório.

Niterói, 25 de junho de 2018.

Patrícia Lusié Velozo da Costa

Relatório de Atividades do Projeto de Iniciação à Pesquisa: *Analisando o impacto socioeconômico e ambiental da hanseníase através de modelos espaciais.*
Aluno: Matheus Camelo dos Santos Araujo
Docente responsável: Patrícia Lusié Velozo da Costa
Instituição: Universidade Federal Fluminense - UFF

Inicialmente o objetivo foi auxiliar as análises estatísticas do trabalho de dissertação do, agora então, mestre Paulo Henrique Leal de Souza, co-orientado pela docente responsável e orientadora desse projeto, Prof. Dra. Patrícia Lusié, no qual fiquei responsável em ajudá-los principalmente na análise exploratória dos dados gerando mapas das taxas de hanseníase no Maranhão.

No início do segundo semestre de 2017 surgiu o interesse em tornar esse auxílio no trabalho do Paulo em Projeto de Iniciação à Pesquisa. Com a aprovação e início do Projeto em setembro do mesmo ano, este me ofereceu um ganho acadêmico na revisão bibliográfica e metodológica em Estatística Espacial, Inferência Bayesiana e criação de algoritmos via métodos de MCMC. E também, é claro, um conhecimento maior em hanseníase, uma das doenças negligenciadas mais estudadas em Epidemiologia atualmente no Brasil, principalmente na região norte do país.

Após a revisão metodológica, o objetivo do trabalho foi avaliar efeitos socioeconômicos e ambiental das taxas de hanseníase nos municípios do Maranhão através de modelos espaciais. Foi utilizado um modelo autorregressivo condicional – CAR, e a principal referência na modelagem espacial foi o livro *Statistics for Spatial Data*, do Noel Cressie. Os dados foram obtidos a partir da dissertação do Paulo.

A partir da modelagem espacial e adequação dos dados ao modelo, verificou-se que a modelagem poderia melhorar caso considerássemos os dados inflacionados em zero. Por questões de tempo de duração do projeto, não foi possível avaliar e ajustar outros modelos que considerassem os dados inflacionados de zero.

Por conta de tudo o que foi dito, é importante destacar o ganho acadêmico e conceitual em metodologias, principalmente em Estatística Espacial, na qual sempre tive interesse em me especializar e aperfeiçoar com trabalhos futuros. Inclusive, esse projeto tornou-se motivação para o meu trabalho de conclusão de curso, possibilitando assim um estudo mais aprofundado na modelagem da hanseníase via modelos espaciais.

Niterói, 25 de junho de 2018.

Matheus Camelo dos S. Araujo

Matheus Camelo dos Santos Araujo

Matheus Camelo dos Santos Araujo

Analisando o impacto socioeconômico e ambiental da hanseníase através de modelos espaciais

Niterói - RJ, Brasil

Junho de 2018

Universidade Federal Fluminense

Matheus Camelo dos Santos Araujo

Analisando o impacto socioeconômico e ambiental da hanseníase através de modelos espaciais

Projeto de Iniciação à Pesquisa

Relatório Final do Projeto de Iniciação à Pesquisa submetido ao Departamento de Estatística da Universidade Federal Fluminense.

Docente responsável pelo Projeto: Patrícia Lusié Velozo da Costa

Niterói - RJ, Brasil

Junho de 2018

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

CAR Modelo Autorregressivo Condicional

MCMC Monte Carlo via cadeias de Markov

SAR Modelo Autorregressivo Simultâneo

SIG Sistema de Informação Geográfica

IDHM Índice de Desenvolvimento Humano Municipal

1 Introdução

Popularmente conhecida como lepra, a hanseníase é uma doença crônica e infecciosa que afeta a pele e troncos nervosos periféricos podendo causar úlceras de pernas e pés, caroços no corpo, febre, edemas e dor nas juntas, entupimento, sangramento, ferida e ressecamento do nariz e dos olhos.

Sua forma de contágio é através do contato com pessoas infectadas com o bacilo *Mycobacterium leprae*, que não estejam sendo tratada. Esse bacilo tem a capacidade de infectar um grande número de indivíduos, mas poucos adoecem. Acredita-se também que fatores como condições de vida e nutrição, insalubridade do ambiente e questões ambientais possam intensificar a propagação da doença.

A hanseníase apresenta um longo período médio de incubação, de 2 a 7 anos, e o diagnóstico dessa doença é essencialmente clínico. E, por isso, espera-se que hajam poucos indivíduos menores de 15 anos com a doença diagnosticada. Sendo assim, um número grande de menores doentes pode ser um indicador de problema grave em uma região.

Há relatos de ocorrências da doença em 600 a.C na Ásia e na África, consideradas o berço da hanseníase. Sem recursos médicos nessa época, a doença se acentuava com graves deformações físicas nas pessoas contaminadas, levando o paciente a marginalização e estigmatização social. Devido aos avanços da medicina, introduziu-se o tratamento de poliquimioterapia tornando a doença curável. Além disso, acredita-se que a redução da pobreza e o crescimento econômico contribuíram para a grande redução no número de pessoas com hanseníase em todo o mundo.

Há ainda algumas regiões consideradas hiperendêmicas. Segundo Who (2012) [1], três países são responsáveis por 83% de todos os casos detectados no mundo: Índia (58%), Brasil (16%) e Indonésia (9%). Sendo assim, o Brasil apresenta a maior prevalência na América Latina. Entre as regiões brasileiras, o Norte, o Nordeste e o Centro-Oeste apresentam as maiores taxas de detecção. Dentres os estados, o Maranhão apresenta a

maior prevalência, a maior taxa de detecção geral e a maior taxa de detecção em menores de 15 anos, considerado como hiperendêmico para os padrões do Ministério da Saúde.

Partindo do pressuposto que a região do Maranhão e seus municípios apresentam altas e diferentes taxas de hanseníase, é possível analisar espacialmente sua influência com o auxílio de dados localmente observados. Essas informações são acessíveis através do Sistema de Informação Geográfica (SIG), que vem se tornando uma grande ferramenta em análises de dados sobre saúde e meio ambiente.

Com a grande redução no número de infectados e a grande concentração de pessoas infectadas sendo de baixa renda, a doença tornou-se negligenciada.

Sendo assim, esse trabalho visa modelar estatísticamente a hanseníase no Maranhão em 2010, descrevendo o comportamento probabilístico dessa doença em indivíduos menores de 15 anos. Para isso, recorreu-se a modelos espaciais. Os parâmetros desconhecidos foram estimados segundo o enfoque bayesiano através dos métodos de Monte Carlo via cadeias de Markov (MCMC).

O presente trabalho encontra-se dividido da seguinte maneira: o Capítulo 2 contém o objetivo desse trabalho; o Capítulo 3 apresenta um resumo bibliográfico de Estatística Espacial, Inferência bayesiana e dos métodos de MCMC; posteriormente, o Capítulo 4 mostra a análise dos resultados encontrados, e por fim, o Capítulo 5 encerra o trabalho apresentando as conclusões sobre o estudo.

2 Objetivos

O objetivo deste trabalho é recorrer a modelagem de dados espaciais a fim de caracterizar a influência de fatores socioeconômicos e ambientais na ocorrência da hanseníase no estado do Maranhão em indivíduos menores de 15 anos, em escala municipal. A inferência sobre os parâmetros será realizada sob o enfoque bayesiano.

3 Revisão bibliográfica

Para um melhor entendimento do estudo, fez-se necessária uma breve revisão bibliográfica em tópicos importantes relacionados a modelagem estatística. Dessa forma, a seguir serão apresentadas seções de Estatística Espacial, Inferência Bayesiana e dos métodos de MCMC.

3.1 Estatística Espacial

Fenômenos observados ao longo do espaço são considerados dados espaciais. A estatística espacial é a área da estatística que busca descrever ou explicar esses fenômenos relacionando-os com o espaço e tem aplicação em diversas áreas tais como economia, epidemiologia, demografia, entre outras.

De acordo com Cressie (1993) [2], dados espaciais podem ser classificados em três grupos: dados de superfícies contínuas (geoestatísticos), padrão de pontos e dados de área.

Dados geoestatísticos são obtidos quando a variável de interesse ocorre de forma contínua no espaço. Apesar de ocorrer de forma contínua no espaço, observa-se apenas um conjunto finito de localizações. O volume pluviométrico em certa região é um exemplo de dados dessa natureza.

Caso o interesse seja modelar a localização (desconhecida) de um evento de interesse (conhecido), então os dados são considerados como padrão de pontos. O estudo de acidentes de trânsito em determinada cidade é um exemplo desse grupo.

Por fim e não menos importante, os dados de área são aqueles agregados em unidades de análises. Dessa forma, é possível avaliar a influência da vizinhança de acordo com a proximidade e analisar seus impactos. Por exemplo: o número de homicídios nos bairros da cidade do Rio de Janeiro. Cada bairro contém um número que representa a quantidade de homicídios que ocorreram em diferentes ruas daquele mesmo bairro.

Este trabalho avalia a hanseníase nos municípios do estado do Maranhão registrando o município de endereço do indivíduo infectado e agrupando esses indivíduos por município. Sendo assim, serão abordados dados de área, comumente utilizados na abordagem médica. Por isso, a seguir serão apresentados os conceitos básicos dessa área.

3.1.1 Dados de Área

No contexto de Estatística Espacial, os dados de área são observações obtidas sob uma região de interesse que pode ser dividida em subregiões regulares (de mesmo comprimento e mesma área) ou irregulares (bairros, cidades, setores censitários, etc). São inúmeros os exemplos para dados dessa natureza tais como: casos de dengue nos bairros do Rio de Janeiro e vendas do produto A nos municípios de São Paulo. Usualmente, esses dados correspondem a contagens, taxas, médias, entre outros.

Os principais objetivos de estudo em dados de área são a detecção e explicação dos padrões espaciais ou tendências encontradas no fenômeno de interesse. Dessa forma, torna-se válido investigar se há uma tendência das observações de regiões mais próximas serem mais semelhantes do que observações mais distantes.

Quando o interesse na modelagem espacial é, por exemplo, relacionar as respostas de uma variável com seus vizinhos, duas especificações de modelos são mais comuns, são elas: o Modelo Autorregressivo Simultâneo (SAR) e o Modelo Autorregressivo Condicional (CAR). Cressie (1993) [2] mostrou que o modelo SAR é um caso específico do modelo CAR e que este último é mais comumente usado em análise espacial de dados de contagem, devido a facilidade computacional.

Comparando algumas propriedades de ambos os modelos e em termos de estimação e interpretação, o CAR é preferível ao SAR (Schmidt et al. (2003) [3]). Uma delas é bastante interessante, a propriedade de que a especificação do CAR fornece diretamente as distribuições condicionais completas a posteriori dos parâmetros do modelo, fator importante para o uso do amostrador de Gibbs em MCMC, que será visto na seção 3.3.1.

Basicamente a ideia do modelo CAR é que a probabilidade do evento de interesse assumir um valor em um local depende do valor desse evento assumido na vizinhança. Assim, supondo Z_i a variável de interesse na região i, o modelo pode ser definido por

$$Z_i = \mu_i + \rho \sum_{j \in S_{-i}} b_{ij}(Z_j - \mu_j) + e_i, \quad i = 1, \ldots, n, \tag{3.1}$$

onde $S_{-i} = 1, \ldots, i-1, i+1, \ldots, n$ é o conjunto de índices que representam todas as regiões excluindo a i-ésima localização, n é o número total de regiões, μ_i é a componente do valor esperado de Z_i que não depende de forma direta dos vizinhos e pode conter por exemplo variáveis explicativas específicas da i-ésima região, ρ é o parâmetro da autocorrelação espacial que determina a dependência da vizinhança, b_{ij} é o efeito do vizinho j na região i e também pode ser visto como uma ponderação e e_i é um efeito aleatório independente. Suponha que esses efeitos sejam independentes e identicamente distribuídos e que possuam a seguinte distribuição normal

$$e_i \overset{iid}{\sim} N(0, V_i). \tag{3.2}$$

Mesmo supondo que e_i tenha uma distribuição normal, tem-se problemas para obter a distribuição conjunta de $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n)'$. Sendo assim, para se obter uma distribuição conjunta válida é necessário cautela para a definição dos efeitos b_{ij} e do parâmetro de autocorrelação espacial ρ. Os efeitos b_{ij} costumam ser definidos através da matriz de vizinhanças $\boldsymbol{W}$ que pode ser representada de diversas formas. Essa matriz indica se as regiões i e j são vizinhas. Para definir isso, pode-se considerar vizinhas se essas regiões dividirem fronteiras ou se elas estiverem no máximo a uma certa distância, por exemplo. Seja W_{ij} o elemento da i-ésima linha e j-ésima coluna da matriz $\boldsymbol{W}$ sendo $W_{ij} = 1$, se as áreas i e j são vizinhas e $W_{ij} = 0$, caso contrário. Seja $W_{i+} = \sum_{j=1}^{n} W_{ij}$ o número de vizinhos da i-ésima região. Suponha então que $b_{ij} = \frac{W_{ij}}{W_{i+}}$. Além disso, suponha que $V_i = \frac{V}{W_{i+}}$, sendo V comum a todas as regiões.

No modelo CAR, a matriz de covariância é dada da seguinte forma

$$\boldsymbol{\Sigma} = VAR(\boldsymbol{Z}) = (\boldsymbol{I} - \rho \boldsymbol{B})^{-1} \boldsymbol{V} \tag{3.3}$$

onde $\boldsymbol{I}$ é a matriz identidade de ordem n, $\boldsymbol{B}$ é a matriz formada pelos elementos b_{ij} e $\boldsymbol{V}$ é uma matriz diagonal formada pelos elementos V_i. Quando $\rho = 0$, tem-se independência e que $Z_i \sim N(0, V/\boldsymbol{W_{i+}})$. Quando $\rho = 1$, é dito ter um modelo autoregressivo íntrinseco e tem-se que o inverso da expressão acima é singular, ou seja, a expressão acima não existe e a distribuição conjunta de $\boldsymbol{Z}$ é imprópria. Pode-se mostrar que se $\rho \in (-1, 1)$, então existe a distribuição conjunta de $\boldsymbol{Z}$ e essa possui a seguinte forma

$$\boldsymbol{Z} \sim N\left(\boldsymbol{\mu}, (\boldsymbol{I} - \rho \boldsymbol{B})^{-1} \boldsymbol{V}\right) \tag{3.4}$$

sendo $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$.

Maiores detalhes sobre esses modelos podem ser vistos em Cressie(1993) [2] e Banerjee e outros (2003) [4].

3.2 Inferência Bayesiana

Inferência estatística consiste em afirmar sobre característica de uma população com base em um subconjunto dessa população chamado de amostra. Sendo assim, considere que $\boldsymbol{\theta}$ seja um vetor de parâmetros populacionais desconhecidos de uma população de tamanho N. A quantidade $\boldsymbol{\theta}$ assume valores no espaço paramétrico e será denotado por $\boldsymbol{\Theta}$.

Seja Z_i uma variável aleatória com i sendo o índice de unidade amostral da população e que pode representar, por exemplo, um indíviduo, um instante de tempo ou uma localidade. Suponha que é obtida uma amostra dessa população de tamanho n e que haja o interesse em inferir sobre a média e/ou a variância da mesma, representadas por μ e σ^2, respectivamente, e então tem-se $\boldsymbol{\theta} = (\mu, \sigma^2)$.

Considere $p(Z_1, \ldots, Z_N|\boldsymbol{\theta})$ a função de distribuição ou de densidade da variável resposta dado um conjunto de parâmetros $\boldsymbol{\theta}$. Após realizar uma amostragem sobre a população, é feita a inferência sobre os parâmetros populacionais.

Em inferência bayesina, diferentemente da clássica, leva-se em consideração um conhecimento prévio sobre os parâmetros, conhecido como *distribuição a priori*. Denota-se essa distribuição por $h(\boldsymbol{\theta})$. Já a informação contida nos dados amostrados é denominada por função de verossimilhança e denotada por $p(\boldsymbol{z}|\boldsymbol{\theta})$, sendo $\boldsymbol{z} = (z_1, \ldots, z_N)$ o valor amostrado da variável aleatória $\boldsymbol{Z} = (Z_1, \ldots, Z_N)$.

Dessa forma, a inferência sobre $\boldsymbol{\theta}$ é dada através da *distribuição a posteriori* $p(\boldsymbol{\theta}|\boldsymbol{z})$, que pode ser obtida a partir do Teorema de Bayes, combinando a função de verossimilhança com a distribuição a priori e com a distribuição marginal dos dados, obtendo a seguinte forma

$$p(\boldsymbol{\theta}|\boldsymbol{z}) = \frac{p(\boldsymbol{z}|\boldsymbol{\theta})h(\boldsymbol{\theta})}{p(\boldsymbol{z})}. \tag{3.5}$$

A distribuição marginal da variável de interesse pode ser obtida da seguinte forma

$$p(\boldsymbol{z}) = \int \ldots \int_{\boldsymbol{\Theta}} p(\boldsymbol{z}|\boldsymbol{\theta})h(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{3.6}$$

Note que a distribuição marginal $p(\boldsymbol{z})$ não varia com o vetor paramétrico $\boldsymbol{\theta}$. Sendo assim, a distribuição a posteriori do vetor paramétrico é proporcional ao produto da função de verossimilhança e da *distribuição a priori*. E, por definição de função de densidade, integrando a distribuição a posteriori com respeito a $\boldsymbol{\Theta}$ essa integral tem que dar 1. Logo, não faz-se necessário calcular a distribuição marginal $p(\boldsymbol{z})$ para obter a distribuição a posteriori. E, portanto, a distribuição a posteriori pode ser reescrita da seguinte forma

$$p(\boldsymbol{\theta}|\boldsymbol{z}) \quad = \quad kp(\boldsymbol{z}|\boldsymbol{\theta})h(\boldsymbol{\theta}), \tag{3.7}$$

sendo $k^{-1} = \int_{\boldsymbol{\Theta}} p(\boldsymbol{z}|\boldsymbol{\theta})h(\boldsymbol{\theta})d\boldsymbol{\theta}$.

Muitas vezes a distribuição a posteriori não possui forma analítica conhecida. Portanto, para inferir sobre o vetor paramétrico desconhecido $\boldsymbol{\theta}$ pode-se obter amostras da distribuição a posteriori recorrendo aos métodos de MCMC. Na seção a seguir, serão apresentados dois desses métodos: o amostrador de Gibbs e o algoritmo de Metropolis-Hastings.

3.3 Métodos de Monte Carlo via Cadeias de Markov

Os métodos de MCMC servem para simular amostras de uma distribuição de interesse $p(\cdot)$ quando essa distribuição possui forma analítica desconhecida ou é custosa de se amostrar diretamente. Para essa amotragem, é necessário que as cadeias de Markov sejam homogêneas, irredutíveis e aperíodicas. Diz-se que uma cadeia de Markov é homogênea se a probabilidade de transição for estacionária, ou seja, se esta probabilidade não depender da iteração. Uma cadeia é irredutível se para um conjunto finito de iterações e com probabilidade positiva, ela se move de um ponto a outro qualquer. E será aperíodica se ela for irredutível e se nenhum de seus estados seja visitado apos n passos com probabilidade igual a um.

Diante dos vários métodos de simulação de amostras, este trabalho irá se concentrar em dois dos principais métodos: o amostrador de Gibbs e o Algoritmo de Metropolis-Hastings. Para mais detalhes consultar Gamerman e Lopes (2006) [5].

3.3.1 Amostrador de Gibbs

O algoritmo amostrador de Gibbs foi proposto por Geman e Geman (1984) [6] e introduzido a comunidade estatística por Gelfand e Smith (1990) [7]. Em inferência

bayesiana, esse algoritmo consiste basicamente em amostrar a partir das distribuições condicionais completas a posteriori, $p(\theta_i \mid \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p, \boldsymbol{z})$, sendo $\boldsymbol{z}$ os valores observados e θ_i o i-ésimo parâmetro desconhecido.

Os passos desse algoritmo, baseado em sucessivas gerações das distribuições condicionais completas a posteriori, pode ser descrito como:

1. Inicialize o contador em $j = 0$ e determine valores arbitrários para

$$\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_p^{(0)})$$

2. Modifique o contador de j para $j + 1$;

3. Obtenha um novo valor para $\boldsymbol{\theta}^{(j)}$ a partir de $\boldsymbol{\theta}^{(j-1)}$ sequencialmente da forma

$$\begin{aligned}
\theta_1^{(j)} &\backsim p(\theta_1 \mid \theta_2^{(j-1)}, \ldots, \theta_p^{(j-1)}, \boldsymbol{z}) \\
\theta_2^{(j)} &\backsim p(\theta_2 \mid \theta_1^{(j-1)}, \theta_3^{(j-1)}, \ldots, \theta_p^{(j-1)}, \boldsymbol{z}) \\
&\vdots \\
\theta_p^{(j)} &\backsim p(\theta_p \mid \theta_1^{(j-1)}, \theta_2^{(j-1)}, \ldots, \theta_{p-1}^{(j-1)}, \boldsymbol{z})
\end{aligned}$$

4. Repita os passos (2) e (3) até que a cadeia convirja.

A convergência das cadeias de Markov é esperada após um número de iterações suficientemente grande e após o período de aquecimento (*burn in*), que são as iterações necessárias até que a cadeia comece a convergir. Importante salientar que os parâmetros amostrados costumam ser altamente autocorrelacionados, característica das cadeias de Markov, desta forma, utiliza-se um espaçamento de ordem k em que seleciona-se uma amostra a cada k interações até que seja corrigida a autocorrelação da cadeia.

3.3.2 Algoritmo de Metropolis-Hastings

O Algoritmo de Metropolis-Hastings foi proposto por Metropolis e outros (1953) [8] e Hastings (1970) [9]. Ele é utilizado quando as distribuições condicionais completas não possuem forma analítica conhecida. Portanto, sem conhecer o núcleo ou a classe de distribuições de $p(\cdot)$, não é posssível amostrar diretamente da distribuição de interesse. Com isso, utiliza-se uma distribuição auxiliar $q(\cdot)$, denominada como distribuição proposta. O algoritmo baseia-se em gerar um valor proposto de $q(\cdot)$ e aceitá-lo na cadeia

a partir de uma condição probabilística de $p(\cdot)$ e $q(\cdot)$. Sob o ponto de vista bayesiano, o método pode ser explicado pelos seguintes passos:

1. Inicialize o contador de iterações em $j = 0$ e determine valores arbitrários para $\boldsymbol{\theta}^{(0)}$;

2. Modifique o contador de j para $j + 1$;

3. Gere um valor proposto $\boldsymbol{\varphi}$ usando uma distribuição conhecida que pode depender do valor amostrado na iteração anterior e essa distribuição será denotada por $q(\boldsymbol{\varphi} \mid \boldsymbol{\theta}^{(j-1)})$. Aceite o ponto gerado com probabilidade

$$\alpha = min\left\{1, \frac{p(\boldsymbol{\varphi}|\boldsymbol{z})}{q(\boldsymbol{\varphi} \mid \boldsymbol{\theta^{(j-1)}})} \frac{q(\boldsymbol{\theta}^{(j-1)} \mid \boldsymbol{\varphi})}{p(\boldsymbol{\theta^{(j-1)}}|\boldsymbol{z})}\right\}. \tag{3.8}$$

   Se o valor for aceito, $\boldsymbol{\theta}^{(j)} = \boldsymbol{\varphi}$, caso contrário $\boldsymbol{\theta^{(j)}} = \boldsymbol{\theta^{(j-1)}}$;

4. Repita os passos (2) e (3) até que a cadeia convirja.

Os critérios de convergência vistos no amostrador de Gibbs também valem para o algoritmo de Metropolis-Hastings, tais como: período de aquecimento (*burn in*) e espaçamento de ordem k.

Uma vez atingida a convergência, torna-se bastante trivial fazer inferência a partir das distribuições *a posteriori* dos parâmetros de interesse.

4 Modelando a Hanseníase

Seja Z_i^* a proporção / taxa de doentes diagnosticados com Hanseníase na região i. Considere que $\boldsymbol{Z} = (Z_1, \ldots, Z_n)'$, sendo $Z_i = \log(Z_i^* + 0,1)$, segue um modelo condicional autoregressivo conforme descrito na Subseção 3.1.1 e dado pela seguinte equação

$$\boldsymbol{Z} \sim N\left(\boldsymbol{X\beta}, (\boldsymbol{I} - \rho\boldsymbol{B})^{-1}\boldsymbol{V}\right), \tag{4.1}$$

sendo $\boldsymbol{X}$ chamada de matriz desenho contendo n linhas nas quais cada linha contém K variáveis relacionadas a i-ésima região. Essa matriz pode conter uma coluna de uns para permitir intercepto na modelagem, variáveis explicativas também chamadas de covariáveis, sazonalidade, entre outros. Além disso, considere que $\boldsymbol{\beta}$ seja um vetor coluna representando os efeitos dessas variáveis na variável resposta, $\boldsymbol{I}$ uma matriz identidade de ordem n, ρ representa o efeito espacial, $\boldsymbol{B}$ sendo uma matriz de ordem $n \times n$ formada pelos elementos $b_{ij} = \frac{W_{ij}}{W_{i+}}$, nos quais $W_{ij} = 1$, se os municípios i e j dividem a mesma fronteira, e $W_{ij} = 0$, caso contrário, e $W_{i+} = \sum_{j=1}^{n} W_{ij}$ sendo o total de regiões que dividem fronteira com a região i. E $\boldsymbol{V}$ uma matriz diagonal de ordem n formada pelos elementos $V_i = \frac{1}{\tau W_{i+}}$ sendo τ um escalar. Considere que o parâmetro de autocorrelação espacial em ρ seja conhecido. Sendo assim, tem-se que o vetor de parâmetros desconhecidos desse modelo é $\boldsymbol{\Phi} = (\boldsymbol{\beta}, \tau)$.

Seguindo o enfoque bayesiano, para inferir sobre o vetor paramétrico $\boldsymbol{\Phi}$ é necessário atribuir uma distribuição a priori para esse vetor. Portanto, considere, a priori, que $\boldsymbol{\beta}$ e τ sejam independentes e que possuam as seguintes distribuições

$$\begin{aligned} \boldsymbol{\beta} &\sim N(\boldsymbol{a}; V_\beta \boldsymbol{I}), \\ \tau &\sim Ga(b, c), \end{aligned} \tag{4.2}$$

sendo $\frac{b}{c}$ e $\frac{b}{c^2}$, respectivamente, a média e a variância da distribuição gama.

Dessa forma, tem-se que a distribuição a posteriori é dada pela seguinte forma

$$p(\boldsymbol{\Phi}|\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{\Phi})p(\boldsymbol{\beta})p(\tau), \tag{4.3}$$

sendo $p(\boldsymbol{Z}|\boldsymbol{\Phi})$ a função de densidade da distribuição dada pela Equação (4.1). Essa distribuição a posteriori não possui forma analítica conhecida e amostras podem ser obtidas através dos métodos de MCMC. Conforme descrito na Seção 3.3, faz-se então necessário obter as distribuições condicionais completas a posteriori. A distribuição condicional completa a posteriori de $\boldsymbol{\beta}$ é uma normal com média $V_p[\boldsymbol{X}^{'}(\boldsymbol{I}-\rho\boldsymbol{B})\boldsymbol{V}^{-1}\boldsymbol{Z}+\tau\boldsymbol{I}\boldsymbol{a}]$ e com a seguinte matriz de covariâncias $V_p = [\boldsymbol{X}^{'}(\boldsymbol{I}-\rho\boldsymbol{B})\boldsymbol{V}^{-1}\boldsymbol{X}+\tau\boldsymbol{I}]^{-1}$. A distribuição condicional completa a posteriori de τ é uma gama com parâmetros $b_{\tau}^{post} = (\frac{n}{2}+b-1)$ e $c_{\tau}^{post} = \frac{1}{2}(\boldsymbol{Z}-\boldsymbol{X}\boldsymbol{\beta})^{-1}(\boldsymbol{I}-\rho\boldsymbol{B})\boldsymbol{V}^{*}(\boldsymbol{Z}-\boldsymbol{X}\boldsymbol{\beta})+c$, onde $\boldsymbol{V}^{*}$ é uma matriz diagonal de ordem n formanda pelos elementos $V_i = \frac{1}{W_{i+}}$.

4.1 Estudo Simulado

Para verificar a capacidade de estimação dos parâmetros e analisar a sensibilidade da modelagem quanto a distribuição a priori, aplicou-se o modelo proposto acima em um conjunto de dados simulados.

Para a simulação dos dados, os parâmetros desconhecidos do modelo foram fixados em valores arbitrários. Suponha que a matriz desenho possui um intercepto e uma variável explicativa e os seguintes valores $\boldsymbol{\beta}^{'} = (-0,5 \quad ; \quad 3)$ e $\tau = 0,5$. Além disso, considere que há uma alta correlação espacial assumindo $\rho = 0,999$.

Com o intuito de analisar a sensibilidade do modelo quanto a distribuição a priori, ajustou-se os dados simulados considerando diferentes escolhas para os hiperparâmetros da distribuição. As escolhas foram realizadas de forma que ora tivesse uma distribuindo a priori informativa e ora tivesse menos informativa. Uma das formas utilizadas para transformar uma distribuição informativa em não informativa é aumentar a variabilidade dessa distribuição. Sendo assim, visando a análise de sensibilidade, a Tabela 1 apresenta as estimativas pontuais, obtidas pelas médias a posteriori, e as intervalares, obtidas pelos intervalos de credibilidade de 95% a posteriori, sob diferentes escolhas para os hiperparâmetros da distribuição a priori. Repare que, mesmo aumentando a variância de V_{β}, as estimativas dos parâmetros a posteriori se mantiveram próximas. Por isso, evidenciou-se que o modelo foi bem ajustado.

Foram realizadas 11000 iterações, com período de aquecimento (burnin) de 1000 e espaçamento de 10, retornando assim amostras a posteriori não correlacionadas de tamanho 1000. A Figura 1 mostra a convergência das cadeias dos parâmetros e também seus histogramas a posteriori utilizando a priori 2 definida na Tabela 1. As linhas em

Tabela 1: Análise de sensibilidade: estimativas a posteriori dos parâmetros sob diferentes escolhas de hiperparâmetros para a distribuição a priori. A estmativa pontual é dada pela média a posteriori e a intervalar pelo intervalo de credibilidade de 95%. Os valores verdadeiros dos parâmetros são $\beta_1 = -0,5$, $\beta_2 = 3$ e $\boldsymbol{\tau} = 0,5$.

Hiperparâmetros					Estimativas a posteriori		
	a	V_β	b	c	β_1	β_2	τ
Priori 1	(0 ; 0)	500	2	0,5	-0,5993	3,0273	0,5768
					(-3,0150 ; 1,7799)	(2,7764 ; 3,2921)	(0,4752 ; 0,6925)
Priori 2	(0 ; 0)	100	0,1	0,1	-0,5652	3,0280	0,5689
					(-3,0097 ; 1,7976)	(2,7631 ; 3,2850)	(0,4611 ; 0,6769)
Priori 3	(0 ; 0)	50	1	0,2	-0,6233	3,0219	0,5750
					(-2,9380 ; 1,6575)	(2,7543 ; 3,2716)	(0,4627 ; 0,6891)
Priori 4	(0 ; 0)	25	1	0,1	-0,6056	3,0208	0,5753
					(-2,8866 ; 1,6396)	(2,7533 ; 3,2703)	(0,4631 ; 0,6896)

verde representam os valores verdadeiros dos parâmetros, já as linhas em vermelho são as estimativas a posteriori dos parâmetros desconhecidos e seus respectivos intervalos de credibilidade de 95% em cor azul. Note que há indícios de convergência, que as médias a posteriori (estimativas pontuais) ficaram próximas dos valores verdadeiros e os intervalos contemplaram os valores verdadeiros.

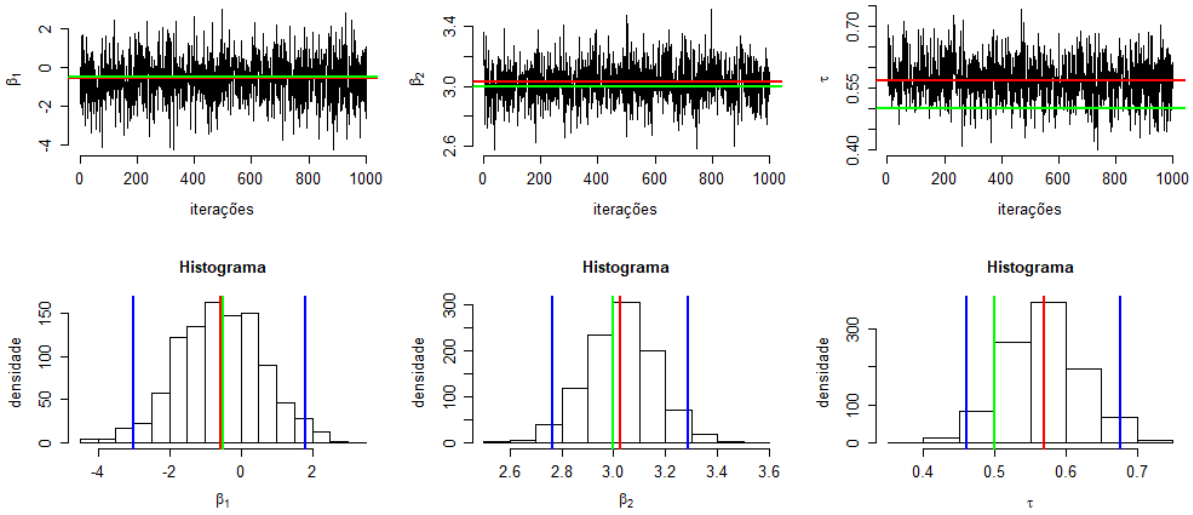

Figura 1: Traços das cadeias e histogramas das amostras dos parâmetros utilizando a priori 2 com dados simulados. As linhas em verde representam os valores verdadeiros dos parâmetros, já as linhas em vermelho são as estimativas a posteriori dos parâmetros desconhecidos e seus respectivos intervalos de credibilidade de 95% em cor azul.

4.2 Dados Reais

Os dados foram fornecidos pelo Paulo Henrique Leal de Sousa que foi orientado pelo Prof. Dr. Iuri da Costa Leite e co-orientado pela Prof. Dra. Patrícia Lusié Velozo da Costa no mestrado profissional em Epidemiologia em Saúde Pública, na Escola Nacional de Saúde Pública Sergio Arouca, na Fundação Oswaldo Cruz, no Rio de Janeiro.

A taxa de detecção de hanseníase em menores de 15 anos possui classificações categóricas diferentes da usuais, uma vez que altos índices nessa faixa etária representam combate inadequado da doença por parte dos orgãos de saúde. Assim, considerando a escala de 100 mil habitantes, a taxa é classificada em: hiperendêmica ($\geq 10,00$); muito alta ($9,99$ a $5,00$); alta ($4,99$ a $2,50$); média ($2,49$ a $0,50$); e baixa ($< 0,50$) (Revista de Saúde Pública (2017)) [10].

A Figura 2 apresenta as taxas de detecção de hanseníase para cada munícipio do Maranhão em 2010 de acordo com a classificação estabelecida desse indicador. Note que as cores predominantes são das categorias baixo e hiperendêmico, ou seja, apesar de muitas regiões apresentarem taxas quase ou totalmente nulas, outras apresentam taxas bastante elevadas. Ademais, percebe-se uma possível correlação espacial entre os municípios.

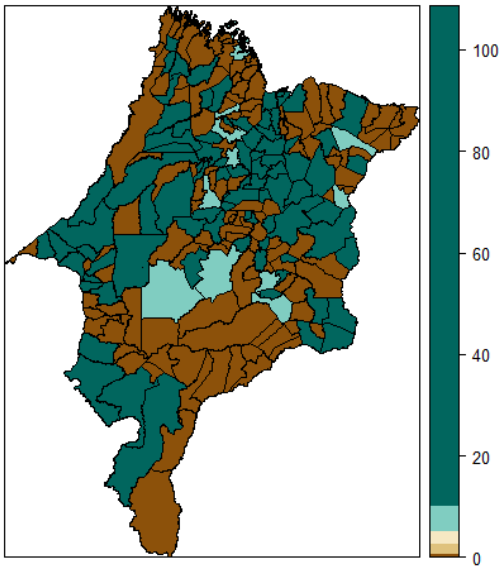

Figura 2: Taxa de Detecção de Hanseníase em menores de 15 anos nos munincípios do Maranhão em 2010.

Além da análise exploratória dos dados pelo mapa coroplético, a correlação espacial entre as regiões pôde ser verificada também por meio de duas medidas: o índice de Moran e o índice de Geary. Verificou-se que, em ambos os casos, há indícios de dependência

espacial entre os munícipios do Maranhão pois os índices de Moran e Geary foram aproximadamente 0,12 e 0,88, respectivamente.

Cerca de 55% das regiões não tiveram registros de infectados por Hanseníase, tendo taxas nulas. Diversos motivos podem ser avaliados, como por exemplo: regiões pouco povoadas, regiões que não notificam os casos ou até mesmo a migração de pessoas para as grandes cidades em busca de tratamento. Assim, esse fato pode implicar diretamente na modelagem das taxas e consequentemente nas estimativas dos parâmetros desconhecidos do modelo.

Considere um modelo com intercepto e uma variável explicativa. Utilizou-se como variável explicativa o Índice de Desenvolvimento Humano Municipal (IDHM) em 2010 de cada município do Maranhão. Além disso, considerando que há uma alta correlação espacial assumius-e $\rho = 0,999$.

Como não crença sob os parâmetros desconhecidos, considere a priori que

$$\begin{aligned} \boldsymbol{\beta} &\sim N(\mathbf{0}; 100\boldsymbol{I}), \\ \tau &\sim Ga(0,1; 0,1), \end{aligned} \tag{4.4}$$

sendo $\mathbf{0} = (0,0)'$.

Foram gerados 11000 valores com burnin de 1000 e espaçamento de 10, retornando assim amostras a posteriori não correlacionadas de tamanho 1000. Para a estimativa dos parâmetros desconhecidos, foram utilizadas a média a posteriori e intervalos de credibilidade de 95%.

A Figura 3 mostra a convergência das cadeias dos parâmetros e os histogramas das distribuições a posteriori. Note que parece ter havido convergência.

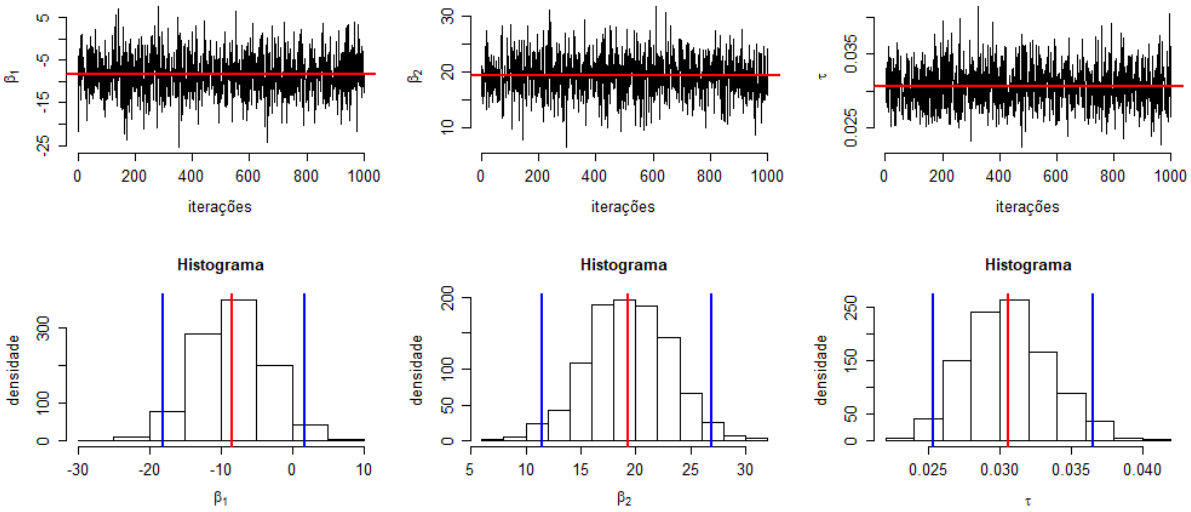

Figura 3: Traços das cadeias e histogramas das distribuições a posteriori usando o conjunto de dados reais.

A Tabela 2 apresenta as estimativas e os intervalos de credibilidade a posteriori dos parâmetros estimados.

Tabela 2: Médias a posteriori e intervalos de credibilidade de 95% para os parâmetros.

Parâmetros	β_1	β_2	τ
Priori 2	-8,4525	15,1894	0,0306
	(-18,0569 ; 1,6520)	(11,5084 ; 19,3378)	(0.0254 ; 0,0365)

A partir das estimativas dos parâmetros na Tabela 2, verificou-se que quão maior for o IDHM, maior deverá ser a taxa de detecção de hanseníase nos municípios do Maranhão. Resultado esse nada trivial, uma vez que esse indicador representa desenvolvimento humano nas áreas de educação, saúde e renda. Como argumentação inicial, essa relação pode estar associada, por exemplo, à subnotificação diferenciada segundo os municípios onde pessoas oriundas de regiões com baixos IDHM são notificadas nos grandes centros urbanos onde apresentam índices mais elevados.

5 Conclusão

Dentre os estados do Brasil, o Maranhão é um dos estados mais preocupante pois possui altas taxas de hanseníase. E isso motivou esse trabalho. Além dos possíveis fatores socio-econômicos associados a doença, esse trabalho analisou e verificou associação espacial entre as regiões do estado. Propos-se um modelo espacial CAR para ajustar uma transformação das taxas.

Para verificar o procedimento de inferência, gerou-se um conjunto de dados e estimou-se os parâmetros desse conjunto. Atribui-se diferentes escolhas para os hiperparâmetros da distribuição a priori do vetor de parâmetros desconhecidos e esse estudo é chamado de análise de sensibilidade. Essa análise se comportou de forma satisfatória e os parâmetros foram bem estimados mesmo sob diferentes escolhas dos hiperparâmetros.

Em seguida, analisou-se o conjunto de dados reais. A partir de uma análise exploratória dos dados e pelos Índices de Moran e de Geary, foi possível verificar que as taxas de detecção de hanseníase apresentaram correlação espacial, ou seja, a taxa de determinada região é influenciada pelas taxas de sua vizinhança. Ademais, através de um modelo espacial CAR e suas covariáveis associadas, verificou-se que o IDHM foi uma covariável significativa, porém indicou que regiões com maiores índices de desenvolvimento humano tendem a ter maiores taxas de haneníase.

O modelo proposto nesse trabalho serve para variáveis respostas contínuas que assumem valores na reta. As taxas de hanseníase são não-negativas. Para levar essas taxas na reta, aplicou-se uma função logarítmica. Porém há muitas taxas nulas indicando que a variável resposta é mista mesmo com a transformação utilizada. Problema esse que pode influenciar negativamente na estimativa e no intervalo de credibilidade dos parâmetros do modelo.

Consequentemente, fica como trabalhos futuros a análise e estudos de modelos mais adaptativos aos dados de hanseníase, levando em consideração principalmente a grande quantidade de taxas iguais a zero.

Referências

[1] ORGANIZATION, W. H. Weekly epidemiological record relevé épidémiologique hebdomadaire. *Weekly Epidemiological Record*, v. 34, p. 317–28, 2012.

[2] CRESSIE, N. A. C. *Statistics for Spatial Data*. [S.l.]: John Wiley & Sons, 1993.

[3] SCHMIDT, A. M.; NOBRE, A. A.; FERREIRA, G. S. Alguns aspectos da modelagem de dados espacialmente referenciados. *Rio de Janeiro*, 2003.

[4] BANERJEE, S.; GELFAND, A. E.; CARLIN, B. P. *Hierarchical Modeling and Analysis for Spatial Data*. [S.l.]: Chapman & Hall/CRC, 2003.

[5] GAMERMAN, D.; LOPES, H. F. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. [S.l.]: CRC Press, 2006.

[6] GEMAN, S.; GEMAN, D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, n. 6, p. 721–741, 1984.

[7] GELFAND, A. E.; SMITH, A. F. M. Samping-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, v. 85, n. 410, p. 398–409, 1990.

[8] METROPOLIS, N. et al. Equation of state calculations by fast computing machines. *The journal of chemical physics*, AIP, v. 21, n. 6, p. 1087–1092, 1953.

[9] HASTINGS, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, v. 57, p. 97–109, 1970.

[10] FREITASI, B. H. B. M. de et al. Tendência da hanseníase em menores de 15 anos em mato grosso (brasil), 2001-2013. *Rev Saúde Pública*, SciELO Public Health, v. 51, p. 28, 2017.

Universidade Federal Fluminense
Instituto de Matemática e Estatística
Departamento de Estatística

Assunto: Avaliação do Relatório Final associado ao Projeto de Iniciação à Pesquisa dos ProfessoresLuis Guillermo Coca Velarde(SIAPE 1282424) e Patrícia LusiéVelozo da Costa (SIAPE 1805333).

A Comissão de Pesquisa, no uso de sua atribuição, avaliou o relatório final do projeto de iniciação à pesquisa entitulado "Modelos hierárquicos bayesianos", submetido pelos docentesLuis Guillermo Coca Velarde(SIAPE 1282424) e Patrícia LusiéVelozo da Costa (SIAPE 1805333), bem como o relato do aluno envolvido Fellipe Carvalho Gomes (matrícula UFF 113054015). Considerando a Instrução de Serviço GET – e entendendo que o projeto cumpriu seu propósito, a Comissão de Pesquisa do GET considera validada a sua execução e indica o aproveitamento do trabalho realizado como atividade complementar para os alunos envolvidos.

Niterói, 03 de julho de 2018

Ana Beatriz Monteiro Fonseca

Eduardo Ferioli Gomes

Luciane Alcoforado

Wilson Calmon

Universidade Federal Fluminense

Instituto de Matemática e Estatística

Departamento de Estatística

Modelos lineares hierárquicos bayesianos

Relatório final

Identificação do Projeto: Projeto de Iniciação à Pesquisa

Docentes responsáveis: Patrícia Lusié Velozo da Costa (SIAPE: 1805333)
Luis Guillermo Coca Vellarde (SIAPE: 1282424)

Aluno vinculado: Fellipe Carvalho Gomes (matrícula: 113054015)
(aluno do curso de Estatística)
carga horária semanal: 20horas

Resumo do plano inicial

Os docentes já haviam inicializado um trabalho com o aluno Fellipe em Projeto Final I em 2017-1. Porém o aluno apresentou muitas dificuldades para inicializar a modelagem hierárquica e, devido a motivos pessoais, resolveu cursar Projeto Final II em 2018-1. Sendo assim, os docentes sentiram a necessidade de criar esse projeto para incentivar o aluno a estudar os conteúdos de inferência bayesiana e modelagem hierárquica. Para isso, pretendeu-se que o aluno trabalhasse com dados simulados a fim de analisar a sensibilidade da distribuição a priori atribuída e aprender sobre como inferir sobre os parâmetros desconhecidos.

Resultados previamente esperados

O aluno apresentava muitas dúvidas sobre Inferência Bayesiana e, em especial, sobre os métodos de Monte Carlo via cadeias de Markov (MCMC). Esperava-se que o aluno aprendesse mais sobre esse conteúdo e sobre modelagem hierárquica. Além disso, esperava-se que o aluno aprendesse a avaliar o modelo proposto usando dados artificiais.

Resumo do projeto executado e resultados efetivamente obtidos

O projeto foi previsto para ser desenvolvido entre setembro de 2017 a fevereiro de 2018, totalizando 6 meses. Porém, o aluno precisou de um pouco mais de tempo para conciliar os estudos do projeto, com os da graduação e o seu estágio. O aluno foi avaliado ao longo do período de execução, citado anteriomente, através de apresentações realizadas sobre o assunto estudado e de implementações de modelos pertinentes ao assunto.

Os docentes esperam ter contribuído na formação acadêmica do aluno uma vez que esse precisou aprender sobre modelagem hierárquica, simulação de dados artificiais e reforçou os conhecimentos sobre inferência bayesiana.

O aluno pode sanar as dúvidas citadas na seção anterior, reforçar os conceitos estudados ao longo da graduação sobre modelos lineares e estudar sobre modelos hierárquicos. Além disso, o aluno estudou alguns modelos diferentes até optarmos pelo modelo mais adequado ao assunto de interesse e pode treinar como fazer as contas da distribuição a priori, da distribuição a posteriori, de distribuições condicionais completas a posteriori e também pode exercitar a implementação do MCMC e de geração de dados artificiais. Além disso, o aluno comparou a amostragem realizada no R com os resultados obtidos no Jags. O aluno preparou um texto relatando o assunto estudado, o modelo implementado e analisando os resultados obtidos.

Niterói, 25 de junho de 2018.

Patrícia Lusié Velozo da Costa (SIAPE: 1805333)

Luis Guillermo Coca Vellarde (SIAPE: 1282424)

Relatório de Atividades - Projeto de Iniciação à Pesquisa

Modelos Hierárquicos Bayesianos

Fellipe Carvalho Gomes

24 junho 2018

O projeto teve início em setembro de 2017 e teve como principal objetivo o estudo sobre a modelagem hierárquica sob a perspectiva de inferência bayesiana pois esta metodologia foi adotada posteriormente em minha monografia, de título "Uma aplicação de modelo hierárquico bayesiano na modelagem da dor em rescém nascidos submetidos à punção de calcâneo" e como objetivo secundário, devido à complexidade da metodologia estudada, foi planejada uma revisão da literatura de modelos de regressão linear simples sob o paradigma clássico e sob o paradigma bayesiano para o aprendizado de novos elementos relacionados ao seu ajuste.

Para isso foram utilizadas referências bibliográficas sugeridas pelos professores Guillermo Coca Velarde e Patrícia Lusié Velozo da Costa para a compreensão da metodologia e toda a implementação da parte computacional referentes à: simulação, manipulação e implementação do algoritmo estudado foram realizados com o uso do software R (versão 1.0.153) que é uma linguagem e um ambiente para programação estatística e além disso todo o texto produzido foi feito em LaTeX, uma linguagem para a produção de artigos científicos.

Toda semana ocorriam reuniões para discutir o projeto, retirar eventuais dúvidas dos cálculos realizados e no decorrer do projeto quando fez-se nescessário as simulações em R esses encontros também envolviam a avaliação do andamento da produção dos códigos computacionais.

Diversos resultados interessantes foram obtidos com o estudo do modelo de regressão linear simples quanto no hicárquico. No modelo de regressão linear simples além dos dados simulados ainda foi utilizado um conjunto de dados reais que contam com registros de tempo e distância até a parada de um carro ao apertar o freio em alta velocidade e no modelo de regressão linear hierárquico todos os parâmetros utilizados para simular a amostra foram recuperados.

De maneira resumida o projeto foi muito benéfico não apenas acrescentando o conhecimento sobre uma nova metodologia para o ajuste de modelos de regressão linear como também gerou diversos benefícios secundários como: aperfeiçoar habilidades de simulação dos dados, treinar a escrita de artigos cientificos com LaTeX, aprimorar a habilidade em programação em linguagem R melhorando assim tanto a minha formação como reconhecendo o verdadeiro valor da educação e da ciência.

Fellipe Carvalho Gomes

MODELOS LINEARES HIERÁRQUICOS BAYESIANOS

Niterói - RJ, Brasil

Fevereiro de 2018

Universidade Federal Fluminense

Fellipe Carvalho Gomes

MODELOS LINEARES HIERÁRQUICOS BAYESIANOS

Projeto de Iniciação à Pesquisa

Relatório Final do Projeto de Iniciação à Pesquisa submetido ao Departamento de Estatística da Universidade Federal Fluminense.

Docentes responsáveis pelo Projeto: Guillermo Coca Velarde e Patrícia Lusié Velozo da Costa .

Niterói - RJ, Brasil

Fevereiro de 2018

Resumo

Este projeto tem como objetivo o estudo sobre a modelagem hierárquica sob a perspectiva de inferência bayesiana. Inicialmente, estudou-se o modelo de regressão linear simples usando o método de Monte Carlo via cadeias de Markov (MCMC) para aprender a estimar os parâmetros de um modelo através da inferência bayesiana os resultados foram comparados com o ajuste do modelo de regressão lienar sobre a perspectiva clássica. Em seguida o projeto envolveu o estudo de modelos lineares hierárquicos. Recorreu-se a dados simulados para analisar a eficiência do procedimento de inferência utilizado e avaliar a sensibilidade da distribuição a priori escolhida.

Sumário

Lista de Figuras

Lista de Abreviações

DCCP distribuições condicionais completas a posterioris

fdp função de densidade de probabilidade

iid independentes e identicamente distribuídas

MCMC Monte Carlo via cadeias de Markov

1 Introdução

Análise de risco de crédito de um cliente, previsão da quantidade de chuva em um dado local e estimativa de erros ou falhas de um novo produto ou serviço são apenas alguns dos exemplos de possíveis assuntos de interesse e nos quais decisões podem ser tomadas, e tais decisões podem ser dadas através da modelagem estatística. Modelar um fenômeno aleatório consiste em realizar afirmações sobre o processo gerador dele e, em Estatística, essas afirmações costumam ser sobre as distribuições das variáveis aleatórias envolvidas na geração do fenômeno de interesse.

A atribuição de uma distribuição pode ser dada em níveis, como, por exemplo, quando as observações pertencem a grupos diferentes e cada grupo tem suas próprias propriedades (média, variância, entre outras). Nesses casos recorre-se a modelos hierárquicos, que também são conhecidos como modelos multiníveis. Aplicações desses modelos podem ser encontradas em várias áreas tais como na Educação, Economia, nas Ciências Sociais e na Saúde. Suponha que demógrafos desejam examinar como diferenças no desenvolvimento da economia nacional podem interferir na relação entre o grau educacional dos adultos e a taxa de fertilidade. Para isso, pode-se utilizar 2 estágios: nível nacional (indicadores econômicos) e nível domiciliar (educação e fertilidade). Ou suponha que o interesse esteja em medir o rendimento escolar dos alunos e, para isso, utiliza-se 4 estágios: os alunos, as turmas, as escolas e os órgãos administradores ou a região.

Muitas vezes, os parâmetros dessas distribuições podem ser desconhecidos e deseja-se inferir sobre eles. Há 2 grandes escolas de inferência: a clássica e a bayesiana. A clássica trata esses parâmetros como quantidades fixas e não atribui distribuições a eles. A estimação dos parâmetros é dada através da função de verossimilhança. A bayesiana atribui uma distribuição, chamada de distribuição a priori, ao conjunto de parâmetros desconhecidos quantificando a sua crença sobre esse conjunto e a estimação dos parâmetros é dada através da distribuição a posteriori, que é proporcional ao produto da função de verossimilhança com a distribuição a priori. A distribuição a priori é proposta por meio de conhecimentos subjetivos que se tenha sobre os parâmetros. É dito ter uma distribuição a

priori informativa, quando há uma forte crença sobre o conjunto de parâmetros. Quando não há crença sobre os parâmetros em questão, distribuições a priori não informativas são utilizadas e, nesse caso, pode-se comparar os resultados obtidos com os da inferência clássica. A inferência bayesiana será o foco desse trabalho.

Quando há mais de um parâmetro desconhecido, a distribuição a posteriori torna-se uma distribuição multivariada. Muitas vezes essa distribuição é desconhecida e/ou muito difícil de ser analisada. O avanço computacional das últimas décadas tem permitido a aplicação de modelos complexos de forma mais realista na representação de fenômenos aleatórios em estudo. Até a década de 80 utilizava-se métodos aproximados de inferência enquanto que na década de 90 os métodos de Monte Carlo via cadeias de Markov (MCMC), e, mais especificamente, o amostrador de Gibbs e o Metropolis-Hastings, revolucionaram as aplicações no contexto bayesiano. Maiores detalhes podem ser vistos em Robert e Casella (2005) [1] e em Gamerman e Lopes (2006) [2].

Modelos de regressão linear explicam a variável resposta através de variáveis explicativas e supõe que dada as variáveis explicativas, as variáveis respostas são independentes. Modelos lineares hierárquicos são generalizações dos modelos de regressão linear pois assumem que as observações das unidades pertencentes ao agregado são dependentes.

Esse trabalho possui a seguinte estrutura: o Capítulo 2 contém os objetivos deste trabalho; o Capítulo 3 apresenta a metodologia utilizada; no Capítulo 4 gerou-se dados simulados a fim de analisar os modelos propostos e, por fim, o Capítulo 5 apresenta as conclusões desse estudo, seguido por referências utilizados para gerar as amostras e a modelagem.

2 Objetivos

O objetivo deste trabalho é estudar sobre a modelagem linear hierárquica bayesiana. Esse estudo consiste em basicamente 2 (duas) etapas: estudar procedimentos de inferência bayesiana e, em seguida, modelagens lineares hierárquicas. Em seguida, trabalhou-se com dados simulados a fim de analisar a sensibilidade da distribuição a priori atribuída e avaliar a inferência sobre os parâmetros desconhecidos.

3 Materiais e Métodos

Este capítulo contém uma breve revisão de alguns conceitos abordados ao longo deste trabalho. Conforme mencionado no Capítulo 1, esse trabalho consiste em modelagem estatísticas e, portanto, distribuições são atribuídas a determinados fenômenos e essas distribuições possuem determinadas quantidades desconhecidas, chamadas de parâmetros, fazendo-se necessário inferir sobre esses parâmetros. Recorreu-se a inferência bayesiana e, portanto, a Seção 3.1 contém uma revisão sobre esse assunto. Em seguida, o interesse consiste em recorrer a modelos lineares hierárquicos e, por isso, a Seção 3.2 introduz o conceito de modelos lineares e, sem seguida a Seção 3.3 extende esses modelos introduzindo hierarquias.

3.1 Inferência bayesiana

Um experimento aleatório é um processo que acusa variabilidade em seu resultado. O conjunto de todos os possíveis resultados desse experimento é chamado de espaço amostral. Os subconjuntos do espaço amostral são denominados de eventos aleatórios. Uma variável aleatória é uma função que associa números reais a cada um dos elementos do espaço amostral. Cada elemento passa a ter um único número real associado a ele. Um mesmo número pode estar associado a mais de um elemento. Todos os elementos do espaço amostral tem que ter um número associado.

Seja Y_i uma variável aleatória. O índice i é chamado de unidade amostral e pode representar, por exemplo, um indivíduo, um instante de tempo ou um grupo de idade. Suponha que tenha-se N unidades amostrais e que haja interesse em inferir sobre a média dessa população, representada por μ, e/ou sobre a variância dessa população, representada por σ^2, por exemplo.

Seja $p(Y_1, \ldots, Y_N|\boldsymbol{\theta})$ a função de distribuição ou de densidade da variável resposta dado um conjunto de parãmetros $\boldsymbol{\theta}$. Após obter uma amostra de tamanho n da variável resposta, pode-se inferir sobre os parâmetros populacionais. Através da inferência

bayesiana, atribui-se uma distribuição a priori para $\boldsymbol{\theta}$. Denote essa distribuição por $h(\boldsymbol{\theta})$. Dessa forma, a inferência sobre o vetor paramétrico é dada através da distribuição a posteriori $p(\boldsymbol{\theta}|y_1, \ldots, y_n)$, sendo y_i o i-ésimo valor amostrado da variável de interesse. Pelo Teorema de Bayes, tem-se que a distribuição a posteriori é dada por

$$p(\boldsymbol{\theta}|y_1, \ldots, y_n) = \frac{h(\boldsymbol{\theta})p(y_1, \ldots, y_n|\boldsymbol{\theta})}{p(y_1, \ldots, y_n)}, \tag{3.1}$$

sendo $p(y_1, \ldots, y_n)$ chamada de distribuição marginal da variável de interesse.

Por exemplo, suponha que $Y_i \overset{iid}{\sim} N(\mu, 1)$ e que o interesse esteja em estimar a média populacional, dada por μ. A priori, suponha que $\mu \sim N(m_\mu, V_\mu)$. Dessa forma, obtem-se que a distribuição a posteriori é dada por

$$p(\mu|y_1, \ldots, y_n) = \frac{h(\mu)\prod_{i=1}^{n} p(y_i|\mu)}{p(y_1, \ldots, y_n)}, \tag{3.2}$$

onde $h(\mu)$ é a função de densidade de probabilidade (fdp) da distribuição normal com média m_μ e variância V_μ, $p(y_i|\mu)$ é a fdp da distribuição normal com média μ e variância 1 e $p(y_1, \ldots, y_n)$ é a distribuição marginal das observações que pode ser obtida integrando o parâmetro μ no numerador, ou seja,

$$p(y_1, \ldots, y_n) = \int_{-\infty}^{+\infty} h(\mu) \prod_{i=1}^{n} p(y_i|\mu) d\mu. \tag{3.3}$$

Note que essa a distribuição dada em (3.3) não depende de μ e que, por definição de fdp, a integral da fdp a posteriori, dada em (3.2), com respeito a μ tem que ser igual a 1. Sendo assim, tem-se que a distribuição a posteriori é proporcional a

$$\begin{aligned} p(\mu|y_1, \ldots, y_n) &\propto h(\mu) \prod_{i=1}^{n} p(y_i|\mu) \\ &\propto \exp\left\{-\frac{1}{2V_\mu}(\mu - m_\mu)^2\right\} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(n + \frac{1}{v_\mu}\right)\left[\mu^2 - 2\mu\left(n + \frac{1}{v_\mu}\right)^{-1}\left(\frac{m_\mu}{V_\mu} + \sum_{i=1}^{n} y_i^2\right)\right]\right\} \end{aligned} \tag{3.4}$$

Integrando a equação acima, obtém-se que

$$\mu|y_1,\ldots,y_n \sim N\left(\left(n+\frac{1}{v_\mu}\right)^{-1}\left(\frac{m_\mu}{V_\mu}+\sum_{i=1}^{n}y_i^2\right),\left(n+\frac{1}{v_\mu}\right)^{-1}\right). \tag{3.5}$$

E, portanto, a inferência sob o parâmetro da média populacional é realizada através dessa distribuição normal. Sendo assim, uma estimativa pontual para o parâmetro μ, sob o paradigma bayesiano, é dada pela média dessa distribuição normal e também pode-se obter estimativas intervalares, que são chamadas de intervalos de crebilidade, no contexto bayesiano.

3.1.1 Distribuição a Priori

Na abordagem bayesiana existem diferentes formas de especificação da distribuição a priori para o vetor paramétrico desconhecido, $\boldsymbol{\theta}$. A distribuição a priori deve representar (probabilisticamente) o conhecimento prévio sobre esse vetor antes de observar os resultados de um novo experimento. Com algum conhecimento probabilístico sobre isso é possível definir uma família paramétrica de densidades. Essa família, muitas vezes, possui parâmetros desconhecidos que são chamados de hiperparâmetros.

A distribuição a priori é subjetiva. Ela pode ser determinada através do conhecimento de um especialista e/ou através de dados experimentais anteriores, por exemplo. Uma outra forma de especificar uma distribuição a priori é escolher uma de forma que a distribuição a posteriori e a priori pertençam a mesma família. Quando ambas as distribuições pertencem a mesma classe de distribuições a atualização do conhecimento que se tem sobre $\boldsymbol{\theta}$ envolve apenas a mudança nos hiperparâmetros e é dito ter uma distribuição conjugada. Um exemplo de distribuição conjugada foi visto anteriormente na Equação 3.5, quando propos-se uma distribuição a priori normal para a média de uma população com distribuição normal e obteve-se uma distribuição a posteriori normal.

A família exponencial é muito importante ao se utilizar distribuições a prioris conjugadas pois através dessa família pode-se encontrar com facilidade a distribuição conjugada e a distribuição a posteriori, tanto para o caso contínuo quanto discreto. Para maiores detalhes, vide Migon (2014)[3].

Definida a família da distribuição a priori, uma outra discussão é como definir os hiperparâmetros. Através disso é possível ter uma distribuição "vaga" ou uma informativa. Se a distribuição estiver concentrada em uma região pequena, é dito ter

uma distribuição informativa. Se a variabilidade da distribuição é muito alta, é dito ter uma distribuição não informativa.

Caso o pesquisador tenha uma crença forte, ele utiliza distribuições informativas. Caso contrário, ele recorre a uma distribuição a priori com efeito mínimo na distribuição a posteriori. Distribuições a priori não informativas podem ser obtidas da seguinte forma: aumentando-se a variância da distribuição a priori, utilizando-se uma distribuição a priori uniforme ou ainda através de distribuições proposta por Jeffreys (1961)[4]. Maiores detalhes podem ser encontrados em Ehlers (2003)[5].

Considere o exemplo em que uma amostra aleatória simples de uma população com distribuição normal com média populacional $\mu = 0$ e variância $\sigma^2 = 1$ é selecionada e dois pesquisadores desejam inferir qual a média populacional, dada por μ. O pesquisador A possui uma forte crença de que a média esteja em torno de 5 e portanto sua distribuição priori contará com uma variância pequena ($\sigma^2 = 2$), de modo que sua distribuição a priori seja informativa. Já o pesquisador B resolve declarar sua distribuição a priori com a mesma média 5 porém sua incerteza o levou a selecionar um alto valor para a variância $\sigma^2 = 20$. A Figura 1 compara as distribuições a priori e a posteriori dos diferentes pesquisadores. Note que a distribuição a posteriori encontrada pelo pesquisador A sofre maior influência da distribuição a priori uma vez que essa é mais informativa e há um tamanho amostral pequeno.

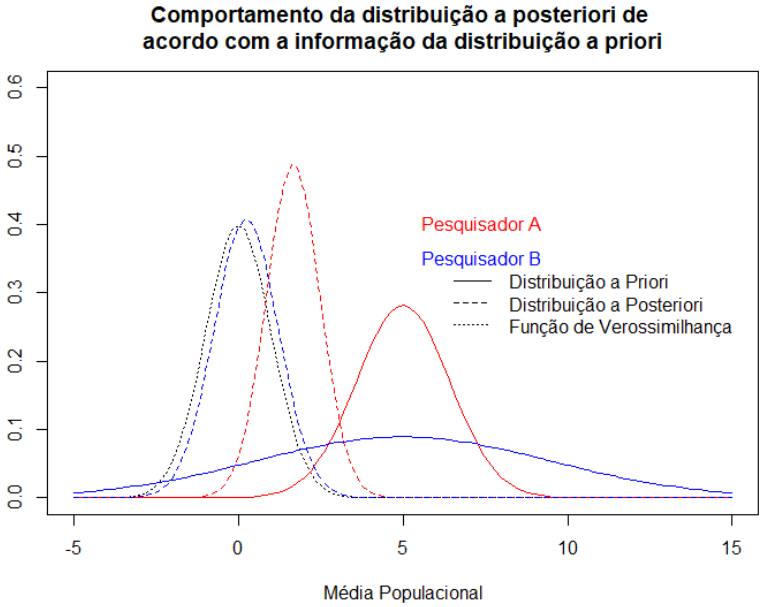

Figura 1: Comparando comportamento da distribuição a posteriori de acordo com a seleção da distribuição a priori

Quanto maior o tamanho da amostra, menor tende a ser a influência da distribuição a priori. Porém, faz necessário ter cautela ao definir uma distribuição a priori e por isso é desejado fazer uma análise de sensibilidade para estudar os impactos disso na inferência sobre os parâmetros.

3.1.2 Amostrador de Gibbs

A distribuição a posteriori de um parâmetro $\boldsymbol{\theta}$, dada pela Equação 3.1 contém toda a informação probabilística a respeito deste parâmetro. Quando a forma analítica dessa distribuição é conhecida, então um gráfico da fdp pode ilustrar o comportamento probabilístico do parâmetro de interesse e auxiliar em alguma tomada de decisão. Porém, quando a forma analítica não é conhecida ou é muito custosa de ser obtida, pode-se recorrer a métodos de simulação tais como os métodos MCMC.

A dependência de Markov é um conceito atribuído ao matemático russo Andrei Andreivich Markov que no início do século 20 investigou o comportamento da alternância de vogais e consoantes no poema Onegin by Poeshkin. Markov desenvolveu um modelo probabilístico onde os resultados sucessivos dependiam em todos os seus predecessores apenas através do antecessor imediato e o modelo permitiu-lhe obter boas estimativas da frequência relativa de vogais no poema. Quase ao mesmo tempo o matemático francês Henri Poincare estudou sequências de variáveis aleatórias que eram de fato Cadeias de Markov, Gamerman (2006)[2].

Uma cadeia de Markov de primeira ordem é um processo estocástico $\{W_0, W_1, ...\}$ de tal forma que a distribuição de W_t, dados todos os valores anteriores $W_0, ..., W_{t-1}$, depende apenas de W_{t-1}, ou seja:

$$p(W_t|W_0, ..., W_{t-1}) = p(W_t|W_{t-1}).$$

Os métodos requerem que a cadeia seja:

- homogênea: as probabilidades de transição de um estado para outro são invariantes.
- irredutível: cada estado pode ser atingido a partir de qualquer outro em um número finito de interações.
- aperiódica: não haja estados absorventes.

A função de transição da cadeia, definida por $P(z|w)$, é a função que indica a probabilidade da cadeia mover-se para o estado z dado que se encontra no estado w

no tempo anterior. Seja uma distribuição $\pi(w), w \in \mathbb{R}^d$, conhecida a menos de uma constante multiplicativa, porém complexa o bastante para não ser possível obter uma amostra diretamente. Para gerar amostras de $\pi(w)$, calcula-se e utiliza-se a função de transição $P(z|w)$ que converge para $\pi(w)$ na k-ésima iteração. O processo é iniciado em um estado arbitrário de w e após um número suficientemente grande de simulações, as observações geradas são aproximadamente iguais a distribuição alvo $\pi(w)$. Robert e Casela (2005)[1]

A convergência da cadeia de Markov acontece depois de um período chamado de aquecimento. Conforme o número de iterações aumenta, os valores iniciais são esquecidos pela cadeia até convergir para a distribuição de equilíbrio $\pi(w)$. Na prática, os valores iniciais são descartados, pois são considerados como uma amostra de aquecimento.

Com os avanços dos métodos de MCMC, surgiu o amostrador de Gibbs, proposto por Geman e Geman (1984)[6] e tornou-se popular por Gelfand e Smith (1990)[7].

Sejam $\pi(\theta)$ a distribuição da qual se tem o interesse de amostrar onde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$, θ_{-j} é o vetor composto por todos os elementos de θ, exceto pelo elemento θ_j, $j = 1, \ldots, d$, e $\pi_j(\theta_j) = \pi(\theta_j|\theta_{-j})$ as distribuições condicionais completas, ou seja, distribuições de cada parâmetro condicionada aos demais parâmetros do modelo.

Portanto o amostrador de Gibbs irá gerar sucessivas amostras das distribuições condicionais completas da seguinte de acordo com o algoritmo descrito abaixo:

1. Determinar um valor inicial para cada θ_j, definindo $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_d^{(0)})$.

2. Iniciar o contador de iteração k=1.

3. Obter um novo valor para $\theta^{(k)} = (\theta_1^{(k)}, \ldots, \theta_d^{(k)})$ pela geração sucessiva das distribuições condicionais completas:

$$\begin{aligned}
\theta_1^{(k)} &\sim \pi(\theta_1|\theta_2^{(k-1)}, \ldots, \theta_d^{(k-1)}), \\
\theta_2^{(k)} &\sim \pi(\theta_2|\theta_1^{(k)}, \theta_3^{(k-1)}, \theta_4^{(k-1)}, \ldots, \theta_d^{(k-1)}), \\
&\vdots \\
\theta_d^{(k)} &\sim \pi(\theta_d|\theta_1^{(k)}, \ldots, \theta_{d-1}^{(k)})
\end{aligned}$$

4. Atualizar o contador $k = k + 1$,

5. Repetir os passos 3 e 4 até que a convergência seja obtida.

Como a convergência ocorre após o aquecimento (ou burn-in), é comum usar os valores de $\theta^{(a)}$, $\theta^{(a+t)}$, $\theta^{(a+2t)}$,... para compor a amostra de θ, sendo $a-1$ o número de iterações iniciais do aquecimento e t o espaçamento utilizado para diminuir a autocorrelação dos parâmetros. Maiores detalhes podem ser vistos em Gamerman (2006)[2].

3.2 Modelo de regressão linear simples bayesiano

Embora o objetivo do projeto seja sobre a abordagem utilizando modelos lineares hierárquicos, é fundamental o entendimento sobre modelos lineares bayesianos simples pois a partir destes modelos que apresentam uma relação linear nos parâmetros entre as variáveis e a função de verossimilhança, uma distribuição a priori para os parâmetros também deve ser declarada e este conceito da declaração da distribuição a priori será aprofundado em seguida ao tratar o tema principal deste projeto.

Inspirado no conjunto de dados disponibilizado por Ezekiel (1930)[8] e que hoje faz parte do conjunto de banco de dados nativos do R [9] (a base de dados pode ser obtida ao escrever "cars"no console) um modelo de regressão linear simples pela perspectiva bayesiana será ajustado e para isso primeiramente alguns cálculos precisam ser realizados para que seja possível a implementação dos algoritmos de MCMC em seguida.

Os dados informam a velocidade dos carros e as distâncias tomadas para parar, esses dados foram registrados na década de 1920 e são de grande utilidade didática até os dias de hoje, sendo assim, considere que a variável aleatória Y corresponda a velocidade seja a de interesse, comumente chamada de variável resposta e que a variável aleatória X que corresponde a distância tomada para parar seja utilizada para explicar a variável Y que comumente é chamada de variável explicativa ou covariável.

Suponha então um exemplo em que a população de interesse tenha distribuição normal com média $\beta_0 + \beta_1 X$, sendo β_0 e β_1 desconhecidos e variância σ^2 desconhecida. Seja $\tau = \frac{1}{\sigma^2}$ o parâmetro chamado de precisão. O parâmetro β_0 é conhecido como intercepto ou coeficiente linear e o β_1 como coeficiente angular. Além disso, suponha que as unidades dessa população sejam independentes e identicamente distribuídas (iid). Dessa forma, tem-se que as unidades dessa população tem a seguinte distribuição:

$$Y_i \overset{iid}{\sim} N\Big(\beta_0 + \beta_1 X_i, \frac{1}{\tau}\Big), \tag{3.6}$$

onde $i = 1, 2, \ldots, N$.

Obtendo-se uma amostra de tamanho n, pode-se inferir sob os parâmetros desconhecidos, $\boldsymbol{\theta} = (\beta_0, \beta_1, \tau)$, através da distribuição a posteriori e para obter essa distribuição faz-se necessário calcular a função de verossimilhança, que pode ser obtida da seguinte forma:

$$\begin{aligned} p(\boldsymbol{y}|\beta_0, \beta_1, \tau) &= \prod_{i=1}^{n} p(y_i|\beta_0, \beta_1, \tau) \\ &= \prod_{i=1}^{n} \frac{\sqrt{\tau}}{\sqrt{2\pi}} exp\Big\{ -\frac{\tau}{2}(y_i - \beta_0 - \beta_1 x_i)^2 \Big\}, \end{aligned} \tag{3.7}$$

onde $\boldsymbol{y} = (y_1, \ldots, y_n)$ é a amostra coletada.

Considere a priori que os parâmetros sejam independentes e que

$$\begin{aligned} \beta_0 &\sim N(m_0, \sigma_0^2), \\ \beta_1 &\sim N(m_1, \sigma_1^2) \text{ e} \\ \tau &\sim G(a, b). \end{aligned}$$

Dessa forma, tem-se que a distribuição conjunta a priori possui a seguinte forma:

$$p(\beta_0, \beta_1, \tau) \propto exp\Big\{ -\frac{1}{2\sigma_0^2}(\beta_0 - m_0)^2 \Big\} exp\Big\{ -\frac{1}{2\sigma_1^2}(\beta_1 - m_1)^2 \Big\} \tau^{a-1} exp\{-b\tau\}. \tag{3.8}$$

Combinando a função de verossimilhança com a distribuição a priori, obtem-se a distribuição a posteriori que é proporcional a:

$$\begin{aligned} p(\beta_0, \beta_1, \tau|\boldsymbol{y}) \propto\ & \tau^{\frac{n}{2}+a-1} exp\left\{ -\frac{\tau}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 - b\tau - \frac{1}{2\sigma_0^2}(\beta_0 - m_0)^2 \right\} \times \\ & exp\left\{ -\frac{1}{2\sigma_1^2}(\beta_1 - m_1)^2 \right\}. \end{aligned} \tag{3.9}$$

Note que essa distribuição é multivariada e não possui forma analítica conhecida. Sendo assim, recorre-se aos métodos de MCMC, descritos na Subseção 3.1.2, para se obter amostras dessa distribuição. E então faz-se necessário obter as distribuições condicionais completas a posterioris (DCCP) de β_0, β_1 e τ.

A primeira DCCP definida aqui será a de τ. Essa distribuição é facilmente obtida reescrevendo a distribuição a posteriori, dada na Equação (3.9), considerando apenas τ como parâmetro desconhecido e todos os outros parâmetros como conhecidos, ou seja,

$$p(\tau|y_1,\ldots,y_n,\beta_0,\beta_1) \propto \tau^{\frac{n}{2}+a-1} exp\Big\{ -\tau\Big(\frac{\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2}{2}+b\Big)\Big\} \tag{3.10}$$

Logo, a DCCP de τ é

$$\tau|y_1,\ldots,y_n,\beta_0,\beta_1 \sim Gama\Big(\frac{n}{2}+a, b+\frac{1}{2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2\Big) \tag{3.11}$$

Já para o cálculo da DCCP de β_0 será considerado apenas β_0 como parâmetro desconhecido e o restante como conhecido, obtendo assim:

$$\begin{aligned} p(\beta_0|y_1,\ldots,y_n,\tau,\beta_1) &\propto exp\left\{ -\frac{\tau}{2}\sum_{i=1}^{n}(\beta_0^2-2y_i\beta_0+2\beta_0\beta_1 x_i)-\frac{1}{2\sigma_0^2}(\beta_0^2-2m_0\beta_0)\right\} \\ &\propto exp\left\{ -\frac{1}{2}(\tau n+\frac{1}{\sigma_0^2})\left\{\beta_0^2-2\beta_0\frac{(\tau\sum_{i=1}^{n}y_i-\tau\beta_1\sum_{i=1}^{n}x_i+\frac{m_0}{\sigma_0^2})}{\tau n+\frac{1}{\sigma_0^2}}\right\}\right\} \end{aligned}$$

e, portanto, tem-se que $\beta_0|y_1,\ldots,y_n,\tau,\beta_1 \sim N(M_0,C_0)$, sendo $C_0^{-1}=\Big(n\tau+\frac{1}{\sigma_0^2}\Big)$ e $M_0=\dfrac{(\tau\sum_{i=1}^{n}y_i-\tau\beta_1\sum_{i=1}^{n}x_i+\frac{m_0}{\sigma_0^2})}{\tau n+\frac{1}{\sigma_0^2}}$.

E por fim, o cálculo da DCCP de β_1 será considerado apenas β_1 como parâmetro desconhecido e o restante como conhecido, obtendo assim:

$$\begin{aligned} p(\beta_1|y_1,\ldots,y_n,\tau,\beta_0) &\propto exp\left\{ -\frac{\tau}{2}\sum_{i=1}^{n}(\beta_1^2x_i^2-2y_i\beta_1 x_i+2\beta_0\beta_1 x_i)-\frac{1}{2\sigma_1^2}(\beta_1^2-2m_1\beta_1)\right\} \\ &\propto exp\left\{ -\frac{1}{2}(\tau\sum_{i=1}^{n}x_i^2+\frac{1}{\sigma_1^2})\left\{\beta_1^2-2\beta_1\frac{\tau\sum_{i=1}^{n}x_iy_i-\tau\beta_0\sum_{i=1}^{n}x_i+\frac{m_1}{\sigma_1^2}}{\tau\sum_{i=1}^{n}x_i^2+\frac{1}{\sigma_1^2}}\right\}\right\} \end{aligned}$$

e, portanto, tem-se que $\beta_1|y_1,\ldots,y_n,\tau,\beta_0 \sim N(M_1,C_1)$, sendo $C_1^{-1}=\Big(\tau\sum_{i=1}^{n}x_i^2+\frac{1}{\sigma_1^2}\Big)$ e $M_1=\frac{\tau\sum_{i=1}^{n}x_iy_i-\tau\beta_0\sum_{i=1}^{n}x_i+\frac{m_1}{\sigma_1^2}}{\tau\sum_{i=1}^{n}x_i^2+\frac{1}{\sigma_1^2}}$.

Ao finalizar essas contas já é possível realizar a implementação do algoritmo do método de MCMC, os resultados desses ajustes serão discutidos na seção 4 de análise e resultados

em 4.1 no momento em que os resultados do amostrados de Gibbs para o modelo linear bayesiano forem apresentados.

3.3 Modelo de regressão linear hierárquico bayesiano

Em muitos casos para a descrição de fenômenos aleatórios complexos não é possível a declaração de um modelo em apenas uma "frase", como foi feito na Seção 3.6. Em diversas áreas do conhecimento é possível notar que existem dados com estrutura hierárquica como por exemplo na Atuária, onde modelos hierárquicos são formulados para análises que envolvem a teoria do risco coletivo (aplicados a seguros e previdência), na Demografia onde os modelos hierárquicos têm utilidade na modelagem da dinâmica populacional, ou mesmo em vário outros domínios de aplicação Estatística como, por exemplo, Avaliação de Desempenho, Curvas de Crescimento, Geoestatística, etc. Migon (2008) [10]

O avanço dos métodos estatísticos em conjunto com o avanço exponencial da tecnologia tem tornado possível a elaboração de modelos altamente estruturados para descrever da maneira mais realista possível tais eventos dos quais muitas vezes os dados se distribuem de maneira diferente e em diferentes níveis

A formulação geral para os modelos hierárquicos utilizando a abordagem bayesiana com o conceito de permutabilidade de Finetti (1972) foi apresentado primeiramente por Lindley e Smith (1972) [11] quando foi mostrado que as estimativas bayesianas podem ser por vezes mais concentradas do que as estimativas da abordagem de mínimos quadrados.

A estrutura hierárquica pode ser concebida de maneiras diferentes como apresentado em Migon (2008) [10], quando existe hierarquia na variável resposta ou no caso da declaração da distribuição a priori, pois em ambos os casos apresenta-se unidades de análise em diferentes níveis.

A abordagem que será utilizada a seguir envolve um exemplo do conceito de priori hierárquica que é essencial na definição dos modelos lineares hierárquicos, Migon e Gamerman (1999) [10] argumentam que este procedimento pode ser utilizado para facilitar sua especificação e descrevem como construir a distribuição priori em estágios, combinando informações estruturais (para divisão dos estágios) com informações puramente subjetivas (para a especificação de cada estágio).

Sendo assim, como no modelo de regressão linear simples bayesiano calculado na Seção 3.6, estes modelos estocasticamente complexos novamente irão demandar o uso de métodos númericos eficientes para a integração e otimização.

Considere que a variável aleatória $Y_{i,j}$ que representa a variável resposta da i-ésima observação ao longo de $T = 5$ intervalos de tempo e seja a variável aleatória X_j numero de dias decorridos ao longo das 5 intervalos de tempo que será utilizada para explicar a variável $Y_{i,j}$ e, comumente, chamada de variável explicativa ou covariável.

Suponha que a população de interesse tenha distribuição normal com média $\alpha_i + \beta_i x_j$, sendo α_i e β_i desconhecidos e variância σ^2 desconhecida. Seja $\tau_c = \frac{1}{\sigma^2}$ o parâmetro chamado de precisão. O parâmetro α_i é o intercepto (ou coeficiente linear) e o β_i é o coeficiente angular. Além disso, suponha que as unidades dessa população sejam iid.

Dessa forma, tem-se que as unidades dessa população tem a seguinte distribuição:

$$Y_{i,j} \quad \sim \quad N(\alpha_i + \beta_i x_j, \tau_c^{-1}) \tag{3.12}$$

Note que o conceito de priori hierárquica será utilizada aqui na definição desse modelo linear hierárquico da seguinte maneira:

$$\begin{aligned} \alpha_i &\sim N(\alpha_c, \tau_\alpha^{-1}) \quad \tau_c \sim G(a_\tau, b_\tau) \\ \beta_i &\sim N(\beta_c, \tau_\beta^{-1}) \quad \tau_\alpha \sim G(a_\alpha, b_\alpha) \\ \alpha_c &\sim N(m_\alpha, V_\alpha) \quad \tau_\beta \sim G(a_\beta, b_\beta) \\ \beta_c &\sim N(m_\beta, V_\beta) \end{aligned} \tag{3.13}$$

onde $m_\alpha, V_\alpha, m_\beta, V_\beta, a_\tau, b_\tau, a_\alpha, b_\alpha, a_\beta, b_\beta$ são parâmetros conhecidos.

Obtendo-se uma amostra de tamanho n, pode-se inferir sob os parâmetros desconhecidos, $\boldsymbol{\theta} = (\alpha_i, \beta_i, \tau_c, \alpha_c, \beta_c, \tau_\alpha, \tau_\beta)$ através da distribuição a posteriori e para obter essa distribuição faz-se necessário calcular novamente a função de verossimilhança deste modelo. Considere então:

$$\boldsymbol{Y} = Y_{i,j} \text{ ; onde: } i = 1, ..., n \text{ e } j = 1, ..., T \tag{3.14}$$

$$\boldsymbol{\alpha} = \alpha_1, ..., \alpha_n \tag{3.15}$$

$$\boldsymbol{\beta} = \beta_1, ..., \beta_n \tag{3.16}$$

onde, o vetor de parâmetros desconhecidos será:

$$\boldsymbol{\theta} = \boldsymbol{\alpha}, \boldsymbol{\beta}, \alpha_c, \beta_c, \tau_c, \tau_\alpha, \tau_\beta$$

Logo, a função de verossimilhança será:

$$p(\boldsymbol{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \tau_c) = \prod_{i=1}^{n}\prod_{j=1}^{n} p(y_{i,j}|\alpha_i, \beta_i, \tau_c) \tag{3.17}$$

$$= (\frac{\tau}{2\pi})^{\frac{nT}{2}} exp\{-\frac{\tau}{2}\sum_{i=1}^{n}\sum_{j=1}^{T}(y_{i,j} - \alpha_i - \beta_i x_j)^2\} \tag{3.18}$$

e seja $\theta_{-\mu}$ o vetor após excluir o elemento μ desse vetor.

Considerando o conceito de priori hierárquica cujo os parâmetros sejam independentes e que possuam as distribuições de probabilidade apresentadas em 4.2, tem-se que a distribuição conjunta a priori possui a seguinte forma:

$$p(\boldsymbol{\theta}) = \prod_{i=1}^{n}[p(\alpha_i|\alpha_c, \tau_\alpha)p(\beta_i|\beta_c, \tau_\beta)]p(\alpha_c)p(\beta_c)p(\tau)p(\tau_\alpha)p(\tau_\beta) \tag{3.19}$$

$$\propto \tau_\alpha^{\frac{n}{2}} exp\{-\frac{\tau_\alpha}{2}\sum_{i=1}^{n}(\alpha_i - \alpha_c)^2 \tau_\beta^{\frac{n}{2}} exp\{-\frac{\tau_\beta}{2}\sum_{i=1}^{n}(\beta_i - \beta_c)^2\} \tag{3.20}$$

$$\times \quad exp\{-\frac{1}{2V_\alpha}(\alpha_c - m_\alpha)^2\} exp\{-\frac{1}{2V_\beta}(\beta_c - m_\beta)^2\} \tag{3.21}$$

$$\times \quad \tau_c^{a_\tau - 1} exp\{-\tau_c b_\tau\} \tau_\alpha^{a_\alpha - 1} exp\{-\tau_\alpha b_\alpha\} \tau_\beta^{a_\beta - 1} exp\{-\tau_\beta b_\beta\} \tag{3.22}$$

Portanto, combinando a função de verossimilhança com a distribuição a priori, obtem-se que a distribuição a posteriori é proporcional a:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{3.23}$$

Assim como em 3.9, essa distribuição é multivariada e não possuirá uma forma análitica conhecida, sendo assim serão utilizados métodos de MCMC, descritos na Subseção 3.1.2, para se obter amostras dessa distribuição. Faz-se necessário obter as DCCP de α_i, β_i e τ_c, α_c, β_c e τ_α, τ_β, portanto veja os cálculos dessas distribuições a seguir:

DCCP de τ_c:

$$\begin{aligned} p(\tau_c|\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\beta},\alpha_c,\beta_c,\tau_\alpha,\tau_\beta) &\propto \tau_c^{\frac{nT}{2}} exp\{-\frac{\tau_c}{2}\sum_{i=1}^{n}\sum_{j=1}^{T}(y_{i,j}-\alpha_i-\beta_i x_j)^2\} &(3.24)\\ &\times \tau_c^{a_\tau-1}exp\{-b_\tau\tau_c\} &(3.25)\end{aligned}$$

Logo,

$$\tau_c|\boldsymbol{y},\theta_{-\tau_c} \sim G(\frac{nT}{2}+a_\tau, b_\tau+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{T}(y_{i,j}-\alpha_i-\beta_i x_j)^2) \quad (3.26)$$

DCCP de α_i:

$$\begin{aligned} p(\alpha_i|\boldsymbol{y},\theta_{-\alpha_i}) &\propto exp\{-\frac{\tau_c}{2}\sum_{j=1}^{T}(y_{i,j}-\alpha_i-\beta_i x_j)^2\}\exp\{-\frac{\tau_\alpha}{2}(\alpha_i-\alpha_c)^2\} &(3.27)\\ &\propto exp\{-\frac{\tau_c}{2}\sum_{j=1}^{T}(-2y^*_{i,j}\alpha_i+\alpha_i^2)-\frac{\tau_\alpha}{2}(\alpha_i^2-2\alpha_i\alpha_c)\} &(3.28)\\ &\propto exp\{-\frac{1}{2}[(\tau_c T+\tau_\alpha)\alpha_i^2-2\alpha_i(\tau_c\sum_{j=1}^{T}y^*_{i,j}+\tau_\alpha\alpha_c)]\} &(3.29)\\ &\propto exp\{-\frac{1}{2}(\tau_c T+\tau_\alpha)[\alpha_i^2-2\alpha_i(\tau_c T+\tau_\alpha)^{-1}(\tau_c\sum_{j=1}^{T}y^*_{i,j}+\tau_\alpha\alpha_c)]\} &(3.30)\\ & &(3.31)\end{aligned}$$

logo,

$$\alpha_i|\boldsymbol{y},\theta_{-\alpha_i} \sim N((\tau_c T+\tau_\alpha)^{-1}(\tau_c\sum_{j=1}^{T}y^*_{i,j}+\tau_\alpha\alpha_c),(\tau_c T+\tau_\alpha)^{-1}) \quad (3.32)$$

DCCP de α_c:

$$
\begin{aligned}
p(\alpha_c|\boldsymbol{y},\theta_{\alpha_c}) &\propto exp\{-\frac{\tau_\alpha}{2}\sum_{i=1}^{n}(\alpha_i-\alpha_c)^2\}\exp\{-\frac{1}{2V_\alpha}(\alpha_c-m_\alpha)^2\} && (3.33)\\
&\propto exp\{-\frac{\tau_\alpha}{2}\sum_{i=1}^{n}(-2\alpha_i\alpha_c+\alpha_c^2)-\frac{1}{2V_\alpha}(\alpha_c^2-2m_\alpha\alpha_c)\} && (3.34)\\
&\propto exp\{-\frac{1}{2}[(\tau_\alpha n+\frac{1}{V\alpha})\alpha_c^2-2\alpha_c(\tau_\alpha\sum_{i=1}^{n}\alpha_i+\frac{1}{V_\alpha}m_\alpha)]\} && (3.35)\\
&\propto exp\{-\frac{1}{2}(\tau_\alpha n+\frac{1}{V_\alpha})[\alpha_c^2-2\alpha_c(\tau_\alpha n+\frac{1}{V_\alpha})^{-1}(\tau_\alpha\sum_{i=1}^{n}\alpha_i+\frac{m_\alpha}{V_\alpha})]\} && (3.36)\\
& && (3.37)
\end{aligned}
$$

logo,

$$\alpha_c|\boldsymbol{y},\theta_{-\alpha_c} \sim N((\tau_\alpha n+\frac{1}{V_\alpha})^{-1}(\tau_\alpha\sum_{i=1}^{n}\alpha_i+\frac{m_\alpha}{V_\alpha}),(\tau_\alpha n+\frac{1}{V_\alpha})^{-1}) \quad (3.38)$$

DCCP de τ_α:

$$p(\tau_\alpha|\boldsymbol{y},\theta_{-\tau_\alpha}) \propto \tau_\alpha^{\frac{n}{2}}exp\{-\frac{\tau_\alpha}{2}\sum_{i=1}^{n}(\alpha_i-\alpha_c)^2\}\tau_\alpha^{a_\alpha-1}exp\{-b_\alpha\tau_\alpha\} \quad (3.39)$$

Logo,

$$\tau_\alpha|\boldsymbol{y},\theta_{-\tau_\alpha} \sim G(\frac{n}{2}+a_\alpha,b_\alpha+\frac{1}{2}\sum_{i=1}^{n}(\alpha_i-\alpha_c)^2) \quad (3.40)$$

DCCP de β_i:

$$p(\beta_i|\boldsymbol{y},\theta_{-\beta_i}) \propto exp\{-\frac{\tau_c}{2}\sum_{j=1}^{T}(y_{i,j}-\alpha_i-\beta_i x_j)^2\}\exp\{-\frac{\tau_\beta}{2}(\beta_i-\beta_c)^2\} \quad (3.41)$$

$$\propto exp\{-\frac{\tau_c}{2}\sum_{j=1}^{T}(\beta_i^2 x_j^2 - 2y_{i,j}\beta_i x_j + 2\alpha_i\beta_i x_j) - \frac{\tau_\beta}{2}(\beta_i^2 - 2\beta_i\beta_c)\} \quad (3.42)$$

$$\propto exp\{-\frac{\tau_c}{2}(\beta_i^2\sum_{j=1}^{T}x_j^2 - 2\beta_i\sum_{j=1}^{T}y_{i,j}x_j + 2\alpha_i\beta_i\sum_{j=1}^{T}x_j) - \frac{\tau_\beta}{2}(\beta_i^2 - 2\beta_i\beta_c)\} \quad (3.43)$$

$$\propto exp\{-\frac{1}{2}(\tau_c\beta_i^2\sum_{j=1}^{T}x_j^2 - 2\tau_c\beta_i\sum_{j=1}^{T}y_{i,j}x_j + 2\tau_c\alpha_i\beta_i\sum_{j=1}^{T}x_j) + \tau_\beta\beta_i^2 - 2\tau_\beta\beta_i\beta_c)\} \quad (3.44)$$

$$\propto exp\{-\frac{1}{2}(\tau_c\sum_{j=1}^{T}x_j^2+\tau_\beta)\beta_i^2 - 2\beta_i(\tau_c\sum_{j=1}^{T}y_{i,j}x_j - \tau_c\alpha_i\sum_{j=1}^{T}x_j + \tau_\beta\beta_c)\} \quad (3.45)$$

logo,

$$\beta_i|\boldsymbol{y},\theta_{-\beta_i} \sim N((\tau_c\sum_{j=1}^{T}x_j^2+\tau_\beta)^{-1}(\tau_c\sum_{j=1}^{T}y_{i,j}x_j - \tau_c\alpha_i\sum_{j=1}^{T}x_j + \tau_\beta\beta_c), (\tau_c\sum_{j=1}^{T}x_j^2+\tau_\beta)^{-1}) \quad (3.46)$$

DCCP de β_c:

$$p(\beta_c|\boldsymbol{y},\theta_{\beta_c}) \propto exp\{-\frac{\tau_\beta}{2}\sum_{i=1}^{n}(\beta_i-\beta_c)^2\}\exp\{-\frac{1}{2V_\beta}(\beta_c-m_\beta)^2\} \quad (3.47)$$

$$\propto exp\{-\frac{\tau_\beta}{2}\sum_{i=1}^{n}(-2\beta_i\beta_c+\beta_c^2) - \frac{1}{2V_\beta}(\beta_c^2 - 2m_\beta\beta_c)\} \quad (3.48)$$

$$\propto exp\{-\frac{1}{2}[(\tau_\beta n + \frac{1}{V\beta})\beta_c^2 - 2\beta_c(\tau_\beta\sum_{i=1}^{n}\beta_i + \frac{1}{V_\beta}m_\beta)]\} \quad (3.49)$$

$$\propto exp\{-\frac{1}{2}(\tau_\beta n + \frac{1}{V_\beta})[\beta_c^2 - 2\beta_c(\tau_\beta n + \frac{1}{V_\beta})^{-1}(\tau_\beta\sum_{i=1}^{n}\beta_i + \frac{m_\beta}{V_\beta})]\} \quad (3.50)$$

$$(3.51)$$

logo,

$$\beta_c|\boldsymbol{y},\theta_{-\beta_c} \sim N((\tau_\beta n + \frac{1}{V_\beta})^{-1}(\tau_\beta\sum_{i=1}^{n}\beta_i + \frac{m_\beta}{V_\beta}), (\tau_\beta n + \frac{1}{V_\beta})^{-1}) \quad (3.52)$$

DCCP de τ_β:

$$p(\tau_\beta|\boldsymbol{y},\theta_{-\tau_\beta}) \quad \propto \quad \tau_\beta^{\frac{n}{2}} exp\{-\frac{\tau_\beta}{2}\sum_{i=1}^{n}(\beta_i-\beta_c)^2\}\tau_\beta^{a_\beta-1}exp\{-b_\beta\tau_\beta\} \tag{3.53}$$

Logo,

$$\tau_\beta|\boldsymbol{y},\theta_{-\tau_\beta} \quad \sim \quad G(\frac{n}{2}+a_\beta, b_\beta+\frac{1}{2}\sum_{i=1}^{n}(\beta_i-\beta_c)^2) \tag{3.54}$$

Como pode ser visto, diferente da modelagem linear simples, a modelagem hierárquica exige que os dados sejam estruturados de forma agrupada em diferentes níveis assim como foi calculado as DCCPs. Em cada nível, as variáveis explicativas estão associadas às outras variáveis dentro do mesmo nível e, possivelmente, às variáveis de níveis inferiores, de tal modo que os níveis mais baixos sejam independentes dos níveis mais altos.

Portanto, com esses resultados calculados será possível implementar os métodos de MCMC tanto para o modelo de regressão linear simples, como mencionados na Seção 3.6 anterior, quanto para o modelo de regressão linear hierárquico apresentado nesta seção.

4 Análise dos Resultados

Neste capítulo, dados artificiais serão gerados e, em seguida, serão ajustados um modelo de regressão linear simples e um modelo linear hierárquico para avaliar o procedimento de inferência empregado. A seção 4.1 contém os resultados para o modelo de regressão linear simples e a Seção 4.1.3 contém os resultados para o modelo linear hierárquico.

4.1 Modelo de regressão linear simples

A Seção 4.1.1 conterá os resultados para os dados simulados e em seguida serão apresentados os dados reais seguindo o modelo de regressão linear simples conforme descrito na Seção 3.2.

4.1.1 Dados simulados

Para o estudo do modelo de regressão linear primeiramente foi utilizado um conjunto de dados simulados conforme o modelo proposto em 3.7, utilizando $N = 1000$, $\beta_0 = 1$, $\beta_1 = 0,5$, $\tau = 2$ e X_i $N(0,1)$ e em seguida os parâmetros deste modelo, denotados por $\boldsymbol{\theta} = (\beta_0, \beta_1, \tau)$ foram estimados usando o paradigma Bayesiano.

Para a estimação foram utilizados os seguintes hiperparâmetros : $m_0 = m_1 = 0$, $\sigma_0^2 = \sigma_1^2 = 100$, $a = 0,1$ e $b = 0,1$. O tamanho da cadeia foi de 30000 simulações e o "burn-in"considerado após o ajuste foi de 15000. A figura 2 apresenta os histogramas junto com as densidades de três cadeias obtidas ao se inicializar o amostrador em pontos diferentes de todos os parâmetros contidos em $\boldsymbol{\theta}$ e uma linha vermelha indicará o valor do real parâmetro utilizado para estimar a cadeia.

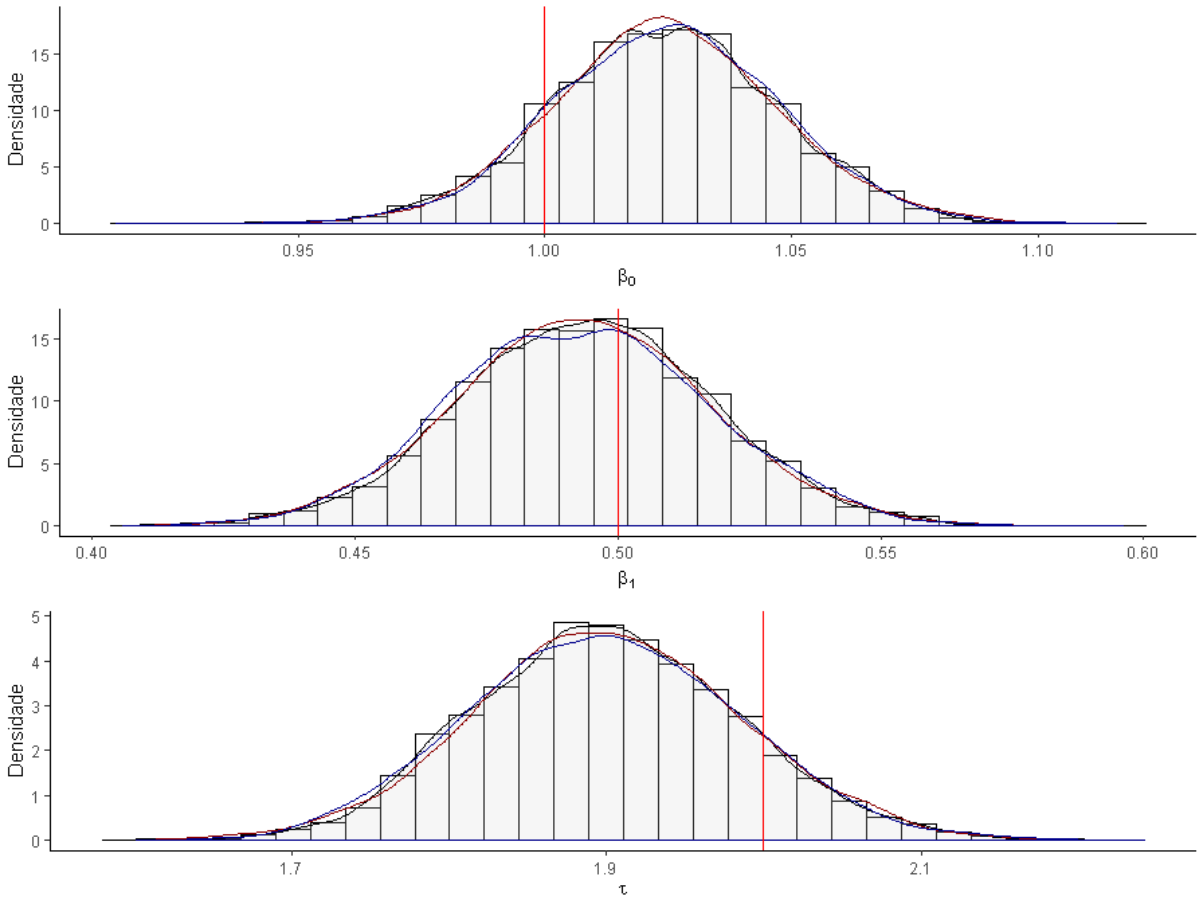

Figura 2: Histogramas e densidades das três últimas cadeias estimadas para modelo de regressão linear simples com dados simulados e destaque para o parâmetro populacional real

A Figura 3 apresenta os traços das cadeias dos parâmetros amostrados exibindo o intervalo de credibilidade com a linha pontilhada em azul e o valor verdadeiro do parâmetro em vermelho. Note que há indícios de convergência.

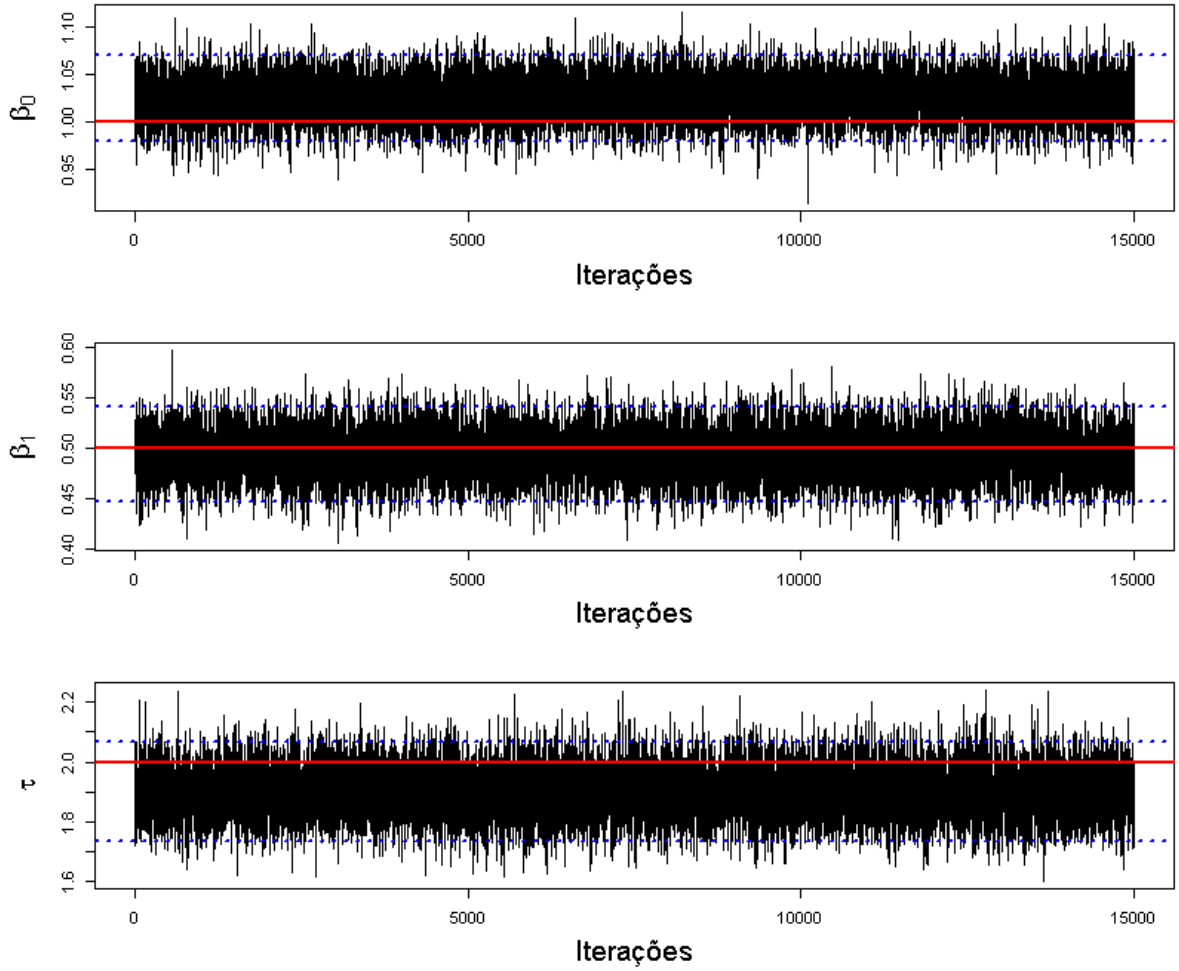

Figura 3: Cadeias estimadas para modelo de regressão linear simples com dados simulados com intervalos de credibilidade em azul e o parâmetro populacional real em vermelho

Ao observar cada uma das figuras é possível notar que todos os intervalos de credibilidade contêm o parâmetro populacional real utilizado para gerar a amostra.

A Figura 4 apresenta os gráficos de autocorrelação, que indicam se houve a influência dos "valores vizinhos"dos parâmetros amostrados. Note que parece haver independência entre as interações.

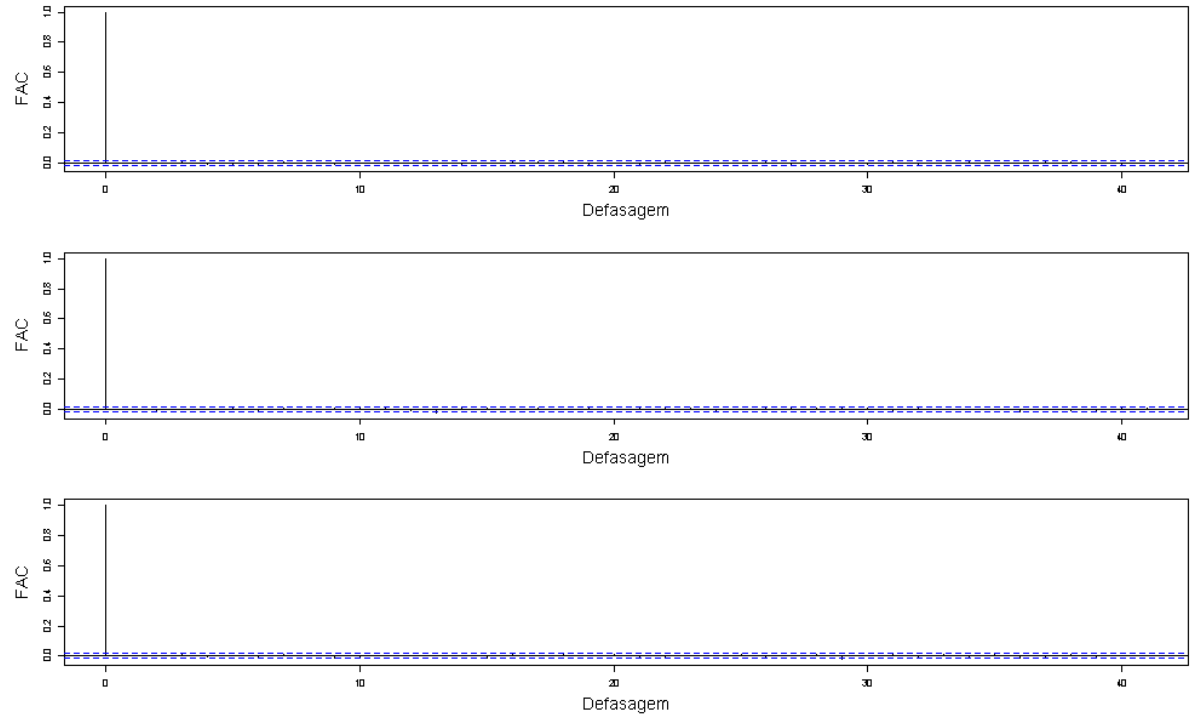

Figura 4: Gráficos de autocorrelação das cadeias estimadas para modelo de regressão linear simples com dados simulados para os parâmetros β_0, β_1 e τ

Segundo a figura é possível notar que nenhuma das cadeias apresentaram estimativas autocorrelacionada, o que somado às análises feitas anteriormente, já permite o estudo sobre as estimativas dos parâmetros através do algoritmo. A Tabela 1 apresenta os resumos a posteriori dos parâmetros amostrados.

Tabela 1: Tabela de estatísticas descritivas dos parâmetros do modelo ajustado para os dados simulados

Parâmetro	Média	Desv. Pad.	$2,5\%$	$97,5\%$	Parâmetro real
β_0	1,02*	0,02	0,98	1,07	1,00
β_1	0,49*	0,02	0,45	0,54	0,50
τ	1,90*	0,08	1,74	2,07	2,00

*: Não contém o zero no intervalo de 95% de credibilidade

Essa tabela mostra que nenhuma das estimativas contém o zero no intervalo de credibilidade e além disso como se trata de uma amostra simulada é possível comparar as estimativas com os valores reais que geraram a amostra e os valores estão muito próximos da média (todos eles estão incluídos no intervalo de credibilidade).

Agora que os resultados sob o paradigma bayesiano já foram conferidos será ajustado um modelo de regressão linear simples pelo método dos mínimos quadrados através da função "lm"(nativa do software [9]) sob o paradigma clássico para comparar com os resultados de um modelo de regressão linear simples sob o paradigma bayesiano utilizando os resultados calculados na Seção 3.2.

O modelo estimado para estes dados sob o paradigma da inferência clássica foi o seguinte: $\hat{y} = 1.0245x + 0,4933$, o que mostra que as estimativas de β_0 e β_1 foram muito parecidas com as estimativas sob o paradigma da inferência bayesiana. A figura 5 apresenta o gráfico de dispersão entre as variáveis da amostra simulada e as retas dos ajustes de ambos os modelos:

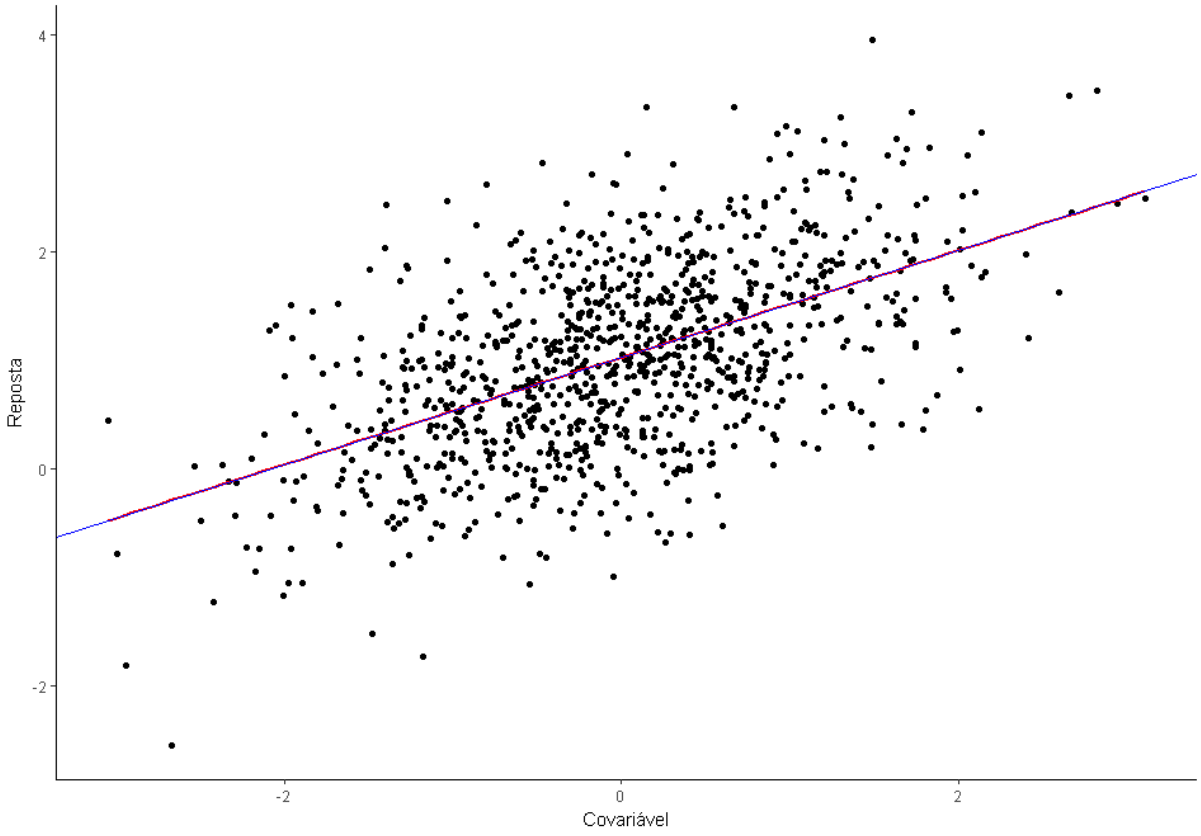

Figura 5: Relação entre a covariável e a variável resposta da cadeia simulada e reta do modelo linear clássico vs bayesiano com dados simulados

4.1.2 Dados reais

Agora que os resultados no algoritmo já foram conferidos e avaliados de maneira satisfatória utilizando os dados simulados, é a vez de fazer o ajuste para dados reais.

O conjunto de dados que será utilizado como exemplo foi disponibilizado por Ezekiel (1930)[8] e hoje faz parte do conjunto de banco de dados nativos do R (a base de dados pode ser obtida ao escrever "cars"no console). Os dados informam a velocidade dos carros e as distâncias tomadas para parar, esses dados foram registrados na década de 1920 e são de grande utilidade didática até os dias de hoje.

Considere que deseja-se modelar a velocidade dos carros de acordo com as distâncias tomadas para parar, portanto a variável resposta será a velocidade e a variável explicativa do modelo será a distância tomada para parar.

A figura 6 exibe os histogramas com as densidades de três cadeias obtidas ao se iniciar o amostrador em pontos diferentes de todos os parâmetros $\boldsymbol{\theta}$ mas dessa vez sem a linha vermelha que indicava o valor do parâmetro real pois agora ele é desconhecido.

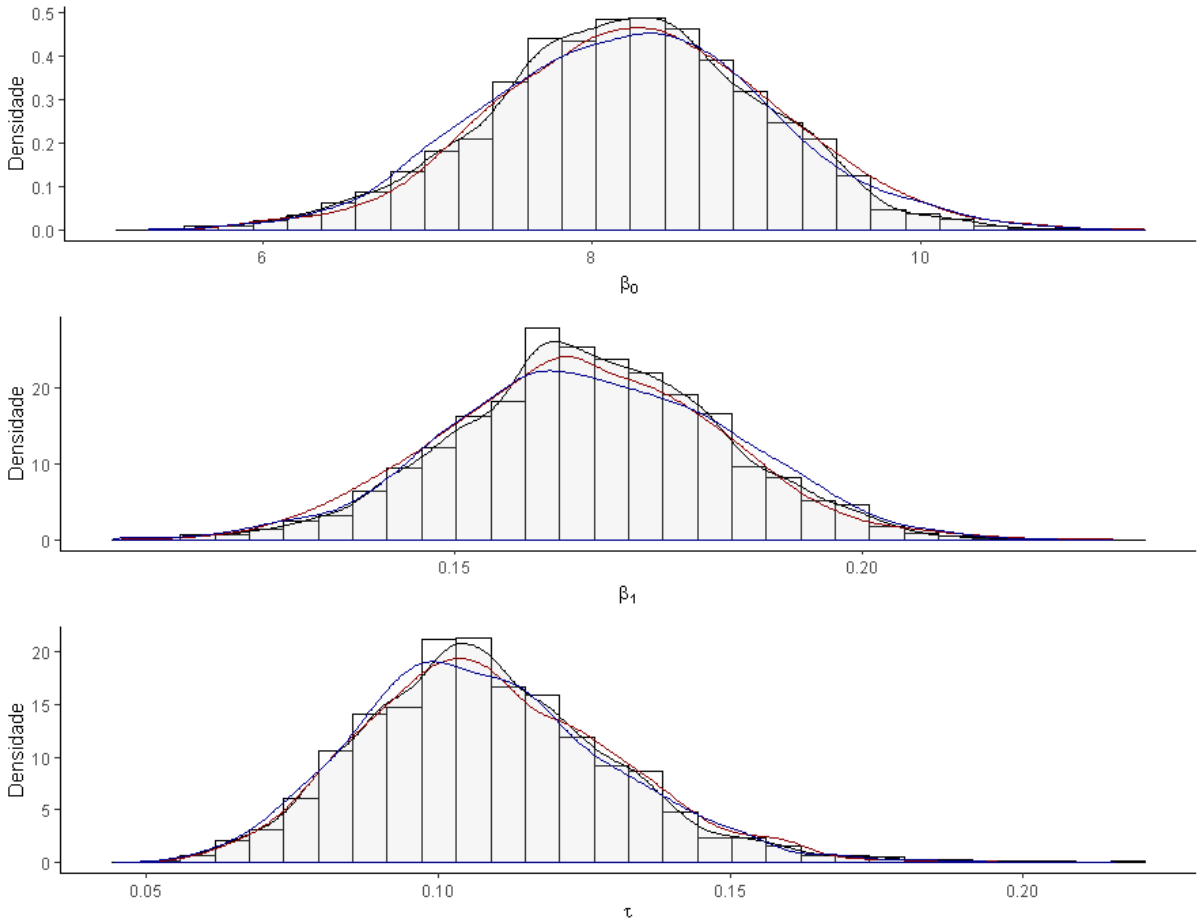

Figura 6: Histogramas e densidades das três últimas cadeias estimadas para modelo de regressão linear simples com base de dados cars

Nota-se que ambas as cadeias convergiram uma mesma distribuição e que as últimas três cadeias apresentaram valores próximos. A figura 7 apresenta os traços das cadeias dos parâmetros amostrados. Note que há indícios de convergência.

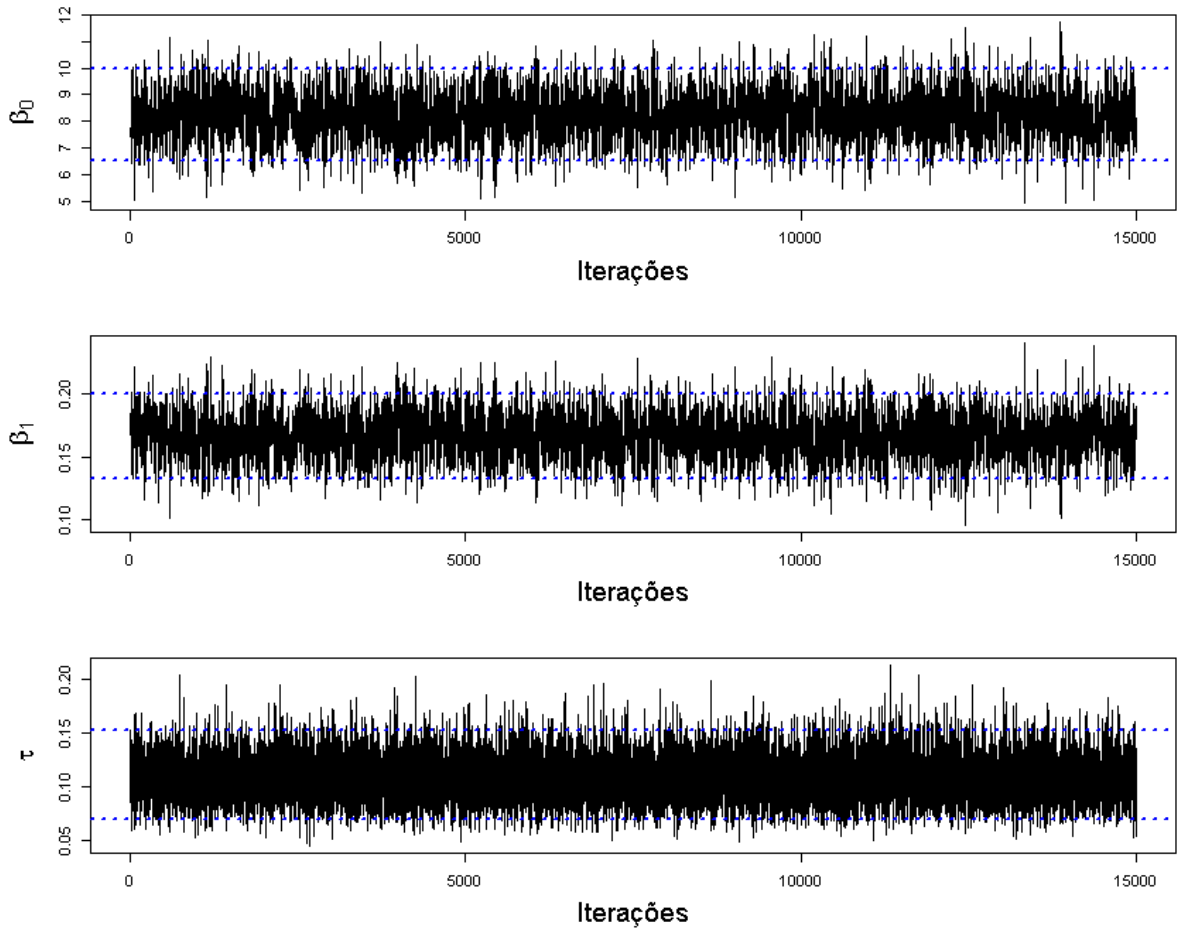

Figura 7: Cadeias estimadas para modelo de regressão linear simples com base de dados cars

A Figura 8 apresenta os gráficos de autocorrelação dos parâmetros amostrados.

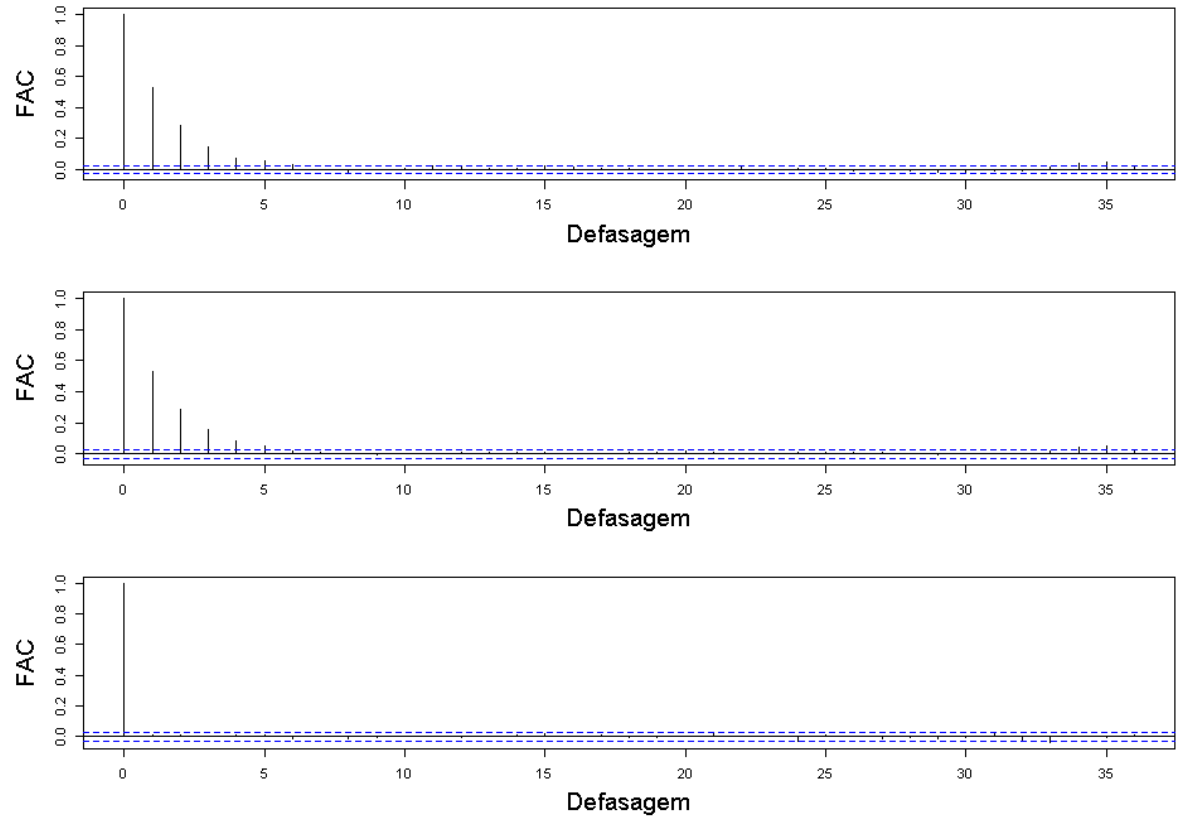

Figura 8: Gráficos de autocorrelação das cadeias estimadas para os respectivos parâmetros β_0, β_1 e τ do modelo de regressão linear simples com base de dados cars

É possível notar que apenas nas primeiras defasagens das cadeias das estimativas para os parâmetros β_0 e β_1 se apresentaram de forma autocorrelacionada e que a partir dessa defasagem o gráfico de autocorrelação se apresentou de forma desejável.

Como todas as características da cadeia gerada foram avaliadas de maneira satisfatória agora será possível conferir o ajuste dos parâmetros de maneira mais segura pois já foi constatada a convergência da cadeia, A Tabela 2 apresenta o resumo a posteriori dos parâmetros estimados da cadeia e note que esta tabela não conta com a coluna dos valores reais como no exemplo anterior.

Tabela 2: Resumo a posteriori dos parâmetros amostrados do modelo ajustado para os dados reais

Parâmetro	Média	Desv. Pad.	$2,5\%$	$97,5\%$
β_0	8,24*	0,84	6,57	9,92
β_1	0,17*	0,02	0,13	0,20
τ	0,11*	0,02	0,07	0,15

*: Não contém o zero no intervalo de 95% de credibilidade

Agora que os resultados sob o paradigma bayesiano já foram conferidos novamente será ajustado um modelo de regressão linear simples pelo método dos mínimos quadrados sob o paradigma clássico para comparar com os resultados do um modelo de regressão linear simples sob o paradigma bayesiano utilizando os resultados calculados na seção 3.2.

O modelo estimado sob este paradigma pode ser escrito da seguinte maneira: $\hat{y} = 8,2839x + 0,1656$, ou seja, os valores de β_0 e de β_1 novamente foram muito próximos dos parâmetros obtidos ao estimar sob o paradigma clássico.

A Figura 9 ilustra o gráfico de dispersão dos dados citados acima, com a intenção de exibir quanto uma variável é afetada por outra, onde no eixo vertical representa a velocidade do carro e no eixo horizontal a distância tomada para parar.

Além do comportamento das variáveis, neste gráfico é exibido também os resultados obtidos do ajuste ao se utilizar o método de mínimos quadrados (representada pela linha em vermelho) para estimar os parâmetros e o ajuste do modelo 3.7 ao se utilizar o método apresentado acima em 3.2 (representada pela linha azul).

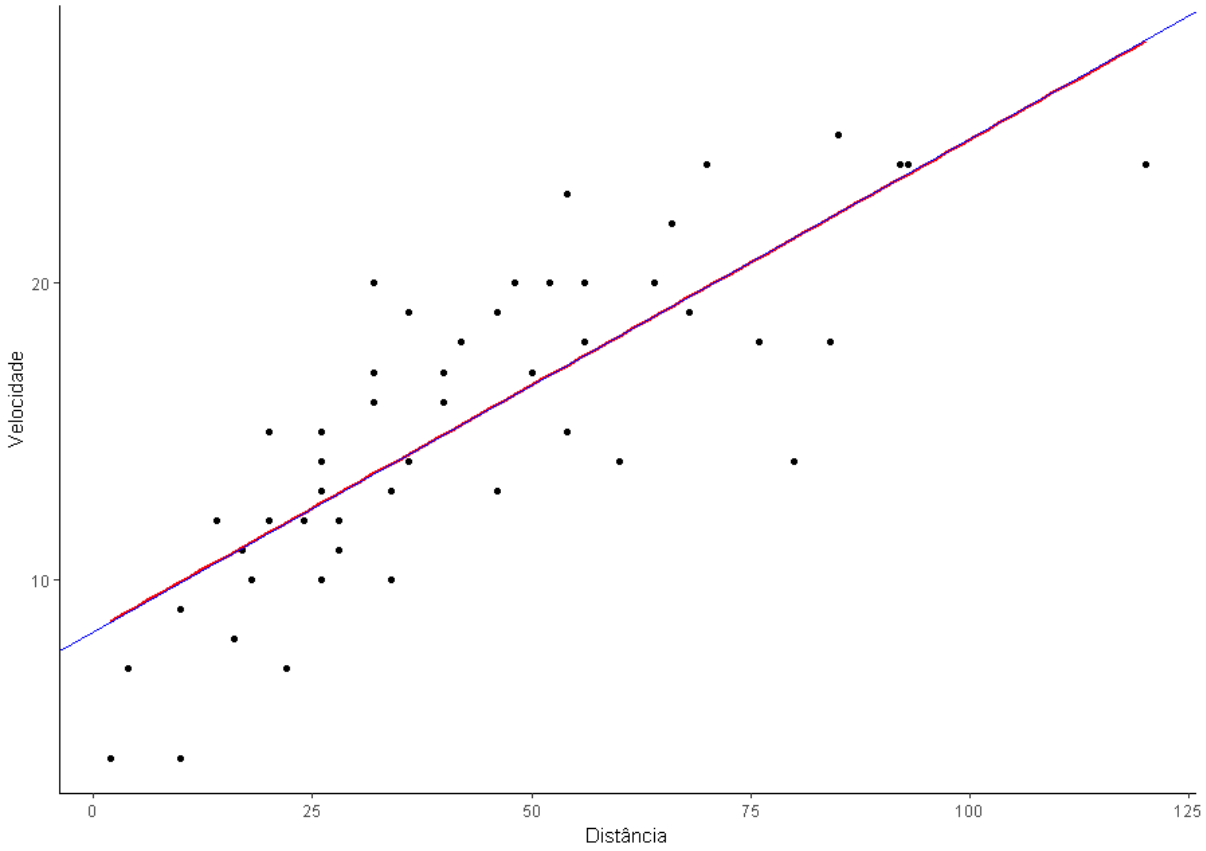

Figura 9: "Relação entre a covariável e a variável resposta da cadeia simulada com reta do modelo linear clássico vs bayesiano com base de dados cars

É possível notar que os coeficientes calculados foram muito parecidos, mesmo apresentando pequenas diferenças decimais no valor dos coeficientes ainda é possível notar que as retas estão basicamente soprepostas, ou seja, os valores estimados em ambas as abordagens foram praticamente os mesmos.

Apesar dos valores dos ajustes terem apresentado basicamente os mesmo resultados, a maneira de se conferir a qualidade do ajuste é diferente em ambas as abordagens. Enquanto sob o paradigma clássico o ajuste do modelo pode ser checado ao avaliar os pre-supostos quanto à distribuição dos resíduos, como recomenda Cordeiro e Demétrio (2008)[12], ao utilizar um método de MCMC faz-se necessário conferir também outros aspectos como por exemplo se houve convergência da cadeias além do comportamento das autocorrelações, vide Migon (2014)[3].

4.1.3 Modelo de regressão linear hierárquico bayesiano

Esta Seção conterá os resultados para os dados simulados seguindo o modelo de regressão linear hierárquico conforme descrito na Seção 3.3.

Todas as contas referentes ao ajuste deste modelo já foram apresentadas na Seção 3.3 na equação 3.12 e agora que todos esses resultados já estão prontos, é possível a implementação do algoritmo computacional. Os dados serão simulados conforme o comportamento dos dados e a estimação dos parâmetros do modelo hierárquico bayesiano e o comportamento da cadeia serão avaliados nessa seção.

Para essa abordagem, os parâmetros desconhecidos deste modelo serão:

$$\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \alpha_c, \beta_c, \tau_c, \tau_\alpha, \tau_\beta)$$

E como foi visto, a distribuição da variável $Y_{i,j}$ que corresponde a variável resposta da i-ésima observação no j-ésimo intervalo de tempo, do qual deseja-se estudar é:

$$Y_{i,j} \quad \sim \quad N(\alpha_i + \beta_i x_j, \tau_c^{-1}) \tag{4.1}$$

Onde $i = 1, ..., 30$ observações e $j = 1, ..., 5$ intervalos de tempo do acompanhamento do estudo.

O conceito de priori hierárquica será utilizado aqui para realizar a simulação dos dados, pois esses dados serão gerados conforme os parâmetros da declaração do modelo linear hierárquico de forma que seja possível recuperar esses parâmetros conhecidos.

Portanto, a seguir veja quais os parâmetros utilizados para gerar os dados e que serão recuperados após avaliar o ajuste do modelo conforme a metodologia proposta:

$$\begin{aligned} \alpha_i \sim N(\alpha_c, \tau_\alpha^{-1}) \quad & \tau_\alpha = \tfrac{1}{0{,}2} \quad \alpha_c = 20 \\ \beta_i \sim N(\beta_c, \tau_\beta^{-1}) \quad & \tau_\beta = \tfrac{1}{0{,}2} \quad \beta_c = 2 \qquad (4.2) \\ & \qquad\qquad\quad \tau_c = 1 \end{aligned}$$

$$(4.3)$$

onde $m_\alpha, V_\alpha, m_\beta, V_\beta, a_\tau, b_\tau, a_\alpha, b_\alpha, a_\beta, b_\beta$ são os parâmetros a priori conhecidos.

Uma amostra de tamanho 30 foi simulada a partir dessas informações. Para inferir sobre os parâmetros desconhecidos, $\boldsymbol{\theta} = (\alpha_i, \beta_i, \tau_c, \alpha_c, \beta_c, \tau_\alpha, \tau_\beta)$ através das distribuições condicionais completas a posteriori e avaliar o comportamento da cadeia

Neste exemplo, serão adotados os seguintes parâmetros a prioris conhecidos:

$$\begin{aligned} m_\alpha &= 0 & m_\beta &= 0 \\ V_\alpha &= \frac{1}{0,0001} & a_\tau &= 0,001 \\ V_\beta &= \frac{1}{0,0001} & b_\tau &= 0,001 \\ a_\alpha &= 0,001 & a_\beta &= 0,001 \\ b_\alpha &= 0,001 & b_\beta &= 0,001 \end{aligned}$$

E além disso o tamanho da cadeia foi de 150000 simulações e após o ajuste 75000 observações descartadas com a finalidade de avaliar o comportamento dos parâmetros após a convergência quando não estiverem mais correlacionados, essa técnica é chamada de "burnin" [10]. A seguir serão apresentados os resultados obtidos após o ajuste da cadeia mas antes de conferir se os parâmetros populacionais conhecidos que geraram a amostra foram recuperados com o ajuste e a implementação do algoritmo será necessário avaliar como foi o comportamento da cadeia novamente através de seus gráficos de densidade, seu gráfico de autocorrelação e se houve convergência na distribuição dos parâmetros amostrados pelo método MCMC.

Seguindo a mesma lógica do método de avaliação da cadeia utilizado na Seção 3.6, inicialmente será conferido o comportamento das cadeias com os histogramas junto com as densidades de três cadeias obtidas ao se inicializar o amostrador em pontos diferentes de todos os parâmetros do primeiro nível ($\tau_\alpha, \tau_\beta, \beta_c$ e α_c) pois elas irão determinar o quanto e em torno de qual valor as estimativas dos parâmetros α_i e β_i do segundo nível (que será avaliado em seguida) estão concentradas.

A figura 10 mostra o histograma junto com as densidades das três últimas cadeias dos parâmetros α_c, β_c, ,τ_α,τ_c e τ_β.

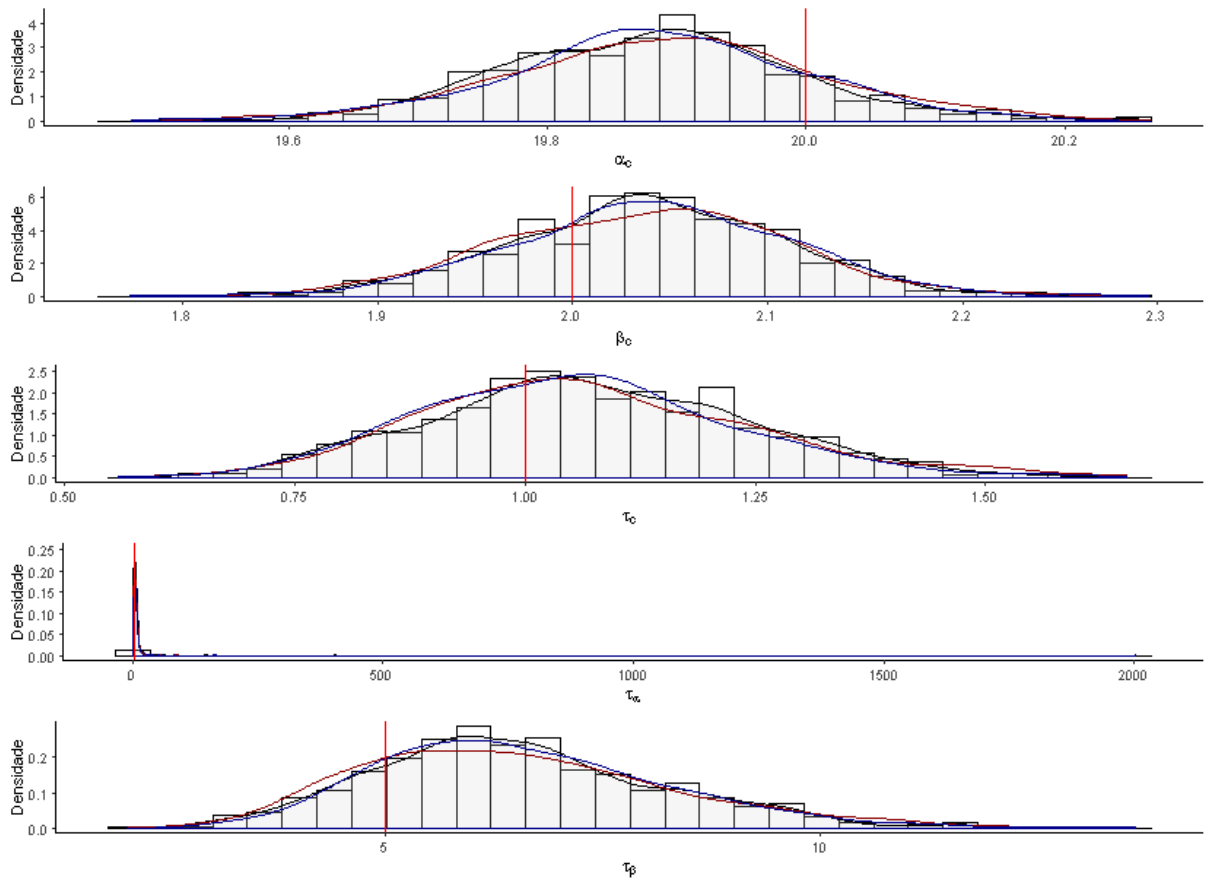

Figura 10: Histogramas e densidades das três últimas cadeias estimadas para o modelo de regressão hierárquico bayesiano com base de dados simulada

A Figura 11 apresenta os traços das cadeias dos parâmetros amostrados. Note que parece ter havido convergência.

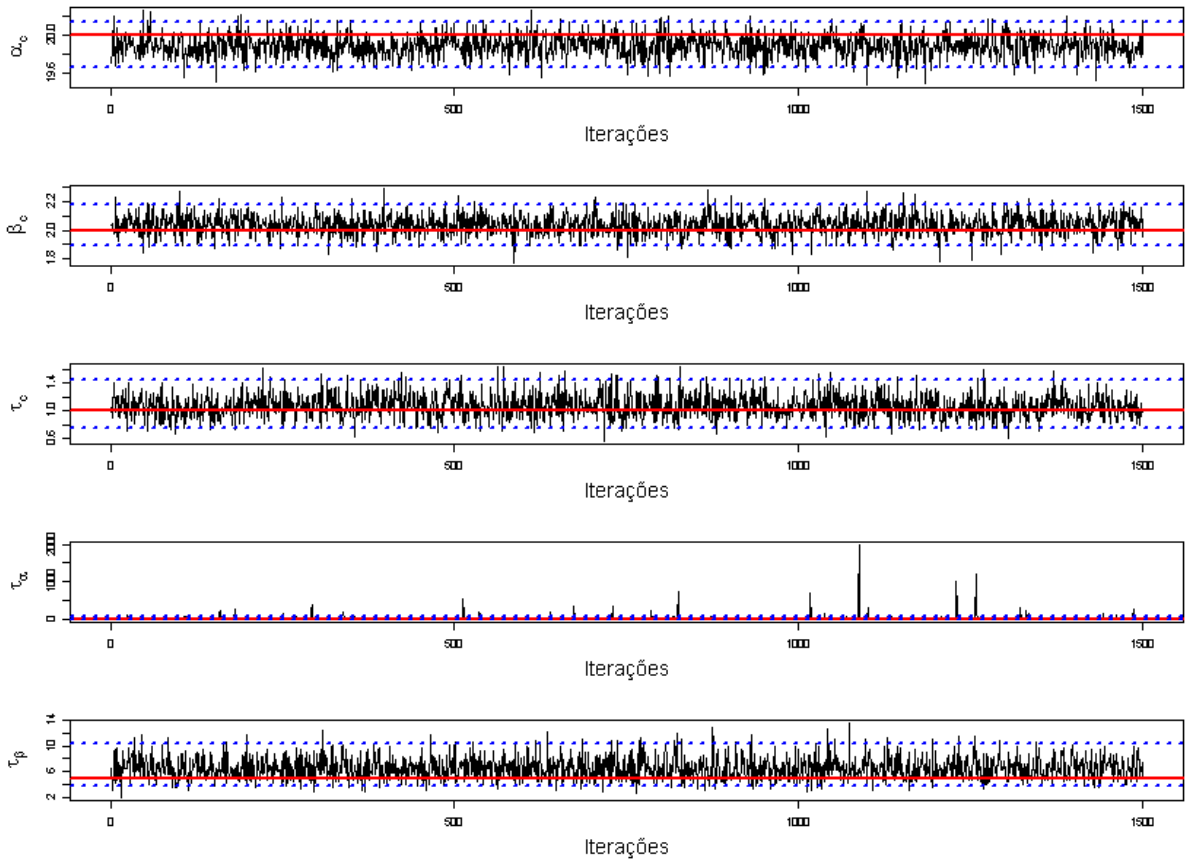

Figura 11: Cadeias estimadas para o modelo de regressão hierárquico bayesiano com base de dados simulada

Novamente a cadeia para o parâmetro τ_α se apresentou um pouco menos estável que as demais porém seus resultados serão melhor avaliados adiante.

A Figura 12 apresenta o gráfico de autocorrelação dos parâmetros do primeiro nível α_c, β_c, ,τ_α,τ_c e τ_β:

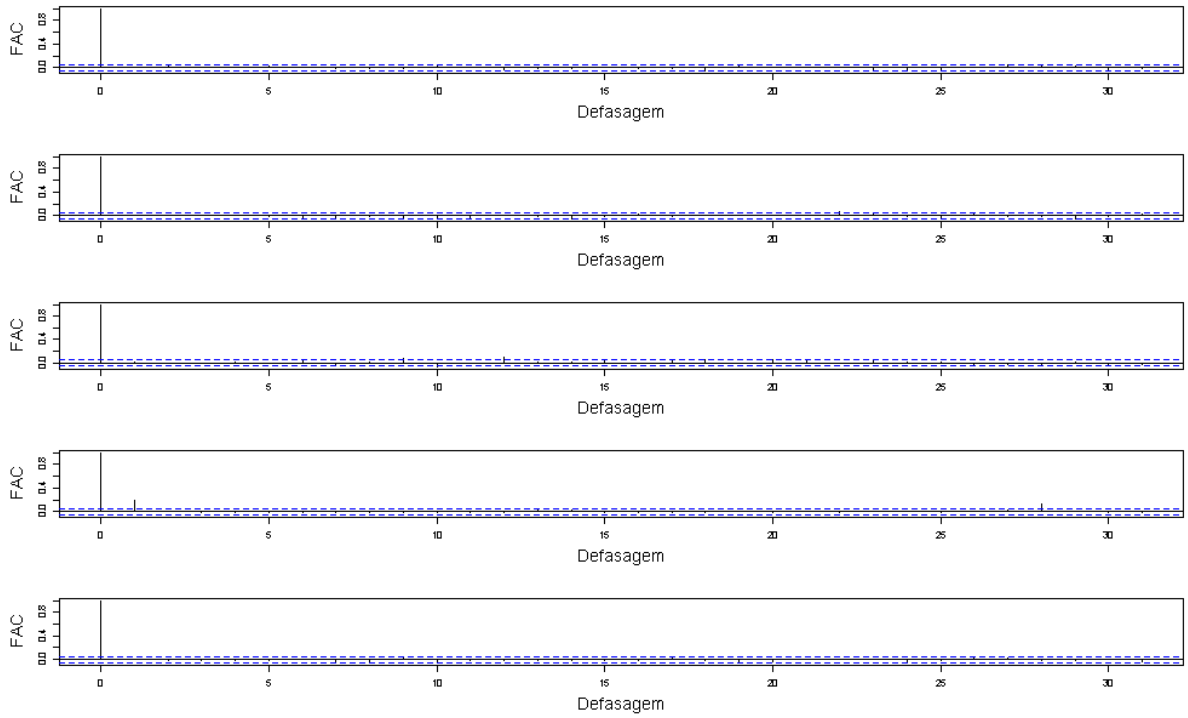

Figura 12: Gráficos de autocorrelação das cadeias estimadas para o os respectivos parâmetros α_c, β_c, ,τ_α,τ_c e τ_β do modelo de regressão hierárquico bayesiano com base de dados simulada

Por fim ao avaliar o gráfico de autocorrelação é possível notar que apenas as estimativas iniciais se apresentaram de forma autocorrelacionada.

Como os resultados gerais da convergência da cadeia já foram avaliados, nessa etapa também serão avaliadas as estimativas de cada um dos i-ésimos α_i e β_i correspondente ao segundo nível do modelo hierárquico.

As Figuras 13 e 14 apresentam as médias a posteriori e o resultado das médias e limites inferiores e superiores dos intervalos de credibilidade de 95% para as cadeias de α_i e para β_i estimadas incluindo o real valor estimado em azul e uma linha tracejada para os reais valores de α_c e β_c.

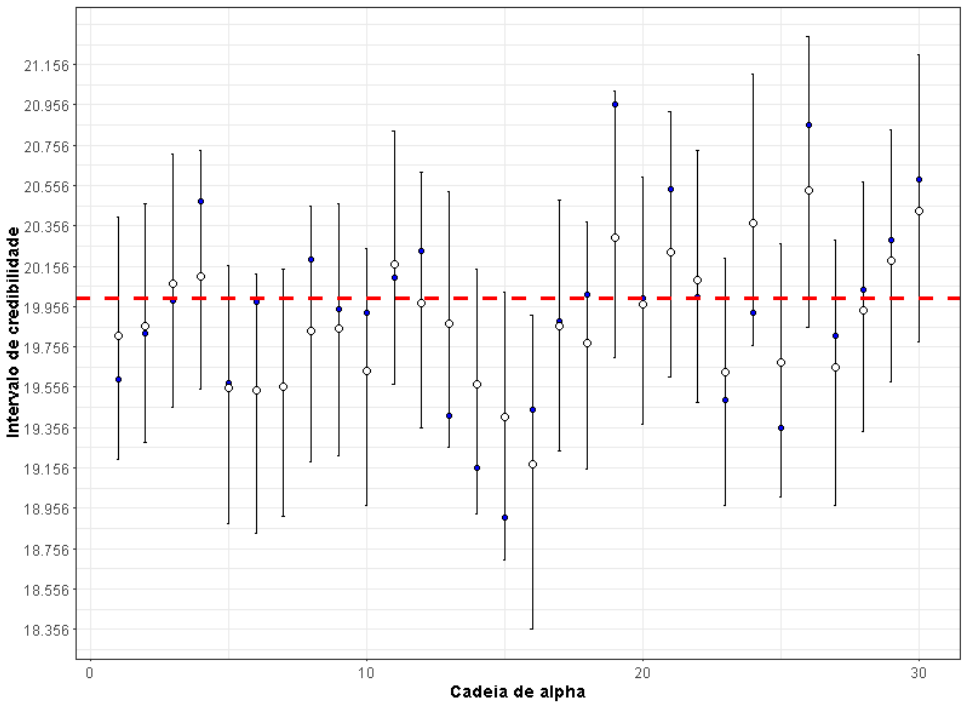

Figura 13: Médias e intervalos de credibilidade para a cadeia de α_i estimada incluindo o real valor estimado em azul e uma linha tracejada para o real valor de α_c

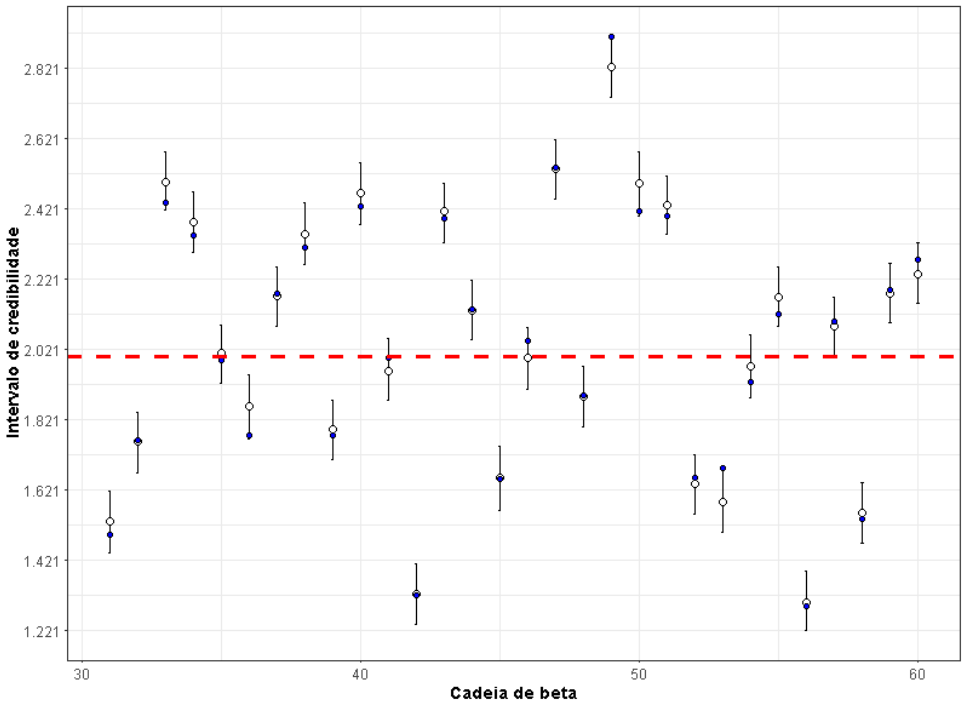

Figura 14: Médias e intervalos de credibilidade para a cadeia de β_i estimada incluindo o real valor estimado em azul e uma linha tracejada para o real valor de β_c

Todos os intervalos de credibilidade contêm o real valor populacional de interesse, o que sugere que as estimativas com a implementação deste algoritmo foram satisfatórias

Assim como nas cadeias anteriores, o comportamento das estimativas para o parâmetro incluído neste exemplo também se apresentou de forma satisfatória, e como todos os parâmetros foram estimados de forma razoavelmente boa a próxima etapa será conferir a Tabela 3 que apresenta os resumos a posteriori dos parâmetros amostrados

Tabela 3: Resumo a posteriori dos parâmetros do modelo ajustado para os dados reais

Parâmetro	Média	Desv. Pad.	$2,5\%$	$97,5\%$	Parâmetro real
α_c	19,65*	0,12	19,89	20,12	20
β_c	1,89*	0,07	2,04	2,19	2
τ_c	0,76*	0,17	1,06	1,41	1
τ_α	1,90*	79,30	16,79	86,14	5
τ_β	3,62*	1,71	6,58	10,42	5

*: Não contém o zero no intervalo de 95% de credibilidade

Mesmo com o alto desvio padrão registrado para a estimativa de τ_α nota-se que este valor não interferiu em todas as outras estimativas, que apresentaram bons resultados pois todas elas incluem o real valor populacional que gerou a amostra em seus intervalos de credibilidade.

5 Conclusão

O uso do algoritmo para simular os dados da implementação do modelo hierárquico bayesiano envolveu diversas etapas. Inicialmente foi necessária a revisão da literatura para a compreensão dos métodos que seriam utilizados na implementação do algoritmo bem como seu desenvolvimento. Essa pesquisa funcionou de maneira muito didática de forma que a cada semana a abordagem pudesse envolver maior grau de complexidade.

Os cálculos realizados para descobrir as distribuições posterioris dos parâmetros foram feitos em diversas passos até que todas as distribuições condicionais completas estivessem calculadas e bem definidas para a implementação do algoritmo.

Durante o estudo diversos valores os parâmetros a priori foram selecionados para que fosse possível avaliar a sensibilidade da qualidade da escolha da distribuição priori. Observou-se que valores elevados para variância a priori (também consideradas como "não informativas", fazendo uma analogia à modelos clássicos) obtiveram melhores ajustes atribuindo maior importância à informação provinda da amostra.

O estudo com dados simulados facilitou o entendimento do algoritmo pois foi possível notar com facilidade a inadequabilidade das escolhas das prioris, que resultavam em estimativas muito distante do parâmetro populacional que gerou a amostra.

Referências

[1] ROBERT, C. P.; CASELLA, G. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005. ISBN 0387212396.

[2] GAMERMAN, D.; LOPES, H. F. *Monte Carlo Markov Chain: Stochastic Simulation for Bayesian Inference*. Second. London: Chapman & Hall, 2006.

[3] MIGON, H. *Statistical Inference: An Integrated Approach*. 2. ed. [S.l.]: CRC Press: Taylor e Francis Group, 2014.

[4] JEFFREYS, H. *Theory of Probability*. 3rd ed.. ed. [S.l.]: Oxford Univ. Press, 1961.

[5] EHLERS, R. S. *Introdução a Inferência Bayesiana*. Second. [S.l.: s.n.], 2003.

[6] GEMAN S. E GEMAN, D. *Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images*. [S.l.]: IEEE, Transactions on Pattern Analysis and Machine Intelligence, 1990.

[7] GELFAND A. E. E SMITH, A. F. M. *Samping-based approaches to calculating marginal densities*. [S.l.]: Journal of the American Statistical Association, 1990.

[8] EZEKIEL, M. *Methods of Correlation Analysis*. [S.l.]: Wiley, 1930.

[9] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2017. Disponível em: <https://www.R-project.org/>.

[10] MIGON, A. D. P. S. e. A. M. S. H. S. *Modelos Hierárquicos e Aplicações*. Second. [S.l.]: ABE- ASSOCIAÇÃo BRASILEIRA DE ESTATÍSTICA, 2008.

[11] LINDLEY D. E SMITH, A. *Bayes estimates for the linear model*. B. [S.l.]: Journal of the Royal Statical Society, 1972.

[12] DEMÉTRIO, G. M. C. e C. G. *Modelos Lineares Generalizados e Extensões*. [S.l.: s.n.], 2008.