Denise de Oliveira Alves Carneiro

Métodos Bayesianos de Seleção de Variáveis

Niterói - RJ, Brasil

19 de dezembro de 2017

Universidade Federal Fluminense

Denise de Oliveira Alves Carneiro

Métodos Bayesianos de Seleção de Variáveis

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador: Prof. Jony Arrais Pinto Junior

Niterói - RJ, Brasil

19 de dezembro de 2017

Universidade Federal Fluminense

Denise de Oliveira Alves Carneiro

Métodos Bayesianos de Seleção de Variáveis

Monografia de Projeto Final de Graduação sob o título *"Métodos Bayesianos de Seleção de Variáveis"*, defendida por Denise de Oliveira Alves Carneiro e aprovada em 19 de dezembro de 2017, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Prof. Dr. Jony Arrais Pinto Junior
Departamento de Estatística - UFF

Prof. Drª. Mariana Albi de Oliveira Souza
Departamento de Estatística - UFF

Prof. Drª. Larissa de Carvalho Alves
Departamento de Estatística - ENCE

Niterói, 19 de dezembro de 2017

C289 Carneiro, Denise de Oliveira Alves
Métodos bayesianos de seleção de variáveis / Denise de Oliveira Alves Carneiro. – Niterói, RJ: [s.n.], 2017.
44f.

Orientador: Prof. Dr. Jony Arrais Pinto Jr
TCC (Graduação de Bacharelado em Estatística) – Universidade Federal Fluminense, 2017.

1. Inferência bayesiana. 2. Métodos de seleção de variáveis. I. Título.

CDD 519.54

Resumo

A quantidade de dados gerados no dia a dia tem aumentado demasiadamente e com isso o interesse em se explicar um determinado desfecho tem-se tornado mais difícil, muitas vezes pela presença da multicolinearidade. Então surge a necessidade de métodos de seleção de variáveis que além de serem eficientes sejam mais rápidos e mais fáceis de se utilizar. Existem vários métodos de seleção de variáveis disponíveis na literatura. Dentre os mais utilizados estão o critério de informação de Akaike, AIC, de informação bayesiana, BIC, e o de informação do desvio, DIC, porém estes métodos são divididos em dois passos, ajustar todos os modelos possíveis e posteriormente calcular uma das medidas citadas para cada modelo e compará-los a fim de saber qual o melhor modelo. Como pode-se observar estes métodos podem ser bastante trabalhosos ou até mesmo inviáveis em alguns cenários. Uma possível alternativa aos métodos clássicos é a utilização de métodos de seleção de variáveis com o enfoque bayesiano. Os métodos bayesianos de seleção de variáveis baseiam-se em um único ajuste e utilizam uma variável indicadora responsável por determinar se uma variável é selecionada ou não e com isso torna possível quantificar a probabilidade de cada variável ser selecionada e a probabilidade de um determinado modelo ser escolhido. Neste trabalho é apresentado um estudo voltado para modelos de regressão linear múltiplo com três métodos bayesianos de seleção de variáveis: método de seleção de variáveis de Kuo & Mallick, métodos de seleção de variáveis de Gibbs e método de seleção de variáveis via busca estocástica. O objetivo é comparar tais métodos, através de um estudo de simulação de dois cenários (com presença/ausência de multicolinearidade) e modificando o valor inicial das variáveis indicadoras. Os três métodos estudados apresentaram bons resultados em ambos os cenários. Para os dois cenários observados o método de seleção de variáveis via busca estocástica se mostrou o mais rápido e o melhor por apresentar a maior probabilidade de o modelo correto ser visitado, sendo visitado 100% das vezes no cenário dependente quando as 4 primeiras variáveis indicadoras assumiram como valor inicial 1. Os valores iniciais adotados para todos os métodos nos dois cenários não influenciou no ajuste do modelo para a probabilidade *a posteriori*, com exceção do método de seleção de variáveis via busca estocástica que sofreu influência quando se inicializava as quatro primeiras covariáveis em zero, caindo absurdamente a probabilidade. Este trabalho utilizou uma abordagem completamente Bayesiana e os resultados computacionais foram obtidos por meio do R e BUGS (**B**ayesian inference **U**sing **G**ibbs **S**ampling).

Palavras-chaves: Métodos de seleção de variáveis, KM, GVS, SSVS, Inferência Bayesiana, OpenBUGS.

Dedicatória

"Paura di cadere, ma voglia di volare"

Agradecimentos

Agradeço primeiramente a Deus que iluminou o meu caminho durante esta caminhada, me dando força e coragem para prosseguir nesta jornada.

A minha mãe, Graça, por fazer todo o possível para me proporcionar uma boa qualidade de ensino, independente do esforço que fosse necessário. Obrigada por ser essa mãe guerreira, corajosa e estar sempre me incentivando a ser melhor.

Ao meu pai, Dejair, que sempre vai ser uma das minhas bases mesmo não estando mais entre nós. Sei que aonde estiver está olhando por mim.

As amigas, Vanessa e Evellyn Francys, irmãs que a UFF me deu. Aos amigos, Fernanda, Gilberto e Luciana, amigos de grupo de estudos, de viagens, da vida. A Fernanda Pedrosa, uma maluquinha que mesmo em ritmo acelerado tá sempre disposta a ajudar e escutar.

Aos amigos, Rapha, Larissa, Gabrielle, Fabrício, Gabriel, Luana, Evellyn, Tuany, Bárbara, Bruno que fazem parte dessa trajetória de alegrias e tristezas.

Ao meu namorado, Igor, por ser paciente, estar sempre comigo na luta do TCC e me ajudar sempre que possível.

Ao Júlio por estar sempre disposto a me ajudar nessa caminha tirando as minhas dúvidas sempre que possível.

Ao Jony por aceitar ser meu orientador luz me ajudando e guiando no mundo da Inferência Bayesiana. Entendendo minhas dificuldades e sendo paciente.

A Mariana por aceitar fazer parte da banca agregando valor ao meu TCC, além de fazer parte da minha formação.

Aos professores da Estatística que de alguma forma fizeram parte da minha formação.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

Com o grande aumento da quantidade de dados gerados nos dias de hoje e a necessidade dos pesquisadores das mais diversas áreas de entender e quantificar o tipo de relação existente entre as variáveis coletadas, vem crescendo o emprego de técnicas estatísticas. Quando tem-se o interesse em explicar um determinado desfecho, por meio de um grande conjunto de fatores, é necessária a utilização de um modelo estatístico. O interesse deste trabalho concentra-se especificamente na classe de modelos de regressão linear múltiplo.

Suponha que um pesquisador tem interesse em analisar como a idade, o consumo médio de bebidas alcoólicas, o nível de cansaço, a velocidade (km/h) e o tempo de experiência no volante (em anos) possuem relação com a quantidade de acidentes fatais no trânsito. Considere agora que o pesquisador possui dois modelos. No primeiro modelo, M_1, estão presentes os fatores: idade, consumo médio de bebidas alcoólicas, velocidade e nível de cansaço. Já o segundo modelo, M_2, é constituído pelos fatores: idade, tempo de experiência no volante e a velocidade. Note que, neste cenário, muitos outros modelos poderiam ser considerados. E o mais natural é se perguntar: como escolher entre os diferentes modelos?

Quando o número de variáveis ou fatores é muito grande, pode-se encontrar problemas para entender as relações de interesse entre as variáveis. Por exemplo, a existência de alta correlação entre dois ou mais fatores pode prejudicar a estimação da variância dos parâmetros, tornando assim as suas estimativas imprecisas. Este problema é bem conhecido na literatura, e é comumente denominado como multicolinearidade (Neter et al. 1996[1]). Com o intuito de contornar este possível problema e achar um modelo com um menor número de fatores importantes para a explicação do desfecho de interesse, pode-se utilizar seleção de variáveis. A partir dos modelos apresentados no exemplo pelo pesquisador, naturalmente surge o seguinte questionamento: Qual é o "melhor"modelo, ou seja, qual o subconjunto de variáveis explica "melhor"a variável resposta?

A literatura propõe vários métodos de seleção de variáveis. Dentre os métodos

clássicos mais utilizados estão o coeficiente de determinação ajustado de Srivastava et al. (1995)[2], R_a^2, o critério de informação de Akaike (1976)[3], AIC, o critério de informação bayesiana de Schwarz et al. (1978)[4], BIC e o critério de informação do desvio de Celeux et al. (2006)[5], DIC. Em resumo, nestes métodos primeiramente é necessário ajustar diversos modelos com subconjuntos de variáveis e, posteriormente, avaliar por meio dos critérios qual destes modelos será o "melhor".

Supondo que foram ajustados os modelos M_1 e M_2, para responder ao pesquisador qual o modelo mais correto, será necessário fazer a comparação dos dois ajustes, calculando, por exemplo, o AIC de cada um. O modelo que apresentar o menor AIC será considerado o "melhor".

Assumindo agora uma situação na qual o pesquisador possua 10 fatores, para saber qual o modelo mais adequado entre todos os possíveis, seria necessário fazer muitos ajustes, 2^{10}, e calcular o AIC para cada um deles. Logo, usar esses métodos clássicos para comparar todas as possibilidades pode ser bastante trabalhoso e até mesmo inviável em alguns cenários. Uma alternativa a esses métodos são os métodos de seleção de variáveis bayesianos baseados em indicadores.

Alguns métodos de seleção de variáveis bayesianas, como o método de Kuo & Mallick (1998)[6], o de seleção de variáveis de Gibbs e o de seleção de variáveis via busca estocástica se utilizam de uma variável indicadora para selecionar variáveis. Na verdade estes métodos utilizam duas variáveis auxiliares: a primeira, é a indicadora, que será responsável por indicar a presença ou ausência de um fator e a segunda é responsável por quantificar o efeito de um fator em um problema de regressão.

O'Hara e Sillanpää (2009)[7] realizaram um estudo de revisão dos métodos de seleção de variáveis bayesianas, incluindo os métodos de Kuo & Mallick (KM), seleção de variáveis de Gibbs (GVS), seleção de variável via busca estocástica (SSVS), entre outros. Nesse estudo, eles fazem uma breve revisão desses métodos, de acordo com suas propriedades, implementando-os no OpenBUGS (2009)[8] em problemas com dados reais e simulados, com o intuito de investigar como esses métodos funcionam na prática.

Este trabalho tem como objetivo estudar os métodos de seleção de variáveis bayesianos de Kuo & Mallick, GVS e SSVS e, por meio de um estudo de simulações, verificar o comportamento destes em cenários distintos, verificando a performance dos mesmos na presença e na ausência de multicolinearidade, por exemplo.

No presente trabalho, no Capítulo 1 consta uma breve introdução sobre o assunto

que será abordado. No Capítulo 2 são apresentados os objetivos a serem alcançados. No Capítulo 3 apresentam-se o modelo de regressão linear múltiplo, alguns conceitos básicos de inferência bayesiana e a estimação dos parâmetros, além dos métodos de seleção de variáveis bayesianas que são o foco deste trabalho. No Capítulo 4 é realizado um estudo de simulação dividido em dois cenários e em 4 casos para cada cenário. Por último, no Capítulo 5, são apresentados as conclusões sobre as simulações realizadas no capítulo anterior.

2 Objetivos

O objetivo principal deste trabalho é estudar os métodos de seleção de variáveis bayesianas de Kuo & Mallick , Gibbs e seleção de variáveis via busca estocástica no contexto de modelos lineares.

Como objetivos secundários tem-se:

- fazer uma breve introdução sobre a fundamentação teórica do modelo de regressão linear múltiplo e inferência bayesiana;
- apresentar os algoritmos dos métodos de simulação de Monte Carlo via Cadeias de Markov ($MCMC$) que serão utilizados para se fazer inferência no modelo estudado;
- estudar os métodos de seleção de variáveis bayesianos de Kuo & Mallick, seleção de Gibbs e o de seleção de variáveis via busca estocástica, identificando diferenças e similaridades nos mesmos e;
- fazer um estudo de simulações sob diferentes aspectos com o auxílio do programa OpenBUGS, incluindo presença/ausência de multicolinearidade.

3 Materiais e Métodos

Neste capítulo serão apresentadas as definições básicas do modelo de regressão linear múltiplo, inferência bayesiana e os métodos de seleção de variáveis bayesianos baseados em indicadoras: Kuo & Mallick, seleção de variáveis de Gibbs e o de seleção de variáveis via busca estocástica que serão abordados neste trabalho.

3.1 Modelo de Regressão Linear Múltiplo

Pela necessidade de se analisar a associação entre várias variáveis em diversas áreas, como, por exemplo, Biologia, Economia, Psicologia, entre outras, tornou-se imprescindível a utilização de modelos estatísticos. Suponha que um nutricionista quer analisar a relação da porcentagem de carboidratos, com as seguintes variáveis: porcentagem de proteínas, peso (kg) e idade de um paciente. Uma das ferramentas que a Estatística dispõe para avaliar e quantificar essas possíveis relações é o modelo de regressão linear múltiplo. O modelo de regressão linear múltiplo é um caso específico de modelos lineares. Esse modelo pode ser definido como um meio formal de expressar a relação estatística entre duas ou mais variáveis que sejam lineares nos parâmetros. Considere um experimento com n unidades experimentais. A variável resposta da i-ésima unidade experimental, também denominada variável dependente, será denotada por Y_i e pode ser predita a partir de um conjunto de p variáveis explicativas, também conhecidas como variáveis independentes, denotadas aqui por $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})'$, $i = 1, 2, ..., n$. Para a descrição do modelo abaixo utilizou-se como suporte Neter et al. (1996)[1]. Essa relação pode ser indicada por:

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i, \ i = 1, 2, ..., n, \tag{3.1}$$

na qual, $\boldsymbol{\beta}$=$(\beta_1, \beta_2, ..., \beta_p)'$ é o vetor de parâmetros que têm-se interesse em estimar, β_j, $j = 1, ..., p$, representa o efeito na variável resposta quando a j-ésima variável explicativa sofre um acréscimo unitário ao mesmo tempo que as outras variáveis explicativas se matem constantes e ϵ_i é o erro aleatório da i-ésima unidade experimental, em que $\epsilon_i \sim N(0, \tau^{-1})$,

no qual τ é uma função da variância σ^2 chamada de precisão, $\tau = 1/\sigma^2$.

Para facilitar a apresentação dos cálculos necessários para se fazer inferência no modelo, pode-se reescrever a Equação 3.1 na forma matricial :

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \boldsymbol{N_n}(0, \tau^{-1} I_n), \tag{3.2}$$

na qual, $\boldsymbol{N_n}$ e I_n denotam, respectivamente, a distribuição normal multivariada de ordem n e uma matriz identidade de ordem n,

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}_{n\times 1}, \qquad X = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ x_{21} & \dots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}_{n\times p}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}_{p\times 1} \quad e \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}_{n\times 1}.$$

No exemplo anteriormente apresentado, a variável resposta é a porcentagem de carboidratos (Y) e as variáveis explicativas são a porcentagem de proteínas ($\mathbf{x}_1$), peso ($\mathbf{x}_2$) e idade ($\mathbf{x}_3$). Assumindo hipoteticamente para este exemplo que $\beta_1 > 0$, $\beta_2 < 0$ e $\beta_3 < 0$, pode-se dizer que β_1 é o aumento esperado na porcentagem de carboidratos ao se aumentar em 1% a porcentagem de proteína, considerando as demais variáveis fixas, β_2 é a diminuição esperada na porcentagem de carboidratos ao se aumentar em 1kg o peso, mantendo as outras variáveis fixas, e β_3 é a diminuição esperada na porcentagem de carboidratos ao se aumentar em 1 ano a idade, mantendo fixas as outras variáveis.

A situação apresentada é um problema clássico e recorrente em diversas áreas. Para se resolver o problema é necessário conhecer as relações entre a variável resposta e as variáveis independentes, isto é, conhecer o vetor $\boldsymbol{\beta}$ e a precisão τ. Neste trabalho, o vetor de parâmetros desconhecidos será denotado por $\boldsymbol{\theta} = (\boldsymbol{\beta}, \tau)'$.

O conjunto de técnicas utilizadas para revelar algumas informações sobre o vetor de parâmetros desconhecido é chamado de inferência. A inferência pode ser abordada sob duas perspectivas, uma com enfoque clássico (ou frequentista) e a outra bayesiana. Este estudo terá uma abordagem completamente bayesiana.

3.2 Inferência Bayesiana

A inferência bayesiana consiste na incorporação de informações subjetivas, que aperfeiçoam o conhecimento sobre algum problema, com a ajuda do Teorema de Bayes. O

vetor de parâmetros de interesse desconhecido será representado por $\boldsymbol{\theta}$, que pode assumir valores em Θ, e é considerado aleatório o que torna possível atribuir distribuições para $\boldsymbol{\theta}$ sob o ponto de vista bayesiano. O conhecimento inicial que se tem sobre o vetor de parâmetros de interesse desconhecido é representado pela distribuição *a priori*, $p(\boldsymbol{\theta})$, e toda outra informação que seja relevante é expressa pela função de verossimilhança. Por meio da combinação da distribuição *a priori* com a função de verossimilhança, torna-se possível achar a distribuição atualizada dos parâmetros de interesse, chamada de distribuição *a posteriori*, via Teorema de Bayes.

Com base em Migon e Gamerman (1999)[9] tem-se:

$$p(\boldsymbol{\theta}|\mathbf{Y}) = \frac{L(\boldsymbol{\theta}|\mathbf{Y})p(\boldsymbol{\theta})}{p(\mathbf{Y})}, \tag{3.3}$$

no qual $p(\boldsymbol{\theta}|\mathbf{Y})$ é a distribuição *a posteriori* de $\boldsymbol{\theta}$, $L(\mathbf{Y}|\boldsymbol{\theta})$ é a função de verossimilhança e $p(\mathbf{Y})$ é a distribuição marginal de $\mathbf{Y}$. Como $1/p(\mathbf{Y})$ funciona como constante normalizadora de $p(\boldsymbol{\theta}|\mathbf{Y})$ e não depende de $\boldsymbol{\theta}$, a Equação 3.3 pode ser reescrita como:

$$p(\boldsymbol{\theta}|\mathbf{Y}) \ \propto \ L(\boldsymbol{\theta}|\mathbf{Y})p(\boldsymbol{\theta}), \tag{3.4}$$

em que o símbolo $\propto$ significa proporcionalidade.

O cálculo da distribuição *a posteriori* de interesse geralmente envolve um grande número de integrações, principalmente no caso em que $\boldsymbol{\theta}$ tem muitas componentes. Este grande número de integrações pode gerar transtornos como, por exemplo, a distribuição *a posteriori* não ter uma forma analítica fechada, tornando impossível resolver o problema de maneira analítica. Logo, uma maneira de contornar esse problema será obter amostras dessa distribuição e recorrer a algum método de aproximação, que são bem úteis nesses casos. Os métodos utilizados neste trabalho serão os métodos de simulação de Monte Carlo via Cadeias de Markov (MCMC), apresentados detalhadamente em Gamerman e Lopes (2006)[10].

Uma cadeia de Markov pode ser definida de maneira simples como um processo estocástico no qual os estados passados e futuros são independentes, dado o estado presente. O MCMC tem como ideia central construir uma cadeia de Markov em que seja fácil gerar uma amostra e tenha distribuição estacionária dada pela distribuição de interesse, que neste caso é a distribuição *a posteriori*. Esta cadeia de Markov deve ser homogênea, isto é, as probabilidades de transição de um estado para o outro não dependem do tempo de iteração; irredutível, ou seja, deve ser possível chegar de um estado à qualquer outro estado em um número finito de iterações; e também aperiódica, não havendo estados ab-

sorventes. Quando a convergência da cadeia for atingida significa que as amostras estão sendo geradas da distribuição estacionária.

Há vários métodos de construção da cadeias de Markov. Neste trabalho serão utilizados o algoritmo de **Metropolis-Hastings** e o amostrador de **Gibbs**, que estão entre os mais empregados.

O algoritmo de **Metropolis-Hastings** foi proposto primeiramente por Metropolis et al. (1953)[11] e futuramente estendido por Hastings (1970)[12]. Este método é utilizado para gerar amostras da distribuição de interesse da qual não é possível obter as condicionais completas na forma analítica fechada. Para cada parâmetro, a distribuição condicional completa é a distribuição deste parâmetro condicionada a todos os outros parâmetros.

Nesta situação, são gerados valores para cada parâmetro segundo uma distribuição auxiliar, chamada de distribuição proposta, e esses valores são aceitos ou não com uma certa probabilidade. Assumindo $p(\cdot)$ como as distribuições condicionais completas e $q(\cdot)$ a distribuição auxiliar, o algoritmo de **Metropolis-Hastings** pode ser representado por:

(i) Inicializa-se o contador $j = 1$ e especifique um valor inicial $\boldsymbol{\theta}^{(0)}$ para $\boldsymbol{\theta}$.

(ii) Gera-se um valor θ' da distribuição proposta $q(\theta'|\theta^{(j-1)})$.

(iii) Aceita-se o valor gerado em (ii) com probabilidade $\alpha(\theta, \theta') = min\{1, \frac{p(\theta')q(\theta^{(j-1)}|\theta')}{p(\theta^{(j-1)})q(\theta'|\theta^{(j-1)})}\}$. Se o valor for aceito, $\theta^{(j)} = \theta'$. Caso contrário, a cadeia não se move e $\theta^{(j)} = \theta^{(j-1)}$.

(iv) Atualiza-se o contador de j para $j+1$ e retorna-se ao passo (ii) até que a convergência seja obtida.

Supondo que a partir da iteração J a convergência foi atingida, pode-se considerar os valores simulados a partir desta iteração como uma amostra aproximada da distribuição *a posteriori* de $\boldsymbol{\theta}$.

O amostrador de **Gibbs**, apresentado por Geman e Geman (1984)[13] e disseminado por Gelfand e Smith (1990)[14], no qual a distribuição proposta é a própria distribuição condicional completa do parâmetro que está sendo amostrado, devendo ser conhecida, fazendo com que a probabilidade de aceitação seja igual a 1, ou seja, todo valor gerado é aceito. Representando as componentes de um vetor paramétrico de $\boldsymbol{\theta}$ como θ_j para todo j, e a distribuição condicional completa como $p(\theta_j|\boldsymbol{\theta}_{-j})$, no qual $\boldsymbol{\theta}_{-j} = (\theta_1, ..., \theta_{j-1}, \theta_{j+1}, ..., \theta_k)'$, o algoritmo amostrador de **Gibbs** pode ser descrito por:

(i) Fixa-se um valor arbitrário inicial para $\boldsymbol{\theta}^{(0)}$ para $\boldsymbol{\theta}$ e inicializa-se o contador $j = 1$.

(ii) Obtém-se um novo valor θ_j a partir de $\theta_{(j-1)}$ através de gerações sucessivas de:

$$\begin{aligned}
\theta_1^{(j)} &\sim p(\theta_1|\theta_2^{(j-1)}, ..., \theta_k^{(j-1)}) \\
\theta_2^{(j)} &\sim p(\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)}, ..., \theta_k^{(j-1)}) \\
\theta_3^{(j)} &\sim p(\theta_3|\theta_1^{(j)}, \theta_2^{(j)}, \theta_4^{(j-1)}, ..., \theta_k^{(j-1)}) \\
&\vdots \\
\theta_k^{(j)} &\sim p(\theta_k|\theta_1^{(j)}, \theta_2^{(j)}, ..., \theta_{k-1}^{(j)}).
\end{aligned}$$

(iii) Atualiza-se o contador de j para $j+1$ e retorna-se ao passo (ii) até que a convergência seja obtida.

Após uma breve descrição de alguns conceitos básicos necessários para a realização da inferência bayesiana, é possível discutir a estimação dos parâmetros desconhecidos do modelo de regressão linear múltiplo apresentado na seção anterior sob o ponto de vista bayesiano.

3.3 Estimação dos parâmetros do Modelo de Regressão Linear Múltiplo

Esta seção abordará a estimação dos parâmetros, $\boldsymbol{\beta}$ e τ, para o modelo de regressão linear múltiplo apresentado anteriormente, por meio da inferência bayesiana, e a implementação do amostrador de Gibbs. Da Equação 3.2 tem-se que:

$$\mathbf{Y} \sim N_n(X\boldsymbol{\beta}, \tau^{-1}I_n), \tag{3.5}$$

ou seja, $\mathbf{Y}$ segue uma distribuição Normal Multivariada de ordem n com valor esperado $E(\mathbf{Y}) = X\boldsymbol{\beta}$ e matriz de covariâncias $Var(\mathbf{Y}) = \tau^{-1}I_n$, em que I_n é a matriz identidade de ordem n.

Pode-se escrever de forma hierárquica o modelo de regressão linear múltiplo com

enfoque bayesiano, da seguinte maneira:

$$\begin{aligned} Y_i &\sim N(\mu_i, \tau^{-1}), \quad i = 1, ..., n, \\ \mu_i &= \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}, \quad i = 1, ..., n, \\ \beta_j &\sim N(A_j, D_j), \quad j = 1, ..., p, \\ \tau &\sim G(c, d). \end{aligned}$$

Note que atribui-se uma distribuição *a priori* Normal Univariada para o efeito da j-ésima covariável (β_j) com média A_j e variância D_j e Gama para o parâmetro de precisão (τ), com parâmetros c e d, em que, A_j,D_j, $j = 1, ..., p$, c e d são conhecidos.

Para fazer inferência no modelo apresentado acima, é necessário conhecer a sua função de verossimilhança, definida por:

$$L(\boldsymbol{\beta}, \tau|\mathbf{Y}) = |\tau|^{\frac{n}{2}}(2\pi)^{-\frac{n}{2}} exp\left\{-\frac{\tau}{2}(\mathbf{Y} - X\boldsymbol{\beta})'(\mathbf{Y} - X\boldsymbol{\beta})\right\} \; I_{\mathbb{R}^p}(\boldsymbol{\beta}) \; I_{(0,\infty)}(\tau), \tag{3.6}$$

na qual $\mathbb{R}^p = (-\infty, \infty)_1 \times ... \times (-\infty, \infty)_p$.

Neste trabalho, será assumida independência *a priori* para $\boldsymbol{\beta}$ e τ, logo, a distribuição *a priori* conjunta para $\boldsymbol{\theta}$ será dada por:

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\beta})p(\tau). \tag{3.7}$$

A partir do Teorema de Bayes (3.3), da função de verossimilhança (3.6) e da distribuição *a priori* conjunta (3.7), apresentados anteriormente, pode-se escrever a distribuição *a posteriori* conjunta, como:

$$\begin{aligned} p(\boldsymbol{\beta}, \tau|\mathbf{Y}) \quad \propto \quad & |\tau|^{\frac{n}{2}} exp\left\{-\frac{\tau}{2}(\mathbf{Y} - X\boldsymbol{\beta})'(\mathbf{Y} - X\boldsymbol{\beta})\right\} \\ & \times exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \mathbf{A})' \; D^{-1}(\boldsymbol{\beta} - \mathbf{A})\right\} \tau^{c-1} e^{-d\tau} \; I_{\mathbb{R}^p}(\boldsymbol{\beta}) \; I_{(0,\infty)}(\tau). \end{aligned} \tag{3.8}$$

Após a especificação da distribuição *a posteriori* conjunta (3.8), pode-se agora escrever as distribuições condicionais completas que serão utilizadas no amostrador de Gibbs. A distribuição condicional completa de $\boldsymbol{\beta}$ será dada por:

$$p(\boldsymbol{\beta}|\tau, \mathbf{Y}) \propto exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}'(\tau X'X + D^{-1})\boldsymbol{\beta} - 2\boldsymbol{\beta}'(\tau X'\mathbf{Y} + D^{-1}\mathbf{A})\right]\right\} \; I_{\mathbb{R}^p}(\boldsymbol{\beta}).$$

Note que, após algumas operações algébricas, $\boldsymbol{\beta}$ tem distribuição condicional completa

conhecida:

$$\boldsymbol{\beta}|\ \tau, \mathbf{Y} \sim N_n\left(\frac{\tau X'\mathbf{Y} + D^{-1}\mathbf{A}}{\tau X'X + D^{-1}},\ (\tau X'X + D^{-1})^{-1}\right).$$

A distribuição condicional completa de τ é dada por:

$$p(\tau|\boldsymbol{\beta}, \mathbf{Y}) \propto \tau^{(\frac{n}{2}+c)-1} e^{-d\tau}\ I_{(0,\infty)}(\tau),$$

e, portanto,

$$\tau|\boldsymbol{\beta}, \mathbf{Y}\ \sim\ G\left(\frac{n}{2} + c, d\right).$$

Depois de todos os cálculos feitos, pode-se facilmente implementar o amostrador de Gibbs, pois as condicionais completas são conhecidas, gerando valores alternados destas distribuições.

Ao se trabalhar com um número grande de variáveis explicativas, isto é, quando p é demasiadamente grande, pode-se enfrentar alguns problemas, como, por exemplo, a presença de multicolinearidade, que consiste em um problema de alta correlação entre duas ou mais variáveis explicativas, tornando assim as estimativas dos parâmetros imprecisa. Para tentar contornar situações problemáticas oriundas de se trabalhar com muitas variáveis explicativas, serão utilizados métodos bayesianos de seleção de variáveis.

3.4 Métodos de Seleção de Variáveis Bayesianos

O principal objetivo dos métodos de seleção de variáveis é selecionar o "melhor"subconjunto de variáveis explicativas, isto é, as variáveis que de fato estão relacionadas com a variável resposta. Mitchell e Beauchamp (1988)[15] citam alguns motivos para a seleção de variáveis: expressar a relação entre a variável resposta e as explicativas da maneira mais simples possível, identificar variáveis explicativas que sejam importantes e insignificantes, reduzir custo e aumentar a precisão das estimativas e previsões estatísticas.

Existem vários métodos para selecionar subconjuntos de variáveis na literatura. Dentre os mais conhecidos e mais utilizados, pode-se citar o critério de informação de Akaike, AIC, proposto por Akaike (1976)[3], definido como:

$$AIC = -2ln(L(\hat{\boldsymbol{\theta}}|\mathbf{Y})) + 2p,$$

e o critério de informação bayesiana, BIC, proposto por Schwarz et al.(1978)[4], definido

como:

$$BIC = -2ln(L(\hat{\boldsymbol{\theta}}|\mathbf{Y})) + pln(n),$$

no qual $L(\hat{\boldsymbol{\theta}}|\mathbf{Y})$ é a função de verossimilhança, p é o número de parâmetros a ser estimado e n o número de observações da amostra. O AIC e o BIC são critérios baseados no máximo da função de verossimilhança. Já o critério de informação do desvio de Celeux et al. (2006)[5], DIC, é definido como:

$$DIC = D(\hat{\boldsymbol{\theta}}) + 2p_D,$$

sendo $D(\hat{\boldsymbol{\theta}})$ a média *a posteriori* do desvio e p_D o número real de parâmetros. O DIC é uma generalização do BIC. Para todos esses métodos, quanto menor for o valor obtido melhor será o modelo. Entretanto, para aplicação destes métodos é necessário fazer todos os ajustes possíveis com subconjuntos de variáveis explicativas e depois calcular as medidas para todos e compará-los para determinar qual o ajuste é mais adequado.

As medidas AIC, BIC e DIC, entre outras, são utilizadas para fazer essas comparações. Suponha que no exemplo inicial existem 3 covariáveis, $p = 3$, e a quantidade de modelos possíveis neste cenário será $2^3 = 8$. Sendo assim, seria necessário fazer 8 ajustes para então escolher o "melhor"modelo dentre todos os possíveis. Se a quantidade de variáveis explicativas fosse maior? Por exemplo, se $p = 10$, existiriam $2^{10} = 1.024$ modelos para serem comparados e vale ressaltar que no cenário com 10 ou mais variáveis é muito comum em áreas como Biologia, Farmácia, Economia, etc.

Quando p é demasiadamente grande, torna-se interessante utilizar métodos de seleção de variáveis bayesianos, que podem ser mais práticos e menos trabalhosos. Alguns exemplos são o método de seleção de variáveis de Kuo & Mallick, o método de seleção de Gibbs, seleção de variáveis via busca estocástica, entre outros. Os métodos de seleção de variáveis bayesianos consistem em se fazer um único ajuste e possuem variáveis indicadoras que sinalizam se a variável explicativa foi ou não selecionada.

Voltando ao exemplo com 3 variáveis independentes e utilizando um dos métodos de seleção de variáveis bayesianos, no decorrer do processo será obtida a seguinte matriz com

amostras da distribuição *a posteriori* das variáveis indicadoras:

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \end{bmatrix},$$

em que a primeira linha $[1 \quad 0 \quad 0]$ indica que somente a variável x_1 foi selecionada. Agora, suponha que o modelo "correto"seja o modelo que inclui as variáveis 1 e 2. Sendo assim, é possível quantificar a porcentagem de vezes que cada variável é selecionada e, além disso, a porcentagem de vezes que o modelo "correto"é visitado.

Note que pode-se pensar no problema de seleção de variáveis como um problema de decidir quais parâmetros β_j's são iguais a zero. No contexto de distinguir entre efeitos grandes e pequenos, deve-se utilizar distribuições *a priori* que expressem a crença de que existem coeficientes grandes.

Alguns dos métodos estudados neste trabalho constroem distribuições *a priori* como uma mistura de duas distribuições, uma concentrada no zero e outra com massa espalhada em um intervalo de valores plausíveis. Elas são conhecidas como distribuições *spike* e *slab*, respectivamente .

Como pode-se ver, estas duas distribuições serão úteis para o propósito de selecionar variáveis, pois elas classificarão os coeficientes da regressão em dois grupos: o primeiro consiste dos regressores importantes e o segundo dos efeitos negligenciáveis.

Para apresentar os métodos que serão estudados neste trabalho,será necessário definir duas variáveis auxiliares para cada uma das variáveis explicativas. A primeira é a variável indicadora citada anteriormente, que será denotada por γ_j. Quando $\gamma_j = 1$, indica-se a presença, e se $\gamma_j = 0$ indica-se a ausência da j-ésima covariável e irá denotar quando a variável está na parte *slab* ou *spike* da distribuição *a priori*. A segunda variável auxiliar quantifica o efeito da j-ésima covariável denotada por ϕ_j. Tem-se que $\phi_j = \beta_j$, se $\gamma_j = 1$, para isso basta definir $\beta_j = \gamma_j \phi_j$.

As diferentes maneiras de gerenciar β_j, ϕ_j e γ_j definem os seguintes métodos de seleção de variáveis.

3.4.1 Kuo & Mallick (KM)

O primeiro método de seleção de subconjuntos de variáveis explicativas no modelo de regressão linear múltiplo que será estudado, trata-se de um método simples, apresentado por Kuo & Mallick (1998)[6] que foi motivado por George e McCulloch (1993)[16] . Esse método define $\beta_j = \phi_j\gamma_j$ e propõe assumir *a priori* que as duas variáveis auxiliares utilizadas para quantificar o efeito são independentes, logo $p(\phi_j, \gamma_j) = p(\phi_j)p(\gamma_j)$. Neste método, ϕ_j será sorteada da distribuição condicional completa, mas se $\gamma_j = 0$, o sorteio será da própria distribuição *a priori*.

Neste cenário, o modelo hierárquico poderá ser escrito como:

$$\begin{aligned}
Y_i &\sim N(\mu_i, \tau^{-1}), \quad i = 1, ..., n, \\
\mu_i &= \phi_1\gamma_1 x_{i1} + \phi_2\gamma_2 x_{i2} + ... + \phi_p\gamma_p x_{ip}, \quad i = 1, ..., n, \\
\phi_j &\sim N(\phi_{0j}, D_j), \quad j = 1, ..., p, \\
\gamma_j &\sim Bernoulli(p_j), \quad j = 1, ..., p, \\
\tau &\sim G(c, d).
\end{aligned}$$

Note que atribui-se uma distribuição *a priori* Normal Univariada para cada variável auxiliar (ϕ_j) com média ϕ_{0j} e variância D_j, logo pode se escrever $\boldsymbol{\phi} \sim N_p(\boldsymbol{\phi_0}, D_0)$, onde $\boldsymbol{\phi_0} = (\phi_{01}, \phi_{02}, ..., \phi_{0p})'$ e $D_0 = diag(D_1, D_2, ..., D_p)$, *Bernoulli* para as indicadoras de presença ou ausência das covariáveis (γ_j, $j = 1, ..., p$) e Gama para o parâmetro de precisão (τ).

Para fazer inferência no modelo apresentado acima, é necessário conhecer a função de verossimilhança, definida por:

$$\begin{aligned}
L(\boldsymbol{\phi}, \boldsymbol{\gamma}, \tau|\mathbf{Y}) &= |\tau|^{\frac{n}{2}}(2\pi)^{-\frac{n}{2}} exp\left\{-\frac{\tau}{2}(\mathbf{Y} - X^*\phi)'(\mathbf{Y} - X^*\phi)\right\} \\
&\quad \times I_{\mathbb{R}^p}(\boldsymbol{\phi})\ I_{\{0,1\}^p}(\boldsymbol{\gamma})\ I_{(0,+\infty)}(\tau),
\end{aligned} \tag{3.9}$$

em que $X^* = [X_1\gamma_1, ..., X_p\gamma_p]$ e $\{0,1\}^p = \{0,1\}^1 \times ... \times \{0,1\}^p$, além da distribuição *a priori* conjunta de $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\gamma}, \tau)'$, assumindo independência *a priori* para τ, $\boldsymbol{\phi}$ e $\boldsymbol{\gamma}$ que é dada por:

$$\begin{aligned}
p(\boldsymbol{\theta}) = p(\boldsymbol{\phi}, \boldsymbol{\gamma}, \tau) &\propto exp\left\{-\frac{1}{2}(\boldsymbol{\phi} - \boldsymbol{\phi_0})'\ D_0^{-1}\ (\boldsymbol{\phi} - \boldsymbol{\phi_0})\right\}\tau^{c-1} \\
&\quad \times e^{-d\tau}\prod_{j=1}^{p}\left[p_j^{\gamma_j}(1-p_j)^{1-\gamma_j}\ I_{\{0,1\}}(\gamma_j)\right]\ I_{\mathbb{R}^p}(\boldsymbol{\phi})\ I_{(0,+\infty)}(\tau).
\end{aligned} \tag{3.10}$$

A partir do Teorema de Bayes (3.3), da função de verossimilhança (3.9) e da distribuição *a priori* conjunta (3.10) apresentados acima, pode-se escrever a distribuição *a posteriori* conjunta como:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{Y}) = p(\boldsymbol{\phi},\boldsymbol{\gamma},\tau|\mathbf{Y}) \quad \propto \quad & |\tau|^{\frac{n}{2}} exp\left\{-\frac{\tau}{2}(\mathbf{Y}-X^*\boldsymbol{\phi})'(\mathbf{Y}-X^*\boldsymbol{\phi})\right\} \\ & \times exp\left\{-\frac{1}{2}(\boldsymbol{\phi}-\boldsymbol{\phi_0})'\ D_0^{-1}\ (\boldsymbol{\phi}-\boldsymbol{\phi_0})\right\}\tau^{c-1} \\ & \times e^{-d\tau}\prod_{j=1}^{p}\left[p_j^{\gamma_j}(1-p_j)^{1-\gamma_j}\ I_{\{0,1\}}(\gamma_j)\right]\ I_{\mathbb{R}^p}(\boldsymbol{\phi})\ I_{(0,+\infty)}(\tau). \end{aligned} \tag{3.11}$$

Depois de especificar a distribuição *a posteriori* conjunta (3.11), pode-se agora escrever as distribuições condicionais completas que serão utilizadas no amostrador de Gibbs.

A distribuição condicional completa de $\boldsymbol{\phi}$ é dada por:

$$p(\boldsymbol{\phi}|\tau,\boldsymbol{\gamma},\mathbf{Y}) \propto exp\left\{-\frac{1}{2}\left[\tau\ (\mathbf{Y}-X^*\boldsymbol{\phi})'(\mathbf{Y}-X^*\boldsymbol{\phi}) + (\boldsymbol{\phi}-\boldsymbol{\phi_0})'\ D_0^{-1}\ (\boldsymbol{\phi}-\boldsymbol{\phi_0})\right]\right\}\ I_{\mathbb{R}^p}(\boldsymbol{\phi}).$$

Após fazer algumas operações algébricas é possível observar que $\boldsymbol{\phi}$ tem distribuição conhecida Normal Multivariada de ordem p, dada por:

$$\boldsymbol{\phi}|\ \tau,\boldsymbol{\gamma},\mathbf{Y} \sim N_p\left(\frac{\tau X^{*'}\mathbf{Y} + D_0^{-1}\boldsymbol{\phi_0}}{\tau X^{*'}X^* + D_0^{-1}},\ (\tau X^{*'}X^* + D_0^{-1})^{-1}\right).$$

A distribuição condicional completa da precisão, τ, é representada por:

$$p(\tau|\boldsymbol{\gamma},\boldsymbol{\phi},\mathbf{Y}) \propto \tau^{(\frac{n}{2}+c)-1} exp\left\{-\tau\left[\frac{1}{2}(\mathbf{Y}-X^*\boldsymbol{\phi})'(\mathbf{Y}-X^*\boldsymbol{\phi}) + d\right]\right\}\ I_{(0,\infty)}(\tau).$$

Note que τ tem distribuição conhecida dada por:

$$\tau|\boldsymbol{\gamma},\boldsymbol{\phi},\mathbf{Y}\ \sim\ G\left(\frac{n}{2}+c,\ \frac{1}{2}(\mathbf{Y}-X^*\boldsymbol{\phi})'(\mathbf{Y}-X^*\boldsymbol{\phi}) + d\right).$$

A distribuição condicional completa de γ_j é expressa por:

$$\gamma_j|\boldsymbol{\phi},\tau,\mathbf{Y} \sim Bernoulli(\ \tilde{p_j}\ ),$$

com $\tilde{p_j} = \dfrac{e_j}{(e_j+f_j)}$, em que:

$$e_j = p_j exp\left\{-\frac{\tau}{2}(\mathbf{Y}-X\boldsymbol{\phi}^*)'(\mathbf{Y}-X\boldsymbol{\phi}^*)\right\}\quad I_{\{1\}}(\gamma_j)$$

e

$$f_j = (1-p_j)exp\left\{-\frac{\tau}{2}(\mathbf{Y}-X\boldsymbol{\phi}^{**})'(\mathbf{Y}-X\boldsymbol{\phi}^{**})\right\} \quad I_{\{0\}}(\gamma_j),$$

no qual $\boldsymbol{\phi}^{*} = \boldsymbol{\phi}|\gamma = 1, \boldsymbol{\phi}^{*} = [\phi_1, ..., \phi_j]$ e $\boldsymbol{\phi}^{**} = \boldsymbol{\phi}|\gamma = 0, \boldsymbol{\phi}^{**} = [0, ..., 0]$.

Observa-se que para o método KM, ϕ_j e γ_j são considerados independentes *a priori*, já para os métodos que serão abordados a seguir isso não ocorrerá, ϕ_j e γ_j serão considerados dependentes.

3.4.2 Seleção de variáveis de Gibbs (GVS)

O segundo método a ser apresentado é conhecido como seleção de variáveis de Gibbs (do inglês, Gibbs Variable Selection - GVS) e foi desenvolvido a partir da ideia de Carlin e Chibi (1995)[17] por Dellaportas et al. (1998)[18]. O método também define $\beta_j = \phi_j\gamma_j$. Carlin e Chibi (1995) sugerem um método em que se faz pequenas modificações no amostrador de Gibbs de acordo com determinadas situações, com objetivo de resolver os problemas de convergência das cadeias de Markov.

Para $\phi_j|\gamma_j = 1$, ϕ_j será amostrado igual ao método citado anteriormente. Para contornar o problema de uma distribuição *a priori* muito vaga para $\phi_j|\gamma_j = 0$ como no método anterior, em que $\phi_j|\gamma_j = 0$ e $\phi_j|\gamma_j = 1$ eram selecionados de uma mesma distribuição *a priori*, o GVS sugere empregar uma pseudo-*priori*; favorecendo saltos entre as iterações do MCMC. Este método assume dependência entre as distribuições *a priori* de ϕ_j e γ_j, ou seja, $p(\phi_j, \gamma_j) = p(\phi_j|\gamma_j)p(\gamma_j)$. Será adotado uma distribuição mista *a priori* para ϕ_j dada por:

$$p(\phi_j|\gamma_j) = (1-\gamma_j)N(\tilde{\mu}_j, S_j) + \gamma_j N(0, \nu^2), \tag{3.12}$$

em que a distribuição *a priori slab* é uma distribuição normal com média 0 e variância ν^2 e a distribuição *a priori spike* é uma distribuição normal com média *a posteriori* de β_j, $\tilde{\mu}_j$, e variância *a posteriori* de β_j, S_j, por exemplo.

Neste cenário, o modelo hierárquico poderá ser escrito como:

$$\begin{aligned}
Y_i &\sim N(\mu_i, \tau^{-1}), \quad i = 1, ..., n, \\
\mu_i &= \phi_1\gamma_1 x_{i1} + \phi_2\gamma_2 x_{i2} + ... + \phi_p\gamma_p x_{ip}, \quad i = 1, ..., n, \\
\phi_j|\gamma_j &\sim (1-\gamma_j)N(\tilde{\mu}_j, S_j) + \gamma_j N(0, \nu^2), \quad j = 1, ..., p, \\
\gamma_j &\sim Bernoulli(p_j), \quad j = 1, ..., p, \\
\tau &\sim G(c, d).
\end{aligned}$$

Note que atribuiu-se as mesmas distribuições *a priori* utilizada no KM para γ_j e τ. Para fazer inferência no modelo apresentado acima, é necessário conhecer a função de verossimilhança que será a mesma do KM, dada por (3.9) e a distribuição *a priori* conjunta de $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\gamma}, \tau)'$, assumindo dependência *a priori* entre ϕ_j e γ_j, dada por:

$$\begin{aligned} p(\boldsymbol{\theta}) \;=\; p(\boldsymbol{\phi}, \boldsymbol{\gamma}, \tau) \;\propto\; & \prod_{j=1}^{p} \left[(1-p_j) \frac{1}{\sqrt{S_j}} exp\left\{ -\frac{1}{2S_j} (\phi_j - \tilde{\mu_j})^2 \right\} I_{\{0\}}(\gamma_j) \right. \\ & \left. + \, p_j \frac{1}{\nu^2} exp\left\{ -\frac{1}{2\nu^2} \phi_j^2 \right\} I_{\{1\}}(\gamma_j) \right] \\ & \times \, \tau^{c-1} \; e^{-d\tau} \; I_{\mathbb{R}^p}(\boldsymbol{\phi}) \; I_{(0,+\infty)}(\tau). \end{aligned} \tag{3.13}$$

A partir do Teorema de Bayes (3.3), da função de verossimilhança (3.9) e da distribuição *a priori* conjunta (3.13) apresentados acima, pode-se escrever a distribuição *a posteriori* conjunta como:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{Y}) = p(\boldsymbol{\phi}, \boldsymbol{\gamma}, \tau|\mathbf{Y}) \;\propto\; & |\tau|^{\frac{n}{2}} exp\left\{ -\frac{\tau}{2} (\mathbf{Y} - X^* \boldsymbol{\phi})'(\mathbf{Y} - X^* \boldsymbol{\phi}) \right\} \\ & \times \prod_{j=1}^{p} \left[(1-p_j) \frac{1}{\sqrt{S_j}} exp\left\{ -\frac{1}{2S_j} (\phi_j - \tilde{\mu_j})^2 \right\} I_{\{0\}}(\gamma_j) \right. \\ & \left. + \, p_j \; \frac{1}{\nu^2} \; exp\left\{ -\frac{1}{2\nu^2} \; \phi_j^2 \right\} I_{\{1\}}(\gamma_j) \right] \\ & \times \, \tau^{c-1} \; e^{-d\tau} \; I_{\mathbb{R}^p}(\boldsymbol{\phi}) \; I_{(0,+\infty)}(\tau). \end{aligned} \tag{3.14}$$

Depois de especificar a distribuição *a posteriori* conjunta (3.14), pode-se agora escrever as distribuições condicionais completas que serão utilizadas.

A distribuição condicional completa de $\boldsymbol{\phi}$ é desconhecida e dada por:

$$\begin{aligned} p(\boldsymbol{\phi}|\tau, \gamma, Y) \;=\; & exp\left\{ -\frac{\tau}{2} (Y - X^* \boldsymbol{\phi})'(Y - X^* \boldsymbol{\phi}) \right\} \\ & \times \prod_{j=1}^{p} \left[(1-p_j) \frac{1}{\sqrt{S_j}} exp\left\{ -\frac{1}{2S_j} (\phi_j - \tilde{\mu_j})^2 \right\} I_{\{0\}}(\gamma_j) \right. \\ & \left. + \, p_j \; \frac{1}{\nu^2} \; exp\left\{ -\frac{1}{2\nu^2} \; \phi_j^2 \right\} I_{\{1\}}(\gamma_j) \right] I_{\mathbb{R}^p}(\boldsymbol{\phi}). \end{aligned}$$

A distribuição condicional completa da precisão, τ, é conhecida e equivale a representada no KM por:

$$\tau|\gamma, \phi, Y \;\sim\; G\left(\frac{n}{2} + c, \; \frac{1}{2} (Y - X^* \boldsymbol{\phi})'(Y - X^* \boldsymbol{\phi}) + d \right).$$

A distribuição condicional completa de γ_j é desconhecida e pode ser expressa por:

$$\begin{aligned} p(\boldsymbol{\gamma}|\phi,\tau,Y) &= exp\left\{-\frac{\tau}{2}(Y-X^*\phi)'(Y-X^*\phi)\right\} \\ &\times \prod_{j=1}^{p}\left[(1-p_j)\frac{1}{\sqrt{S_j}}exp\left\{-\frac{1}{2S_j}(\phi_j-\tilde{\mu_j})^2\right\}I_{\{0\}}(\gamma_j)\right. \\ &\left.+\ p_j\ \frac{1}{\nu^2}\ exp\left\{-\frac{1}{2\nu^2}\ \phi_j^2\right\}I_{\{1\}}(\gamma_j)\right]. \end{aligned}$$

3.4.3 Seleção de variáveis via busca estocástica (SSVS)

O método de seleção de variáveis via busca estocástica (do inglês, Stochastic Search Variable Selection - SSVS) foi inicialmente apresentado por George and McCulloch (1993)[16] e adaptado para vários outros tipos de modelos, inclusive para os modelos de regressão múltiplos por Brown et al. (1997)[19]. Diferentemente dos métodos apresentados anteriormente, defini-se $\beta_j = \phi_j$, mas a dependência entre ϕ_j e γ_j é assumida, ou seja, $p(\phi_j, \gamma_j) = p(\phi_j|\gamma_j)p(\gamma_j)$. Será adotada uma distribuição mista *a priori* para ϕ_j dada por:

$$p(\phi_j|\gamma_j) = (1-\gamma_j)N(0,\psi^2) + \gamma_j N(0, g\psi^2), \tag{3.15}$$

no qual a primeira densidade é a distribuição *a priori spike* e a segunda a distribuição *a priori slab*. Pela construção da distribuição *a priori* acima, percebe-se que ψ^2 deve assumir valores pequenos e g é o responsável por "espalhar"a distribuição. Em contrapartida com o GVS, neste método os valores *a priori* dos parâmetros afetam *a posteriori* quando γ_j for igual a zero.

No SSVS como μ_i não depende da variável indicadora, quando assumi-se $\gamma_j = 0$ a primeira parte da distribuição *a posteriori* que é a função de verossimilhança, continua sendo a mesma com a presença da variável auxiliar do efeito ϕ. Já para o GVS a função verossimilhança depende da variável indicadora e quando esta assumi o valor 0 a variável responsável por quantificar o efeito some, fazendo com que os valores *a priori* dos parâmetros não afetem *a posteriori* quando γ_j for igual a zero.

Neste cenário, o modelo hierárquico poderá ser escrito como:

$$\begin{aligned} Y_i &\sim N(\mu_i, \tau^{-1}), \quad i = 1, ..., n, \\ \mu_i &= \phi_1 x_{i1} + \phi_2 x_{i2} + ... + \phi_p x_{ip}, \quad i = 1, ..., n, \\ \phi_j|\gamma_j &\sim (1-\gamma_j)N(0, \psi^2) + \gamma_j N(0, g\psi^2), \quad j = 1, ..., n, \\ \gamma_j &\sim Bernoulli(p_j), \quad j = 1, ..., n, \\ \tau &\sim G(c, d). \end{aligned}$$

Note que atribuiu-se as mesmas distribuições *a priori* utilizadas nos métodos anteriores para γ_j e τ. Para fazer inferência no modelo apresentado acima, é necessário conhecer a função de verossimilhança que diferentemente dos outros métodos será:

$$\begin{aligned} L(\boldsymbol{\phi}, \boldsymbol{\gamma}, \tau|\mathbf{Y}) &= |\tau|^{\frac{n}{2}}(2\pi)^{-\frac{n}{2}} exp\left\{-\frac{\tau}{2}(\mathbf{Y} - X\boldsymbol{\phi})'(\mathbf{Y} - X\boldsymbol{\phi})\right\} \\ &\quad \times I_{\mathbb{R}^p}(\boldsymbol{\phi})\ I_{\{0,1\}^p}(\boldsymbol{\gamma})\ I_{(0,+\infty)}(\tau). \end{aligned} \tag{3.16}$$

Com base na função de verossimilhança dada por (3.16) e a distribuição *a priori* conjunta de $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\gamma}, \tau)'$, assumindo dependência *a priori* entre $\boldsymbol{\phi}$ e $\boldsymbol{\gamma}$, dada por:

$$\begin{aligned} p(\boldsymbol{\theta}) = p(\boldsymbol{\phi}, \boldsymbol{\gamma}, \tau) &\propto \prod_{j=1}^{p}\left[(1-p_j)\frac{1}{\psi^2} exp\left\{-\frac{1}{2\psi^2}\phi_j^2\right\} I_{\{0\}}(\gamma_j)\right. \\ &\quad \left. + p_j \frac{1}{\psi^2\sqrt{g}} exp\left\{-\frac{1}{2g\psi^2}\phi_j^2\right\} I_{\{1\}}(\gamma_j)\right] \\ &\quad \times \tau^{c-1}\ e^{-d\tau}\ I_{\mathbb{R}^p}(\boldsymbol{\phi})\ I_{(0,+\infty)}(\tau). \end{aligned} \tag{3.17}$$

A partir do Teorema de Bayes (3.3), da função de verossimilhança (3.9) e da distribuição *a priori* conjunta (3.17) apresentados acima, pode-se escrever a distribuição *a posteriori* conjunta, como:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{Y}) = p(\boldsymbol{\phi}, \boldsymbol{\gamma}, \tau|\mathbf{Y}) &\propto |\tau|^{\frac{n}{2}} exp\left\{-\frac{\tau}{2}(\mathbf{Y} - X\boldsymbol{\phi})'(\mathbf{Y} - X\boldsymbol{\phi})\right\} \\ &\quad \times \prod_{j=1}^{p}\left[(1-p_j)\frac{1}{\psi^2} exp\left\{-\frac{1}{2\psi^2}\phi_j^2\right\} I_{\{0\}}(\gamma_j)\right. \\ &\quad \left. + p_j \frac{1}{\psi^2\sqrt{g}} exp\left\{-\frac{1}{2g\psi^2}\phi_j^2\right\} I_{\{1\}}(\gamma_j)\right] \\ &\quad \times \tau^{c-1}\ e^{-d\tau}\ I_{\mathbb{R}^p}(\boldsymbol{\phi})\ I_{(0,+\infty)}(\tau). \end{aligned} \tag{3.18}$$

Depois de especificar a distribuição *a posteriori* conjunta (3.18), pode-se agora escrever as distribuições condicionais completas que serão utilizadas.

A distribuição condicional completa de $\boldsymbol{\phi}$ é dada por:

$$\begin{aligned} p(\boldsymbol{\phi}|\tau,\boldsymbol{\gamma},\mathbf{Y}) &= exp\left\{-\frac{\tau}{2}(\mathbf{Y}-X\boldsymbol{\phi})'(\mathbf{Y}-X\boldsymbol{\phi})\right\} \\ &\quad \times \prod_{j=1}^{p}\left[(1-p_j)\frac{1}{\psi^2}exp\left\{-\frac{1}{2\psi^2}\phi_j^2\right\}I_{\{0\}}(\gamma_j)\right. \\ &\quad \left. + p_j\frac{1}{\psi^2\sqrt{g}}exp\left\{-\frac{1}{2g\psi^2}\phi_j^2\right\}I_{\{1\}}(\gamma_j)\right]\ I_{\mathbb{R}^p}(\boldsymbol{\phi}). \end{aligned}$$

A distribuição condicional completa da precisão, τ, é conhecida e dada por:

$$\tau|\boldsymbol{\gamma},\boldsymbol{\phi},\mathbf{Y}\ \sim\ G\left(\frac{n}{2}+c,\ \frac{1}{2}(\mathbf{Y}-X\boldsymbol{\phi})'(\mathbf{Y}-X\boldsymbol{\phi})+d\right).$$

A distribuição condicional completa de γ_j é desconhecida e pode ser expressa por:

$$\begin{aligned} p(\boldsymbol{\gamma}|\boldsymbol{\phi},\tau,\mathbf{Y}) &= \prod_{j=1}^{p}\left[(1-p_j)\frac{1}{\sqrt{S_j}}exp\left\{-\frac{1}{2S_j}(\phi_j-\tilde{\mu_j})^2\right\}I_{\{0\}}(\gamma_j)\right. \\ &\quad \left. + p_j\ \frac{1}{\psi^2}\ exp\left\{-\frac{1}{2\psi^2}\ \phi_j^2\right\}I_{\{1\}}(\gamma_j)\right]. \end{aligned}$$

Após a apresentação dos 3 métodos para modelos de regressão múltiplo, pode-se perceber que nos métodos KM e GVS define-se $\beta_j = \phi_j\gamma_j$, para o KM ϕ_j e γ_j são considerados independentes, já no GVS ϕ_j e γ_j são considerados dependentes e utiliza-se uma pseudo-*priori* para $\phi_j|\gamma_j = 0$ que não influência na distribuição *a posteriori*. Por outro lado no SSVS defini-se $\beta_j = \phi_j$ e isso influência na distribuição *a posteriori*. No GVS e no SSVS a distribuição *a priori* para $\phi_j|\gamma_j$ tem uma distribuição mista composta pela distribuição *a priori spike* e a distribuição *a priori slab*, já no KM não se tem isso. Para os três métodos a distribuição *a priori* para $\phi_j|\gamma_j = 1$ é a mesma.

4 Análise dos Resultados

Neste capítulo serão apresentados os resultados obtidos a partir de um estudo de simulação realizado para os métodos de seleção de variáveis bayesianos do modelo de regressão linear múltiplo apontado no capítulo anterior.

4.1 Definição dos cenários

Definiu-se dois cenários com a presença/ausência de multicolinearidade, com n=100 e os erros normalmente distribuídos.

Primeiramente determinou-se dois cenários de interesse a fim de se investigar a eficiência dos métodos de seleção estudados no capítulo anterior. Os dois cenários investigados são caracterizados pela presença/ausência de dependência entre as variáveis explicativas. Foram feitas 100 replicações, em cada cenário, com $n = 100$ em que os erros são normalmente distribuídos. Utilizou-se uma cadeia, com um burning de 3000 e uma amostra final de 2000.

A variável resposta (Y_i) foi gerada como a seguir:

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_{10} x_{i10} + \epsilon_i,\ i = 1, 2, ..., n,$$

em que, $\boldsymbol{\beta}$=(2; 1; 1, 5; 0, 5; 0; 0; 0; 0; 0; 0), $p = 10$ e $\epsilon_i \sim N(0, \tau^{-1})$. O valor assumido para τ foi 1.

No primeiro cenário, chamado aqui de **independente**, x_i, $i = 1, ..., 10$, foram gerados de uma $N(0, 1)$. Já no segundo, chamado aqui de **dependente**, o vetor de covariáveis foi gerado por meio de uma distribuição normal com correlações entre x_i e x_j, dadas por $0, 5^{|i-j|}$, $i, j = 1, ..., 10$.

Com o intuito de verificar se existia influência dos valores inicias escolhidos para a variável indicadora, investigou-se 4 casos para cada método sendo o 1º caso com todos os valores iniciais para $\boldsymbol{\gamma}$ iguais a zero, o 2º com todas iguais a 1, o 3º com $\gamma_1,\ \gamma_2,\ \gamma_3\ e\ \gamma_4$

iguais a zero e o resto igual a um e o último caso com $\gamma_1,\ \gamma_2,\ \gamma_3$ *e* γ_4 inicializando em um e os outros em zero. Logo, defini-se:

$$
\begin{aligned}
1^{o})\ &\gamma_j = (0,0,0,0,0,0,0,0,0,0)';\\
2^{o})\ &\gamma_j = (1,1,1,1,1,1,1,1,1,1)';\\
3^{o})\ &\gamma_j = (0,0,0,0,1,1,1,1,1,1)';\\
4^{o})\ &\gamma_j = (1,1,1,1,0,0,0,0,0,0)'.
\end{aligned}
$$

Foi avaliado também o tempo de iteração para cada caso a fim de avaliar qual o método mais rápido.

4.1.1 Convergência

Para verificar a convergência da cadeia analisou-se para cada cenário em cada método uma cadeia simulada para a variável auxiliar responsável por quantificar o efeito conhecida como ϕ_j, sendo que observou-se ϕ_1 e ϕ_5, cujos valores verdadeiros são 2 e 0 respectivamente, e para τ o valor verdadeiro é 1. Observa-se para o cenário independente as figuras a seguir:

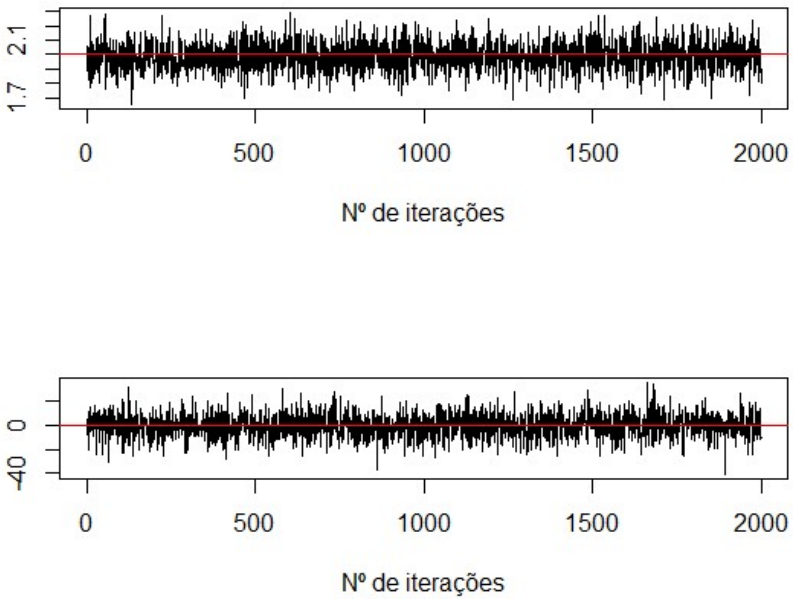

Figura 1: Cenário Independente - Cadeia para ϕ_1 e ϕ_5 no método KM

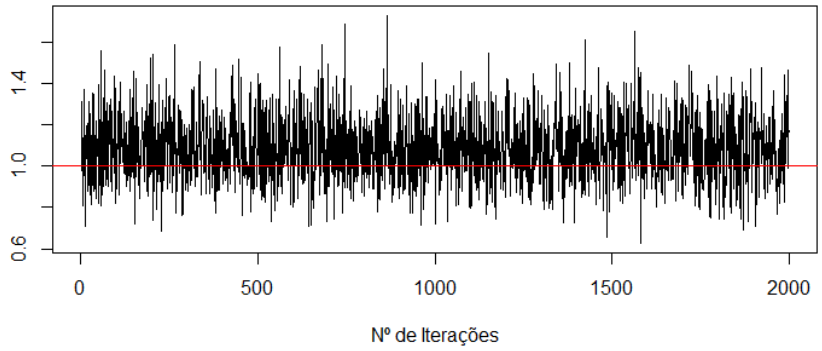

Figura 2: Cenário Independente - Cadeia para τ no método KM

A partir da Figura 1 percebe-se que ambos, ϕ_1 e ϕ_5, convergem para os seus valores verdadeiros e isso se repete para todos os métodos. Já na Figura 2 apesar de convergir e o método ter superestimado τ, o intervalo de credibilidade simétrico de 95% para τ contém o seu verdadeiro valor (1), $[0,797\ ;\ 1,42]$. Como assumiu-se o mesmo valor de τ o comportamento foi o mesmo no cenário dependente.

Agora olhando para o cenário dependente tem-se:

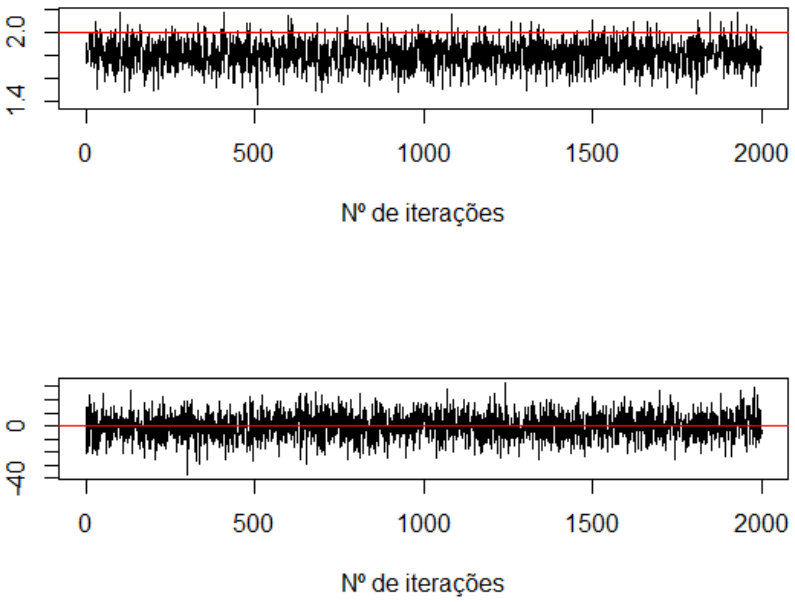

Figura 3: Cenário Dependente - Cadeia para ϕ_1 e ϕ_5 no método KM

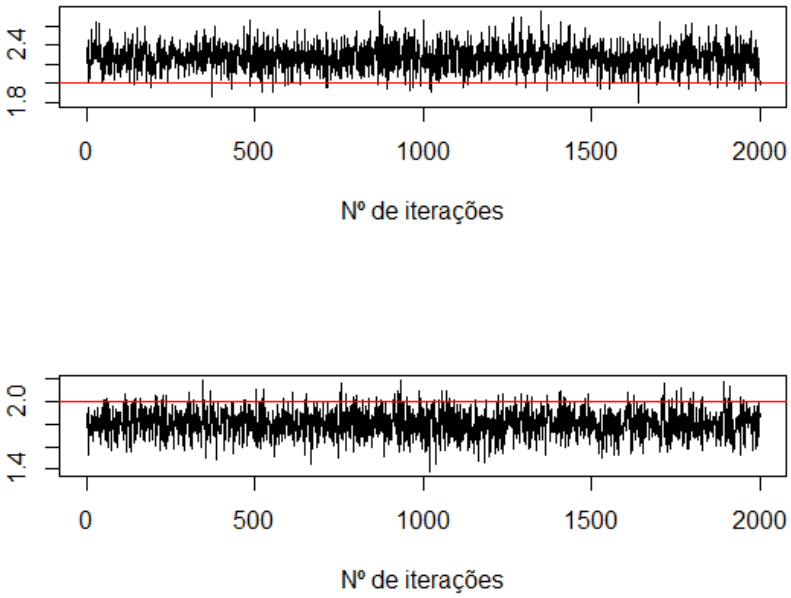

Figura 4: Cenário Dependente - Cadeia para ϕ_1 no método SSVS para o 1º e 2º caso

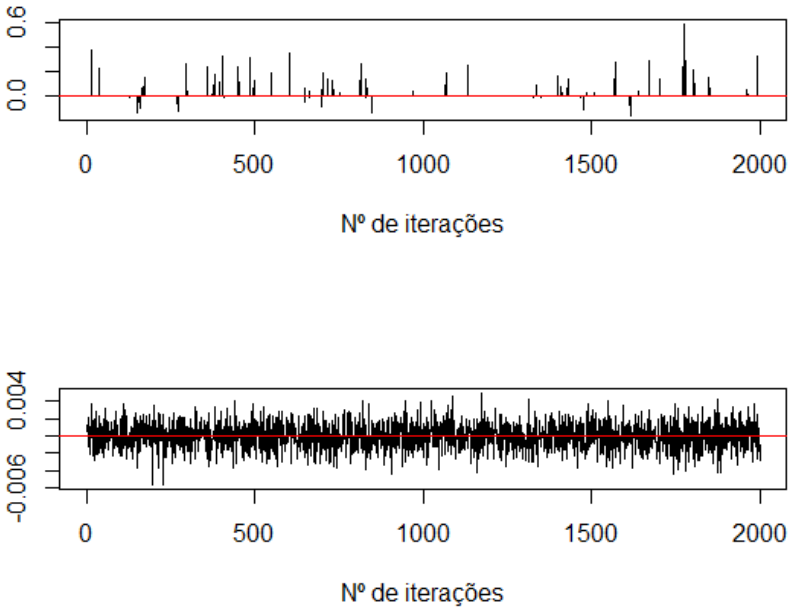

Figura 5: Cenário Dependente - Cadeia para ϕ_5 no método SSVS para o 1º e 2º caso

Observando a Figura 3 percebe-se que há convergência e uma subestimação de ϕ_1, mas o intervalo de credibilidade de 95% contém o valor verdadeiro, $[1,579\ ;\ 2,031]$. Já para ϕ_5 a cadeia converge para o valor verdadeiro, 0. Para todos os casos de inicialização

de γ_j e para o método GVS o comportamento é semelhante.

Agora, olhando pra Figura 4 e 5 tem-se que para o método SSVS, quando inicializa-se a variável indicadora em 0 há uma superestimação e quando se inicializa em 1 tem-se uma subestimação de ϕ_1 e o intervalo de credibilidade simétrico de 95%, $[2,011\ ;\ 2,535]$, não contém o valor verdadeiro de ϕ_1 no primeiro caso, porém apresentou um limite inferior próximo a este valor. Já no segundo caso para ϕ_1 o intervalo de credibilidade simétrico de 95% é $[1,573\ ;\ 2,037]$ e contém o valor verdadeiro. Quanto ao ϕ_5 percebe-se que ele converge em ambos os casos para o seu valor verdadeiro.

4.2 Distribuição *a priori* utilizada nas análises

Para todos os métodos foram assumidas as seguintes distribuições a *priori*:

$$\begin{aligned} \phi_j|\gamma_j = 1 &\sim N(0,100), \quad j = 1,...,10, \\ \gamma_j &\sim Bernoulli(0,5), \quad j = 1,...,10, \\ \tau &\sim G(10^{-4}, 10^{-4}). \end{aligned} \tag{4.1}$$

Considerou-se uma *Bernoulli* de parâmetro 0,5 para as variáveis indicadoras seguindo a sugestão de George e McCulloch (1998), tornando todos os modelos equiprováveis.

A distribuição *a priori* de $\phi_j|\gamma_j = 0$ depende do método de seleção de variáveis que for utilizado.

4.3 Implementação dos modelos no OpenBugs

As simulações neste trabalho foram implementadas no *OpenBugs*3.2.3, em que, utilizou-se o pacote "*R2OpenBUGS*" no software *R* para integrar o R com o *OpenBugs*.

O modelo implementado para o método de seleção de variáveis de **Kuo & Mallick** utiliza uma indicadora γ_j que assume os valores 0 ou 1 e ajusta $\beta_j = \gamma_j\phi_j$. Em que:

$$\beta_j = \begin{cases} 0, & se\ \gamma_j = 0, \\ \phi_j, & se\ \gamma_j = 1. \end{cases}$$

Assumindo a mesma distribuição *a priori* para ϕ_j independente do valor assumido para γ_j, $\phi_j|\gamma_j \sim N(0,100), \quad j = 1,...,10.$

Já o modelo implementado para o método de seleção de variáveis de **Gibbs**, quando

$\gamma_j = 0$ é necessário utilizar-se uma pseudo-*priori* que neste estudo será $\phi_j|\gamma_j = 0 \sim N(0; 0,25)$,no qual fixou-se os parâmetros do mesmo modo que feito em O'Hara e Sillanpää (2009)[7].

Para o modelo implementado para o método de seleção **Via Busca Estocástica** quando $\gamma_j = 0$ a distribuição *a priori* de ϕ_j foi construída assim como em O'Hara e Sillanpää [7] de maneira que a $P(|\phi_j| < k) < 0,01$ esteja a 3 desvios padrões da média, ou seja, $\phi_j|\gamma_j = 0 \sim N(0; (3 \times 0,0005)^2)$. Já para $\gamma_j = 1$, $\phi_j|\gamma_j = 1 \sim N(0; g\psi^2)$, em que $g\psi^2 = 100$.

4.4 Resultados

Nesta seção serão apresentados os resultados separados por cenário. Em cada cenário serão avaliados os 4 casos para os valores iniciais de γ_j apresentados anteriormente.

4.4.1 Cenário Independente

Neste cenário todas as variáveis explicativas (x_i) são consideradas independentes. Na Figura 6, é apresentada a probabilidade *a posteriori* das variáveis indicadoras para os três métodos e na Tabela 1 algumas medidas descritivas de interesse.

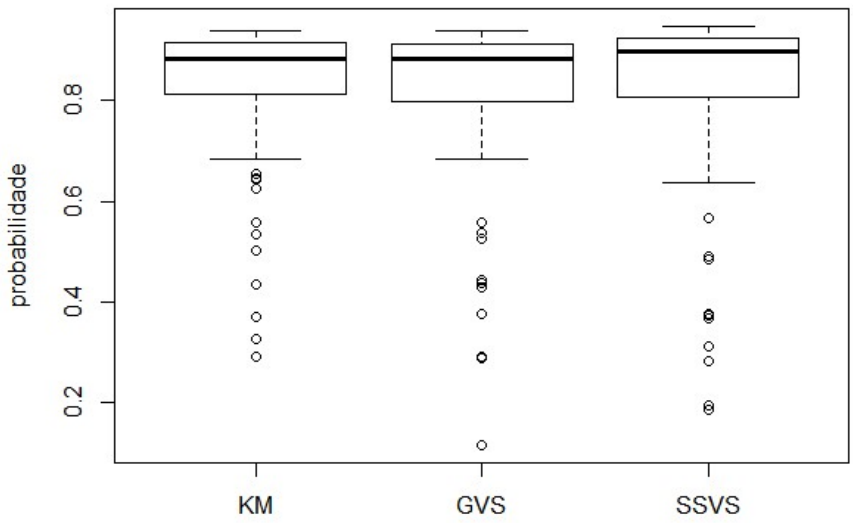

Figura 6: Cenário Independente - Box-plot da probabilidade *a posteriori* das variáveis indicadoras para o 4º caso.

Pode-se observar na Figura 6 e na Tabela 1 que os 3 métodos apresentam alta probabilidade *a posteriori* para o modelo correto e que o SSVS foi o método que apresentou

Tabela 1: Cenário Independente - Medidas Descritivas da probabilidade *a posteriori* das variáveis indicadoras do 4º caso.

Métodos	**probabilidade - 4º caso**		
	1º Quartil	Mediana	3º Quartil
KM	0,8161	0,8838	0,9159
GVS	0,8005	0,8815	0,9120
SSVS	0,8101	0,8975	0,9231

resultados levemente superiores, ou seja, 89,75% das vezes o método escolheu o modelo correto, isto é, somente com as quatro primeiras covariáveis. Vale ressaltar que para todos os casos da inicialização de γ_j adotados o comportamento foi o mesmo, logo, pode-se dizer que os valores iniciais não influenciam na proporção de vezes que o modelo correto é visitado pela cadeia de *Markov* e por isso foi apresentado somente o 4º caso. Observe agora na Figura 7 o tempo de iteração em segundos para a probabilidade *a posteriori* das variáveis indicadoras em todos os métodos e os diferentes casos de γ_j:

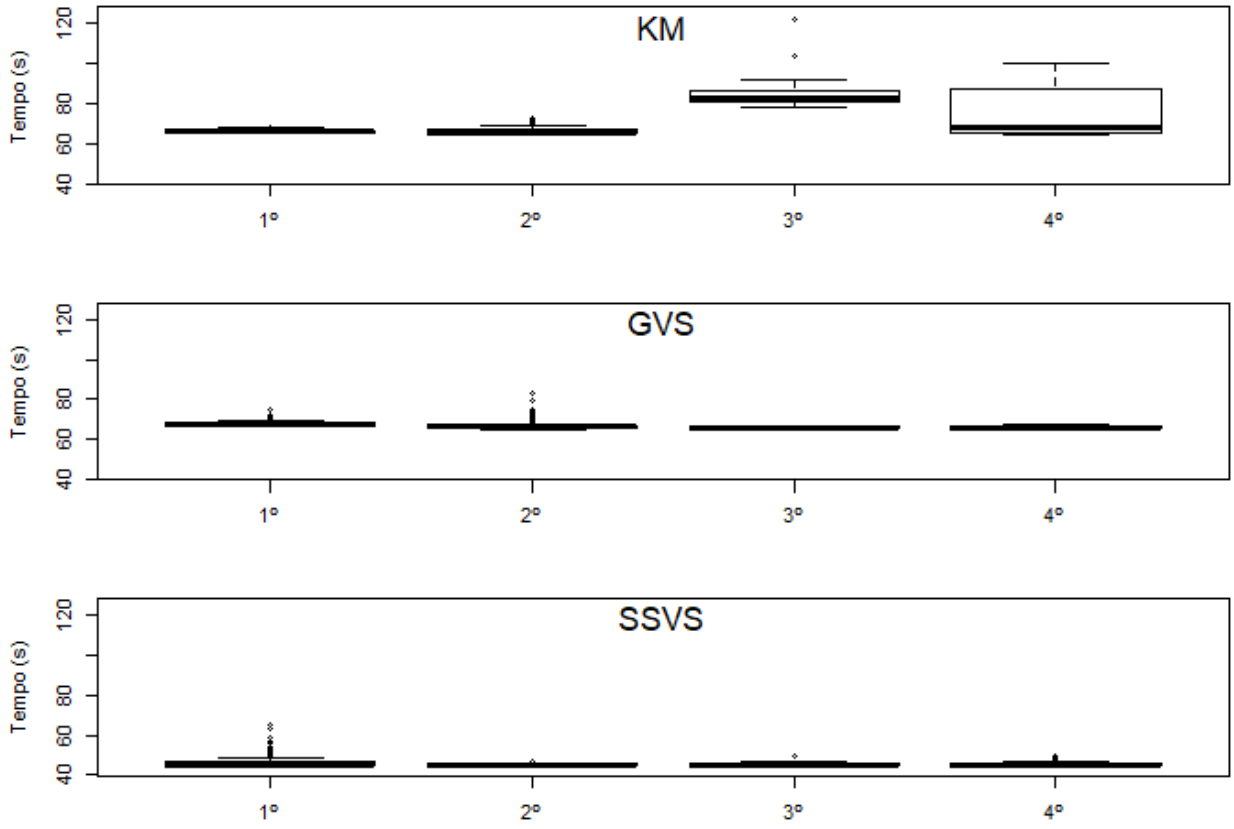

Figura 7: Cenário Independente - Box-plot do Tempo de iteração (em segundos

.

Ao comparar os tempos entre os casos para cada método representado na Figura 7, percebe-se que existe uma variação no tempo com relação aos valores iniciais utilizados,

sendo mais significativo entre os casos no método Kuo & Mallick. Ao analisar o gráfico percebe-se claramente que o método do SSVS é o mais rápido dentre os métodos comparados, diminuindo consideravelmente o tempo de ajuste do modelo com relação ao KM, principalmente no 3º e 4º casos.

4.4.2 Cenário Dependente

Neste cenário existe multicolinearidade, ou seja, há correlação entre as variáveis. Este cenário foi pensado com o intuito de verificar se a presença de multicolinearidade afeta na escolha dos métodos estudados. Na Figura 8, assim como no cenário anterior é apresentada a probabilidade *a posteriori* das variáveis indicadoras para os métodos e na Tabela 2 algumas estatísticas descritivas.

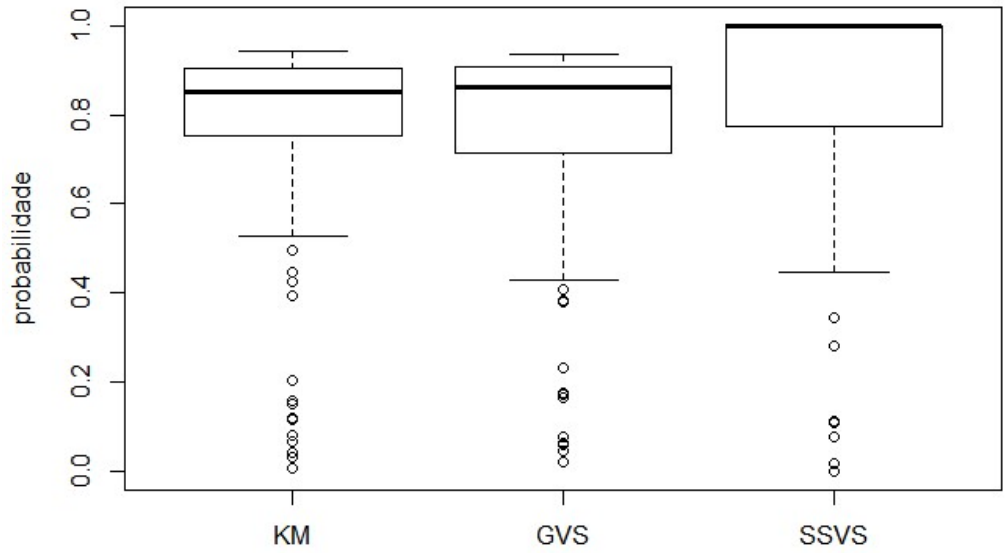

Figura 8: Cenário Dependente - Box-plot da probabilidade *a posteriori* das variáveis indicadoras para o 4º caso

Tabela 2: Cenário Dependente - Medidas Descritivas da probabilidade *a posteriori* das variáveis indicadoras para o 4º caso.

Métodos	**probabilidade - 4º caso**		
	1º Quartil	Mediana	3º Quartil
KM	0,7599	0,7490	0,9051
GVS	0,7147	0,86	0,9075
SSVS	0,7799	1	1

Pode-se observar na Figura 8 e na Tabela 2 que a probabilidade *a posteriori* do modelo correto ser escolhido para os três métodos continua sendo alta, porém com uma leve queda para o KM e o GVS, entretanto melhorou para o SSVS. Logo para este cenário

o método de seleção de variáveis via busca estocástica também foi considerado o melhor método, escolhendo 100% das vezes o modelo correto. Logo, observa-se que a presença de multicolinearidade influência na probabilidade *a posteriori*, principalmente no SSVS.

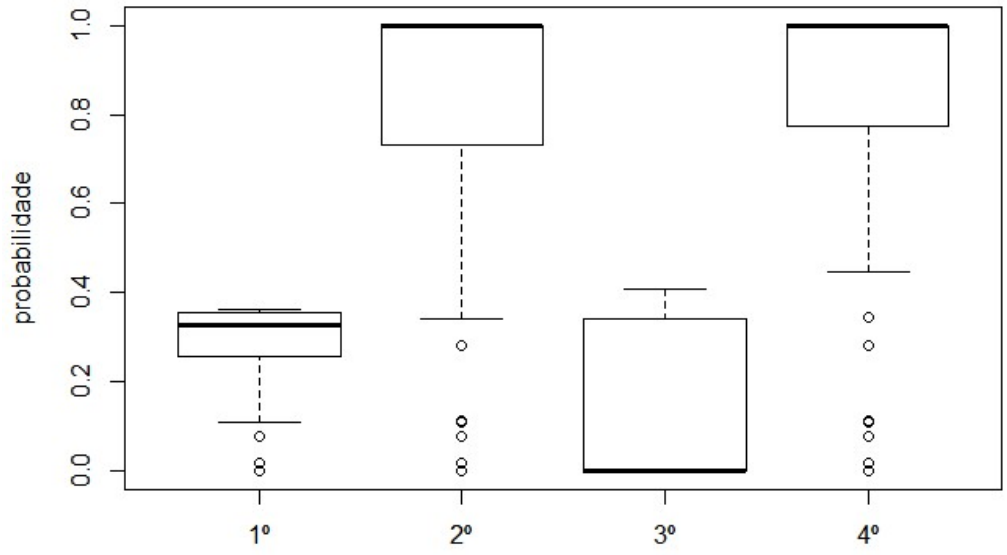

Figura 9: Cenário Dependente - Box-plot da probabilidade *a posteriori* das variáveis indicadoras em todos os casos no SSVS

Curiosamente para o SSVS os valores inciais influenciam na chance do modelo certo ser visitado neste cenário. Como pode-se observar na Figura 9, que quando assumi-se que γ_1,γ_2, γ_3 e γ_4 iniciam-se em zero a chance de selecionar o ajuste adequado diminui, enquanto quando iniciam-se em 1 a mediana da probabilidade é de 100%.

Para os outros métodos a probabilidade *a posteriori* do modelo independe do valor da inicialização da indicadora e para a análise foi considerado os valores iniciais indicados no 4º caso.

A Figura 10 mostra o tempo de ajuste dos modelos para cada método em todos os casos.

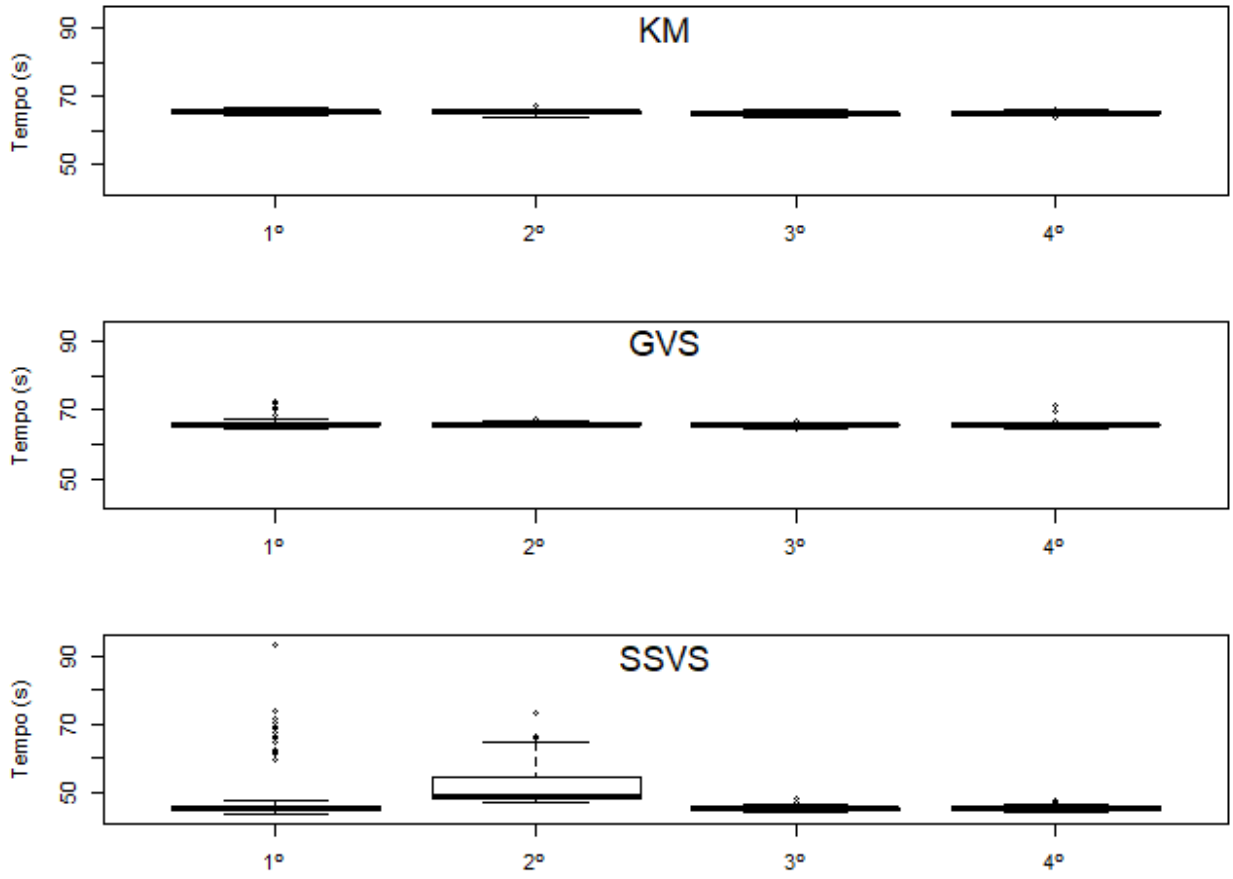

Figura 10: Cenário Dependente - Box-plot do Tempo por segundo

Ao comparar os tempos entre os ajustes para cada método presente na Figura 10, percebe-se que existe uma variação no tempo com relação aos valores iniciais utilizados, não muito significativos na maioria dos casos. Nota-se que o método de seleção de variáveis via busca estocástica é o mais rápido dentre os métodos comparados e o que apresentou maior oscilação em relação ao valor inicial de γ_j.

5 Conclusão

O objetivo deste trabalho foi apresentar e comparar por meio de um estudo de simulação três métodos de seleção de variáveis com enfoque bayesiano para modelos de regressão linear múltiplo: método de seleção de variáveis bayesianas de Kuo & Mallick, seleção de variáveis de Gibbs e seleção de variáveis via busca estocástica.

Os métodos bayesianos de seleção de variáveis tornam-se interessantes quando se tem um número exorbitante de variáveis explicativas, pois eles podem ser mais práticos e menos trabalhosos que os métodos clássicos de seleção de variáveis. A escolha do modelo é feita por meio de uma medida com uma interpretabilidade fácil e acessível, a probabilidade de escolha do modelo correto.

O diferencial dos métodos de seleção de variáveis bayesianos reside no fato de se realizar somente um ajuste e por meio de variáveis indicadoras apontar se a variável foi ou não selecionada. O que torna possível quantificar a porcentagem de vezes que cada variável é selecionada e mais do que isso, a porcentagem de vezes que o cada modelo é visitado, indicando o modelo que deve ser considerado e sendo mais prático que os métodos clássicos, em que é necessário ajustar todos os modelos possíveis e calcular uma medida de qualidade de ajuste para cada modelo e compará-los, a fim de descobrir o modelo mais adequado.

Com intuito de melhor compreender e comparar os métodos aqui estudados foi realizado um estudo de simulação para dois cenários, independente e dependente em modelos de regressão linear múltiplo, a fim de analisar o comportamento destes, além de se empregar valores diferentes para a inicialização das variáveis indicadoras.

No estudo realizado, observou-se que os 3 métodos apresentaram excelentes resultados, pois obtiveram altas probabilidades para o modelo correto. No cenário independente os diferentes valores adotados para inicialização das variáveis indicadoras não influenciaram de nenhuma maneira os modelos ajustados.

Logo a probabilidade *a posteriori* de se selecionar o modelo correto se manteve muito

próxima para os 4 casos de inicialização. Já para o tempo de iteração, existe uma variação mais significativa principalmente para o método de seleção de Kuo & Mallick. Olhando para os métodos percebe-se que o método de seleção de variáveis via busca estocástica apresentou uma probabilidade um pouco maior para o modelo correto e além disso levou menos tempo para fazer o ajuste do modelo, sendo considerado o melhor método neste cenário.

No cenário dependente notou-se que a probabilidade do modelo correto ser visitado diminui para o KM e o GVS e aumenta para o SSVS, com isso o método de seleção de variáveis via busca estocástica também é considerado o melhor método analisando a probabilidade do modelo escolhido. Entretanto, observou-se que somente o método de seleção de variáveis via busca estocástica sofreu influência decorrente da inicialização da variável indicadora, diminuindo bruscamente a probabilidade quando inicia-se as 4 primeiras covariáveis em zero, uma maior investigação sobre a influência dos valores iniciais neste método deveria ser realizada. Com relação ao tempo o método de seleção de variáveis via busca estocástica continuou sendo o mais rápido dentre os três métodos apesar de ter sofrido um leve aumento.

Após observar os dois cenários, tem-se que, em ambos, o método de seleção de variáveis via busca estocástica é considerado o melhor, por ser o mais rápido e possuir maiores probabilidades de escolher o modelo correto, especialmente se as variáveis indicadoras são inicializados em 1. Logo, o método de seleção de variáveis via busca estocástica será o método mais indicado em ambos os casos, lembrando que para o cenário em que há multicolinearidade deve-se atribuir o valor inicial 1 para a variável indicadora.

Como trabalhos futuros seria interessante investigar o impacto do número de variáveis no modelo e um possível impacto do tamanho de amostra. Ainda, poderia-se avaliar para o método de seleção de variáveis de Gibbs o uso dos valores de $\tilde{\mu}_j$ e S_j provenientes do ajuste do modelo completo.

Referências

[1] NETER, J. et al. *Applied linear statistical models*. [S.l.]: Irwin Chicago, 1996.

[2] SRIVASTAVA, A. K.; SRIVASTAVA, V. K.; ULLAH, A. The coefficient of determination and its adjusted version in linear regression models. *Econometric reviews*, Taylor & Francis, v. 14, n. 2, p. 229–240, 1995.

[3] AKAIKE, H. An information criterion (aic). *Math Sci*, v. 14, n. 153, p. 5–9, 1976.

[4] SCHWARZ, G. et al. Estimating the dimension of a model. *The annals of statistics*, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978.

[5] CELEUX, G. et al. Deviance information criteria for missing data models. *Bayesian analysis*, International Society for Bayesian Analysis, v. 1, n. 4, p. 651–673, 2006.

[6] KUO, L.; MALLICK, B. Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, JSTOR, p. 65–81, 1998.

[7] O'HARA, R. B.; SILLANPÄÄ, M. J. et al. A review of bayesian variable selection methods: what, how and which. *Bayesian analysis*, International Society for Bayesian Analysis, v. 4, n. 1, p. 85–117, 2009.

[8] LUNN, D. et al. The bugs project: evolution, critique and future directions. *Statistics in medicine*, Wiley Online Library, v. 28, n. 25, p. 3049–3067, 2009.

[9] MIGON, H.; GAMERMAN, D. Statistical inference: An integrated approach, arnold. *London, UK*, 1999.

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

[11] 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.

[12] HASTINGS, W. K. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, Biometrika Trust, v. 57, n. 1, p. 97–109, 1970.

[13] 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.

[14] GELFAND, A. E.; SMITH, A. F. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, Taylor & Francis Group, v. 85, n. 410, p. 398–409, 1990.

[15] MITCHELL, T. J.; BEAUCHAMP, J. J. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, Taylor & Francis, v. 83, n. 404, p. 1023–1032, 1988.

[16] GEORGE, E.; MCCULLOCH, R. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, v. 88, n. 423, p. 881–889, 1993.

[17] CARLIN, B. P.; CHIB, S. Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 473–484, 1995.

[18] DELLAPORTAS, P.; FORSTER, J. J.; NTZOUFRAS, I. On bayesian model and variable selection using mcmc. 1998.

[19] BROWN, P. J.; VANNUCCI, M.; FEARN, T. Multivariate bayesian variable selection and prediction. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, [Royal Statistical Society, Wiley], v. 60, n. 3, p. 627–641, 1998. ISSN 13697412, 14679868.