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.

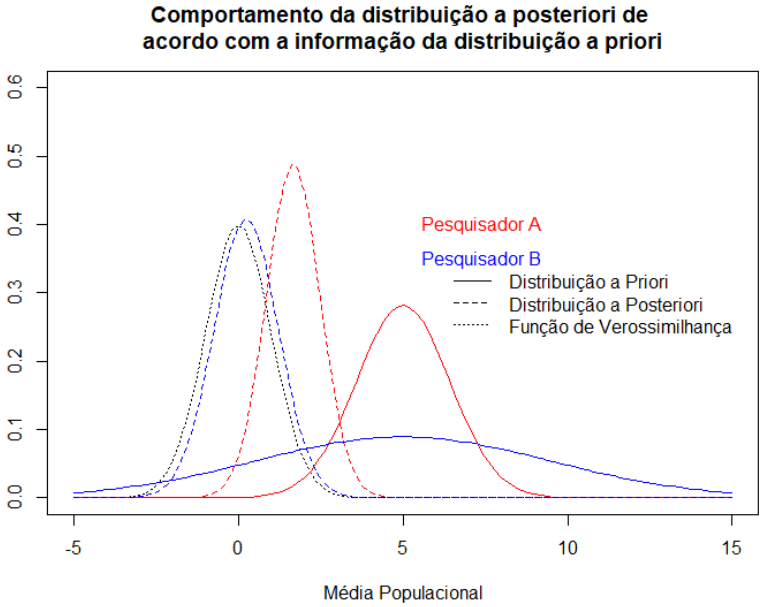

Figura 1: Comparando comportamento da distribuição a posteriori de acordo com a seleção da distribuição a priori

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.

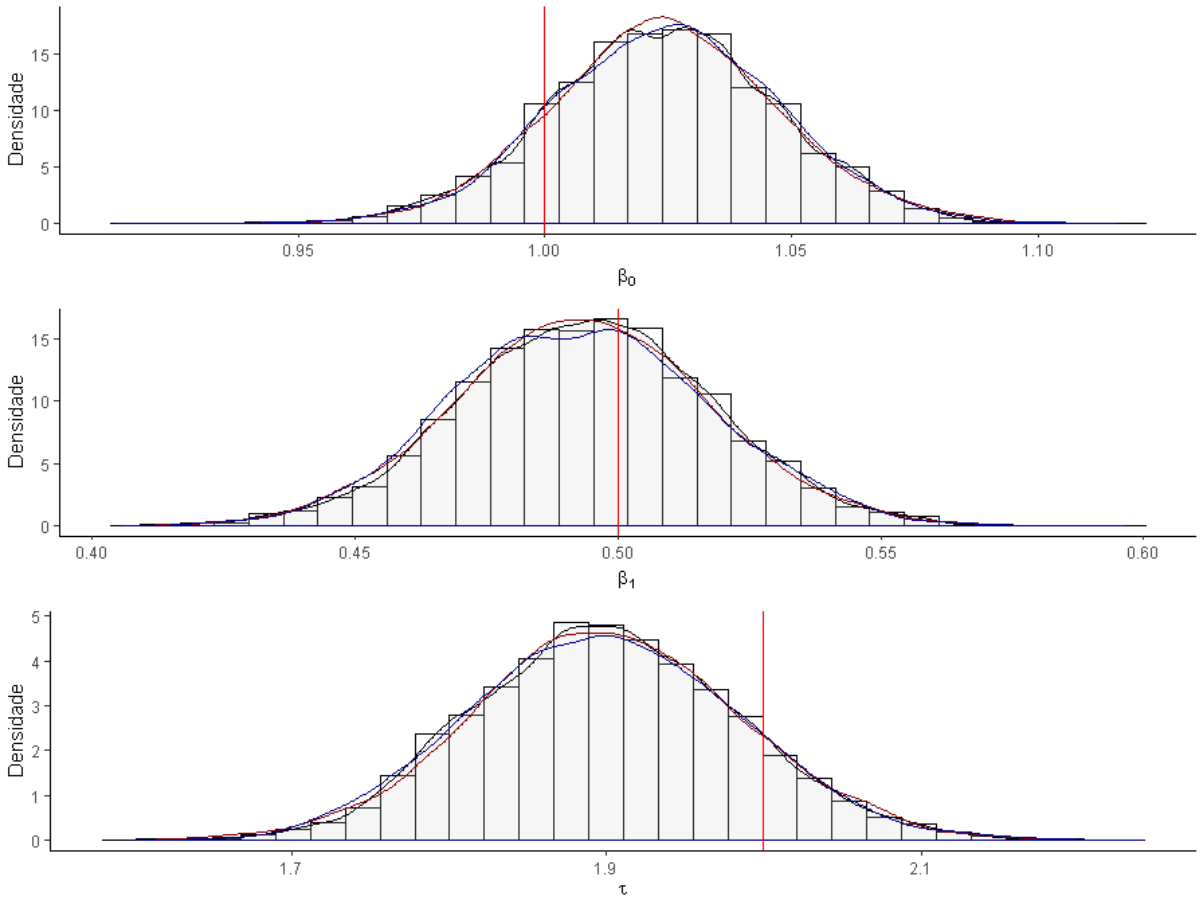

Figura 2: Histogramas e densidades das três últimas cadeias estimadas para modelo de regressão linear simples com dados simulados e destaque para o parâmetro populacional real

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.

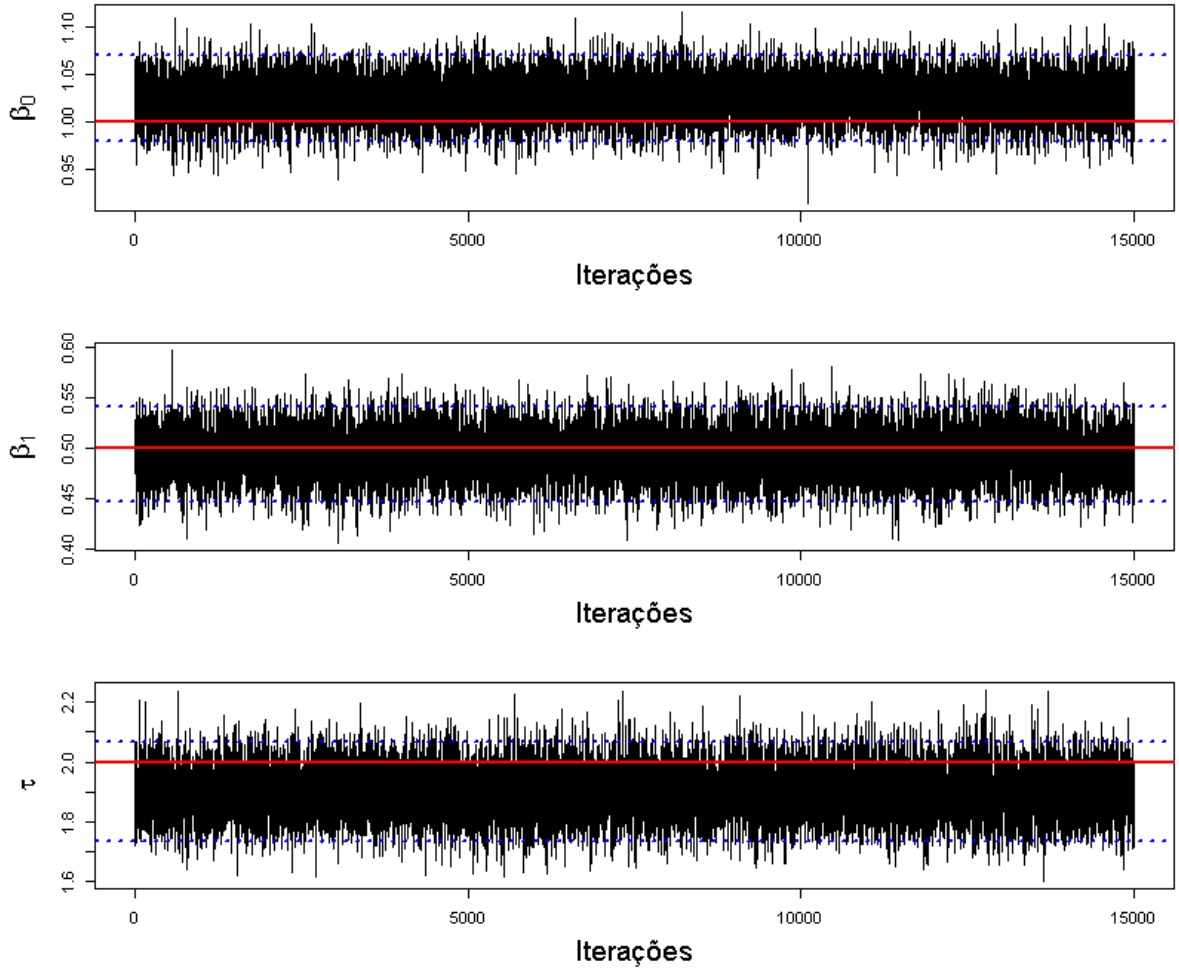

Figura 3: Cadeias estimadas para modelo de regressão linear simples com dados simulados com intervalos de credibilidade em azul e o parâmetro populacional real em vermelho

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.

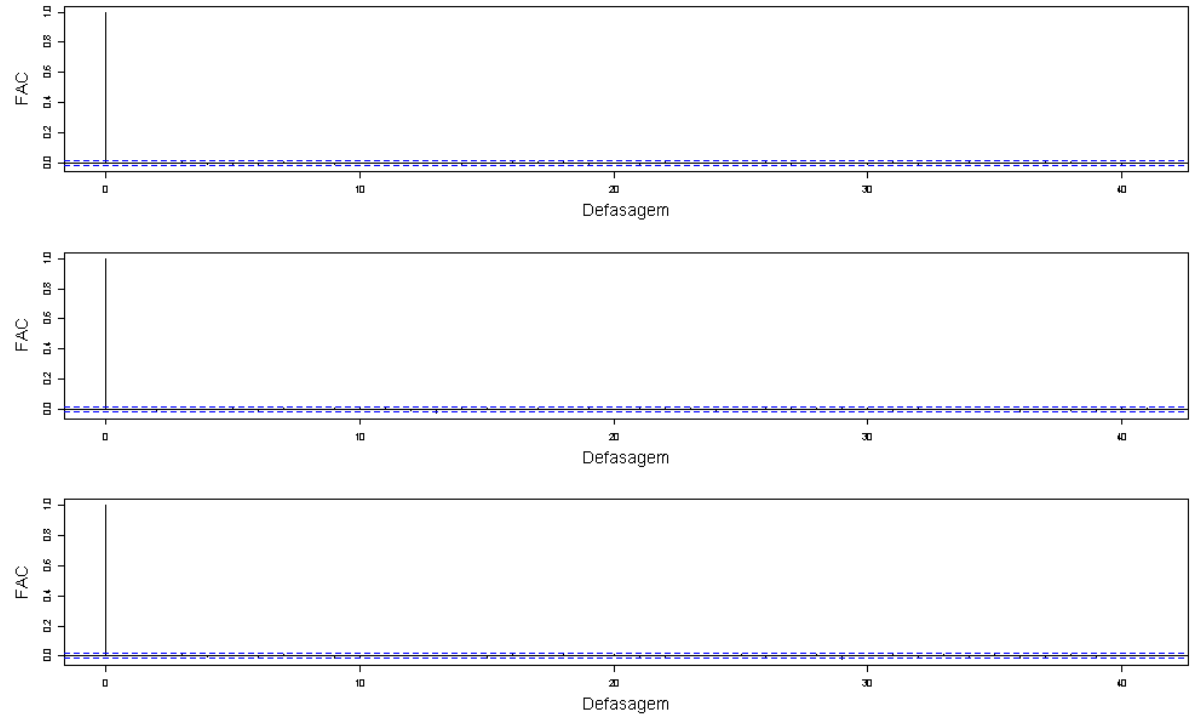

Figura 4: Gráficos de autocorrelação das cadeias estimadas para modelo de regressão linear simples com dados simulados para os parâmetros β_0, β_1 e τ

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:

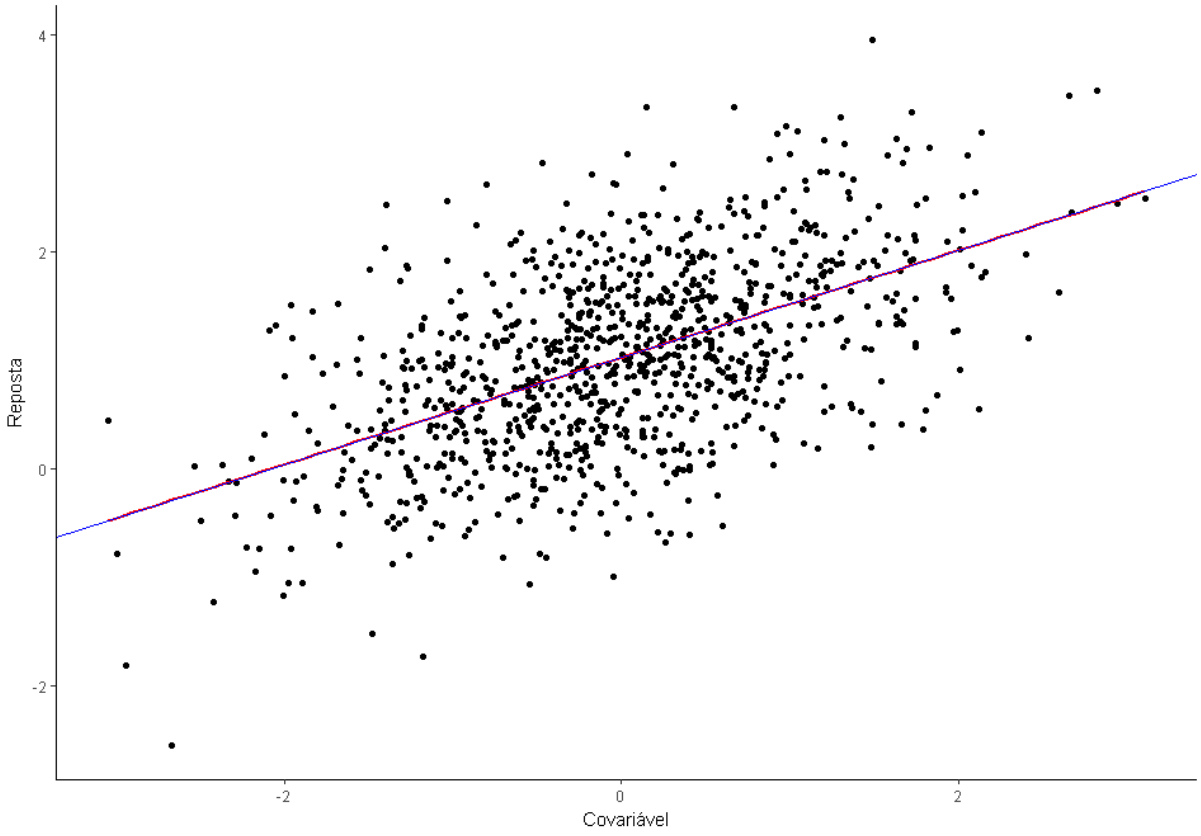

Figura 5: Relação entre a covariável e a variável resposta da cadeia simulada e reta do modelo linear clássico vs bayesiano com dados simulados

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.

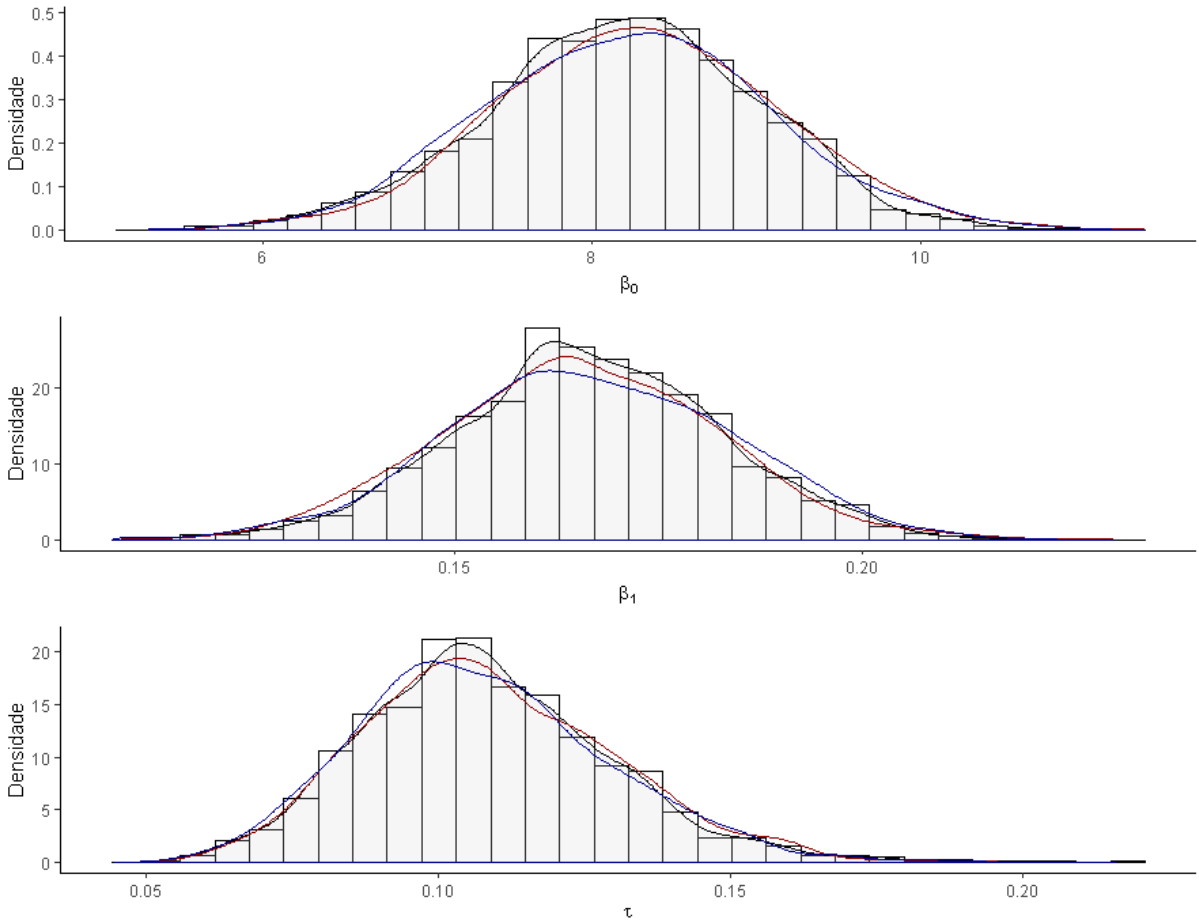

Figura 6: Histogramas e densidades das três últimas cadeias estimadas para modelo de regressão linear simples com base de dados cars

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.

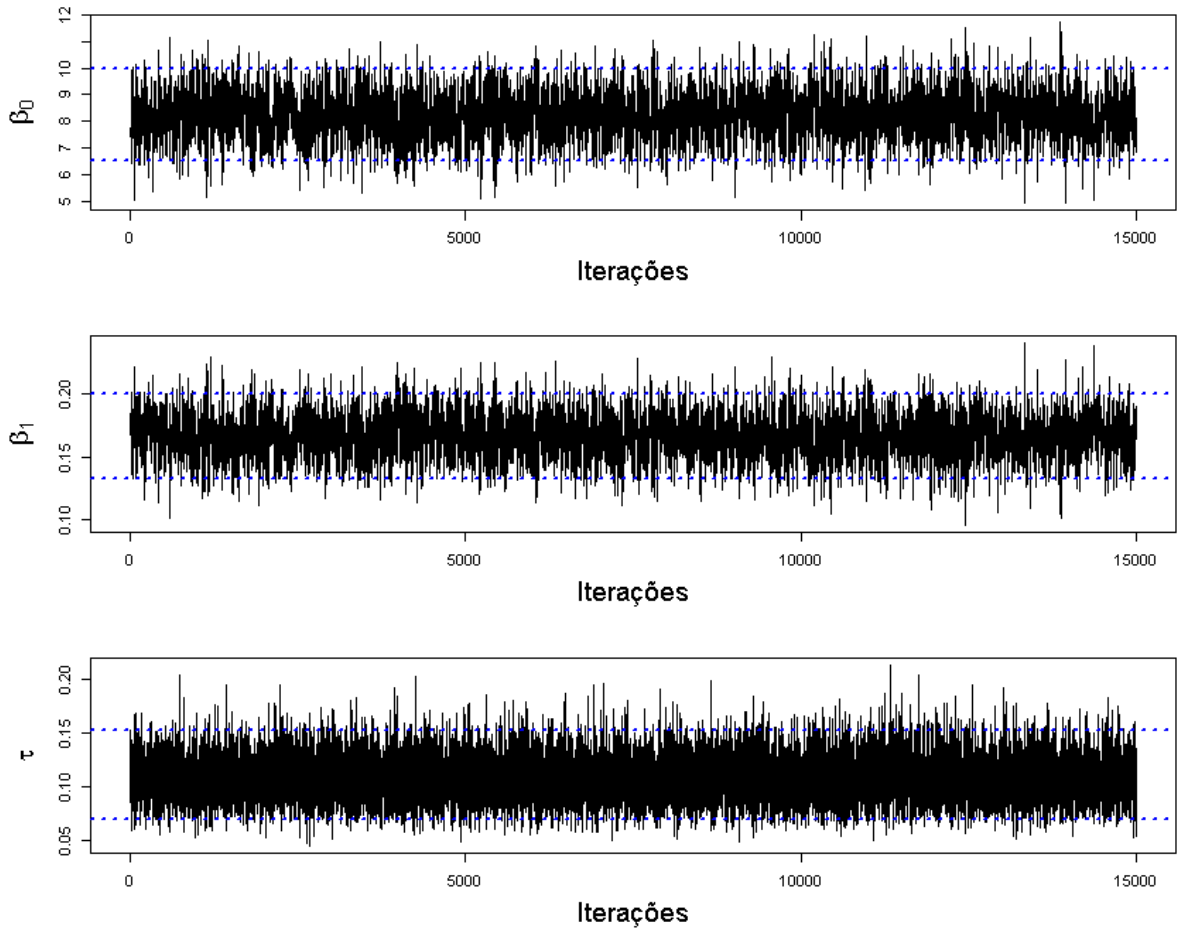

Figura 7: Cadeias estimadas para modelo de regressão linear simples com base de dados cars

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

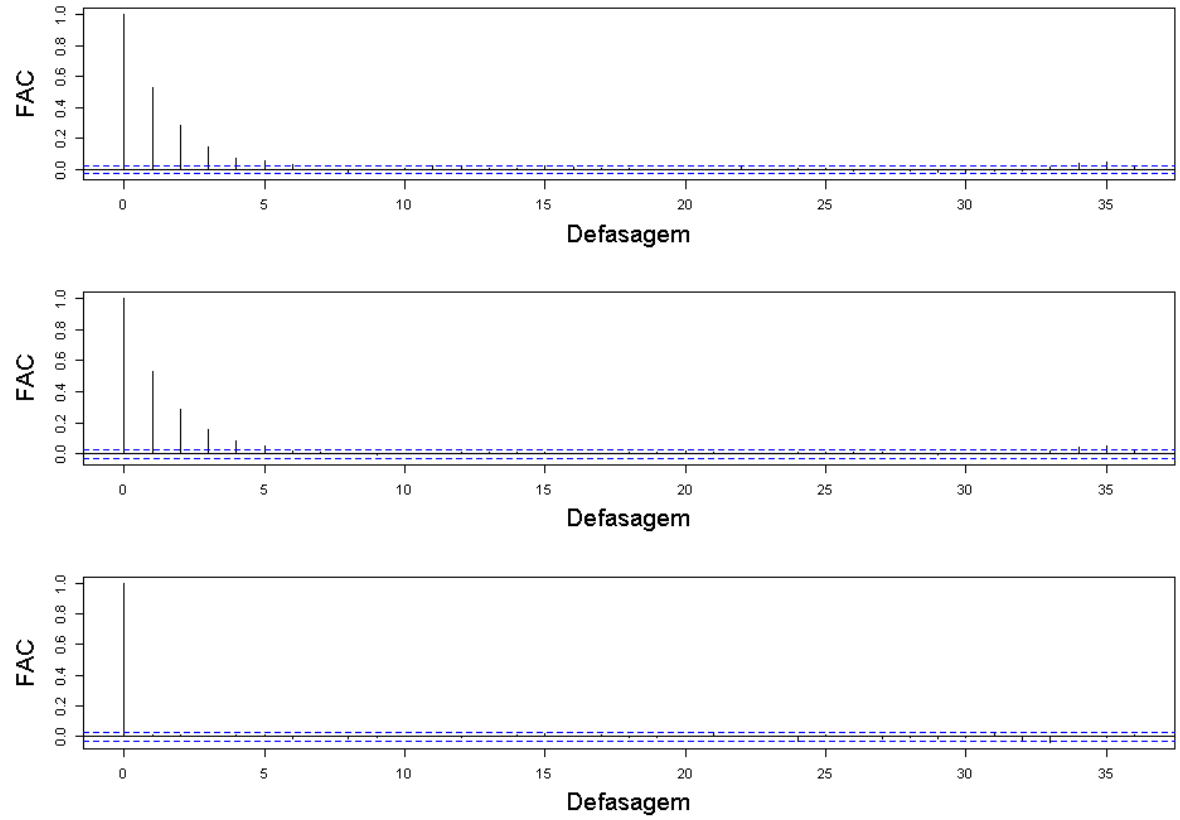

Figura 8: Gráficos de autocorrelação das cadeias estimadas para os respectivos parâmetros β_0, β_1 e τ do modelo de regressão linear simples com base de dados cars

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).

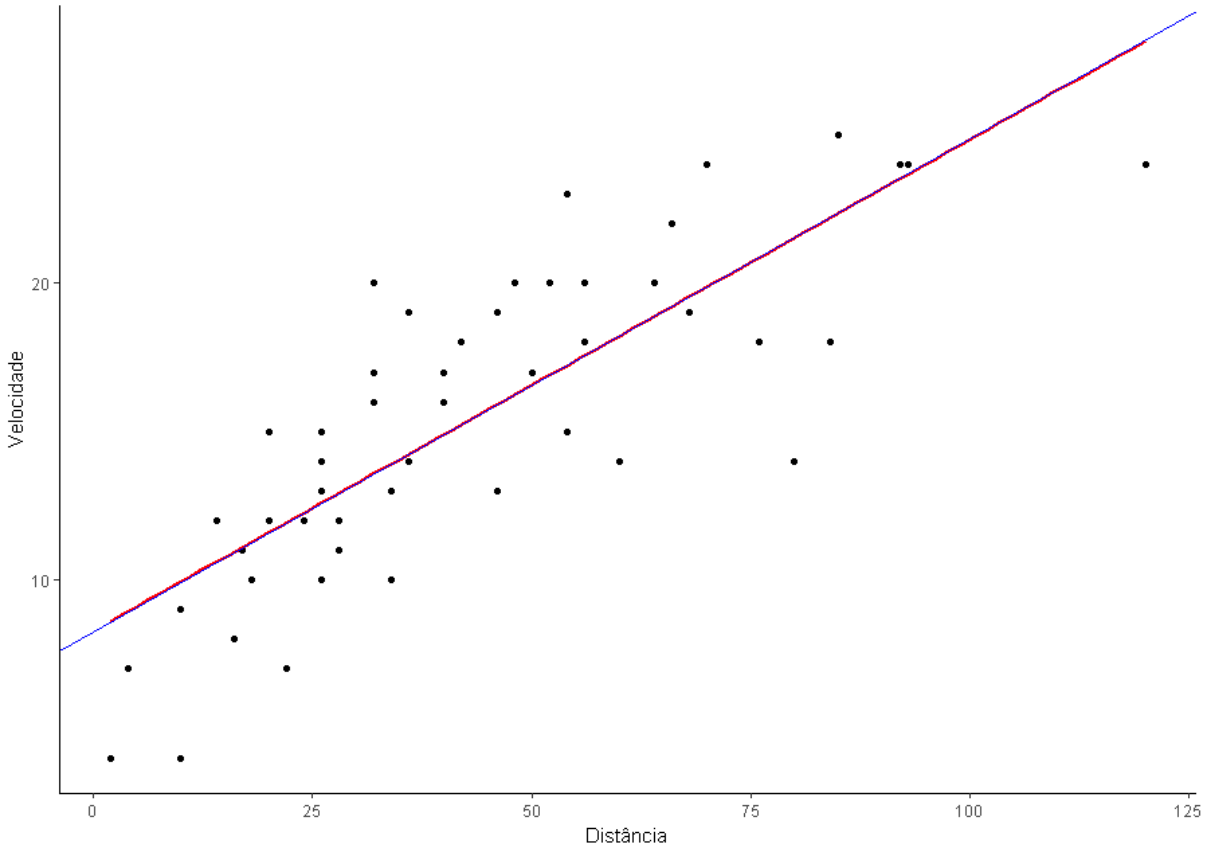

Figura 9: "Relação entre a covariável e a variável resposta da cadeia simulada com reta do modelo linear clássico vs bayesiano com base de dados cars

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.

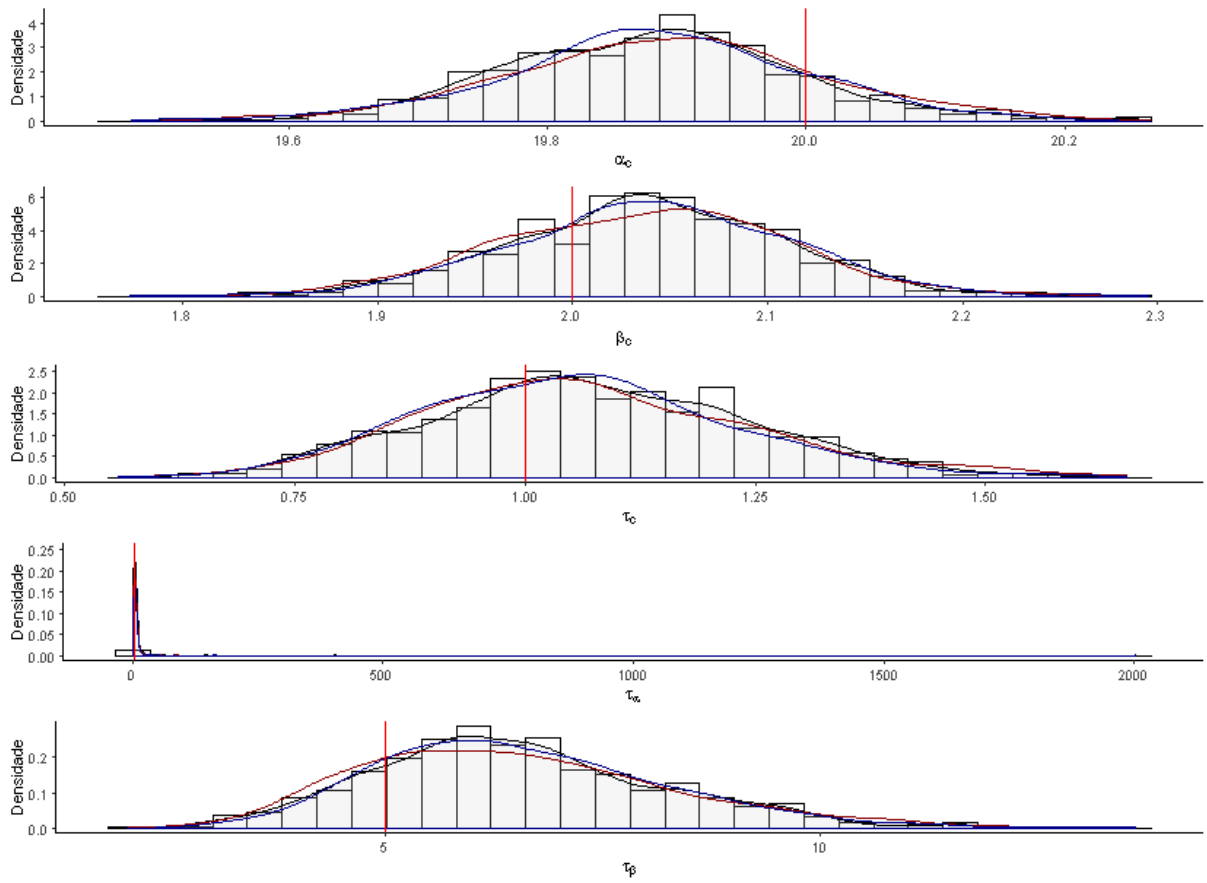

Figura 10: Histogramas e densidades das três últimas cadeias estimadas para o modelo de regressão hierárquico bayesiano com base de dados simulada

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

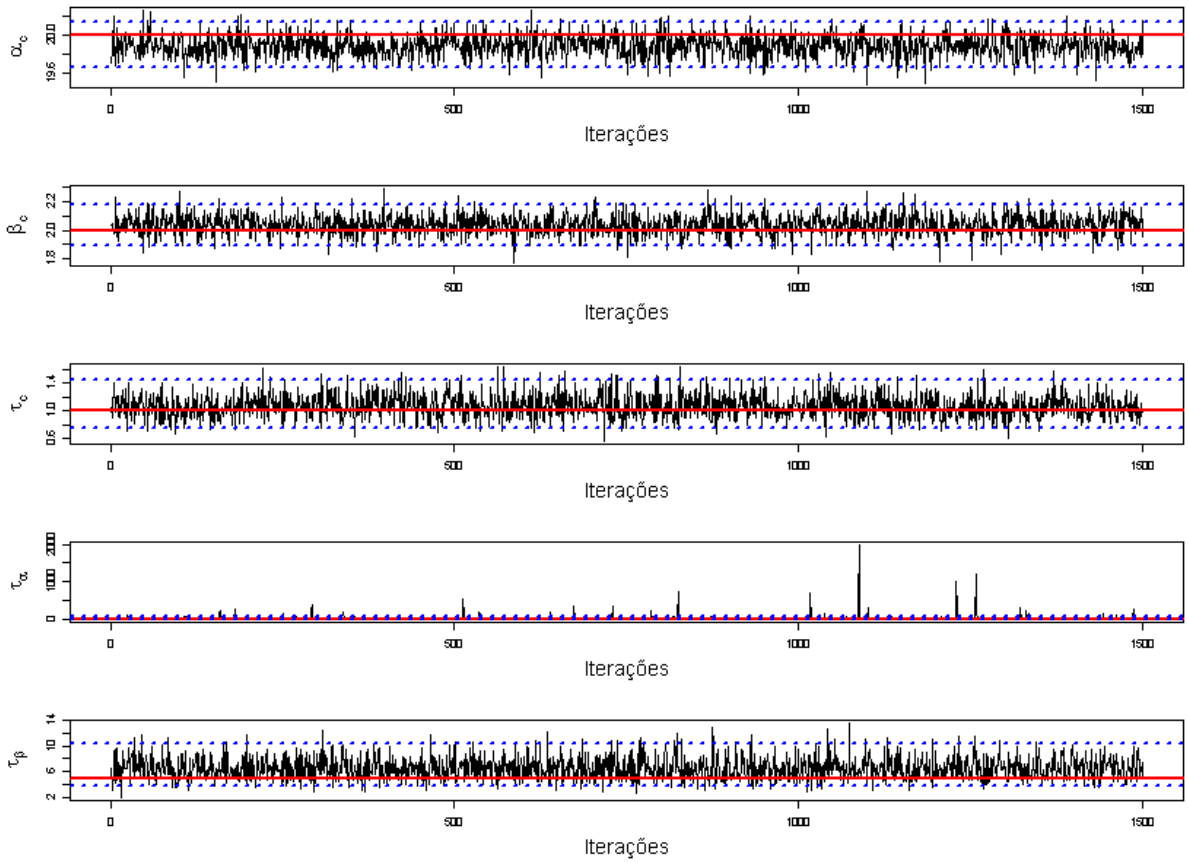

Figura 11: Cadeias estimadas para o modelo de regressão hierárquico bayesiano com base de dados simulada

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:

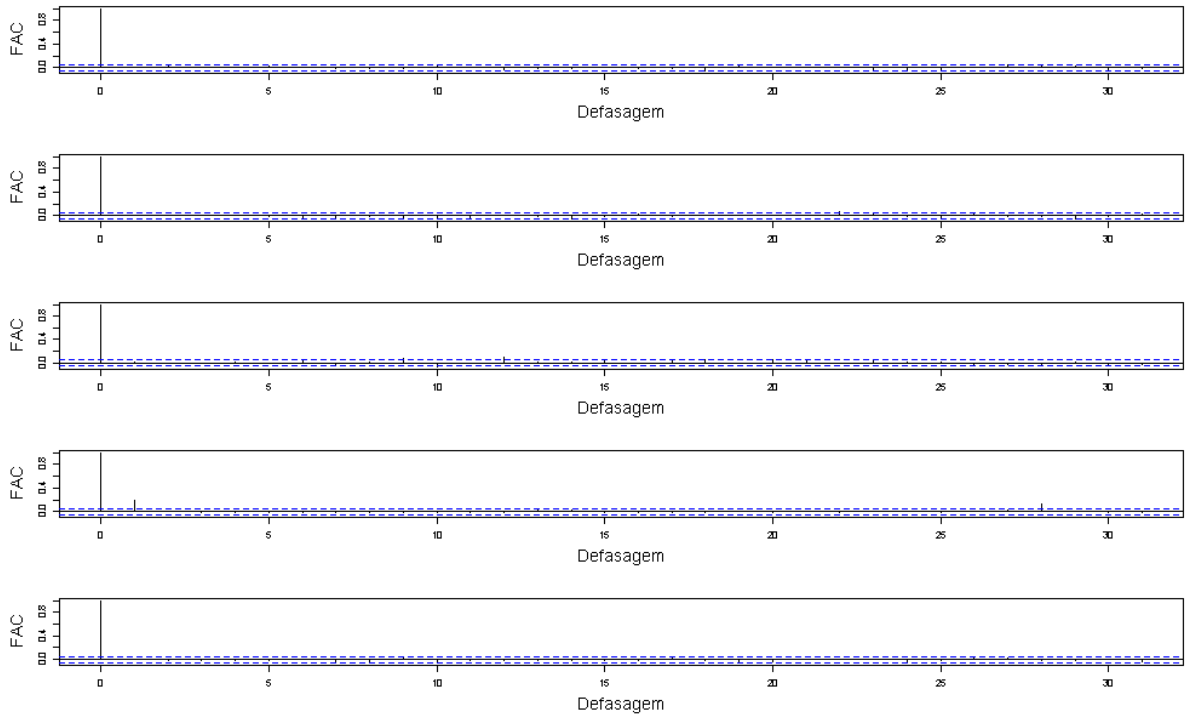

Figura 12: Gráficos de autocorrelação das cadeias estimadas para o os respectivos parâmetros α_c, β_c, ,τ_α,τ_c e τ_β do modelo de regressão hierárquico bayesiano com base de dados simulada

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.

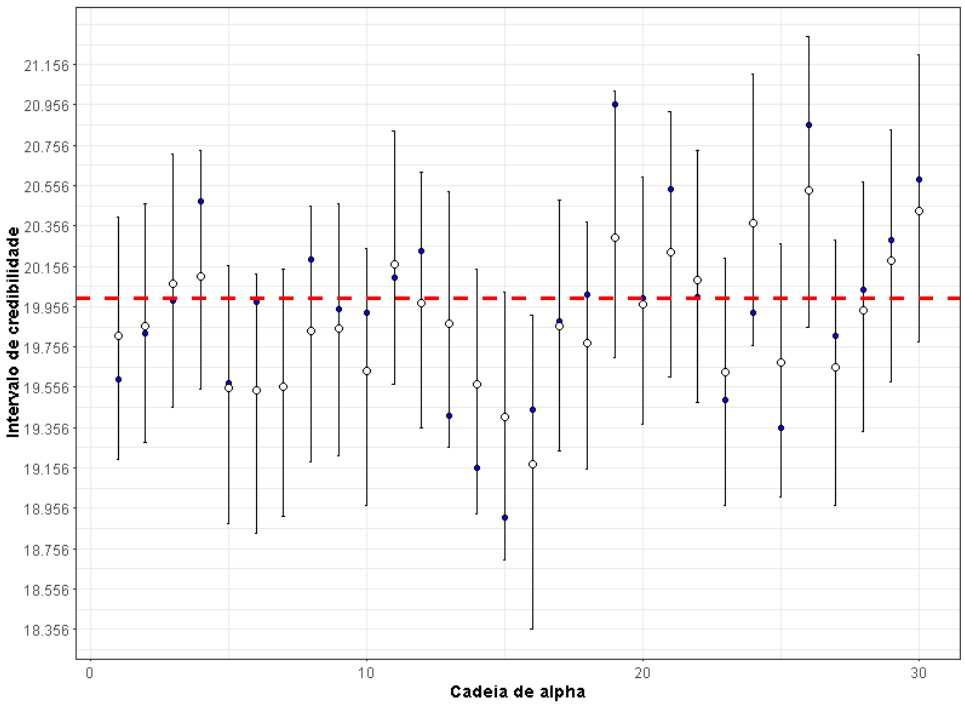

Figura 13: Médias e intervalos de credibilidade para a cadeia de α_i estimada incluindo o real valor estimado em azul e uma linha tracejada para o real valor de α_c

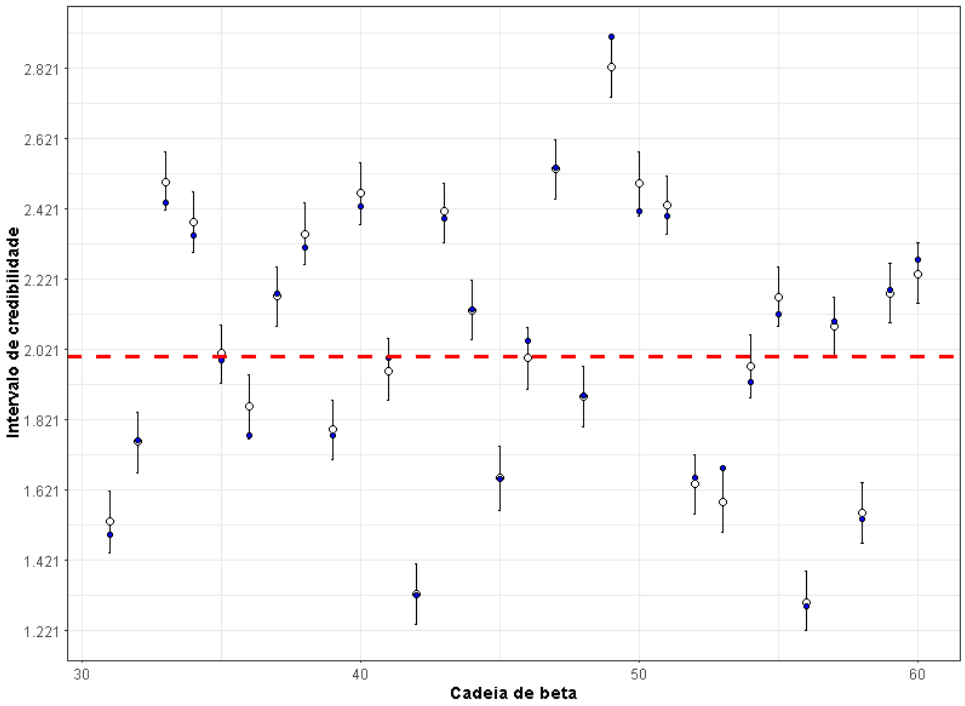

Figura 14: Médias e intervalos de credibilidade para a cadeia de β_i estimada incluindo o real valor estimado em azul e uma linha tracejada para o real valor de β_c

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.